بسم الله الرحمن الرحيم

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

Machine Learning & Data Science (ENCS5341)

Project Report

Prepared by : **ID :**

Momen Salem **1200034**

Basheer Arouri **1201141**

Supervised By : Dr. Yazan Abu Farha

Section : 2 **Date : 25-1-24**

Place : Ramallah

Table of contents

List of Figures

List of Tables

Introduction

There are many medical problems in our world and if we can classify some of crucial diseases then we add benefit to people. We decide to choose binary classification task about heart disease which was chosen and studied properly before apply machine learning model on it. This choice was because the heart diseases is the most diseases that causes death for long time. We try to train different models which are: (KNN), SVM, AdaBoost and MLP and then see their behavior using evaluation metrics like: accuracy, recall, specificity, and F1-score, but our main evaluation metric is **recall** because in medical classification we need to minimize the false negative as possible.

Dataset

The data source collected from:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations
Duplicated: 272 observations

Final dataset: 918 observations

In this project, we used a realistic dataset that represents the **Heart Disease**. Here is a reference for our dataset: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?fbclid=IwAR1RXmrmwJgUyG0mBrl6C71YpohiWaYB_VqJQ2tBGD_-0hB9D2P4GdBa-ik

In our dataset, there are 11 features (and the target) that we worked on. Here is a brief explanation for them:

Attribute Information

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

Figure 1. Features Description

As we see, there are some features that are continuous features: **Age, RestingBP, Cholesterol, MaxHR, Oldpeak**, and there are nominal features: **Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope**. We showed some statistical analysis and applied exploratory data analysis, we will show them later on.

As we said, our dataset consists of 11features (and the target) and 918 examples. Each example represents one realistic scenario that this person for specific features is injured or not.

	Age	Sex	ChestPainType	...	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	...	0.0	Up	0
1	49	F	NAP	...	1.0	Flat	1
2	37	M	ATA	...	0.0	Up	0
3	48	F	ASY	...	1.5	Flat	1
4	54	M	NAP	...	0.0	Up	0
..	...	..	...	...	...	...	...
913	45	M	TA	...	1.2	Flat	1
914	68	M	ASY	...	3.4	Flat	1
915	57	M	ASY	...	1.2	Flat	1
916	57	F	ATA	...	0.0	Flat	1
917	38	M	NAP	...	0.0	Up	0

[918 rows x 12 columns]

Figure 2. The Dataset

The count of label instances value is:

```
The Label (Heart Disease) counts:
HeartDisease
1    508
0    410
Name: count, dtype: int64
```

Figure 3. Label count

We saw that people with heart disease = 508 and without = 410

We showed here some statistical and quantitative measures, for continuous and nominal features.

	count	mean	std	min	max	Skewness	Mode	Correlation
Age	918.0	53.510893	9.432617	28.0	77.0	-0.195933	54.0	0.282039
RestingBP	918.0	132.396514	18.514154	0.0	200.0	0.179839	120.0	0.107589
Cholesterol	918.0	198.799564	109.384145	0.0	603.0	-0.610086	0.0	-0.232741
MaxHR	918.0	136.809368	25.460334	60.0	202.0	-0.144359	150.0	-0.400421
Oldpeak	918.0	0.887364	1.066570	-2.6	6.2	1.022872	0.0	0.403951

Figure 4. Quantitative Measures for Continuous Features

```
Correlation between the nominal features and the output 'HeartDisease' ----->
Correlation between Sex and HeartDisease: 0.30275622279754005, Mode for this feature = M
Correlation between ChestPainType and HeartDisease: 0.5403815716169053, Mode for this feature = ASY
Correlation between FastingBS and HeartDisease: 0.26470001743528276, Mode for this feature = 0
Correlation between RestingECG and HeartDisease: 0.10912341027519352, Mode for this feature = Normal
Correlation between ExerciseAngina and HeartDisease: 0.4920494053821572, Mode for this feature = N
Correlation between ST_Slope and HeartDisease: 0.6226642132252813, Mode for this feature = Flat
```

Figure 5. Quantitative Measures for Nominal Features

As shown in previous figures, in case of continuous features, we displayed mean, standard deviation, min, max, skewness, mode and correlation for each feature. Correlation and mode in case of nominal features. As we know, when the correlation coefficient approach 1 or -1, that means this feature has a strong relationship with other feature. In our case, we evaluated the correlation between each feature with the target label, and we can conclude from the previous fact, that the feature '**ST_Slope'** has the strongest relationship with the target label

than other features. We visualized some features, so that the reader can extract the information in a good manner.

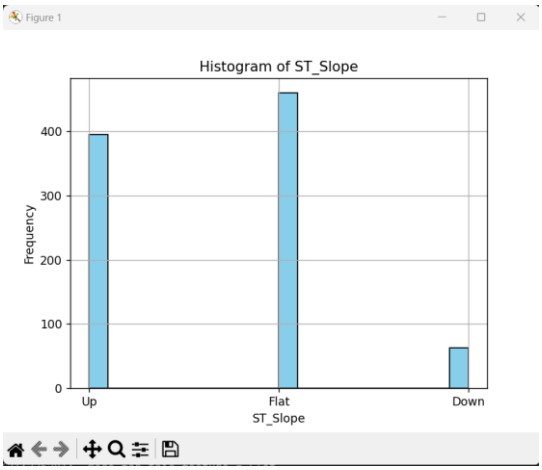

Figure 6. Histogram of ST_Slope

This is a scatter histogram for ST_Slope feature, each bar represents the frequency for those kinds of it. **It is noticeable that the ST_Slope with kind 'Flat' is the most commonly to be according to the people, about 580 people have Flat ST_Slope.**

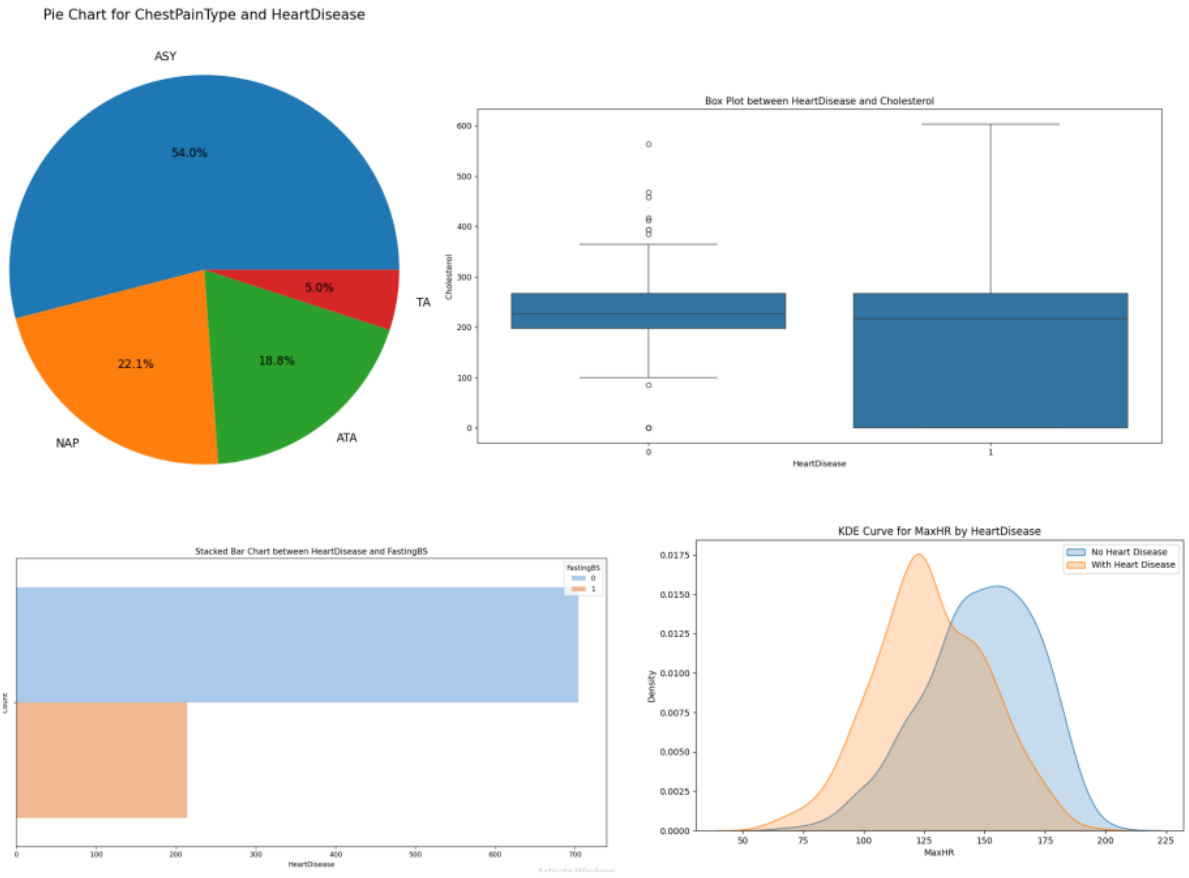

Figure 7. Pie Chart, Box Plot, Stacked Bar Chart and KDE Curve Visualizations

We also plotted and demonstrated some of the other visualization techniques to represents each feature we have. First, we plotted the pie chart for the ChestPainType, and observed how much each kind of it represents from the whole dataset, we can notice that they are as the following: 54% Asymptotic, 5% Typical Angina, 18.8% Atypical Angina and 22.1%

Non-Anginal Pain. Second, we plotted a box plot to illustrate the relationship between heart disease and Cholesterol feature, we can notice that, 50% of injured people with HeartDisease, their Cholesterol are about 220mm/dl, and there is 7 people which they have a highly amount of Cholesterol (higher than the upper wheisker limit, which they are outliers in our case) and 2 people who have a low amount of Cholesterol (lower than the lower wheisker limit, which they are also outliers in our case). Third, we plotted a Stacked Bar Chart to show the count of the whole people who has a positive FastingBS (FatingBS=1) or negative (FatingBS=0). For both: injured and not injured with HeartDisease, we can notice that we have about 700 people with negative FatingBS and 218 without positive FatingBS. Finally, we plotted a KDE Curve which represents the distribution of the MaxHR feature along with HeartDisease. We plotted the curve for each point of the MaxHR, it has a density of the whole dataset. We can notice that the MaxHR for HeartDisease people is about 125 and has the density of 0.0175. Also, we can notice that the MaxHR for people without HeartDisease is about 160, and has the density of 0.0150. If you want to find as a number of people, you can multiply the density with 918 (Density * 918).

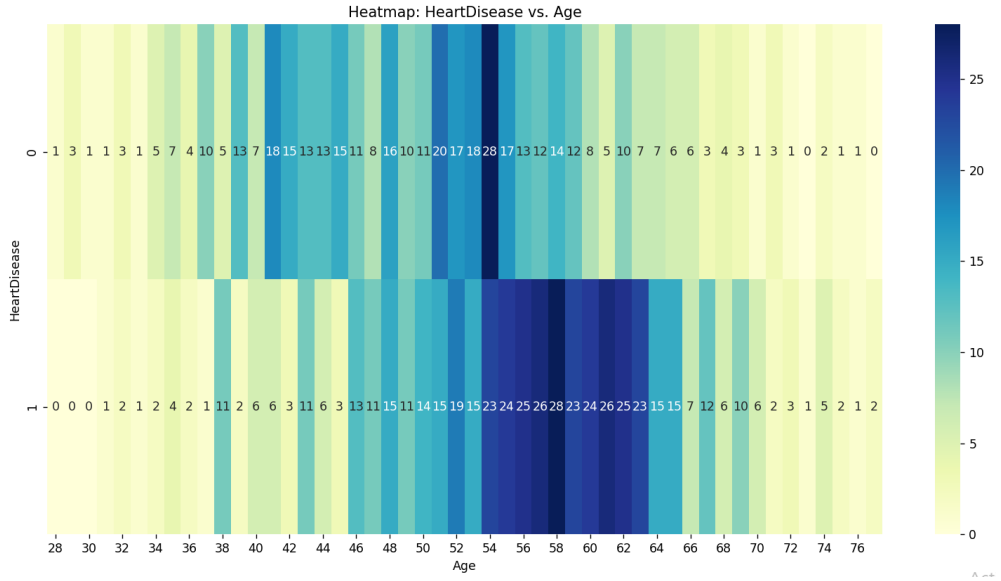

Figure 8. Heat-map For Age Feature

This is a Heatmap for the age feature versus the HeartDisease label. **It is noticeable that the Age with range [54-63] is the most commonly age that the people are suffering from the HeartDisease, which is sensible for our life, that the adult people are more likely to infected with this Disease.**

Experiments And Results

Baseline models

We split our data randomly to training and testing sets (20% of overall data as testing set) and then performed (1 and 3) nearest neighbors as a baseline model with Euclidean distance as our distance metric, choosing Euclidean distance is more suitable when features are scaled as in our case. For 1-NN we have recall = 75.7% and for 3-NN = 84.1%, which evaluated on testing set. It is a good and expected results, because if we have increased the k for NN we expect increase in performance but for specific k (if we increase k after this threshold the

model seems to overfit the training examples). To gain more performance recall, we decided to perform more models on data and see if we could improve our classification recall.

AdaBoost Model

T o improve our classification results, we decided to train dataset using boosting and we choose AdaBoost algorithm for this purpose. For hyper-parameter, we tuned a number of estimators (number of tree stumps) to be trained on K-fold validation, with 5 folds, we saw that on cross-validation, the best estimator number = 201, with recall on testing set = 86%. As a conclude, the number of estimators is good and big, recall metric improved. We can conclude that our data is not linearly separable, because the number of estimators is more than one (data need more than one line to separate positive and negative label class).

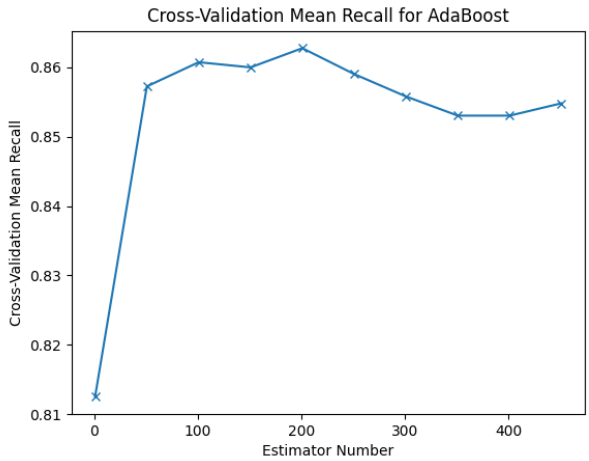

Figure 9. estimator num vs mean recall

Support Vector Machine (SVM) Model

We first started with Support Vector Machine (SVM) Model. For hyper-parameter, we tuned a number of degrees (0 up to 15) to be trained on K-fold validation, with 5 folds, we can notice that at degree of 4, it gained the highest mean recall, which is the most important metric in our case, and this mean recall is about 83.2%, which can be considered as a good recall. For the regularization parameter C, we evaluated the performance (which is the recall in our case), by changed manually the value of C as [0.01, 0.1, 1, 10, 100], and concluded that the best C = 0.1.

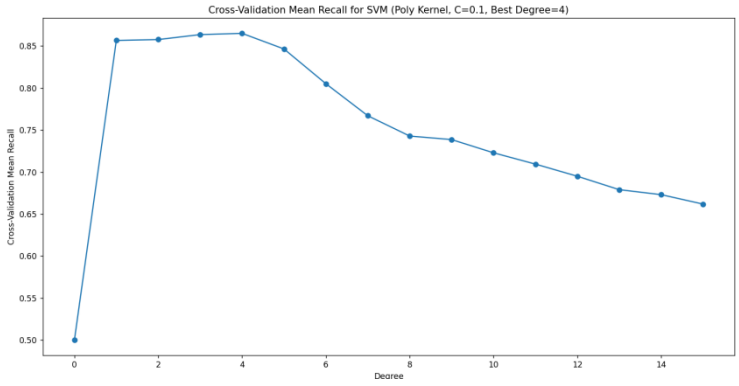

Figure 10. Mean Recalls For SVM With degrees from 0 up to 15

Multi-Layer Perceptron (MLP) Model

We built MLP with two hidden layers, the first one with 5 neurons, and the second with 4, we chose these values by fixing the number of second hidden layer perceptron's and changed the first one, and see the recall on testing set, then concluded that the best value is 5 perceptron's. We decided to have 2 hidden layers because we did not need to spent more time on computing the classification output, also, we wanted a model not be very deep. For optimization algorithm (solver attribute) we chose the default one Adaptive Moment Estimation (Adam for short) and we chose it by experiment another solvers like: stochastic gradient descent (sgd). For activation function, we experimented ReLU, tanh, and logistic (sigmoid) and chose the best one on testing set that was **ReLU**. Our chosen hyper-parameter is the learning rate we tune 4 different values using cross-validation $=$ $[0.001, 0.01, 0.1, 1, 10]$ and observed that when learning rate $=$ 0.001 we have best recall on testing set $=$ 91.6% which is the best value have seen over all models. From the figure, learning rate $=$ 0.001 is the best value obtained on cross-validation for MLP model. It seems that our data not converge quickly, sot, the learning rate is small and the changes of parameters are proportional to this learning rate.

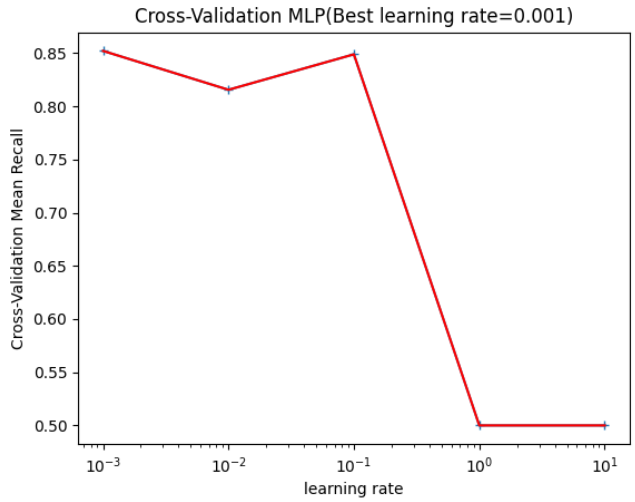

Figure 11. learning rate vs mean recall

Also, the model not overfit because the recall on training set not 100%.

Analysis

From previous parts, we concluded that MLP was the best model on our data set, and to study this model more efficient, we obtained these results:

```
The Metrics on Training set using MLP :
True Positive = 372, False Positive = 49
True Negative = 284, False Negative = 29
Recall = 92.8%, Precision = 88.4%, Accuracy = 89.4%, Specificity =
85.3%, F1 = 90.5%
------------------------------------------------------------------------
The Metrics on Testing set using MLP :
True Positive = 98, False Positive = 9
True Negative = 68, False Negative = 9
Recall = 91.6%, Precision = 91.6%, Accuracy = 90.2%, Specificity =
88.3%, F1 = 91.6%
```

So the MLP was wrong in classified 9 examples as false negative (which is bad in our medicine situation), after this, we traced the data and saw the index of these examples which are :

```
The index of misclassified examples from test set are :
[30, 655, 759, 375, 662, 120, 652, 684, 211]
```

And we print each example with its feature to analyze any pattern or interesting feature.

Table 1 Misclassified examples

Ex\Feat.	Age	Sex	Chest PainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope
30	**53**	**M**	**NAP**	**145**	**518**	**0**	**Normal**	**130**	**N**	**0.0**	**Flat**
655	**40**	**M**	**ASY**	**152**	**223**	**0**	**Normal**	**181**	**N**	**0.0**	**UP**
759	**54**	**M**	**ATA**	**192**	**283**	**0**	**LVH**	**195**	**N**	**0.0**	**UP**
375	**73**	**F**	**NAP**	**160**	**0**	**0**	**ST**	**121**	**N**	**0.0**	**UP**
662	**44**	**M**	**ASY**	**110**	**197**	**0**	**LVH**	**177**	**N**	**0.0**	**UP**
120	**47**	**F**	**NAP**	**135**	**248**	**1**	**Normal**	**170**	**N**	**0.0**	**Flat**
652	**59**	**M**	**TA**	**160**	**273**	**0**	**LVH**	**125**	**N**	**0.0**	**UP**
684	**47**	**M**	**NAP**	**108**	**243**	**0**	**Normal**	**152**	**N**	**0.0**	**UP**
211	**50**	**F**	**NAP**	**140**	**288**	**0**	**Normal**	**140**	**Y**	**0.0**	**Flat**

We can see from examples that when oldpeak = 0.0 then the model misclassified the result, and this affected the performance of model. We can also see that the feature ExerciseAngina when it is almost (N) the model misclassified the result. And this is also occurs when FastingBS almost equal (0). For ST_Slope if the value is (Down) then the model will not misclassify the label (maybe because the number of examples ST_Slope = Down are little).

Table 2 Recap Metrics (Testing Metrics)

Model	Recall	Precision	Accuracy	F1score
KNN (1-NN)	**75.7%**	**84.4%**	**77.7%**	**79.8%**
KNN (3-NN)	**84.1%**	**90.9%**	**85.9%**	**87.4%**
SVM	**83.2%**	**91.8%**	**85.9%**	**87.3%**
AdaBoost	**86%**	**92%**	**87.5%**	**88.9%**
MLP	**91.6%**	**91.6%**	**90.2%**	**91.6%**

Conclusions And Discussion

In this project, we learned many new things and added them into our knowledge for our life. From constructing a model from scratch, judging on a performance of a specific model is good or bad, deciding the best model among a number of models, and finally, analyzing and finding an important and richable patterns in the data. If we want to illustrate it in our journey, we can explain it as the following: we have worked on 3 models (in addition with the KNN, with 1-NN and 3-NN), which are SVM (Support Vector Machine), AdaBoost and MLP (Multi-layer Perceptron). We constructed them, and tuned each one with it's corresponding hyper-parameter, which they are (degree for SVM, number of estimators in AdaBoost , learning rate in MLP).There are might some other hyper-parameters, we tuned them manually, such as number of layers in MLP and number of neurons in the second layer(explained briefly in ***Experiments and Results***). After that, we entered them into testing set, from the highest recall, which is the base metric in our case, because we want to eliminate the False Negative (the model return that this person has not a heart disease, but actually, this person has a heart disease), we chose the best one, which is the MLP in our case.

We then analyzed some important patterns found in the miss-classification examples, that the MLP classified. According to why did the models, SVM and AdaBoost have failed to obtain a high recall, this can be according to many reasons such as that the data is not linearly separable in SVM, or there might be a highly distribution of the outliers, according to AdaBoost, might be the number of estimators is not enough to gain a highly recall, such as the recall on the MLP, we did not use a high number of estimators, because our PCs have limited size and speed. According to the MLP, we explained it briefly why did it failed to model some examples in right way in (***Analysis Part***). We wish in the future to deal with deep learning techniques such as CNN, to gain a higher performance on our data. To add, if we want to improve our data we can tune more parameters for each model and also have more accurate data examples for our medicine problem.