A Comparative Study on Different Machine Learning Algorithms for Achieving Accurate Prediction for Heart Diseases

BY

Pronab Ghosh
ID: 151-15-4844

Khobayeb Ahmed
ID: 151-15-5200
AND

Madhob Karmaker
ID: 151-15-5331

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

Supervised By

Mr. Shaon Bhatta Shuvo
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Anup Majumder
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

NOVEMBER 2018

APPROVAL

This Project titled **"A Comparative Study on Different Machine Learning Algorithms for Achieving Accurate Prediction for Heart Diseases"**, submitted by Pronab Ghosh, Khobayeb Ahmed and Madhob Karmaker 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 Last Week of November 2018.

BOARD OF EXAMINERS

_____

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

_____

Dr. Sheak Rashed Haider Noori **Internal Examiner**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

_____

Md. Zahid Hasan **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

_____

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

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Mr. Shaon Bhatta Shuvo, Senior Lecturer, Department of CSE** in Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Mr. Shaon Bhatta Shuvo
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

_____

Mr. Anup Majumder
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

_____

Pronab Ghosh
ID: 152-15-4844
Department of CSE
Daffodil International University

_____

Khobayeb Ahmed
ID: 151-15-5200
Department of CSE
Daffodil International University

_____

Madhob Karmaker
ID: 151-15-5331
Department of CSE
Daffodil International University

Acknowledgement

First of all, we express our heartfelt thanks and gratitude to our Creator for our divine blessings, which helped us to successfully complete this thesis.

We express gratitude to **Mr. Shaon Bhatta Shuvo, Senior Lecturer** of CSE Department at Daffodil International University. Our supervisor's deep interest in deep knowledge and data mining affected us to manage this thesis. His endless endurance, scholarly instruction, continuous encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading of many inferior drafts and correcting them at all levels made it possible to complete this thesis. Without his guidance, it was totally impossible for us to complete, no doubt. And also to our parents and friends for their help and support.

We would like to express my heartiest gratitude to **Mr. Anup Majumder, Senior Lecturer**, Department of CSE, **Professor Dr. Syed Akhter Hossain, Head,** Department of CSE, Daffodil International University, Dhaka and for their kind help to finish my project.

We would like to express our gratitude to the men and women who were willing to share their experiences with us and take part in the study. Lastly, we are thankful to all of the faculty members of Daffodil International University who have inspired and motivated us throughout the entire undergraduate program.

Abstract

Over the years, heart diseases have become one of the most common causes related to death. Most of the time heart diseases are detected at the very last stage; therefore, an accurate prediction may reduce the catastrophe related to heart diseases. Heart-related diseases have a significant relationship with various health features including age, sex, heartbeat rate, blood pressure, cholesterol etc. In this context, four machine learning algorithms (e.g. Multiple Linear Regression, Decision Tree, Random Forest and Support Vector Machine) are applied on Cleveland heart disease dataset to analyze the comparative performance for achieving accurate prediction. The dataset contains thirteen health features, which have significant relations to heart disease. The best prediction has been achieved by the Random Forest algorithm, which is an ensemble version of the Decision Tree algorithm. To recapitulate the Random Forest algorithm outperformed other three algorithms followed by Support Vector Machine algorithm by providing a satisfactory prediction on 303 patient's data.

TABLE OF CONTENTS

LIST OF FIGURE

LIST OF TABLE

TABLES **PAGE**

CHAPTER 1

INTRODUCTION

1.1 Introduction

The heart is considered as one of the fundamental as well as the essential part for every human being to be alive. Malfunction of the heart may lead to damage of some other essential parts of the body such as kidney and brain. There are several categories of heart diseases. If a newborn baby born with heart disease then it is called ‘congenital heart disease’. On the other hand, if someone achieves heart disease after reaching infancy then it is called ‘acquired heart disease’. Nowadays, most of the case studies related to heart diseases fall in the categories of acquired heart disease. Most common heart related diseases are cardiovascular diseases, heart attack, coronary heart disease and Stroke. Stroke is the term of heart disease that is occurred by narrowing, blocking, or hardening of the blood vessels that go to the brain or by high blood pressure [1, 2].Various factors are responsible for various acquired heart diseases. Almost 17.3 million people died in 2008 only for being affected by various heart diseases. WHO provided an idea that based on current situation, heart diseases will be responsible for almost 23.6 million deaths in 2030. Various heart related data are generating every day. By analyzing these data we may find out in which stage now we are belonging, whether we are going to be affected in near future or not. With the advancement of technology, plenty of techniques have been developed for diagnosing heart diseases. However, accuracy in terms of prediction has not been able to reach up to the mark yet. In this paper, we shall see the comparative studies of

various machine learning algorithms used in the prediction of heart disease by using different data mining tools. This paper has separate parts to explain the overall procedure.

1.2 Motivation

The rate of heart diseases is increasing at an exponential rate. The busy lifestyle of people in this era with all the fast food in the lunch break and getting back to sitting and working has pushed as over the edge. Along with this people today have a lack of exercise and are less active. For most of them recreation is just another movie in bed or anything technology based. Physical activities have reduced drastically. These factors boosted the rate of heart diseases to an unfortunately high percentage. In a developing country like us the rate of heart diseases has the same effect. The annual mortality rate per 100,000 people from cardiovascular diseases in Bangladesh has increased by 128.9% since 1990, an average of 5.6% a year [3].

Prediction of heart diseases is a difficult and risky task. Since it is directly dependent on people's health, accuracy is a major factor. If not predicted accurately it can be disastrous. This research therefore focuses on the comparison of different data mining techniques to predict it. It shows the comparative analysis of the different methods. Cross validation error is used to compare the techniques. We have chosen Decision Tree, Random forest, Multiple Linear Regression and Support Vector Machine as they are the most widely used techniques in determining diseases.

1.3 The rationale of the study

Now- a- days, heart disease is counted as one of the most serious diseases. Many peoples are being died due to heart disease. It considered as a most attractive disease for becoming related to the

heart. After searching and analyzing we chose heart disease as our research topic. For becoming a large number of dead on the heart disease, the research topic has been selected. Finally, the paper has been working on this to provide a better suggestion that helps us to reduce the dead number for our modern age peoples.

1.4 Research Questions

Already, several harmful diseases have been detected for a human being. Although each disease has a solution for prevention it's not possible for everyone due to only for unconsciousness. Everyone wants to lead a happy life in where a disease is the only obstacle. Any kind of disease prevention is possible if that in remain primary stage. For that reason, we built a prediction system that helps to identify the disease stage and provides us the result that he or she is affected or not. All of the diseases, heart disease is considered one of the leading diseases. Many peoples are died due to this disease. Heart disease is the biggest killer of both men and women in the United States, England, Wales, and Canada. Finally, we selected it as our research topic for our satisfaction.

1.5 Expected Output

Heart disease prediction system is a system that helps to generate an expected result based on the given dataset. From this system, we used 70% of the training to get more accurate predictions. How accurate is it, it depends entirely on the training dataset. After completing all the needed procedure of the proposed system, our system has been ready for preparing output on the given dataset. We have applied various strategies to achieve our desired results. We got a 91% accuracy from the random forest (RF) among all that we have used.

1.6 Report Layout

These have shown in different chapters according to the following instruction: Chapter 2 has been given for the background of this research. The Research methodology has been illustrated in chapter 3 where its implementation process has been demonstrated and the predictable data source has been also used. Experimental results and discussion that has been provided in chapter 4 for showing our expected outcome basis on the given dataset. Finally, we have considered some aspects through the Summary, conclusion, recommendation and implication for future research that has been also explained in chapter 5.

CHAPTER 2

BACKGROUND

2.1 Introduction

Several works have been done for getting accurate prediction on heart disease prediction system. Due to the leading cause of death various approach has been applied for detecting prediction. In this segment, we have discussed Relative works, Research Summary, Scope of the Problem and Challenges of the proposed system architecture that is given one after another.

2.2 Related Works

Already numerous works have been done for getting disease prediction. They applied different types of techniques according to their proposed system. There are several research paper has been published using just one data mining technique for diagnosis of heart disease as given in Shadab et al [4], Carlos et al [5] etc. Frank Le Duff et al [6] have made a decision tree for treatment with the patient database. But in Ms. Ishtake et al [7], MA. Jabbar et al [8], Shantakumar et al [9] has used more than one data mining techniques in their research paper. Naive Bayes, Neural Network and Decision Trees has been shown through Sellappan Palaniappan et al [10] for displaying better prediction in their research paper. Niti Gaur et al [11] applied neural networks on Blood Pressure and Sugar for getting prediction of heart disease. They have

used 13 attributes for training and testing of data through neural networks. Latha Parthiban et al [12] proposed the neural network capabilities with the fuzzy logic and genetic algorithm. Kiyong Noh and et al [13] used a classified method that collected from various features and ECG. The dataset had 670 people and they were distributed to two groups. T.J. Peter et al [14] used pattern recognition and various types of data mining techniques include the Naive Base, Decision Tree, KNN and Neural Networks. But Naive Bayes technique outperformed than other used techniques T. J. Peter et al [14]. S. B. Patil et al [15] used K-means clustering algorithm to do extract necessary data, and MAFIA (Maximal frequent Item set algorithm) algorithm applied to determine the weightage. M. Jabber et al [16] has proposed a new system based on the serial number and cluster transaction data set, which is followed by the strategy of mining strategy for implementation by C programming. S. U. Amin et al [17] includes age, family history, diabetes, obesity hypertension, high cholesterol, smoking and alcohol intake for prediction and have used genetic algorithm for initialization of neural network weights.

2.3 Research Summary

Heart disease is a disease that attacks the heart. Without a doubt, the heart is a very important part of every human being. Therefore, if we want to lead a healthy life, we have to be cautious. If we find it early in the initial stage we can easily overcome it. Otherwise, we must bear its bad effects for our future life. After making the decision based on the current situation, we wanted to establish a system that provides better performance due to disease and understand the situation of affected patients. Finally, we touch our expected goal for God's blessing, which we have thought to implement.

2.4 Scope of the Problem

Scopes of the Problems in my thesis was including data collection, missing value selection, data preprocessing and implementation process and choses appropriate algorithm for getting accurate prediction accuracy of the given system.

The thesis topic has some own criteria with respect to its conditions. After completing of implementation process it generates a result that helps to predict which stage a heart disease patients are, somehow it predicts accuracy less than 70 then it counts as an unusable because of low prediction. In that time, we have to choose another topic that we can't think in our dream. Because of getting lowest accuracy it would be great danger for us. So we just divide our thesis work in various parts that we can easily handle it and face any problem then we overcome it. The time limit is given to the page:

Time scheduling	Month
Data collection	2 months
Preprocessing	1 month
Feature scaling	1 month
Implementation	1 month
Testing	1 days
Total	5 months and 1 days

The whole task has been completed with respect to given times. But we faced several circumstances when we tried to complete it. At last we became success and touch our goal in time.

Target of my thesis work was prediction. How much accurate prediction of my system will provide that depends on the dataset.

After all, i did the tasks and share this idea to my friends and some younger brother they are excited to accept this type of system, I also share this thinking to many of our respected teachers and the enjoyable things was that they encouraged me to do this thesis complete.

2.5 Challenges

Data collection is one of the big challenges for getting predicting accuracy. Without data, the prediction is not possible and it can't predict. After that, another challenge is preprocessing. After doing preprocessing our data set has no null value and helps us to get a good prediction. Next, Feature scaling helps to take all feature values into the same scale with respect to value. Therefore different algorithm has been applied to the proposed architecture. Finally, the implementation process has been established to get accurate predicted value. There were several challenges rising according to the working procedure.

Chapter 3
RESEARCH METHODOLOGY

3.1 Introduction

Today, many hospitals can easily find out their needed information from healthcare. For example, the system has a lot of information for intelligent medical diagnostics, so they need hidden information too. There are 13 reasons for heart disease in this system. Some of them include age, sex, blood pressure and cholesterol. First, the dataset has been collected from the UCI repository. After that, we got the missing value and resolved using the intermediate method. To get the proper prediction, we've already completed the feature scaling process and installed our datasets to forecast. Datasets are used for training and testing purposes and here are some of them included algorithm Decision Tree (DT), Multiple Linear Regression (MLR), Support Vector Machines (SVM) and Random Forest (RF).The appropriate scenario has been given based on the working procedure.

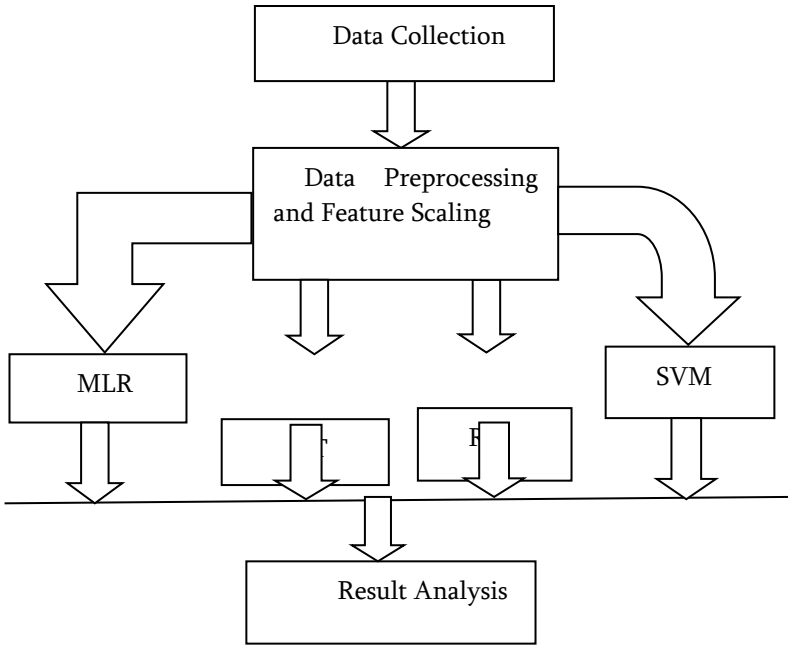

Figure. 1. Architecture of our proposed system.

3.2 Research Subject and Instrumentation

In recent years, the popularity of machine learning algorithms is rising exponentially. Machine Learning Algorithms provide computers with the ability to learn from data with the help of statistical approaches. A machine can find out the internal data pattern and produce a decision or predictive knowledge as an outcome without the help of explicit coding is considered as the most interest part. Therefore, the same algorithm can be applied to datasets of different domains without having a modification of its internal structures. There are different types of machine learning algorithms, but we have used some of them to our system. The algorithm details are given below:

3.2.1 Decision Tree

Decision tree method is used as the most powerful tool for learning the machine because it helps to get effective results as soon as possible. Decision tree has different types of algorithms: Cart, ID3, C 4.5, CHH and H48. Among them J48 is used and it is very popular algorithm.J48 uses pruning method for building a tree. This algorithm continues to be a recursive process until the expected results are found. It provides good accuracy and flexibility. This formula is made available from the following equations in Han, j. et al [18], Shouman et al [19], K.L. Jaiswal et al [20].

$E = \sum_{i=1}^{K} P_i \, log_2 \, P_i$(1)…….. (1)

From equation (1), E
K defines the number of classes of target attributes,
Pi defines the number of occurrences of class,
i is divided by the total number of instances.

3.2.2 Multiple Linear Regression

Multiple linear regression is a model that creates the relationship between one dependent variable and two or more independent variables and fits it through the linear regression for observing data. Each value of y is dependent on each value of the variable x. By analyzing the relation and direction of information, fitting the line and providing better scalability plot for better understanding. Multiple Linear Regression equation follows Durga Prasad et al [21]:

$y=\beta_0+ \beta_1 X_1+ \beta_2 X_2+.....+ \beta_i X_i$……. (2)

From equation (2) we see,

X is called explanatory variables. y is changing according to the variable of X. The suited values b0, b1… bi has allotted the parameters$\beta_0,\beta_1....\beta_i$ of the patient's regression line. Between the regression line and the single data point, the variance is not possible to explain. Therefore, unexplained variation is called the residual.

3.2.3 Random forest

This is another tool that helps to get the expected result. Random forests are considered as an ensemble learning methods for classification, regression, and probability. A large number of decision trees are used to create a random forest algorithm. Random forest algorithm came from Tin Kam Ho [22] using random subspace method

that constructs a multitude decision trees at training time and provides the output based on the mode for classification or means prediction for regression for each and every trees. [22]Random forests can easily find the missing values from large datasets and it can provide a more accurate value to the decision tree.

3.2.4 Support Vector Machine (SVM)

Support vector machines (SVMs) [23] is also known as support vector networks and supervised learning models associated with machine learning algorithms. It also analyzes data using for regression and classification and the working process has been described by V. D. Bhagile et al [24].

$$Y = \text{sign}(\sum_{i=1}^{N} y_i \alpha_i \, (\mathrm{x} * x_i) + \mathrm{b}) \quad \text{..... (3)}$$

From equation 3,

$(\mathrm{x} * x_i)$ Is known as labeled training and works as an input vector.

$$Y = \text{sign}(\sum_{i=1}^{N} y_i \alpha_i K \, (\mathrm{x}, x_i) + \mathrm{b}) \quad \text{..... (4)}$$

From Equation (3), we see

$\mathrm{K}(\mathrm{x}, x_i)$ is called kernel function and constructs machine along with different types of non-linear decision surfaces in the input space for generating the inner products.

3.3 Data Collection Procedure

The most important part of this paper is to create intelligent heart disease forecasting systems that help diagnose heart disease by using our Cleveland Heart Disease Dataset. UCI [17] has many database of cardiovascular diseases in the machine learning repository [17], among which we have taken the database of Cleveland Heart Disease which has 303 records. The data set consists of 3 types of attributes: Input, Key & Predictable attribute which has been given according to list in below:

3.3.1 Input attributes

The Cleveland data set has 13 features these are:
1. Age in Year
2. Sex - (value 1: Male; value 0: Female)
3. Cp - chest pain type (value 1: typical angina, value 2: atypical angina, value 3: non-angina pain, value 4: asymptomatic)
4. Trestbps - resting blood pressure (in mm Hg on admission to the hospital)
5. Chol - serum cholesterol in mg/dl
6. Fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. Restecg - resting electrocardiographic results (value 0: normal, value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. Thalach - maximum heart rate achieved
9. Exang - exercise induced angina (1 = yes; 0 = no)
10. Oldpeak - ST depression induced by exercise relative to rest
11. Slope - the slope of the peak exercise ST segment (value 1: unsloping, value 2: flat, value3: downsloping
12. Ca - number of major vessels (0-3) colored by fluoroscopy
13. Thal - 3 = normal; 6 = fixed defect; 7 = reversible defect
14. Num - diagnosis of heart disease (angiographic disease status) (value 0: < 50% diameter narrowing, value 1: > 50% diameter narrowing)

3.3.2 Key attribute

Patient ID: Patient's Identification Number

3.3.3 Predictable attribute

Diagnosis: Value 1 = < 50 % (no heart disease) Value 0 = > 50 % (has heart disease)

3.4 Statistical Analysis

This figure illustrates which one are the most suitable output for this proposed system. From this figure, we can easily understand that Random forest is the best solution among of them.

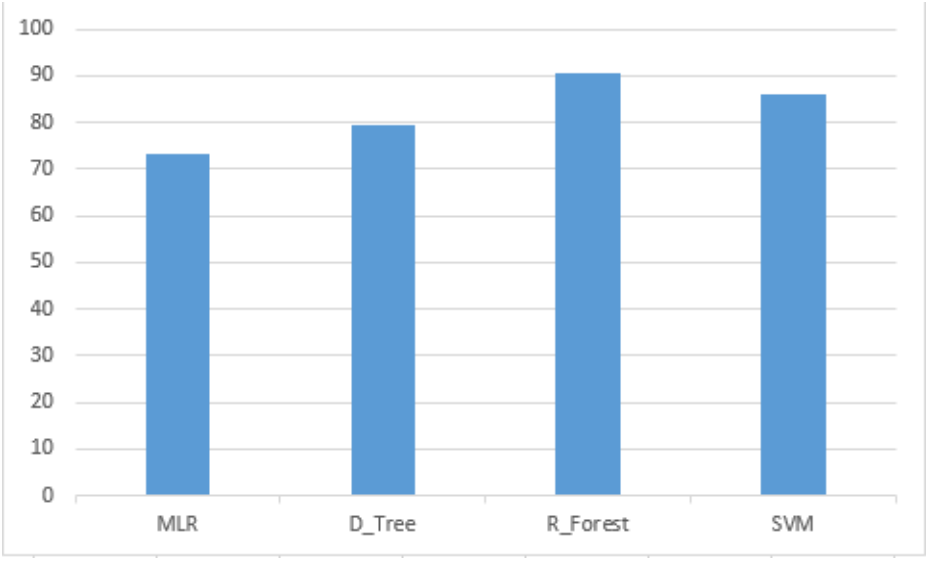

Figure. 2. Applied algorithms of our proposed system.

3.5 Implementation Requirements

Implementation is considered as a fundamental sector for making any system. Anaconda is an environment that consist of python and all deep learning packages. We implement our system using Spyder compiler that helps to terminate our system. Python version 3.0.5 has been used and considered as a latest version of python. Various types of library Function has been used for implementation. We collected our system dataset from UCI Repository and downloaded it as a CSV file. After downloading dataset has been applied to the coding segment through the pandas library function. All of them library function has been given that we used in the implementation segments:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import sklearn.metrics
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import RandomForestRegressor
from sklearn.cross_validation import train_test_split
```

Chapter 4
Experimental Results and Discussion

4.1 Introduction

Many researchers have implemented their system according to their proposed procedure due to cardiovascular disease. Prophecy or outcome is the only means of detecting the stage of the disease. Therefore, depending on how good the results will be, on the given data set and on how much the correct value is given in the given dataset.

4.2 Experimental Results

In this study, we have implemented a variety of strategies that help get the right results based on the given dataset. In the proposed dataset we used four different methods, they included multiple linear regression (MLR), random forest (RF), decision trees (DT) and support vector machine (SVM). Each system has its own criteria according to their limitations. On our systems, random forests help us to get the highest accuracy among them.

Table 1: Accuracy comparison for Heart Disease Data set

Serial - Number	Approach	Accuracy
1.	Multiple Linear Regression	73%
2.	Decision Tree	79.25%
3.	Random Forest	90.50%
4.	Support Vector Machines	86%

4.3 Descriptive Analysis

Although, data preprocessing is so much needed to get expected accuracy. Accuracy takes an important part for doing every prediction. Our data is ready for the algorithm because of completing the preprocessing method. Before the applied algorithm, we divided our data into two parts like training dataset (75%) and testing dataset (25%). After that, we applied 4 algorithms. Firstly, Multiple Linear Regression provides us 73% accuracy on the given dataset. Secondly, Decision tree gives us about 79% accuracy. Thirdly, Random forest provides us more than 90% that experienced was so much good and finally from Support Vector Machines we have achieved 86% prediction. Overall, Random Forest provides the highest accuracy between all of them.

4.4 Summary

Overall, our system has several approaches for finding the accurate predicted results. However, researchers applied their system according to their knowledge based on the Cleveland heart disease dataset. Where approximately 369 dataset diseases are taken for forecasts. Although,

the prediction values are not always provided with the correct values. . However, our forecast system provides better accuracy than other existing systems and it gives us about 91% on the basis of dataset. Finally, we overcome all types of obstacles for doing accurate prediction and arrives a stage that helps to identify the situation of disease.

Chapter 5
Summary, Conclusion, Recommendation and Implication for Future Research

5.1 Summary of the Study

This research paper is designed for people who do not have the ability to pay huge amounts. Besides, if the diagnosis of the disease is in normal condition, it is possible to return to good position without paying a lot of money. However, we tried to implement this system thinking about eliminating cost, time and stress and enjoying a happy life.

5.2 Conclusions

We have analyzed some machine learning algorithms to predict heart diseases in this paper. Since we are fully dependent on the dataset collected with disease techniques, the arrangement cannot be explained as final. Here, we have analyzed for thirteen features. Although, prediction varies with the dataset. If we provide a huge amount of data

in the training value it can easily provide better prediction from our present. However, we are pleased with the results of our work and our results can be considered well enough compared to other relevant work. After that, the accuracy might change if we apply the more features to a large dataset.

5.3 Recommendations

This system is designed for affected patients, who have no idea about the true signs of heart disease. Through this method, they can easily manage their happy life without any panic. They will gain knowledge of all types of existing signs that are harmful for healthy living. Eventually, the person affected by heart disease will be able to detect the disease in the initial stage.

5.4 Implication for Further Study

In modern times the heart disease is considered a formidable disease from all kinds of diseases. The ratio of death value is maximum than any other disease. There are several unknown symptoms observed in the patients. The effects of the disease are increased to destroy the human heart. For this reason in the future, we want to add more features according to that situation and will try to provide a better solution from the existing system capabilities.

REFERENCES

[1] T.H.G Dent, “Predicting the risk of coronary heart disease”, PHG foundation publisher, 2010.

[2] World Health Organization, “Global status report on no communicable diseases”, 2014.

[3] Shadab Adam Pattekari and Asma Parveen, “Prediction System for Heart Disease Using Naive Bayes”, International Journal of Advanced Computer and Mathematical Sciences, vol. 3, pp. 290-294, 2012,

[4] Carlos Ordonez, Edward Omiecinski, Mining Constrained Association Rules to Predict Heart Disease, IEEE. Published in International Conference on Data Mining (ICDM), pp. 433-440, 2001.

[5] Franck Le Duff, Cristian Munteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.

[6] Ms. Ishtake S.H, Prof. Sanap S.A., “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, 2013.

[7] Ma.jabbar, Dr.prirti Chandra, B.L.Deekshatulu," cluster based association rule mining for heart attack prediction", Journal of Theoretical and Applied Information Technology, 2011.

[8] Shantakumar B. Patil, Dr.Y.S. Kumaraswamy," Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", (IJCSNS) International Journal of Computer Science and Network 228 Security, 2009.

[9] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, 2008

[10] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1, 2007.

[11] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological and Life Sciences, Vol. 3, No. 3, 2008.

[12] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721- 727, 2006.

[13] T. J. Peter and K. Somasundaram, "AN EMPIRICAL STUDY ON PREDICTION OF HEART DISEASE USING CLASSIFICATION DATA MINING TECHNIQUES," 2012.

[14] 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.

[15] 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.

[16] 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, no. Ict, pp. 1227–1231, 2013

[17] UCI Machine Learning Repository [homepage on the Internet]. Arlington: The Association; 2006 [updated 1996 Dec 3; cited 2018 October 7]. Available from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[18] Han, j. and M. Kamber, “Data Mining Concepts and Techniques”. 2006: Morgan Kaufmann Publishers. Lee, I.-N., S.-C. Liao, and M. Embrechts, Data.

[19] Mai Shouman, Tim Turner, Rob Stocker,” Using Decision Tree for Diagnosing Heart Disease Patients”, Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia

[20] Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen,” A Heart Disease Prediction Model using Decision Tree”, IOSR Journal of Computer Engineering (IOSR-JCE)e-ISSN: 2278-0661, p-ISSN: 2278-8727 Vol. 12, PP 83-86 ,2013

[21] K.Polaraju, D.Durga Prasad, “Prediction of Heart Disease using MultipleLinearRegression Model”, Vol.05, ISSN: 2321-9939

[22] Tin Kam Ho, "The Random Subspace Method for Constructing Decision Forest"(PDF).IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 832–844, 1998.

[23] Cortes, Corina; Vapnik, Vladimir N, "Support-vector networks" .Machine Learning, pp. 273–297, 1995.

[24] Shaikh Abdul Hannan, V. D. Bhagile, R. R. Manza, R. J. Ramteke,” Diagnosis and Medical Prescription of Heart Disease Using Support Vector Machine and Feed forward Backpropagation Technique ” ,International Journal on Computer Science and Engineering Vol. 02, No. 06, pp. 2150-2159,2010

APPENDIX

Appendix A: Research Reflection

The segment are used here for showing the reflection of our thesis that means the actual outcome of our thesis. After completing this research, we all get clear concept about this. We are introduced how to face challenges and overcome them. We did various types of individual project for completing our semester. But we have worked on the Research for the first time. This was so much interesting for all of us and we faced lots of problems around us. After that, we complete it, enjoy it and learn many things from it. We can understand that proper planning is the fundamental key for getting any kind of success. Without a hint of doubt, we completely satisfied for our working process only for follow some basic steps that will be so much helpful for our future life. How to maintain time and how to use it. Finally, we have completed all our works within the times and become affected by the research paper that will help us to become a researcher.

Appendix B: Related Issues

 # Multiple_Linear_Regression

```
#import the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

#import the dataset

Dataset = pd.read_csv("knowledge.csv")

dataset.loc[dataset["num"]==2,"num"]=1
dataset.loc[dataset["num"]==3,"num"]=1
dataset.loc[dataset["num"]==4,"num"]=1

X=dataset.iloc[:, :-1].values
df=pd.DataFrame(X)
Y=dataset.iloc[:, 13].values

#import the missing values

X[:, 11] =dataset["thal"].fillna(method="ffill")
X[:, 12] =dataset["ca"].fillna(method="ffill")


#train_test_split

from sklearn.cross_validation import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.25,rand
om_state=0)



#feature scalling
```

```
from sklearn import preprocessing
mimmax_scaler=preprocessing.MinMaxScaler(feature_range=(0,1))
X_train=mimmax_scaler.fit(X_train).transform(X_train)
X_test=mimmax_scaler.transform(X_test)


from sklearn import preprocessing
X_train=preprocessing.normalize(X_train, norm='l1')
X_test=preprocessing.normalize(X_test, norm='l1')

from sklearn.preprocessing import StandardScaler
sc_X=StandardScaler()
X_train=sc_X.fit_transform(X_train)
X_test=sc_X.transform(X_test)

#Backward Elimination sector
import statsmodels.formula.api as sm
X=np.append(arr=np.ones((303,1)).astype(int),values=X,axis=1)

X_optimal=X[:, [0,1,2,3,4,5,6,7,8,9,10,11,12,13]]
regressor_OLS=sm.OLS(endog = Y, exog = X_optimal).fit()
regressor_OLS.summary()

X_optimal=X[:, [0,2,3,4,5,6,7,8,9,10,11,12,13]]
regressor_OLS=sm.OLS(endog = Y, exog = X_optimal).fit()
regressor_OLS.summary()

X_optimal=X[:, [0,2,3,4,6,7,8,9,10,11,12,13]]
regressor_OLS=sm.OLS(endog = Y, exog = X_optimal).fit()
regressor_OLS.summary()


X_optimal=X[:, [0,2,3,4,7,8,9,10,11,12,13]]
regressor_OLS=sm.OLS(endog = Y, exog = X_optimal).fit()
regressor_OLS.summary()

X_optimal=X[:, [0,2,3,4,8,9,10,12,13]]
regressor_OLS=sm.OLS(endog = Y, exog = X_optimal).fit()
```

```
regressor_OLS.summary()


X_optimal=X[:, [2,3,4,8,9,10,12,13]]
regressor_OLS=sm.OLS(endog = Y, exog = X_optimal).fit()
regressor_OLS.summary()

X_optimal=X[:, [2,3,8,9,10,12,13]]
regressor_OLS=sm.OLS(endog = Y, exog = X_optimal).fit()
regressor_OLS.summary()
```

……

#Decision_Tree_Classifier

```
#import the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import sklearn.metrics
from sklearn.tree import DecisionTreeClassifier


#import the dataset

dataset=pd.read_csv("knowledge.csv")

dataset.loc[dataset["num"]==2,"num"]=1
dataset.loc[dataset["num"]==3,"num"]=1
dataset.loc[dataset["num"]==4,"num"]=1

X=dataset.iloc[:, :-1].values
df=pd.DataFrame(X)
Y=dataset.iloc[:, 13].values

#import the missing values
```

```
X[:, 11] =dataset["thal"].fillna(method="ffill")
X[:, 12] =dataset["ca"].fillna(method="ffill")

#train_test_split

from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3,rando
m_state=0)


from sklearn import preprocessing
X_train=preprocessing.normalize(X_train, norm='l1')
X_test=preprocessing.normalize(X_test, norm='l1')

#feature scalling
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)


#Training the Algorithm


tree = DecisionTreeClassifier(criterion = "entropy", random_state =
100,max_depth=3, min_samples_leaf=5)
tree.fit(X_train,Y_train)

print('Accuracy on the training subset:
{:.3f}'.format(tree.score(X_train, Y_train)*100))
print('Accuracy on the test subset: {:.3f}'.format(tree.score(X_test,
Y_test)*100))
```

…….

#support_vector_Machines

```
#import the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.svm import SVC


#import the dataset

dataset=pd.read_csv("knowledge.csv")

dataset.loc[dataset["num"]==2,"num"]=1
dataset.loc[dataset["num"]==3,"num"]=1
dataset.loc[dataset["num"]==4,"num"]=1

X=dataset.iloc[:, :-1].values
df=pd.DataFrame(X)
Y=dataset.iloc[:, 13].values

#import the missing values

X[:, 11] = dataset["thal"].fillna(dataset["thal"].median())
X[:, 12] = dataset["ca"].fillna(dataset["ca"].median())

#train_test_split

from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3,ran
dom_state=0)


from sklearn import preprocessing
X_train=preprocessing.normalize(X_train, norm='l1')
X_test=preprocessing.normalize(X_test, norm='l1')

#feature scalling

from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
```

```
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)


#Training the Algorithm
svm=SVC()
svm.fit(X_train,Y_train)

print('Accuracy on the training subset: {:.3f}'.format(svm.score(X_train, Y_train)*100))
print('Accuracy on the test subset: {:.3f}'.format(svm.score(X_test, Y_test)*100))


……
#Random_Forest

from sklearn.neural_network import MLPClassifier
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.model_selection import train_test_split
import sklearn.metrics



#import the dataset

Dataset = pd.read_csv("knowledge.csv")

dataset.loc[dataset["num"]==2,"num"]=1
dataset.loc[dataset["num"]==3,"num"]=1
dataset.loc[dataset["num"]==4,"num"]=1

X=dataset.iloc[:, :-1].values
df=pd.DataFrame(X)
Y=dataset.iloc[:, 13].values

#import the missing values
```

```
X[:, 11] = dataset["thal"].fillna(dataset["thal"].median())
X[:, 12] = dataset["ca"].fillna(dataset["ca"].median())

#train_test_split

X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.25,ra
ndom_state=42)

from sklearn import preprocessing
mimmax_scaler=preprocessing.MinMaxScaler(feature_range=(0,1))
X_train=mimmax_scaler.fit(X_train).transform(X_train)
X_test=mimmax_scaler.transform(X_test)


from sklearn import preprocessing
X_train=preprocessing.normalize(X_train, norm='l1')
X_test=preprocessing.normalize(X_test, norm='l1')


#feature scalling
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

from sklearn.ensemble import RandomForestRegressor
forest=RandomForestRegressor()
forest.fit(X_train,Y_train)


mlp = MLPClassifier(random_state=42)
mlp.fit(X_train, Y_train)

print('Accuracy on the training subset:
{:.3f}'.format(mlp.score(X_train, Y_train)*100))
```

```
print('Accuracy on the test subset: {:.3f}'.format(mlp.score(X_test, Y_test)*100))

print('The maximum per each feature:\n{}'.format(X.max(axis=0)))
```

…….....