Heart Disease Prediction Using ML model

Abstract

The main objective of this study is to build a predictive model to predict whether a person has heart disease or not based on feature values. The research dataset named “heart.csv” is taken from Kaggle which is loaded in the Jupyter Notebook in order to do the data analysis. After collecting the data, I have done data preprocessing and correlated each feature of the dataset. The statistical measure of the dataset is done and column distributions was plotted for the dataset and checked the duplicate values. I have also plotted the death and non-death occurrence event, heatmap, histogram, boxplot, and a scatterplot of the dataset. Encoding of categorical and continuous values of the dataset is done along with their graph representations. The feature scaling of the dataset is done as well. The data was splitted into training and test data and fed the training data in an ML model using different algorithms like Linear regression, Support Vector Machine, K nearest neighbor, Decision tree, Random Forest method, Gradient Boosting, Neural Networks, etc. and performance comparison is done and accuracy score is checked.

At last, confusion matrix is created and accuracy is calculated along with its graphical representation. Building a predictive model take different input values like ‘Age’,’ Sex’,’ Cholesterol present’, ‘thalassemia’, and other details or features of the patient data and successfully predict whether a person has heart disease or not.

Introduction

More than 12 million people worldwide pass away each year as a result of cardiovascular disease, often known as heart disease, according to the World Health Organization. The term Heart disease is a broad term that refers to a variety of illnesses that all directly impact a person's heart and arteries. Heart disease affects even young people, those with a lifespan of 20 to 30 years. Young people may have a higher risk of developing heart disease due to poor eating habits, lack of sleep, restlessness, depression, and a host of other risks, including obesity, a poor diet, family history, high blood pressure, high cholesterol, sedentary behavior, smoking, and hypertension.

The most challenging task in medicine is the diagnosis of heart problems, which is both crucial and challenging in and of itself. The doctor analyzes and comprehends the patients through manual check-ups at regular intervals of time, taking into account all the listed factors.

Data mining is the process of obtaining crucial decision-making data from a pool of historical records in order to analyze or anticipate the future. The details could be obscured and not without the aid of data mining, distinguishable. One data mining technique is classification, which uses the available historical data to predict future outcomes or outcomes. The integration of classification algorithms and the provision of automated training on the dataset makes it feasible to explore the hidden patterns in medical data sets, which are then utilized to predict the patient's future state. As a result, it is possible to provide information about a patient's past via medical data mining.

The supervised machine learning idea is used in this study work to make predictions. Analyzing different data mining classification techniques in comparison to make predictions, Nave Bayes, Decision Tree, and Random Forest are employed. The study is carried out at various degrees of and using various percentage split evaluation methodologies. When the heart disease dataset isutilized for training, the predictions are made using the classification model that is created fromthe classification methods. Any sort of heart illness can be predicted using this final model.

Lastly a predictive model is created to predict whether a person has heart disease or not based on the feature values.

Motivation

The primary reason for conducting this study is to propose a model for predicting the development of heart disease. Additionally, the goal of this research is to determine the optimum classification method for detecting cardiac disease in a patient. Different classification algorithms, including Naive Bayes, Decision Tree, KNN, Logistic Regression, and Random Forest, are applied at various levels of evaluation in a comparative study and analysis to support this work.

Although these machine learning methods are widely utilized, predicting cardiac disease is a crucial task requiring the highest level of accuracy. Consequently, a variety of levels and assessment strategy types are used to evaluate the algorithms. This will enable scientists and medical professionals to better understand and help them find out a better solution for predicting the heart diseases.

The provision of high-quality services at fair costs is a major issue facing healthcare organizations (hospitals, medical institutions). High-quality facilities point to patient diagnosis correctly and control medicines that work. Unsatisfactory outcomes can stem from poor clinical decisions, which is why they are unacceptable. Clinical test costs should be kept to a minimum by hospitals. They can achieve these results by leveraging appropriate PC-based data as well as selective emotionally supportive networks.

The most important organ in our body is the heart. Effective cardiac function is essential to life. Incorrect heart function will have an impact on the human body's other organs, including the kidney, the brain, and so forth.

Due to the wide availability of excellent information and the requirement to transform this readily available vast amount of information into useful knowledge data mining is used for information mining. Recently, KDD (learning disclosure in the database) and information mining have gained popularity. Since the size of the available information increases is too great to be physically analyzed, the popularity of information mining and KDD (information revelation in databases) shouldn't come as a surprise. Additionally, the techniques for programmed information investigation based on established insights and machine adaptation frequently pose problems when preparing large, dynamic information increases consisting of complex items.

Information mining has specific phases that we must get familiar with, including exploration, pattern discovery, and implementation.

Models and Algorithms

In order to predict heart disease based on an ML model , first different libraries were imported.

Data collection and preprocessing is done in the dataset. Missing values, redundancies, outliers, and inconsistencies compromise the integrity of the dataset. Data preprocessing is the concept of changing raw data into a clean data set. The dataset is preprocessed in order to check the missing values and other inconsistencies before executing it to the algorithm.

Feature selection is done on the dataset as well. A different analysis is made on the dataset to get information about the dataset which are very crucial.

Feature values such as 'age', 'sex','trestbps','chol','fbs' distribution is shown in the dataset.

Missing values are checked and removed from the dataset. Duplicate value checking is alsodone as part of the project.

Correlation matrix evaluation shows correlation between two variables. It is used to summarize the data. Key decision was made while creating the correlation matrix as a choice of correlationstatistic, coding of the variables etc, treatment of the missing values.

This matrix is used to get clear view of the relationship among the variables at different stage.

Identifying the pattern and trends become easier using it.

Statistical measure of the dataset is done. This is a descriptive analysis technique which is used here to summarize the characteristics of the dataset. This measures are classified as measure of central tendency and measure of spread.

Countplot is plotted using sns.countplot() for visualizing data. It is used to show the observational count in different category-based bins with the help of bars. Here the counterplot is plotted based target and it's value as '0' and '1'.

Heatmap plotting is also used in order to show the events within the dataset and also help viewers towards the area on which data visualization matters the most.

Histogram is plotted to provide visual representation of data distribution. It shows the large amount of data and the frequency of the data values. The median and the distribution of the data is determined by the histogram.

Scatterplot is done to show different data points that are measured in two variables. It helps to find out the cluster or pattern in the dataset. Here scatterplot is made for disease or non disease detected among patients based on maximum heart rate and their age.

Boxplots are created for different numerical features to give a good graphical image of the concentration of the data. It shows how far the extreme values are from most of the data. This boxplot is constructed from five values. They are : median, the third quartile, the minimum value, first quartile and maximum value.

Filtering of the dataset is done based on positive and negative heart disease patients. The positive patients ST depression(0.58) is less than negative heart disease patients ST depression (1.58) analysis. Encoding of categorical and continuous attributes are done and plotted.

Feature scaling is done to change the data model and give it a better version. It is used to normalize the features in the dataset into a finite range.

Dataset is then splitted into training and testing data set to make a clear analysis and test datais taken as 0.2. Percentage of people having a heart disease is more and in male which is recorded in the dataset.

Different classification algorithms are used and their accuracy scores are measured and compared in order to get a better view of which algorithm suits the model the best.

Different algorithms like logistic regression, KNN, Naïve bayes, random forest, decision tree, SVM etc. is used and found random forest to be the best algorithm for the dataset.

Random forest is supervised classification algorithm which is used for large datasets. It can build prediction models using regression trees which are unpruned to give strong predictions. Here by applying random forest algorithm the accuracy score has come upto 85.2% comparedto other algorithms being used.

Naïve bayes and SVM are two algorithms which gave the closest result to the Random forest algorithm which is 85.2% and 83.6% respectively.

Lastly confusion matrix is created and plotted the graph. The accuracy score of confusion matrix is 78.6%

Confusion matrix allows viewers to see at the glance the results of using a classifier or other algorithm. After applying classification to split the data into categories the model may not perform well. Confusion matrix measure the performance well.

A predictive model is built to predict the target as ‘0’ or ‘1’ to check whether heart is defective or not for a particular patient when input values are given.

Conclusion

This in depth analysis via different machine learning algorithm will not only help to get the best algorithm to be applied in the predictive model but also with the help of the suitable algorithmthis can predict the disease quiet well in future.