

PERFORMANCE ANALYSIS OF HEART DISORDER PREDICTION USING MACHINE LEARNING APPROACHES

BY

**MD. EMTIAZ AHMED
ID: 181-15-1839**

**NAZMUL HASAN SANY
ID: 181-15-1784**

**MASUM BILLAH
ID: 181-15-1732**

The report is presented in Partial compliance with the Qualifications
Requirements for Computer Science and Engineering.

Supervised By

Ohidujjaman

Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mushfiqur Rahman

Senior Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

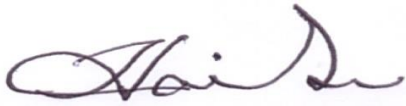
DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project titled “**Performance Analysis of Heart Disorder Prediction Using Machine Learning Approaches**”, submitted by **Md. Emtiyaz Ahmed**, ID No: 181-15-1839; **Nazmul Hasan Sany**, ID No: 181-15-1784 & **Masum Billah**, ID No: 181-15-1732 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 18.01.2022.

BOARD OF EXAMINERS



Sheak Rashed Haider Noori

Associate Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Ohidujjaman

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Dr. Mohammad Shorif Uddin

Professor

Department of Computer Science and Engineering

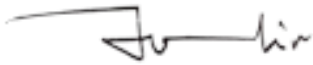
Jahangirnagar University

External Examiner

DECLARATION

We therefore make declaration that this work has been done by us under the watchful eye of **Ohidujjaman, Assistant Professor** in the Department of CSE, **Daffodil International University**. We also announce that neither this research nor any part of this research has been relocated to be awarded any degree or diploma.

Supervised by:



Ohidujjaman

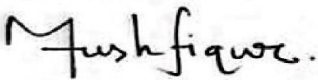
Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Co-Supervised by:



Mushfiqur Rahman

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Submitted by

Md. Emtiyaz Ahmed.

Md. Emtiyaz Ahmed

ID: 181-15-1839

Department of CSE

Daffodil International University

Nazmul Hasan.

Nazmul Hasan Sany

ID: 181-15-1784

Department of CSE

Daffodil International University

Masum Billah

Masum Billah

ID: 181-15-1732

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First of all we express our deepest gratitude and gratitude to Almighty God for His divine blessing enabling us to successfully complete our final year research.

We are very grateful and wish our deepest debt to **Ohidujjaman, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. In-depth knowledge and in-depth interest of our manager in the field of “Machine Learning” To undertake this research. His unwavering patience, expert guidance, constant encouragement, unwavering supervision and enthusiasm, constructive criticism, valuable advice a lot of low draft learning and correction at all stages made it possible to complete the research.

We would like to express our deepest gratitude to our parents, our family, and the Head of the CSE Department “**Professor Dr. Touhid Bhuiyan**”, for his kind assistance in completing our research and for the other members of the faculty and staff of the CSE department of Daffodil International University.

We would like to thank the entire study of our partner at Daffodil International University, who participated in the discussion while completing the course work.

Finally, we must respectfully acknowledge the support and support of our parent’s patients. We would like to thank every one of our Daffodil International University classmates who involved in this discussion while completing this course work.

Finally, Our parents continuous support and patience must be appreciated with due respect.

ABSTRACT

Machine learning, Data mining are fundamental in health care also in health care information and identification are essential. Machine learning approaches has recently been utilized to detect and forecast a variety of major health hazards, including diabetes prediction, brain tumor detection, renal problem prediction, and Covid-19 identification, among others. The part of heart is precious organ of our body and if it has any problem then the impact is more dangerous to our body. According to the Centers for Disorder Control and Prevention (CDC) Trusted Source, heart disorder is the leading cause of death worldwide. We use a few attributes to check our heart disorder analysis, and this attribute is one of the most common causes of heart disorder. As a consequence, 6 machine learning classifiers are employed to evaluate the data using Google Collaboratory: Naive Bayes (NB), Logistic Regression (LG), K Nearest Neighbor, Bagging, Decision Tree (DT), and Random Forest (RF). Using the Seaborn distplot, we extract all attributes' features. Here, Applying Random Forest Algorithm (RF) we get the best accuracy, which is 99.18 %. We have the biggest value of the ROC (receiver operating characteristic) curve of any other algorithm.

Keyword: Bagging Classifier, Decision Tree, Logistic Regression, K Nearest Neighbor, Naive Bayes, Random Forest.

MATERIALS

LIST OF MATERIALS	PAGE NO.
BOARD OF EXAMINERS	I
DECLARATION	II-III
ACKNOWLEDGEMENT	IV
ABSTRACT	V

CHAPTER	PAGE NO.
CHAPTER 1: INTRODUCTION	1-5
1.1 Background	2-4
1.2 Motivation behind the research	4
1.3 Statement of the Problem	5
1.4 Query for Research	5
1.5 Research Scope	5
1.6 Research Organization	5
CHAPTER 2: RELATED WORK	6-8
CHAPTER 3: METHODOLOGY	9-21
3.1 Description of Data	9-11
3.2 Description of Algorithm	11-14
3.3 Planned Model	15-21
CHAPTER 4: RESULT ANALYSIS	22-27
4.1 Confusion Metrics Analysis	22-24
4.2 ROC Analysis	25-27
CHAPTER 5: CONCLUSION AND FUTURE WORK	28-36
5.1 Conclusion	28
5.2 Future Work	28
REFERENCES	29-30
APPENDICES	31
PLAGIARISM REPORT	32-35

LIST OF TABLE

LIST	PAGE NO.
Comparative analysis from existing method	8
A comparison between the prior model and our model	16

LIST OF FIGURES

LIST	PAGE
Figure 1: Metrics of Correlation from Target to Feature Attribute	15
Figure 2: Correlation Metrics among Target to Feature Attribute	17
Figure 3: Seaborn distplot for showing a single- variate allocation of data using histogram	18
Figure 4: sns pairplot according to among 18 attributes	19
Figure 5: Structure of Confusion Metrics	21
Figure 6: Confusion Matrix for Logistic Regression	22
Figure 7: Confusion Matrix for Decision Tree Classifier	22
Figure 8: Confusion Matrix for Random Forest	23
Figure 9: Confusion Matrix for K-Nearest Neighbor	23
Figure 10: Confusion Matrix for Naïve Bayes	24
Figure 11: Confusion Matrix for Bagging Classifier	24
Figure 12: ROC for Logistic Regression	25
Figure 13: ROC for Decision Tree	26
Figure 14: ROC for Random Forest	26
Figure 15: ROC for K Nearest Neighbor	27

CHAPTER 1

INTRODUCTION

Machine learning is a powerful area of study for research. Many statistical and machine learning techniques are often used in various fields. Machine learning can also be used in sectors such as marketing, health and medical disorders, weather prediction, socioeconomic activity analysis, and so much more. Many disorders in the medical sector may be discovered or predicted by machines applying machine learning techniques. Heart disorder is a big concern to the world's health in the twenty-first decade. Because of the high rate of heart disorder, has massive repercussions for a country health and socio - economic growth. In similar ways, there is a high risk of high-speed growth. Mainly poor people faced it. Also such disorders as diabetes, hypertension, and others. These are regularly we face in nowadays. There is a lack of heart disorder, mainly heart disorder, the impact of the world health community's focusing into heart disorders, and its awareness. It is extremely crucial that developed nations focus on this and develop stronger inclusive, values, and prevent heart disorder.

Many hospitals keep information on heart disorder patients in their systems. Various patterns can be established by analyzing those information, which will help in prediction. Applying data mining methods on all those information, it is possible to discover lot of information and apply this information to anticipate disorder. Heart disorder which affects a large number of people. The large number of people in Bangladesh are unconcerned about disorder. Because of it, the number of persons influenced by the disorder is raising every moment of a day. It might be controlled if individuals could identify or forecast whether they are afflicted or are about to be impacted.

People must take special care to avoid being affected. Predictive can be used to forecast the disorders. Classification, regression, and indexing are some of the approaches that could also be performed. Many individuals believe that categorization is the preferred approach. After the research has been completed, it will definitely help in the prediction of Heart Disorder. People will be aware of the illness as well as their own condition of body.

Primary aims of this work for anticipating the condition using machine learning algorithm, to warn if anyone has proclaim the disorder to assess following situation using different to evaluate which machine learning algorithm performs the best, researchers used machine learning algorithms.

1.1 Background

1.1.1 Heart Mechanism

In terms of the heart's structural anatomy, our mechanism of heart, which is a connection between blood arteries that supplies blood to all parts in our bodies, revolves around the heart. Blood carries oxygen and other critical nutrients to all human organs, allowing them to stay healthy and function effectively. The pumping of blood throughout our circulatory system is the job of our heart, which is a muscle. The right and left sides of our hearts pump in opposite directions. The right side of our heart pumps oxygen-depleted blood from our veins to our lungs, where it takes up oxygen and exhales carbon dioxide. Our hearts' left half transfers oxygen-rich blood from our lungs to the rest of our bodies through our arteries. When a fatty material called plaque builds up in our arteries, it causes heart disorder. Plaque hardens and narrows our arteries over time. When plaque clogs an artery, it acts like a clogged drainpipe, causing less blood to flow through. We grow fatigued and our legs may feel weak if our heart does not pump enough oxygen-rich blood to our primary organs and muscles. Weight increase, swelling in our ankles, legs, and tummy.

1.1.2 Heart Disorder classifications:

The following are some of the most common kinds of cardiac disorder affects the heart's blood flow and can lead to a heart attack:

- **Arrhythmia:**

Heartbeats that are too fast (tachycardia), too slow (bradycardia), or irregular are caused by a shift in the heart's electrical impulse sequence (palpitations).

- **Heart valve disorder:**

When a valve in the heart is damaged or sick, the condition is known as heart valve disorder.

- **Failure of the Heart:**

A syndrome whereby the heart sometimes doesn't operate as well as it should, causing fluid to accumulate. Heart failure does not imply that the heart has ceased to beat.

- **Coronary Artery Disorder:**

The most frequent type of heart illness is coronary artery disorder (CAD). This disorder is characterized by a hardness or constriction of the arteries leading to the heart.

- **Heart Attack:**

A heart attack, unfortunately, is often the initial symptom for cardiovascular disorder among various person. Blood circulation is disturbed when the arteries leading to the heart get clogged, causing together in heart problem.

1.1.3 Reason behind Heart Disorder

Many persons with heart disorder (HD) do not realize difficulties or problems until their condition has proceeded to a critical stage.

The following symptoms have been observed:

1. Discomfort in the chest,
2. Anorexia, dyspepsia, heartburn, or stomach pain
3. Arm Pain,
4. You're dizzy or light-headed
5. Jaw or Throat Pain,
6. You are rapidly exhausted,
7. Snoring ,
8. Sweating,
9. A Continuous Cough,
10. Swollen Legs, Feet, and Ankles
11. Increase in heart rate

1.2 Motivation behind the research

Heart disorder is a life-threatening disorder; however, if we have proper knowledge about it and how to prevent it, we can save our lives from being affected by this disorder. Because there is no early warning sign of this disorder. The only way to predict the outcome of a test for this disorder is to look for the possibility of HD in blood, angiogram, ECG, ECO, urine. Patients will benefit from this disorder if it is identified early. The forecast for the Test effect is used for early detection. This is our justification for using a computer to predict HD.

1.3 Statement of the Problem

This paper, dataset for Heart Disorder is taken from hospital and online data. Choosing the best algorithm model for precision, accuracy, f-measure, recall and learning rate. Models are created by using various classifiers for machine learning and evaluating the results after preprocessing and converting data.

1.4 Query for Research

- The primary aims of this work are to anticipate that disease have used a machine learning technique, to notify whenever a patient is suffering of the disorder and to compare the results of multiple algorithms to determine how this following algorithm works efficiently. To allow for proper reduction of the property perspective on the used dataset.
- To develop a more precise statistical machine-learning algorithm for determining a lower FN count.

1.5 Research Scope

Machine learning algorithms focused on classification will predict or not heart disorder. Using machine learning methods, we find greater performance. In the future, we will work as neural networks in deep learning to make decisions.

1.6 Research Organization

The structure of the paper residue is as follows:

- Chapter 2 discusses the linked works.
- The full process is depicted in Chapter 3.
- The findings are presented in detail in Chapter 4.
- Chapter 5 presents the conclusions and future efforts.

CHAPTER 2

RELATED WORK

Ayon et al. applied seven soft computing method to predict heart problems to use the Detection dataset & Clinic heart disorder datasets. They attained an accuracy of 98.15 percent using the Statlog dataset, whereas SVM produced overall efficiency of 97.36 percent. [1]

Morey et al. have taken a demographic dataset from Korea and proposed a HDCDSS model for a clinical decision support system. They have taken two datasets. Firstly, they have found 95.90% for the first dataset and secondly, 98.40% for the second dataset. [2]

Rajdhan et al. predicted heart disorder by ML and found the best accuracy of 90.16% in random forest,. In the decision tree, Logistic Regression and Native Bayes found 81.97%, 85.25%, and 85.25% accuracy, respectively. [3]

Singh et al. have used a ML approach using heart disorder prediction founding percentage 87% in KNN, rather than in DT (79%), LR (78%) and SVM (83%). [4]

Dutta et al. used a multi-stage model to predict heart disorder using 37,079 imbalanced clinical data. CHD cases are predicted with 77.3 percent accuracy, while non-CHD cases are predicted with 81.8 percent accuracy. [5]

Andres et al. employed an ML technique to predict heart disorder and obtained 99 percent accuracy for Hungarian, 98.7 percent accuracy for Cleverland, and 99.4 percent accuracy for CH by utilizing a bigger dataset with 74 characteristics. [6]

Ali et. al. proposed a new model. They found 83.5% accuracy in heart disorder prediction. They also found low accuracy in DT, LR, NB, SVM, RF, MLP. [7]

Rani et.al. predicted a system of heart disorder predicting using the RandomizedSearchCV() method by using the dataset. They have used NB, SVM, LR, and RF algorithms using ML. They have found the best accuracy in the Random Forest at 86.60%. [8]

Katarya et. al. used machine learning approaches to predict cardiac disorder and reported 95.60 percent accuracy in Random Forest. They found that logistic regression had 90.40 percent accuracy, Nave Bayes had 90.10 percent accuracy, SVM had 92.30 percent accuracy, KNN had 71.42 percent accuracy, Decision Tree had 81.31 percent accuracy, ANN had 92.30 percent accuracy, DNN had 76.92 percent accuracy, and MLP had 5.42 percent accuracy. [9]

Almustafa et. al. used ML to predict heart illness, The K-NN (K = 1), Decision Tree J48, and JRip classifiers, respectively, had 99.7073, 98.0488, and 97.2683 percent classification accuracy. [10]

Bharti et. al. employed ML techniques to predict cardiac disorder using two types of datasets. Using the Random Forest technique, they found 91.60% accuracy in Cleveland dataset and 97.0 percent accuracy in the people's dataset. [11]

Comparative analysis from existing method

In this section comparing among the others research work as reference before. From this easily can see the difference between among algorithms and shows which one is doing better preforms hereby.

TABLE 1: SOME COMPARATIVE ANALYSIS OF PREVIOUS MODEL AND OUR MODEL

Authors	Models	Accuracy
Amirgaliyev [1]	SVM	93.1 %
Charleonnann [2]	SVM, LR & KNN	98.3%,96.55% & 94.8%
Sinha [3]	KNN & SVM	73.75% & 78.75%
Sharrma [4]	DT	98.6%
Khan [5]	NB,LR,MLP,J48,SVM & NBTree	95.75%,96.50%,97.25%,97.75%, 98.25% & 98.75%
Polat [6]	SVM	98.5%
Radha [7]	SVM & KNN	98% better than KNN
Serpen [8]	Decision tree	98.25%
Ahmad [9]	Classification modeling	98.34%
Hosseinzadeh [10]	SVM, Decision tree (J48), MLP, and NB strategies	97%
Our Model	Random forest classifier	99.18 %

After comparing, Random Forest gives the better result in this section and thus this report make impressive piece of work. It gives 99.18% Accuracy, Best among the other models.

CHAPTER 3

METHODOLOGY

There are several components to this chapter. Let us have a conversation about it.

3.1 Description of Data

We used 1203 datasets from the UCI ML repository into our paper. There are 19 components in this dataset, 18 of which are predictive factors, and one characteristic for the target class. So we'll go over all of our characteristics in a little more detail.

Age: This attribute specifies a person's age in years. It's a predictive variable with a numeric value.

Sex: This attribute determines whether a person is female or male. It's a numeric-valued predictive variable.

CPT: This attribute specifies a person's chest pain type. Whether it is typical angina, denotes with 1. Whether it is atypical angina, denotes with 2. When it is non-angina pain, denotes with 3. When it is asymptomatic, denotes with 4. It's a predictive variable with a numeric value.

RBP: This attribute indicates Resting Blood Pressure. It's a predictive variable with a numeric value.

Chol: This attribute refers serum cholesterol in mg/dl sugar. It's a predictive variable with a numeric value.

Fbs: This attribute indicates Fasting Blood Sugar. It's a predictive variable with a numeric value.

Restecg: This attribute indicates resting electrocardiographic results. It's a predictive variable with a numeric value.

MHR: This attribute indicates Maximum Heart Rate achieved. It's a predictive variable with a numeric value.

EXGINA: This attribute indicates that Exercise included Angina. It denotes 1 for Yes and 0 for No. It's a predictive variable with a numeric value.

Oldpeak: This attribute refers depression induced by exercise relative to rest. It's a predictive variable with a numeric value.

Slope: This attribute indicates the slope of the peak exercises. It's a predictive variable with a numeric value.

NMV: The number of main vessels colored with 0-3 by fluoroscopy is referred to by this property. It's a predictive variable with a numeric value.

Diabates: This attribute refers the diabetes history of a person. It's a predictive variable with a numeric value.

Thal: This attribute refers 3 for normal and 6 for fixed. It's a predictive variable with a numeric value.

Hypertension: This attribute refers the hypertension history of a person. It's a predictive variable with a numeric value.

Stroke: This attribute refers the stroke history of a person. It's a predictive variable with a numeric value.

Smoking history: This attribute refers the smoking history of a person. If yes it refers 1 and if not it become 0. It's a predictive variable with a numeric value.

BMI: This attribute refers the BMI (Body Mass Index) history of a person. It's a predictive variable with a numeric value.

Target: It's the responsive feature. Here is a list of people who have or do not have a heart disorder. This variable is also of the nominal kind.

3.2 Description of Algorithm

3.2.1 Naive Bayes:

In Naive Bayes, the Bayes' Theorem is applied, signifying that the all classifiers be impartial. Putting differently, the position between one element in a class has no bearing on the inclusion of another.

The following equation expresses Bayes' theorem mathematically:

$$P(A|x) = P(x|A) P(A) / P(x) \quad (1)$$

$$P(A|x) = P(x_1 | A) \times P(x_2 | A) \times \dots \times P(x_n | A) \times P(A) \quad (2)$$

$P(A|x)$ represents the class's posterior probability based on the predictor (x) (A). $P(A)$ represents the class's prior probability, $P(x)$ represents the predictor's prior probability, $P(x|A)$ reflects the chance of the predictor for specific class (A).

Naive Bayes implies that all variables contribute equally, in order to consider their independence. It's crucial to remember this.

3.2.2 Logistic Regression:

When the dependent variable is dichotomous, logistic regression is the best regression method to use (binary). Logistic regression, like other regression studies, is a predictive study. To describe data and explain the connection between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, logistic regression is utilized.

A significant machine learning algorithm is logistic regression. The purpose is to simulate a random variable's probability. Given experimental results, being 0 or 1.

The following equation expresses formula of Logistic Regression:

The constant (s_0) in logistic regression moves into the curve left and right, whereas the slope (s_1) determines how steep the curve is. Any number of numerical and/or categorical variables can be handled using logistic regression. So, the probability equation for Logistic Regression will be-

$$P = 1 / (1 + e^{-(S_0 + S_1 X_1 + S_2 X_2 + S_3 X_3 + \dots + S_n X_n)})^0 \quad (1)$$

Between linear regression and logistic regression, there are a number of analogies. In the same way that simple In linear regression, least square regression is used to estimate coefficients for such best fit lines, whereas maximum likelihood estimation (MLE) is used in LR to estimate model coefficients that connect predictors to the destination. After this initial function has been assessed, the process is repeated until the LL (Log Likelihood) does not change significantly.

$$V^1 = V^0 + [X^T S X]^{-1} \cdot X^T (y - \mu) \quad (2)$$

Here,

V is a logistic regression vector.

S is Square Matrix.

μ is length of vector

3.2.3 Bagging Classifier:

Bagging is a supervised learning approach for improving performance by combining the results of many classifiers. These approaches operate by splitting the training set and running it through a variety of machine-learning models, then aggregating their predictions when they return to offer an overall estimate for each occurrence in the original data. Bagging is a common strategy used in machine learning for classification challenges when using decision trees or artificial neural networks as part of a boosting ensemble. Although bagging may be used to tackle regression issues, classification is more effective.

3.2.4 Decision Tree:

Decision The classification of trees is based on tree-like structures. The root nodes represent the criterion, while the child nodes represent the class label.

The repercussions of the events are shown by the branches of the root nodes.

The following Entropy $E(S)$ can be represented as –

$$E(S) = \sum_{i=1}^c - p_i \log_2 p_i \quad (1)$$

While S is the current condition, and P_i denotes the likelihood of occurrence i occurring in initial State, or the fraction of class I in a nodes in state S .

3.2.5 Random Forest Classifier:

RF is a supervised learning approach. It may be used for both classification and regression. It's also the most versatile and easy-to-use algorithm. A forest is made up of trees. The more trees a forest contains, the stronger it is supposed to be. Random forests construct decision trees from randomly selected data samples, derive predictions from each tree, and then vote on the best alternative. It also acts as a strong signal of a feature's significance. Random forests are used in a variety of applications, including recommendation engines, image classification, and feature selection. It may be used to classify loyal loan applicants, identify fraud, and predict illness.

$$\text{MSE} = 1/N \sum_{i=1}^c (f_i - y_i)^2 \quad (1)$$

Here,

N is the data points number, f_i is the returned value through the model and y_i is the genuine value for data point i .

This algorithm determines the distance between each node's expected real value and the predicted actual value, assisting you in determining which branch is the better choice for your forest. The value of the data point you're testing at a particular node is y_i , and the value given by the decision tree is f_i .

3.2.6 K Nearest Neighbor:

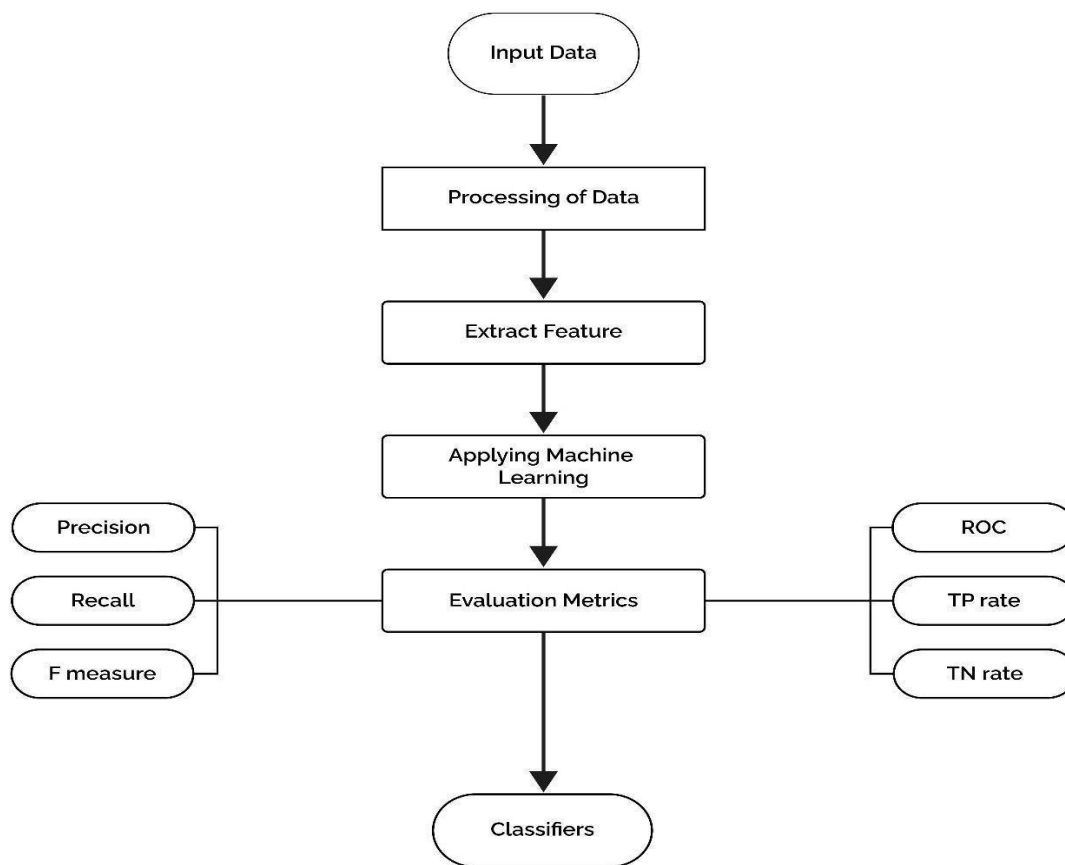
K Nearest Neighbor is a basic algorithm that keeps all available data and classifies fresh data or facts which is based on statistical method. That is often being used for define a piece of data depending on how its neighbors are defined.

3.3 Planned Model

Going through the execution processes. The study was conducted using Python and the sklearn library. Hereby proposed a model for the among works mapping through a flow diagram. Here contains some attributes and they have their own value to present the work memoranda.

Firstly insert data and processing the data to extract features. Then applying the machine learning algorithm through evaluation matrix used Precesion, Recall, F measure, ROC, TP rate and TN rate.

In figure 1, creating a flowchart for the suggested model's method.



Figure_1: Metrics of Correlation from Target to Feature Attribute

A comparison of the existing methods

TABLE 2: A comparison between the prior model and our model

Author	Model	Accuracy
Ayon[1]	Statlog, SVM	98.15%,97.36%
Morey[2]	HDCDSS	95.90% & 98.40%
Rajdhan[3]	RF,DT,LR,NB	90.16%,81.25%,85.25%,85.25%
Singh[4]	KNN,DT,LR,SVM	87%,79%,78%,83%
Dutta[5]	CHD, Non-CHD	77.3%,81.8%
Andres[6]	Hungarian,Cleveland,CH	99%,98.7%,99.4%
Ali[7]	Proposed Model	83.5%
Rani[8]	Random Forest	86.60%
Katarya[9]	RF,LR,NB,SVM,KNN, DT,ANN,DNN,MLP	95.60%,90.40%,90.10%,92.30%, 71.42%,81.31%,92.30%,76.92%,75.42%
Almustafa[10]	KNN,DT,JRip	99.7073%,98.0488%,97.2683%
Bharti[11]	Random Forest	91.63%,97%
Our Model	RF,LR,DT,KNN,BC,NB	99.18%,87.65%,87.65%,80.65%,26.34%, 13.60%

Here, comparing among the others research work as reference before. From this, easily can see the difference between among algorithms and shows which one is doing better preforms hereby. Our proposed model performs the algorithms and shows the best accuracy among others in Random Forest Classifier, Then in order Logistic Regression, Decision Tree, K Nearest Neighbor Performs well. But in Naïve Bayes and Bagging classifier it shows the worst value among others due to numeric value in the dataset. For nicely categorized we select only numerical value to perform the machine learning algorithms.

3.3.1 Data Input

A total of 1203 patient data were collected for this study in order to forecast an analysis and determine performance. This data was taken from online.

3.3.2 Preprocessing of Data

A total of 19 characteristics are utilized in this dataset, 18 of which factors are predictive and 1 of which is a variable that responds. Some of the 18 predictive qualities are nominal, while others are numerical. Our data collection is now filled with numeric values. Furthermore, the dataset was partitioned: 80 percent for & percent for testing.

3.3.3 Extraction of Feature

Figure 2 depicts the feature extraction process using principal component analysis.

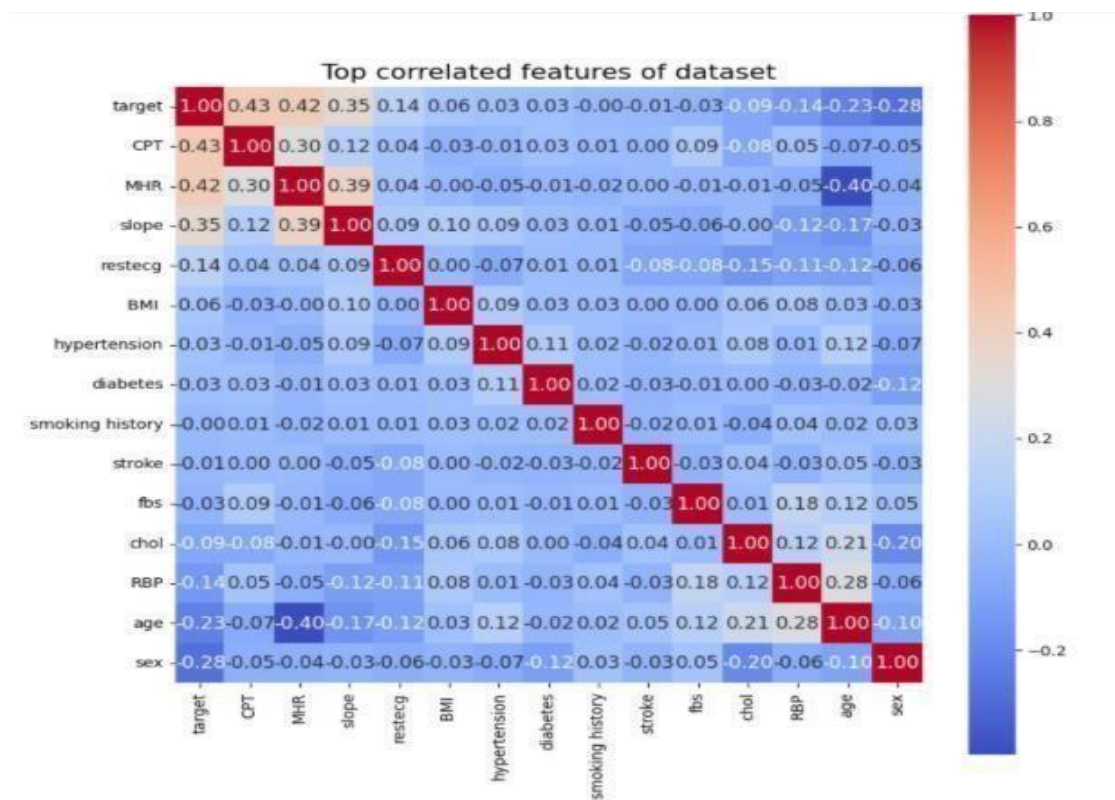


Figure 2: Correlation metrics among target to feature attribute

A histogram with a line may be shown using Seaborn distplot. This will be demonstrated in a variety of ways. Seaborn is used in collaboration with matplotlib, a Python charting library. A distplot is a graph showing a single-variate allocation of data.

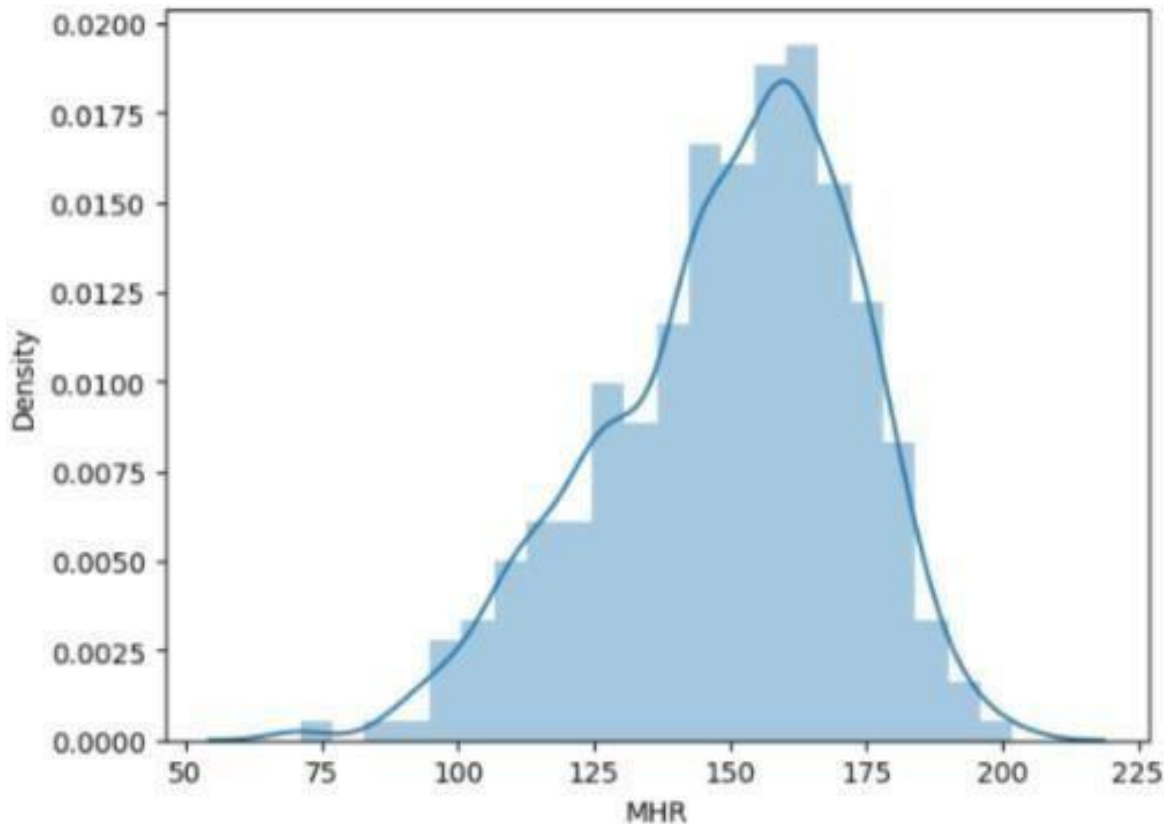


Figure 3: Seaborn distplot for showing a single-variate allocation of data using histogram

In this figure we will show among our 18 attributes and their Density-MHR (Maximum Heart Rate) ratio using histogram graphical representation. And the highest density among 150 – 175 rates.

Seaborn Pairplot, shown in Figure 4, is being used for determine the relationship between all of the variables in a Pandas DataFrame. It's similar to a seaborn scatter plot, however it only plots two variables, whereas sns paiplot shows several features/variables in a grid style. It clearly shows that variations among internal data attributes with each other components. That might help us to see the variation among the attributes and their multi purposes accommodation process instead of data hazard.

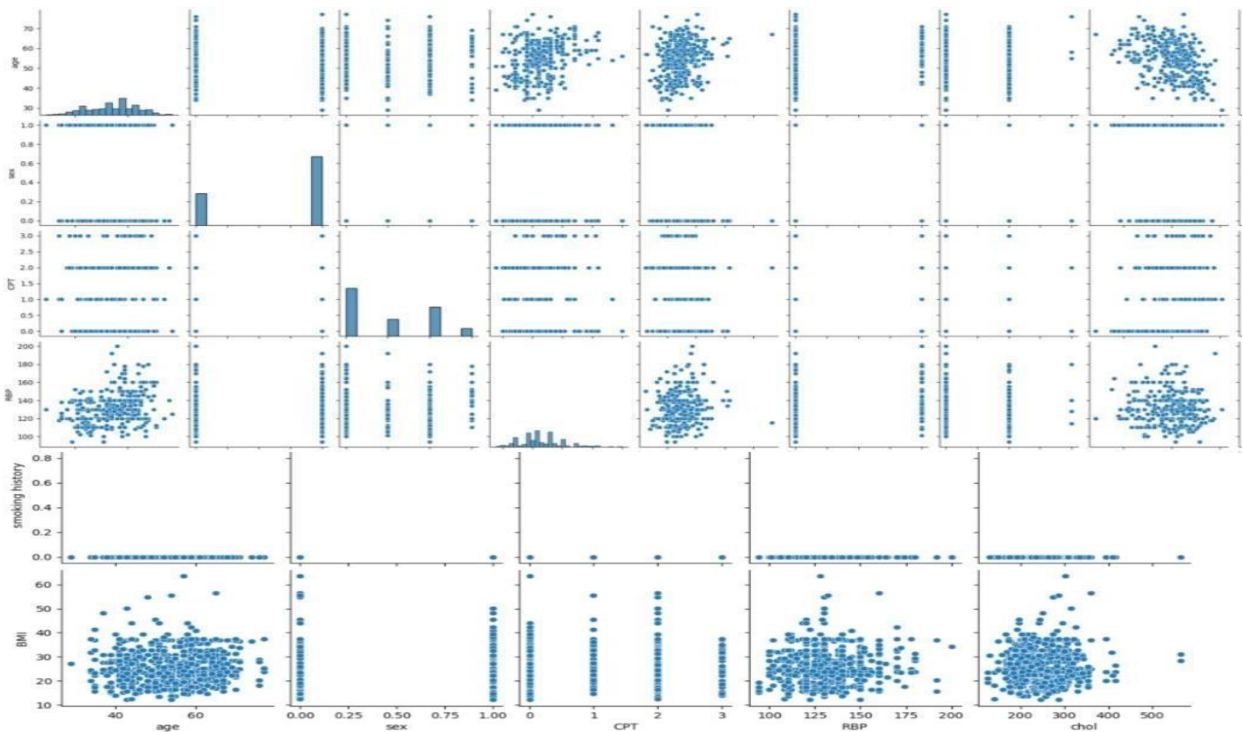


Figure 4: sns pairplot according to among 18 attributes

3.3.4 Machine Learning Classifier

Here uses total six machine learning algorithm to predict the heart disorder or non-heart disorder prediction. The following algorithms: -

- RF,
- DT,
- NB,
- LR,
- BC,
- KNN.

Every algorithms are being discussed previously in 3.2 Description of Algorithms part.

Every algorithm performs their best and helps us to take decision on which algorithms accuracy is better than the others. After applying all of the algorithm our code returns a simplest the best outcome among them.

3.3.5 Metrics of Evaluation

In this following part, we'll learn about several main insights from our research. We'll go through them shortly from here.

- **Precision:**

Precision or positive predictive value = $\text{TRUE POSITIVE} / (\text{TRUE POSITIVE} + \text{FALSE POSITIVE})$

- **Recall:**

Recall = $\text{TRUE POSITIVE} / (\text{TRUE POSITIVE} + \text{FALSE POSITIVE})$

- **F Measure:**

F-1 Score = $(2 \times \text{TRUE POSITIVE}) / (2 \times \text{TRUE POSITIVE} + \text{FALSE POSITIVE} + \text{FALSE NEGATIVE})$

- **Accuracy:**

Acc = $(\text{TRUE POSITIVE} + \text{TRUE NEGATIVE}) / (\text{TRUE POSITIVE} + \text{TRUE NEGATIVE} + \text{FALSE POSITIVE} + \text{FALSE NEGATIVE})$

3.3.6 Finding Best Model

The dataset's degree of precision is determined in order to demonstrate the efficiency of various algorithms. Finally we found the best accuracy among models is “Random Forest”.

That gives us 99.17% accuracy that was enough larger accuracy than the other accuracy.

3.3.7 Confusion Metrics

- † TP indicates true positive = We expected HD and it is the same when a label is successfully predicted.
- † FP indicates the false positives = When a label is anticipated wrongly, we assumed it was HD, but it's not.
- † FN indicates the false-negative = We expected that when an expected label is absent, it's not HD, but it's also.
- † TN indicates true negative = We predict that a label is not HD when another label correctly predicts it.

		Actual Class	
Predicted Classes		FP	TP
		TN	FN

Figure 5: Structure of Confusion Metrics

Spam detection, churn prediction, emotion analysis, dog breed recognition, and other categorization examples are just a few instances.

Suppose, a hospital has doctor and patient. Some patient feel sick but they have no disease (TN).

Again some patient feel sick but they have disease (TP). Some patient feel comfortable but they have no disease (FP). Some patient feel comfortable but they have disease (FN).

CHAPTER 4

RESULT ANALYSIS

Here represents confusion metrics gaining from the performance analysis among the algorithms. The confusion matrix representation of classifying values. The number of correct and incorrect predictions is totaled and divided by class using count values. When the confusion matrix generates predictions, this key to it becoming confused.

4.1 Confusion Metrics Analysis

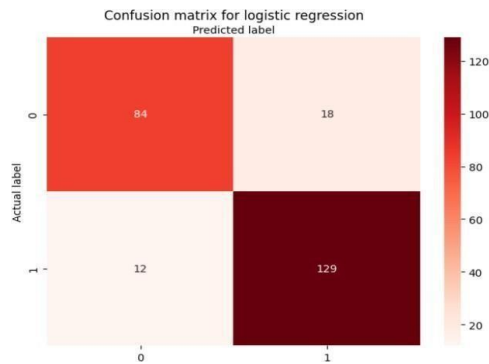


Figure 6: Confusion Matrix for Logistic Regression For Logistic Regression the value of confusion metrics are:
TP = 18; TN = 12; FP = 84; FN = 129

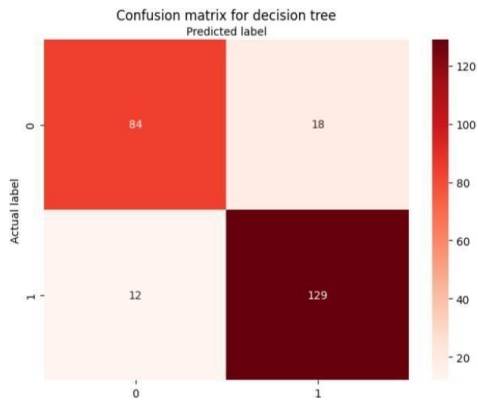


Figure 7: Confusion Matrix for Decision Tree Classifier

For Decision Tree Classifier the value of confusion metrics are:

TP = 18; TN = 12; FP = 84; FN = 129

Also there has been some other algorithms analysis for confusion matrix and found outcomes like below-

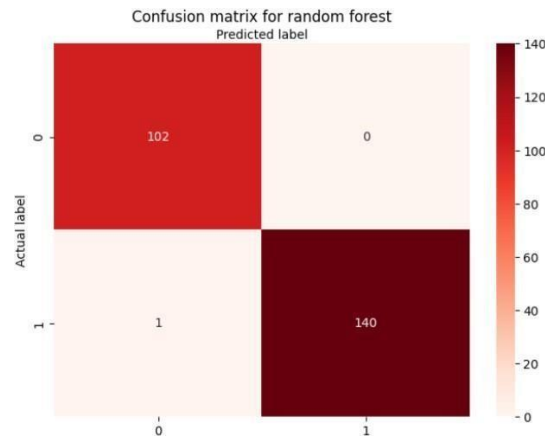


Figure 8: Confusion Matrix for RF

For RF Classifier the value of confusion metrics are:

TP = 0; TN = 1; FP = 102; FN = 140

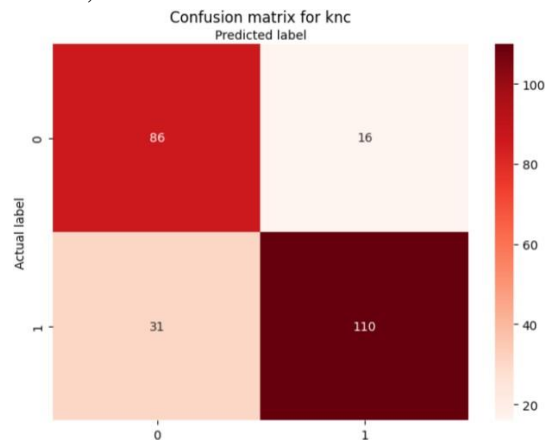


Figure 9: Confusion Matrix for KNN

For KNN Classifier the value of confusion metrics are:

TP = 16; TN = 31; FP = 86; FN = 110

Here seen a minimal ratio for Naive Bayes and Bagging classifier's confusion matrix values

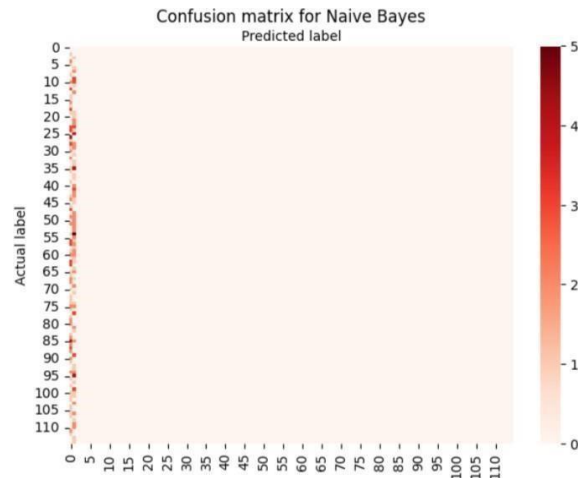


Figure 10: Confusion Matrix for Naïve Bayes

As we get low accuracy among Naïve Bayes and Bagging Classifier, the estimated Confusion Metrics looks like abnormal. Others we get a quality accuracy to identify the confusion metrics properly with figure shown before.

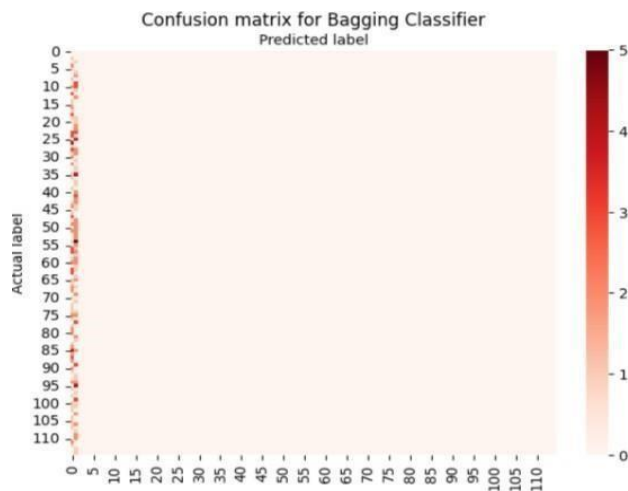


Figure 11: Confusion Matrix for Bagging Classifier

4.3 ROC Analysis

A receiver operating characteristic (ROC) curve refers to graphical presentation that indicates the detectability of a binary classifier system as its threshold of detection varies. ROC curve compares the TPR (true positive rate) of various threshold settings to the FPR (false positive rate).

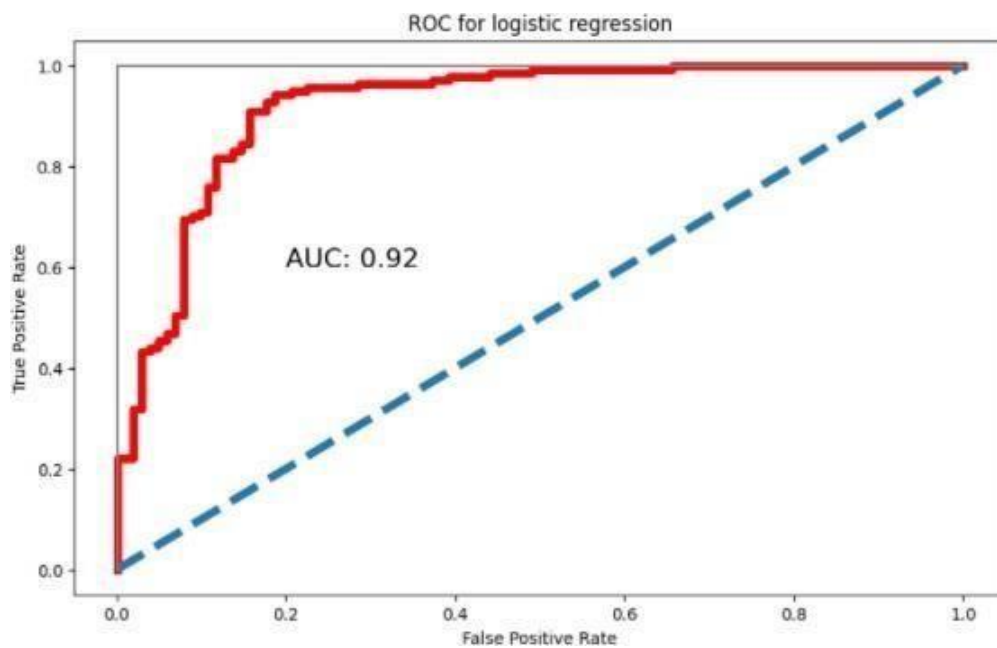


Figure 12: ROC for Logistic Regression

In this ROC analysis for Logistic Regression, found that the accuracy is 92% according to True Positive-False Positive curve.

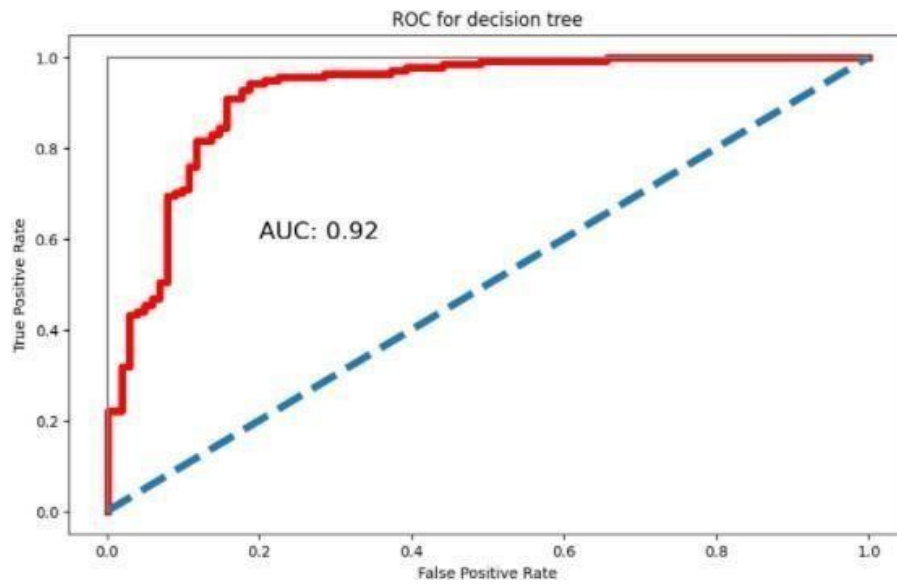


Figure 13: ROC for Decision Tree

In this ROC analysis for Decision Tree, found that the accuracy is 92% according to True Positive-False Positive curve.

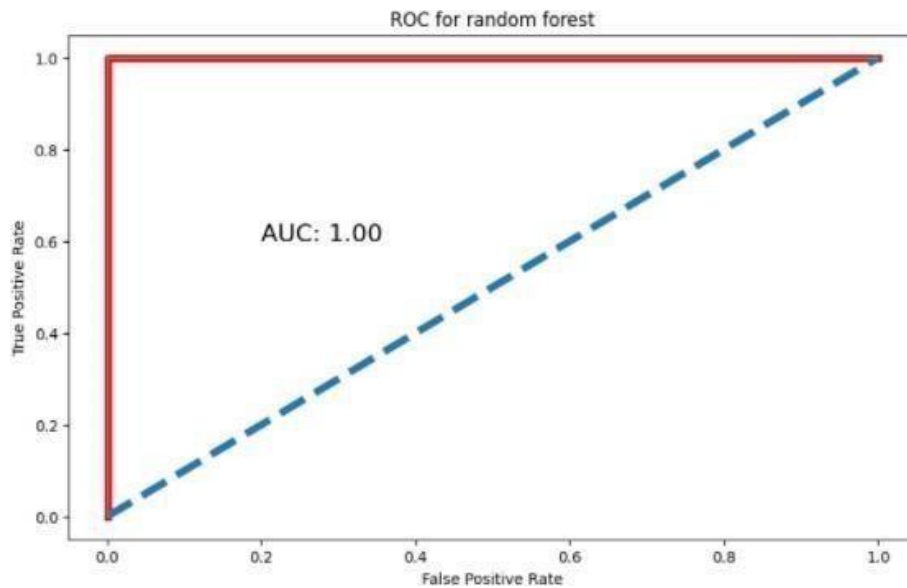


Figure 14: ROC for Random Forest

In this ROC analysis for Random Forest Classifier, found that the accuracy is 100% according to True Positive-False Positive curve.

In this ROC analysis for K Nearest Neighbor, found that the accuracy is 91% according to True Positive-False Positive curve.

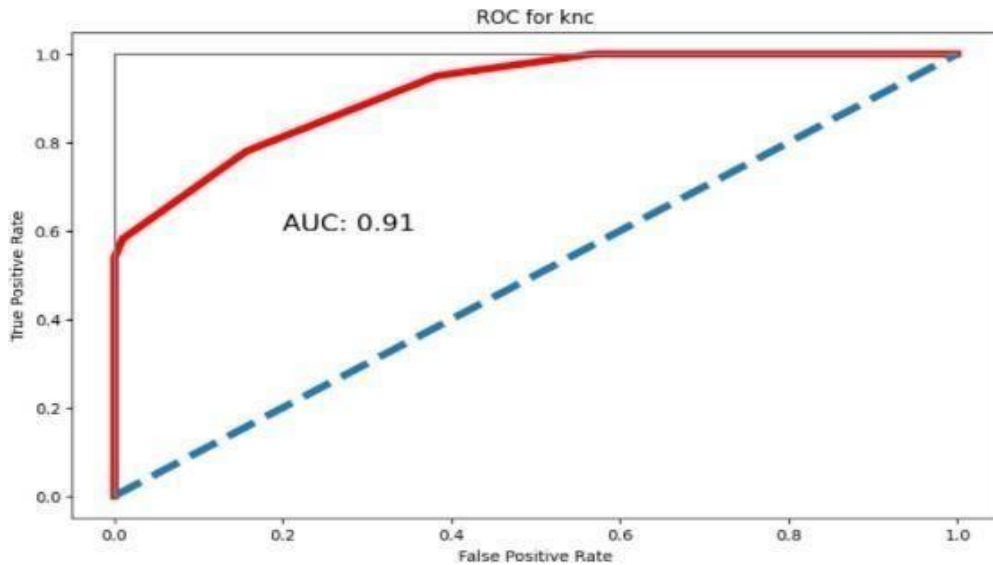


Figure 15: ROC for K Nearest Neighbor

We hereby mentioned that our prediction system predicts 4 algorithms with very high accuracy. So, we've found 4 perfect figure for ROC Analysis. And we see that Random Forest Classifier Shows the best level Accuracy among four algorithms.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Heart disorder is becoming the top cause of mortality all over the world. When a heart condition rises to an advanced level, dangerous quantities of electrolytes, wastage, and fluid can remain in our bodies. There is a shortage of heart disorder, particularly heart disorder, as a result of the world health community's, they are focused on cardiac problems, as well as a lack of awareness. It is critical that developed countries focus on this and establish stronger inclusive ideals in order to reduce heart disorder. This will be a dangerous cause of death for a person. Therefore people should be aware for his/her health and also avoid those work from happening in the very initial stages of his/her life. Several machine learning classifiers were executed in this work to determine the optimal accuracy, recall, ROC, f measure and precision. Random forest, on the other hand, has a 99.18 percent accuracy rate and a ROC value of 100.

5.2 Future Work

For a data mining process, data is everything. More data, hidden patterns, and dimensions will be delivered with greater precision. This dataset had 1203 records, which is just enough for a better prediction, but we will add additional data to the dataset in the future. So that we can compute and get a satisfactory level of accuracy. More information can be applied to this dataset in the future and the most significant variety of data would be a plus point for further analysis. Furthermore, deep learning and other related approaches will be employed to apply our findings.

REFERENCES

- [1] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2020). Coronary artery heart disorder prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, 1-20.
- [2] Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: an effective heart disorder prediction model for a clinical decision support system. *IEEE Access*, 8, 133034-133050.
- [3] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disorder prediction using machine learning. *International Journal of Research and Technology*, 9(04), 659-662.
- [4] Singh, A., & Kumar, R. (2020, February). Heart disorder prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.
- [5] Kumar, P., VA, S. P., Maheshwari, R., & Gowda, S. D. (2021). A Comparative Study of Machine Learning Techniques in Heart Disorder Detection. *Perspectives in Communication, Embedded-systems and Signal-processing-PiCES*, 4(10), 264-272.
- [6] Escamilla, A. K. G., El Hassani, A. H., & Andres, E. (2019). Dimensionality Reduction in Supervised Models-based for Heart Failure Prediction.
- [7] Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disorder prediction based on ensemble deep learning and feature fusion. *Information Fusion*, 63, 208-222.
- [8] Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disorder prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, 1-13.
- [9] Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disorder prediction: a comparative study and analysis. *Health and Technology*, 11(1), 87-97.
- [10] Almustafa, K. M. (2020). Prediction of heart disorder and classifiers' sensitivity analysis. *BMC bioinformatics*, 21(1), 1-18.
- [11] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disorder prediction using machine learning techniques. *SN Computer Science*, 1(6), 1-6.
- [12] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7, 81542-81554.

- [13] Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, 63, 208-222.
- [14] Khan, M. A. (2020). An IoT framework for heart disease prediction based on MDCNN classifier. *IEEE Access*, 8, 34717-34727.
- [15] Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering and Systems*, 12(1), 242-252.
- [16] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In 2019 16th international bhurban conference on applied sciences and technology (IBCAST) (pp. 619-623). IEEE.
- [17] Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network. *IEEE Access*, 7, 34938-34945.
- [18] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 2018.
- [19] Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., & Singh, P. (2021). Latest trends on heart disease prediction using machine learning and image fusion. *Materials Today: Proceedings*, 37, 3213-3218.
- [20] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, 16, 100203.

APPENDICES

Abbreviation

- HD = Heart Disorder
- ROC = Receiver Operating Characteristic
- NB = Naive Bayes
- KNN = K Nearest Neighbor
- RFC = Request For Comments
- DTC = Decision Tree Classifier

Appendix: Reflections of Research

Start our project, having limited expertise with ML methods for observing and planning. Supervisor sir was quite pleasant and helpful to us. Sir provided with good advice and was quite helpful. During this study period, we learned a lot of new knowledge, approaches, and how to apply new algorithms, as well as how to deal with diverse methodologies. We encounter several challenges when we first begin working with this, but we progressively grow more acquainted with these methods.

Ultimately, by finishing this paper, we were able to get a great deal of knowledge as well as abilities.

Project ID: FL20D365 for "Performance Analysis of Heart Disorder Prediction Using Machine Learning Approaches"

ORIGINALITY REPORT

11%

SIMILARITY INDEX

8%

INTERNET SOURCES

5%

PUBLICATIONS

6%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Daffodil International University

Student Paper

1%

2

Submitted to The University of the South Pacific

Student Paper

1%

3

Submitted to Higher Education Commission Pakistan

Student Paper

1%

4

dspace.daffodilvarsity.edu.bd:8080

Internet Source

1%

5

Minhaz Uddin Emon, Al Mahmud Imran, Rakibul Islam, Maria Sultana Keya, Raihana Zannat, Ohidujjaman. "Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches", 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021

Publication

1%

6

Md Shahin Ali, Md. Khairul Islam, Jahurul Haque, A Arjan Das, D S Duranta, Md Ariful

1%

Islam. "Alzheimer's Disease Detection Using m-Random Forest Algorithm with Optimum Features Extraction", 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), 2021

Publication

7	link.springer.com Internet Source	1 %
8	Submitted to Chiang Mai University Student Paper	1 %
9	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
10	www.coursehero.com Internet Source	<1 %
11	dokumen.pub Internet Source	<1 %
12	www.datasklr.com Internet Source	<1 %
13	dspace.bracu.ac.bd Internet Source	<1 %
14	Submitted to Anglia Ruskin University Student Paper	<1 %
15	Submitted to Cardiff University Student Paper	<1 %

16	scholar.uwindsor.ca Internet Source	<1 %
17	huggingface.co Internet Source	<1 %
18	publisher.unimas.my Internet Source	<1 %
19	dspace.sctimst.ac.in Internet Source	<1 %
20	ramanroshana.wordpress.com Internet Source	<1 %
21	Milija Suknovic, Boris Delibasic, Milos Jovanovic, Milan Vukicevic, Dragana Becejski-Vujaklija, Zoran Obradovic. "Reusable components in decision tree induction algorithms", Computational Statistics, 2011 Publication	<1 %
22	Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès. "Classification models for heart disease prediction using feature selection and PCA", Informatics in Medicine Unlocked, 2020 Publication	<1 %
23	dergipark.org.tr Internet Source	<1 %
24	ebin.pub Internet Source	<1 %

25 medium.com <1 %
Internet Source

26 www.slideshare.net <1 %
Internet Source

27 zoektdelle.com <1 %
Internet Source

28 ijream.org <1 %
Internet Source
