Project Milestone 3

Kiran Chowdary Komati

College of Science and Technology, Bellevue University

DSC680-T301: Applied Data Science (2233-1)

Dr. Catherine Williams

January 10, 2022

Proposal:

Topic: The topic I have selected as my first project is “Heart Attack prediction”.

Business Problem:

The goal of this project is to help identify if a person is prone to heart attack or not using different machine learning algorithms.

Background/History:

Heart Attack prediction.

In the current day fast paced modern world, the no. of cases of heart attack are on a rise. There are several factors that contribute to it. Using machine learning models, we’ll analyze different attributes of persons to predict if a person is prone to heart attack or not. This could help in identifying the risk and the patient can take necessary precautions to minimize the risk there by even extending the life expectancy of the individuals.

Data Explanation:

Data sets:

I’m using the data set “Heart Attack Analysis & Prediction Dataset “from Kaggle. It contains one csv file heart.csv. This has 303 records and 14 different attributes of a patient which are listed below.

age - Age of the patient

sex - Sex of the patient

cp - Chest pain type ~ 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic

trtbps - Resting blood pressure (in mm Hg)

chol - Cholestoral in mg/dl fetched via BMI sensor

fbs - (fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False

restecg - Resting electrocardiographic results ~ 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy

thalachh - Maximum heart rate achieved

oldpeak - Previous peak

slp - Slope

caa - Number of major vessels

thall - Thalium Stress Test result ~ (0,3)

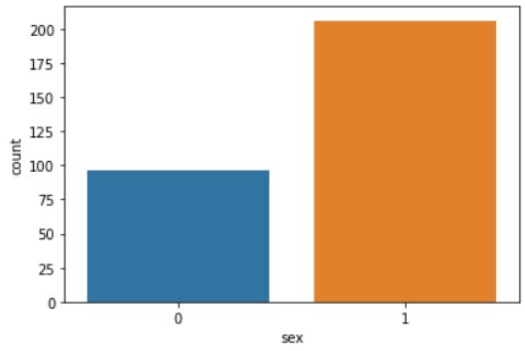
exng - Exercise induced angina ~ 1 = Yes, 0 = No

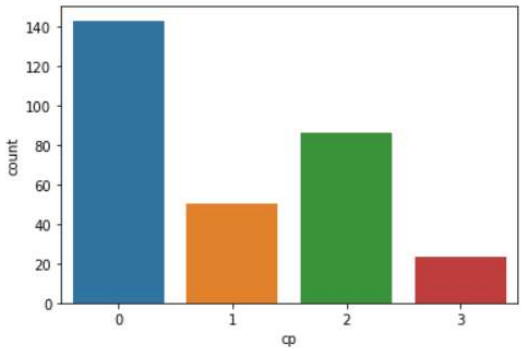
output - Target variable

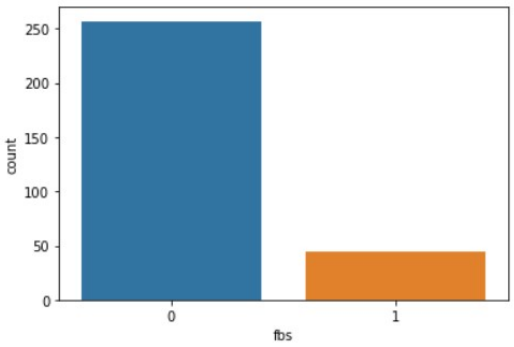
target variable is “target” with values “0” and “1” where “0” represents less chance of heart attack and “1” represents more chance of heart attack.

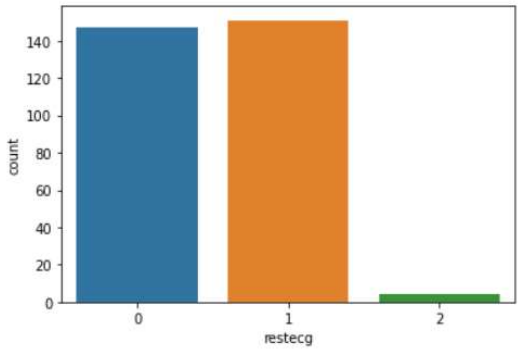
Methods:

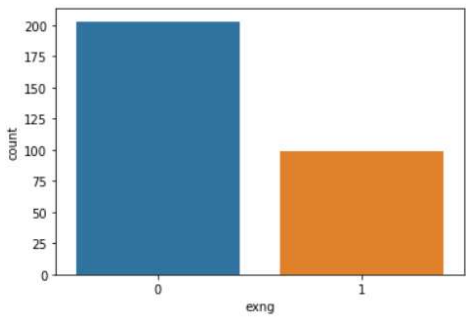
Exploratory Data Analysis was performed on the data set using Python. We initially checked to see if there are any records that has NULL values which signifies incomplete information of the patient. We checked for the duplicates as well to make sure that we are not dealing with duplicate data. We found only one duplicate record in the entire set, and we deleted the duplicate by keeping the last occurrence. Later we plotted different graphs to better understand the distribution and the data.

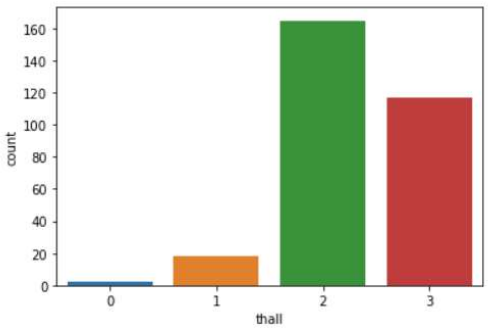




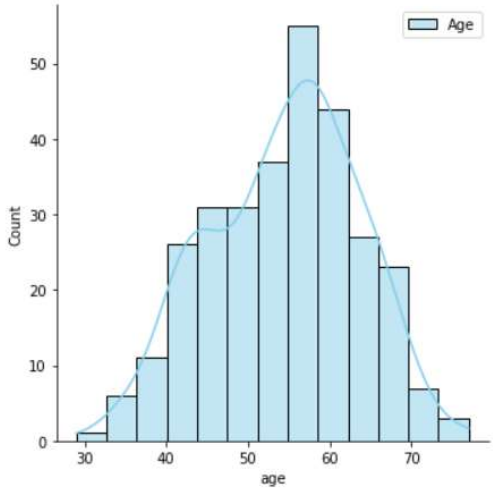




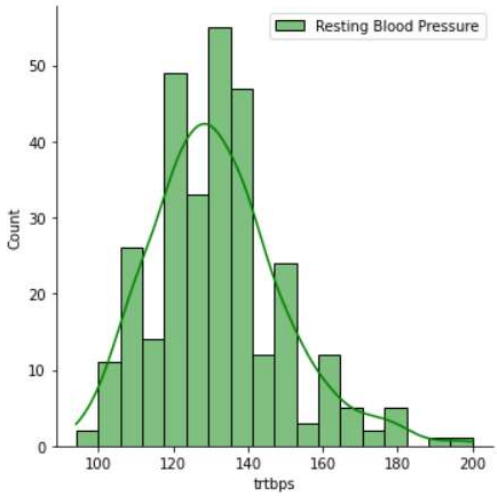




Age Distribution:



Resting Blood pressure:



Analysis:

The data is split into train and test data sets in the ration 70:30. We used a Standard Scaler to perform the task of Standardization. Our dataset contains variable values that are different in scale. As these columns are different in scale, they are Standardized to have a common scale while building the models. in this case the gender class is not balanced, and hence I used SMOTE on the test set to balance them.

We used logistic regression and Random Forest on the training data set and used these models to predict outcomes for the test data sets. We then compared the results with the actual test sets. We got an accuracy of 82% while using the logistic regressions and a slightly reduced accuracy of around 80% while using random forest.

Conclusion:

Logistic regression yielded a slightly better accuracy when compared to the Random Forest algorithm and hence I'm leaning towards the logistic regression as the final model to use on this data set. Both the models seem to take same time.

Limitations:

Although the data sets did not contain any NULL data, there could be some other variables that may be better in predicting the heart attack. Also, this model needs to be tested on a larger data set to make sure that the accuracy of the model is not affected. Also, we used SMOTE method to balance the data and this method has its own limitations. SMOTE's procedure is inherently dangerous since it blindly generalizes the minority area without regard to the majority class. This strategy is particularly problematic in the case of highly skewed class distributions since, in such cases, the minority class is very sparse with respect to the majority class, thus resulting in a greater chance of class mixture**.**

Challenges:

In the real world, it could be a little difficult to gather the data to related to the patients due to the PII restrictions although the data I found from Kaggle was publicly available. Also, it could be a little challenge to gather data from multiple sources if it needs different types of tests to be performed on the patient.

Future uses/Additional Applications:

A thorough Analysis of the models using vast amount of real data will help in prediction of the heart health of the patients and could be used to take a better caution of the health of the risk is high.

Recommendations:

Based on the available information and the analysis performed on the limited data available, this model has a better accuracy of predicting if a person is at risk of having a heart attack. But as this involves health of person, much more analysis needs to be performed when the larger data sets are available.

Implementation Plan:

Once we can find the large data sets, this model needs to be trained and tested to make sure that the accuracy is not affected. Backup options or models needs to be developed as well to make sure that we have a working model that is better than a random guess incase the original model doesn't workout with the larger data sets.

Ethical Assessment:

As this project is related to the health of individuals and as it contains sensitive information, I made sure to not use any PII information such as individual names. Also, the data needs to be presented in accurate form without any modifications or misrepresentations. Also, steps must be taken to avoid any bias towards any gender. The users of this model also need to be aware of the limitations of the model and not use it for self-medications and should seek doctor incase of any discomfort even though their result may suggest otherwise using the model.

References:

Rahman, Rashik (2020). *Heart Attack Analysis & Prediction Dataset* Retried from https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Dharmaraj (May 16, 2022). *Logistic Regression with StandardScaler-From the Scratch* Retrieved from

https://medium.com/@draj0718/logistic-regression-with-standardscaler-from-the-scratch-

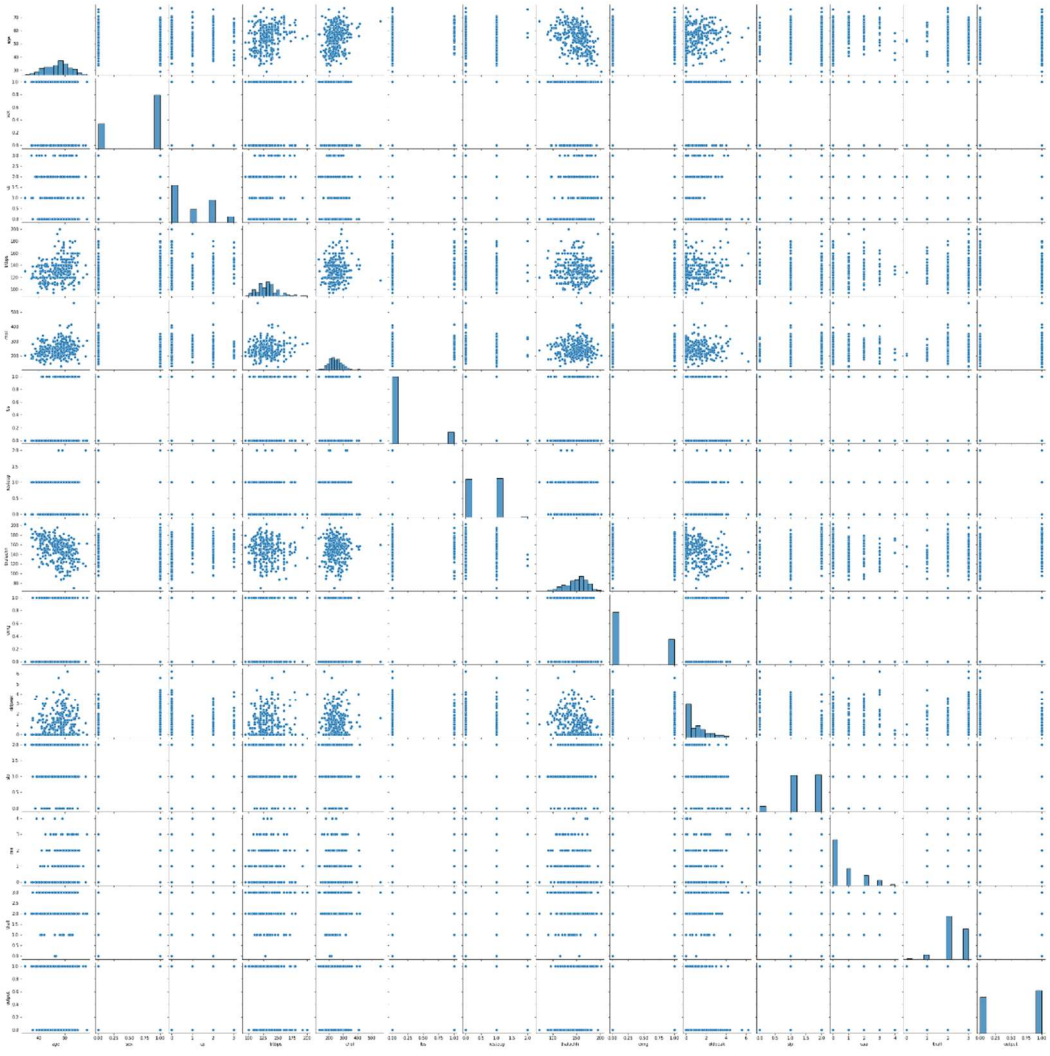
ec01def674e8

Genesis(June 26 2018). *SMOTE (Synthetic Minority Oversampling Technique)* Retrieved from

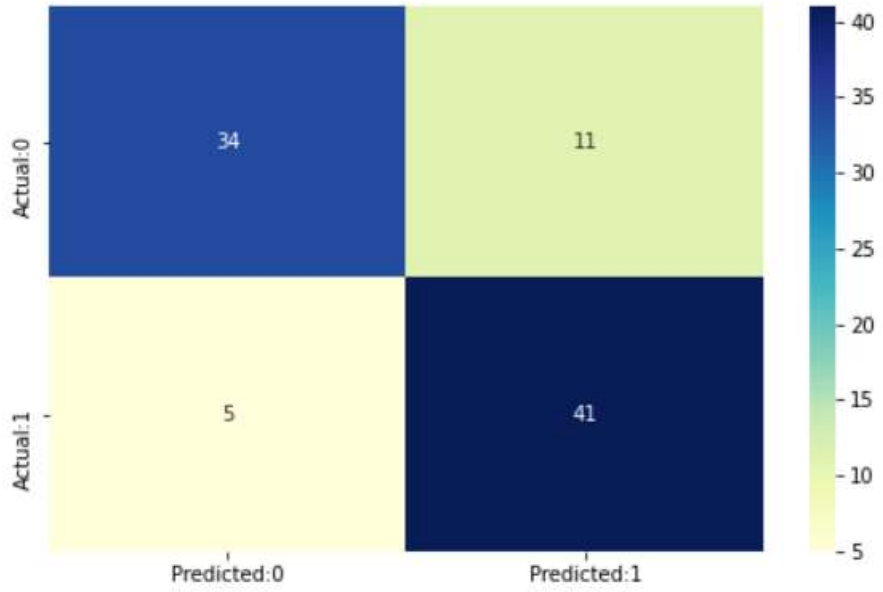
https://www.fromthegenesis.com/smote-synthetic-minority-oversampling-technique/

Appendix:

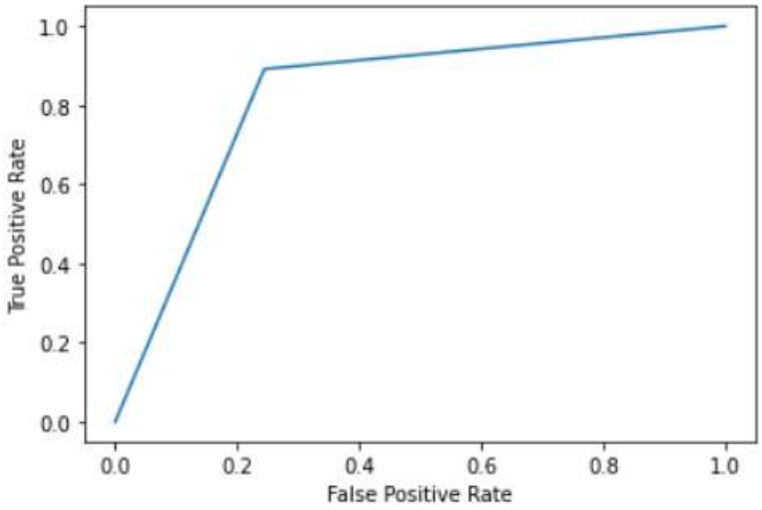
Pair Plots:


age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
target

Confusion Matrix:



ROC Curve:



Questions:

1.Why do we need to predict the Risk of heart diseases?

Predicting the risk of heart disease will help patients to take precautions there by extending the life expectancy of individuals with risks.

2. How does this project helps in predicting the heart disease?

The attributes from the data set will be used for Analysis using the Machine learning models Logistic regression and Random Forest to predict the Risk.

3. Why is the balancing done? Will it have any affect on the original outcome?

Imbalanced data can have major impact on the model accuracy. Hence balancing needs to be done to make sure that the classes have equal representation. For eg: if there are two classes and one represents 90% of the data, model can learn to predict this class 100% of the time. But the accuracy will still be 90% which is called accuracy paradox.

4. What parameters are used in the model?

There are 14 different parameters that we have used in this model including Age, sex, blood sugar to name a few.

5. What are the different models that are built as part of this project?

We built Logistic regression and Random Forest Algorithms as part of this project. We will explore other options as well as backup once we have a large dataset.

6.Are there any new features built from the existing features?

No. We didn't have to build any new features from the existing features but having said that there could be some unidentified features that may have strong outcome on the accuracy of the data models.

7. Can adding any other features affect the model accuracy?

There could be some features which are not part of the attributes in the dataset that may affect the accuracy.

8. is there any additional data to supplement the dataset used?

We currently do not have any supplement data set for our data set. Also, it could be a little difficult to get it owning to the PII information that it may contain. We need to take necessary steps to mask any sensitive content and then can be used as a supplement to our data set.

9. Why is accuracy a poor measure for imbalanced data sets?

An imbalanced dataset could lead our model into deceptive accuracy results because of the accuracy paradox. Confusion matrix can be used for better metrics as they take false positives and false negatives into account.

10. Can this methodology be used in other areas of medical sciences?

The same concept can be applied in the different areas of medical sciences but thorough analysis and testing needs to be done as different areas may have different datasets before concluding a model.