# **6006CEM**

# MAchine learning and related applications

# coursework

# Github: https://github.coventry.ac.uk/guscan/6006CEM\_2021s1\_8367996\_NG

# DEADLINE: 04/12/2020

# by Nikita Gusca

# Background

In the global context, heart disease is one of the main and common disease among the various individuals. However, based on the modern lifestyle, significance number of attributes are used to explain the nature of the heart disease. Heart is one of the main components of anyone’s life and therefore, it’s a noteworthy attempt to get the proactive identification on the heart problems. Based on the latest estimates by the world health organization, in every year there are more than 17.9 million deaths occur in global level because of the cardiovascular disease. (Seckeler & Hoke, 2011) and this statistics is gradually increasing in every year. Since, the heart disease is the major aspect to deal with, however, it’s not easy to proactivity identify the nature of the heart problems and latest technologies in analytics and data science will be applied. Machine learning in the medical science domain has come across a significance journey by innovating novel products to get the predictions and insight. Therefore, predicting heart issue will be a value addition process for better improvement of the medical science domain. Diagnosis can be done proactively based on the machine learning predictions and once the prediction on heart problem diagnosis has been derived, the necessary action will be taken based on the and this will enhance the unnecessary problem with heart issues or problems. In order to do that, finding critical attribute and initiate the data collection will be a significance process and however, with the help of the secondary data, a power full data source has been found and the following is the brief explanation on that.

# Dataset

The data set has been downloaded for the UCI machine learning repository and this has been prepared with 76 attributes, however, most of the published experiments have been completed using a subset of 14 of them. Therefore those 14 variables have been used for this analysis and model fitting process.

Data source link : https://archive.ics.uci.edu/ml/datasets/heart+disease

Attribute Information:

Only 14 attributes used:

1. (age): Age in years

2. (sex): (Male 1 and Female 0)

3. (cp)

cp: chest pain type

Value 1: typical angina

Value 2: atypical angina

Value 3: non-anginal pain

Value 4: asymptomatic

4. (trestbps)

Resting blood pressure (in mm Hg on admission to the hospital)

5. (chol) : serum cholesterol in mg/dl

6. (fbs) : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7. (restecg) :

resting electrocardiographic results

Value 0: normal

Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. (thalach) : Maximum heart rate achieved

9. (exang) : Exercise induced angina (1 = yes; 0 = no)

10. (oldpeak) : ST depression induced by exercise relative to rest

11. (slope) : The slope of the peak exercise ST segment

12. (ca) : Number of major vessels (0-3) colored by flourosopy

13. (thal) : 3= normal; 6 = fixed defect; 7 = reversable defect

14. (target) (the predicted attribute) : Diagnosis of heart disease yes(1) and No (0)

# Data Preprocessing:

As the first step, the data variables have been investigated and based on that, target variable has “balanced nature” of binary classification. Because, it has 55% of 1s and 45% of 0s. And, in this data set there is a mix of categorical and continuous data and based on that necessary actions have been taken based on the data pre-processing.

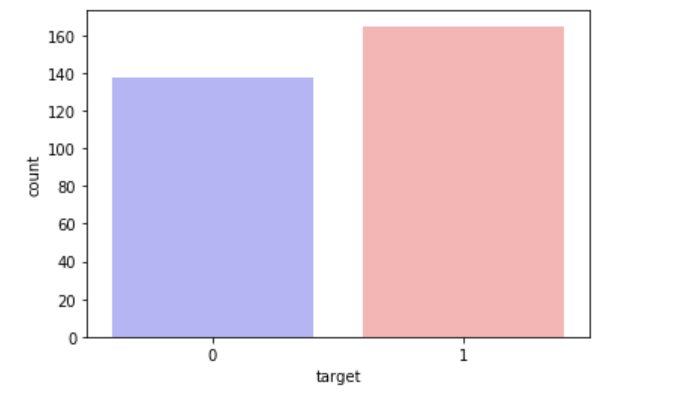


Figure 1: Investigating the Target Variable

**Create Dummy Variables:**

Basically, categorical variables have been identified as a very important set of variables which hide and carry lot of useful information in the data set. Therefore, there should be a method to deal with categorical variables before fitting the machine learning models. If this process has not been done before the developing of the machine learning models, then as the results of that several important variables will be missed in the model fitting process. In other words, the model which has been trained with numerical variables only does not give a perfect result in the model evaluation process.

**Dummy Coding with Python.**

This has been identified as the common method to convert the categorical variable into the numerical variable. This approach creates another attribute which has one level representation of the categorical variable. The presence of that level represents by 1 and otherwise 0. Likewise, if one categorical column has n levels this will create n number of new attributes which has only 0s and 1s as the representation. In this way, this will convert the categorical attributes into the numerical attributes. ( Brownlee, 2020)

To apply this to this study, the dummy approach has been used as a pre-processing step.

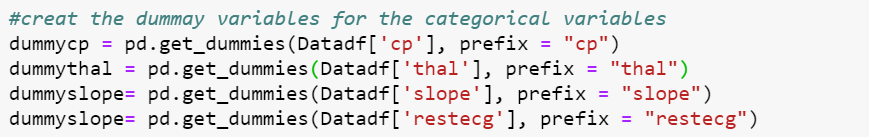


Figure 2: Approach to create dummy

The variables which has the categorical nature have been applied the dummy approach to achieve the above explained task. The variables are “cp”, “thal”, “slope” and “restecg” and at the initial variables explanation all those variables are at categorical nature and now with this dummy approach those have been converted into the numerical features.

get\_dummies() function in the pandas in python has been used to this purpose. And as a value addition to the model fitting process, prefix method has been used and as the results of that, it will create the new column to the data frame which has a prefix related to the level of that categorical column.

The following is the example got after the above process:

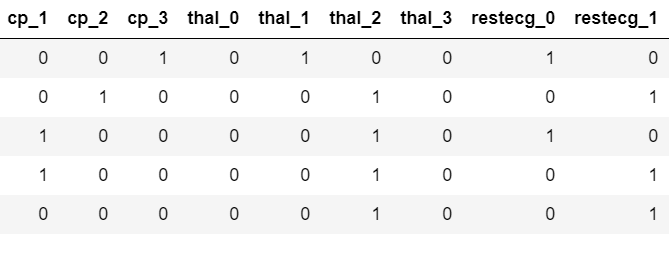


Figure 3: Dummy variables creation

Once the dummy approach has been performed, the initial categorical variables have been dropped.

**Normalizing the Data:**

This is a value addition process for machine learning model fitting. Because, this approach will be put all the numerical variables into a common scale. Because, the scale and measurement of a column have a significance impact for the model fitting process. For an example, the age column has values with the measurement years and weight column has values with the measurement with kg (Kilograms). Therefore, the value 50 in those two variables are not equal. Therefore, we need to do the scaling method for this.

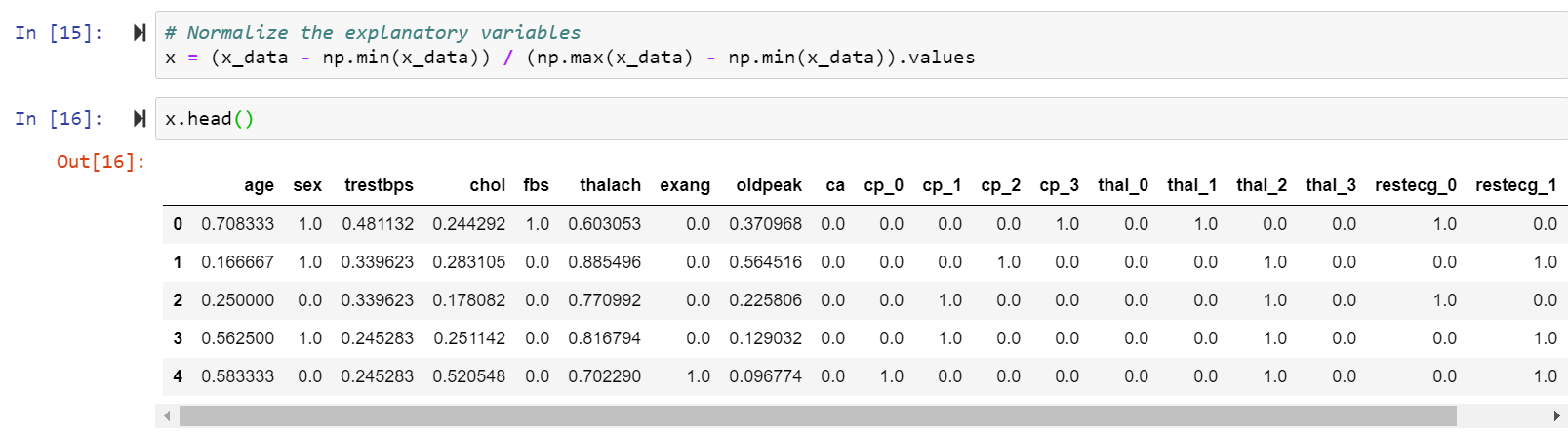


Figure 4. Process to create normalized variables

In order for it to work, min max approach has been used.

Min-max normalization approach is one of the most significance approach to do the normalization of the data. In there, minimum values get transformed into the 0 and maximum value gets transformed into the 1. The values other than that get transformed into the decimal values between 0 and 1s. Then at the end of this process, all the values are in between 0 and 1s.

**Train and Test Split:**

In machine learning, the initial pre-processed data set will be divided into train and test to data set. The model will be fitted based on the training data and the model evaluation process will be done based on the test data. In order to do that, the following approach has been added. However, firstly the main data set has been divided into the x which contains all the explanatory variables and y which contains all the dependent variable.



Figure 5. Step to split the data

Then, we have applied the above functions and in that, 20% has been allocated to the test data set. And 80% of the initial data, will be allocated to the training data.

# Model Fitting:

This is a classification technique in machine learning because, the dependent variable here is the “Target” variable and which has binary labels called 1s and 0s. Based on that, as the initial step logistic regression model and gradient boosting algorithm have been implemented.

**Logistic Regression Model**

Logistic regression is used for binary classification problems and this is using sigmoid functions to get the classification. (Brownlee , 2020)

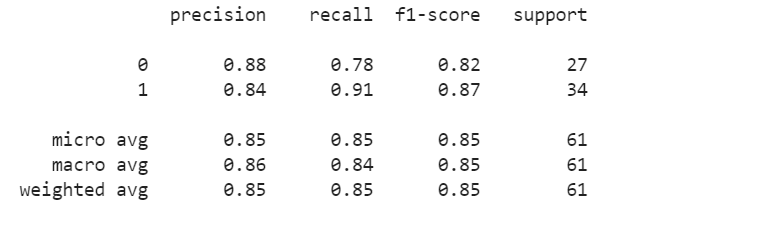


Figure 6. Model results for logistic regression

Based on the above results, the accuracy is at 85% and by looking at the precision of the 0s is nearly 0.88 and for recall this is at 0.78 and for 1s precision is at 0.84 and recall is at 0.91 based on that, logistic regression has done a good job to classify the labels in the data set.

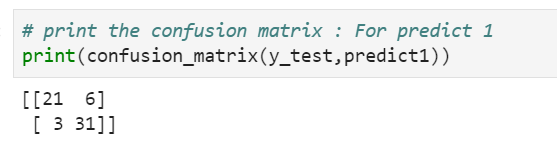


Figure 7 . Confusion Matrix: Results

The following are the calculated results for model evaluation process.

Accuracy = (21+31) / (21+31+6+3) = 85%

Precision = 31/ (31+6) = 83%

Recall = 31 / (31+3) = 91%

As a result, the logistic regression has done a good job since, not only accuracy is high but the precision and recall as well.

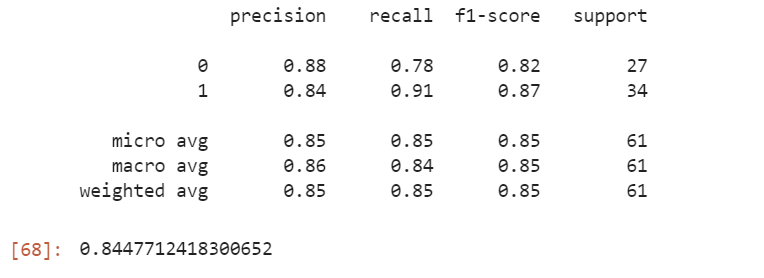
As the next step, ROC AUC measurement has been taken and this is one of the most important measurement for machine learning classifications. In theory, if this is close to 1, the perfect classification will be occurred. However, based on the model fitting process of logistic regression, this measurement is nearly 0.84 and this is a good score of this ROC AUC measurement.



Figure 8 : ROC AUC results

**Hyper Parameter Tuning:**

Hyperparameters of the logistic regression has been done. However, similar results to the default model training has been derived. Therefore, we can conclude that the default model taring had the best hyperparameter set.



**Developing the Gradient Boosting Classifier**

This is one of the best machine learning algorithms that based on the ensemble learning and in gradient boosting it uses the sequence of model building process to get better outcomes. In gradient boosting algorithm, it combines the many weak learning models together to create a strong predictive model. The based is decision tree and gradient boosting used tree-based method to get the outcome. (Aliyev, 2020)

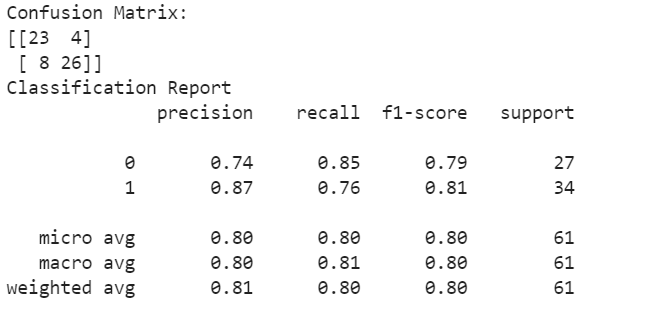


Figure 9. Model results for gradient boosting

The following are the model evaluation results:

Accuracy = (23+26) / (23+26+4+8) = 80%

Precision = 26/ (26+4) = 86%

Recall = 26/ (26+8) = 76%

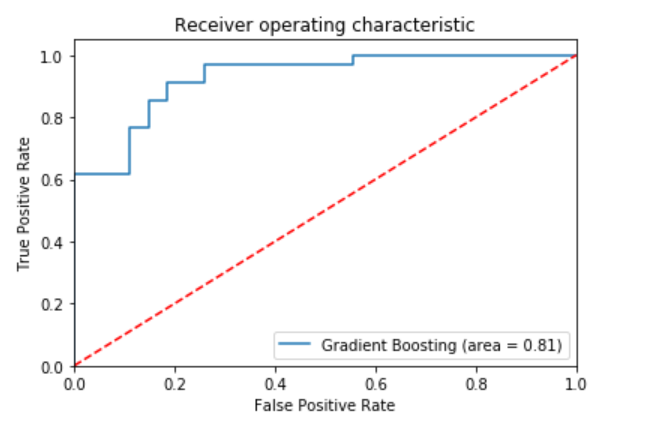
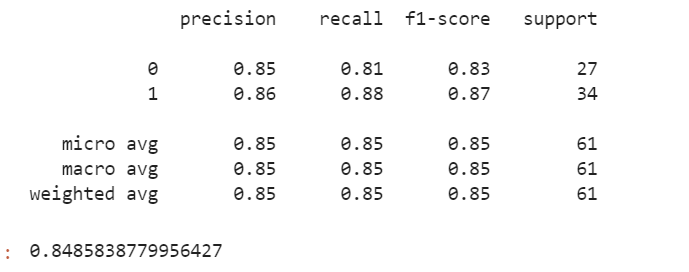


Figure 10. ROC AUC Results: Gradient boosting

Based on the above results, the accuracy is at 80% and by looking at the precision of the 0s is nearly 0.74 and for recall this is at 0.85 and for 1s precision is at 0.87 and recall is at 0.76 based on that, gradient boosting has done a good job to classify the labels in the data set, however the model performance is not better than logistic regression (default and hyperparameter optimized) . Here, we can conclude that this data set can be well classified using the logistic regression model. However, this gradient boosting model has been fitted with default model paraments and now we need to optimize the model hyperparameters parameters with grid search.

**Hyperparameter Tuning for the Gradient Boosting**

Hyper parameter has been tested with grid search option and based on that, we got improved results than default model building process. And, this results for precision and recall and f measurement is better than logistic regression model for both default and parameter optimized.



# Conclusion:

Predicting the risk of the heart disease is a binary classification problem and this has been done based on the two main classification algorithms. The two methods are at two extreme situations that logistic regression is using sigmoid function to get the classification based on the regression approach and gradient boosting tree is using tree-based approach to get the predictions. Based on that, gradient boosting model has performed well to get the better results for the test data set. Then we can conclude that, optimized (hyperparameter) gradient boosting model is a better classifier to predict the heart failure risk based on this data set.

# References

Brownlee, J., 2020. *Ordinal and One-Hot Encodings for Categorical Data.* [Online]   
Available at: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/

Aliyev, V., 2020. *Gradient Boosting Classification explained through Python.* [Online]   
Available at: https://towardsdatascience.com/gradient-boosting-classification-explained-through-python-60cc980eeb3d

Brownlee , J., 2020. *Logistic Regression for Machine Learning.* [Online]   
Available at: https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Seckeler, M. & Hoke, . T., 2011. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. *Clinical Epidemiology,* 3(1).

Terence , S., 2020. *Understanding the Confusion Matrix and How to Implement it in Python.* [Online]   
Available at: https://towardsdatascience.com/understanding-the-confusion-matrix-and-how-to-implement-it-in-python-319202e0fe4d