Heart Diseases Prediction

ABUSUFIYAN ATHANI
PES1UG19CS024
SEC:A

NAVEEN KUMAR
PES1UG19CS292
SEC:E

SAMMED D TAPAKIRE
PES1UG19CS426
SEC:G

ROHIT METRE
PES1UG19CS392
SEC:F

INTRODUCTION

Heart is the main organ of human and is the major cause of mortality in today's world. It is the most common disease in all age groups and in males and females. The diagnosis of these cardiovascular diseases requires more accuracy and precision as a little mistake can cause fatigue problem or death. There are a lot of advancement in machine learning, artificial intelligence and data science, which helps in decision making and predictions on large amount of data from the health industries. The ML approaches have helped to find the trends and patterns of human body using the algorithms and data analytic models. There are many features like age, gender, blood pressure, fasting, cholesterol, exercise etc which leads to heart diseases. In this paper we find the accuracy of the machine learning algorithms which are used to predict the heart diseases by using logistic regression and support vector machine (SVM) by using Kaggle dataset for

training and testing. Since Python programming Anaconda (jupytor) notebook is best tool and has many libraries, header file which makes the work precise.

Previous work

There are numerous works has been done related to heart disease prediction systems using different data mining techniques and machine learning algorithms such as Logistic regression, Random forest classifier etc. There are some limitations to prediction using these techniques such as:

1. Prediction of cardiovascular disease results is not accurate and not fully reliable.

2. Data mining techniques does not help to provide effective decision making.

3. Various factors must be considered (with some factors having more importance than others) for building the model.

Numerous approaches have been made to overcome these limitations such as:

1. Using Ensemble models - combining multiple models to improve prediction accuracy.

2. Data pre-processing - Normalisation, Standardization.

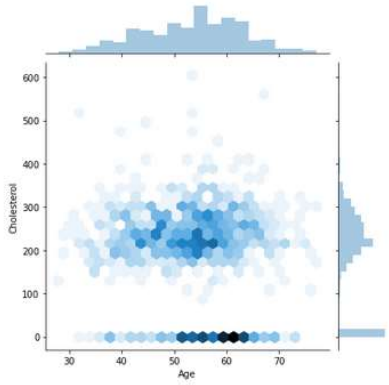
3. Assigning weights – weights are assigned to attributes according to their importance and effect on prediction.

To predict the heart disease we have built two models using SVM of machine learning and logistic regression after pre-processing the data.

Proposed solution:

Pre-processing of the data : there are various methods in pre-processing of data to analyse the data clearly. Finding the null values of each column and replacing them if any. In the dataset there are more zero values in cholesterol column (cholesterol value cannot be zero) , which are observed in the dataset by **jointplot** method from seaborn module between cholesterol and age . Replacing zero values of a cholesterol column with mean of the cholesterol values of the particular gender and age as cholesterol is different for male and female based on gender.

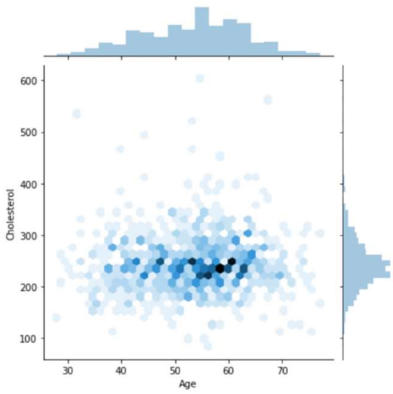
Before processing



	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Building a model for a dataset and training the model to get best accuracy of the predications will depend on how the build model is built and train it well.

After processing



Converting categorical data of every column to ordinal and binary data as SVM uses numerical data.

As SUPORT VECTOR MACHINE is a simple supervised machine learning algorithm used for classification. In SVM a hyper plane is created as the boundary between types of data. For 2 dimensional this hyper plane is a line.

Modelling using SVM of machine learning, building a model using LINEAR SVC function , train the model by 80% of the source data. The accuracy of predicting training data is 0.75 where as error is 0.18 (by using mean square error). For the test data (20% of source data) the prediction accuracy, error was 0.49 and 0.51 respectively . As error rate is too high for SVM model ,a new model has to be build here it is logistic regression .

Logistic regression is also a supervised classification algorithm, this model builds regression model for predicting the probability of the given data entry belongs to category numbered as 1. Logistic regression models using sigmoid function ,it's a classification technique only when decision threshold is brought into picture.

By training the logistic regression model by 80% of source dataset , the accuracy is 0.87 and error is 0.12 for training dataset , where as for testing dataset it is 0.76 for accuracy and 0.24 for error.

svm model. For logistic regression model the prediction accuracy is 0.76.

From the above insights we get to know that SVM has less accuracy and more error-rate as it fails for most of the cases, where as other model has more accuracy and less error-rate than SVM .

Reasons for the SVM to fail:

- May not suitable for large dataset.
- When data has more noise i.e. target classes are overlapping .
- It may fail when the number of features for each datapoint are more than the number of training data samples.

Experimental results:

Here we have built two models , the first one is support vector machine and logistic regression where both the models are supervised algorithm based models of machine learning which are used for classification and/or regression.

First model we built is SVM ,the data has to be split into 80% and 20% of source data for training and testing respectively. The model is trained with 80% data , accuracy for predicting training data is 0.75 where as for the model that is logistical regression the accuracy is 0.87 .

For testing data set that is other 20% of source data ,the prediction has the accuracy of 0.48 which is too low for

So we choose logistic regression as our dataset is linearly separable , logistic regression performs well on linearly separable data.

Our data has no multicollinearity between independent variables. Logistic regression is suitable when the variable being predicted for is a probability on a binary range from 0 to 1.

Limitations of our model :

- It may fail when number of observations is lesser than the number of features of data, it may lead to overfitting.

- When data set has multicollinearity between independent variables.
- Logistic regression doesn't perform well when the dataset has complex relationship.

Conclusion:

A heart disease prediction model can be used to predict the risk or likelihood of a heart disease in a patient in advance so that precautions can be taken and any further complications can be avoided. Such technologies help in advance prediction which can result in reduction of serious complications, treatment expenses and help in better planning of diagnosis.

Two models have been built using Support Vector Machine (SVM) and Logistic Regression methods. These models aim on predicting whether a patient is at risk of a heart disease by analysing the patients medical/health parameters such as resting blood pressure, cholesterol level and many more other factors. These models consider a total of 12 parameters to determine if a patient is at risk of a heart disease and provides a binary output(0 for no and 1 for yes).

Even though these models use complicated machine learning algorithms and huge amount of data from previous patients, they cannot be fully trusted as they make wrong predictions, also the provided data required to predict may not be sufficient and may require even more information for accuracy. It also varies with person to person , place to place based on lifestyles and habits.

Contributions:

Abusufiyan Athani : 27%

Naveen Kumar : 27%

Sammed D Tapakire : 27%

Rohit Metre : 20%