CARDIOVASCULAR DISEASE RISK PREDICTION USING DATA MINING TECHNIQUES

BY

ISTYAK AHAMED
ID: 161-15-6984
AND

SHOHANUR HOSSAIN
ID: 161-15-7444

This Report Presented in Partial Fulfilment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised by

Saiful Islam
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH

DECEMBER 2019

APPROVAL

This Project titled "**Cardiovascular Disease Risk Prediction Using Data Mining Techniques,**", submitted by Istyak Ahamed, ID: 161-15-6984, Shohanur Hossain, ID: 161-15-7444 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on December 5, 2019.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Nazmun Nessa Moon — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Gazi Zahirul Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare, this project has been done under the supervision of **Saiful Islam, Senior Lecturer, Department of CSE,** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised By:

Saiful Islam
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Shah Md. Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University

Submitted By:
Istyak Ahamed

Istyak Ahamed
ID: 161-15-6984
Department of Computer Science and Engineering
Daffodil International University
Shohanur Hossain

Shohanur Hossain
ID: 161-15-7444
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for his divine blessing makes us possible to complete the final year thesis successfully.

We really grateful and wish our profound our indebtedness to **Saiful Islam, Senior Lecturer,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining and Machine Learning*" to carry out this thesis. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Prof. Dr. Syed Akhter Hossain, Professor and Head,** Department of CSE, for his kind help to finish our thesis and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mates in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

At present, cardiovascular disease has become the leading cause of death worldwide. Particularly in the South Asian countries have a tremendous risk of cardiovascular disease at an early age than any other ethnic group. Most often it's challenging for medical practitioners to predict cardiovascular disease as it requires experience and knowledge which is a complex task to accomplish. This health industry has enormous amounts of data which is useful for making effective conclusions using their hidden information. Using appropriate results and making effective decisions on data, some superior data mining techniques are used such as Logistic Regression, Decision Tree, Nave Bayes , SVM. By using some properties like (age, gender, bp, stress etc) we can be predicted the chances of cardiovascular disease.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE**

LIST OF TABLES

TABLES **PAGE**

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cardiovascular disease is the number one cause of death globally, more people die annually from cardiovascular diseases than from any other cause. It accounts for nearly one in every three deaths worldwide annually. Accounting for 15.5% of all deaths worldwide cardiovascular disease is the world's leading cause of death.

At present growing concern for Bangladesh is a cardiovascular disease with patients suffering it topping the list of people with non-communicable diseases. According to the National Health Bulletin, the top cause of hospital admission, morbidity, and mortality in the country is also cardiovascular disease. “Cardiovascular disease became the leading cause of death among the non-communicable diseases in Bangladesh,” said the World Health Organization in its latest Non-Communicable Diseases Country Profiles 2018.

According to (WHO) [7], deaths caused by cardiovascular disease increased manifold in Bangladesh over recent years. Cardiovascular identification could be a sophisticated and necessary task that must be dead accurately and expeditiously. Supported a doctor’s expertise & information the identification is usually created. This results in unwanted results & excessive medical prices of treatments provided to patients. Therefore, a computerized medical diagnosis system would be extremely propitious.

But using the data mining technique we can explore the hidden patterns in the medical domain. These patterns can be utilized for clinical diagnosis. An online-based prediction system can make these diagnoses easy and affordable especially for the financially insecure people.

This paper intends to research the various prognostic descriptive data processing techniques introduced in recent years to predict the possibilities of upset.

1.2 Motivation: A Silent Story of Bangladesh

Three out of four individuals in Bangladesh run the chance of developing viscus diseases [20]. Cardiovascular disease is the leading reason for vas and urinary organ diseases and over one. Five billion are expected to be stricken by cardiovascular disease by 2025; the case being particularly adverse in South Asia only if the prevalence of cardiovascular disease is already at 40 %.

High-level officers from DGHS and policymakers in Asian countries showed their enthusiasm for the elapid snake strategy because it is aligned with the operational set-up of the fourth Health, Population, and Nutrition Sector Program (2017-2022) [20].

In Bangladesh, vast population living under the poverty level does not have sufficient access to the desired medical care.

The public medical sector that completely run by the national fund does not have the capability or the proper medical resources to include this large amount of financially unprivileged people to the medical sector.

As a result, the chances of cardiovascular disease are not diagnosed with these people.

Again cardiovascular disease often does not show any symptoms, as a result, these chronic phase transitions turn quickly into the end-stage and require diagnosis (ECG, ECHO, etc) which is unaffordable to many people or quickly becomes unaffordable after few months into the treatment.

In most cases, patients develop cardiovascular disease without recognizing they even had a disease in the first place. The lack of governmental funding, a dull process in healthcare and overall disappointing service in public health institutes had forced the general public to seek private healthcare. This consequently has increased the private sector but at the price of high medical costs. So a huge amount of people affected by cardiovascular disease for lack of proper treatment and diagnosis which often leave the patient and the family in unbearable sufferings and humanitarian crises

The first step towards treatment is to get diagnosed first. The advancement of medical technology and the capacity of storing medical data in digital form has paved the idea for medical automation. An automated system helps to predict the chances of cardiovascular disease can make things easy to be careful.

1.3 Can Data Mining help?

At present technology is moving towards AI automation, computerized mechanization. Raw data are being produced in real-time. This data allows us to analyze new data mining techniques.

For that maintenance of patient's data is becoming more and more common as far as the medical data is treated more computerized. Using the data mining and machine learning can describe the potential solution of those problems instead of conventional statistics which may give us answers like "how" and "why". At present data mining is being practiced frequently for analysis. By using these techniques hidden patterns can also be explained.

In Bangladesh, cardiovascular disease chance prediction can be classified using appropriate data. The public sector deals with a huge number of the patient simultaneously. That's why an automated system can give a pre-suggestive result in which a patient can be wide-awake or be cautious. This testing platform can be implemented for a certain period during which further data will be collected. Custom modification of the model can be done. Further training of the model using real-time and actual clinical data will be a leap forward to knowing how much can we rely on an automated system. There is a good chance to predict more diseases using these.

1.4 Rationale For the study

Here are some of the reasons and arguments in favor of the study on the chances of cardiovascular disease.

- Researchers always fascinated by the potential of an automated system for the classification of disease.
- This automated system has the potential to get implemented, although the clinical acceptance for this has not got too much attention.
- Cardiovascular disease can be deceptive in nature as the symptoms come often at a very late stage.
- An automated system can constantly check the risk percentage of patients in general.
- Often patients get tested for numerous other medical conditions. But this system can help to erase the chances of unnecessary test.

- The unusual number of people that seek medical services in the public medical hospitals in Bangladesh has to face endless pain to go through the process. An automated system can certainly save those people a lot of money and time. People can filter out only those who are genuinely at risk of the disease. This will save time, resources and money.

  Bangladesh has a massive people suffering from cardiovascular disease.

1.5 Research Question

In this research we tried to accomplish the following:

- Can this project be classified with a Convincing level of clinical accuracy?
- Applying different classifiers to the dataset.
- Enlisting and comparing the accuracy rate of each classifier.
- As the dataset is quite extensive we will work with various sub set of the dataset.

1.6 Expected Outcome

After completing this research we hope to find these outcomes:

- By analyzing the dataset the patient should be classified with comprehensive accuracy.
- Comparing different Machine Learning classification methods.
- Finding out the correlation of different attributes in the dataset.

1.7 Report Layout

We have discussed our objective and what our expected outcome was.

- In the first chapter, we overviewed the project with our motivation. We also discussed our objective and what our expected outcome was.
- In the second chapter, we have extensively discussed our background study on cardiovascular disease and literature survey.
- The third chapter is about the research methodology. The classifier algorithms are also discussed here.
- The fourth chapter is about a detailed description of our experimental results and comparative studies of the classifier accuracies.

- The fifth chapter is about the research summary along with the future goal of the study with detailed discussion about more areas for study in a duplicate field.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

Here we will review the cardiovascular disease and its definition. Also, we will try to explore the risks and effects of cardiovascular disease. We tried to discuss the literature survey in related fields.

2.2 Cardiovascular Disease

In this section, we will discuss some description of cardiovascular disease.

2.2.1 Definition

Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.

2.2.2 Causes and Risk Factors

While the disease can confer with entirely completely different heart or vessel problems, the term is sometimes accustomed mean hurt to your heart or blood vessels by induration of the arteries, a buildup of fatty plaques in your tracks. Plaque increase thickens and stiffens artery walls, which could inhibit blood run through your arteries to your organs and tissues. Cardiovascular disease is to boot the foremost common reason for disorder. It's caused by correctable problems, like associate unhealthy diet, being overweight lack of exercise and smoking.

There are various risk factors for the disorder. Some you will be ready to predict, others you can’t. Some preservation can control this disease if we take this early stage.

Ones that can’t be controlled include:

- Gender (males square measure at larger risk)
- Age
- A family history
- Being post-menopausal

Still, creating some changes in your modus vivendi will scale back your probability of getting cardiopathy. manageable risk factors include

- Smoking
- High LDL, or "bad" cholesterol, and low high-density lipoprotein, or "good" cholesterol
- Uncontrolled cardiovascular disease (high blood pressure)
- Physical inactivity
- Obesity
- Uncontrolled polygenic disorder
- Uncontrolled stress and anger

The following figure 2.2.2 shows plaque on blood vessel

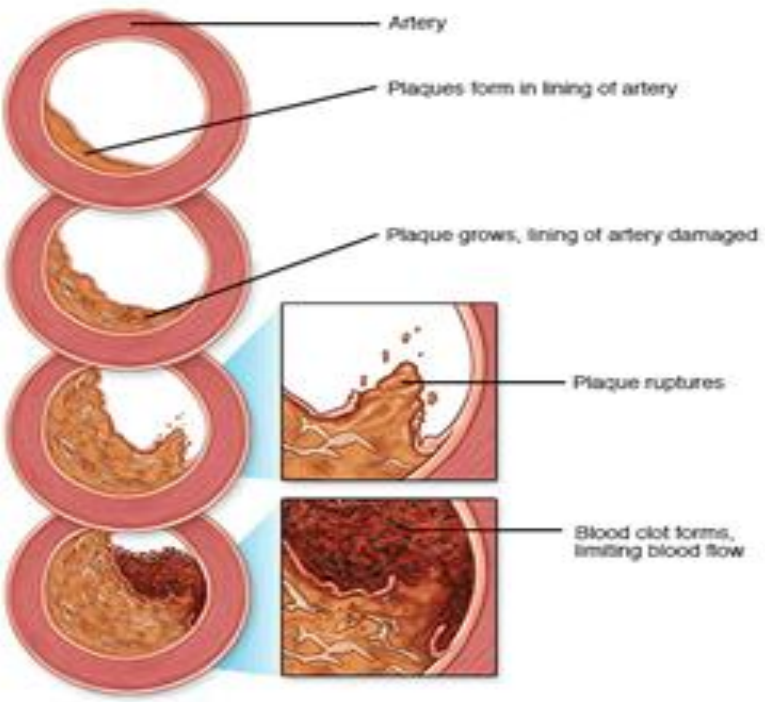


Figure 2.2.2: Plaque on blood vessel

2.2.3 The Effects in Human Body

- With less blood flow, your heart doesn't get the element it desires, which will cause hurting, known as angina, particularly after you exercise or do significant labor.
- It can also affect however well your heart pumps and build the remainder of your body short on the element, too.
- Without it, your cells won't work still as they ought to, and you will require breath or feel tired than usual.
- If the plaque breaks off associated whole blocks an artery, you've got an attack.

2.3 Comparative studies and Related works

Though the medical data is unstructured and its nature is geographically diverse that's why it is difficult to automate the classification of the disease that much complicated.

This data has racial, geographical and cultural partiality. Often clinical symptoms for the disease vary over different regions.

Despite these difficulties, there have been multiple studies in the field which attempted to classify cardiovascular disease with a range of data mining techniques. Many of these have been breaking new grounds and has brought new ideas to classify cardiovascular disease. Some of those are reviewed below.

Pattern recognition and data mining methods in predicting models in the domain of cardiovascular diagnoses are used by the researchers [4]. The research uses classification algorithms like Naïve Bayes, Decision Tree, K-NN and Neural Network and results showed that Naïve Bayes technique executed well then other used techniques [4].

The researchers [5] also suggested a new approach for association rule mining based on the sequence number and clustering transactional data set for cardiovascular disease predictions. the implementation was in the C programming language and reduced main memory requirement by considering a small cluster at a time to be scalable and efficient [5].

To predict the circumstances of coronary heart disease simulated in MATLAB tool the researchers [6] proposed a layered neuro-fuzzy strategy. In performing analysis for coronary

heart disease happens the neuro-fuzzy integrated approach produced an error rate very low and a high work efficiency [6].

The researchers uses a heart disease warehouse extract data relevant to heart disease by K-means clustering algorithm, and to calculate the weight of the frequent patterns significant to heart attack predictions the MAFIA (Maximal Frequent Itemset Algorithm) algorithm is used [9].

For predicting and analyzing heart disease from the dataset the researchers used the data mining algorithms decision trees, Naïve Bayes, neural networks, association classification, and genetic algorithm [10].

The researchers experimented on a dataset produced a model using neural networks and hybrid intelligent algorithm and the results show improved accuracy of the prediction using the hybrid intelligent technique improved accuracy of the prediction [11].

This research paper describes a prototype using Naïve Bayes and a weighted associative classifier (WAC) to predict the probability of patients receiving heart attacks [12].

A web-based intelligent system is developed by the researchers using naïve bayes algorithm to answer complex queries for diagnosing cardiovascular disease and help medical practitioners with clinical decisions [13].

Decision trees, naïve bayes, and neural networks are used by the researchers to predict heart disease with 15 popular attributes as risk factors listed in the medical literature [16].

Association rules using feature subset selection is created to predict a model for heart disease. Classification predicts the class in the patient dataset where else Association rule determines relations amongst attributes values [18].

The researchers uses global optimization interest of genetic algorithm for the initialization of neural network weights implementing a hybrid system [19].

2.4 Research Summary

As the previous study and literature survey demonstrates that there have been numerous studies in this field. The studies have been fairly successful in their way. This type of computer-aided classification problem has been studied with many other diseases. From studying different algorithms to making re-optimization to the existing algorithm to find better results, researchers have gone through many different ways. The noticeable factor is that although the accuracy has been quite good, yet we have not seen any real implementation of these processes. Probably the idea of consulting a computerized diagnosis system for a disease isn't as convincing as consulting a doctor for the public. But with more accuracy and some experimental periods, a fully automated diagnosis probably would be as normal as consulting a doctor.

2.5 The scope of this problem

The scope of this problem is to classify our dataset using different machine learning algorithms which include training and testing sets. We will try to explore the similarity between the dataset attributes to find out their dependency on each other for the development of cardiovascular disease prediction.

In Bangladesh, an automated diagnosis system could reduce the lengthy process of health care. With improved symptoms analyzing algorithms, the system can suggest diagnostic tests to the users hence reducing time and cost in big hospitals.

2.6 Challenges

The initial challenge for this thesis is to collect data on cardiovascular disease in Bangladesh. The dataset we used is well pre-processed. In contrast, finding this type of dataset is quite difficult in Bangladesh. The same patient's data are not kept in a structured way. The data is in incomplete form. So with this kind of data results in insufficient or biased training which will result in lower accuracy.

Quite often the patient affected in cardiovascular disease came to know at a very late stage and so the model is designed with the data for which most of the patients can check initially without any medical tests.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Data Mining is a technique where large volume pre-existing immature data in a database is processed, or modified to requirements and analyzed to reveal useful patterns and new relations among attributes for achieving various goals. Data mining is also called knowledge discovery in databases known as KDD.

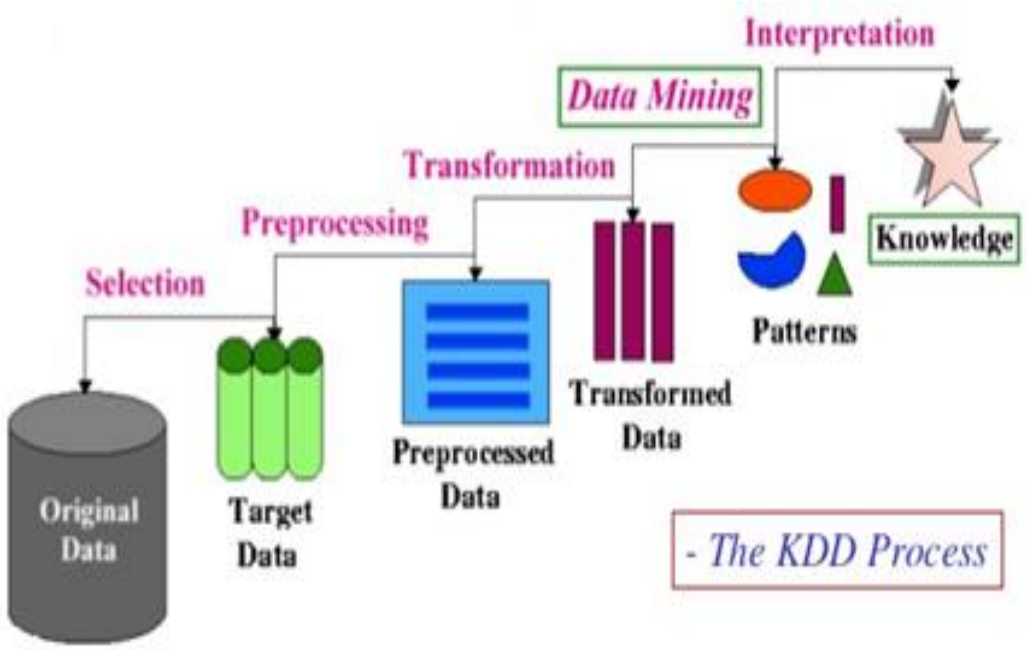


Figure 3.1.1: Steps associated with KDD

Traditional data analysis techniques through a statistical approach has been used for a long time. This approach has been very useful and will be used in the foreseeable future. Accumulating and preserving various transactional and other types of data became more convenient as the storage capacity of modern computers increased.

Inevitably the size and diversity of the data grew larger and traditional data analysis techniques began to be less effective and inefficient for such large amounts of data warehouses traditional data analysis techniques began to be less effective and inefficient

and also for Inevitably the size and diversity of the data grew. So among large companies and researchers data mining and machine learning techniques gained popularity

The data mining and machine learning has became a “Gold Mine” for extracting new patterns and useful knowledge in medical advancement due to latest explosion of Medical data through machine automation and use of computerized technology in diagnosis and treatment of disease.

Although the acceptance of automated classification of disease is still not popular and desirable among the medical community, it is still a research area of the enormous potential for data scientists and researchers around the globe.

So we will try to explore this concept of data mining to help automation of the classification of Cardiovascular disease chance prediction. This is one of the best way to find the chances. By the help of this technology many data could be mined and the accuracy will be better for this. At present we have many tools to use and for that it’s easy to predict the future by considering some attributes and their values.

3.2 About Dataset

We have collected data using google from. We spread the from among every group of people through the internet and got more than 350 data. The dataset has 12 attributes including sex, age, bc, bp, hereditary, smoking, alcohol, diabetes, exercise, diet, obesity, stress. These twelve attributes are used here. We have used about 301 data from different age group people and the data is well decorated.

3.3 Data Description and Preprocessing

In table 3.3.1 we have listed the attributes in the data set with their encodings.

Table 3.3.1: Attributes in the data and encodings

Attributes	Attribute Description	Encodings
sex	SEX	Male (1), Female (0)
age	AGE(YEARS)	20-34 (-2), 35-50 (-1), 51-60 (0), 61-79 (1), >79 (2)
bc	BLOOD CHOLESTROL	Below 200 mg/dL - Low (-1) 200-239 mg/dL - Normal (0) 240 mg/dL and above - High (1)
bp	BLOOD PRESSURE	Below 120 mm Hg- Low (-1) 120 to 139 mm Hg- Normal (0) Above 139 mm Hg- High (-1)
hereditary	HEREDITARY	Family Member diagnosed with HD -Yes (1) Otherwise –No (0)
smoking	SMOKING	Yes (1) or No (0)
alcohol	ALCOHOL INTAKE	Yes (1) or No (0)
exercise	PHYSICAL ACTIVITIES	Low (-1), Normal (0) or High (-1)
diabetes	DIABETES	Yes (1) or No (0)
diet	DIET	Poor (-1), Normal (0) or Good (1)
obesity	OBESITY	Yes (1) or No (0)
stress	STRESS	Yes (1) or No (0)

The dataset is quite extensive and detailed in its patient data with some of the preprocessing already done. We did some preprocessing which we will be discussing in the next section.

3.4 Screenshot of the dataset

In figure 3.4.1 we can see the dataset along with its mean max value and standard deviation.

```
df.describe()
```

	age	sex	bc	bp	hereditary	smoking	alcohol	exercise	diabetes	diet	obesity	stress	3
count	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	
mean	43.684385	0.534884	214.754153	133.172757	0.209302	0.372093	0.521595	0.043189	0.445183	-0.189369	0.425249	0.481728	
std	14.028901	0.499612	34.353156	16.777268	0.407488	0.484168	0.500365	0.684199	0.504465	0.721586	0.495204	0.500498	
min	21.000000	0.000000	105.000000	87.000000	0.000000	0.000000	0.000000	-1.000000	-1.000000	-1.000000	0.000000	0.000000	
25%	33.000000	0.000000	188.000000	124.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-1.000000	0.000000	0.000000	
50%	41.000000	1.000000	216.000000	131.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	54.000000	1.000000	239.000000	143.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	
max	79.000000	1.000000	290.000000	183.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

Figure 3.4.1: Dataset

3.5 Classification Algorithms

In ML classification the aim is to predict the target class by analyzing the training dataset by Finding the actual boundaries from every target class. By using the training dataset to get better boundary states which could be used to determine each target class. Whenever the boundary is determined, the next task is to predict the target class. And this process is called classification. Here to predict the class level we had used some classification methods.

3.5.1 Logistic Regression

There are some machine learning algorithms from the area of statistics. Also known as the go-to technique for classification problems in machine learning. Logistic Regression and Linear Regression are a little bit similar because both have the goal of estimating the values for the parameters or coefficients. After train a model in machine learning we find out the relation between training and testing data.

3.5.2 Decision Tree

It is a popular classifier that is easy and simple to implement. It can handle high dimensional data with no domain knowledge or parameter setting. It's easy to read and interpret the results obtained from Decision Trees. The drill-through feature to access detailed patients" profiles is only available in Decision Trees.

3.5.3 Naive Bayes

Naive Bayes is a classifier or a machine learning algorithm which uses the Bayes theorem with independent assumptions between features. One dimensional Naive Bayes classifier computes the ratio of the log probabilities of the features belonging in all the classes. Naive Bayes does not consider the correlation between attributes. Naive Bayes is a very scalable classifier but it can create bias towards one or more attributes which often result inaccuracy.

3.5.4 SVM

A Support Vector Machine (SVM) could be a discriminative classifier formally outlined by a separating hyper plane. In different words, given labelled coaching information (supervised learning), the algorithmic program outputs associate degree optimum hyper plane that categorizes new examples. In area, this hyper plane could be a line dividing a plane into two elements whereby every category lay on either aspect.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

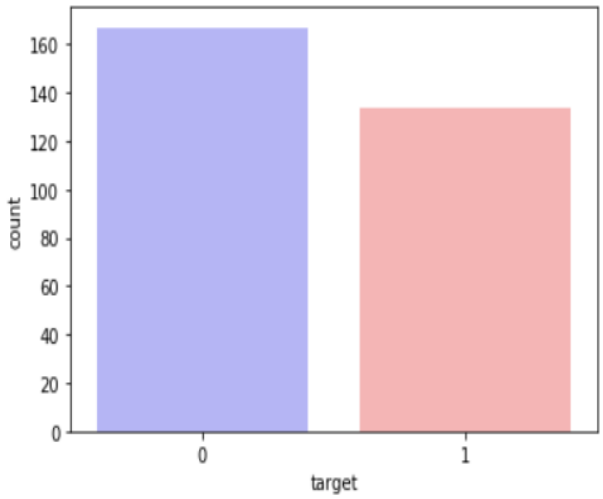
In this chapter, we will discuss the results of the conducted experiment. We will compare and explore different classifier performance and accuracy. We will show the result in a graph and also in tables.

4.2 Experimental Results

Following Section, we explore the data with different attributes. Then we came to know many facts about the data set.

4.2.1 Data Exploration

In figure 4.2.1 we will know about the percentages of heart disease patients.

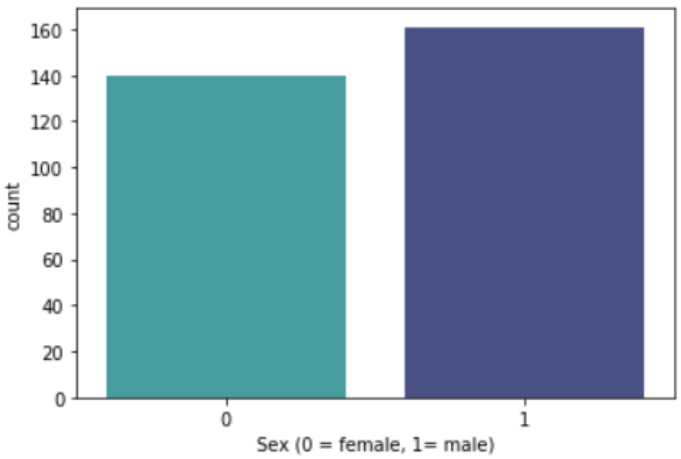



```
countNoDisease = len(df[df.target == 0])
countHaveDisease = len(df[df.target == 1])
print("Percentage of Patients Haven't Heart Disease: {:.2f}%".format((countNoDisease / (len(df.target))*1
print("Percentage of Patients Have Heart Disease: {:.2f}%".format((countHaveDisease / (len(df.target))*10
```

```
Percentage of Patients Haven't Heart Disease: 55.48%
Percentage of Patients Have Heart Disease: 44.52%
```

Figure 4.2.1: Heart disease percentages

In figure 4.2.2 we can see that 46.51% female and 53.49% male heart patients are there.




```
countFemale = len(df[df.sex == 0])
countMale = len(df[df.sex == 1])
print("Percentage of Female Patients: {:.2f}%".format((countFemale / (len(df.sex))*100)))
print("Percentage of Male Patients: {:.2f}%".format((countMale / (len(df.sex))*100)))
```

```
Percentage of Female Patients: 46.51%
Percentage of Male Patients: 53.49%
```

Figure 4.2.2: Heart disease percentages

In figure 4.2.3 we see that at the age of 27 to 37 have the highest frequency

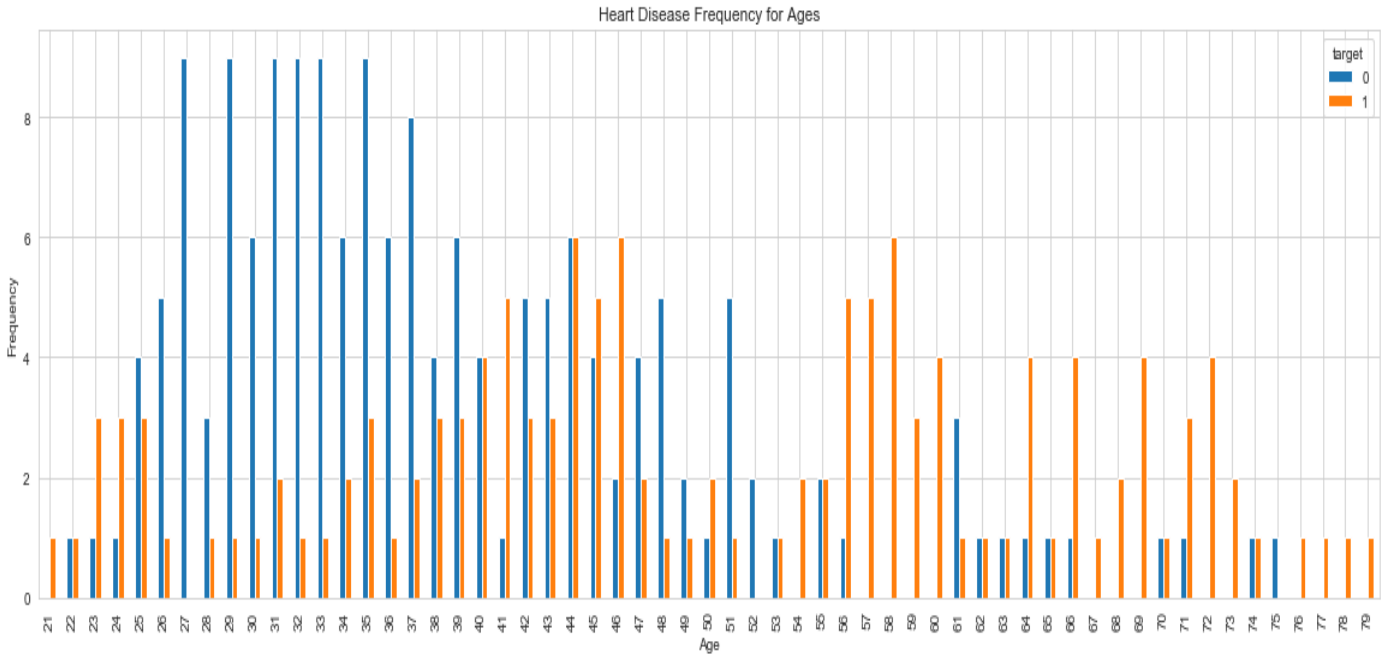


Figure 4.2.3: Heart disease frequency for ages

In figure 4.2.4 we see that in female 40 up was the highest frequency and 80 was for the men.

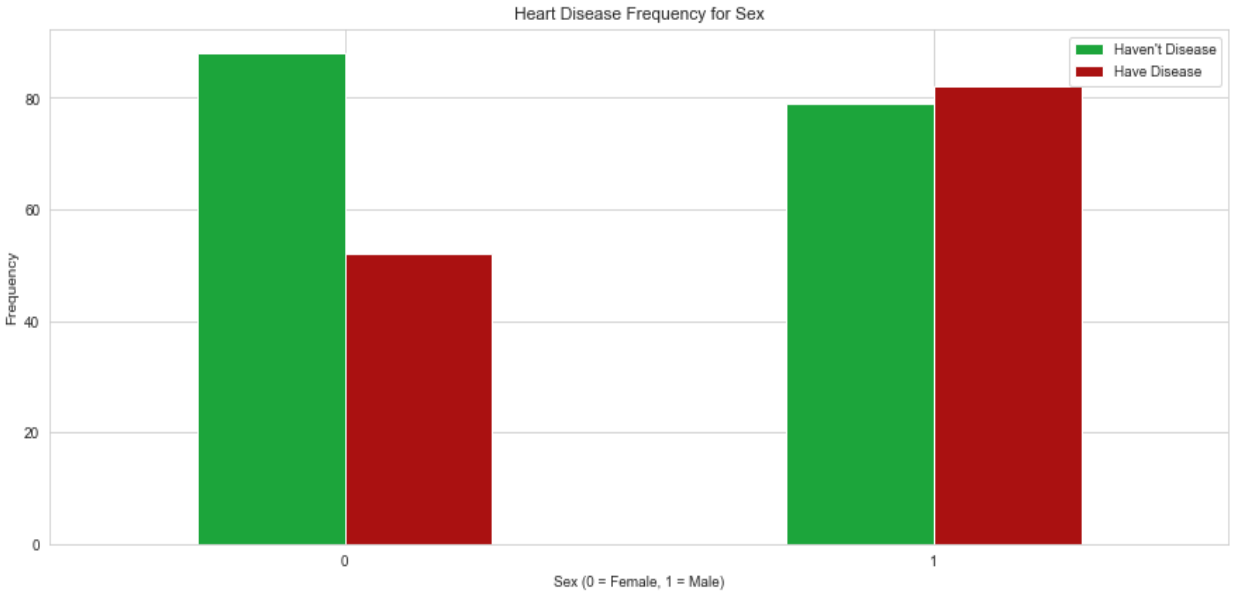


Figure 4.2.4: Heart disease frequency for sex

In figure 4.2.5 we see that patients who follow normal diet have the maximum frequency 70 and with high diet have the lower frequency.

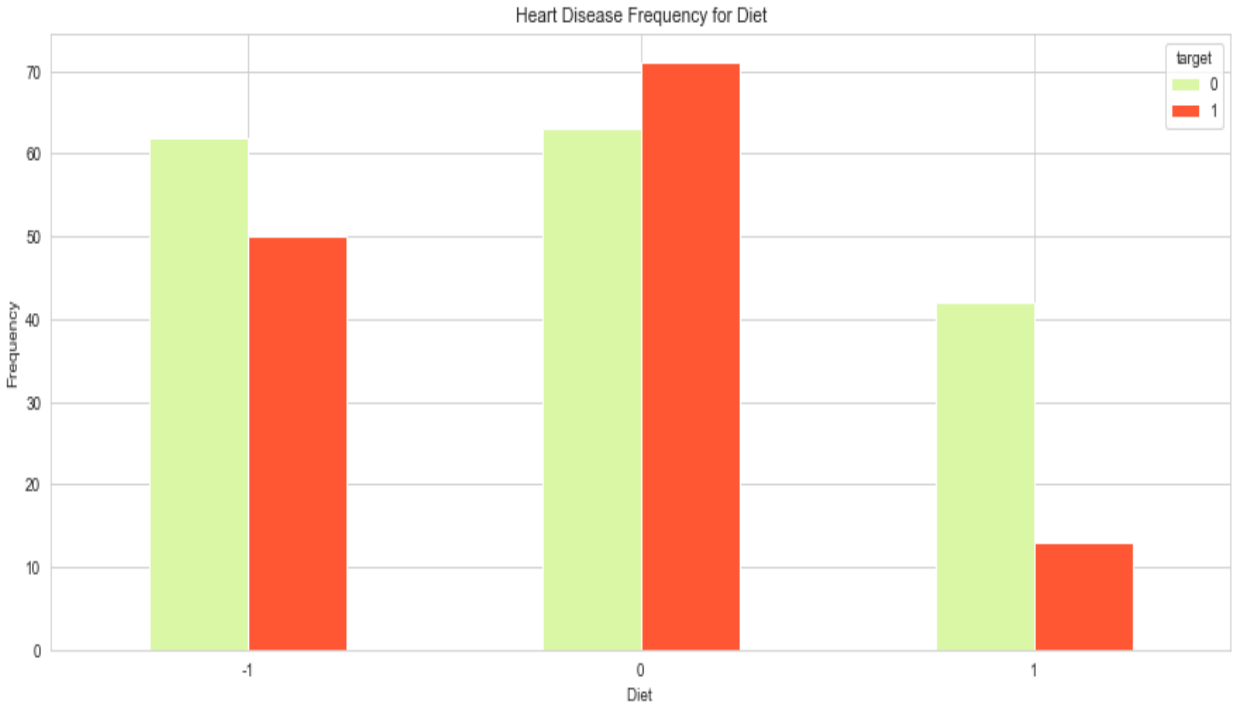


Figure 4.2.5: Heart disease frequency for diet.

In figure 4.2.6 we see that the people who follow normal exercise have the highest frequency and who follow low exercise have the lowest frequency.

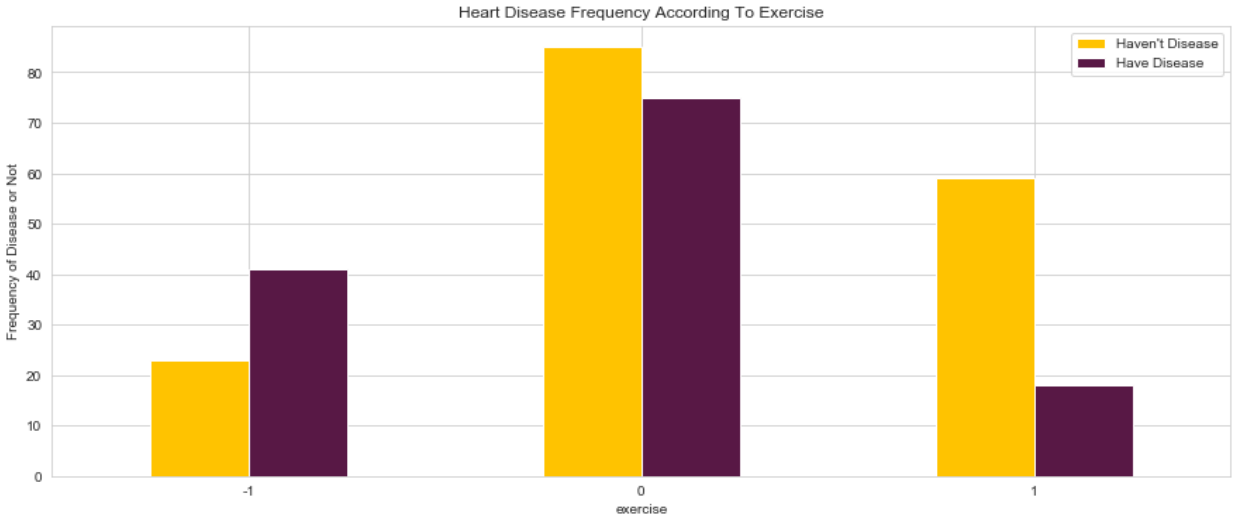


Figure 4.2.6: Heart disease frequency for exercise.

In figure 4.2.7 we see that the frequency according to the blood pressure.

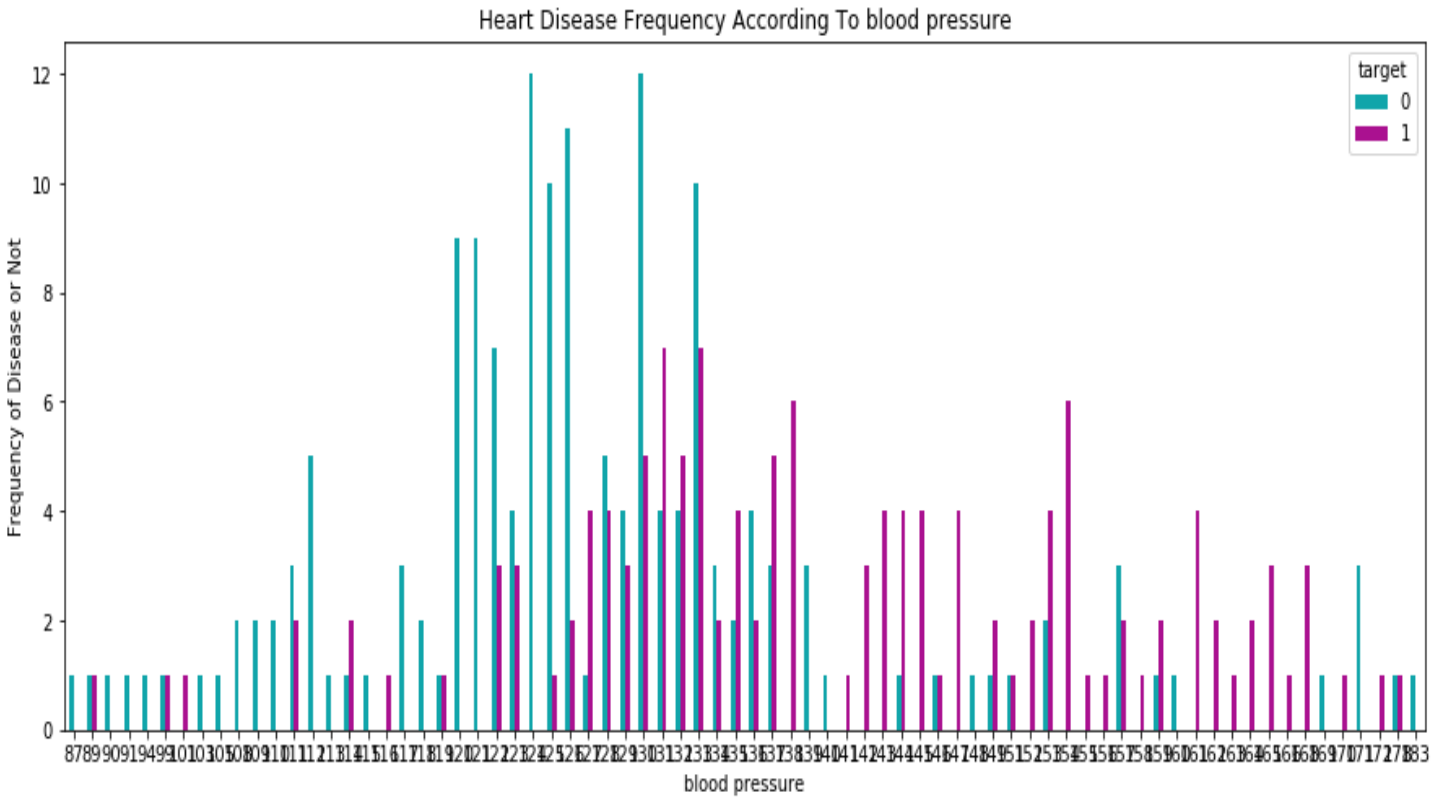


Figure 4.2.7: Heart disease frequency for blood pressure.

In figure 4.2.8 we made categorical variables into dummy variables.

	age	sex	bc	bp	hereditary	smoking	alcohol	exercise	diabetes	diet	obesity	stress	target	exercise_-1	exercise_0	exercise_1	diet_-1	diet_0
0	69	1	197	101	1	0	1	1	1	0	0	0	1	0	0	1	0	1
1	61	0	250	144	0	1	0	0	1	-1	1	1	0	0	1	0	1	0
2	35	0	182	123	0	0	1	0	0	1	0	0	1	0	1	0	0	0
3	29	1	170	131	0	0	1	1	0	0	0	0	1	0	0	1	0	1
4	38	0	192	123	1	0	1	1	1	0	0	1	1	0	0	1	0	1

```
df = df.drop(columns = ['exercise', 'diet' ])
df.head()
```

	age	sex	bc	bp	hereditary	smoking	alcohol	diabetes	obesity	stress	target	exercise_-1	exercise_0	exercise_1	diet_-1	diet_0	diet_1
0	69	1	197	101	1	0	1	1	0	0	1	0	0	1	0	1	0
1	61	0	250	144	0	1	0	1	1	1	0	0	1	0	1	0	0
2	35	0	182	123	0	0	1	0	0	0	1	0	1	0	0	0	1
3	29	1	170	131	0	0	1	0	0	0	1	0	0	1	0	1	0
4	38	0	192	123	1	0	1	1	0	1	1	0	0	1	0	1	0

Figure 4.2.8: Creating dummy variables.

4.3 Different Classification Algorithm

Here we will use different classification algorithm and try to understand which one gives the best accuracy. We have used six classification algorithms to check this accuracy.

4.3.1 Logistic Regression with Confusion Matrix

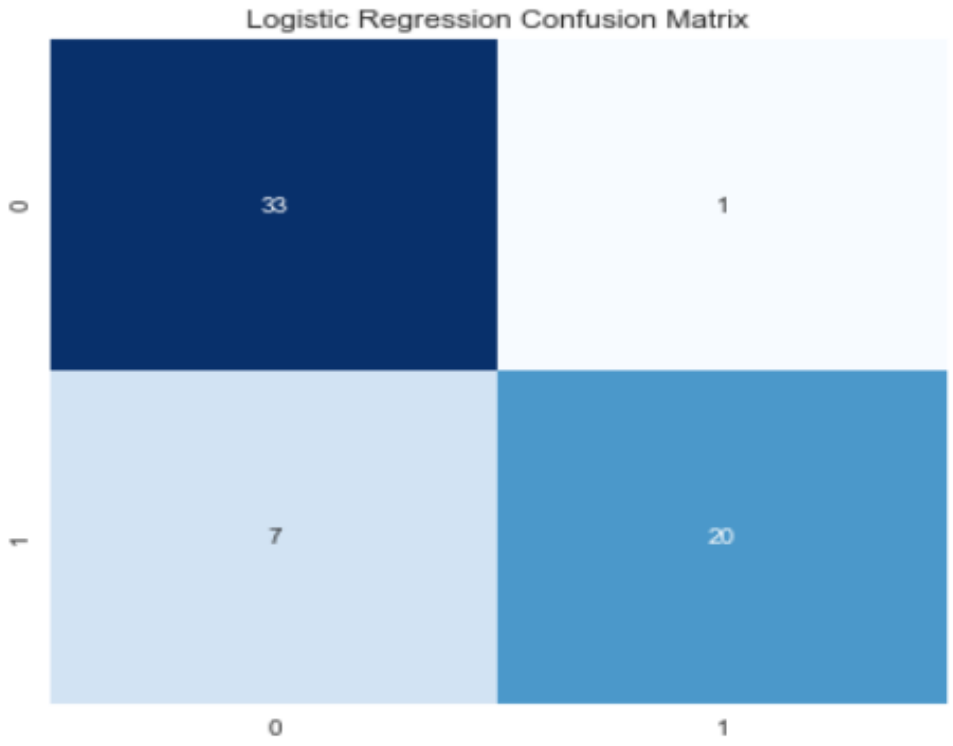
In figure 4.3.1 we have used logistic regression and the accuracy was 86.89%.

```
lr = LogisticRegression()
lr.fit(x_train.T,y_train.T)
print("Test Accuracy {:.2f}%".format(lr.score(x_test.T,y_test.T)*100))
```

```
Test Accuracy 86.89%
```

Figure 4.3.1: Logistic Regression

Table 4.3.1: Confusion Matrix of Logistic Regression



4.3.2 Naive Bayes with Confusion Matrix

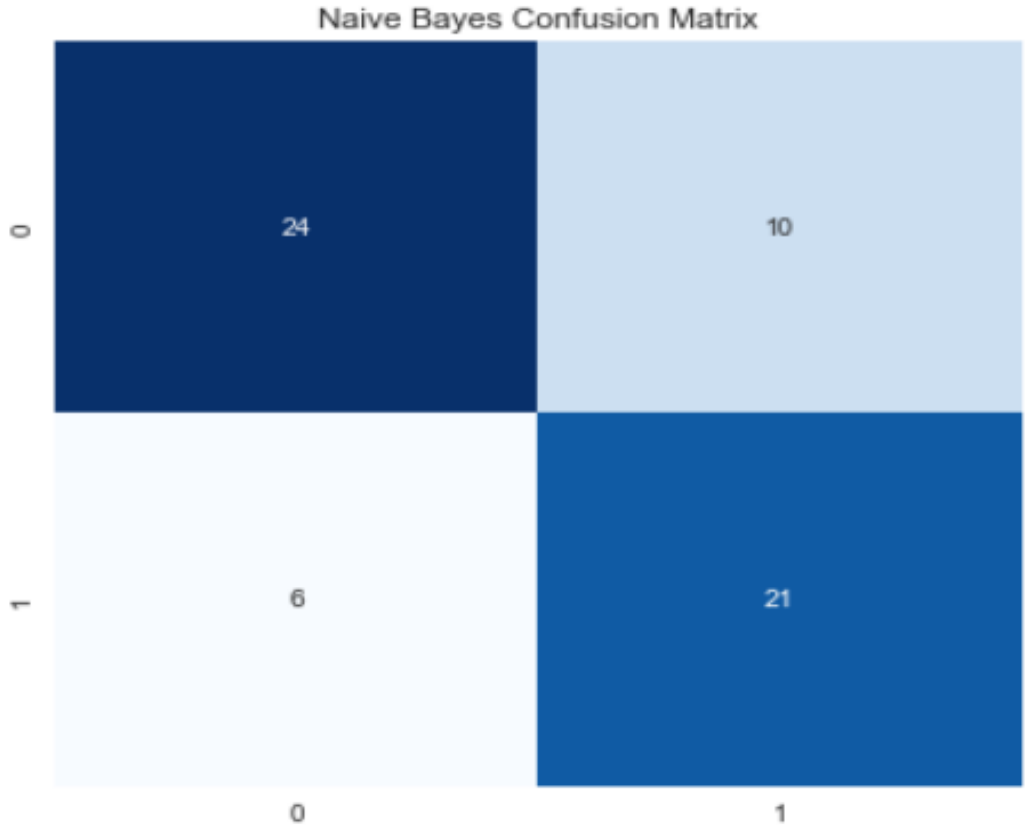
In figure 4.3.2 we have used Naive Bayes code and the accuracy is 73.77%.

```
from sklearn.naive_bayes import GaussianNB
nb = GaussianNB()
nb.fit(x_train.T, y_train.T)
print("Accuracy of Naive Bayes: {:.2f}%".format(nb.score(x_test.T,y_test.T)*100))
```

```
Accuracy of Naive Bayes: 73.77%
```

Figure 4.3.2: Naive Bayes

Table 4.3.2: Confusion Matrix of Naive Bayes



4.3.3 SVM with confusion matrix

In figure 4.3.3 we have used SVM code and the accuracy is 83.61%.

```
from sklearn.svm import SVC
```

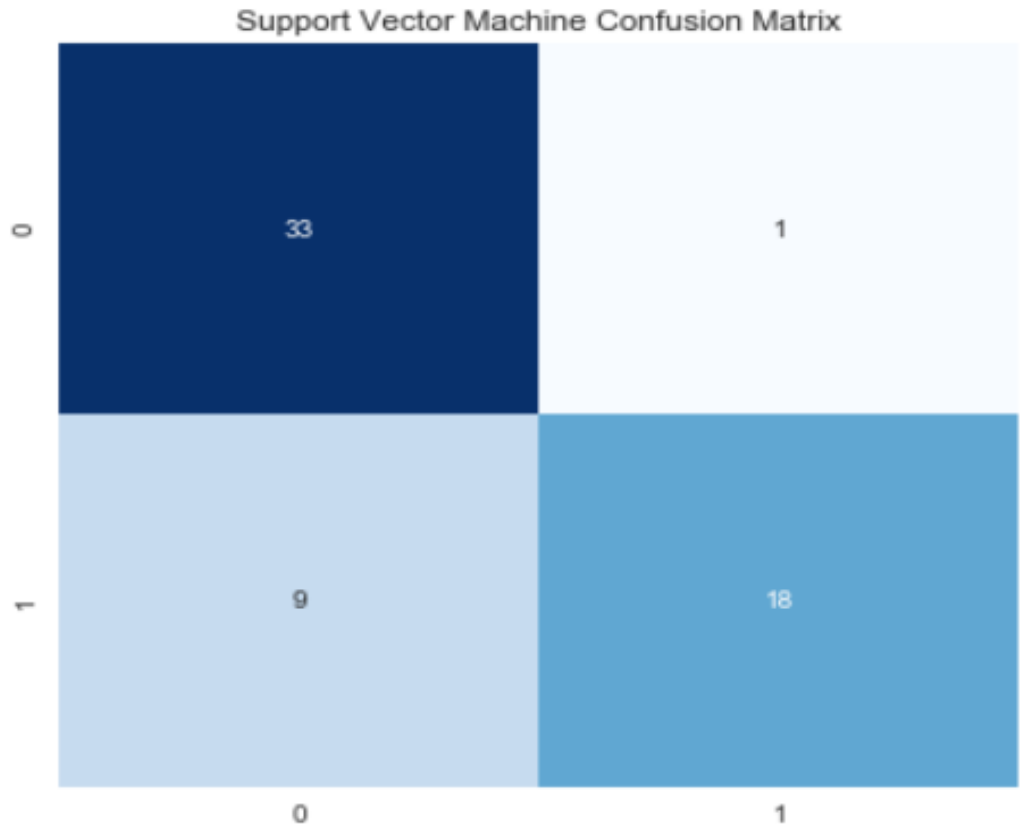
```
svm = SVC(random_state = 1)
svm.fit(x_train.T, y_train.T)
```

```
print("Test Accuracy of SVM Algorithm: {:.2f}%".format(svm.score(x_test.T,y_test.T)*100))
```

```
Test Accuracy of SVM Algorithm: 83.61%
```

Figure 4.3.3: SVM

Table 4.3.3: Confusion Matrix of SVM



4.3.4 Decision tree with confusion matrix

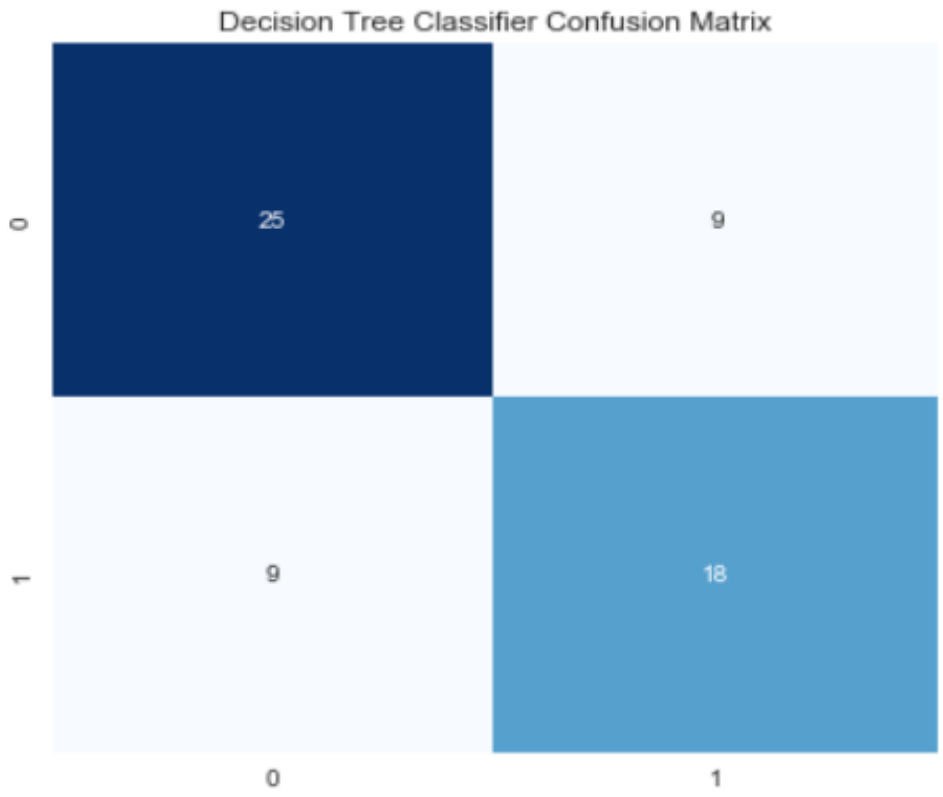
In figure 4.3.4 we have used decision tree code and the accuracy is 70.49%.

```
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()
dtc.fit(x_train.T, y_train.T)
print("Decision Tree Test Accuracy {:.2f}%".format(dtc.score(x_test.T, y_test.T)*100))
```

```
Decision Tree Test Accuracy 70.49%
```

Figure 4.3.4: Decision tree

Table 4.3.4: Confusion Matrix of decision tree



4.4 Potential Future Improvement

The study shows that an automated system can be implemented for predicting the chances of cardiovascular disease. With more efficient algorithms and training data, this can be a real-life implementation for clinical prediction in Bangladesh.

Bangladesh's medical system can gather more clinical data in a structured and organized way where each patient’s medical data can be acquired which can be later used in many different studies for other diseases and also consolidate with the existing system for automated classification and detection of those diseases.

4.5 Summary

In brief, the Logistic Regression had the most accurate detection of the disease which is 86.89%. Due to insufficient and partial data the result is pretty high numerically.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the study

We have used six algorithms the accuracy of all modes. The result is given below.

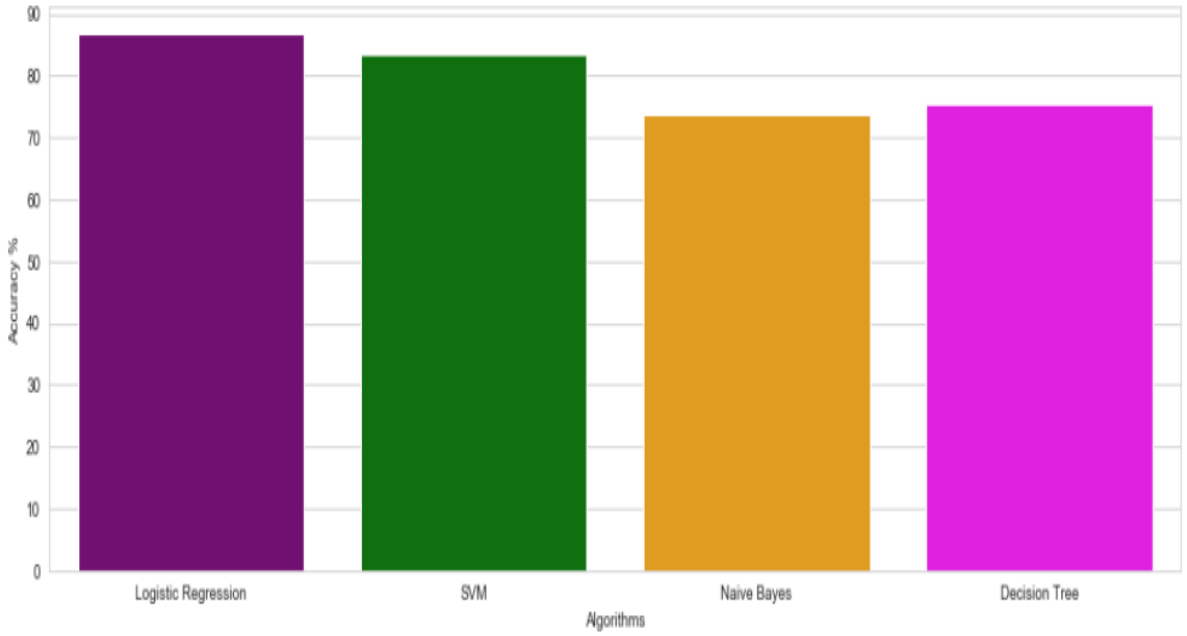


Figure 5.1.1: Accuracy of all classifiers

5.2 Conclusion

Hence to conclude it all, we have used six different classifiers among which the Logistic Regression classifier had the highest level of accuracy(86.89%). Although the other classifier also gave quite close and accurate results compared to Logistic Regression Classifier. So we can say that for measuring the risk of cardiovascular disease logistic regression will give the best accuracy using these kinds of data set.

5.3 Recommendation

This model can be used with acceptable accuracy for a Clinical testing period to find out its probability and sustainability in practical use. More clinical data needs to be accumulated with required data organization which should be used to train the model order for it to be used as a truly medically large platform for automation.

5.4 Implication for further studies

Further studies can be undertaken on numerous other diseases using similar techniques and more data on other clinical health problems should be accumulated for similar studies. Further studies in this field require for achieving clinical accuracy and reliability.

APPENDICES

Appendix A: Project Reflection

First I would like to thank our supervisor Saiful Islam for his help. We make this project for every health conscious people who are very conscious about their health. Everyone can use this and know about his situation or percentage. If the percentage is very high then he may consult with a doctor. But unnecessary medical teste can be removed by modifying this project with more attributes and relevant data.

Appendix B: Related Diagrams

Decision Tree

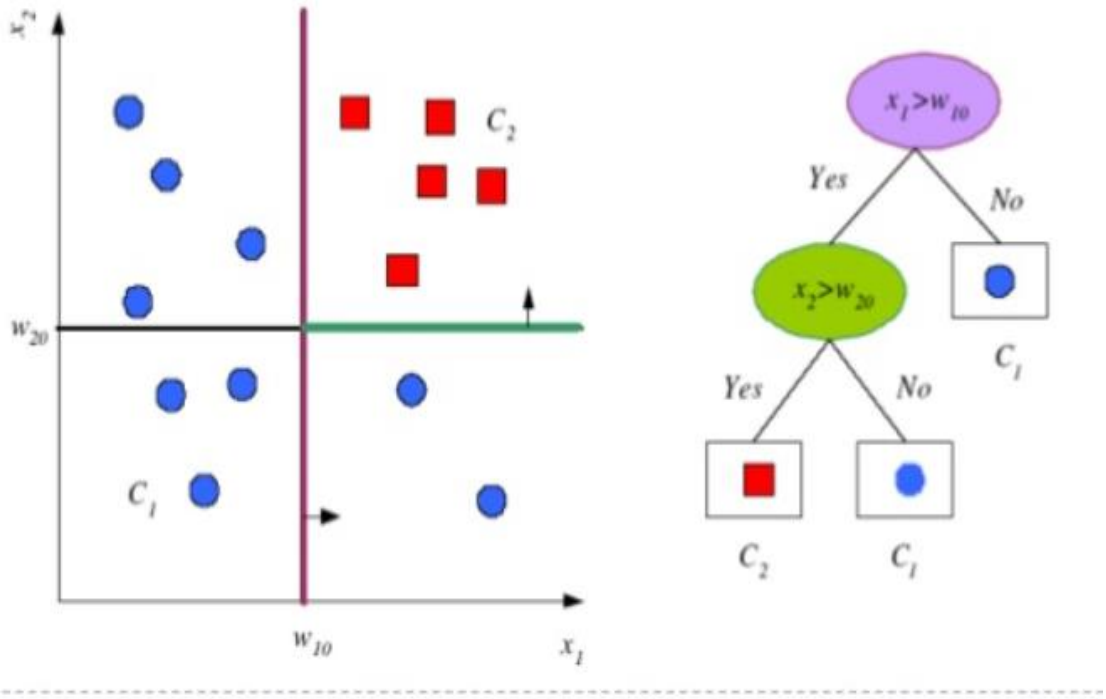


Figure 6: Decision Tree

SVM

SVM stands for Support vector machine. It's also a classification technique in machine learning. The figure shows the technique how it works.

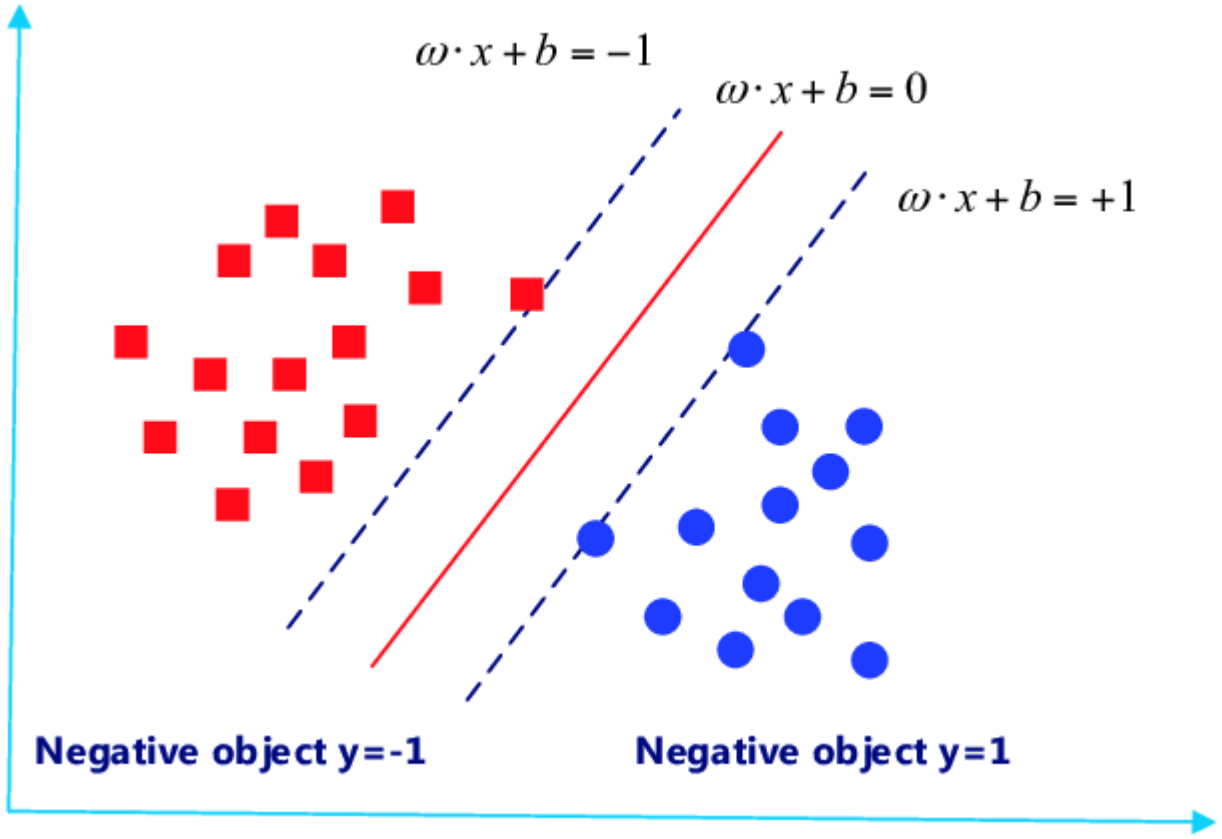


Figure 7: SVM

Our Project Link: https://github.com/istyak/hd

REFERENCES

[1] “Global atlas on cardiovascular disease prevention and control”, available at << https://www.who.int/cardiovascular_diseases/publications/atlas_cvd/en/>>, last accessed on 02-03-2019 at 1:00 PM..

[2] S. .Ishtake and S. .Sanap, “‘ Intelligent Heart Disease Prediction System Using Data Mining Techniques ” International Journal of healthcare & biomedical Research, vol. 1, no. 3, pp. 94–101, 2013.

[3]“Heart Disease Rates By Country”, available at <<https://www.worldatlas.com/articles/countries-with-highest-rates-of-cardiovascular-disease-deaths.html/>>, last accessed on 12-04-2019 at 11:00 AM..

[4] S.DANGARE, C. AND S. APTE, S. “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques” In-text: (S.Dangare and S. Apte, 2012)
Your Bibliography: S.Dangare, C. and S. Apte, S. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications, 47(10), pp.44-48.

[5] M. Jabbar, P. Chandra, and B. Deekshatulu, “CLUSTER BASED ASSOCIATION RULE MINING FOR,” Journal of Theoretical & Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.

[6] A. K. Sen, S. B. Patel, and D. P. Shukla, “A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level,” International Journal of Engineering and Computer Science, vol. 2, no. 9, pp. 1663–1671, 2013.

[7] “Cardiac patients rising in Bangladesh, 2.5 lakh die annually”, available at <<http://www.newagebd.net/article/51904/cardiac-patients-rising-in-bangladesh-25-lakh-die-annually/>>, last accessed on 17-04-2019 at 8:00 PM..

[9] S. B. Patil and Y. S. Kumaraswamy, “Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction,” International Journal of Computer Science and Network Security (IJCSNS), vol. 9, no. 2, pp. 228–235, 2009.

[10] In-text: (World Heart Federation, 2019)
Your Bibliography: World Heart Federation. (2019). Global Atlas on CVD Prevention and Control - World Heart Federation. [online] Available at: https://www.world-heart-federation.org/resources/global-atlas-cvd-prevention-control/ [Accessed 20 Nov. 2019].

[11] R. Chitra and V. Seenivasagam, “REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES,” Journal on Soft Computing (ICTACT), vol. 3, no. 4, pp. 605–609, 2013.

[12] N. A. Sundar, P. P. Latha, and M. R. Chandra, “PERFORMANCE ANALYSIS OF CLASSIFICATION DATA MINING TECHNIQUES OVER HEART DISEASE DATA BASE,” International Journal of Engineering Science & Advanced Technology, vol. 2, no. 3, pp. 470– 478, 2012.

[13] S. A. Pattekari and A. Parveen, “PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES,” International journal of Advanced Computer and Mathematical Sciences, vol. 3, no. 3, pp. 290–294, 2012.

[16] K. Srinivas, K. Raghavendra Kao, and A. Govardham, Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques,” in The 5th International Conference on Computer Science & Education, 2010, pp. 1344– 1349.

[18] P. Chandra, M. . Jabbar, and B. .Deekshatulu, “Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection,” in 12th International Conference on Intelligent Systems Design and Applications (ISDA, pp. 628– 634), 2012.

[19] S. U. Amin, K. Agarwal, and R. Beg, “Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors,” in Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), no. Ict, pp. 1227– 1231, 2013.

[20] “3 out of 4 people in Bangladesh run the risk of developing cardiac diseases” “<https://www.dhakatribune.com/health/2018/10/02/3-out-of-4-bd-individuals-run-the-risk-of-developing-cardiac-diseases/>>, last accessed on 19-04-2019 at 9:00 PM..

[21] Kelly BB, Fuster V. Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health. Washington, DC: National Academies Press (2010).

[22] Finks SW, Airee A, Chow SL, Macaulay TE, Moranville MP, Rogers KC, Trujillo TC. "Key articles of dietary interventions that influence cardiovascular mortality". Pharmacotherapy. (April 2012)

[23] Jump up to:a b Micha R, Michas G, Mozaffarian D. "Unprocessed red and processed meats and risk of coronary artery disease and type 2 diabetes--an updated review of the evidence". Current Atherosclerosis Reports. 14 (6): 515–24, (December 2012)

[24] Mendis S, Puska P, Norrving B (2011). Global Atlas on Cardiovascular Disease Prevention and Control. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization.. Archived from the original on 2016-05-06.

[25] Jump up to:a b Ciaccio EJ, Lewis SK, Biviano AB, Iyer V, Garan H, Green PH. "Cardiovascular involvement in celiac disease". World Journal of Cardiology (Review). 9(8): 652–666(August 2017).

[26] Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, et al. "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study"(2004).

[27] Jump up to:a b McPhee S. Current medical diagnosis & treatment. New York: McGraw-Hill Medical. p. 430(2012).

PLAGARISM REPORT

Turnitin Originality Report

Processed on: 31-Oct-2019 13:53 +06
ID: 1204146711
Word Count: 4474
Submitted: 1

Heart Disease By Ishtyak Ahmed

Similarity Index	Similarity by Source	
15%	Internet Sources:	N/A
	Publications:	N/A
	Student Papers:	15%

3% match (student papers from 07-Apr-2018)
Submitted to Daffodil International University on 2018-04-07

1% match (student papers from 28-Apr-2016)
Submitted to North East Surrey College of Technology, Surrey on 2016-04-28

1% match (student papers from 23-Sep-2017)
Submitted to Victoria University on 2017-09-23

1% match (student papers from 08-Oct-2019)
Submitted to Indiana University on 2019-10-08

1% match (student papers from 18-Jun-2017)
Submitted to Vel Tech University on 2017-06-18

1% match (student papers from 05-Apr-2018)
Submitted to Daffodil International University on 2018-04-05

1% match (student papers from 24-Nov-2018)

Score: 15% similarity.