CSA Open Day Data Science Hackathon 2020

Team : Optimizers

Team Members: Akash Kumbhar (ed18smail.iitm.ac.in)

Siya Das (siyadas1509@gmail.com)

PROBLEM: Heart Attack Prediction

Problem Description.:

The prevalence of heart disease and stroke has increased by over 50% from 1990 to 2016 in India, with an increase observed in every state. The contribution of these diseases to total deaths and disease burden in the country has almost doubled in the past 25 years. Heart disease now is the leading individual cause of disease burden in India, and stroke is the fifth leading cause. Effective and tailored medical treatment can be developed using technologies that use Big Data to predict and manage heart attack. The medical treatment can be tailored according to the needs of the individual and the results will guide providers, healthcare organizations, nurses, and other treatment providers in estimating, analyzing and thus preventing any future mishaps.

The major reason for heart attack includes blockage of arteries due to formation of plaque. This restricts the blood flow to the heart. The other reasons include age, sex, smoking, family history, cholesterol, poor diet, high blood pressure, obesity, physical inactivity, and alcohol intake. Apart from this eating habits, physical inactivity, and obesity are also considered to be major risk factors.

Proposed Solution :

One third of the deaths in the world are due to heart diseases. 24% of deaths in India are due to heart ailments. Hence predicting the heart attack for a particular patient a priori will help us to take necessary precautions to avoid the heart attack in future. Machine learning involves artificial intelligence, and is used in solving many problems in data science. One common application of machine learning is the prediction of an outcome based upon existing data. Hence we will be using machine learning technique to build a model which will predict whether a patient is prone to heart attack or not. The most important element in building the model is collecting the dataset. Dataset must contain most of the reasons for the occurences of heart attack. These reasons we will be used as input features for the model. The output will be the value of ‘0’ or ‘1’ depending upon the percentage of heart blockage due to plaque. Generally, the threshold will be 50%. Hence it is a classification task.

Data collection will be done from hospitals and clinics. After data collection, we will pre-process the data which will replace the unknown information of a particular patient with suitable values based on statistical measure like mean, called as imputation process. Once our dataset is imputed, we will divide it into train set and test set. Train dataset will be used to train the model by most of the machine learning algorithms like SVM, Logistic Regression, Naïve Bayes Classifier, Linear Discriminant Analysis and Neural Network etc. The model with maximum accuracy or f1 score will be selected as final model. The brief overview of the solution process is given in figure 1. Hence we build a ML model which will predict the heart attack a prior and thus we save the life of many!

Flowchart of Solution Process :

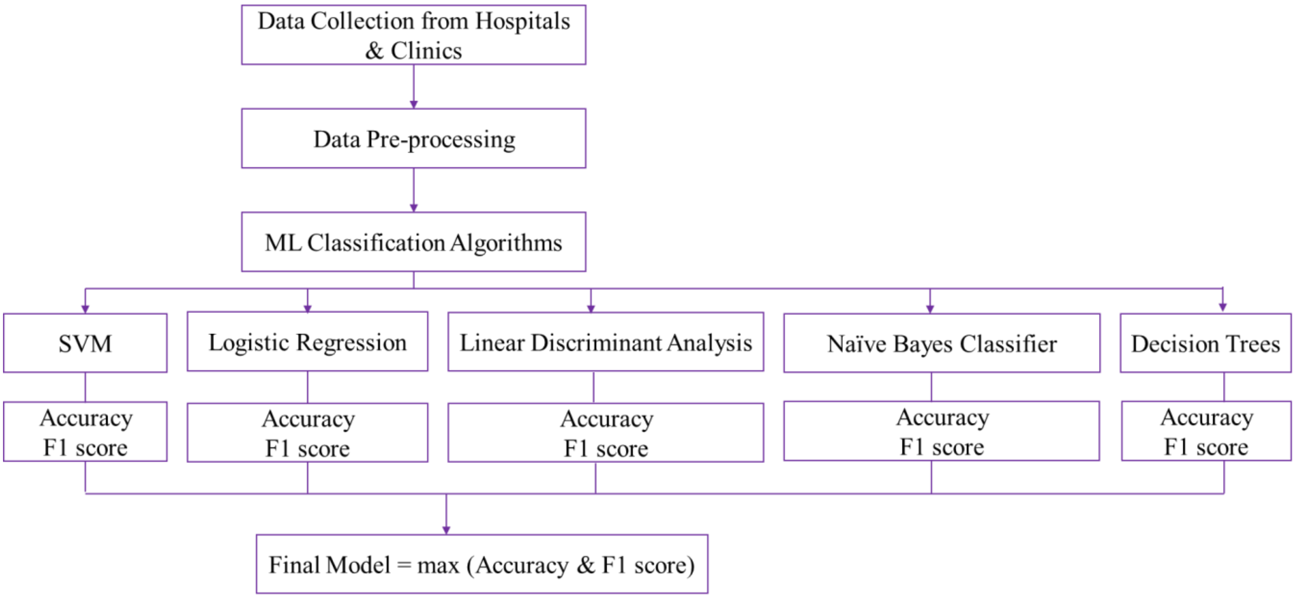

Figure 1: Block Diagram of Solution

Dataset Description:

As a prototype, dataset from Cleveland Heart Disease Database of UCI Machine Learning Repository is used for this study. The dataset includes 14 attributes for 294 patients. The 14th attribute represents the risk of heart attack to be high, if percentage of blockage is greater than 50%, denoted by value 1 and if it is less than 50%, it is denoted by value 0. It is named as '**num**' in the dataset.

The first 13 attributes are described as follows:

SR NO.	Variable Name	Description
1	Age (age)	numerical
2	Sex. (sex) (0=Female,1=Male)	categorical
3	Chest Pain Type (cp)	categorical
4	Resting blood pressure (trestbps)	numerical
5	Cholesterol (chol)	numerical
6	Fasting blood sugar (fbs)	ordinal
7	Resting electrocardiographic results (restecg)	numerical
8	Maximum heart rate achieved (thalach)	numerical
9	Exercise induced angina (exang)	ordinal
10	ST depression induced by exercise relative to rest (old peak)	numerical
11	The slope of the peak exercise ST segment (slope)	categorical
12	Number of major vessels (0-3) coloured by fluoroscopy (ca)	numerical
13	Thal	categorical

Note : Some of the categorical variables has category of '1', '2','3','4' and some has category of '0' and '1' only . Hence categorical variables are converted into dummy variables to have uniformity in their values throughout the dataset.

Data Analyzing:

In classification task, the objective of data analyzing is to look for the distribution of classes in the dataset, the number of missing values in a particular feature and variation of output with respect to each dataset. In this dataset, number of patients having less risk for heart attack is denoted by class ‘0’ are more in number than the number of patients with more risk denoted by class ‘1’. The number of patient with class 0 are around 194 and number of patient with class 1 are around 100. This can be illustrated from figure 1. Hence the dataset is skewed towards class 0.

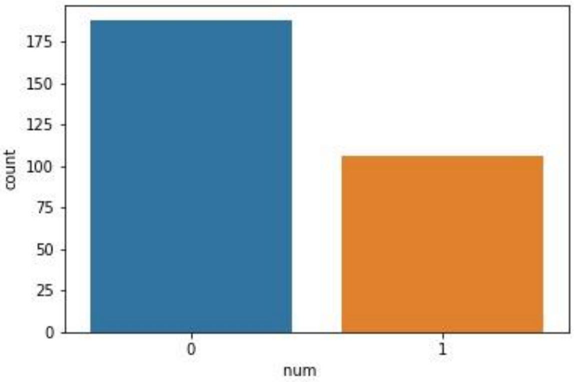

Figure 1. Distribution of output classes

From figure 2, it is observed that the age variable is normally distributed, thus people between age 40 and 60 are more in number. Among the people from class 1, people with the age between 40 and 60 are having high risk of heart attack than the people below 40 and above 60, denoted by orange colour bar in the figure.

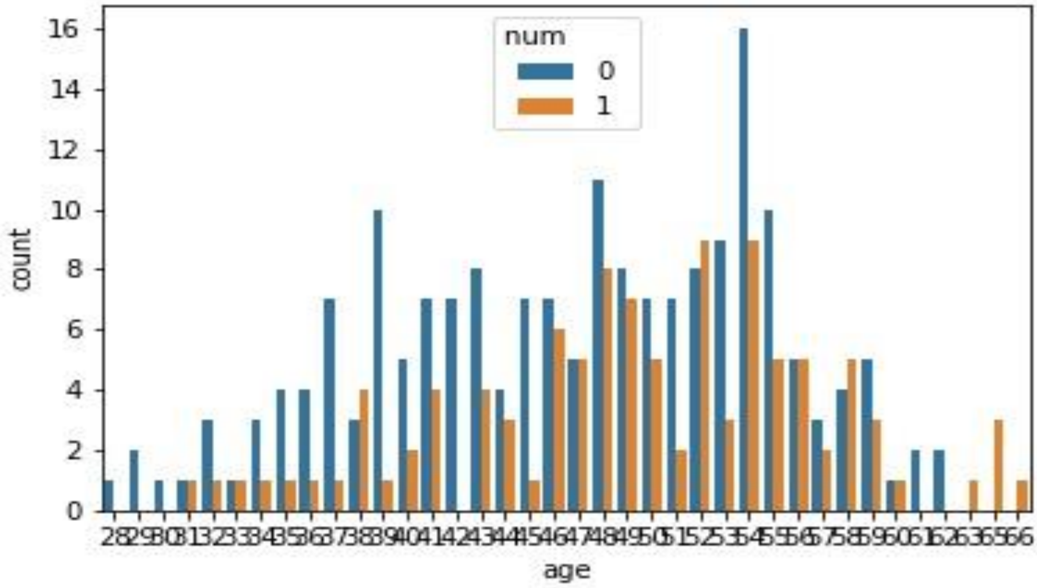

Figure 2: Distribution of variable ‘age’ with respect to output classes

From figure 3, it is observed that among the people having high risk of heart attack denoted by class 1, males are more in number than females and the same observation goes to class 0. Hence it is inferred that males are more prone to risk of heart attack than females.

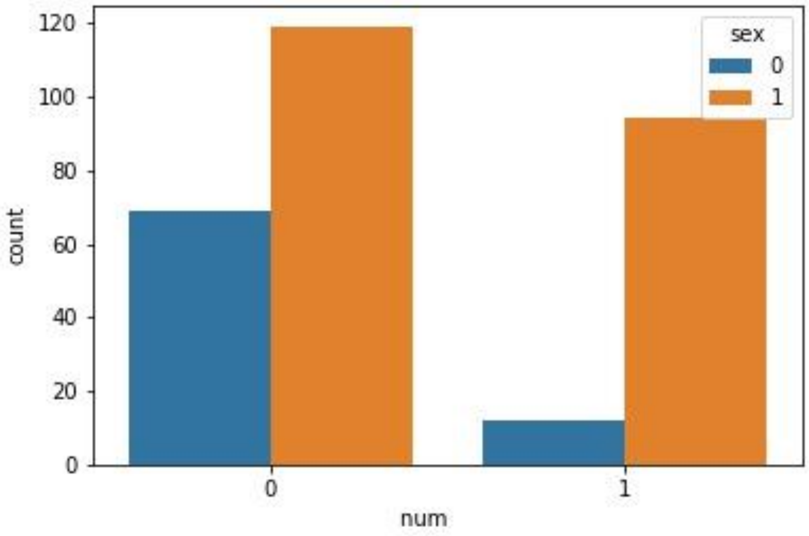

Figure 3. : Distribution of variable ‘sex’ with respect to output classes.

Data Cleaning

It is found that the variables ‘slope’, ‘ca’ and ‘thal’ have a large number of unknown information and demonstrated by plotting a heatmap of all the variables as shown in figure 4. The white colour represents unknown information in the data and the black colour represents the known values. Hence it is safe to neglect these variables and consider the remaining 10 variables for building the model.

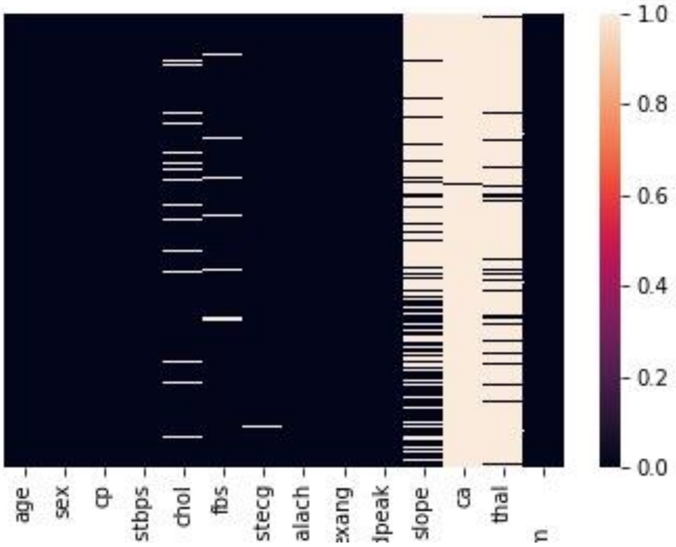

Figure 4.: Heat map of all the variables in the data.

For the remaining variables, the unknown values are computed by taking the mean of the known values. This is called data imputation. The categorical variables are transformed into dummy variables to have uniformity in the values of categorical variables. Heat map after data imputation is shown in figure 5., which demonstrates no 'na' or 'NaN' values the data. Hence our data is cleaned and ready for training the model.

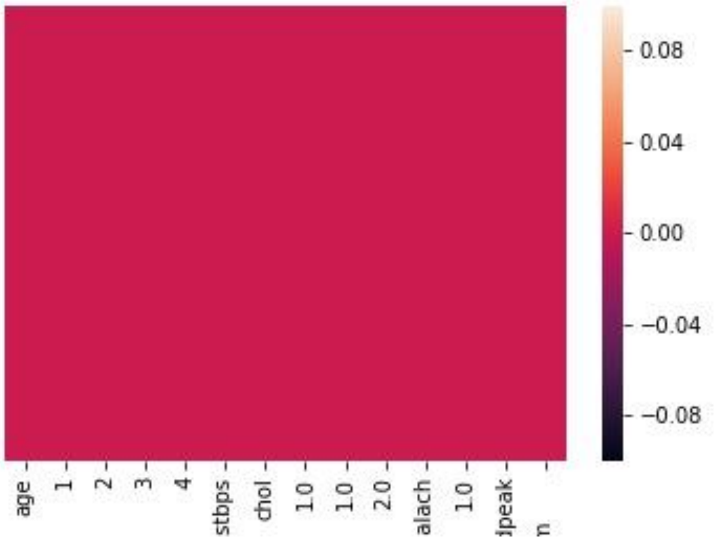

Figure 5. : Heat map after data imputation

Building the model :

For training the model we divided the dataset in ratio 0.8:0.2. 80% of the data will be used for training and 20% of the data will be used for testing the model. Machine Learning algorithms used for training the model is stated in table 1 with their test accuracy and F1 score after training the model. Test Accuracy and F1 score are the metrics of performance of the model . Test accuracy provides information of percentage for correctly predicted outputs in the dataset. F1 score gives us the information about balance between precision and recall. Hence **high F1 score and high test accuracy will decide which model to choose.**

Logistic regression model has highest test accuracy and F1 score hence it is choosen to be the final model for prediction.

SR NO.	Model	Test Accuracy	F1 Score (0)
1	**Logistic Regression**	**0.78**	**0.83**
2	KNN Classifier	0.71	0.77
3	Support Vector Machine (SVM)	0.69	0.74
4	Linear Discriminant Analysis (LDA)	0.71	0.77
5	Naïve Bayes Classifier	0.68	0.72
6	Neural Networks	0.69	0.76
7	Decision Trees	0.75	0.82
8	Bagging Ensemble of Decision Tree	0.75	0.81
9	Boosting Ensemble of Decision Tree	0.71	0.77
10	Random Forest	0.76	0.81

Table 1. Machine Learning Models with performance metrics

Performance Metrics of Logistic Regression Model :

Confusion Matrix

	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Confusion Matrix represents the number of correct classification and misclassification of the model. The rows represent the ground truth and columns represents the predicted output. In our case study, we wish to have true positive to be high because it is predicting that a patient has disease and it is predicting it correctly. While False Negative should be very low as it predicts that a patient doesn't have a heart attack but in reality he does have. He is it has to be as low as possible.

In our case study, Confusion Matrix is represented as follows :

	Positive	Negative
Positive	31	6
Negative	7	15

Precision : It gives us the information about exactness of the classifier. In our case, it comes out to be **0.82**

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

Recall : It tells us about the sensitivity of the classifier. In our case, it comes out to be **0.84.** Its formula is given as follows :

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

F1 score : It is a balance between precision and recall. In our case, it is **0.83**. Its formula is given as follows :

$$\text{F1 Score} = \frac{\text{2*Precision*Recall}}{\text{Precision+Recall}}$$

Training Accuracy: It is accuracy of the model when tested on train data. Our model has training accuracy of about **0.8554**

Testing Accuracy: It is accuracy of the model when tested on test data. Our model has test accuracy of about **0.7791.**

Hence we have a prototype of Machine Learning Model which works on logistic regression algorithm that will be used to predict the heart attack and be useful to take necessary precautions to avoid its occurences in the future.