HEART DISEASE PREDICTION USING DATA MINING

BY

SADIA TAMIM DIP
ID: 171-15-8638

KANIJ FATEMA NAMMI
ID: 171-15-8871

AND

IBRAHIM RAYHAN
ID: 171-15-8561

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

Supervised By

Gazi Zahirul Islam
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Abbas Ali Khan
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

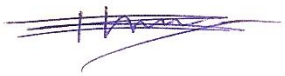
DHAKA, BANGLADESH

31 MAY 2021

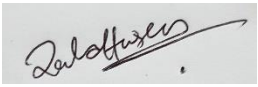
APPROVAL

This Project titled "**Heart Disease Prediction Using Data Mining**", submitted by **Sadia Tamim Dip, ID No: 171-15-8638, Kanij Fatema Nammi, ID No: 171-15-8871 and Ibrahim Rayhan, ID No: 171-15-8561** 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 **31 May, 2021**.

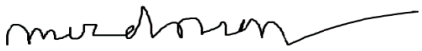
BOARD OF EXAMINERS



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



Zahid Hasan — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



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

Dr. Md Arshad Ali — **External Examiner**
Associate Professore
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Gazi Zahirul Islam, Assistant Professor, 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:

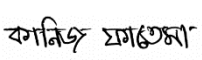
Mr. Gazi Zahirul Islam
Assistant Professore
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

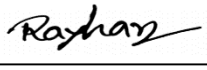
Md. Abbas Ali Khan
Sr. Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Sadia Tamim Dip
ID: 171-15-8638
Department of Computer Science and Engineering
Daffodil International University



Kanij Katema Nammi
ID: 171-15-8871
Department of Computer Science and Engineering
Daffodil International University



Ibrahim Rayhan
ID: 171-15-8561
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 project successfully.

We really grateful and wish our profound our indebtedness to **Gazi Zahirul Islam**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. 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 draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Prof. Dr. Touhid Bhuiyan, Professor, and 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

CVDs(Cardiovascular disorders) are the primary health problem, with 17.9 million deaths every year (World Health Organization). Heart disease has been the primary cause of death on a global scale for the last 20 years. It has become more difficult to diagnose illness and have adequate care at the right time as the population and disease have grown. However, medical research has advanced to the point that we can see a glimpse of hope. We primarily address it in this article. We looked at various data mining approaches, including Decision Tree Classification, Random Forest Classification, and K-Nearest Neighbor Classification, and we used a good data set of random attributes and values to achieve the highest accuracy. We are only attempting to forecast the progression of heart disease in this article. These Data Mining techniques require less time and have higher accuracy. It is used to monitor and examine the outcome of heart disease patients, with a current diagnosis ranging from in decent form to good shape. Using various data mining methods, the proposed study forecasts the likelihood of Heart Disease and classifies patients' risk levels. As a result, this report provides a comparative analysis of the success of various Data mining algorithms. As opposed to other data mining algorithms, the trial results suggest that the Random Forest and Decision tree algorithms have the best accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Background Study

Heart diseases are considered the most prevalent among all fatal diseases. It describes a range of the state that has the effects on the heart. The heart condition becomes the number one explanation for the enervation, and the death worldwide in men and girls over the old cardinal and currently, most countries face high and increasing rates of the heart condition. The heart condition is viewed as the second epidemic, replacing infectious diseases because of the leading the explanation for the death in several countries (Gale Nutrition reference, 2011).

Traditionally, it had been thought that cardiomyopathy was the matter of developed countries; however, it is currently changing into a priority for developing countries. It is particularly depilatory for the developing countries as they are doing not have enough health care. As noted by the workplace on Women's Health (2006), cardiomyopathy was thought about to be a man's drawback; however, currently, it is recognized because the favored killer of ladies, even as it's of men. One in 3 adults worldwide has raised force per unit area — a condition that causes around half all deaths from stroke and cardiomyopathy (The World Health Statistics 2012) [1]. Heart-related issues cause thirty-one p.c of all deaths worldwide [2].

Data mining plays a significant role in the healthcare industry and helps to find different skills and best practices, and makes maximum effort to make the health care system use different data. As a result, health care is increased, and costs are reduced. Opportunities to simultaneously improve care and reduce costs should account for 30% of overall healthcare (Wurz & Takala, 2006).

The significant application of knowledge mining in apparent fields like e-business, marketing, and retail has semiconductor diode to its application in alternative industries and sectors. Among these sectors, simply discovering is tending. The tending atmosphere remains knowledge poor." At intervals, the tending systems, wealth of knowledge on the

market. However, to find hidden relationships and trends within African genres, practical analysis tools are scarce.

The likelihood of someone being in danger of illness, heart condition, cardiomyopathy, and cardiovascular disease will be cut back the death rate by early-stage detection of the disease and predicting. Medical data processing techniques the square measure employed in medical knowledge to compose meaty patterns and data. Medical knowledge has abundance, the multi-attribution, unity, and complex relationship with the time. Downside victimization, the massive quantity of information effectively becomes a severe problem for the health sector.

Data mining provides the methodology and the technology to convert these knowledge mounds into helpful decision-making information. This prediction system for cardiovascular disease would facilitate Cardiologists in making faster choices so that many patients will receive treatments at intervals the shorter amount, leading to saving lots of the life [3].

1.2 Introduction

Finding the antecedents many patterns and current in databases and victimization that info to make prophetical models is the method of information mining. Data processing meld applied math analysis, machine learning, and info technology to drag out invisible patterns and relationships from giant databases. Cardiomyopathy, conjointly called disorder (CVD), encloses many conditions that influence the center — not simply heart attacks.

Heart disease was the foremost necessary clarification for casualties in many countries and the People's Republic of Bangladesh. Every 34 seconds at intervals, the USA dies because of sickness cardiovascular disease, cardiomyopathy upset; there have some categories of cardiomyopathy's like Coronary cardiovascular disease, cardiovascular disease, and disorder. One in three adults universal has increase force per unit space — therefore as that causes on every aspect 1/2 all deaths from stroke and cardiomyopathy (The World Health Statistics 2012) [1]. The term “cardiovascular disease” includes a decent vary of conditions that have a sway within the center and conjointly the blood vessels, and also the means blood is pumped-up and circulated through the body. Identification could also be a

challenging and necessary task that should be dead accurately and efficiently. The identification is usually created, supported the doctor's experience and information. This finishes up in unwanted results and extraordinarily medical costs of treatments provided to patients. Therefore, it's going to be passing helpful if there has associate degree automatic identification system.

1.3 Motivation

In the recent era, cardiovascular disease prediction is one of the foremost tortuous tasks in a medical field. The Designation is complicated and necessary task that has to be performed accurately and with efficiency. Supported doctor's expertise and information, the designation is created. This conduct to unwanted results and further medical prices of treatments provided to patients. Therefore, an associate degree automatic diagnosis system would be extraordinarily helpful. Our work experiments to gift the careful study concerning various data processing techniques which might be enlarged in these machine-controlled systems.

1.4 Objective

This study predicts whether the patient is affected by cardiovascular disease or the not victimization of the professional dataset completely different machine learning algorithms. Conclude the correlations between completely different attributes [4]. We will get the transparent plan of our planned data processing techniques, analyze the result, and compare the results of various data processing techniques. We will analyze our techniques if there is any chance to bring improvement to our results [5].

1.5 Expected Outcome

The primary goal of our study is to create an Intelligent Heart Disease Prediction System that uses a historical heart database to diagnose heart disease; as the range of people dying of heart disease increases, it is becoming increasingly important to create a system that can successfully and reliably predict heart disease. The study's goal was to discover the most effective machine learning algorithm for detecting heart diseases.

CHAPTER 2

Literature Review

2.1 Literature Review

In this chapter, we discuss previous work on heart disease and different types of machine learning algorithms. Our project will predict weather there has any heart disease or not. For this, we use different types of algorithms for prediction. Because it leads us to a better output, if we use one algorithm, we cannot compare it, so we cannot tell that it is a probable. We have chosen to use: 1. KNN, 2. Random Forest, 3. Decision Tree algorithms for our work. Now we are talking about some risk factors and previous work on heart disease.

Right, Bundle Branch Block (RBBB) may is cardiac arrhythmia within the right bundle part of the conductivity framework. Throughout accurate pack half block, the proper ventricle is not squarely initiated by driving forces researching the proper cluster branch. It's an internal organ assay. A male patient walks on an assay treadmill to possesses his heart's reason checked [13].

The implantable cardiac monitor could be a tiny device ingrained subcutaneously into the left aspect of the chest, presenting three to four years of battery life. ICMs have the advantage of long heart condition observation, letting patients self-express and record indicative events [14].

Unrecognized Myocardial Infarction is understood to represent a considerable. CAD is that the pathology directs clinical management and affects prognosis [15].

Women have different causes and risks than men in the case of HA, and for example, women have a 50% higher chance of HA than men due to depression. In 2014, about 50 thousands of women died from HA [4]. Every year 17.3 million people died from Heart Disease (HD) which ranked HD to number 1 position as a reason for the global cause of death [16].

Pneumonia is a special problem to heart patients. Even, respiratory illness could cause complications, together with microorganism respiratory illness, or the worsening of

chronic heart problems. It is a respiratory organ infection that forestalls your lungs from obtaining enough blood element, making a strain on the heart [19].

During delivery, about 1,061 women had a heart attack involved with labor works; during their pregnancy, 922 had heart attacks, and after giving birth 2,390 women had a heart attack [10].

If the harm is big enough, a heart failure will cause fast and severe left atrioventricular valve mention- abnormality of the guts muscle (cardiomyopathy). Over time, sure things, like a high vital sign, will make your heart figure more challenging, step by step enlarging your heart's ventricle. Severe disseminated multiple sclerosis (SMS) will have an effect on vas operate in numerous of how that result in abnormalities in vital sign, regular recurrence, heart rate, left cavum pulsation to operate, etc. [17].

Heart sickness is one of the diseases because that death can happen essentially, and in step with the global health organization, the rates is a lot of for that. Thus heart condition is determined for the extensive knowledge approach, and as massive knowledge is considered thus used ‘Hadoop Map’ cut back platform. For classification Improved Decision Tree and the cluster purpose K-Means algorithm and ID3 is used in the hybrid approach [18].

“This paper prediction model using their techniques and few of them also attempted by combining multiple techniques by making hybrid models in order to reach the accuracy. This research paper is to consider the work done by combining two techniques to make the hybrid model in order to predict the heart disease” [20].

Analyzing data from various perspectives and combining it into useful information is the process of data mining. This method is used for find out heart disease. The heart diseases can be defined very quickly based on risk factors. The primany purpose of this work is to evaluate different classification techniques in heart diagnosis [21].

Now, cardiopathy prediction is one of the foremost complicated tasks in the medical field. Recently, nearly one person dies per minute because of cardiopathy. This paper makes use of the cardiopathy dataset on the market in Kaggle UCI. The raised work predicts the probabilities of cardiopathy and kinds of patient's risk level by implementing totally

different data processing approach like Logistic Regression, Naive Bayes, Random Forest and Decision Tree [22].

Heart disease embraces the heart and blood vessels of individuals unique to the planet. Science Citation Index- distended (SCI-E) was accustomed pull out all papers indexed as a subject of CVD throughout 2001- 2010. Describe information showed that the number of publications within the space of vessel has inflated steadily. In search resulted in a complete range of 98143 papers within the sort of book, biography, editorial, review, meeting, article, correction, abstract, other, bibliography, news, and letter. Analysis of knowledge extracted from the info of SCI-E reported that the amount of publication within the field of CVD has inflated linear throughout 2001-2010 [23].

The main objective of this analysis paper is to summarize the recent analysis with comparable results that have been done on heart condition prediction and create analytical conclusions. From the discussion, its seen that Naive Thomas Bayes, with Genetic algorithms, call Trees, and Artificial Neural Networks techniques, improve the accuracy of the guts illness prediction system in numerous situations. Their complexities square measure summarized and unremarkably used data processing and machine learning techniques to predict the result during this paper. We'll see using Decision Tree how much better prediction we are able to get [24].

This paper's object is to observe whether or not patients have cardiovascular disease or not by a given variety of options from patients. The motivation of this work is to avoid wasting human resources in medical centers and improve the accuracy of designation. This paper use SVM, Logistic Regression, Naïve Bayes, Artificial neural network and Random Forest methods to discover heart disease. We can see how the project work with using Random Forest Algorithm [25].

2.2 Overview of Data Mining

Data mining techniques convert a large quantity of info to helpful information thanks to the complete convenience of a large quantity of information. Data processing has become well-liked in recent years. The recognition knowledge of information mining should not be a surprise since the scale of the offered data collections is much overlarge to be examined

manually. The method for automatic information analysis supported classical statistics, and machine learning typically faces issues once it processes giant, dynamic information collections consisting of complicated objects. The multiplicity of information, coupled with the necessity for powerful information analysis tools, has been explicit as a data-rich, however, information-poor scenario.

The invasive, vast quantity of knowledge, collected and hold on in massive and various knowledge repositories, transcend our human ability for comprehension while not powerful tools. As a result, the gathering of information in massive data repositories become “data tombs” that square measure seldom visited. So, vital selections square measure usually created primarily based not solely on the information-rich knowledge held on in knowledge repositories but also on call maker’s intuition, just because the decision-maker doesn't have the tools to substance the precious data embedded within the enormous amounts of knowledge.

Moreover, consider knowledgeable system technologies, which generally estimate users or domain specialists to manually input information into information bases. Sadly, this technique is one-sided biases and errors and is extremely long. Data processing tools redact knowledge analysis and will reveal necessary knowledge patterns, causative greatly to business ways, information bases, and scientific and medical analysis.

CHAPTER 3

Methodology

3.1 Methods

The training stage of data mining is followed by feature engineering, the selection of different attribute combinations and classification modeling, and the construction of data mining-based prediction models. The function selection and modeling were repeated for all combinations of attributes [6]. Latest papers, journals including articles under the parameters of computer information, science and engineering, processing of data, and cardiovascular problem have been the key methods used for our work [7]. During this work, we experimented with three algorithms: KNN, Random Forest, and Decision Tree.

3.1.1 K Nearest Neighbor

K Nearest Neighbor (KNN) is an algorithm for non-parametric machine learning. A supervised method of learning is the KNN algorithm. The KNN algorithm preserves all the cases present and classifies new ones using a resemblance measure [8]. This suggests that all the information is classified, and the algorithm learns from the input data to estimate the output K-nearest Neighbor (KNN) grouping, a more sophisticated method, considers a band of k objects from training set in order that most closely related to the test. The mark distribution is based on the primacy of a social class in this neighborhood [9]. For model building and testing, the train set is used. The classification of data is formed on different algorithms of supervised machine learning, including, KNN has been presented [10].

3.1.2 Random Forest

Algorithm of Random Forest is a user-friendly along with scalable machine learning algorithm that, in many cases, produces spectacular results without hyper-parameter tuning. It is also one of the most commonly used algorithms due to its simplicity and variety. It is suitable for classification as well as regression. Random Forest aims to synthesize leaning models, and it converts the weak model into a powerful and robust learning model [11]. Random Forest provides additional uncertainty to the model as the number of trees is increased. Instead of aiming for the most important function when

splitting a node, it searches for the best feature in a random subset of features. as a consequence, there is a wide range of options, which leads to a more robust paradigm overall.

3.1.3 Decision Tree

One of them is a decision tree, and this approaches for viewing an algorithm. It is a classic algorithm for machine learning. There are multiple of heart disease, such as nicotine, BP, cholesterol, weight, etc. The decision tree's difficulty lies in choosing the root node. It is a schematic; in this spot, the internal node indicates the characteristics of the dataset, and even the outer nodes impact on it [12]. The data must be specifically classified by this element included in the root nodes. As the root node, we make use of age. It is quick to interpret the decision tree. They are non-parametric, and function collection is implicitly achieved. This is one way to show an algorithm that only includes statements for conditional control.

3.2 Design of System

Design of system defines what information needs to be processed and how the components of data interlink. Here we describe the whole process of our work. From preprocessing to finding the best result of our work.

3.2.1 Dataset

There are 303 documents and 14 features compiled from Kaggle in this dataset. It is a traditional dataset. Our dataset link-

https://www.kaggle.com/ronitf/heart-disease-uci?fbclid=IwAR096kaZb6a7HYvmJXsNaY3f1upxrAjFn0TLqHBCz8P4RxnC9Eo3bUkhPO4#__sid=js0

3.2.2 Preprocessing

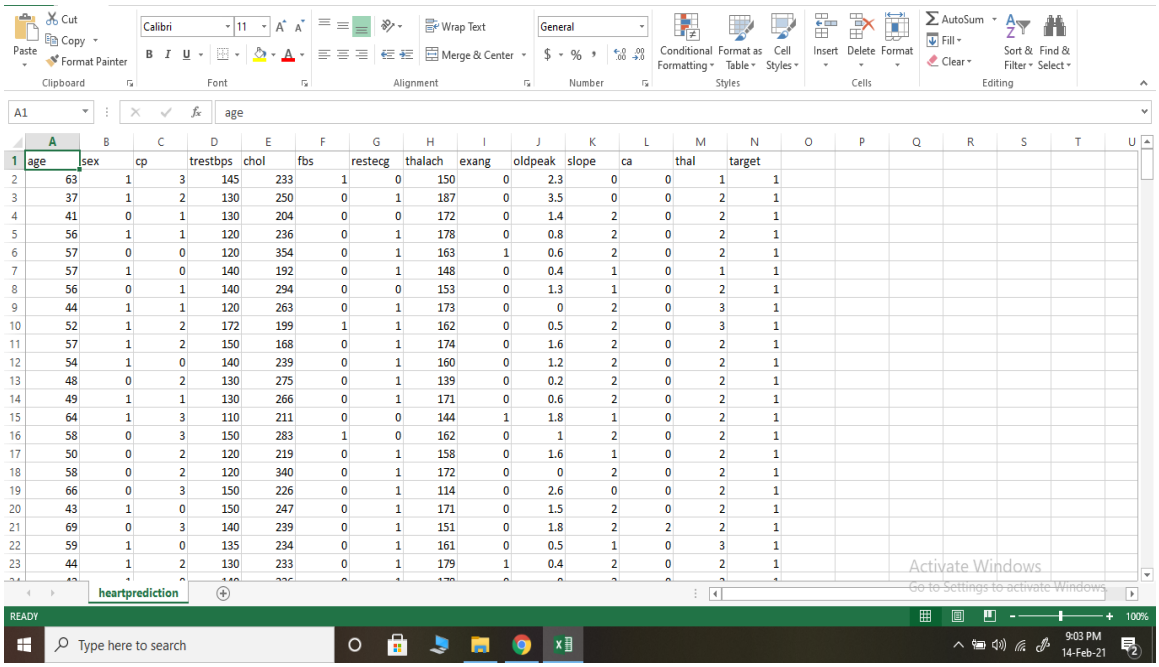
It must clean unnecessary any data from raw data and place it in a formatted manner before doing some operation with data. It has typically been used for a data mining method as an initiation stage. In our data, there is no repeat data or any NULL data. For this, we use our full dataset.

3.2.3 Load Data

Data load is the duplicate and loading data or complete sets of data to a database or related program from a source file, folder, application. It is typically used to duplicate this digital this data from any source and then paste or load it into a utility for its collection or working.

3.2.4 Analyze Feature

In our dataset, there are 14 features,11 are symptoms. All symptoms are different from each other. It also varies by person's age and gender.



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1

Figure 3.1: Showing all features

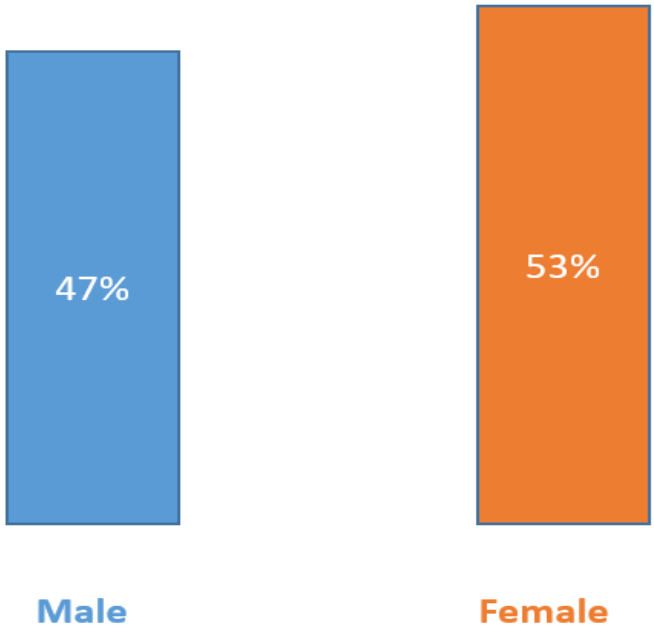


Figure 3.2: Showing percentage of gender

In our total data there have 47% of male and 53% of female (Figure 3.2.4.2).

TABLE 3.1: ATTRIBUTE AND SYMPTOMS

Attribute	Value	Symptoms
Age	29-77	Age in years
Sex	0 – male, 1- female	gender
cp	1-typical angina; 2-atypical angina 3-non-anginal pain; 4-asymptomatic	chest pain type
trestbps	Numeric value(140mm/Hg)	resting blood pressure in mm/Hg
chol	Numeric value(289mg/dl)	serum cholesterol in mg/dl
fbs	1-true, 0-false	fasting blood pressure>120mg/dl
restecg	0-normal, 1-having ST-T, 2-hypertrophy	resting electrocardiographic results
thalach	-	maximum heart rate achieved
exang	1-yes, 0-no	exercise induced angina
oldpeak	Numeric value	ST depression induced by exercise relative to rest
slope	1-upsloping, 2-flat, 3-downsloping	the slope of the peak exercise ST segment
ca	0-3 vessels	number of major vessels colored by flourosopy
thal	3-normal, 6-fixed defect, 7-reversable defect	thalassemia

3.2.5 Modeling and Prediction with Data Mining

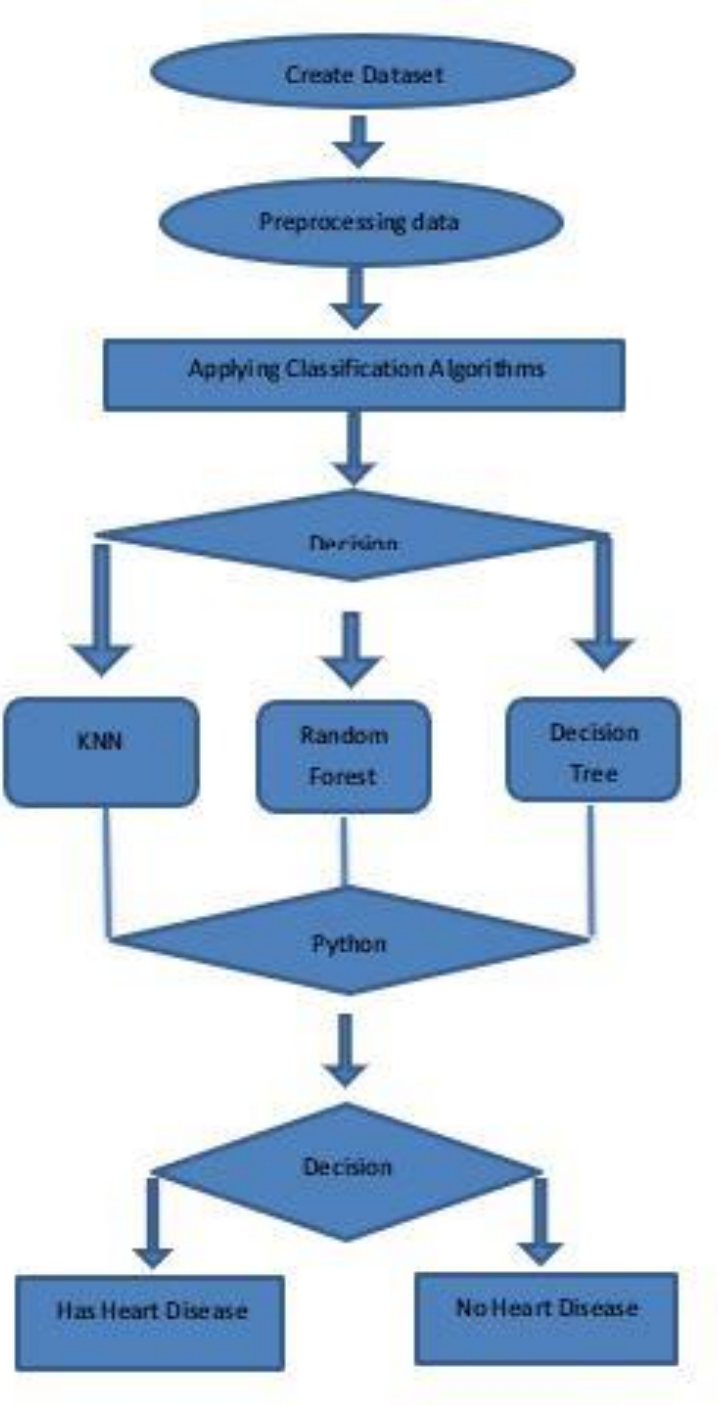


Figure 3.3: Prediction Flow Chart

3.2.6 Finding the Result

This section displays the results of using KNN, Random Forest and Decision Tree. TP (prediction is True and end is positive), FP (prediction is False and end is positive), TN (prediction is True and end is negative), FN (prediction is False and end is negative) [27].

- True Positive: predicted patient has disease and result is positive.
- False Positive: predicted patient has not disease and result is positive.
- True Negative: predicted patient has disease and result is negative.
- False Negative: predicted patient has not disease and result is negative.

The precision metric gives an accurate measure of positive analysis. The term "recall" refers to the number of accurate real positives.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

TABLE 3.2: CONFUSION MATRIX DERIVED FOR ALGORITHMS

Algorithm	**True Positive**	**False Positive**	**False Negative**	**True Negative**
KNN	22	111	16	108
Random Forest	0	127	0	130
Decision Tree	0	127	0	130

The experiments are carried out using a pre-processed dataset, in addition to the algorithms listed above are studied along with applied. The confusion matrix is used to retrieve the performance metrics discussed above. The model's performance is described by the Confusion Matrix. Table 3.2.6.1 shows the proposed model's confusion matrix for different algorithms.

CHAPTER 4

Result and Analysis

4.1 Result and Analysis

Accuracy score, Recall, Precision including F-measure are the metrics used to evaluate the algorithm's performance. The precision metric gives an accurate measure of positive analysis. The term "recall" refers to the number of accurate real positives. The F-measure is a kind of precision test [27].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + Fales\ Negative} \quad (4)$$

$$\mathrm{F-Measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

TABLE 4.1: KNN ALGORITHM ANALYSIS

Name	precision	recall	f1-score	support
Negative Level	0.87	0.83	0.85	130
Positive Level	0.83	0.87	0.85	127
Macro avg	0.85	0.85	0.85	257
Weighted avg	0.85	0.85	0.85	257

TABLE 4.2: RANDOM FOREST ALGORITHM ANALYSIS

Name	precision	recall	f1-score	support
Negative Level	1.00	1.00	1.00	130
Positive Level	1.00	1.00	1.00	127
Macro avg	1.00	1.00	1.00	257
Weighted avg	1.00	1.00	1.00	257

TABLE 4.3: DECISION TREE ALGORITHM ANALYSIS

Name	**precision**	**recall**	**f1-score**	**support**
Negative Level	1.00	1.00	1.00	130
Positive Level	1.00	1.00	1.00	127
Macro avg	1.00	1.00	1.00	257
Weighted avg	1.00	1.00	1.00	257

Model evaluation metrics such as precision and recall are fundamental. While precision mentions to the percentage of appropriate results correctly classified by algorithm, recall represents the percentage of total appropriate results correctly classified by algorithm.

4.2 Accuracy of Model with All Features

Accuracy, precision, and recall are the three major metrics used to assess a classification model (discuss section 4.1). The proportion of accurate projections for the test data is known as accuracy. It's simple to figure out by dividing the overall number of predictions by the number of accurate predictions The accuracy of a model is defined as the number of accurate predictions divided by the total number of records. Simultaneously, dataset is unbalanced; also, accuracy is not a valid indicator of model efficiency.

$$\text{KNN}= \begin{bmatrix} 108 & 22 \\ 16 & 111 \end{bmatrix} \quad (6)$$

$$\text{Random Forest}= \begin{bmatrix} 130 & 0 \\ 0 & 127 \end{bmatrix} \quad (7)$$

$$\text{Decision Tree}= \begin{bmatrix} 130 & 0 \\ 0 & 127 \end{bmatrix} \quad (8)$$

4.3 Feature Engineering

Using domain information and data mining techniques, feature engineering is the practice of extracting functionality from raw data. Machine learning algorithms may benefit from these characteristics. Feature engineering is a form of data mining that is used in real-world situations. It entails combining the use of domain knowledge with data mining methods to extract characteristics from raw data. Feature engineering is a type applied to increase the performance of algorithms.

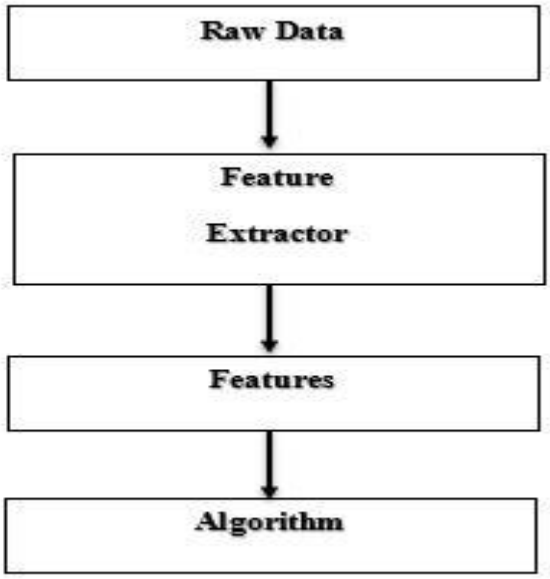


Figure 4.1: Feature engineering flow

4.4 Feature Importance

This mentions to methods for assigning a score based on their usefulness to input features in forecasting a target variable - the significance of features in a predictive modeling problem. Which features are valid is determined by feature importance. Using feature selection will aid in an exceedingly higher perception of the resolved error and, on occasion, lead to model improvements. To input features in a predictive model, it is a series of tools for assigning scores to showing how important each function is when making a prediction. For problems involving numerical value prediction, known as regression, and problems involving class mark prediction, known as classification, the scores may be computed.

In predictive modeling problem, the scores are useful in a variety of scenarios. Including:

- A higher understanding of the knowledge.
- Gaining a stronger understanding of a model.
- The variety of input options is being reduced.

Feature importance:

- Come up with information into the dataset: The relative scores will indicate which features will be most significant to the goal and are least important. A domain expert should understand this and use it as a starting point for collecting more or different data.
- Provide insight into the model: The dataset has fitted to a predictive model used to quantify the most importance scores. Inspecting the value score when making a prediction adds insight into the specific model and which features are most appropriate and weakest to the model. For those models that support it, this is a model that describes what it should do.
- Predictive model enhance: This is done by determining which features to delete based on the significance scores (those with the lowest scores) and which features to retain (those with the highest scores). The model's reliability will improve by this kind of feature choice as well as this may help to alter the matter being modeled (dimensionality reduction) speed up the modeling method and, in some things.

Feature importance assigns a score to each of our data's features; the higher the score, the more significant or significant the feature is to our output variable.

4.5 Accuracy of Model with Selected Features

Model accuracy is a metric based on training data or input data to determine between identifying correlations and relationships which model is best in dataset. When taking statistical samples, accuracy and precision are important. A measurement's precision refers to how accurate it is to its true value. This is significant because errors in outcome may result from faulty equipment, poor data processing, or human error.

One metric for evaluating classification models is accuracy. Informally, precision refers to the percentage of correct predictions made by our model. The following is the formal definition of accuracy. Accuracy = The total number of accurate estimates Number of predictions in total.

TABLE 4.4: MODEL TECHNIC ACCURACY

Model	Training Accuracy
KNN	0.855140
Random Forest	1.00
Decision Tree	1.00

4.6 Cross-Validation

Cross-validation is a collection of model validation methods for evaluating how well the outcomes of a statistical study will generalize to a particular set of data. It is also known as rotation calculation or out-of-sample checking. It's a tool for assessing the applicability of a statistical approach to a particular collection of data. It's a method of assessing machine learning models that involves training some models based on subsets of the input data and then assessing them on the complementary. It aims to see if the model can predict new data that was not used in the estimation process, as well as to see if the model can generalize to a particular dataset to find problems like overfitting or selection bias.

Cross-validation has a number of advantages:

- Out-of-sample precision can now be estimated more accurately.
- Every observation is used for both training and testing, resulting in a more "effective" use of data.

Cross-validation consists of three steps:

- The percentage of sample data set should be set aside.
- Train the model with the rest of the data.
- Use the data-reserve set's part to test the model.

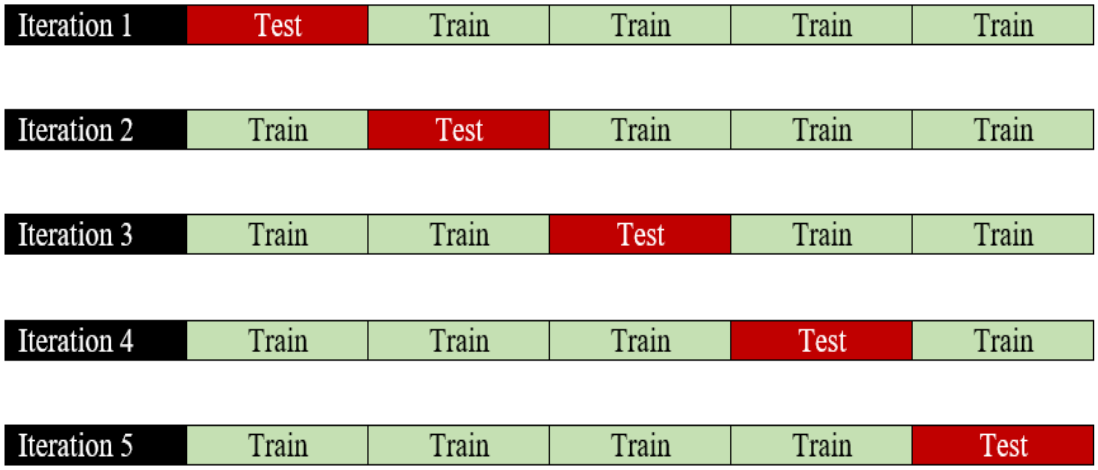


Figure 4.2: Cross-Validation

Overfitting, like data mining in general, has the disadvantage of not being able to predict how well a model can do on new data before it is tested. We will address this by dividing our initial dataset into training and trial subsets. Cross-Validation Techniques come in a multitude of shapes and sizes, but the fundamental idea is the same.

- To divide the data into several subsets.
- The model on the remaining set will train, while holding out a set at a time.
- On the held-out set, run the model.

Then repeat the process for each subset of the dataset.

In our work, cross-validation result is 0.7431906614785992

4.7 Analysis

The analysis phase ensures that we understand the project's vision and establishes a clear scope. This will aid in making decisions about “nice-to-have” features that may come up along the way. The algorithm allows us to classify all of the data according to the priorities that we previously recognized, making the analysis far more precise. It can analyze databases with a large amount of information.

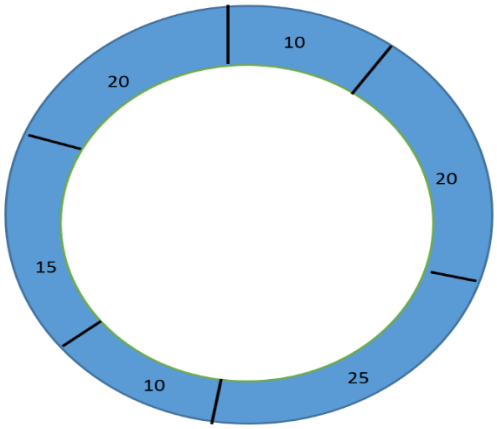


Figure 4.3: Process to Analysis our work in Chart

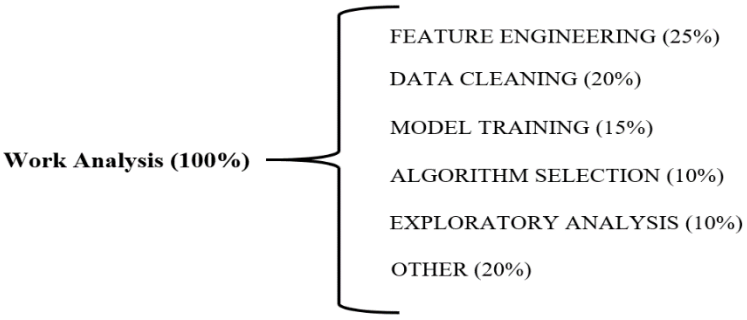


Figure 4.4: Showing all percentage of Work Analysis

In figure 4.3, it shows full analysis in the chart and figure 4.4 shows name of the work analysis including each parts percentage.

CHAPTER 5

Conclusion and Future Scope

5.1 Future Scope

Any condition that affects the heart is referred to as heart disease. There are many varieties, some of which can be avoided. Heart disease affects only the heart, as opposed to cardiovascular disease, which affects the whole circulatory system. The Heart Disease Prediction System provides a massive amount of data that is used to extract undisclosed information to make intelligent medical diagnoses.

In the future, we will convert this research project into an App for this,and we make a site. Futhermore, we are still working on this. Here is some snap of our site.

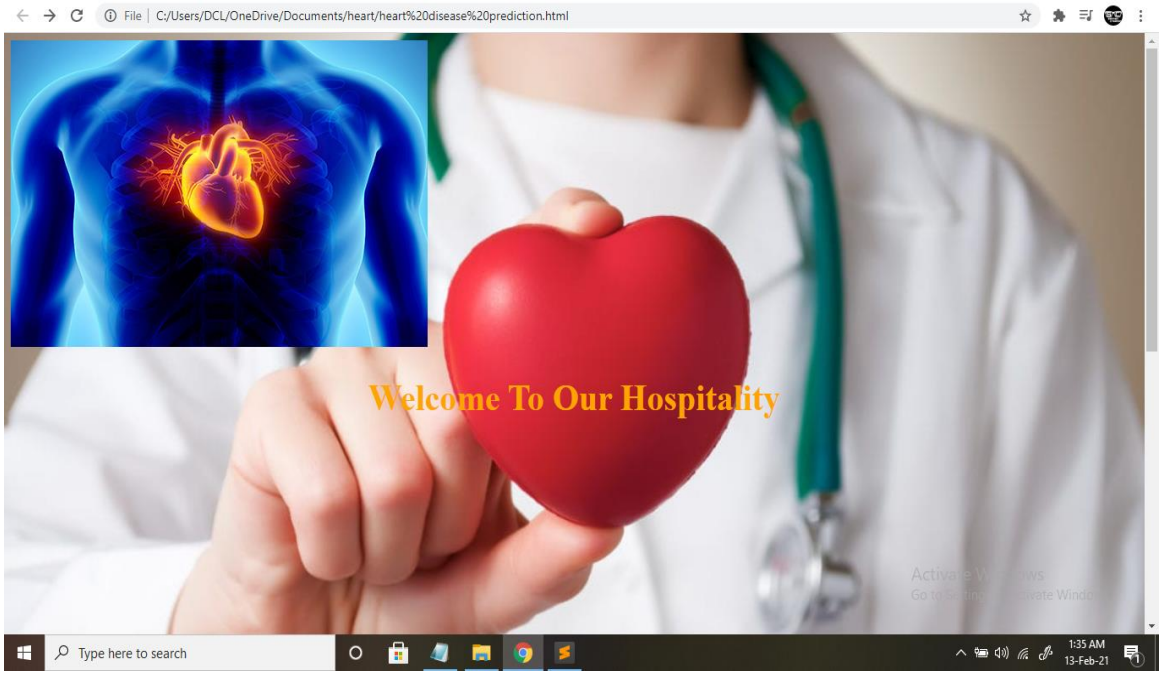


Figure 5.1: Homepage

Figure 5.1 shows home page of our site that we made. This page is our first page. When any one enter this site it shows “Welcome to Our Hospitality”.

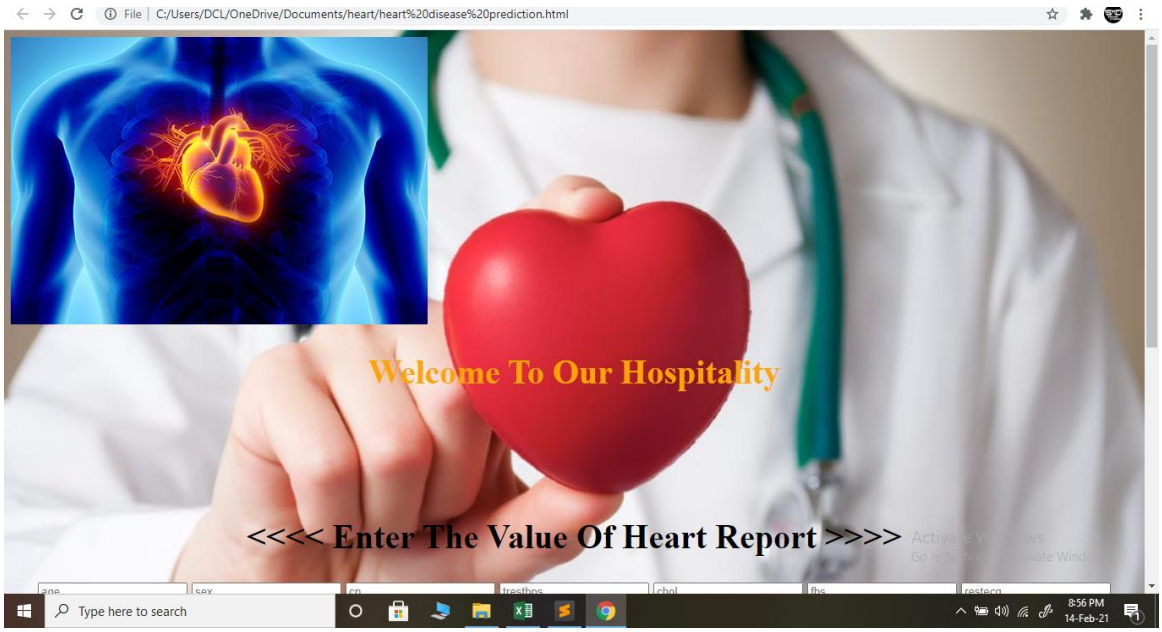


Figure 5.2: Enter the Value of Heart Report

Figure 5.2, here site visited people are going click the “Enter the Value of Heart Report”.

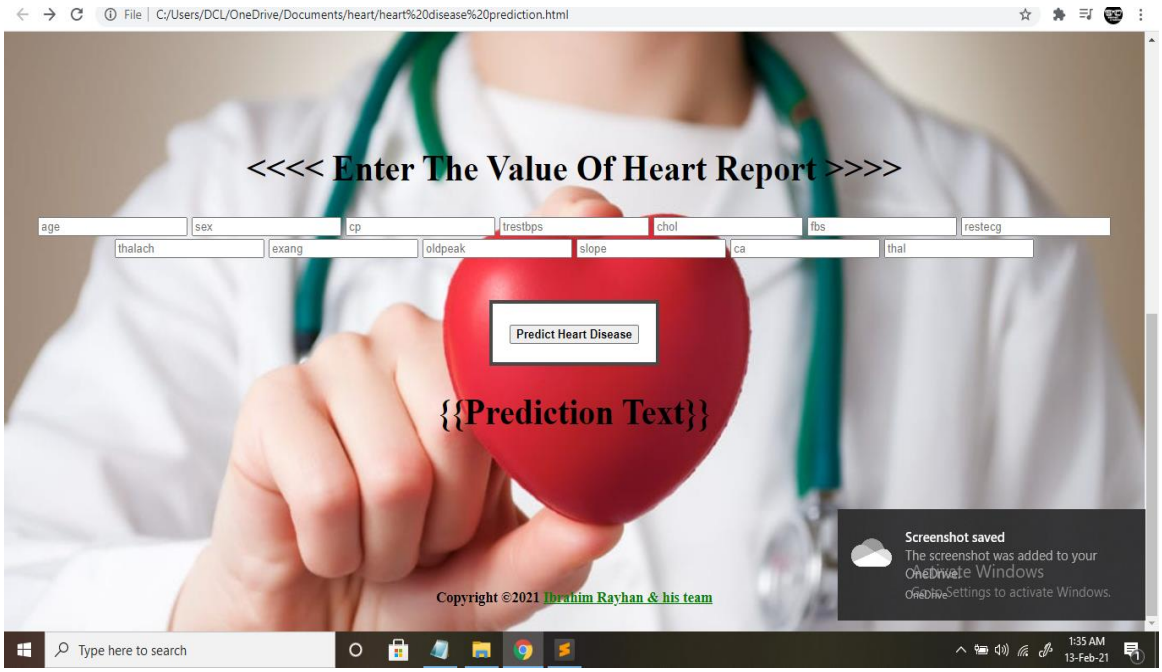


Figure 5.3: Prediction Text

Figure 5.3, they will enter all value according to their report. Then click the “Predict Heart Disease”.

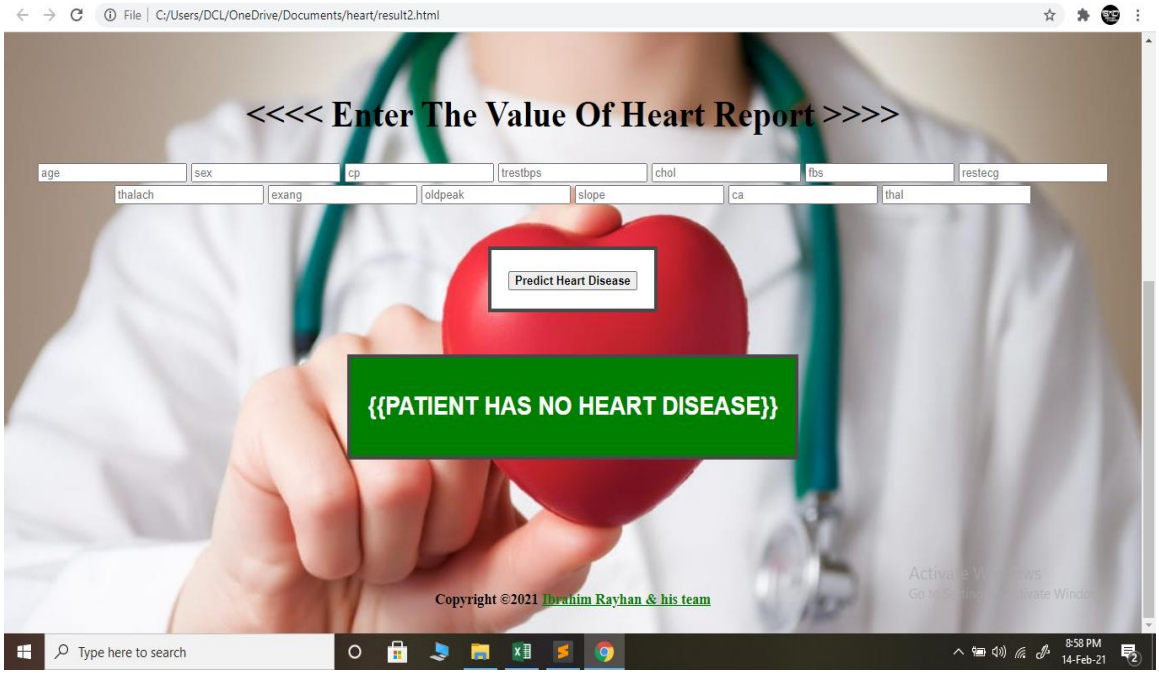


Figure 5.4: Showing Patient has no Heart Disease

Figure 5.4, after click “Predict Heart Disease” then the green box will show the no heart disease result.

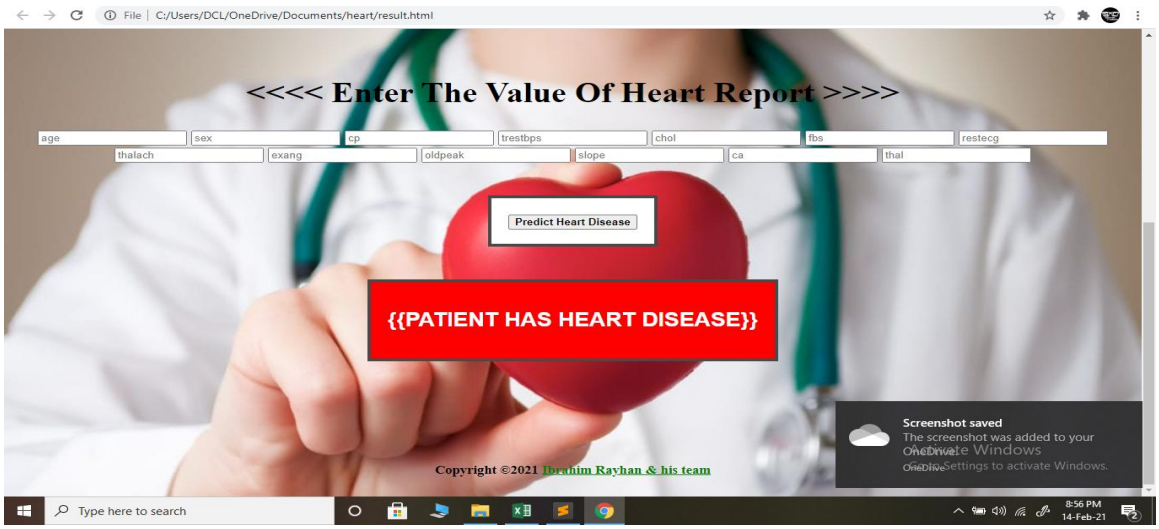


Figure 5.5: Showing Patient has Heart Disease

Figure 5.5, after click “Predict Heart Disease” then the red box will show the has heart disease result.

Through developing a web application, we will improve this work in the future based on the Decision Tree and Random Forest algorithms, and in this analysis, we use not such an extensive dataset.However, we will plan to implement a more extensive dataset, leading to the output of accurate results and the successful and reliable prediction of heart disease by health professionals. Because the heart database has high dimensionality, identifying and selecting important traits for improved heart disease diagnosis is a difficult task for future study.

5.2 Conclusion

Our research investigates various data mining methods used in heart disease prediction approach that are automatic. Our work describes numerous techniques and data processing classifiers that have gained considerable attention for diagnosing heart disease efficiently and effectively. Using the Kaggle repository dataset, this study compares the precision scores of KNN, Random Forest, and Decision Tree algorithms for forecasting heart disease. According to the findings of this research, the Random Forest and Decision Tree algorithm is the most effective algorithm for predicting heart disease. Following the experiments, we discovered that the Random Forest and Decision Tree algorithms provide the best test accuracy, which both are 100 %.

REFERENCE

[1] K. Prince, A. Himanshu and K. Pankaj, "Early Heart Disease Prediction Using Data Mining Techniques", Computer Science & Information Technology (CS & IT), pp. 53-59, 2014.

[2] J. C. Gold and D. J. Cutler, "Cumulated Index Medicus", Volume 41, 2000, pp. 28459.

[3] Caroline Arnold, Heart disease, Franklin Watts, 1990, Heart disease, pp. 111

[4] R. M. Carney and Freedland. "Psychotherapies for depression in people with heart disease". Depression and Heart Disease, K.E. pp. 145–168. 2010.

[5] Y. Choi & J. Choi, "Hypertension Prediction Using Machine Learning Technique", International Journal of Strategic Decision Sciences, 11(3), 52–62, 2020.

[6] J. Soni, S. Soni et al.," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Predictionǁ", International Journal of Computer Applications (0975 – 8887), Volume 17, pp. 8, March 2011.

[7] S. Rajathi, and G. Radhamani, "Prediction and analysis of Rheumatic heart disease using KNN classification with ACO", International Conference on Data Mining and Advanced Computing (SAPIENCE), pp. 6, March 2016.

[8] T. Abidin and W. Perrizo, "SMART-TV: A Fast and Scalable Nearest Neighbor Based Classifier for Data Mining", Proceedings of ACM SAC-06, Dijon, France, ACM Press, New York, NY, pp. 536-540, April 23-27, 2006.

[9] N. Bhatla and K. Jyoti, "An analysis of heart disease prediction using different data mining techniques", International Journal of Engineering Research & Technology (IJERT), Vol. 1, pp. 4, October 2012.

[10] Heart Disease Detection Project Report - Noiselab@UCSD, available at << https://www.google.com/search?q=Heart+Disease+Detection+%22Project%22+Report+Group72+Member:+Yangguang+He,+Xinlong+Li,+Ruixian+Song&sa=X&ved=2ahUKEwjr4bOGr9rvAhUCyjgGHROdBqkQ5t4CMAF6BAgDEAs&biw=1536&bih=754 >>, last accessed on Date 31-03-2021 at 05:06 PM.

[11] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi and P. Ghuli. "Heart Disease Prediction using Machine Learning". International Journal of Engineering Research & Technology (IJERT), vol. 9, 4 April 2020.

[12] Coronary Artery Disease | CAD | Medline Plus, Medlineplus.gov, 2019. [Online]. Available: https://medlineplus.gov/coronaryarterydisease.html. [Accessed: 21- Dec2018].

[13] Atherosclerosis, 2019. [Online]. Available: https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/atherosclerosis. [Accessed: 17- Nov- 2018.

[14] R. Khayat, A. Pederzoli and W. Abraham, "Central Sleep Apnea in Heart Failure", Uscjournal.com, 2009. [Online]. Available: https://www.uscjournal.com/articles/central-sleep-apnea-heart-failure. [Accessed: 12- Nov- 2018].

[15] "Silent Ischemia and Ischemic Heart Disease", www.heart.org, 2019. [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/about-heartattacks/silent-ischemia-a

[16] MyHeart.NET, available at << https://myheart.net/articles/nstemi/>>, last accessed on 31-03-2021 at 4.50 PM.

[17] U. Tejaswini and N. Narhe, "Smart heart disease prediction system using Improved K-Means and ID3 on Big Data", 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), pp. 7, Feb 24-26, 2017

[18] First-degree atrioventricular block, En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/First-degree_atrioventricular_block. [Accessed: 09- Dec2018].

[19] M. J. A. Junaid and R. Kumar, "Data Science and Its Application in Heart Disease Prediction," 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 2020, pp. 396-400.

[20] R. Kasabe and G. Narang, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT), vol. 9, 8 August 2020.

[21] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi & P. Ghuli, "Heart Disease Prediction using Machine Learning", International journal of engineering research & technology (IJERT), Volume 09, Issue 04, April 2020.

[22] M. H. Biglu, M. Ghavami and S. Biglu, "Cardiovascular diseases in the mirror of science", Journal of Cardiovascular and Thoracic Research (JCVTR).

[23] A. H. Seh, "A Review on Heart Disease Prediction Using Machine Learning Techniques", International Journal of Management, IT & Engineering, Vol. 9, pp. 18, 4, April 2019.

[24] R. Rettner, "Why Short People May Have Higher Risk of Heart Disease", Live Science, 2015. [Online]. Available: https://www.livescience.com/50429-short-heightheart-disease-genes.html. [Accessed: 10- Oct-2018].

[25] A. R. Milan S. Avi, D. Ravi, P. Ghuli, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 9 Issue 04, April-2020.

Heart Disease prediction using Data Mining

ORIGINALITY REPORT

15%	9%	7%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	www.ijrte.org Internet Source	2%
3	Submitted to Higher Education Commission Pakistan Student Paper	2%
4	Submitted to University of Melbourne Student Paper	1%
5	Submitted to Liverpool John Moores University Student Paper	1%
6	www.ijert.org Internet Source	1%
7	Mohammed Jawwad Ali Junaid, Rajeev Kumar. "Data Science And Its Application In Heart Disease Prediction", 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020 Publication	1%