DV2578-Assignment 3

Patamsetti Naga Vishnu Chakravarthi

napa18@student.bth.se

Prediction of Heart Failure Using Machine Learning Classification.

Abstract:

The use of machine learning approaches in the diagnosis of diseases has been steadily growing. In this study, a machine learning system like this was used to diagnose heart disease, which is a quite common and severe disease. To design the best performing machine learning model (classifier) to classify whether the person affected by Cardiovascular disease or not is the main aim of the project.

Keywords—Machine learning model; medicaldata; heart failure diagnoses

1. INTRODUCTION

Disorders of heart and blood vessels is defined as Cardiovascular disease, which includes several other symptoms like hypertension, heart attack, cerebrovascular disease(stroke). Heart failure is a chronic condition which means once's if a person is attacked with heart failure there is 50-50% of chance to live or die. If the patient has resisted the heart failure, he/she should always be under the supervision of a medical supervisor. Heart failure in simple words means that the heart muscle which is present in the body is unable to send enough blood to the body to meet the need of the body like oxygen and blood. Narrowing or blockage of coronary arteries is also a common cause for heart failure. This disease is caused by the buildup of fatty deposits in arteries which in return will decrease the blood flow which ultimately causes heart attack/ heart failure.

Approximately there are 26 million people in and around the world are affect with heart disease. In the present generation, young people are affected from heart failure/ heart attack due to development of obesity and diabetes in earlier age is the reason why youngsters are affected with heart diseases. If a person is once's attacked with cardiac arrest it will be stressing throughout the life by which the person can again get heart attack or a heart stroke. In this project we can predict the heart failure using machine learning classifiers.

2. RELATED WORK

We can save lives of the heart failure patient's by introducing a way to improve the diagnosis of the patient on the bases of their pervious medical history [6]. As most of the time, physical work, additional finical charges are overburden to the patient when he/ she is getting tests related to heart. As discussed inthe introduction the reasons behind heart failure are unhealthy food habit's, excessive sugar, overweight (fatty deposits). Where the common symptoms of heart failure are chest, arms contraction and breathlessness.

There were some studies in which machine learning models were used to automatethe process of predicting the heart failure using the patient's information like food habits,tobacco, age and so on. In the study [7] the authors had used multiple machine learning models to detect the presence of detect the heartdisease using the data collected from different medical databases. By studying some more research papers in which algorithms are used for prediction of heart failure where the results are trained and evaluated based on the comparison of algorithms. Logistic regression, SVM, Random forest algorithms were compared with K nearest neighbors' classification. KNN algorithm can be used for classification and regression problem. KNN algorithm uses 'feature similarity' to predict thevalues of the data set given which in will give the result. Finally, KNN is concluded as thebest because it gives the best efficiency.

The effectiveness of the KNN algorithm is a combination of six similar heterogenous data sets and the output is calculated with Euclidean distance and four combination of similar measures are effective in handling binary and numerical features all together. There are four different distance function which can be to calculate the effectiveness of KNN classifier, which are Euclidean distance, cosine similarity measure, Minkowski correlation and chi square.

3. DATA SET

The dataset used for this project is taken from Kaggle [1]. This dataset has 13 attributes along with the class attribute which tells whether the patient has heart disease or not. The different features present in this data set are Age, decrease of red blood cells or hemoglobin, Level of the CPK enzyme in the blood, If the patient has diabetes, Percentage of blood leaving the heart at each contraction, If the patient has hypertension,

Platelets in the blood, Level of serum creatinine in the blood Level of serum sodium in the blood, Woman or man, If the patient smokes or not, Follow-up period and DEATH_EVENT.This dataset has a total of 299 instances.

4. IMPORTING LIBRARIES

Import the libraries which are required to evaluate the heart failure data. Libraries such as train test split, pandas, logistic regression, gradient booster, random forest, svm, k nearest neighbours, andaccuracy score are imported

5. ALGORITHM SELECTION

Here the data has two value so we can say that it the problem is binary classification problem. So supervised classification algorithms are suitable for this problem. Logistic, SVM, Random Forest and K-Nearest- Neighbors are the best performing algorithms when it comes to classifying two classes either yes or no. The performance of an algorithm depends on the size of the data and the number of independent features it consists of. So, each algorithm has its pros and cons.

Logistic regression (LR): It is one of the best classification algorithms and outperforms the remaining algorithms when it comes to binary classification. The reason for selecting the Logistic regression is it performs better on linearly separable data. Moreover, it also has a verry low training and inference time. It is also easy to implement. It uses regression analysis to learn and predict the parameters in the data set. The processes of learning and prediction are focused on calculating the likelihood of binary classification.

SVM: In this study, the other machine learning algorithm that I used is support vector machine (SVM). Similarly, to liner regression, SVM also works well on linearly separable data. It performs well on high dimensional data [1]. This works by categorizing the objects according to the predefined classes in the dataset provided. The main reason for selecting the SVM algorithm is its better performance while dealing with high dimensional and linearly separable data.

Random Forest (RF): The next model selected for this project is Random Forest. This model also belongs to the family of classifications algorithms. In random forest is an ensemble of decision trees which is constructed using the training data, given the input instance each tree in the ensemble will give prediction for the target label. The final output of the Random forest classifier is the output label which is given by majority of the trees in the ensemble of trees which are trained on the training data. The main reason for selecting this algorithm is its efficiency while dealing with overfitting and missing values in the training data.

K-Nearest neighbor (KNN): It is a classifier that is simple, lazy, and nonparametric. When all the features are continuous, KNN is preferred. KNN is also known as case-based reasoning and has been used in many applications, such as pattern recognition, statistical estimation and so on [2]. To determine the class of an unknown sample, classification is obtained by identifying the nearest neighbor conclude the best one based on metrics such as accuracy for each algorithm. The main reason for selecting the KNN is its verry low training time.

6. METHOD

Loading the data:
First import the data, then load the data into pandas data frame using the function pandas.read_csv().

Correlating test:
Generally, while training a classification model it is better to use the attributes which are highly corelated to the class attribute and often it is better to omit the attributes which have verry less correlation with the class attribute, this will reduce the training time and also improve the classification performance. The figure 4.1 is the correlation heat map which shows how the attributes are corelated to each other.
Only the attributes which are corelated to theclass attribute are selected for training the models. The attributes which have the correlation less than or equal to 0.1 are omitted and only the attributes which have the correlation greater than 0.1 are selected for training the models. From the correlation heat map, itis evident that the attributes ‘age’,‘ejection fraction’, serum_vretinine’, ‘serum_sodium’, ‘time’ are the attributes which have the correlation greater than 0.1with the class attribute.

Training and Testing the Models:

Now we will divide the data into two parts for training (80%) and testing (20%). Now we will train all the four classification algorithms namely LR, SVM, KN N and RF on the training data. Now we will test these models on the test set during which the metrics accuracy, precision, recall and f1-score are extracted while testing the models.

7. RESULTS

All the models are trained on the training set and the trained models are tested on test set during which the fallowing metrices accuracy, and f1-score are extracted.

	LR	RF	SVM	KNN
Accuracy	88.33	85.00	90.67	93.33
F1-score	75.86	68.97	75.00	83.33

Table 4.1. Accuracy and F1-score table

8. CONCLUSION

After analyzing the results form the table we can conclude that K Nearest Neighbors Classifier (93.33) has the highest accuracy fallowed by SVM(90.67), LR(88.33) and then RF(85.00)

When it comes to f1-score KNN(83.33) has the highest f1-score fallowed by LR(75.86) and SVM(75.00) which have similar f1-scores and then fallowed by RF(68.97) which has the lest f1-score.

9. REFERENCES

1. "K. Polat, S. Şahan and S. Güneş, " Automatic detectionof heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k -nn (nearest neighbour) based weighting preprocessing ", Expert Syst. Appl., vol. 32, pp. 625-631, 2007."

2. "H. Yang and J. M. Garibaldi, 'A hybrid model for automatic identification of risk factors for heart disease',J. Biomed. Inform., vol. 58, pp. S171-S182, Dec. 2015."

3. "Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei and A. A. Yarifard, 'Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm', Comput. Methods ProgramsBiomed., vol. 141, pp. 19-26, Apr. 2017."

4. E. O. Olaniyi, O. K. Oyedotun and K. Adnan, 'Heart diseases diagnosis using neural networks arbitration', Int. J. Intell. Syst. Appl., vol. 7, no. 12, pp. 72, 2015."

5. "S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K.Bashir, ―Improving Heart Disease Prediction Using Feature Selection Approaches,‖ in 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2019, pp. 619–623."

6. "Implementation of Machine Learning Model to PredictHeart Failure Disease."

7. "Prediction of heart disease using k-nearest neighbor andparticle swarm optimization." .

8. "Heart Failure Prediction." https://kaggle.com/andrewmvd/heart-failure-clinical-data (accessed Mar. 21, 2021).

10. APPENDEX

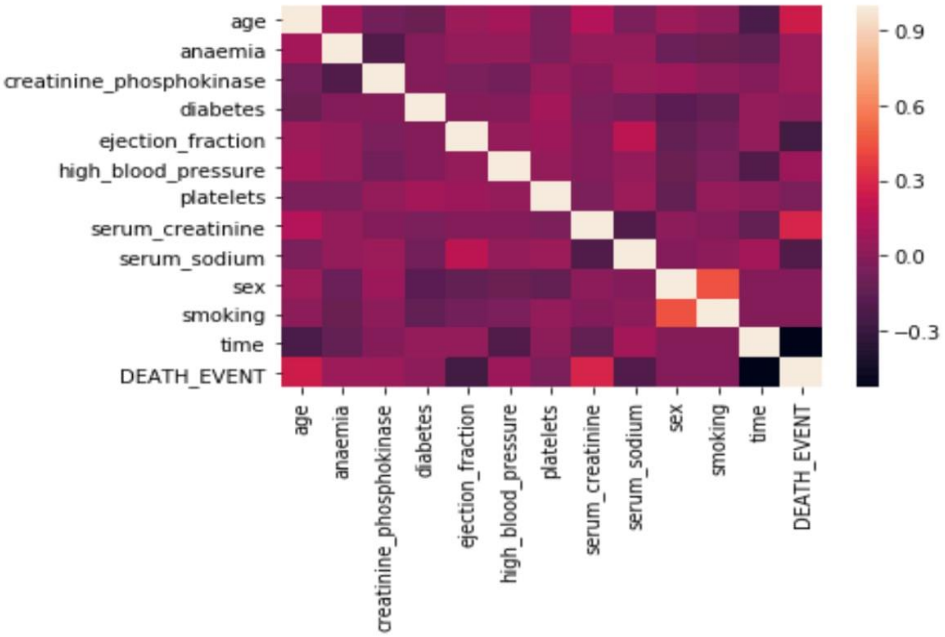

Fig 4.1. Correlation heat map.