COMP-1801-M01-2021-22 Machine Learning

Sohel Arsad Salam - 001166291 - ss9992k

Abstract

This document[1] provides a basic idea of using a dataset to predict heart attack. In living creatures, the heart serves an important function. Diagnosis and forecast of cardiac illnesses need more precision, perfection, and accuracy because even a minor error can result in exhaustion or death. There are countless mortality instances connected to the heart, and the number is growing exponentially day by day.We need to have a mechanism in place to be able to recognise the signs of a heart stroke early and so avoid it, given the rapid growth in heart stroke rates among children and adolescents, seeking to uncover connections between the variables in our investigation.As a result, developing an app that can forecast the vulnerability of a cardiac illness based on fundamental symptoms such as age, sex, pulse rate, and so on. Employing typical Machine Learning algorithms to analyse the different properties in the data-set and then applying them effectively in predicting the likelihood of heart disease result reveals that, when compared to other machine learning approaches, Random Forest is the most effective provides the prediction more accuracy in less time

1. Introduction

As the heart is one of the most vast and crucial organs in the human body, it requires special attention. The majority of diseases are linked to the heart so it is necessary to predict heart diseases, which necessitates a comparative study in this field. Today, most patients die because their diseases are detected at an advanced stage due to instrument inaccuracy, so there is a need to learn about more efficient algorithms for disease prediction.There is no shortage of documents pertaining to the medical symptoms of people who have had a heart attack. However, their potential to help us predict similar outcomes in otherwise healthy persons is largely undetected.

Machine Learning is a very effective testing method that is based on training and testing. It is a branch of Artificial Intelligence (AI), which is a large field of learning in which machines mimic human skills. Machine learning is a subset of AI.The goal is to collect relevant data on all aspects of our field of research, train the data using the suggested machine learning method, and forecast how likely a patient is to develop heart disease. We recommend using the readily available sensors in watches and mobile phones to assess simple elements for the purpose of patients providing data.Around the world, heart disease is responsible for over 31 percent of all fatalities. Because of the paucity of diagnostic centres, competent doctors, and other resources that impact the precise prognosis of heart disease, early identification and treatment of various cardiac illnesses is extremely difficult, especially in poor countries.

2. Methods

The system's processing begins with data collecting, for which I utilised a data-set from kaggle.com that has been thoroughly vetted by a number of academics.

2.1. SYSTEM METHODOLOGY

A. Collection of Data:

The first stage in developing a prediction system is gathering data and settling on a training and testing dataset. In this research, we used 73 percent of the training data-set and 37 percent of the testing dataset to develop the system.

B. Selection of Attributes:

Attributes of datasets are properties of datasets that are utilised for systems, particularly for the heart, many attributes include the person's heart bit rate, gender, age, kind of chest discomfort, resting blood pressure, resting ECG, and so on.

C. Data Processing:

Prepossessing is required for the machine learning algorithms to provide renowned results. I used "getdummies" function to create different columns for different categories. I converted the binary categorical to numerical category using "one hot encoding".

D. Balancing the Data:

[1]This template and document is based on ICML 2021 LaTeX style file (https://media.icml.cc/Conferences/ICML2021/Styles/icml2021_style.zip)

Balancing the data is essential when you are expecting an accurate result so I used the "SMOTE" function from the "imblearn" library to balance the classes.

2.2. Models Used

A. Random Forest Classification Model:

Random Forest is a group of classification-based trees that haven't been trimmed. It performs admirably in a variety of real-world issues since it is unaffected by noise in the dataset and the danger of overfitting is minimal. It is quicker than many other tree-based algorithms and enhances accuracy for testing and validation data. The aggregate of the predictions of individual decision tree algorithms is known as random forests. When creating a random tree, there are several options for tuning the random forest's performance.

For Gini Index:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{1}$$

Determining from Entropy:

$$Entropy = \sum_{i=1}^{C} -p_i * log_2(p_i) \tag{2}$$

B. Logistic Regression Model:

The purpose of Logistic Regression is to find a relationship between a set of attributes and the probability of a particular event. The response variable contains two values: pass and fail, for example, when predicting whether a student passes or fails an exam based on the amount of hours spent studying. A Logistic Regression model is similar to a Linear Regression model, but instead of using a linear function, it uses a more complicated cost function known as the "Sigmoid function" or "logistic function."

$$log \frac{P(x)}{1 - P(x)} = \sum_{j=0}^{K} b_j x_j \tag{3}$$

C. Support Vector Machine (SVM):

Support Vector Machine [11] is a machine learning classification tool for analysing data and discovering patterns in classification and regression analysis. When data is classified as a two-class issue, SVM is usually considered. Data is described in this technique by determining the optimum hyper plane that isolates all data points from one class from the other. The greater the separation or edge between the two classes, the better the model is thought to be. Support vectors are data points that are located near the margin's edge. The mathematical approaches used to construct complicated real-world situations form the foundation of SVM.

3. Experiments

In the Experiments section, students are required to describe the dataset, experimental settings, evaluation criteria, results, and discussion.

3.1. Experimental settings

Basically, I build three models with my dataset i.e. Random Forest, Logistic Regression and Support Vector Machine. After building the models, I deduced that the Random Forest Classification model has the highest accuracy percentage. The evaluation matrices would be Precision, Recall and F1-Score to evaluate the accuracy of the models.

3.2. Evaluation criteria

A. Precision Matrix:

The accuracy is the number of TP multiplied by the number of TP 'Plus' the number of FP. False positives occur when a model is wrongly classified as positive when it is actually negative.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

B. Recall Matrix:

The number of genuine TP separated by the TP '+' FN can be used to determine the recall.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

C. F1-Score Matrix:

Precision and Recall are two factors that influence F1. When there is an unequal class distribution and you want to find a balance between Precision and Recall, you'll require the F1 Score.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

3.3. Results

After examining all of the models, the Random Forest model was determined to be the best match for the given situation. The logistic model had an accuracy of 86 percent, while the Random Forest had an accuracy of 88 percent. When compared to the other two models, the random forest is obviously the victor in terms of evaluation metrics.

Logistic Model Matrix Matrix

Confusion Matrix	0	1
Logistic Model	113	18
	17	106

Random Model Matrix

Confusion Matrix	0	1
Random Forest	111	20
	12	111

3.4. Discussion

Because random forest is an ensemble approach that incorporates many trees working on various characteristics for the same target variable, and the final classification is based on the majority of the comparable classifications, it may be a better performer here than logistic regression. Unlike the logistic regression model, which is a linear model in and of itself, the characteristics included in the model development process stay consistent throughout. Unlike random forest, where all features are included and the model can accommodate large changes in input, random forest does not employ all features.

4. Conclusion

Because the heart is such an important and critical organ in the human body, and the prediction of heart problems is also a major worry for people, algorithm accuracy is one of the parameters used to evaluate algorithm performance.The dataset used for training and testing purposes determines the accuracy of machine learning algorithms.In addition, based on the input supplied by the consumers, the algorithm generates a nearby reliable output. If the number of people who use the system grows, more people will be aware of their present heart condition, and the number of people dying from heart disease will decrease. As a result, after examining these possibilities, I chose this option, which is easier for me to deal with.For the Future Scope, more machine learning techniques will be used to better analyze the heart disease and to predict the disease earlier so that the risk of death can be reduced by being aware of the diseases.