Department of Electrical and Computer Engineering

North South University

Heart Disease Prediction using Machine Learning

Submitted by:

Ishrat Jahan 1921909042

Maria Rahman Jui

192236642

Faculty Advisor:

Sarnali Basak (SLB)

Department of Electrical and Computer EngineeringNorth South University

Section: 07

Submission Date: 10/06/23

CORRECTIONS

- Mention the number of Data points
- Remove Heart Disease from Categorical Features
- Implement SMOTE
- Correction of Confusion Matrix to match the number of test data
- Implement SVM model

Abstract- Heart disease is the leading cause of death worldwide therefore an early detection and prevention is critical. The project mainly focuses on predicting which patient has heart disease and which one has a healthy heart based on several health parameters and machine learning algorithms.

Key words: Heart disease; classification; machine learning; algorithms; detection

I. INTRODUCTION

Heart disease, also known as cardiovascular disease, is a group of conditions that affect the heart and blood vessels. Heart disease is a leading cause of death worldwide, with risk factors including high blood pressure, high cholesterol, smoking, obesity etc. Early detection of heart disease is important because it allows for timely intervention and management of the condition. The aim of our project is to predict if a person has Heart Disease using Machine Learning algorithms based on several health parameters. The following parameters will be counted as input and based on the values the algorithm will predict the result. It'll eventually help a person with the diseased heart to know abouthis/her heart condition and seek proper medical care and treatment.

II. LITERATURE REVIEW

In **[1]** R.Radhika and S. Thomas Geaorge applied machine learning algorithms like Logistic Regression, KNN, SVM, Naïve Bayes, Decision Tree and Random forest to classify heart disease. Indexes such as age, gender, chest pain were the attributes of the dataset they used. According to their models, among every one of the calculation KNN and Random Forest had the best accuracy of 88.52%.

In **[2]** data mining techniques were used by the authors along with three machine learning algorithms to predict the heart disease. The dataset they used had the attributes and the algorithm that worked best was KNN and achieved accuracy was 88.52%.

In a study **[3]** published by Alotalibi the author utilized a dataset from Cleveland Clinic and implemented various machine learning techniques. The models used were Decision Tree, Logistic regression, Random Forest, Naive Bayes and SVM. It used 10 fold cross validation and obtained a highest accuracy of 93.19% accuracy. SVM on the other hand had the second highest accuracy of 92.30%.

Another study was conducted by Shah et al aimed**[4]** to develop a model by using machine learning for cardiovascular diseases detection which used a heart disease dataset having 17 attributes. The author used supervised classification method, Naive Bayes, KNN, Random Forest, Decision Tree. On this project KNN achieved the highest accuracy of 90.8%.

III.METHODOLOGY

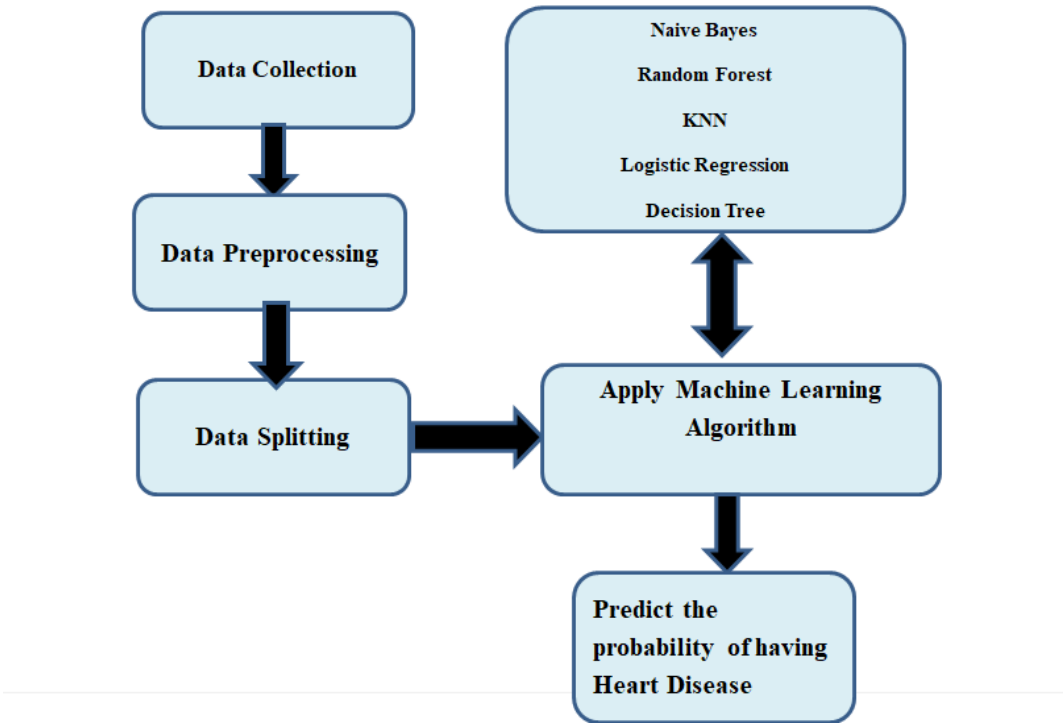


Figure-1: System diagram

- **Data Collection:** The input for this system will be a set of data taken from dataset provider kaggle. This dataset contains different health parameters which are used for detection of Heart Disease.

- **Data Preprocessing:** Preprocessing is required to improve quality of data so that models can analyze it more efficiently. In our system, we used SMOTE which is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier. We also checked for any null values/outliers and deleted the duplicated ones.

- **Data splitting:** Train-test split is required for understanding how a model will perform on new data. The training portion of the dataset will be used for model training and the testing portion will be used for evaluating the performance of trained models. We will use 80% of our dataset for training and 20% for testing purposes.

- **Apply Machine Learning algorithm :** In our system, we used the following machine learning algorithms:

  Logistic Regression: Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. **[5]**

  SVM: Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane**. [6]**

  Random Forest: Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that *is* used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. **[7]**

  KNN: The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another. **[8]**

  Naïve Bayes: The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes. **[9]**

  Decision Tree: A decision tree is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm. It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems. It is one of the very powerful algorithms. **[10]**

- **Prediction:** Based on the trained model, the system will provide output indicating the target class. Model accuracy will be evaluated by measuring how much accurately the model has classified the information along with F1-score, Recall and Precision.

- **Select most optimum Algorithm:** From all the algorithms applied on the dataset, we will select the algorithm with most robust performance. For evaluating performance, the algorithm providing high accuracy will be considered as more efficient.

IV. DATA SET

For our analysis, we have collected from the open-source dataset provider Kaggle **[11].** Pre- processing is required to check for any nulls values in the dataset. It removes missing or inconsistent data values which will help improve the accuracy and quality of dataset. It'll make the data more consistent.

A. Dataset Description

The dataset titled "Heart Disease Prediction" [1] includes 18 attributes, including the predicted attribute which predicts whether the person has heart disease or not based on the values of the attributes. The attributes which the algorithms take as input includes **BMI** (body mass Index), **Smoking**: Whether the user smoked at least 100 cigarettes during the course of their life. **Alcohol Drinking**: If the user is a heavy user of alcohol. **Stroke**: Any prior history of stroke. **Physical Health**: Includes physical illness and injuries, for how many days during the past 30days when physical health was not good? (0-30 days). **Mental Health**: For how many days during the past 30 days had the user's mental health declined? (0-30 days). **DiffWalking**: Any significant difficulty walking or climbing stairs? **Sex**: male or female? **Age Category**: Fourteen-level age category (then calculated the mean.) **Race**: Imputed race or ethnicity value. **Diabetic**: Do you have diabetes or not? **Physical Activity**: Engaged in physical activity or exercise occurred during the past 30days? **GenHealth**: Is the user's health generally good? **Sleep Time**: On average, how many hours of sleep the user get in a day? **Asthma**: Do you suffer from asthma? **Kidney Disease**: Not including kidney stones, bladder infections, or incontinence, did the user have any other kidney disease? **Skin Cancer**: Have any history of skin cancer? **Target**: Whether the user has heart disease or not?

Overall the number of data points used in the project is 1999 with 18 columns and 0 duplicated rows and two unique target values Yes/No which are to be predicted. It is the most recent datasetwhich includes data from 2020 and have been through some cleaning so it would be usable for machine learning projects. The data collected are from annual CDC survey for people of most races in the US. The classes are not balanced so the weights need to be fixed.

The most relevant factors for predicting the specific output would be any prior history of stroke, smoking and alcohol drinking as these often leads to deterioration in heart functions.

target	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
No	16.6	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes	Very good	7	No	No	No
No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No
Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No	Fair	12	No	No	No
No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes	Fair	4	Yes	No	Yes

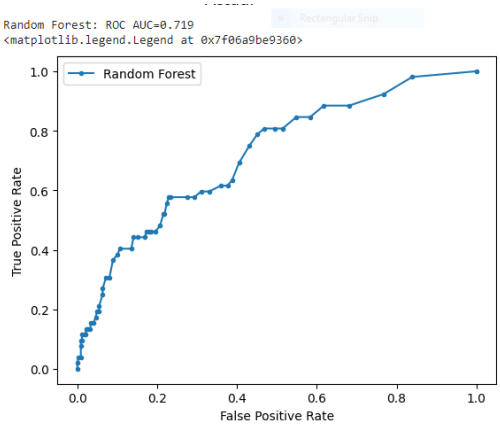
Figure-2: Detail of all instances and target of the dataset

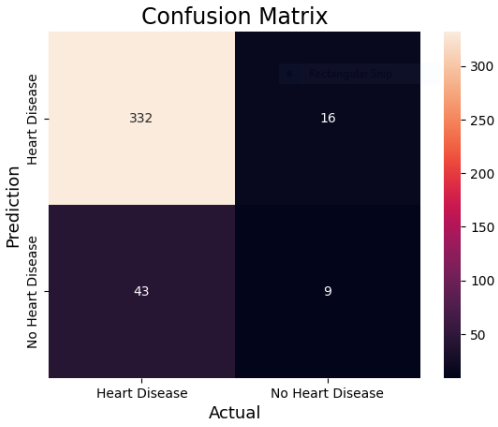
V. Result and Analysis

Utilizing binary classification techniques, we have used different machine learning models for the classification whether a person has Heart Disease or not. We have implemented the Logistic Regression, Random Forest, KNN, Naïve Bayes and Decision tree algorithms. For all the algorithms, we have split our dataset by keeping 80% for training and 20% for testing. After that, we preprocessed our dataset checking for null or duplicate values. The performance of each algorithm in terms of accuracy, precision recall, confusion matrix, F1 score, and ROC described below.

- **Random Forest**

For the dataset Random Forest achieves an accuracy of 85.25%.



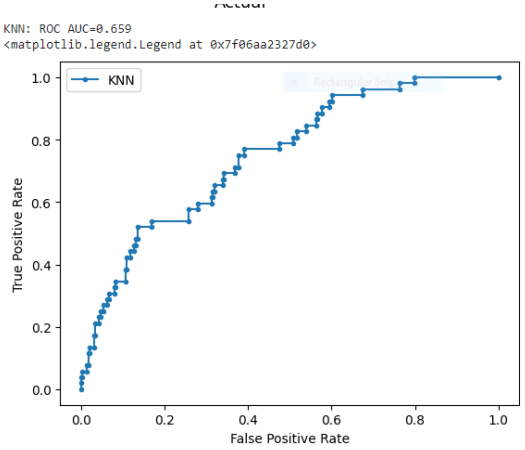


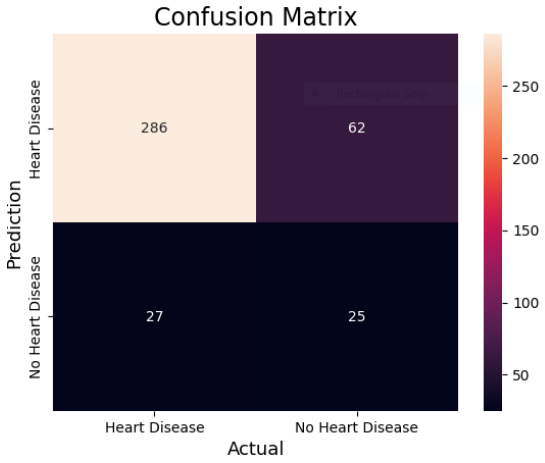
1(a) **1(b) Confusion Matrix**

In 1(a) the curves depicts that Random Forest achieves an ROC AUC of 71.9%. In 1(b) the confusion matrix for Random Forest depicts the algorithm is able to classify 332 as TP, 9 as TN, 43 as FN and 16 as FP from the 400 test data available.

- **KNN**

For the dataset KNN achieves an accuracy of 77.5%.



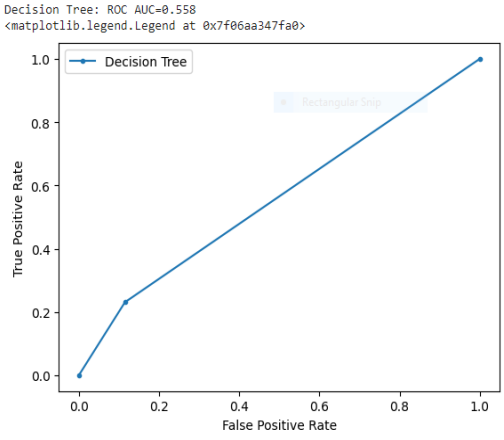


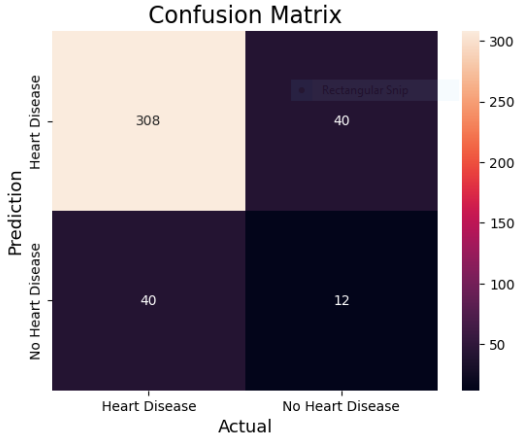
2(a) ROC **2(b) Confusion Matrix**

In 2(a) the curve depicts KNN achieves an AUC of 65.90%. In 2(b) the confusion matrix for KNN depicts the algorithm is able to classify 286 as TP, 25 as TN, 27 as FN and 62 as FP from the 400 test data available.

- **Decision Tree**

For the dataset Decision Tree achieves an accuracy of 80.00%.



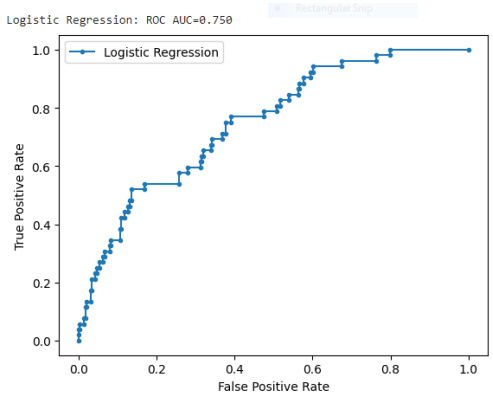


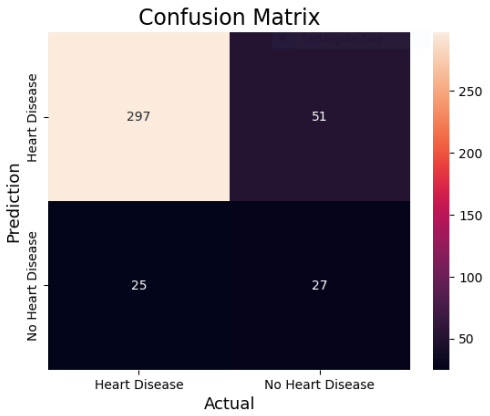
3(a) ROC **3(b)Confusion Matrix**

In 3(a) the curve depicts Decision Tree achieves an ROC AUC of 55.8%. In 3(b) the confusion matrix for Decision Tree depicts the algorithm is able to classify 308 as TP, 12 as TN, 40 as FN and 40 as FP among the 400 test data available.

- **Logistic Regression**

For the dataset random forest achieves an accuracy of 81.00%.



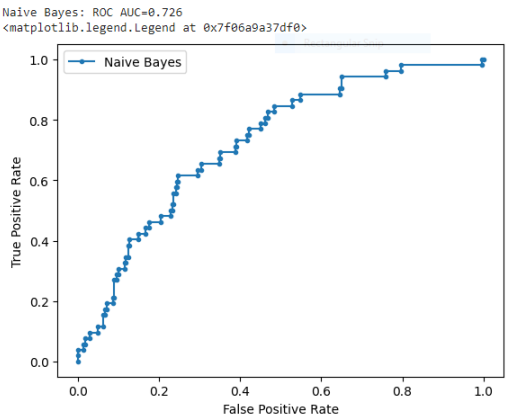


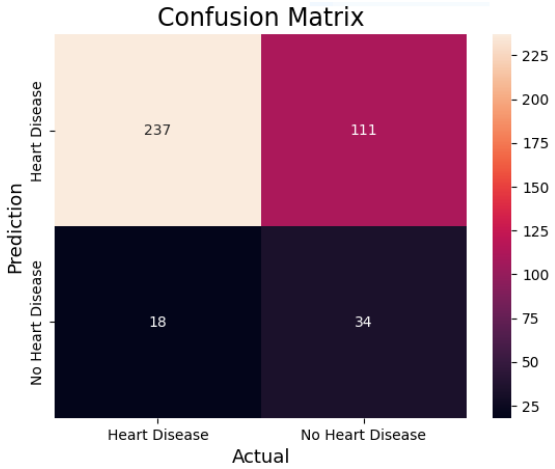
4(a) ROC **4(b) Confusion Matrix**

In 4(a) the curve depicts Logistic Regression achieves an ROC AUC of 75.0%. In 4(b) the confusion matrix for Logistic Regression depicts the algorithm is able to classify 297 as TP, 27 as TN, 25 FN and 51 as FP among the 400 test data available.

- **Naïve Bayes**

For the dataset Naïve Bayes achieves an accuracy of 67.75%.

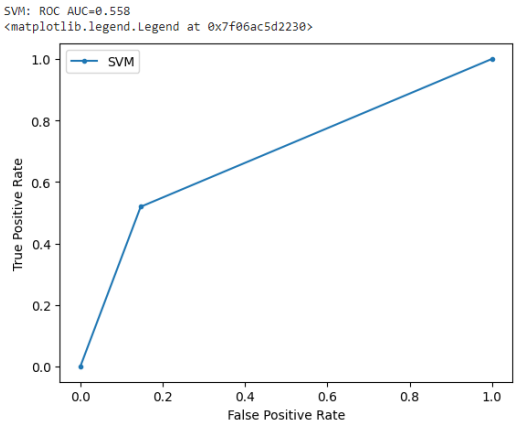


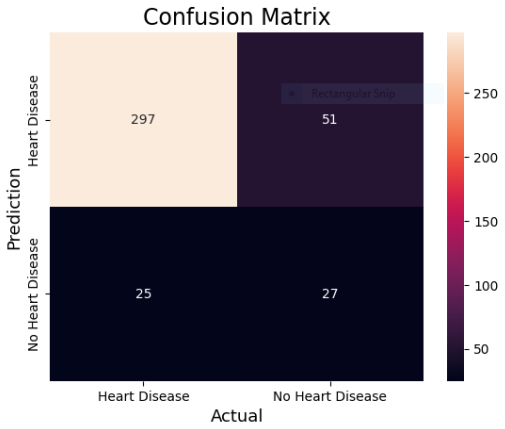


5(a) ROC **5(b)Confusion Matrix**

In 5(a) the curve depicts Naïve Bayes achieves an ROC AUC of 72.6%. In 5(b) the confusion matrix for Naïve Bayes depicts the algorithm is able to classify 237 as TP, 34 as TN, 18 as FN and 111 as FP among the 400 test data available.

- **SVM**





6(a) ROC **6(b)Confusion Matrix**

In 6(a) the curve depicts SVM achieves an ROC AUC of 55.8%. In 6(b) the confusion matrix for SVM depicts the algorithm is able to classify 297 as TP, 27 as TN, 25 as FN and 51 as FP among the 400 test data available.

Model Evaluation

The performance analysis of the algorithms is evaluated based on accuracy, precision, recall, and F1-score. The performance of the proposed algorithms were assessed using the terms true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Precision is the opposite of recall. The F1-score is a combined measure of precision and recall, which shows how often the predicted value is accurate. It is also known as the harmonic mean of p and r in mathematics. These equations are given below. Accuracy is a measure of how well a model works for all classes. Below are the mathematical formulas for calculating the parameters.

Accuracy: It shows the overall performance of the classification model and can be calculated by the formula given below:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall: It is the ratio of classified heart patients to the total patients having heart disease. It means the model prediction is positive and the person has heart disease. The formula forcalculating recall:

$$re - call = \frac{TP}{TP + FN}$$

Precision: It is the ratio of the actual positive score and the positive score predicted by the classification model. Precision can be calculated by the following formula:

$$precision = \frac{TP}{TP+FP}.$$

F1-score: It is the weighted measure of both recall and precision. Its value ranges between 0 and 1.If its value is close to one then it means the good performance of the classification algorithmand if its value is close to 0 then it means the bad performance of the classification algorithm

$$\text{F1-score} = 2pr/(p+r)$$

The model evaluation of accuracy, precision, recall, and F1-score of Random forest, KNN, Decision Tree, Logistic Regression, Naive Bayes and SVM are given in Table 1.

Table 1: Model evaluation

Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	0.8525	0.820	0.850	0.830
KNN	0.775	0.830	0.780	0.800
Decision Tree	0.800	0.800	0.800	0.800
Logistic Regression	0.810	0.850	0.810	0.830
Naïve Bayes	0.6775	0.840	0.680	0.730
SVM	0.810	0.850	0.810	0.830

From the Table:1, we can see among all machine learning algorithms Random Forest achieves the highest accuracy of 85.25% for binary classification on the given dataset. It also has the high F1-score of 0.830 along with Logistic Regression and SVM along with the highest recall among all other algorithms among all algorithms so it can be declared the most optimum model for the given dataset. Apart from Random Forest, SVM and Logistic Regression performs well with an accuracy of 81% and 81% respectively. Among all the algorithms, Naive Bayes performs poorly achieving the lowest accuracy of 67.75%.

VI. Conclusion

The project's main objective was to categorize heart disease using various machine learning models and a real-world dataset. In order to discover the best model with the highest accuracy, various machine learning algorithms were applied to a dataset of patients with heart disease. According to the data, Random Forest had the highest accuracy of 85.25%. These results show the machine’s ability to predict heart disease and suggest that the algorithm may be a useful tool in establishing specialized approaches for the detection of Heart Conditions in the health care sector.

VII. REFERENCES

[1] R. Radhika and S. Thomas George, "HEART USING MACHINE LEARNING TECHNIQUES," *Journal of Physics: Conference Series*, vol. 1937, no. 1, p. 012047, Jun. 2021, doi: https://doi.org/10.1088/1742-6596/1937/1/012047.

[2] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath, "Heart disease prediction using machine learning algorithms," *ResearchGate*. Heart disease prediction using machine learning algorithms (accessed May 2022).

[3] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci.* **2020**, *1*, 345.

[4] Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 261–268.

[5] G. Lawton, "What is Logistic Regression? - Definition from SearchBusinessAnalytics," *SearchBusinessAnalytics*, Jan. 2022. https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

[6] JavaTPoint, "Support Vector Machine (SVM) Algorithm - Javatpoint," *www.javatpoint.com*. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[7] S. E R, "Random Forest | Introduction to Random Forest Algorithm," *Analytics Vidhya*, Jun. 17, 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[8] IBM, "What is the k-nearest neighbors algorithm? | IBM," *www.ibm.com*. https://www.ibm.com/topics/knn

[9] IBM, "What is Naïve Bayes | IBM," *www.ibm.com*. https://www.ibm.com/topics/naive-bayes

[10] GeeksForGeeks, "Decision Tree - GeeksforGeeks," *GeeksforGeeks*, Oct. 16, 2017. https://www.geeksforgeeks.org/decision-tree/

[11]D. Lapp, "Heart Disease Dataset," *Kaggle*, 2020. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download&select=heart.csv