Heart Disease Prediction

DATA1030 Final Report – Yash Bafna

Supervised by Prof. Andras Zsom

Section 1. INTRODUCTION

Heart diseases have become extremely common these days due to changing lifestyle. High blood pressure, diabetes, hyper tension and cholesterol are some of the major factors that lead to heart attacks. With the advancement of technology, we can now predict using some important factors how likely a person is to get a heart disease such as coronary-artery disease, cardio-vascular, stroke, heart failure and much more, thereby helping the person to prevent it and save their life. Heart-related diseases are responsible for nearly a third of all deaths worldwide and disproportionately affects lower socioeconomic groups. Hence, in my opinion, this is a very important and an interesting topic to work upon.

The dataset I will be working on is from UCI's Machine Learning repository. This dataset consists of thirteen features representing different parameters which will be used to predict if the person is likely to have a heart related disease or not and a target column titled 'target' which informs us whether the person has a heart disease or not. The target variable only holds values zero and one, where zero indicates no presence of heart disease and one indicates that the person has a heart disease. This dataset has three hundred and three datapoints. It is a classification-based problem where we can classify a given person into one of the two options, either has a disease or doesn't have one. The fourteen columns can be classified into two types, continuous columns and categorical columns.

Following are categorical columns and the values they can take:
sex - sex (1 = male; 0 = female)
cp - chest pain; 1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
fbs - fasting blood sugar > 120 mg/dl: 1 = true; 0 = false
restecg - resting electrocardiographic results; 0: normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
exang - exercise induced angina; 1 = yes; 0 = no
slope - the slope of the peak exercise ST segment; 1: upsloping 2: flat 3: downsloping
ca - number of major vessels (0-3) colored by flourosopy
thal - 1 = normal; 2 = fixed defect; 3 = reversable defect

Following are categorical columns and the values they can take:
age - age in years
trestbps - resting blood pressure (in mm Hg on admission to the hospital)

chol - serum cholestoral in mg/dl
thalach - maximum heart rate achieved
oldpeak - ST depression induced by exercise relative to rest

Literature Survey:

1. Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization - Muhammad Saqib Nawaz, Bilal Shoaib, and Muhammad Adeel Ashraf [1]
https://doi.org/10.1016/j.heliyon.2021.e06948

This paper talks about the simulation results of Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization, NB, SVM, K-NN, RF, and ANN classifiers and concluded that Intelligent CVD Prediction Empowered with GDO achieved maximum accuracy (98.54%) which they could achieve through optimization algorithms.

2. Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction - Emrana Kabir Hashi and Md. Shahid Uz Zaman [2]
https://doi.org/10.33736/jaspe.2639.2020

This paper talks about the use of data mining algorithms like NB, DT, and Random-Forest on the same dataset. These performance techniques showed that the Random Forest Algorithm produced the highest accuracy for heart disease prediction, which is 90.16%

Section 2. EXPLORATORY DATA ANALYSIS

This section contains some of the graphs that were created during exploratory data analysis.

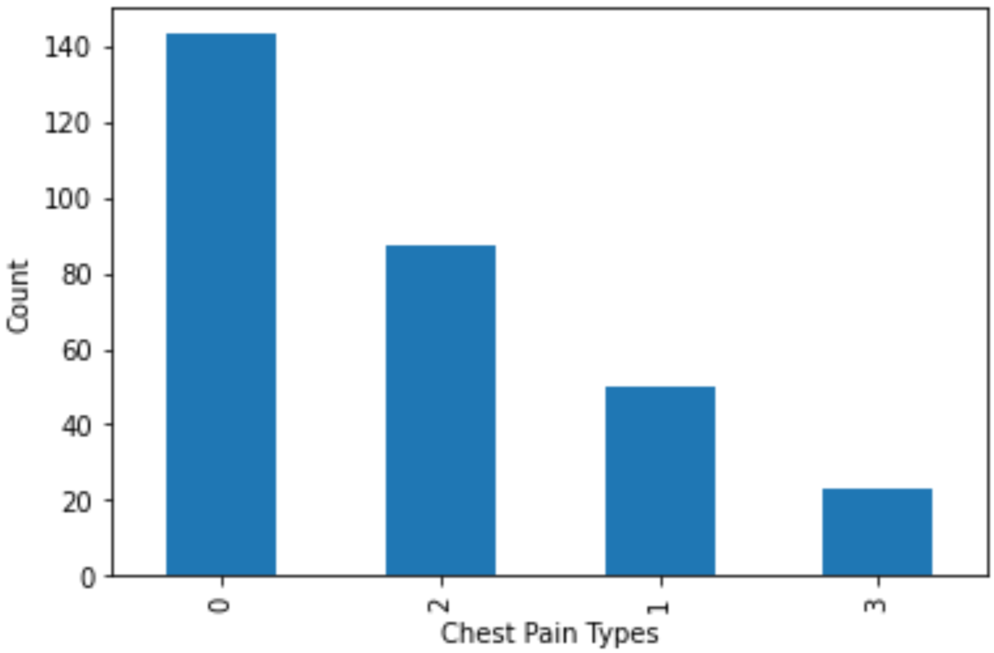


Figure 1: The above bar graph shows the total count for each of the four types of chest pains. As mentioned earlier, there are four types of chest pains: 0: no pain 1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic. As we can see, most people do not have any chest pain.

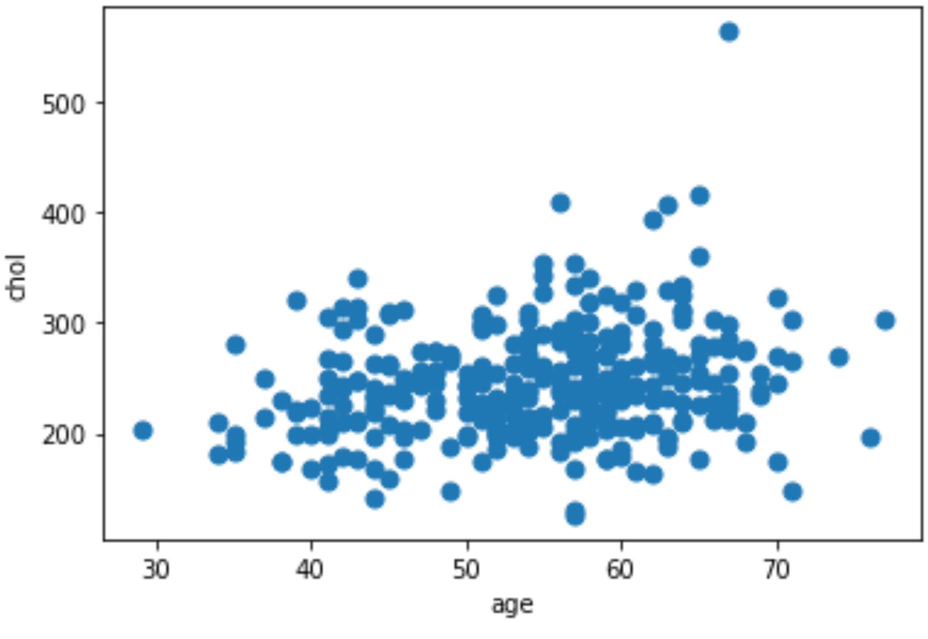


Figure 2: This scatter plot shows the relation between the age and cholesterol levels. Both the columns are continuous in this case. If we were to draw conclusions from this scatterplot, we could conclude that with increasing age, there is an overall slight increase in cholesterol levels.

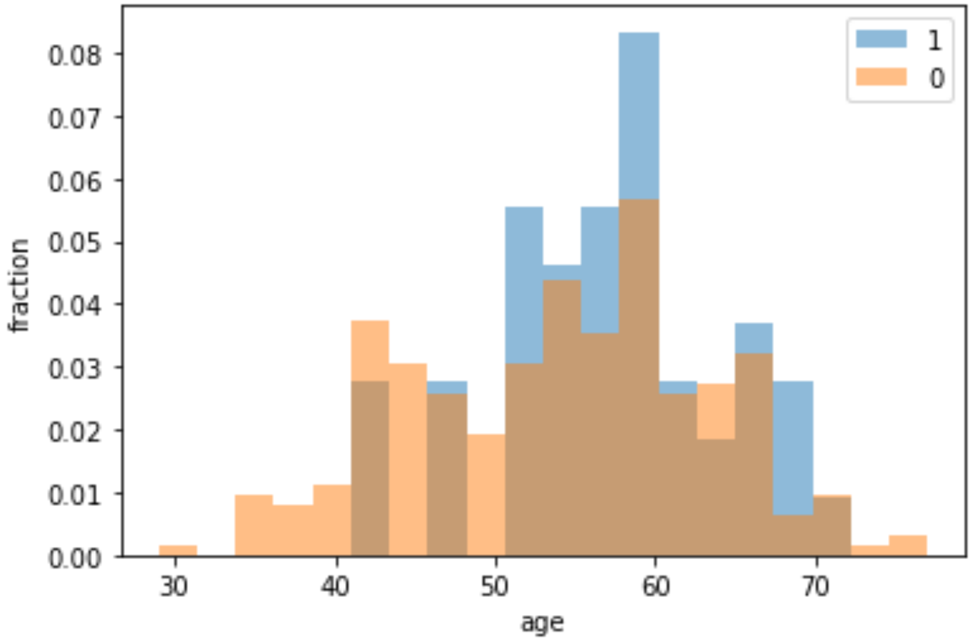


Figure 3: This is category specific histogram graph which plots the columns age, which is continuous and fasting blood sugar, which is categorical. Orange ones indicate people with fasting blood sugar greater than 120 mg/dl and blue ones indicate people with fasting blood sugar less than 120 mg/dl.

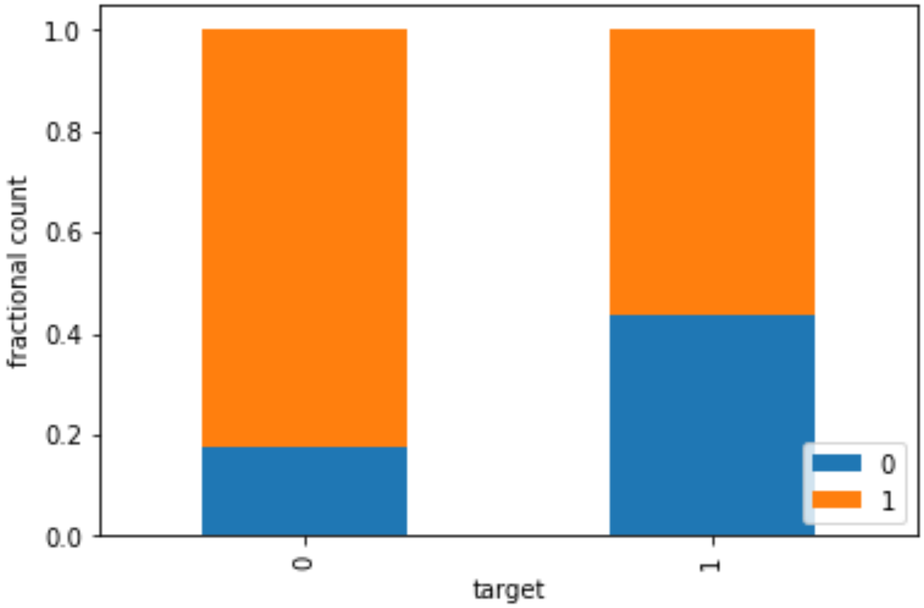


Figure 4: This stacked bar graph plots two categorical columns which are sex and target columns. 1 indicates males and 0 indicates females for sex column. For target column, which is plotted across the x axis, 0 indicates that the person does not have disease whereas 1 indicates that the person has disease.

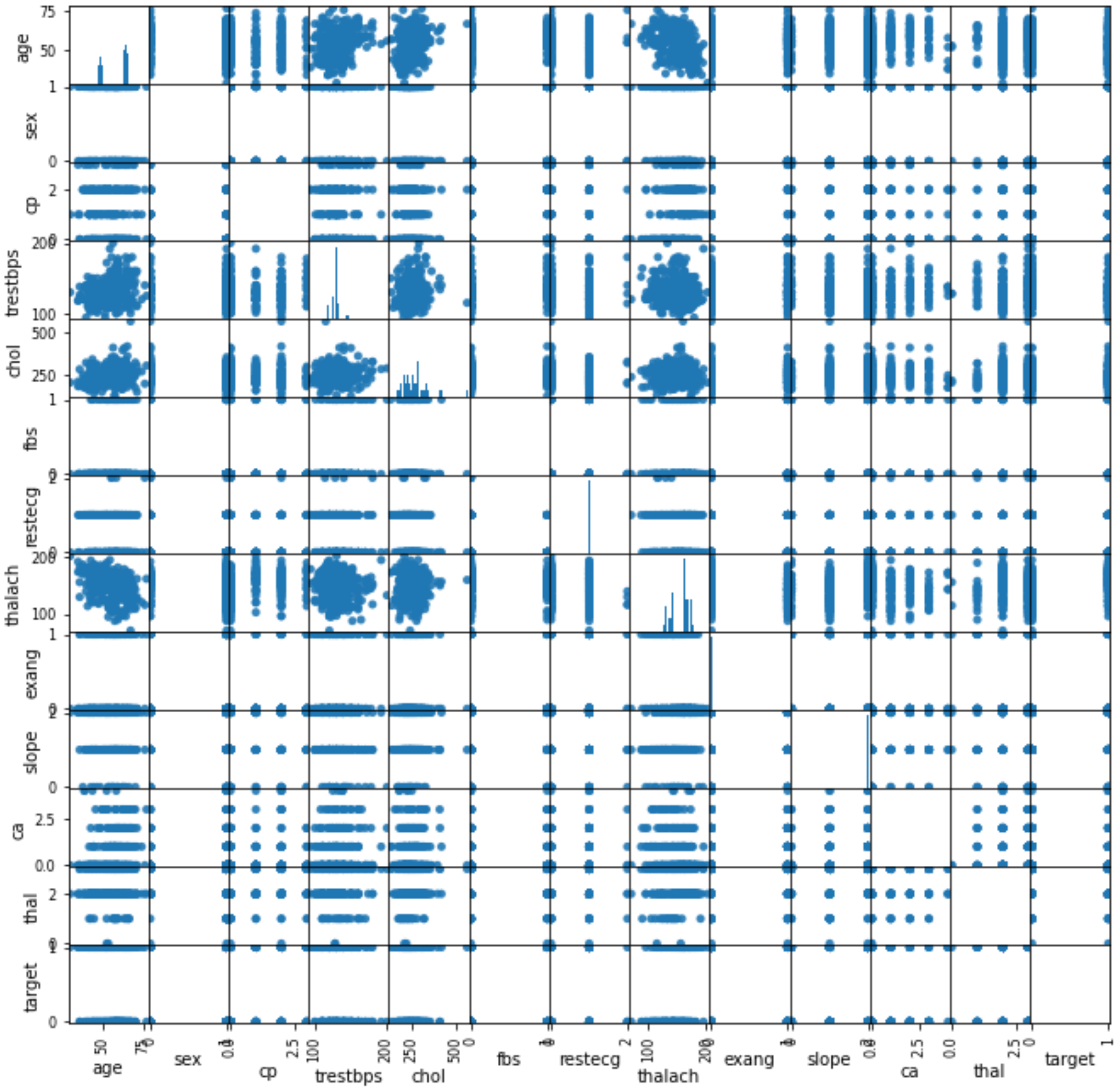


Figure 5: This is a scatter matrix for the given dataset. It compactly plots all the numeric variables we have in a dataset against each other one, and thereby giving us and overall visualization of how each attribute in the dataset stand against each other.

Section 3. METHODS

Section 3.1 DATA SPLITTING & PREPROCESSING

The dataset for this project is independent and identically distributed (IID) dataset. We used the basic train_test_split function for splitting the dataset where 20% of the dataset was allocated to testing. We used stratifiedKFold for cross validation and the reason for choosing this is that stratification ensure that same percentage of feature of interest is given to both training and testing datasets while utilizing the power of kfolds. For this project, we have set the value of n_folds to three. The stratifiedkfold takes care of the remaining parts of the dataset. This dataset does not have any null values. It also does not follow group structure and is not a time-series data. We applied OrdinalEncoder on categorical features since it is a medical dataset, the values should be ordered in my opion and MinMaxEncoder on continuous features since continuous feature values for all the columns are reasonably bounded. List of categorical and continuous columns are given in Section 1.

Section 3.2 MODEL SELECTION

We worked with four different classification algorithms for this project. We used GridSearchCV to determine the optimal parameter values for any given model and constructed a machine learning pipeline using the same. The four models we used are Support Vector Classifier, Random Forest, K nearest neighbors and Decision Tree. We passed and tailored many parameters for every algorithm to allow GridSearchCV to identify the best parameters for a given model. The function we created accepts six arguments, the feature matrix, target variable, random state value, number of folds, the parameter grid for the algorithm and the name of the algorithm itself. The metric used to compare different models is accuracy.

The models were trained for three different random states and the best random state with the corresponding parameters were drawn for each model and compared with other based on accuracy score. As we played around with the parameters and random state values, there we many cases of underfitting and overfitting observed.

MODEL	**PARAMETERS**

SVC	param_grid = { 'algo__C': [0.001, 0.01, 1, 10, 100],
	'algo__gamma': [0.001, 0.01, 0.1, 1, 10, 100, 1000],
	'algo__probability':[True],
	'algo__kernel':['linear']}

Random Forest	param_grid = { 'algo__max_depth': [1, 10, 15, 30, 100],
	'algo__max_features': [0.5,0.75,1.0],
	'algo__n_estimators': [10,50,100, 200, 300, 1000]}

K-Nearest Neighbors	param_grid = {'algo__n_neighbors':[1,2],
	'algo__leaf_size': [1,2,3]}

Decision Tree	param_grid = {"algo__criterion": ["gini", "entropy"],
	"algo__min_samples_split": [2,4,6,8],
	"algo__max_depth": [1, 5, 10, 15]}

Figure 6: This is table summarizing all the parameters used for each of the algorithms.

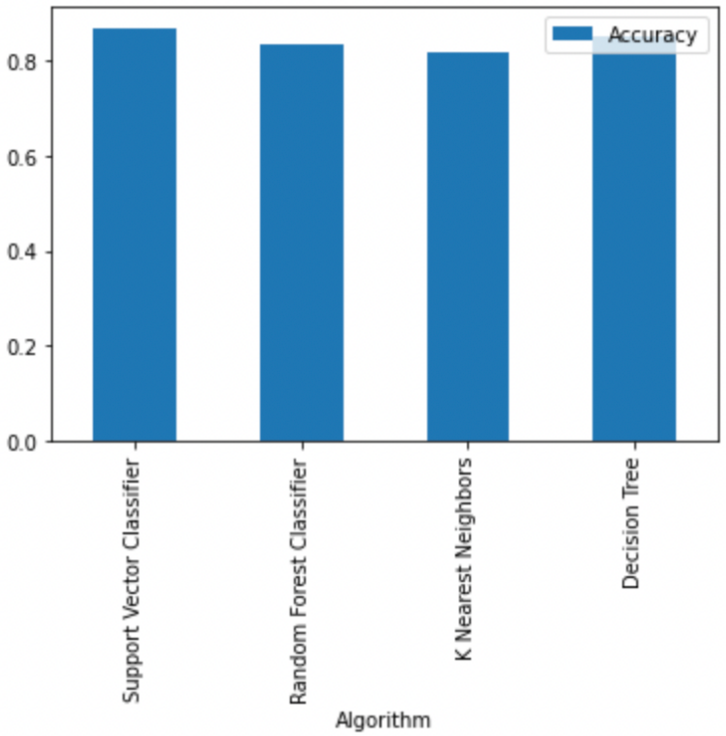


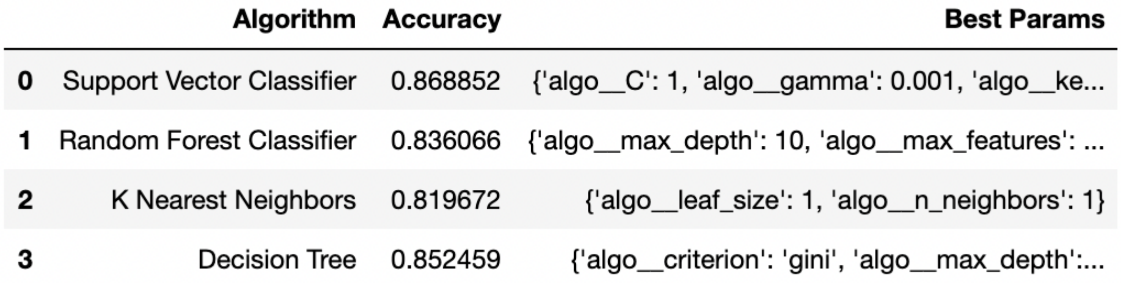
Figure 7: This figure is a histogram which shows the accuracy achieved by the four models we are working with for the best combination of random state value and parameter values. Although all the models come very close, Support vector classifier seems to be the winner.

From the results, we can conclude that Support Vector Classifier does the best job, Decision Tree coming in second. GridSearchCV played with all the possible combinations of the parameters to identify which combination works the best. We then fit the other dataset we obtained after splitting and pass the Kfold splitter as an argument in GridSearchCV. We go ahead with Support Vector Classifier and retrain the model using the best parameters and get an accuracy of 86.88%. The best parameter and corresponding random state values were stored for the best versions of all the algorithms.

Section 4. RESULTS

Section 4.1 EVALUATION OF MODELS

The accuracy score and the baseline score were computed for all the models. As mentioned above, Support Vector Classifier emerges victorious out of the four with the best accuracy being 86.88% with standard deviation of 0.02 and baseline accuracy of 47.5% with standard deviation of 0.007. Of all the models, K nearest neighbors has performed the worse with an accuracy of 81.9%.



	Algorithm	Accuracy	Best Params
0	Support Vector Classifier	0.868852	{'algo__C': 1, 'algo__gamma': 0.001, 'algo__ke...
1	Random Forest Classifier	0.836066	{'algo__max_depth': 10, 'algo__max_features': ...
2	K Nearest Neighbors	0.819672	{'algo__leaf_size': 1, 'algo__n_neighbors': 1}
3	Decision Tree	0.852459	{'algo__criterion': 'gini', 'algo__max_depth':...

Figure 7: This table shows the accuracy achieved by the four algorithms and their corresponding best parameters.

Section 4.2 INTERPREATION OF FINDINGS

Global feature importance and local feature importance were computed for the best model using the permutation technique, coefficient technique and SHAP. Following are the results after implementing the above discussed techniques:

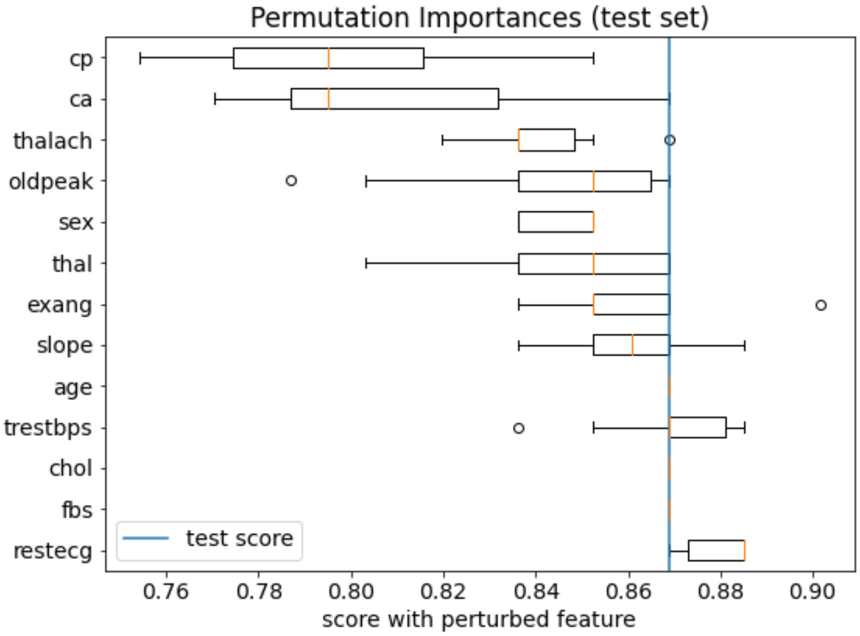


Figure 10: This figure shows permutation feature importance on the test set. From this figure, we can see that restecg and trestbps are probably the most important features while cp seems to be the least important one.

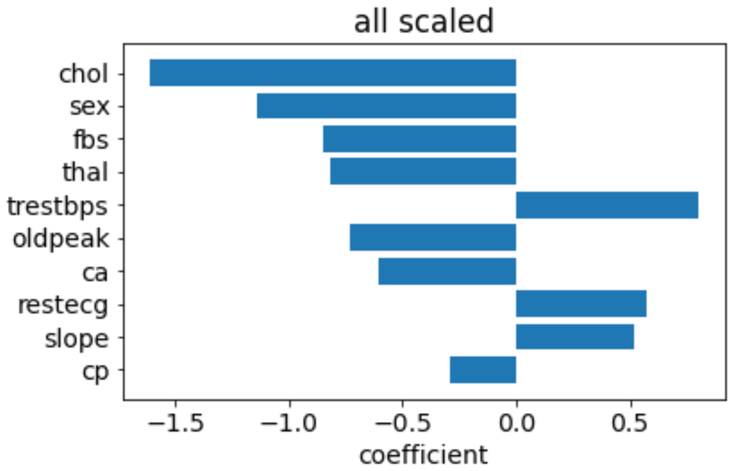


Figure 9: This figure shows the absolute coefficient values for ten of the total features. Chol seems to be the most important feature while cp seems to be the least important one.

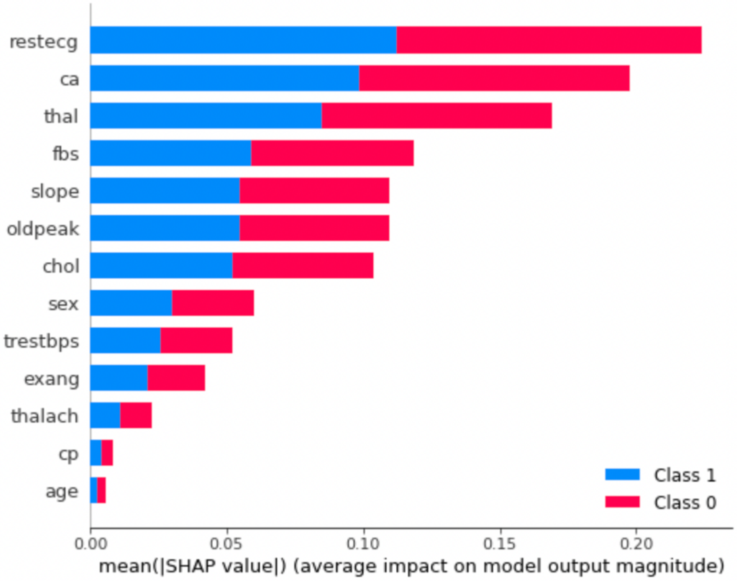


Figure 10: This figure is the mean shap plot. From the above results, we can conclude that restecg seems to be the most important feature, followed by ca and thal in second and third positions respectively. Age and cp, the ones at the bottom, seem to be the least important features.

Section 5. OUTLOOK

The topic for this project is of great importance to the society. There are many ways I can think of which will make this project much better and yield better performance. The most important enhancement to this project would be to get more data. The dataset we are working with here is a very small dataset, we just have three hundred and three datapoints, which is probably not enough. More the datapoints, better the machine learning models and more confidence we will have with our findings. Second improvement that can be done is to experiment with wider range of parameter values and possibly implement and apply more machine learning algorithms which may produce better results than the existing ones. Also, we can try other evaluation metric instead of accuracy. AUC(Area Under Curve) is probably the most popular metric when it comes to working with binary classification which also could be used in this project. There are various other metrics which can be considered for this project and may give us better insights.

Section 6. REFERNCES

1. Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization - Muhammad Saqib Nawaz, Bilal Shoaib, and Muhammad Adeel Ashraf
https://doi.org/10.1016/j.heliyon.2021.e06948
2. Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction - Emrana Kabir Hashi and Md. Shahid Uz Zaman
https://doi.org/10.33736/jaspe.2639.2020
3. Heart Disease Detection Using Machine Learning - Chithambaram T, Logesh Kannan N and Gowsalya M
https://doi.org/10.21203/rs.3.rs-97004/v1
4. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis – Ankur Gupta, Rahul Kumar, Harkirat Singh Arora and Balasubramanian Raman
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8943993
5. Prediction of Heart Diseases using Random Forest - Madhumita Pal and Smita Parija
https://iopscience.iop.org/article/10.1088/1742-6596/1817/1/012009/pdf

Section 7. GITHUB

GitHub Link: https://github.com/yashbafna23/DATA1030_HeartDiseasePrediction