HEART DISEASE CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

SUBMITTED BY

Mahfuzur Rahman
ID: 0242310005103043

SUPERVISED BY

Dr. Md Zahid Hasan
Associate Professor
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

JULY, 2024

APPROVAL

This Project/Internship titled **"Heart Disease Classification Using Machine Learning Algorithms"**, submitted by **Mahfuzur Rahman,** ID No: **0242310005103043;** 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 M.Sc. in Computer Science and Engineering (Major In Data Science) and approved as to its style and contents. The presentation has been held on **6 July 2024**.

BOARD OF EXAMINERS

Narayan Ranjan Chakraborty
Associate Professor and Associate Head — **Chairman**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Sadekur Rahman
Assistant Professor — **Internal Examiner**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

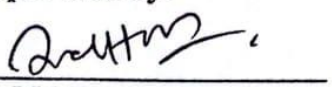
Dr. Ohidujjaman, PhD
Assistant Professor — **Internal Examiner**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Sadat Hossain
Data Scientist — **External Examiner**
Risk Management Division,
BRAC Bank Limited

DECLARATION

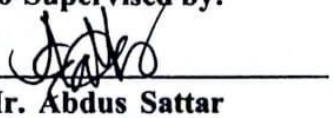
I hereby declare that, this project has been done by us under the supervision of **Dr. Md Zahid Hasan, Associate Professor, Department of CSE Daffodil International University**. I also declare thatneither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



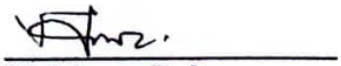
Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:



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

Submitted by:



Mahfuzur Rahman
ID: 0242310005103043
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish our profound our indebtedness to **Dr. Md. Zahid Hasan, Associate Professor, Department of CSE, Daffodil International University**. Deep Knowledge & keen interest of oursupervisor in the field of "**Machine learning**" 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 possibleto complete this project.

I 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.

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

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

Abstract:

Heart illness, or cardiovascular disease, is one of the main medical concerns at the moment. The purpose of my research, " Heart Disease classification using Machine learning Algorithms" is to draw attention to the risk factors for heart disease. To reach the maximum accuracy in heart disease prediction, I analyze numerous patient variables using multiple algorithms. Logistic Regression, K-Nearest Neighbours Classifier, Support Vector machine, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier are some of the algorithms that were used. With an accuracy of 97.99%, the Random Forest Classifier was the most accurate of them. People are able to make well-informed judgments about their next actions for more successful therapy because of this high predictive accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction:

Heart disease is a global ailment that is both common and concerning. I conducted a thorough analysis to enhance the heart disease prediction. A healthy heart is essential to human health, and heart attacks can result from a number of cardiovascular conditions. A stroke can happen after a heart attack because of uneven or ineffective blood flow. My study included a wide range of characteristics, such as age, sex, type of chest pain (cp), serum cholesterol levels (chol in mg/dl), fasting blood sugar (fbs > 120 mg/dl, with > 126 mg/dl indicating diabetes), Oldpeak, the slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), results of the thalium stress test (thal), and the target variable indicating the presence of disease (1 = yes, 0 = no).

Heart disease-related chest pain can appear as a typical angina (not related to the heart), non-anginal discomfort (mostly caused by esophageal spasms), classic angina (caused by decreased blood flow to the heart), and asymptomatic. High blood pressure and various associated disorders, as well as high serum cholesterol, are important risk factors for heart disease.

1.2 Motivation:

In the medical community, diagnosing heart disease is difficult and expensive, especially for those from lower middle-class backgrounds. Our goal is to use Machine Learning techniques to solve this problem. Unaware that they have cardiovascular disease, many people suffer from it. It is difficult to identify cardiac disease before symptoms appear. Early detection is further complicated by the fact that some people are born with congenital cardiac abnormalities. Many tests are usually needed to diagnose cardiac disease, which can be costly and challenging. In order to address this, I will accurately forecast heart illness by compiling a complete set of data on the symptoms of cardiovascular disease. For individuals who require therapy, this will make it easier for them to get it.

1.3 Rationale of the Study:

- I must compile thorough statistics regarding cardiac illness.
- I will use algorithms to predict the danger of heart disease.
- Patients who understand their illness and seek appropriate treatment can benefit from this information.

1.4 Expected Outcomes:

- Achieve the best prediction accuracy for cardiac disease.
- Have a solid understanding of the stages of cardiac disease.
- Analyze the outcomes of heart disease prognoses.
- Inform those who have been predicted about heart disease.

CHAPTER 2

BACKGROUND

2.1 Terminologies:

With confidence, I can say that my main goal is to forecast heart disease in individuals and determine whether or not they have the ailment. When patients attend the hospital informed about the stage of their cardiac illness, they can receive more precise information and, as a result, receive better care.

2.2 Related Works:

1. J. Vijayashree and N. Ch. Sriman Narayana Iyengar investigated a range of hybrid intelligence methods and data mining algorithms. They produced a basic database methodology and gave an overview of the predictions made using various techniques. They also contrasted the accuracy of previous studies' predictions. [1]

2. Ossama Embarak and Monther Tarawneh created a hybrid model that combined multiple data mining strategies. KNN (77.0%), J 4.8 (84.0%), Genetic Algorithm (84.0%), Decision Tree (78.0%), ANN (85.0%), SVM (86.0%), and Naïve Bayes (87.0%), were the components of the hybrid algorithm. Their hybrid technique yielded an accuracy of 89.20% in the end. [2]

3. Among the methods used were Sequential Minimal Optimization (SMO) (84.070%), Bayes Net (BN) (81.110%), Multi-Layer Perceptron (MLP) (77.40%), and Naive Bayesian (NB) (89.770%) by Anjan Nikhil Repaka, Sai Deepak Ravikanti, and Ramya G. Franklin. For data protection, they also used the Advanced Encryption Standard (AES). The Naive Bayesian approach was chosen for the final forecast since it produced the best accuracy, at 89.770%. [4]

2.3 Comparative Analysis and Summary:

According to the aforementioned work [1], J. Vijayashree and N. Ch. SrimanNarayana Iyengar only summarizes a number of studies and compares various kinds of work in this area; they don't analyze any data sets after that.Although Monther Tarawneh and Ossama Embarak's hybrid model claims an accuracy of 89.20%, our suggested study will yield a higher result. [3] Anjan Nikhil Repaka, Sai Deepak Ravikanti, and Ramya G. Franklin found that the Navies Bayesian algorithm yielded the best results (89.77%), where my Random Forest Classifier produced the best results (97.994%).

2.4 Scope of the Problem:

There are several opportunities for data collection in the field of medicine. But for this research, I used data from the sizable data resource Kaggle. In order to produce precise predictions, I carefully selected and arranged a variety of data kinds. I've completed multiple training courses to obtain unprocessed data from medical websites. I'm still determined to keep working toward getting real-time data in spite of obstacles.

2.1 Challenges:

- Dwelling extensively on the subject of data mining.
- Seeking the best possible accuracy when forecasting results using a lot of data.
- Making the switch to implementing the complete system through a mobile-apps.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation:

Heart disease is a serious issue that primarily affects the elderly. I did a lot of study on the subject to help spread awareness about heart illness and encourage healing. I made a number of efforts to inform people about cardiac issues in order to make treatment easier. As a result, I created a website and a heart disease prediction system. I implemented everything on a Python platform and processed all the data related to heart illness using Google Colab to get greater accuracy. In order to improve the system's usability for patients, I also developed a mobile application.

.

3.2 Data Collection:

Typically, I use the Kaggle platform to gather worldwide data. I was unable to obtain information directly from hospitals due to medical constraints. As a result, I was able to get from Kaggle over 1,387 heart disease records. Because my data pattern is varied and based on reliable sources, I was also unable to collect data online utilizing tools like Google Forms.

3.3 Detailed Methodology:

I used roughly six methods in order to explore my dataset.

- Logistic Regression
- K-Nearest Neighbours Classifier
- Support Vector machine
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier

3.3.1 Data Processing:

I divided the data into groups based on whether the patient had heart disease or not after first merging it from the Kaggle website. I have balanced my dataset by include data for 691 people with cardiac disease and 637 people without it.

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

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

data.target.value_counts().plot(kind="bar", color =["red", "green"])

<Axes: xlabel='target'>
```

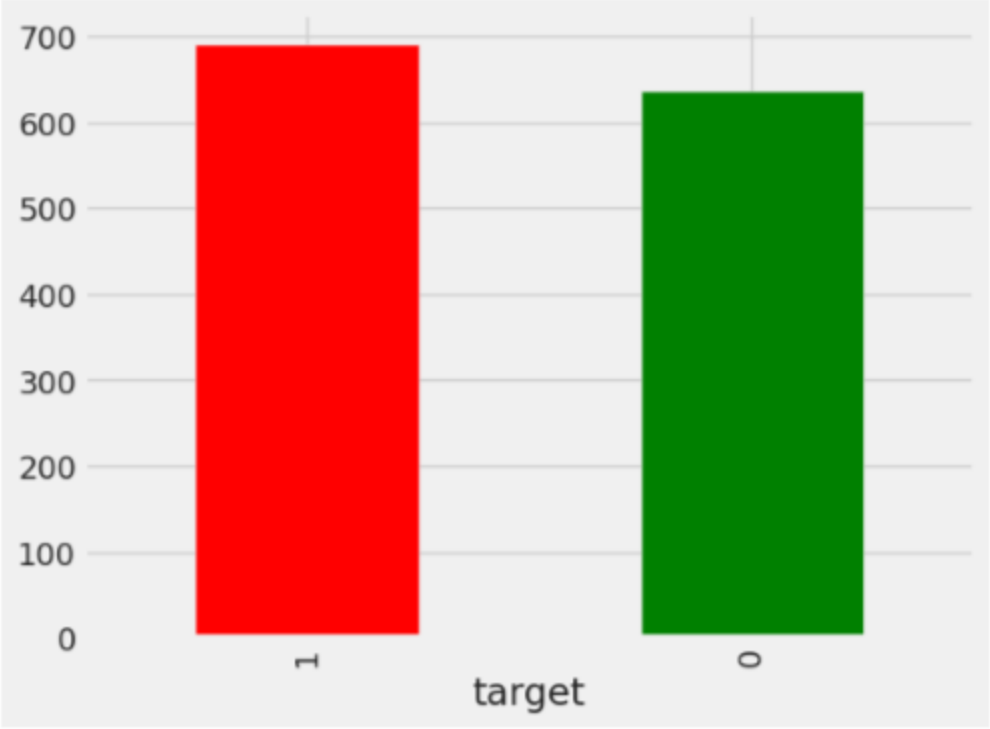


Fig: 3.1: Target column graph (yes-1, no-0

In this analysis, I defined a goal value of '1' for cardiac disease and '0' for no disease. Examining the dataset, I saw that I needed to scale all of the values and make some category variables dummy variables before training the models based on machine learning. Using the `get_dummies` method, I will first create dummy columns for the category variables. I split the training trial before that. I combined all of our unique characteristics into a single variable, which I then converted into the anticipated class.

I processed the data using six main approaches, using each one step-by-step to improve the forecast accuracy. I discovered that Random Forest Classifier produced the best accuracy, at 97.994%, after processing the data using every method available.

3.1 Statistical Analysis:

I performed a visual analysis of the data in this area. The data is shown in red for those having heart disease and green for those who do not.

```
[15] plt.figure(figsize=(14, 14))

     for i, column in enumerate(continous_val, 1):
         plt.subplot(3, 3, i)
         data[data["target"] == 0][column].hist(bins=34, color='green', label='Have Heart Disease = NO', alpha=0.5)
         data[data["target"] == 1][column].hist(bins=34, color='red', label='Have Heart Disease = YES', alpha=0.5)
         plt.legend()
         plt.xlabel(column)
```

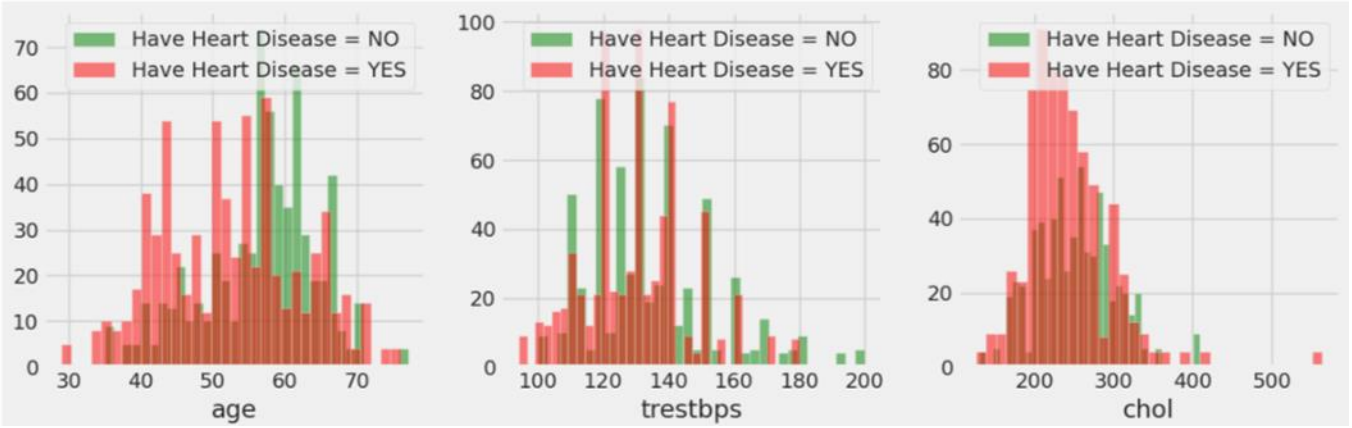


Fig 3.2: Target and input value comparison in a bar chart – 1

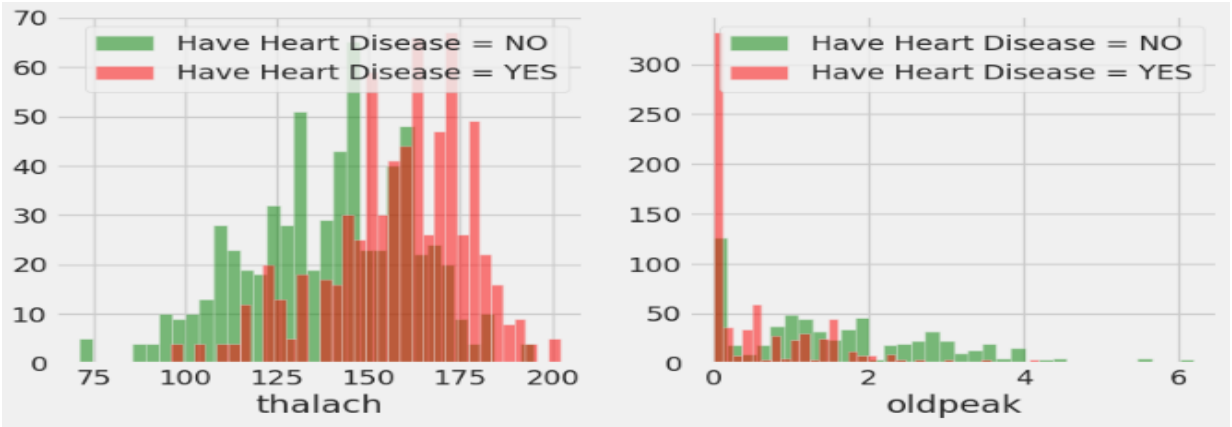


Fig 3.2: Target and input value comparison in a bar chart – 2

Six parameters were analyzed in this analysis: age, thalach, trestbps, oldpeak, chol and others.

Trestbps (resting blood pressure in millimeter-Hg): Readings between 130 and 140 are seriously concerning and may necessitate hospitalization.

Chol (serum cholesterol in milligrams/dl): Values more than 200 should be taken seriously.

Thalach (highest heart rate achieved): People who have a higher risk of heart disease if their maximum heart rate is greater than 140.

Oldpeak: Exercise puts stress on the heart, with diseased hearts under higher stress, as evidenced by ST depression brought on by activity compared to rest.

Age in Relation to Max. Heart disease occurrence rate

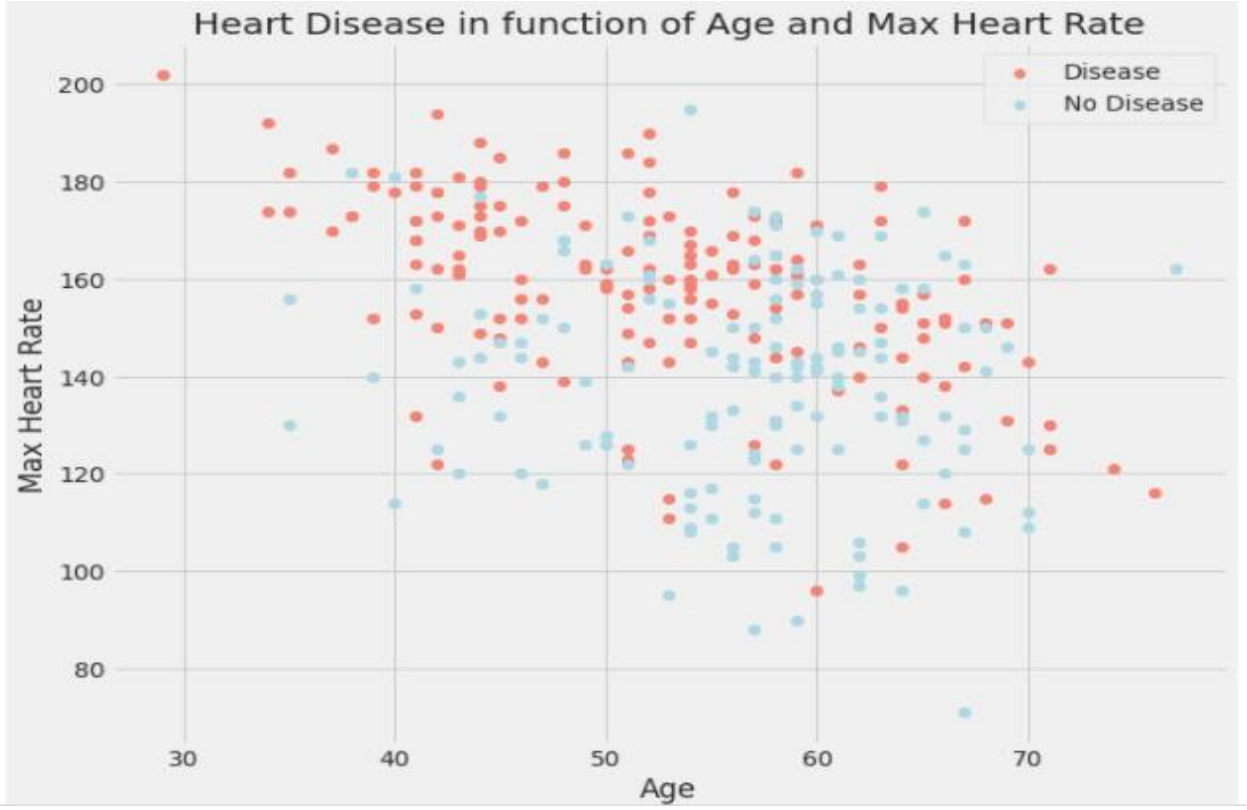


Fig 3.4. Point graphs that compare the maximal heart rate to the total output

According to this data, people in the 40–50 age range have a noticeably higher incidence of heart disease than people in other age groups. Every graphical analysis compares various metrics with the number of patients. I formatted all of the data as input for different techniques after I had analyzed it all. Ultimately, I used Random Forest Classifier to get a prediction accuracy of 97.994%.

```
results_df
```

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	87.836383	84.210526
1	K-nearest neighbors	97.524220	88.972431
2	Support Vector Machine	96.986006	93.233083
3	Decision Tree Classifier	100.000000	96.992481
4	Random Forest Classifier	100.000000	97.994987
5	XGBoost Classifier	100.000000	96.992481

Figure: 3.5: All algorithm results

CHAPTER 4

REQUIREMENT SPECIFICATION

4.1 Business Process Modelling:

Figure 4.1 I refer to the Dataflow Diagram. This flowchart is used to explain every stage of the procedure, including the inputs, data processing, outputs, and internal project workflow.

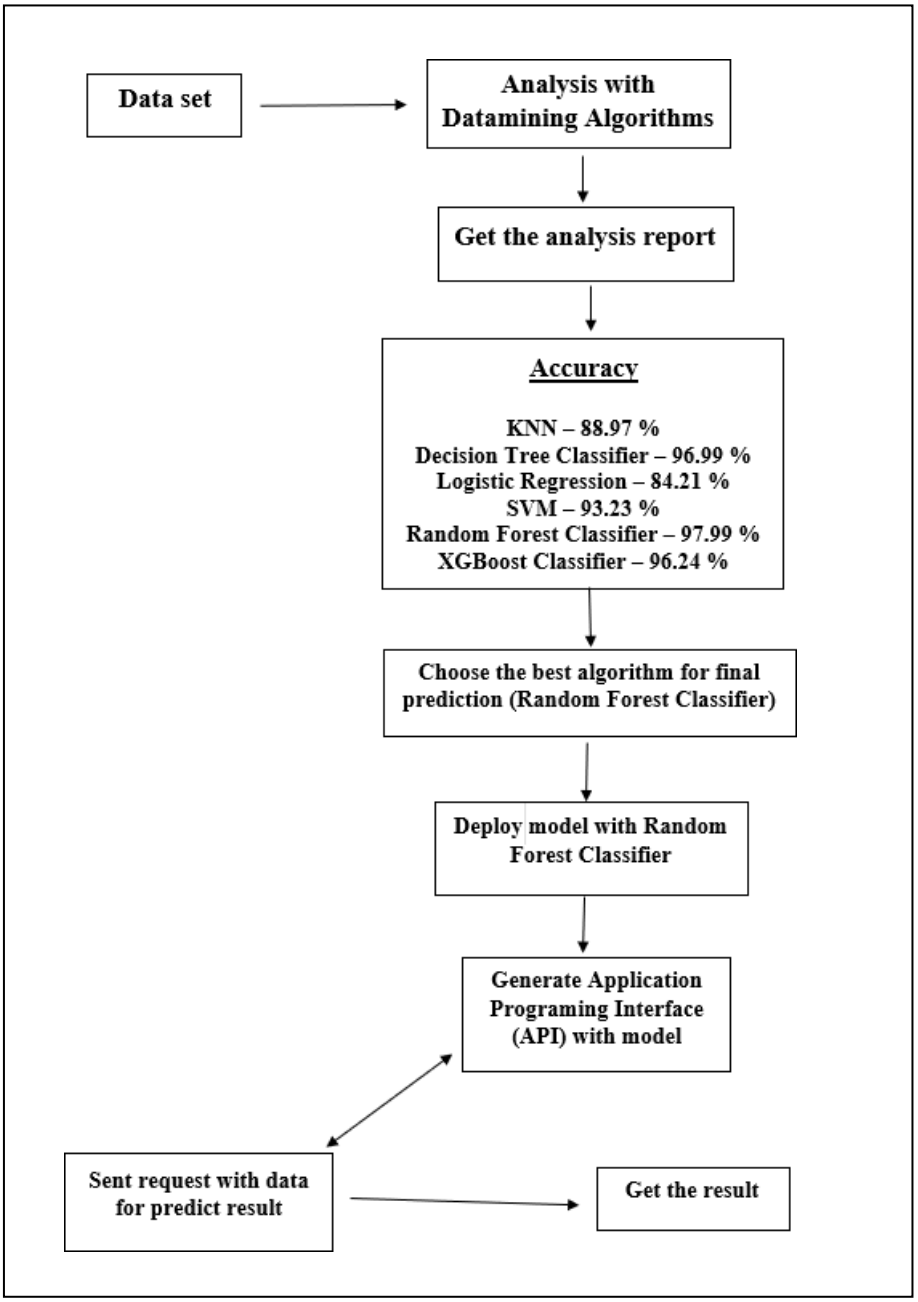


Figure: 4.1: Data Flow Diagram

4.2 Use case Diagram:

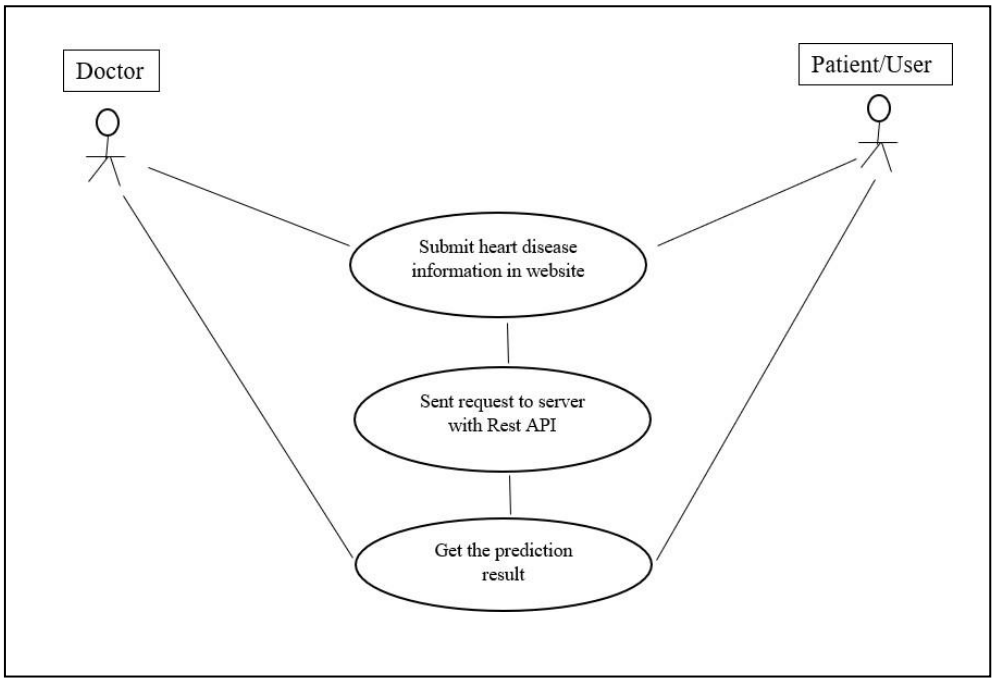


Figure: 4.2: Use case diagram

4.3 Design Requirements:

The most important prerequisite for data analysis is the tested patient medical records. Health tests are required to supply certain fields in the dataset, such as blood sugar and cholesterol readings.

CHAPTER 5

DESIGN SPECIFICATION

5.1 Front-End Design:

HTML, CSS, and JavaScript are used in front-end design. where users can enter data to forecast the outcome.

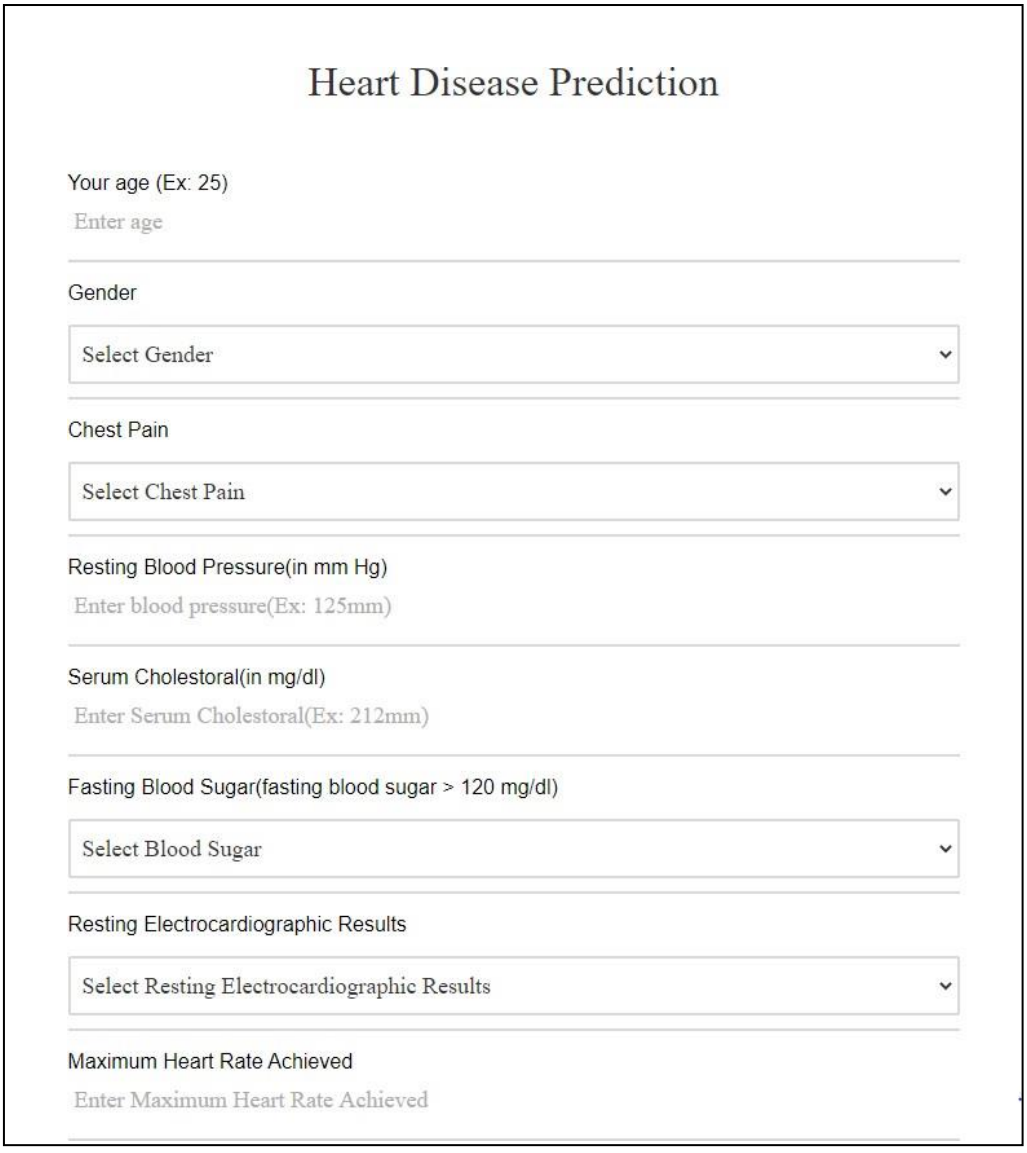


Figure 5.1: Design screen for a web page: 1

Figure 5.2: Design screen for a web page: 1

5.2 Back-End Design:

Using the "Random Forest" algorithm, my model gets trained. I can observe from statistical analysis that the "Random Forest" algorithm has the highest accuracy

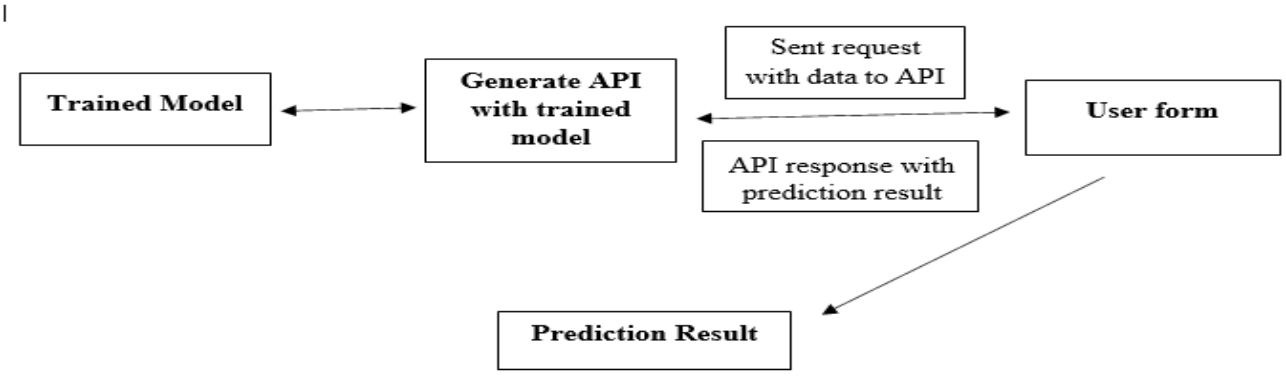


Figure: 5.2: Back end flow

5.3 Implementation Requirements:

I needed "Google Colab" or a comparable platform that supports Python for training models in order to create this solution. Visual Studio Code was also required in order to construct the server connection API. After that, a webpage was created to communicate with the API and forecast the outcomes.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE OF DEVELOPMENT

6.1 Discussion and Conclusion:

Many variables can lead to heart disease, but the primary ones include excessive blood pressure, high cholesterol, and smoking. Poor eating behaviors are often the cause of high blood pressure and cholesterol.I created an algorithm to precisely forecast the existence of heart disease, which will help identify the condition and inform people about their status. Many people today lead unhealthy lifestyles that include drinking alcohol, eating poorly, and having trouble sleeping. In order to prevent heart disease, people must give up these dangerous habits. I really think that when you take care of yourself, your heart will take care of you.

6.2 Scope of Further Developments:

Using Python, I finished the entire project to forecast heart illness. For precise use, I got all of my data from Kaggle. I was unable to get information directly from hospitals due to medical restrictions, but hospital data gathering may be possible in the future. Even though the Random Forest Classifier was my best predictor, there's still room to improve with a more user-friendly mobile application and a web-based solution for heart disease prediction. Additionally, by adding different variables recommended by medical professionals, the technique of predicting cardiac disease can be validated and turned into a business model.

References

1. J. Vijayashree and N. Ch. Sriman Narayana Iyengar, “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.

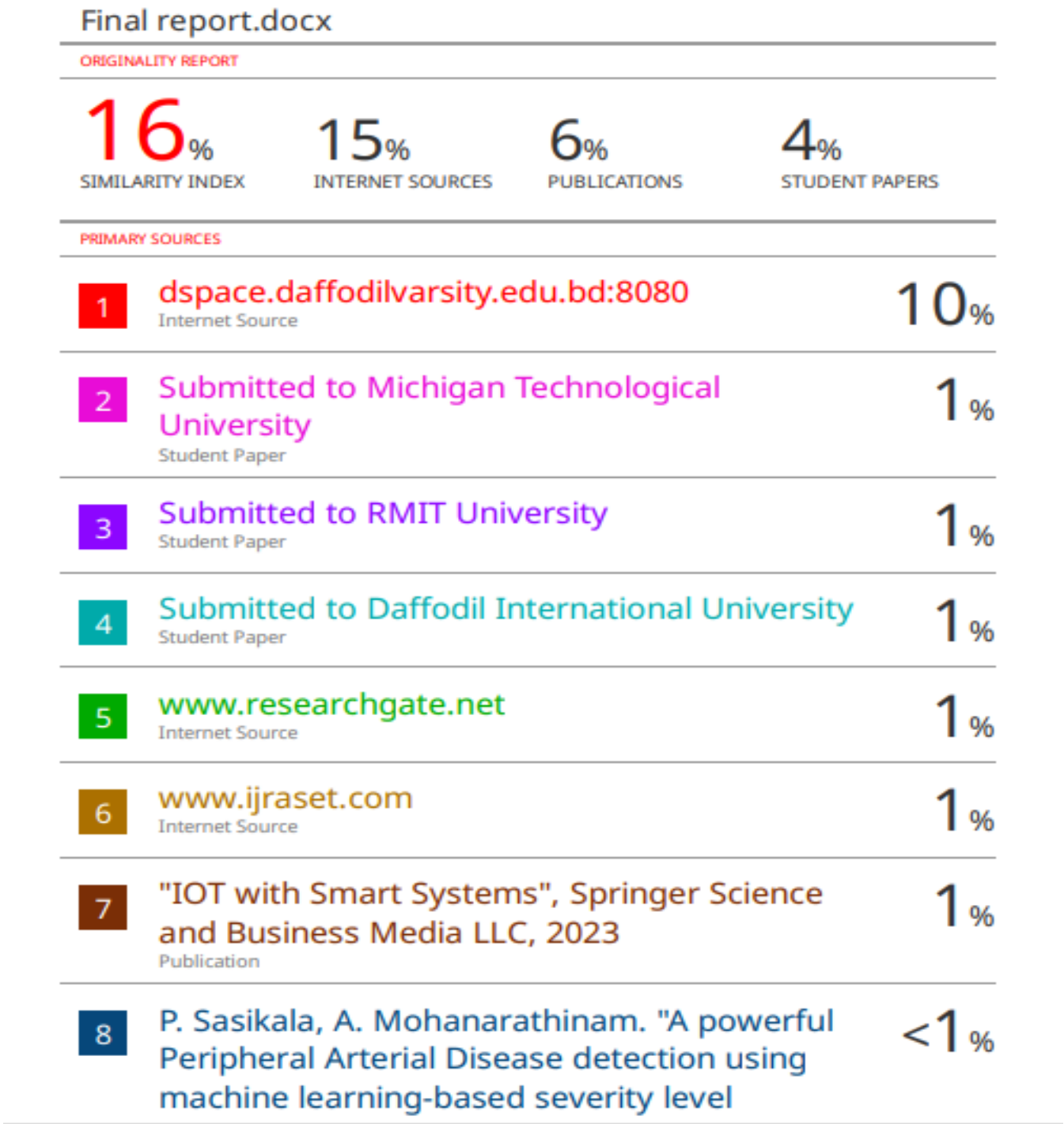
3. Ajit Solanki and Mehul P. Barot, “Study of Heart Disease Diagnosis by Comparing Various Classification Algorithms,” International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958, Volume-8, Issue-2S2, January 2019.

4. Anjan Nikhil Repaka, Sai Deepak Ravikanti, and Ramya G. Franklin, “Design and Implementing Heart Disease Prediction Using Naive 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



Final report.docx

ORIGINALITY REPORT

16%	15%	6%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	10%
2	Submitted to Michigan Technological University Student Paper	1%
3	Submitted to RMIT University Student Paper	1%
4	Submitted to Daffodil International University Student Paper	1%
5	www.researchgate.net Internet Source	1%
6	www.ijraset.com Internet Source	1%
7	"IOT with Smart Systems", Springer Science and Business Media LLC, 2023 Publication	1%
8	P. Sasikala, A. Mohanarathinam. "A powerful Peripheral Arterial Disease detection using machine learning-based severity level	<1%

Figure: Plagiarism result