Project Report On Heart Disease Prediction

Content:

Introduction:

Need of heart disease prediction-

Predicting and diagnosing heart disease is the biggest challenge in the medical industry and it is based on factors like physical examination, symptoms and signs of the patient. Machine learning algorithms play a vital and accurate role in predicting heart disease.

Importance of heart disease prediction-

Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

Understanding and Define the problem:

Problem statement-

1. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.
2. The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease.
3. Attributes considered form the primary basis for tests and give accurate results more or less.
4. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease.

Objective-

1. Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.
2. Reduce the cost of medical tests.

Scope of the project-

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome.

Limitations-

1. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently.
2. The automation of the same would be highly beneficial.Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the

database.

Dataset Preparation and Perprocessing:

Data collection-

The analysis is carried out using a publicly available data for heart disease prediction . The dataset holds 4240 records with 15 features such as age, gender, current smoker, cigarettes per day , heart rate etc.

Data Visualization-

Two data visualization techniques can be used

1. Matplotlib - Matplotlib supports all the popular charts (lots, histograms, power spectra, bar charts, error charts, scatterplots, etc.) right out of the box.
2. Seaborn- unique feature of Seaborn is that it is designed in such a way that you write way lesser code to achieve high-grade visualizations.

Labelling-

In this stage,the data points are labelled. Target values of the data are known.

Data Preprocessing-

1. The purpose of preprocessing is to convert raw data data into a form that is useful in training and testing the ML model.
2. Missing values in the data are filled with median & mode.
3. The outliers in the data are removed or corrected.

4. Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values.

Dataset Splitting-

The given dataset is split into two parts: training and testing sets. The ratio od training and testing set is typically 80 to 20 percent.

Model Training:

In this stage, the training data is fed to the ML algorithm to bulid and train a model. The purpose of training is to develop a model.

Model Testing and Evaluation:

Confusion matrix-

The comparison a confusion matrices, it is the summary for the prediction of the result which we classified. Based on the classification of attributes the correct and incorrect predictions are marked with count values.A confusion matrix,which is represented in table format will explains about the performance of characterization model on the trained dataset.The most of the performance measures are calculated using this confusion matrix.

Accuracy-

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

It works well only if there are equal number of samples belonging to each class.

Model Development:

To initiate with the work we can use different types of techniques and algorithms. In this paper, machine learning techniques are used to increase the accuracy rate. In machine learning technique we can use the following algorithm

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest
4. KNN
5. SVM

Logistic Regression-

The logistic regression is also known as sigmoid function which helps in the easy representation in graphs. It also provides high accuracy. In this algorithm first the data should be imported and then trained. Logistic regression algorithm is represented in the graphs showing the difference between the attributes.From the training data we have to estimate the best and approximate coefficient and represent it.

Decision Tree-

Decision tree learning is one of the most widely used and practical machine learning method for inductive inference.Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent classification results and branches represent conjunctions of features that point to those classification results. The importance of decision trees lies in their ability to build interpretable models, which is a decisive factor for implementation.

Random Forest-

Random forest is an ensemble learning method for classification, it operates by constructing a multitude of decision trees at training time and outputting the class that have mean prediction.

KNN-

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.

SVM-

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training set (supervised learning), the algorithm outputs an optimal hyper planewhich categorizes new examples.SVM does the mapping from input space to feature spaceto support nonlinear classification problems. The kernel trick is helpful for doing this which makes a linear classification in the feature space equivalent to nonlinear classification in the original space. SVMs do these by mapping input vectors to a higher dimensional space where a maximal separating hyper planeis constructed.

Conclusion and Further Development:

The diagnosis of heart disease is difficult as a decision relied on grouping of large clinical and pathological data. The main idea behind this work is to study diverse prediction models for the heart disease and selecting important heart disease feature using genetic algorithm. In this work, different prediction models were studied and the experiments are conducted to find the best classifier for predicting the heart disease. Five classifiers Logistic Regression, Random Forest, KNN, Decision Tree, Support Vector Machine(SVM) were used for prediction of patients with heart diseases. Observation shows that in most of the cases Logistic Regression performance is having more accuracy compared with other three classification methods with respect to all the dataset .