HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

Project report

By

AKHILA DURGEMPUDI	-	Z643Z855
KISHORE POOJARI	-	G397M563
PAVAN KUMAR PUJARI	-	Y927P247
RAKESH KUMAR DONGARI	-	N254V945
SAI KRISHNA CHAPALAMADUGU	-	P949S628
SHIVAJI REDDY SAMA	-	F454W858

Abstract

The goal of our project is to use a variety of patient data to determine whether or not patients have heart disease. The motivation behind this project is to conserve human resources in medical centers while also improving diagnosis accuracy. We use a variety of algorithms to detect heart disease in our project, including Logistic Regression, SVM, Decision Trees, and Random Forest. Random Forest, out of all of these algorithms, has the best accuracy of 87.6 %.

Introduction

Our objective is to use the data with various features from the heart dataset to predict whether or not a person has heart disease. This is critical in the medical industry. We can not only avoid incorrect diagnoses but also save human resources if such a prediction is accurate enough. When a patient without a heart condition is diagnosed with one, he will experience unnecessary worry, and when a patient with a heart condition is not diagnosed with one, he will miss the best chance to cure his disease. Both patients and hospitals suffer because of such inaccurate diagnoses. We can avoid unneeded difficulties by making precise predictions.

Furthermore, if we can utilize the machine learning technology to predict medical outcomes, we will save human resources by eliminating the need for a complex diagnosis process in hospitals. Our algorithms receives 12 attributes with numerical values and categorical values as input. To classify whether a person has heart disease or not, we use various algorithms such as Logistic Regression, SVM, Decision Tree, and Random Forest.

DATASET AND FEATURES

This dataset contains 918 observations and 12 features where 7 are numerical features and 5 are categorical features.

Our Predictor (Y, Positive or Negative diagnosis of Heart Disease) is determined by 12 features (X):

• age: age in years

• sex: sex (M = male; F = female)

• ChestPainType

-- TA: typical angina

-- ATA: atypical angina

-- NAP: non-anginal pain

-- ASY: asymptomatic

• RestingBP: resting blood pressure (in mm Hg on admission to the hospital)

• Cholesterol: serum cholesterol in mg/dl

• FastingBS: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

• RestingECG: resting electrocardiography results (Normal = 552; LVH = 188; ST = 178)

• MaxHR:

• ExerciseAngina: exercise induced angina (Y = yes; N = no)

• Oldpeak: ST depression induced by exercise relative to rest

• ST_Slope: Slope of the peak exercise ST segment (Up = up sloping; Flat = flat; Down = down sloping)

• Target: Heart Disease (0 = no, 1 = yes)

Pre-processing of Data

The dataset does not contain any missing values. So, we then created dummy variables for the categorical variables such as Sex, ChestPainType, RestingECG, ExcerciseAngina, ST_Slope where the value 1 indicate the presence of that feature and value 0 indicate the absence of that feature. We have split the dataset into 70% observations for training and 30% observations for testing and did normalization on the dataset to avoid overfitting.

METHODS

The four classification algorithms used for this project are Logistic Regression, Decision Trees, SVM and Random Forest.

1)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 categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, like in our data where 1 indicates that a person has a heart disease and 0 indicates no heart disease.

2)Decision Tree:

Decision Tree is a Supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

3) SVM(Support vector machine):

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 hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

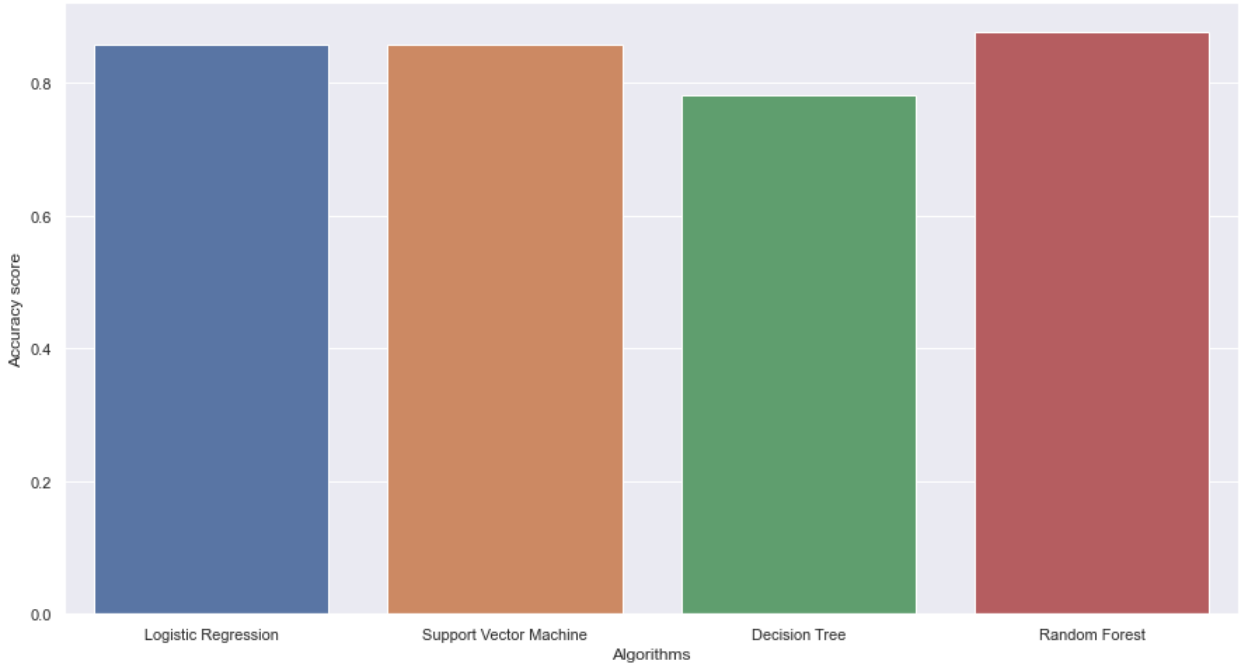
4)Random Forest:

Random Forest is an ensemble learning method for classification and regression by constructing multiple decision trees in training and outputting the classification or prediction(regression). The goal of Random Forest is to combine weak leaning models into a strong and robust leaning model. The algorithm of Random Forest can be summarized in 4 steps: Step 1:Randomly draw M bootstrap samples from the training set with replacement. Step 2: Grow a decision tree from the bootstrap samples. At each node: Randomly select K features without replacement and split the node by finding the best cut among the selected features that maximizes the information gain. Step 3:Repeat the steps 1 and 2 T times to get T trees; Step 4:Aggregate the predictions made by different trees via the majority vote.

RESULTS

Since our project is a classification problem, we use accuracy, precision, recall and F1 score to evaluate the models.

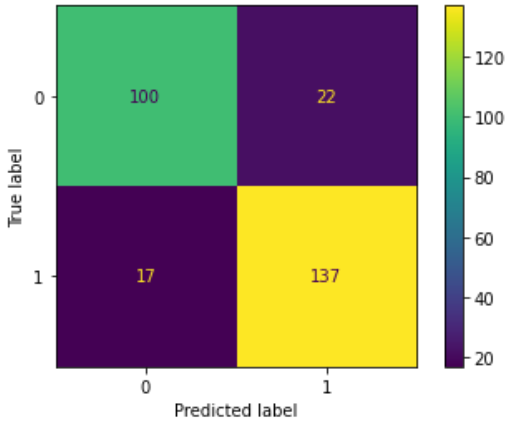
Methods	Accuracy	Precision	Recall	F1 score
Logistic Regression	85%	0.86	0.85	0.86
Decision Trees	78%	0.78	0.78	0.78
Random Forest	87%	0.88	0.87	0.87
Support Vector Machine	85%	0.86	0.85	0.85



ACCURACY SCORES PLOT

1)Logistic Regression

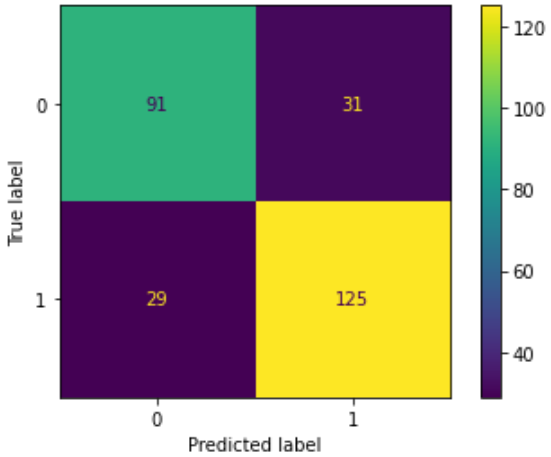
The confusion matrix of the Logistic Regression:



The Logistic regression Model gave us an accuracy of 85.86%. The Logistic Regression has the advantage of using less processing resources and being extremely interpretable. As a result, applying Logistic Regression is simple and sufficient. However, Logistic Regression has a drawback in that it assumes linearity between the dataset's features. In the real world, data is rarely separable, and our dataset is no exception, thus it's possible that's why we can't achieve high accuracy.

2)Decision Tree

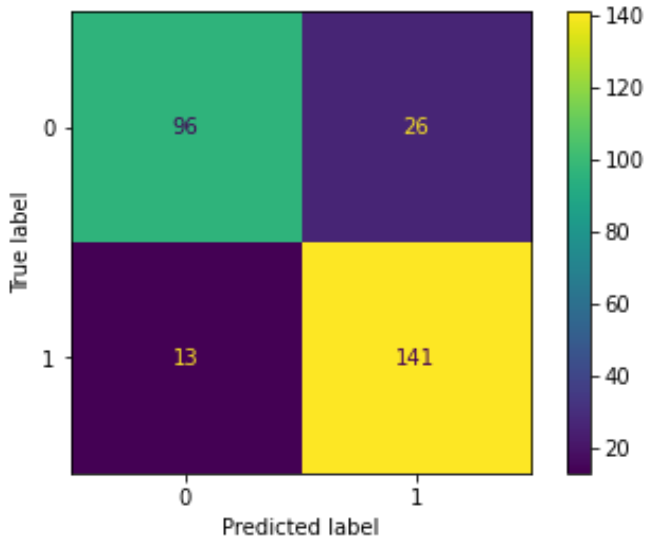
The confusion matrix of Random Forest is:



The Decision Tree model has an accuracy score of 78.26%. When compared to other models, it performs poorly. The advantage of Decision Trees is that they can make predictions with a small amount of training data, but the disadvantage is that they assume all features are mutually independent. In reality, we rarely encounter a dataset with mutually independent attributes, which may explain why we can't achieve very high accuracy.

3)SVM

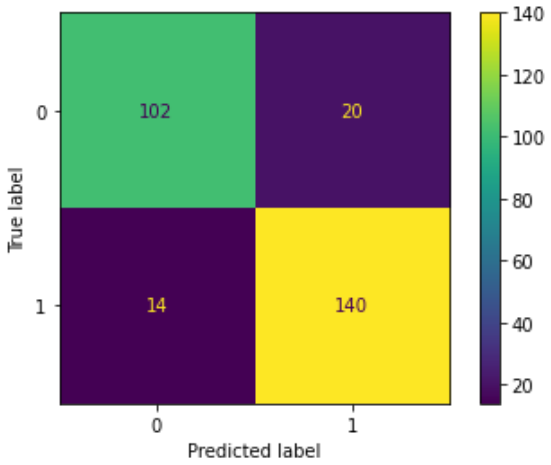
The confusion matrix of Random Forest is:



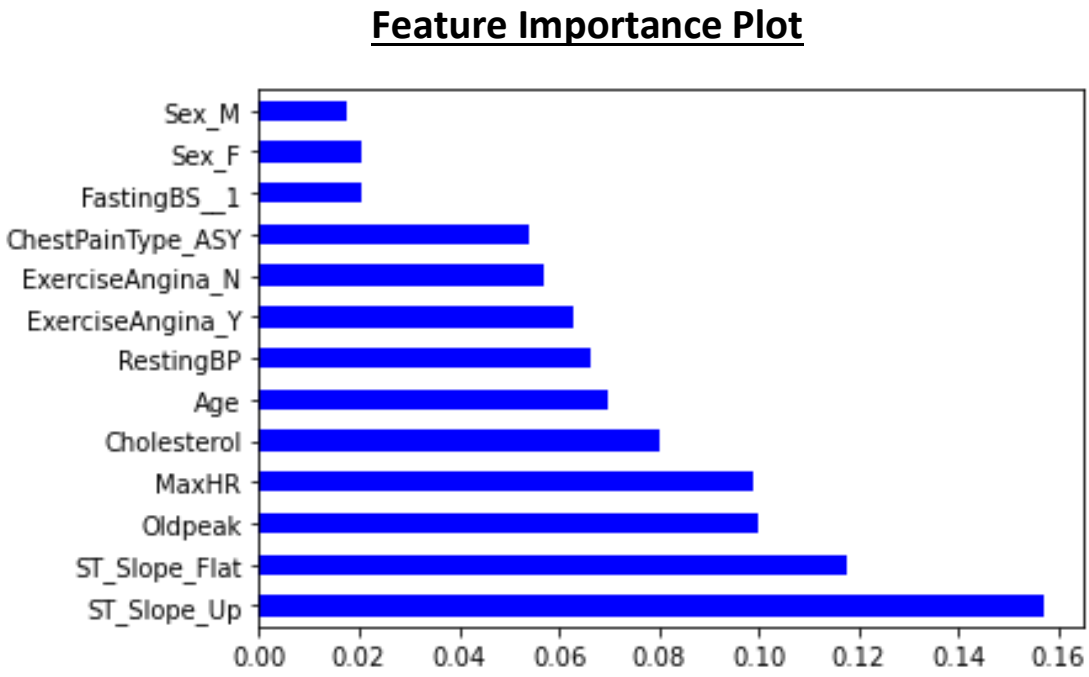
The Support Vector Machine model has an accuracy score of 85.86 percent. SVM has the advantage of being particularly efficient in high-dimensional spaces. The biggest disadvantage is that the SVM has a large number of parameters that must be selected appropriately in order to produce the best results. We simply utilized the default SVM parameters, and the accuracy of this model is similar to the Logistic Regression model but they differ in confusion matrix values.

4)Random Forest

The confusion matrix of Random Forest is:



The Random Forest model has an accuracy score of 87.68 percent. Random Forest does have the advantage of being able to deal with datasets with a lot of features, balance the variance, and not be affected by data noise. In comparison to the other models, Random Forest has the greatest accuracy score, as well as good precision, recall, and F1 scores.



Conclusions

As the Random Forest algorithm yields the highest accuracy of 87.68% and best precision, recall and F1 scores, we choose this as the best classification model for the heart disease dataset. The reason why Random Forest outperforms others is that it is not limited to the property of the dataset. Out of the 12 features we examined, the top 4 significant features that helped us classify between a positive & negative Diagnosis were ST_Slope, Old peak, MaxHR, Cholesterol. Therefore, our machine learning algorithm can now classify patients with heart disease and can properly diagnose patients and get them the help they need to recover. By diagnosing detecting these features early, we may prevent worse symptoms from arising later.