Heart Disease Classification Report

Introduction

Heart disease has risen to become one of the leading causes of death all over the world. Ac- cording to the World Health Organization, cardiac illnesses claim the lives of 17.7 million people each year, accounting for 31% of all fatalities worldwide. Heart disease has become the top cause of death in India as well. As a result, it is essential to be able to forecast heart-related disorders in a reliable and precise manner. Data on various health-related concerns is compiled by medical institutions all over the world. These data can be used to gain significant information utilizing a variety of machine learning techniques. However, the amount of data collected is enormous, and it is frequently noisy.

We analyze the various machine learning algorithms and find the best to predict the presence or absence of heart disease. The tar- get we will be exploring is binary classification which is 0 to show the absence of heart disease and 1 to show the presence of heart disease.

We are going to use various machine learning algorithms to predict the target. We will be using a number of different features about a person to predict whether they have heart dis- ease or not. The dependent variable is whether or not a patient has heart disease, while the independent variables are the patient's many medical characteristics. The various machine learning algorithms used for our model will be Logistic Regression, K-Nearest Neighbours, and Random Forest. We will compare the scores of all these models by splitting our data into training and testing in an approximate 80:20 ratio. We will also tune the hyper parameters for all these models to yield the best results. And finally conclude the best prediction model for our heart disease dataset.

Methodology Implementation

We have collected data from various reliable sources from the internet. After analyzing various factors, we have reached a conclusion that 13 independent variables will determine 1 target variable. To do this we will have to split the target variable from the rest. If we can reach 96% accuracy at predicting whether or not a patient has heart disease during the proof of concept, we'll pursue this project.

Training and Testing Dataset

The train and split procedure is used to divide the data into two halves.

1. Train split
2. Test split

The model designed will first train on the train split where it tries to learn the patterns in the data. Then based on the patterns it has learnt it will be tested on the test split. In this entire process choosing the test split size is also very important. A rule of thumb is to use 80% of your data to train on and the other 20% to test on.

Machine learning Models

Machine learning models are majorly classified as supervised and unsupervised. If the model is supervised, it is divided into two categories: regression and classification. We will focus on the following machine learning models:

1. ***Logistic Regression:*** It is a basic classification algorithm which predicts the probability of a target variable.

2. ***K-nearest Neighbors:*** It's a ma- chine learning algorithm that's supervised. The idea behind nearest neighbor methods is to find a predetermined number of training samples that are closest in distance to the new point and use them to predict the mark. It makes no assumptions about the data and is typically used for classification tasks where little to no prior knowledge of the data distribution is available. Finding the k closest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the identified data points to it is the aim of this algorithm.

3. ***Random Forest:*** Random forest is a supervised machine learning algorithm that can be used to solve problems in both classification and regression. It builds decision trees out of data samples, then gets predictions from each of them before voting on the best solution.

4. Naive 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.

5. 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.

6. Support Vector: 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.

we will find the other metrics for the logistic regression model:

A. ROC Curve

The metric compares the true positive rate with the false positive rate.

The True Positive Rate (TPR) is defined as follows:

$$TPR = \frac{TP}{TP+FN}$$

The False Positive Rate (FPR) is defined :

$$FPR = \frac{FP}{FP+TN}$$

It also provides us with AUC scores which denotes the area underneath the ROC curve

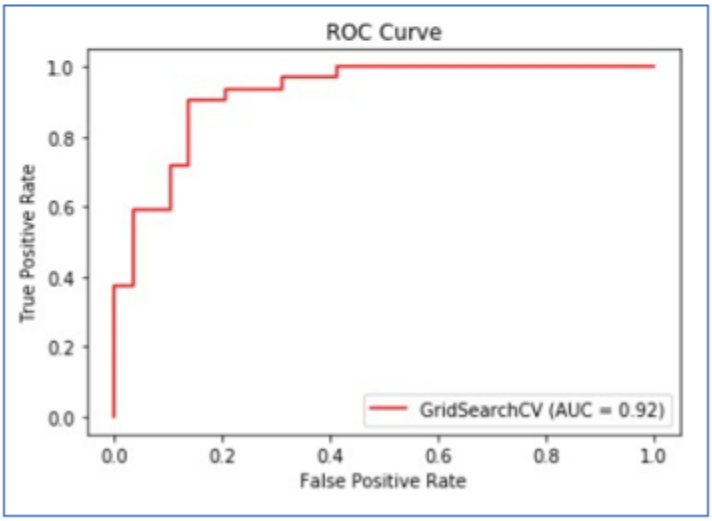

Fig 8-ROC Curve

B. Confusion Matrix

A confusion matrix is a table that is used to describe the output of a classification model/classifier by comparing the true values of the training and test datasets. It is divided into four parts, each of which is defined as follows:

1. True positives (TP): These are cases in which we expected yes (they have the disease) and they do.
2. Real negatives (TN): We predicted they wouldn't have the disorder, and they don't.
3. False positives (FP): We expected that they will have the disease, but they don't. (This is often referred to as a "Type I error.")
4. False negatives (FN): We expected that they will not have the disorder, but they do. (This is often referred to as a "Type II error.")

C. Classification Report

The Classification report is used to find the quality of predictions from a classification algorithm. It helps us to find how many predictions are correct and how many are wrong. More specifically, it gives us an understanding of True negatives and False Negatives, True Positives and False Positives, and uses them to predict the metrics of a classification. The main metrics found by the Classification report are accuracy, precision, recall, and f1- score.

D. Feature Importance

It refers to the techniques that assign a score to the input attributes/features with respect to the fact that which feature has the highest contribution in predicting the results for a given ma- chine learning model. For finding it we will use the coef_ attribute .The coef_ attribute is the coefficient of the features in the decision function. We can note that negative coef_ attribute denotes the presence of negative correlation.

Future Scope

In the future, the work could be improved by creating a web application premised on the logistic regression algorithm and by using a larger dataset than the one used in this study, which would help to provide better outcomes and aid health professionals in predicting heart disease efficiently and effectively.

Conclusion

With the rising number of deaths due to heart disease, it is becoming increasingly important to build a system that can effectively and accurately forecast heart disease. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of KNN, Logistic Regression and Random Forest for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Logistic regression algorithm is the most efficient algorithm with an accuracy score of 89% for prediction of heart disease. Accuracy of the algorithms in ma- chine learning depends upon the dataset that is used for training and testing purposes.