CARDIOVASCULAR(HEART) DISEASE PREDICTION USING MACHINE LEARNING MODELS

Ish Phanda, 3CS12, 102016098

LAB INSTRUCTOR- DR ANJULA MEHTO

I. INTRODUCTION

Cardiovascular or heart disease is a broad term used to refer to diseases and conditions that affect the heart and circulatory system. It is considered one of the biggest causes of death among the world's population. According to World Health Organization, heart related diseases are responsible for taking 12 million lives every year. A lot of research has been done in an effort to identify the most influential factors for heart disease as well as accurately predict the overall risk. Early diagnosis of heart disease plays a vital role in making decisions about lifestyle changes in high-risk patients, thus reducing complications [1].

Machine learning has proven to be effective in helping to make decisions and predictions from the large amount of data produced by the healthcare industry. Although heart disease can occur in many different forms, there is a common set of underlying risk factors that influence whether or not a person is ultimately at risk of developing heart disease. By collecting data from different sources, this data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very huge, and more often than not, this data can be very noisy. Thus, these algorithms have become very useful, in recent times, for accurately predicting the presence or absence of heart-related diseases.

A. Motivation

The main motivation for doing this project is to predict future heart disease by analyzing patient data that classifies whether or not they have heart disease using a machine learning algorithm. Moreover, the work of this project aims to define the best classification algorithm for determining the probability of developing heart disease in a patient. It also aims to increase the accuracy with which heart disease risk can be predicted. This work was justified by conducting a comparative study and analysis using five classification algorithms namely, Logistic Regression, SVM, Random Forest, K-Nearest Neighbors and Decision Tree used in different levels of assessments.

B. Problem Definition

The main challenge in cardiology is its discovery. There are tools available that can predict heart disease but they are either too expensive or ineffective to calculate the chance of heart disease in humans. Early detection of heart disease can reduce mortality and general complications. However, it is not possible to monitor patients every day accurately in all cases and 24-hour patient consultation by a physician is not available as it requires more wisdom, time and experience. Since we have a large amount of data in today's world, we can use this data to predict whether or not a patient has heart disease based on different criteria and different machine learning algorithms to analyze the data [1].

II. DATA DESCRIPTION

The heart dataset provides information about heart disease. The dataset contains 1025 samples and 13 input features as well as one output label. Traits describe the personal and health characteristics of patients such as age, gender, fasting blood sugar, and maximum achieved heart rate. There are two data types for all features which are integer and float. The target label is the decision category variable that indicates the presence of heart disease in the patient. The goal here is to classify the target variable to the value 0 for no disease and 1 for disease, using a different machine learning algorithm and finding out the appropriate algorithm for this data set.

The names of the features in the data set are abbreviated and the meaning of these features is difficult to understand. The full medical/technical names are hard to understand for most of us let alone their abbreviated form. To better understand the meaning of the features, I have the responsibility to explain some of the attributes of dataset as follows [4]:

- age: age in years
- sex: sex (1 = male; 0 = female)
- cp: chest pain type (0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic)
- trestbps; resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg: resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- thalach: maximum heart rate achieved
- exang; exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment (0 = upsloping; 1 = flat; 2 = downsloping)
- ca: number of major vessels (0-3) colored by flourosopy
- thal: (0 = error (in the original dataset 0 maps to NaN's); 1 = fixed defect; 2 = normal; 3 = reversable defect)
- target (the lable): (0 = no disease; 1 = disease)

III. DEVELOPMENT ENVIRONMENT

I’ve used Python because it has scikit-learn package.
Scikit-learn is a free software machine learning library for the Python programming language. Simple and efficient tools for predictive data analysis. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means, and is designed to interoperate with the Python numerical and scientific libraries NumPy for array vectorization, Matplotlib and plotly for plotting, SciPy and many more [6].

IV. DATA PREPROCESSING

Data preprocessing is an important step for creating a machine learning model. Initially, I downloaded a dataset from Kaggle [2], and checked for any missing values in the dataset, it turned out that we had no missing values in our database, then I separated the features from the response, then I split the dataset into 80% (820 instances) for the training set and 20% (205 instances) for the test set. Then I noticed that I need to standardize scaling all the values before training the machine learning models.

V. FLOWCHART

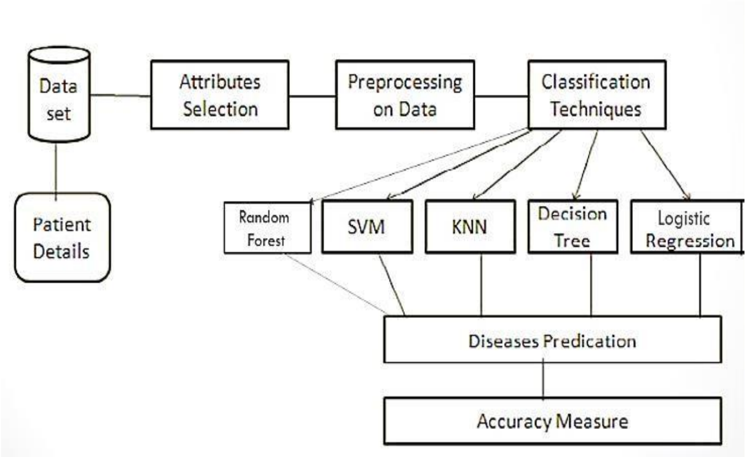

Fig. 1. Architecture of Prediction System

VI. METHODS

This is a binary classification problem (has-heart disease or no-heart disease cases). Scikit-learn offers a wide range of classification algorithms and is often the starting point for most traditional machine learning challenges. In this project, we have tried 5 algorithms for experiments and they are Logistic Regression, SVM, Random Forest, K-Nearest Neighbors and Decision Tree.

A. Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the output of a categorical dependent variable using a given set of independent variables.

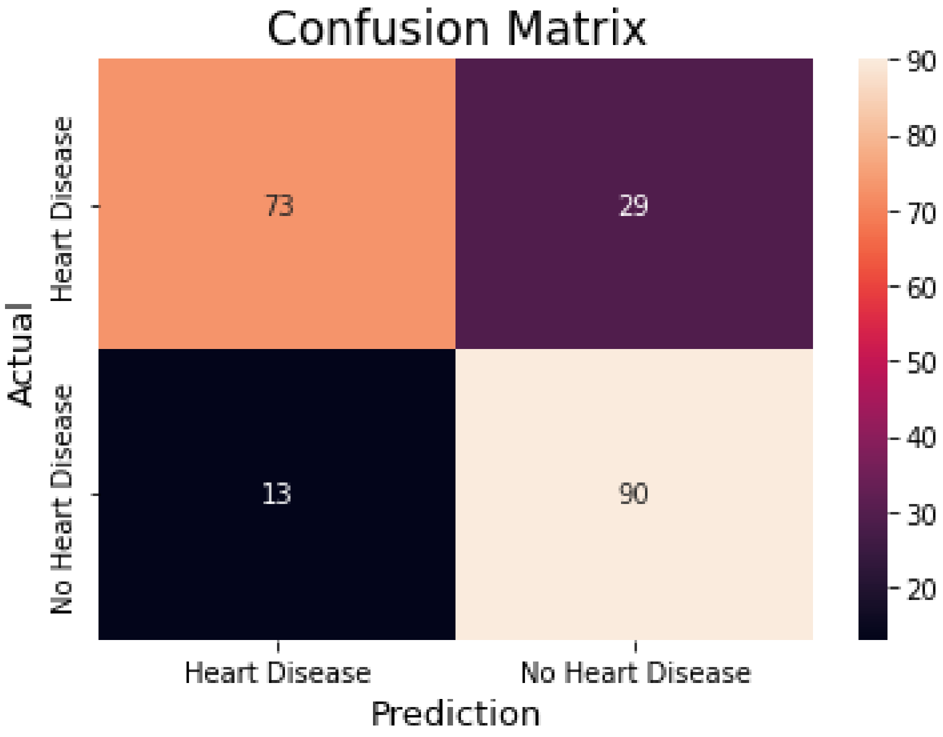

Fig. 2. Confusion Matrix for Logistic Regression

After implementing a machine learning approach for testing and training, we found that the train accuracy is 87.19% and the test accuracy is 79.51%. It works fine but not the best for us. The advantage of logistic regression is that it

does not require a lot of computational resources and is highly explainable. So, it is easy and sufficient to apply logistic regression. However, the limitation of logistic regression is that it assumes that there is linearity between the features of the data set. In the real world, data is rarely separable, nor like our dataset. That is why we cannot reach a very high accuracy of 90% [7]. Accuracy can be calculated with the support of confusion matrix of the Logistic Regression as shown in Fig.2 here number of count of TP, TN, FP, FN are given and using the accuracy equation that is the ratio of the number of correct predictions to the total number of inputs in the dataset [1].

B. Support Vector Machine (SVM)

Support Vector Machines are powerful and flexible supervised machine learning algorithms, which can be used to solve both linear and nonlinear classification as well as regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can separate n- dimensional space into classes so that we can easily put the new data point into the correct class in the future [1][3].

I used the polynomial kernel of degree 3, coef0 parameter =2 and C parameter =8. Training accuracy is 100% and test accuracy is 98.53% which is better than logistic regression and the advantage of SVM is that it is very effective with high dimensional distances. The main drawback is that SVM has many parameters that must be chosen correctly for best performance [7].

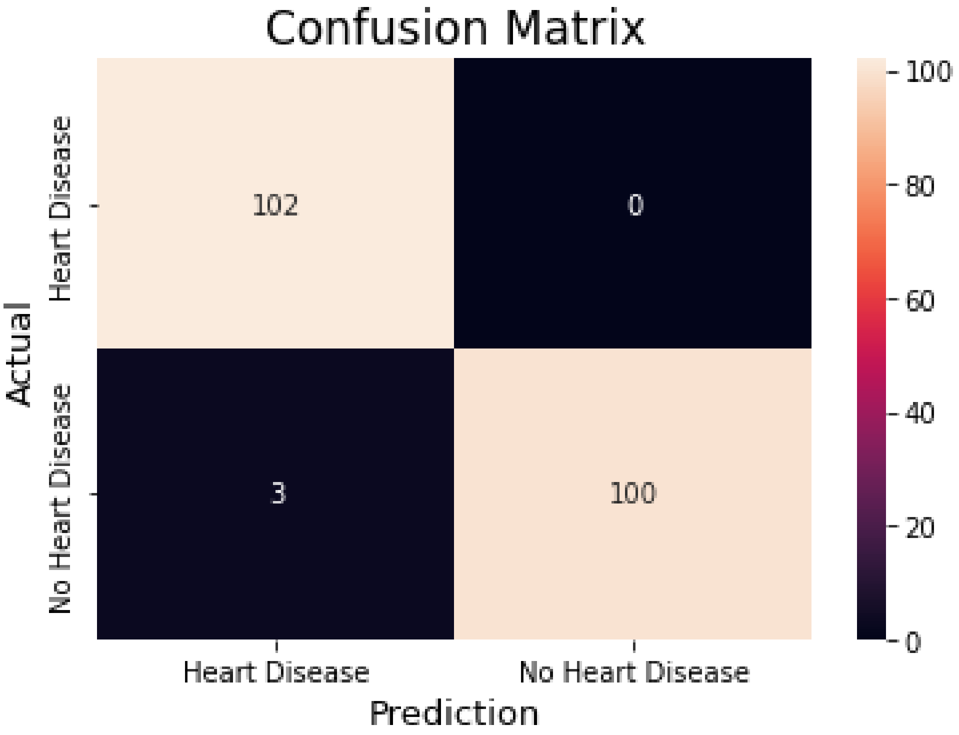

Fig. 3. Confusion Matrix for Support Vector Machine

C. Random Forest

Random Forest is a supervised learning algorithm. It can be used for both regression and classification . It is also very flexible and easy to use algorithm. As the name implies, “Random Forest is a classifier that contains a number of decision trees in different subsets of a given data set and takes the average to improve the predictive accuracy of that data set.” Instead of relying on a single decision tree, a random forest takes the prediction from each tree and bases the majority of the predictions votes on, and predicts the final output. A larger number of trees in a forest results in higher accuracy and prevents the problem of overfitting [1][8].

I used the parameters (n_estimators = 39 which means the number of trees in the forest is 39 and max_depth = 9 that is the maximum depth of the tree). We get 100% train accuracy and also 100% test accuracy. The advantage of Random Forest is that it can handle data set with high features and variance balance and is not sensitive to data noise [7]. As shown in Fig. 4, this graph shows the confusion matrix of Random Forest.

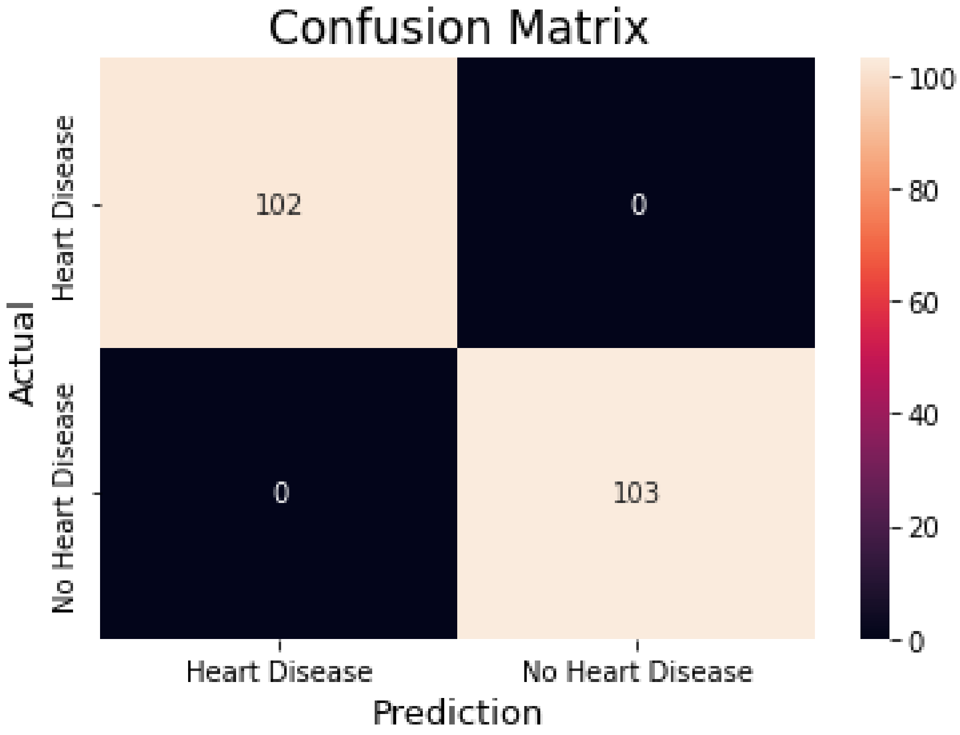

Fig. 4. Confusion Matrix for Random Forest

D. K-Nearest Neighbors

K-Nearest Neighbor is a classification method that uses an imaginary boundary to classify data. When new data points are received, the algorithm will try to predict them as closely as possible from the boundary line. It works on the basis of the distance between the location of the data and on the basis of these distinct data they are categorized with each other. All other data set is called "neighbor of each other" and the number of neighbors is determined by the user which plays a very important role in analyzing the dataset [4].

To evaluate the accuracy, we used only one neighbor (k = 1), and we found that the training accuracy is 100% and the test accuracy is 98.53% which is the same as the accuracy for Support Vector Machines. We can see in Fig.5 the confusion matrix of KNN.

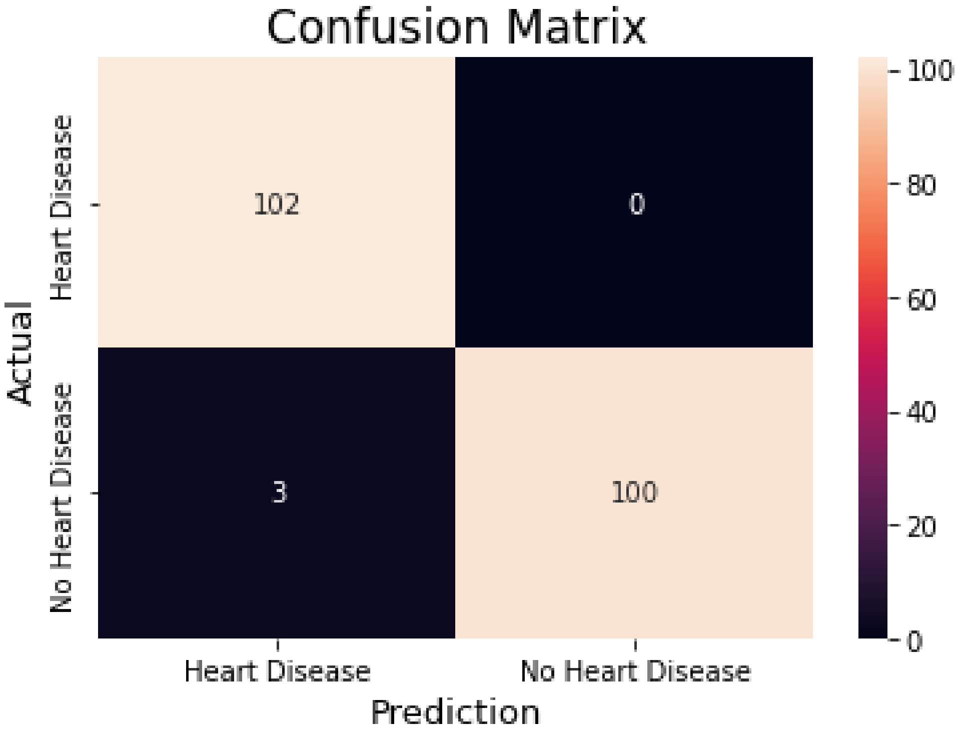

Fig. 5. Confusion Matrix for K-Nearest Neighbor

E. Decision Tree

Decision tree is a supervised learning technique that can be utilized for both classification and regression problems. In a

decision tree, there are two nodes, which are the decision node and the leaf node. Decision nodes are used to make any decision and have multiple branches, whereas leaf nodes are the product of those decisions and do not contain any branches. It is a tree-structured classifier, where the inner nodes represent the features of the dataset, the branches represent the decision rules and each leaf node represents the result. Decisions or testing are performed based on the features of the selected dataset [1].
I used the parameters (max_depth = 10 which is the maximum depth of the tree, min_samples_leaf = 1 that is the minimum number of samples required to be at a leaf node and min_samples_split = 2 which means the minimum number of samples required to split an internal node), we achieved 100% training accuracy and 98.53% test accuracy, which is the same accuracy to Support Vector Machines and K-Nearest Neighbors. Here, in Fig.6 is the confusion matrix of the Decision Tree.

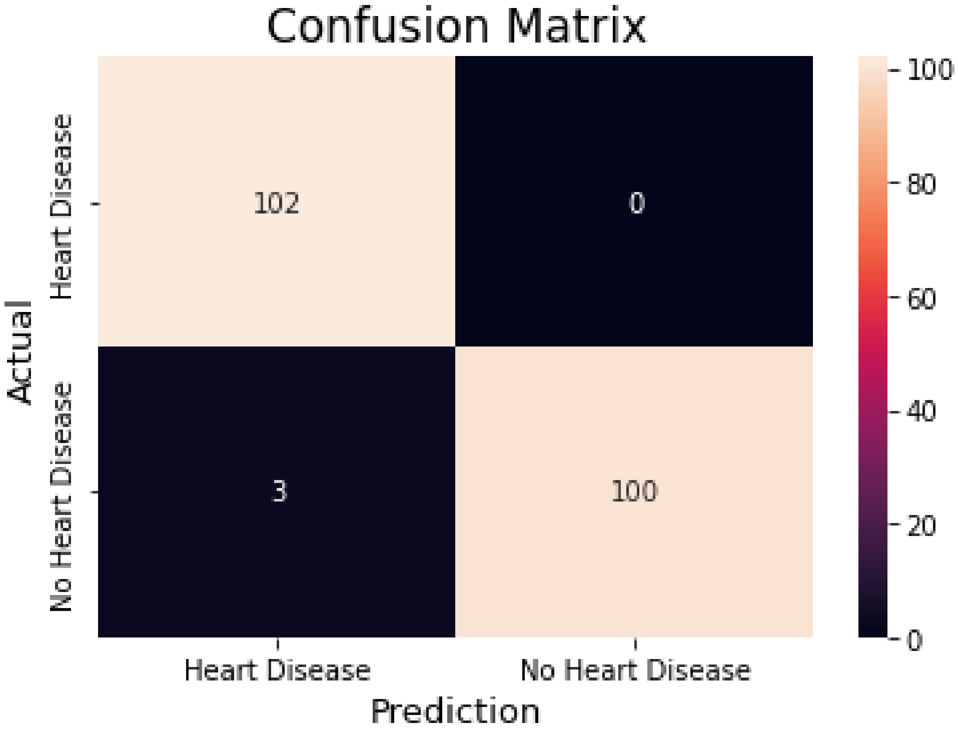

Fig. 6. Confusion Matrix for Decision Tree

VII. RESULT AND DISCUSSION

Now we've seen all the precision values for classifiers, it's time to make the decision to choose the best possible classifier algorithm. We found that Random Forest gives us the highest accuracy of 100%, followed by 98.53% for SVM, KNN and Decision Tree, these three algorithms give us the same accuracy.

Comparing the results for the different algorithms we used in this project. Therefore, the best algorithm we can use to predict heart disease is Random Forest which is the more efficient compared to other algorithms and gives us the best test accuracy, which is 100%. The reason why it outperforms the others is because it is not limited to the property of the dataset.

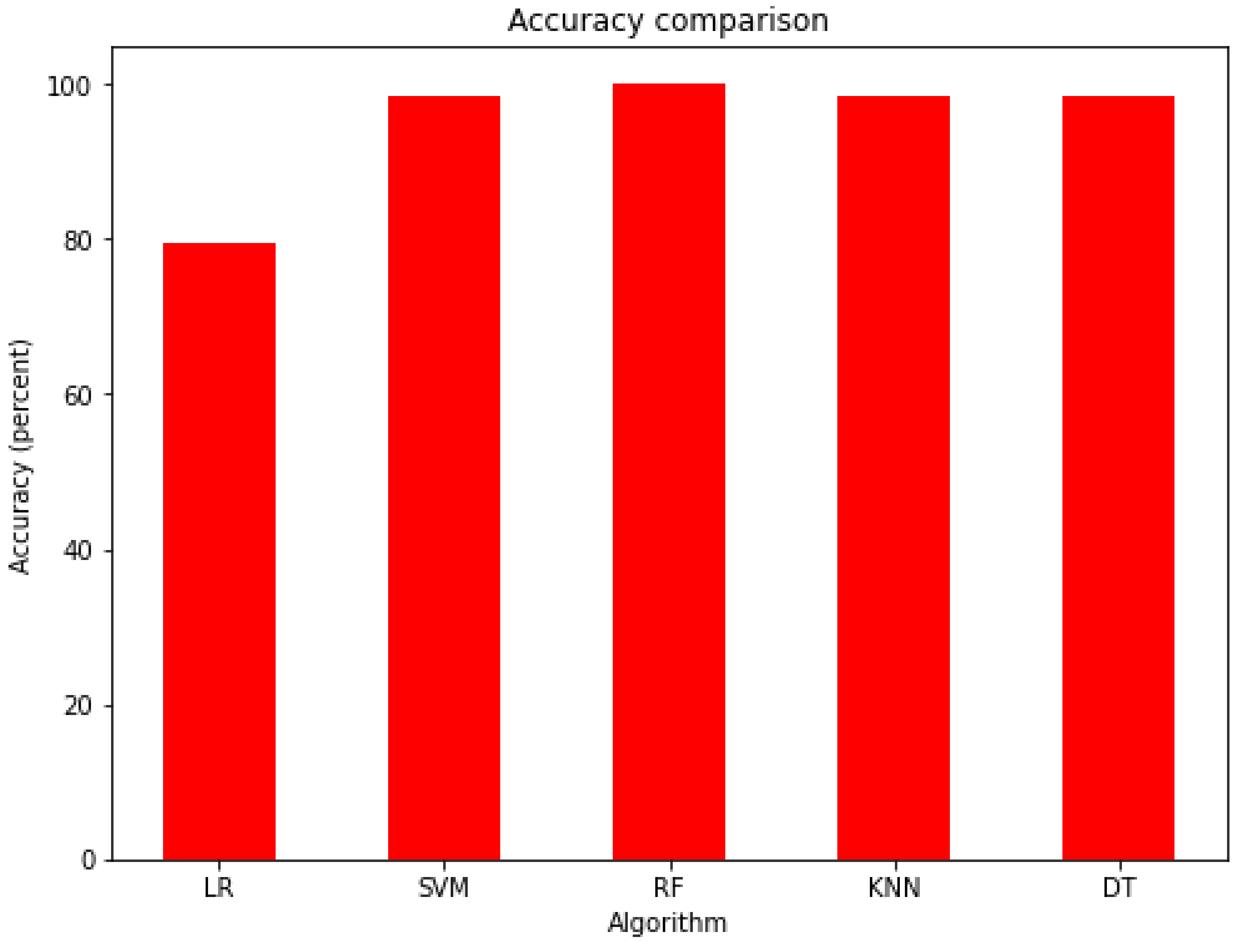

Fig. 6. Accuracy Comparison of all the Models

Finally, the algorithm that gives us the least accuracy is a logistic regression of 79.51%, as shown in Fig. 7

VIII. CONCLUSION

Finally, heart diseases are a major killer in the world, application of promising technology such as machine learning to the initial prediction of heart disease will have a profound impact on society and could be a milestone in the field of medicine. In this project, 5 different machine learning algorithms used to measure performance are SVM, Decision Tree, Random Forest, K-Nearest Neighbor, and Logistic Regression applied to the dataset. After the experiments, the Random Forest algorithm gives us the highest test accuracy, which is 100%. Although we get a good result, this is not enough because it does not guarantee that a misdiagnosis will not occur. To improve accuracy, we hope to require more of the dataset because 1025 instances of the dataset are not enough to do an excellent job.

In the future, machine learning approach will be used for better analysis of heart disease and for early disease prediction so that death rate can be reduced through disease awareness and an intelligent system can be developed that can lead to the selection of appropriate treatment modalities for a patient diagnosed with heart disease. To predict disease we want to try different diseases like lung cancer by using image detection. In this way, the data set becomes complex and we can apply the convolutional neural network to make accurate predictions [4][7].

IX. REFERENCES

[1] Pranitha, Gunturu, Cherukuri , Koruprolu, and Kesuboyina. "Heart Disease Prediction Using Machine Learning Algorithms.", 2017-2021.

[2] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[3] Shruti Patil and Mrunal Annadate. “Implementation of Machine Learning Model to Predict Heart Problem.”, January 2022.

[4] Archana Singh and Rakesh Kumar. “Heart Disease Prediction Using Machine Learning Algorithms.”, 2020.

[5] Wikipedia, " Python (programming language).", 21 May 2022.

[6] Wikipedia, " scikit-learn .", 14 January 2022.

[7] Yangguang He, Xinlong Li and Ruixian Song. “Heart Disease Detection Project Report.”.

[8] Baban Uttamrao Rindhe. “Heart Disease Prediction Using Machine Learning.”, May 2021.

X. APPENDIX

Source Code

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
https://www.geeksforgeeks.org/bar-plot-in-matplotlib/
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

My Code

https://github.com/IshPhanda/CardioVascular-Disease-Prediction-Model-Machine-Learning-Project

***