MACHINE LEARNING AND DATA SCIENCE

ENCS5341.

ASSIGNMENT#3.

PREPARED BY:

SALEH KHATIB – 1200991.

SECTION 1.

DR: YAZAN ABU FARHA.

DATE: 26/1/2023.

Table of Contents

Table of figures

Introduction

Accurate and timely prediction tools are critical in the healthcare industry, particularly for life-threatening conditions like heart attacks. The objective of this project is to create a strong classification module for Heart Attack Analysis & Prediction in order to address this pressing issue. The main goal is to develop a predictive model that, given pertinent medical data, can accurately determine the risk of a heart attack.

This predictive module was built by studying a number of machine learning models. The nearest neighbor algorithm is baseline model with k=1 and k=3. The AdaBoost algorithm and decision tree were also selected as more advanced models.

The decision tree and AdaBoost algorithm were chosen due to their computational efficiency, suitability to the given data set, and simplicity. Given the relatively small size of the data set, these algorithms are ideal for efficient analysis without sacrificing predictive accuracy. Most importantly, its simplified parameter configurations facilitate implementation, accelerating development and deployment.

It came down to dropping support vector machines (SVM) and logistic regression to properly analyze the dataset. Strong relationships between factors were found to be less suitable for using SVM or logistic regression, making decision tree algorithms more suitable

Detailed experimental parameters for evaluating the performance of the models are described. Specificity, recall, F1 score, area under the ROC curve (AUC), and confusion matrix generation are some of these metrics. Using this multifaceted approach, the project hopes to provide an accurate understanding of the strengths and weaknesses of the models in classification, as well as an accurate prediction of heart disease.

Dataset

```
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trtbps    303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalachh  303 non-null    int64
 8   exng      303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slp       303 non-null    int64
 11  caa       303 non-null    int64
 12  thall     303 non-null    int64
 13  output    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
None
```

Figure 1: Dataset information.

As we see, in this dataset we have: 13 features, 303 examples (rows), with the data types in the figure.

The description of the features:

- Age: Age of the patient
- Sex: Sex of the patient
- exang: exercise induced angina (1 = yes; 0 = no)
- ca: number of major vessels (0-3)
- cp: Chest Pain type chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
- trtbps: resting blood pressure (in mm Hg)
- chol: cholesterol in mg/dl fetched via BMI sensor
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- rest_ecg: 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
- thalach: maximum heart rate achieved
- target: 0= less chance of heart attack 1= more chance of heart attack

now for dataset analysis:

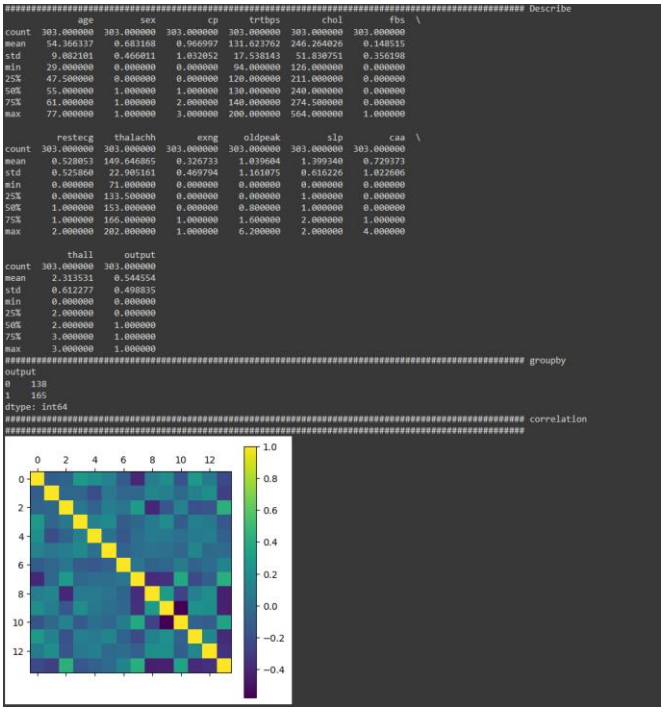

Figure 2: Dataset analysis

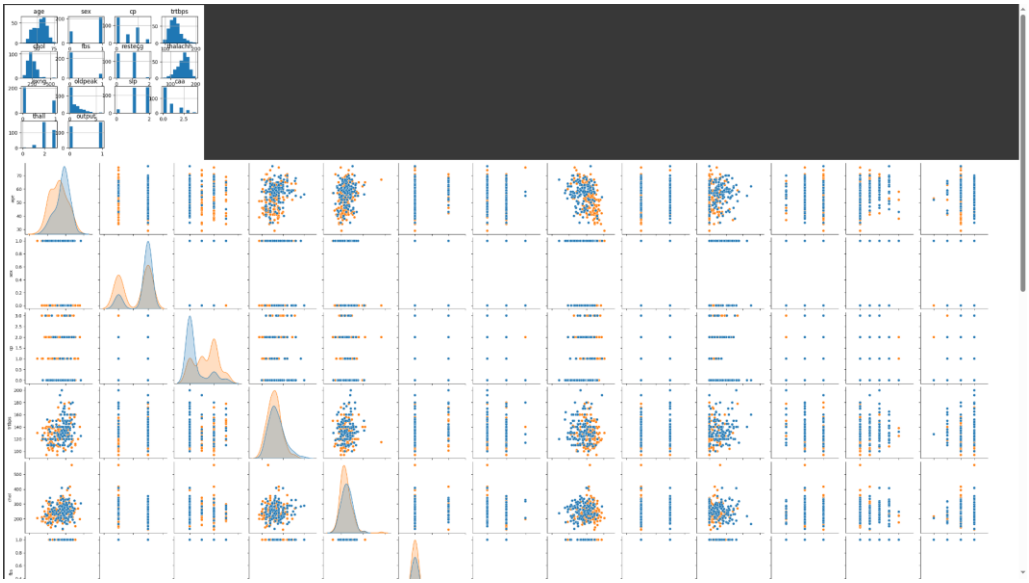

Figure 3: histograms and Scatter plots for the Features.

It is evident that there is noise in the features and that not all of the features have a normal distribution, and also there is no an excellent correlation between the features. On the other hand, it has a good set of features for my task and is good for algorithms like DT and Ada because of the above and binary classification task, it doesn't have any null or incorrect values and doesn't require preprocessing or values encoding. It's a solid data set, then, with a few weak points.

Experiments and Results

Our experiments will be as follow: nearest neighbor with k=1, k=3, Decision Tree, hyper-parameter (n_estimators) selection for AdaBoost algorithm, and finally AdaBoost algorithm with the best n_estimators.

Nearest Neighbor with k=1:

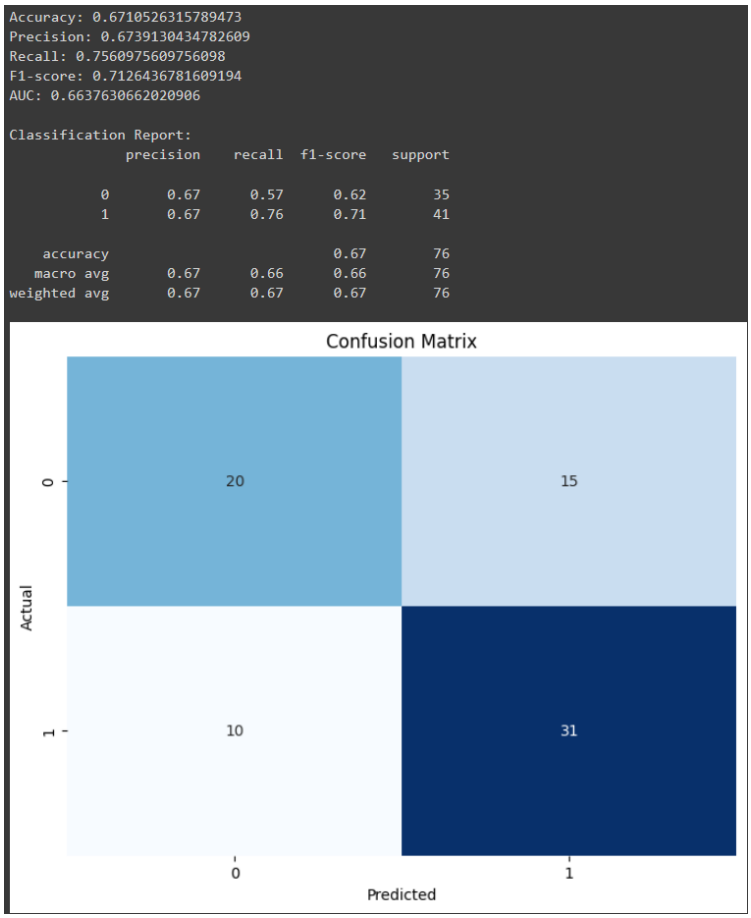

Figure 4:Nearest Neighbor with k=1 experiment.

As we see, this module is bad in term of evolutions metrics, with accuracy = 67.1%. but it takes just 0.016 second.

Nearest Neighbor with k=3:

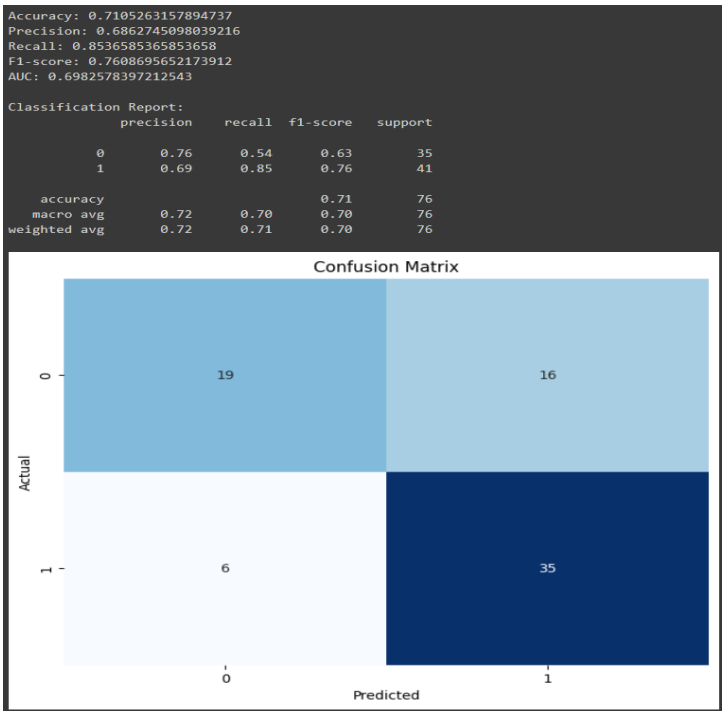

Figure 5: Nearest Neighbor with k=3 experiment.

As we see, this module is also bad in term of evolutions metrics except Recall which is not bad. Accuracy = 71.05%. but it takes just 0.013 second, so K =3 is much better than K = 1.

Decision Tree:

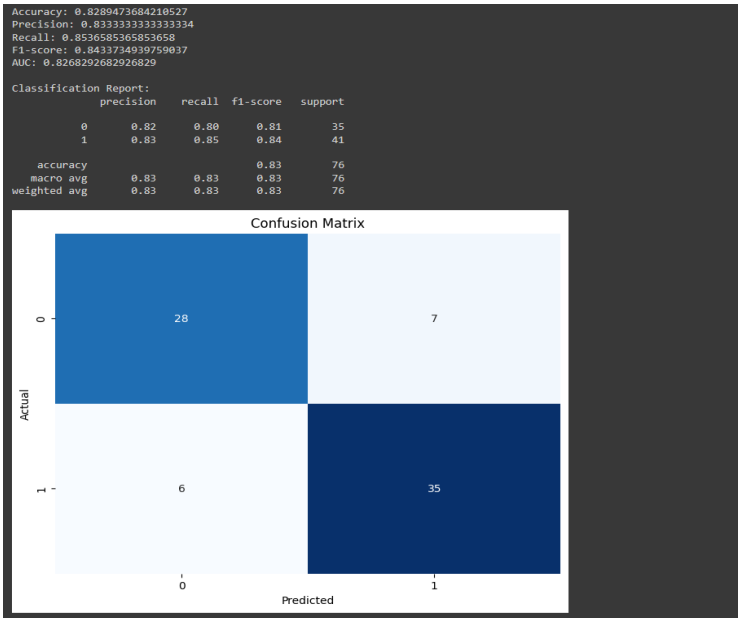

Figure 6: Decision Tree experiment.

As we see, this module has improved much better in the metrics than K-NN, but it takes 0.303 second.

AdaBoost algorithm hyper-parameter:

Figure 7: AdaBoost algorithm hyper-parameter experiment.

n_estimators is the number of weak learners, so I tried twenty values in this formula: N_E = 10 + i*3. Which (i) is the value, and I get that i=10 is the best one (n_estimators = 40). And I even test the learning rate, and gives me that the best value is the based value = 1.

AdaBoost algorithm:

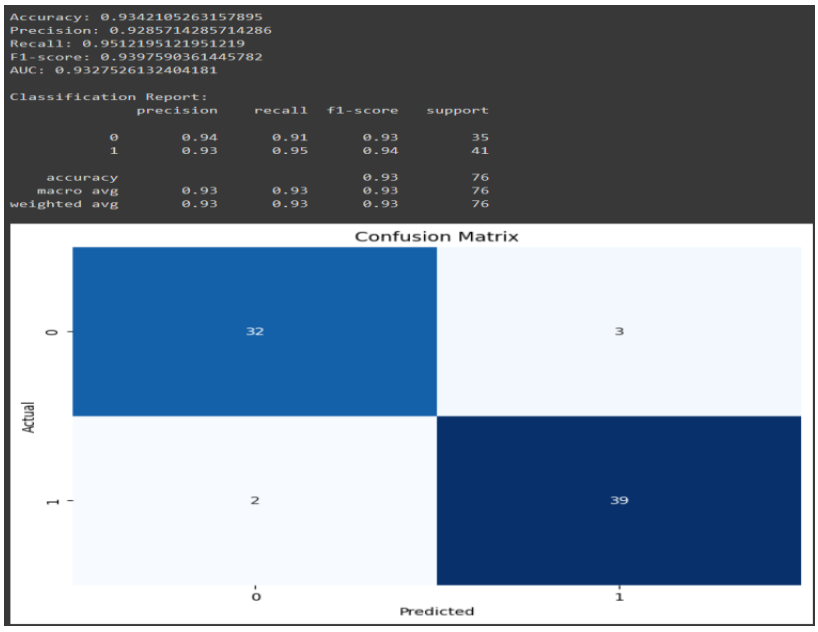

Figure 8: AdaBoost algorithm experiment.

As we see, AdaBoost algorithm with n_estimators = 40, is the best model I have with accuracy = 93.42%, which is an excellent one, compere to K-NN. the it take some time = 0.346 second, which is the longest.

Analysis

The performance of my classification model was evaluated using various metrics, such as accuracy, precision, recall, F1 score, area under the curve and I started the analysis as an initial model, using K-NN with k=1 and k=3 among. Unfortunately, the amplitude curse caused by the presence of 13 elements caused these models to perform poorly.

To increase complexity, I went to a decision tree model. Although it showed significant improvements over K-NN, the game was still less than I wanted. This is because the sensitivity of decision trees to overfitting is evident, as it depends on the rules derived from the training set.

Undeterred, I used the AdaBoost algorithm to achieve higher difficulty. Tuning parameters such as n_estimators and learning rate were the most successful model in this study with significant improvements in performance in all evaluation parameters but it should be noted that some shortcomings remain These differences can be explained in by the presence of similar examples. Despite this drawback, the AdaBoost algorithm proved to be the most efficient solution for the given classification problem.

Conclusions and Discussion

Analysis of the Heart Attack Predictive Analysis Project gained insight into the effectiveness of machine learning models for heart attack prediction based on multiple clinical symptoms The primary goal of the trials was to develop a reliable classifier so, the AdaBoost algorithm, . decision trees and nearest-neighbor algorithm

Limitations: Curse of Design: The presence of 13 features made it more difficult to use neighbor models and nearest decision trees, which negatively impacted their performance

The decision trees felt too convenient, which hampered their ability to generalize the processing of additional data points.

Difficult to distinguish between similar samples: Although very efficient, the AdaBoost algorithm had trouble distinguishing between samples with similar features, resulting in some residual errors

The evaluation metrics have provided a comprehensive understanding of model performance, leading to an in-depth analysis of the advantages and disadvantages of different areas of classification

The project emphasizes the importance of selecting advanced models and adjusting their designs to achieve the best results.

AdaBoost is a promising tool for cardiac prediction because it shows remarkable accuracy when parameters are carefully selected.

The limitations point to the need for continued research and improvement of the model to meet the challenges posed by complexity.

For future: in addition, enhance prediction skills, future research might also have a look at ensemble strategies or extra superior deep getting to know strategies.

To overcome difficulties in distinguishing between comparable samples, it could be vital to acquire greater statistics or use state-of-the-art engineering techniques.

In precis, even though the task has produced a strong class module for coronary heart assault evaluation, it's far critical to apprehend the constraints of the mission and put into effect further development of the model to enhance predictive accuracy and reliability in real-international healthcare situations