

Early Identifying Heart Disease by Prediction Using Data Mining

SUBMITTED BY

MD. Khairul Islam
ID: 172-15-9838

MD. Nabir Hossain
ID: 172-15-9679

AND

MD. Mahfuzur Rahman
ID: 172-15-10030

SUPERVISED BY

Ms. Sharmin Akter
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

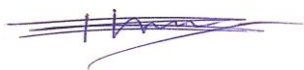
DHAKA, BANGLADESH

APRIL, 2021

APPROVAL

This Project/Internship titled “**Early Identifying Heart Disease by Prediction Using Data Mining**”, submitted by **Md. Khairul Islam**, ID No: **172-1-9838**; **Md. Nabir Hossain**, ID No: **172-1-9679**; **Mahfuzur Rahman**, ID No: **172-15-10030**; 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 **1 June 2021**.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan

Chairman

Professor and Head

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



Raja Tariqul Hasan Tusher

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Dr. Dewan Md. Farid

Associate Professor

Department of Computer Science and Engineering

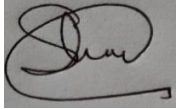
United International University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ms. Sharmin Akter, Lecturer, Department, of CSE Daffodil International University**. We 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:



Sharmin Akter

Lecturer

Department of CSE

Daffodil International University

Co-Supervised by:



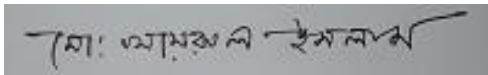
Mr. Abdus Sattar

Assistant Professor

Department of CSE

Daffodil International University

Submitted by:



MD. Khairul Islam

ID: 172-15-9838

Department of CSE

Daffodil International University

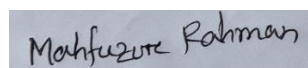


Md. Nabir Hossain

ID: 172-15-9679

Department of CSE

Daffodil International University



Mahfuzur Rahman

ID: 172-15-10030

Department of CSE

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 project/internship successfully.

We really grateful and wish our profound our indebtedness to **Ms. Sharmin Akter, Lecturer, Department, of CSE Daffodil International University**. Deep Knowledge & keen interest of our supervisor in the field of “Data Mining” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

Abstract:

Nowadays Cardiovascular disease (heart disease) is a hot theme of medical field. Under the title of “Early Identifying Heart Disease by Prediction by Using Data Mining” our paper will expose pre-condition about heart disease. Whole works are done by some attributes of patients by using some algorithm to express better accuracy of heart disease prediction. K-NN, Decision Tree Classifier, Regression, Support vector Machine, an

d Random Forest Classifier are used. Using every algorithm, we found out best accuracy from Random Forest Classifier. And the accuracy is 97.99%. So, after better predicting we/people can take decision our/their next steps for better treatment.

TABLE OF CONTENTS

CONTENTS	PAGES
Approval Page	i - ii
Declaration	iii - iv
Acknowledgement	v
Abstract	vi
List of Figures	ix
Chapter 1: Introduction	01 - 02
1.1 Introduction	01
1.2 Motivation	01
1.3 Rational of the Study	01
1.4 Expected Outcome	02
Chapter 2: Background	03 - 04
2.1 Preliminaries/Terminologies	03
2.2 Related Works	03
2.3 Comparative Analysis and Summary	03
2.4 Scope of the problem	04
2.5 Challenges	04
Chapter 3: Research Methodology	05 - 10
3.1 Research Subject and Instrumentation	05
3.2 Data Collection Procedure/Dataset Utilized	05
3.3 Detailed Methodology	05
3.4 Data Processing	06
3.5 Statistical Analysis	07 - 10
Chapter 4: Requirement Specification	11 – 12
4.1 Business Process Modeling	11
4.2 Use Case Modeling and Description	12
4.3 Design Requirements	12

Chapter 5: Design Specification	13 - 15
5.1 Front-End Design	13 - 14
5.2 Back-End Design	14
5.3 Implementation Requirements	15
Chapter 6: Conclusion and Future Scope of Development	16
6.1 Discussion and Conclusion	16
6.2 Scope of Further Developments	16
References	17
Plagiarism Result	18

LIST OF FIGURES

FIGURES	PAGE NO
Figure: 3.1: Target column graph (yes-1, no-0)	05
Figure: 3.2: Bar chart comparison between target & input value - 1	06
Figure: 3.3: Bar chart comparison between target & input value - 2	06
Figure: 3.4: Point graphs between Max heart rate vs final output.	07
Figure: 3.5: Result of all algorithm.	08
Figure: 4.1: Data Flow Diagram	09
Figure: 4.2: Use case diagram	10
Figure 5.1: Web page design screen - 1	11
Figure 5.2: Web page design screen – 2	12
Figure: 5.2: Back end flow	12
Figure: Plagiarism result	16

CHAPTER 1

INTRODUCTION

1.1 Introduction:

Heart disease is common concern and frightening disease in the world. To get better accuracy of heart disease prediction, we worked about it precisely. Human being is dependent on a proper functioning of heart. There may happen heart attack by different type of cardiovascular diseases. After heart attack, there may be brain stroke for imbalance or inefficient blood circulation. Averagely, a lot of attributes are used here. Such, age in years, sex, cp- chest pain, trestbps- resting blood pressure(anything above 130-140 is typically cause concern), chol- serum Cholesterol in mg/dl, fbs - fasting blood sugar (> 120 mg/dl , >'126' mg/dl signals diabetes), restecg – resting electrocardiographic results, thalach – maximum heart rate achieved, exang – exercise induced angina, Oldpeak, slope –the slope of the peak exercise ST segment , ca – number of major vessels colored by fluoroscope, thal –thallium stress result, Target –have disease or not (1=yes; 0 = No ; prediction attributes). In heart disease chest pain in different type. Like, typical angina: chest pain may happen decrease blood supply to the heart, Atypical angina: Not related to heart, non –anginal pain: typically, esophageal spasms, and asymptomatic. Excessive serum cholesterol another major problem to carry out heart disease, and high blood pressure and others.

1.2 Motivation:

In the medical field area, there is no easy way to identify heart disease. It's a little bit costly specially, lower middle-class people. We want to solve this problem using data mining. People are suffering cardiovascular disease surreptitiously of their mind. Prior this happens, it's difficult to identify cardiovascular disease (heart disease). Some people are infected congenital heart disease. To identify heart disease patient, have to follow various test concurrently. It's a little bit hard and costly to catch up the syndrome. That's why we will find out the collection of cardiovascular disease related syndrome data to predict heart disease perfectly. So, that people can easily get treatment.

1.3 Rationale of the Study:

- We have to collect huge data about heart disease.
- Using algorithm there will be given predict about heart disease prediction.
- Let patient know their condition about the disease and can take proper treatment.

1.4 Expected Outcomes:

- Find out heart disease prediction with best accuracy.
- Knowing the stage of heart disease properly.
- Perceiving about heart disease prediction result.
- Increase aware about heart disease after prediction to the people.

CHAPTER 2

BACKGROUND

2.1 Preliminaries/Terminologies:

For sure we can say, our primary task is to quest heart disease prediction from patient whether they have heart disease or not. As though, patient can obtain precise information. And finally, patient can get better treatment going to hospital knowing after their stage of heart disease prediction.

2.2 Related Works:

01. J.Vijayashree and N.Ch.SrimanNarayanaIyengar discussed different types of data mining algorithm and Hybrid intelligent technique. And they try to give a concept about a summary of prediction result with different types of technique. And they also discover a basic methodology of database. And they also compare prediction percentage of several previous work.

02. Monther Tarawneh and Ossama Embarak build a hybrid model based on some data mining algorithm. In thy hybrid algorithm they used KNN (77%), J 4.8(84%), Genetic Algorithm (84%), Decision Tree (78%), ANN (85%), SVM(86%), Naïve Bayes(87%) algorithms. And the final hybrid Technique result is 89.2%

03. Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin used different types of algorithm such as, Sequential Minimal Optimization (SMO) (84.07%), Bayes Net (BN)(81.11%), Multi-Layer Perception (MLP)(77.4%), Navies Bayesian (NB)(89.77%) algorithms. And they also used Advanced Encryption Standard (AES) for data security technique. As the best accuracy comes from the Navies Bayesian algorithm (89.77%) so the final prediction comes from Navies Bayesian algorithm.

2.3 Comparative Analysis and Summary:

From the above mention work [1]J.Vijayashree and N.Ch.SrimanNarayanaIyengar they just give summary of several work and the compare different types of work in this field but then don't analysis any data set.[2] Monther Tarawneh and Ossama Embarak their hybrid model gives the accuracy is 89.2% but our proposed analysis will give more than that. [3] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin they get most accuracy from the Navies Bayesian algorithm (89.77%) but we got the most accuracy from the Random Forest Classifier (97.99%).

2.4 Scope of the Problem:

In medical field there is a large scope to collect data. But in this paper work, we obtained data from Kaggle a large website of data. We collect different type of data and sorted out them to find precise prediction. We trained so many times to collect raw data from medical site, but we fell to the ground because of covid-19. But we are keep trying our best to collect real time data.

2.5 Challenges:

- To derive new data in covid-19 situation.
- Learning about data mining field deeply.
- To predict best accuracy on large scale of data.
- Turn to set up whole system into based and mobile phone application.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation:

Recently, the name of alarmism is heart ailment suffered aged people literally. To be healing from heart disease and to arouse consciousness among people about it. We worked in large quantities on it. Getting cure from it we need to launch some steps to introduce about heart diseases. Thereby we inaugurated heart disease prediction system alongside a website based on it. We used google colab allowed us to execute all data about heart disease to lead better accuracy. And everything were performed by python language platform.

We retained it also into mobile base platform (application) to do more user friendly to patients.

3.2 Data Collection:

Normally, we collected Global data from kaggle website. We couldn't collect data from hospital due to covid-19. That's why we collected over 1387 data about heart disease from kaggle site. Our data pattern also diverse (depend on testified) from general, thereby we also couldn't collect data using online base way like google form.

3.3 Detailed Methodology:

On the purpose of exploring our dataset, we adopted almost six algorithms.

- K-NN (K- nearest neighbors)
- Decision tree classifier
- Logistic regression
- Support vector machine
- Random forest classifier
- XGBoost classifier

3.3.1 Data Processing:

Firstly, we merged data from kaggle website, then we arranged data with heart disease patients and non-heart disease patients. We have 691 person/patients with heart disease and 637 person/patients without disease. So, we can say our dataset is balanced.

```
[ ] df.target.value_counts()
```

```
1    691
0    637
Name: target, dtype: int64
```

```
[ ] df.target.value_counts().plot(kind="bar", color=["salmon", "lightblue"])
```

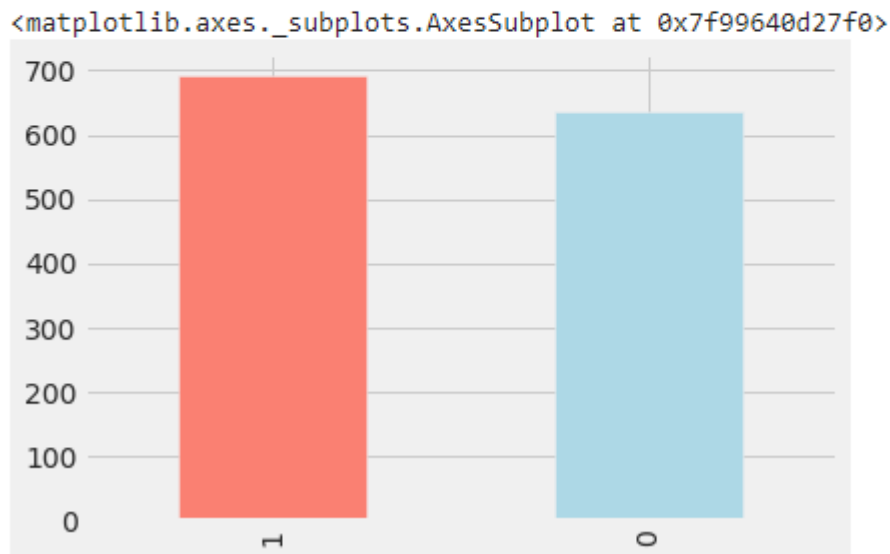


Figure: 3.1: Target column graph (yes-1, no-0)

Here, we set up target value '1' for disease and '0' for non-disease.

After exploring the dataset, I observed that I need to convert some categorical variables into dummy variables and scale all the values before training the Machine Learning models. First, I'll use to get dummies method to create dummy columns for categorical variables. Before that we attempted to commit training trial split. And before committing training trial split we took all our distinct peculiarities at single volatile and finally all distinct peculiarities turned into predicted class.

Mainly we used six methodology to process data to get better accuracy of prediction. And then we applied these methods step by step properly. After using data processing by all method, we got the best accuracy by Random Forest Classifier and which is 97.99%.

3.4 Statistical Analysis:

In this section we analyzed data with graphically. Data are explained with heart disease (Orange color) and without heart disease (Blue color).

```
[ ] plt.figure(figsize=(15, 15))
```

```
for i, column in enumerate(continous_val, 1):
    plt.subplot(3, 2, i)
    df[df["target"] == 0][column].hist(bins=35, color='blue', label='Have Heart Disease = NO', alpha=0.6)
    df[df["target"] == 1][column].hist(bins=35, color='red', label='Have Heart Disease = YES', alpha=0.6)
    plt.legend()
    plt.xlabel(column)
```

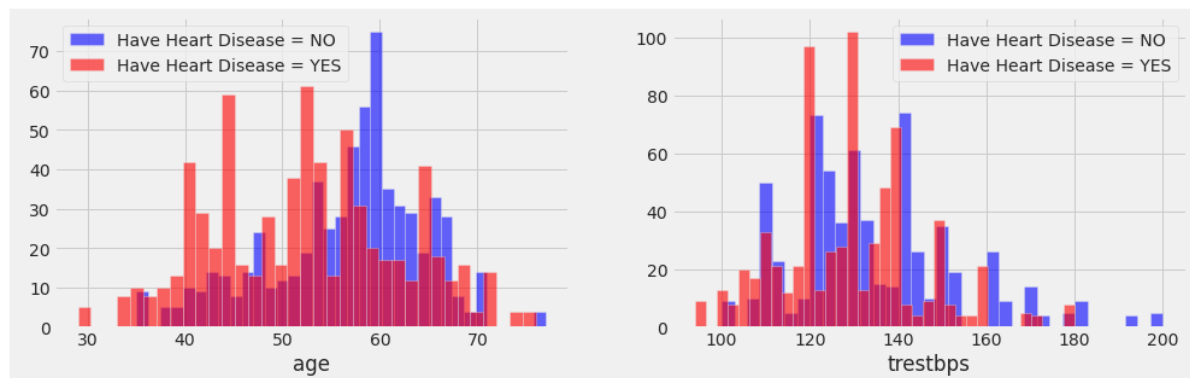


Figure: 3.2: Bar chart comparison between target & input value - 1

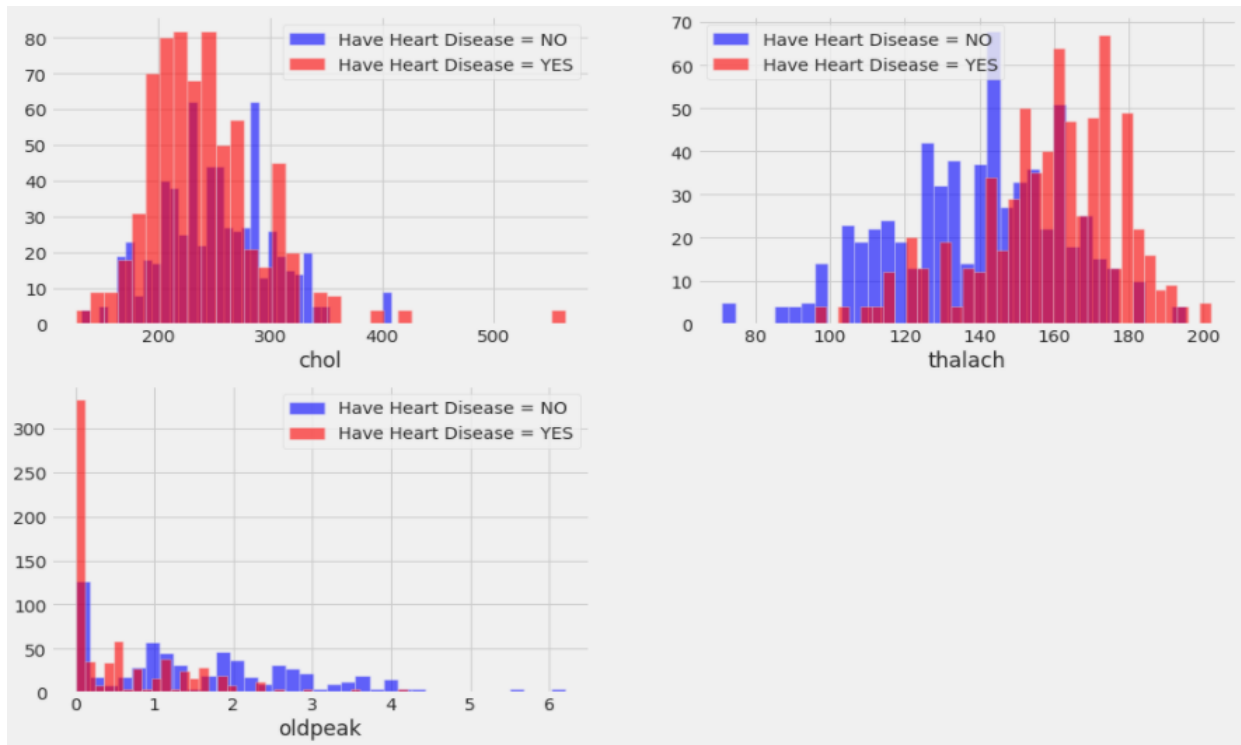


Figure: 3.3: Bar chart comparison between target & input value - 2

Here, we analyzed with six parameters such, age, trestbps, chol, thalach, oldpeak etc.

Trestbps: resting blood pressure (in mm Hg should be hospitalized) if above 130-140 is materially cause for anxiety.

Chol {serum cholesterol in mg/dl}: above 200 is cause for anxiety.

Thalach {maximum heart rate achieved}: People who attained a maximum more than 140 are more likely to have heart disease.

Oldpeak: ST depression induced by exercise relative to rest looks at stress of heart during exercise unhealthy heart will stress more.

Age vs. Max. Heart disease rate for Heart disease

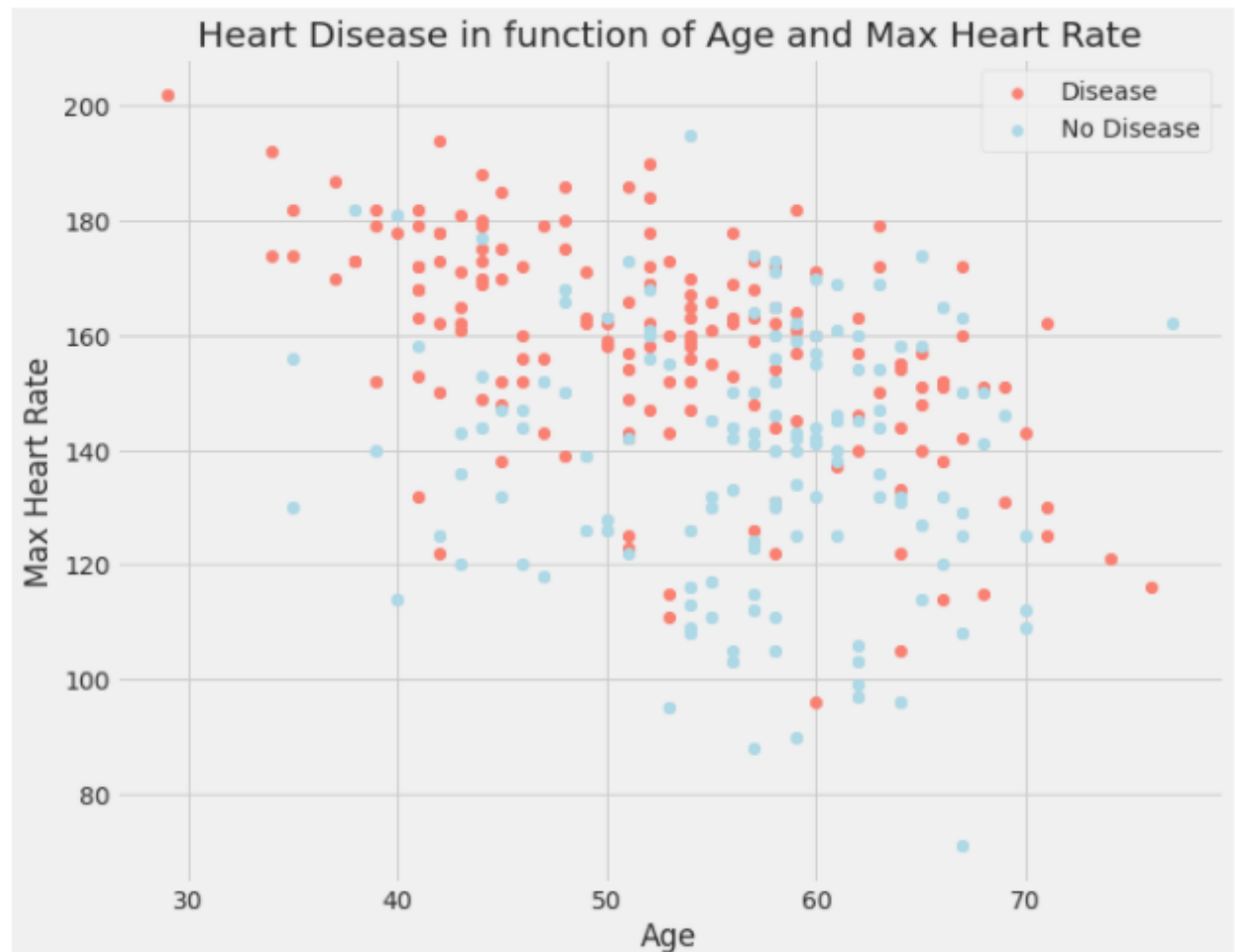


Figure: 3.4: Point graphs between Max heart rate vs final output.

Herein, contrast of 40-50 age max. Heart disease is too high than others. Every graphical analysis based on number of patients vs. Different types of parameters. After analyzing all data, we formed it a shape as input to different type of methods. And finally, we obtained 97.99% prediction by using Random forest classifier.

```
[ ] results_df
```

	Model	Training Accuracy %	Testing Accuracy %
0	K-nearest neighbors	97.52	88.97
1	Decision Tree Classifier	100.00	96.99
2	Logistic Regression	87.84	84.21
3	Support Vector Machine	96.99	93.23
4	Random Forest Classifier	100.00	97.99
5	XGBoost Classifier	98.82	96.24

Figure: 3.5: Result of all algorithm.

CHAPTER 4

REQUIREMENT SPECIFICATION

4.1 Business Process Modelling:

Figure 4.1 known as Dataflow Diagram. In this diagram used to describe every steps of this work. It is describe the inputs, data processing, result and the internal workflow of this work.

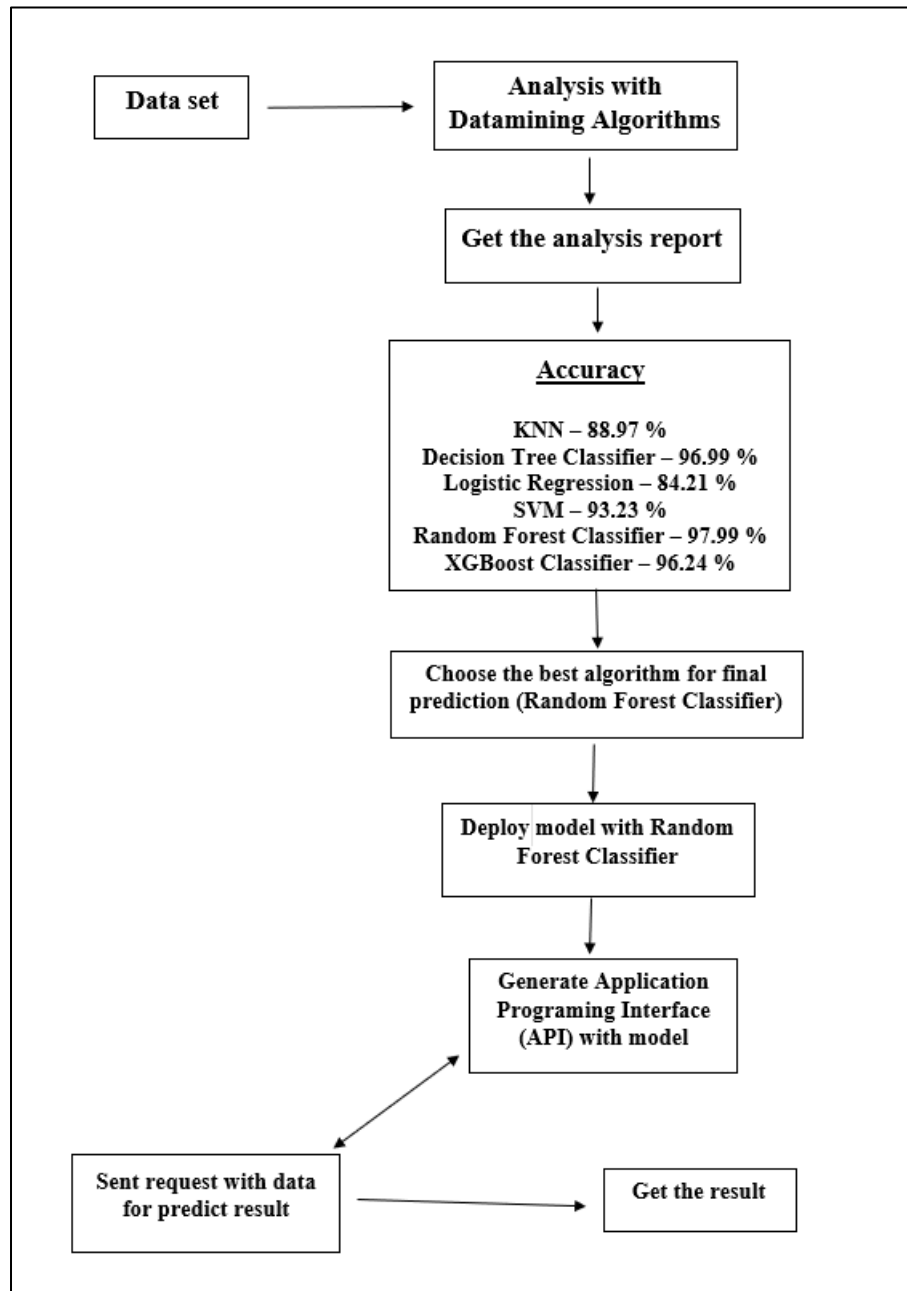


Figure: 4.1: Data Flow Diagram

4.2 Use case Diagram:

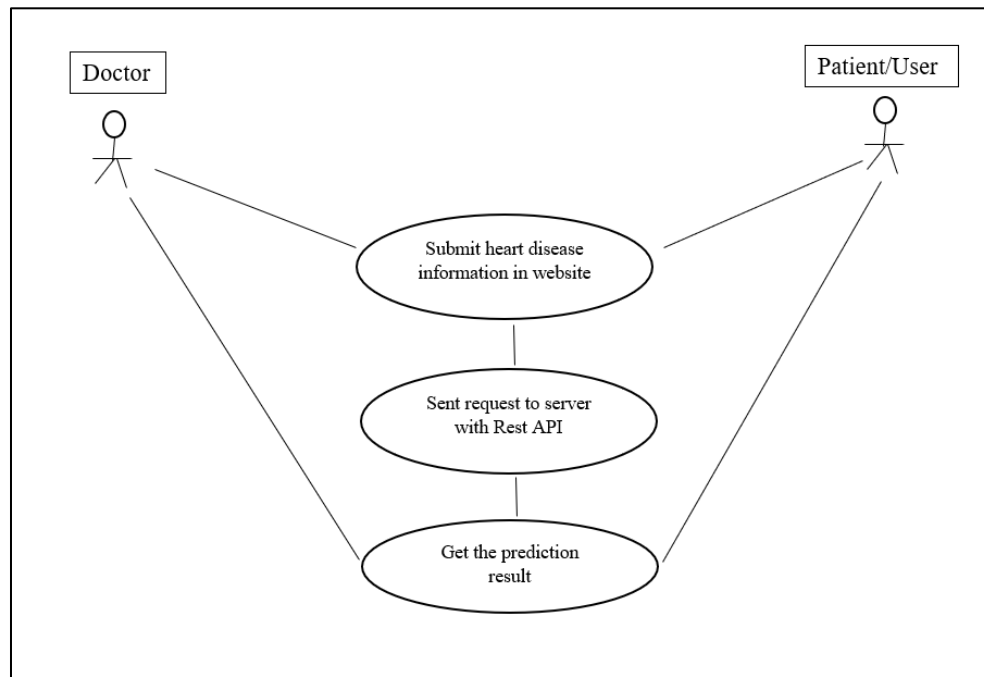


Figure: 4.2: Use case diagram

4.3 Design Requirements:

For analysis the data need the tested report data of patient. This is the most important requirement. There are some field of this data set which can not give without test health like cholesterol, blood sugar level etc.

CHAPTER 5

DESIGN SPECIFICATION

5.1 Front-End Design:

For front end design we use html, CSS, java script. Where user can submit the inputs for predict the result.

Heart Disease Prediction

Your age (Ex: 25)
Enter age

Gender

Select Gender

Chest Pain

Select Chest Pain

Resting Blood Pressure(in mm Hg)
Enter blood pressure(Ex: 125mm)

Serum Cholestoral(in mg/dl)
Enter Serum Cholestoral(Ex: 212mm)

Fasting Blood Sugar(fasting blood sugar > 120 mg/dl)

Select Blood Sugar

Resting Electrocardiographic Results

Select Resting Electrocardiographic Results

Maximum Heart Rate Achieved
Enter Maximum Heart Rate Achieved

Figure 5.1: Web page design screen - 1

Exercise induced angina

Select Exercise induced angina

ST depression induced by exercise relative to rest

Enter value (Ex: 1.0,3.1)

The slope of the peak exercise ST segment

Select slope of the peak exercise ST segment

Number of major vessels

Select Number of major vessels

Displays the thalassemia

Select thalassemia

Submit

Figure 5.2: Web page design screen - 2

5.2 Back-End Design:

Our model is trained by “Random Forest” Algorithm. From Statistical Analysis we can see that the “Random Forest” Algorithm is given the best accuracy.

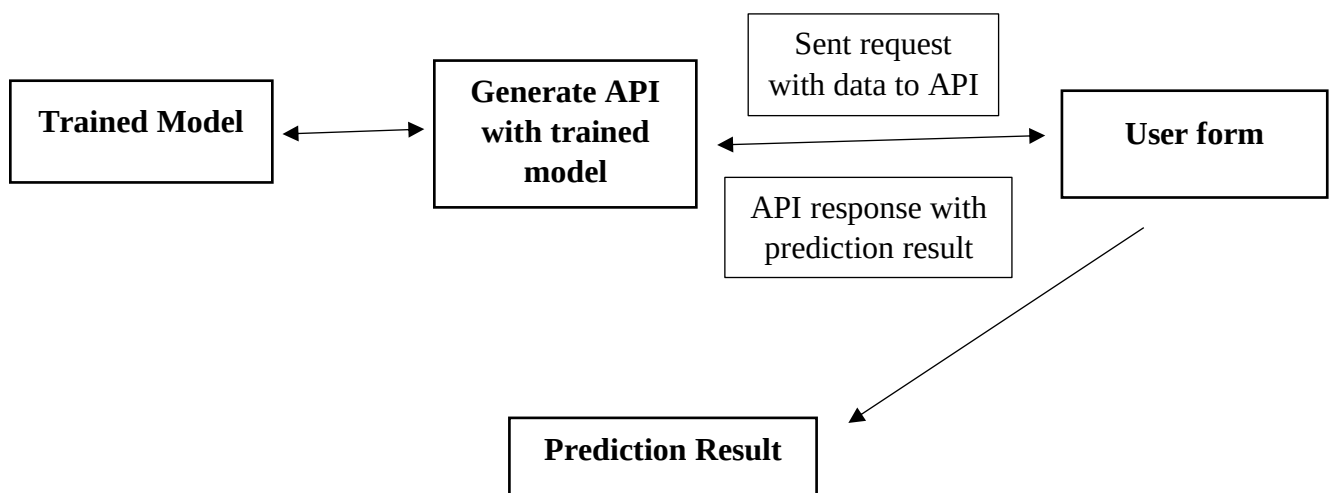


Figure: 5.2: Back end flow

5.3 Implementation Requirements:

To implement this system, we required “Google Colab” or same kind of system which support python for create the trained model. And required Visual Studio Code for generate API for connect with server. And Build a web page which keep connect with API to predict the result.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE OF DEVELOPMENT

6.1 Discussion and Conclusion:

There may be multiple type of reasons for the heart diseases. In that, major citable matters are high blood pressure, high cholesterol level, and smoking etc. And the bad food habitudes responsible behind the high blood pressure and high cholesterol level. For the sake of identifying heart disease we equipped an algorithm to predict heart disease accuracy. As they can be notified about the disease level. Nowadays, people are adamant about their badly lifestyle also, such sleeping, poor dieting, drinking alcohol etc. So, people must be abandoned these bad habitudes as soon as possible to get rid of the disease. Concededly I can say if you looked after yourself properly, your heart also looked after you.

6.2 Scope of Further Developments:

We accomplished whole work by python to predict the disease. We gathered all data from Kaggle to utilize rightly. We didn't collect data physically from the hospital due to covid-19 but in future there have a scope to collect data from the hospital. Though we turned our best prediction methodology (Random forest classifier) into a web-based system to know the prediction of heart disease. But in future there have a scope to make it into a mobile application also to rehash more user friendly. Furthermore, after testifying all the process of heart disease prediction, we can give a shape it to a business category adding the different type of features by the doctors.

References

1. J.Vijayashree and N.Ch.SrimanNarayanaIyengar, “ Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review”, International Journal of Bio-Science and Bio-Technology(2016)
2. Monther Tarawneh* and Ossama Embarak, “Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques”, Acta Scientific Nutritional Health 3.7 (2019): 147-151.
3. Ajit Solanki, Mehul P. Barot, “Study of Heart Disease Diagnosis by Comparing Various Classification Algorithms”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-2S2, January 2019
4. Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, “Design and Implementing Heart Disease Prediction Using Naives Bayesian”, Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8
5. Sujata Joshi and Mydhili K. Nair, “Prediction of Heart Disease Using Classification Based Data Mining Techniques”, Springer India 2015. L.C. Jain et al. (eds.), Computational Intelligence in Data Mining - Volume 2, Smart Innovation, Systems and Technologies 32, DOI 10.1007/978-81-322-2208-8_46
6. M. Raihan, Saikat Mondal, Arun More, Md. Omar Faruqe Sagor, Gopal Sikder, Mahbub Arab Majumder, Mohammad Abdullah Al Manjur and Kushal Ghosh, “Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design”, 19th International Conference on Computer and Information Technology, December 18-20, 2016, North South University, Dhaka, Bangladesh

Plagiarism Result

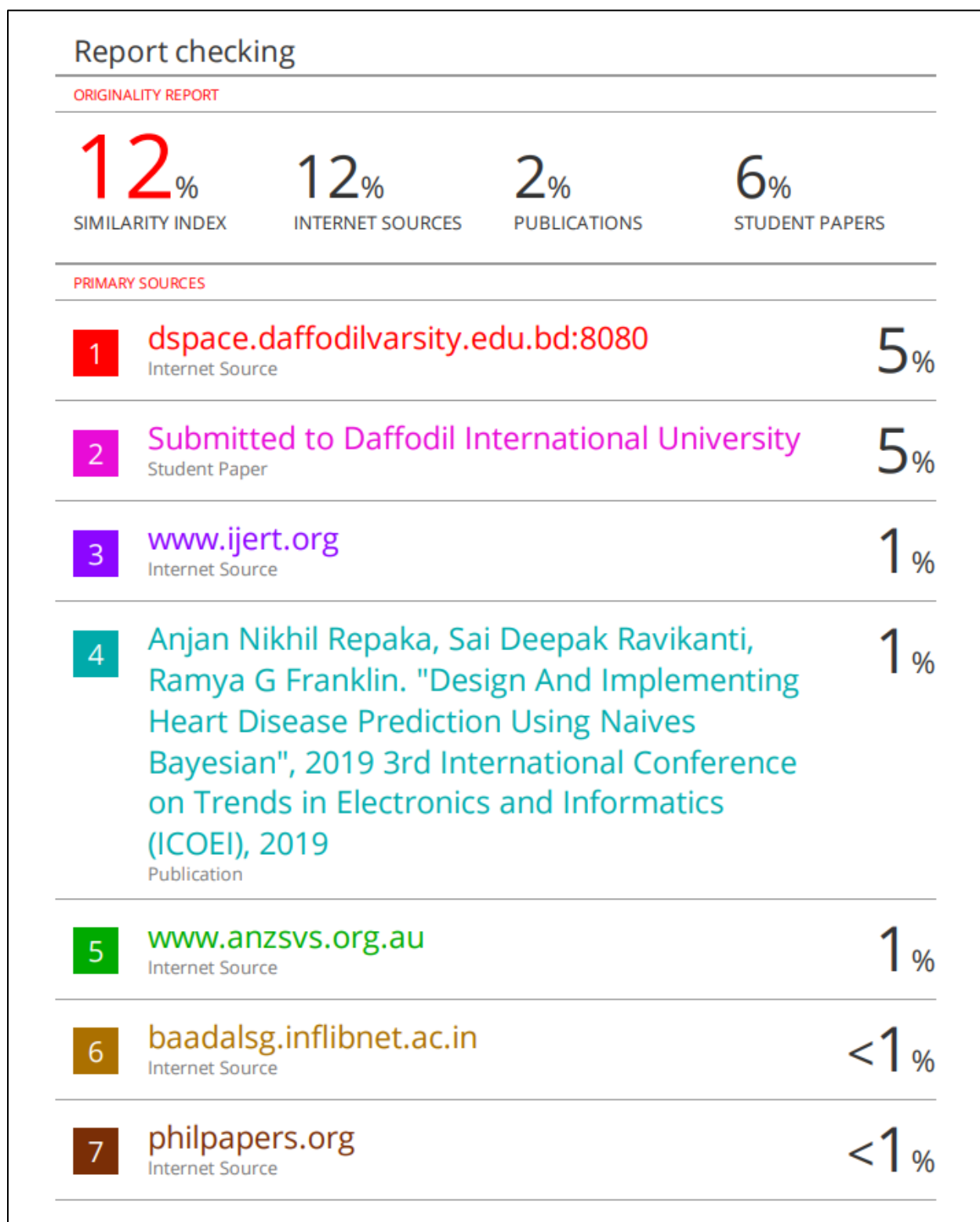


Figure: Plagiarism result