Manoj Kumar Nallamala

Heart Failure Detection

**Introduction**

Healthcare organizations are faced with challenges to provide cost-effective and high-quality patient care. An effective heart failure prediction is developed using different machine learning algorithms for predicting the risk level of heart disease. This is critical to enhance the effectiveness of disease treatment and preventions. The system uses 12 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. Having either high LDL cholesterol (“bad” cholesterol) or low HDL cholesterol (“good” cholesterol) or both is one of the best predictors of your risk of heart disease. A blood lipid profile measures both your cholesterol numbers and your triglycerides, another type of fat in the blood that is a risk factor.

**Problem Definition**

Way to recognize patients heart failure rate by applying data mining and machine learning techniques on patient treatment history. Collecting Dataset from Kaggle or any other Open-Source platform. Data Pre-processing using techniques such as data cleaning, Encoding, Normalizing the columns to avoid bias. Applying algorithms and techniques and predicting the heart failure rate. We also perform feature extraction and find the important feature on which model is highly dependent on. We will compare the accuracy on each of the algorithms and conclude on the best algorithm after analysis.

**Predictive Model Selection**

Heat map depicting the correlation between attributes for the feature in the considered dataset.

As our goal is to predict heart failure which is a classification problem we have selected below models

Logistic Regression

Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. This model uses sigmoid function as our criteria to map the obtained real values to map it to 0 or 1. Accuracy we got here is 84.17%.

Linear Regression

Though regressors are not relevant to the problem we are solving we have applied multiple linear regressor for predicting the values for heart failure rate between 0 and 1. Linear regression models are used to identify the relationship between a continuous dependent variable and one or more independent variables. Here we have kept a threshold value to separate the two classes once the continuous values for dependent variable are ready which is 0.5 in our case. Accuracy we got here is 89%. It is quite surprising as it out performed logistic regression.

K-Nearest Neighbor

K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most like the available categories. For Optimal K value we have computed and plotted Error rate vs K and picked the optimal value which is giving minimal error rate. We have selected K as ‘2’ and built the K-NN classifier which gave us the accuracy of 80.83%

Support Vector Machine

SVM is a supervised machine learning algorithm that can be used for both classification and regression challenges. we plot each data item as a point in n-dimensional space (where n is several features). Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. We have used linear kernel which generally transforms the training set of data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. Accuracy we got here is 80%.

Decision Tree

A decision tree is a popular classification structure used in machine learning. It is a map of the possible outcomes of a series of related choices. A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a treelike shape. We have selected entropy as the criterion which is a measure of chaos within the node. And chaos, in the context of decision trees, is having a node where all classes are equally present in the data.

Random Forest

A random forest is a machine learning technique that’s used to solve regression and classification problems. Multiple decision trees can be constructed on the same dataset. Random forest aggregates all possible trees than can be constructed and decides the outcome based on the majority vote of the individual trees. Picking the best tree from the constructed trees will lead us to a even better model for predicting heart failure. We have constructed 10 different decision trees, which can be passed as value for argument “n - estimators”.

**Analysis**

**** From various models that we used, Random Forest algorithm gave the best accuracy (95.8%) for the predictions. Decision tree models have second best performance among the remaining models with an accuracy of 93.3%. So as per our analysis as the features are large in number decision trees seems to be best bet. Time and ejection fraction have best splits among all the features in the constructed decision trees which indicates that those features play a prominent role in predicting heart failure. We have also constructed decision tree taking time and ejection factor and achieved a healthy split with even distribution which suggest a better heart.

**Conclusion**

As observed from the results of various models that we used, Random Forest algorithm gave the best accuracy (95.8%) for the predictions. Decision tree models have second best performance among the remaining models with an accuracy of 93.3%. It is obvious as to why Random Forest performed better than Decision Tree model because it builds on decision trees and overcomes the shortcomings of individual trees. Random Forest also performs better on large datasets compared to many other algorithms in machine learning.

**Recommendations**

A much effective heart disease prediction system can be developed using neural network for predicting the risk level of heart disease. Using more than 12 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction and more data sets can help provide more accuracy. The following are the recommendations can be used for deep learning models to improve accuracy, add More Layers, Transfer learning, Increase Epochs. Also, we could also try collecting more data samples ranging different regions and age groups. For larger datasets neural network can be a viable solution for this heart disease classification.