Pattern Recognition - End sem Project

Heart Disease Classification

Team:
S20180020225 - Nanubala Gnana Sai
S20180020227 - Palaash Agarwal
S20180020235 - Pranjal Sahu
S20170010106 - Parth Patwa

I. Introduction

Heart diseases are on the rise and can be fatal. It is very important to detect heart diseases at an early stage. Machine learning can help in early diagnosis. Early detection will save a lot of time, money and most importantly it can save lives. In this project we explore how useful machine learning is to predict heart disease. If we are predicting something as important as a heart disease, it is crucial that the model should be explainable. We also attempt to explain the learnings of our model.

II. Problem description and dataset:

The formal description of the problem is that given a feature vector, predict whether the person has heart disease or not (yes/no).

DataSet - The dataset has the following attributes:

- **age**: The person's age in years
- **sex**: The person's sex (1 = male, 0 = female)
- **cp:** The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- **trestbps:** The person's resting blood pressure (mm Hg on admission to the hospital)
- **chol:** The person's cholesterol measurement in mg/dl
- **fbs:** The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- **restecg:** Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- **thalach:** The person's maximum heart rate achieved

- **exang:** Exercise induced angina (1 = yes; 0 = no)
- **oldpeak:** ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.)
- **slope:** the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
- **ca:** The number of major vessels (0-3)
- **thal:** A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)
- **target:** Heart disease (0 = no, 1 = yes)

III. Data Analysis:

From figure 1 we can see that the data is slightly skewed towards yes.

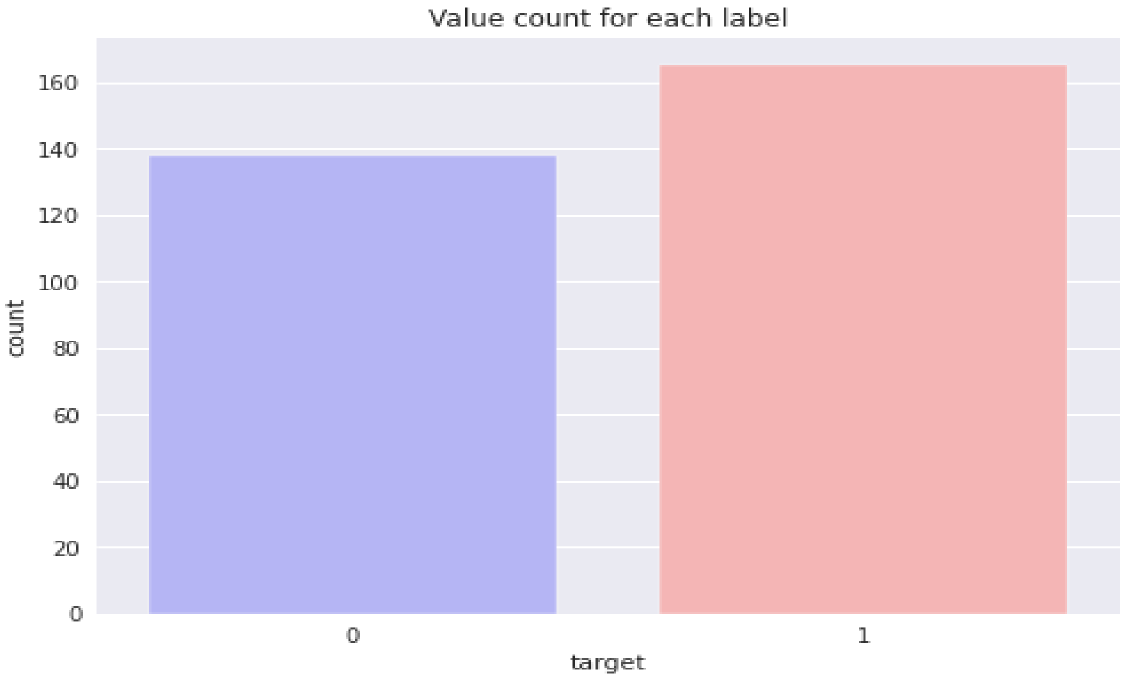

Figure 1: Plot of the count of data in each class.

From figure 2, we can see that There are more females than males in the data.

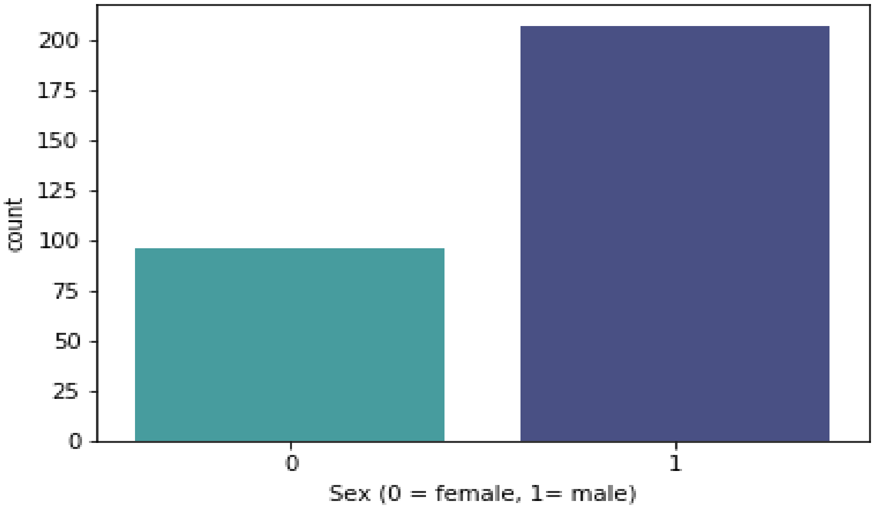

Figure 2: gender distribution of the data.

Figures 3 and 4 give heart disease frequency for age and gender respectively. The ratio of heart disease generally increases with age and is more in females.

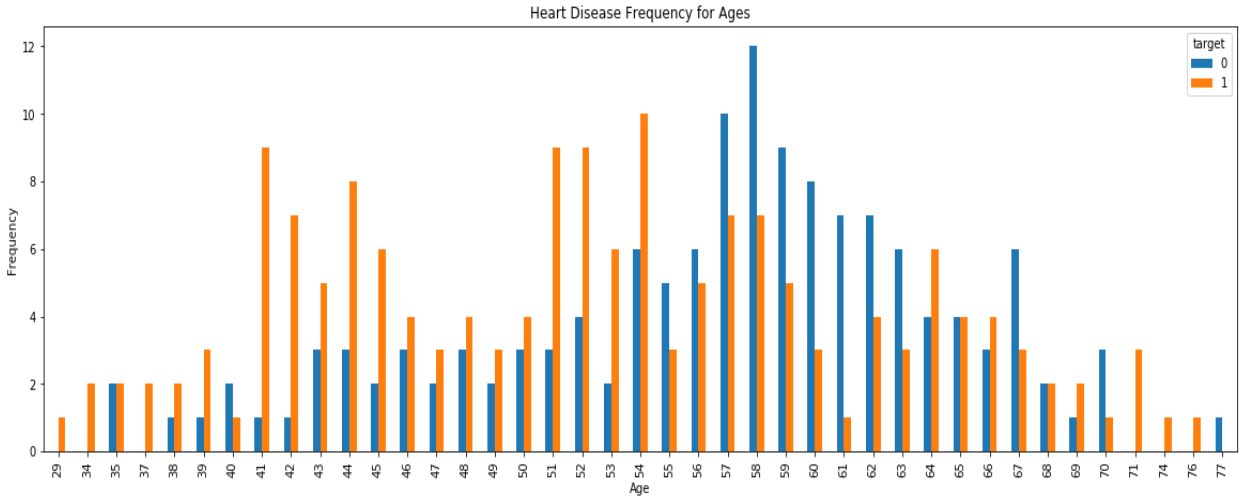

Figure 3 : disease frequency by ages

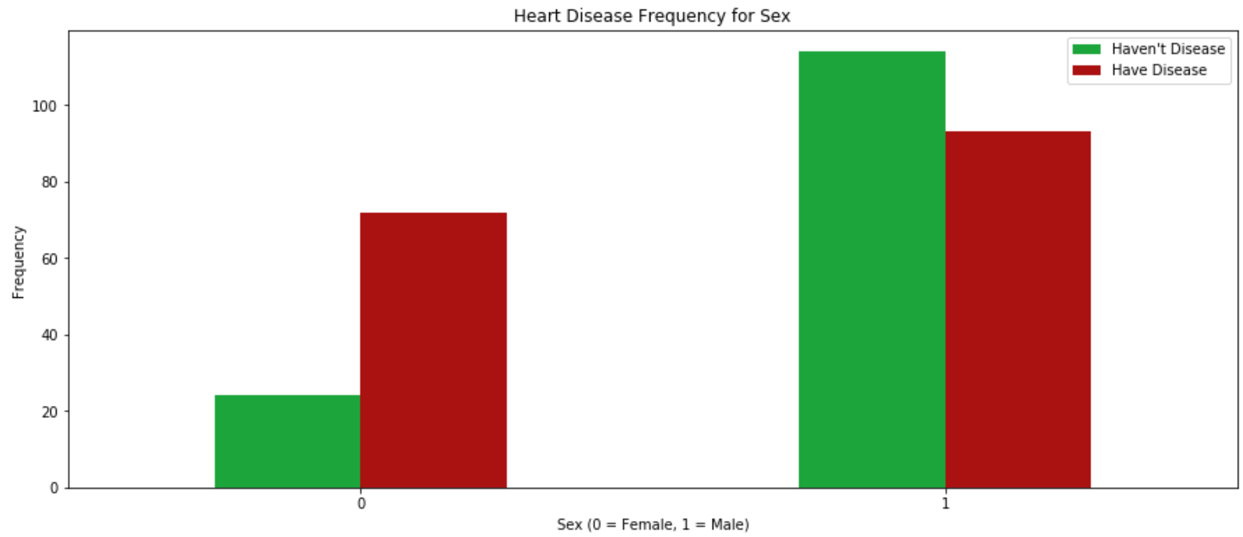

Figure 4 : disease frequency by gender

From figure 5, we can conclude that a high heart rate at a young age leads to heart disease.

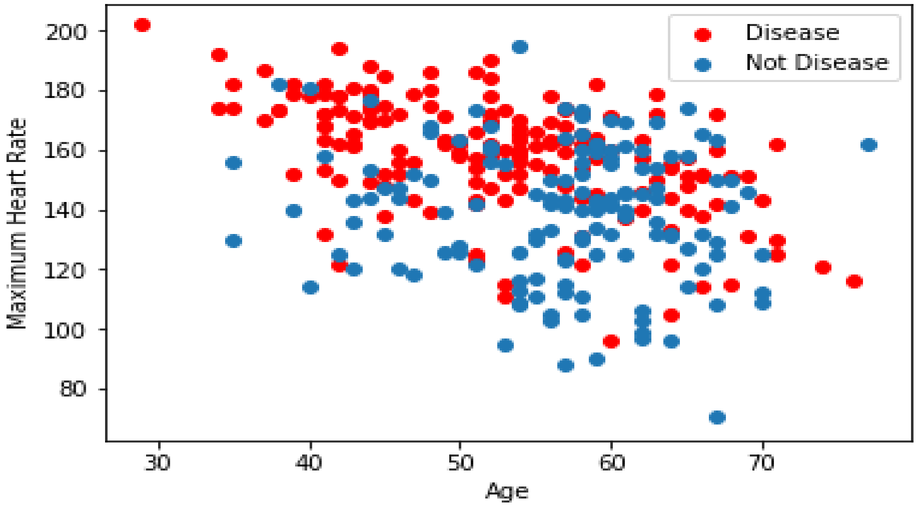

Figure 5 : disease plotted against heart rate and age

IV. Hypothesis

Diagnosis: The diagnosis of heart disease is done on a combination of clinical signs and test results. The types of tests run will be chosen on the basis of what the physician thinks is going on , ranging from electrocardiograms and cardiac computerized tomography (CT) scans, to blood tests and exercise stress tests.

Possible important heart disease risk factors:

High cholesterol, high blood pressure, diabetes, weight, family history and smoking

The major factors that can't be changed:

Increasing age, male gender and heredity.

Major factors that can be modified are:

Smoking, high cholesterol, high blood pressure, physical inactivity, and being overweight and having diabetes.

Other factors: stress, alcohol and poor diet/nutrition. We see no reference to the 'number of major vessels', but given that the definition of heart disease, it seems logical the *more* major vessels is a good thing, and therefore will reduce the probability of heart disease. Given the above, we hypothesize that, if the model has some predictive ability, we'll see these factors standing out as the most important.

V. Methodology:

We begin by forming a hypothesis on what could contribute to a heart disease. This could be backed by medical professionals and/or industry professionals. We've used internet sources for this matter. This is followed by visualizing the dataset to build a basic intuition about the dataset. Next, we employ an array of models which can best fit the data. State of the art models have been used in comparison to naive algorithms. Once the data has been fit, we use model interpretive tools to confirm our initial hypothesis thus concluding the project.

SVM

We begin by applying the standard linear and RBF kernel to the dataset and measure the performance.

Explicit Feature Mapping:

A major problem with kernel based predictors is that calculating the Kernel matrix has computation complexity of O(n^3) where n is the number of training samples. Standard kernel methods **do not scale well** to large datasets. To solve this, we compute the feature map explicitly by approximating them to a lower dimension and computing the solution in primal form.

Nystroem Mapping:

It works by approximating the Kernel matrix(K) to a lower dimension. Random rows/ and columns of K are chosen whose eigenvectors act as the basis. This creates the feature map Φ such that $\overline{K} = \Phi\Phi^T$.

Explicit feature mapping works particularly well when num_features << num_samples. Once we've computed the feature map, we can use Linear SVM to gain accuracy without sacrificing speed.

Random Forest

Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes. Random decision forests, correct for the decision tree's habit of overfitting the training set. This results in improved performance relative to naive trees.

Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Explainability

Once a model has been fit, it'd be beneficial if we could interpret the model. We'd like to know how much a feature contributes in the final prediction, it also serves as a sanity check against fitting to noise.

The following tools help us in that regard:
PDP: A partial dependence (PD) plot depicts the functional relationship between a small number of input variables and predictions. They show how the predictions partially depend on values of the input variables of interest.

Shap: SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction we would make if that feature took some baseline value.

IV. Results:

All experiments are implemented in python, sklearn , pandas.

- **SVM:**

Confusion matrix:

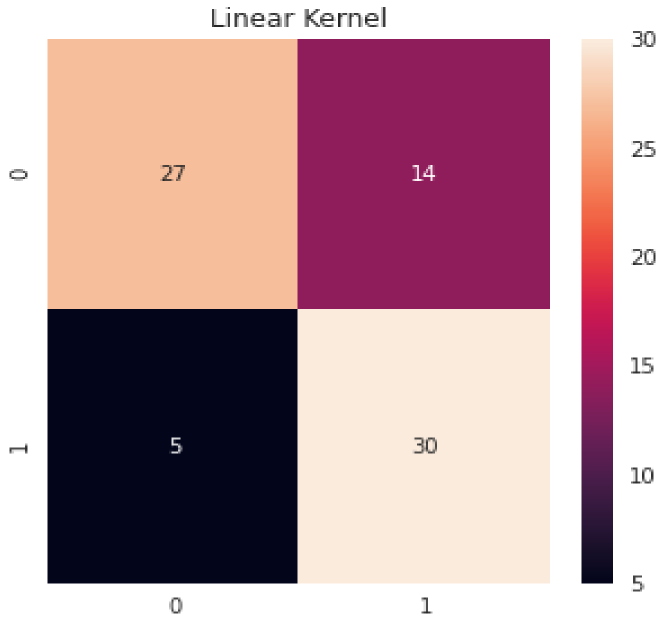

Figure 6: Confusion matrix using linear kernel

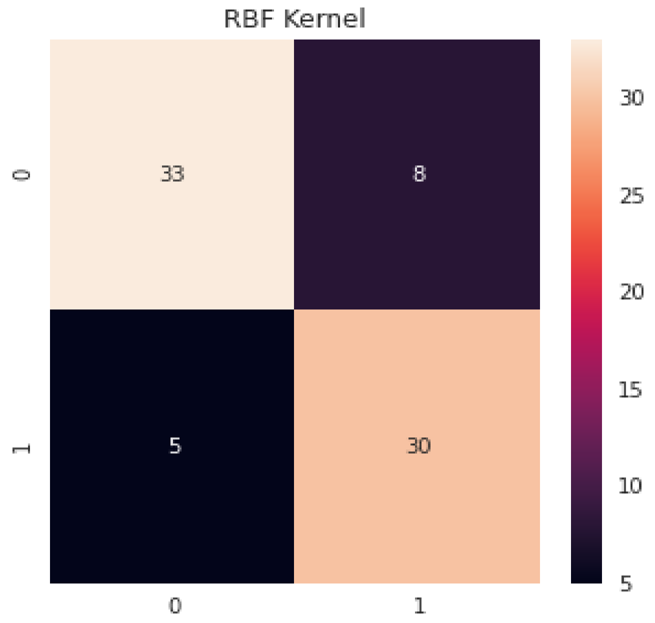

Figure 7: Confusion matrix using RBF kernel

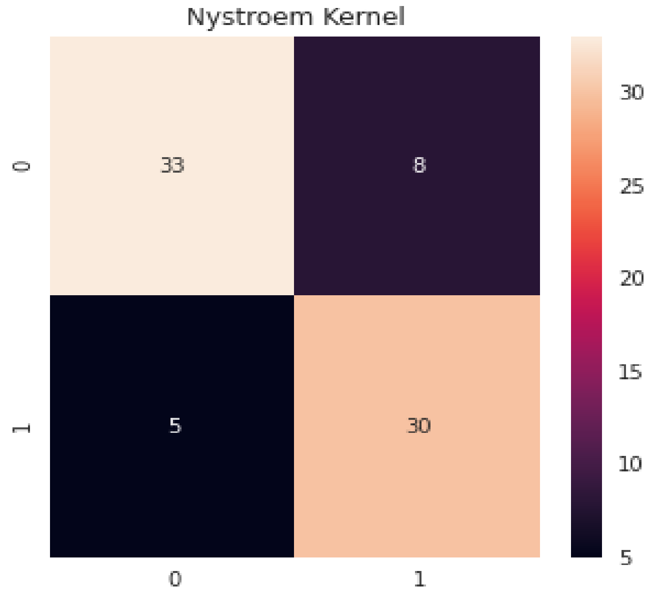

Figure 8: Confusion matrix using Nystroem kernel

Accuracy : We employ 10 fold cross validation strategy on this dataset and check the score for each of the above classifier:

Linear SVM	75%
Kernel SVM	82.89%
Nystroem SVM	82.8%

- **Random Forest Classifier:**

Confusion matrix:

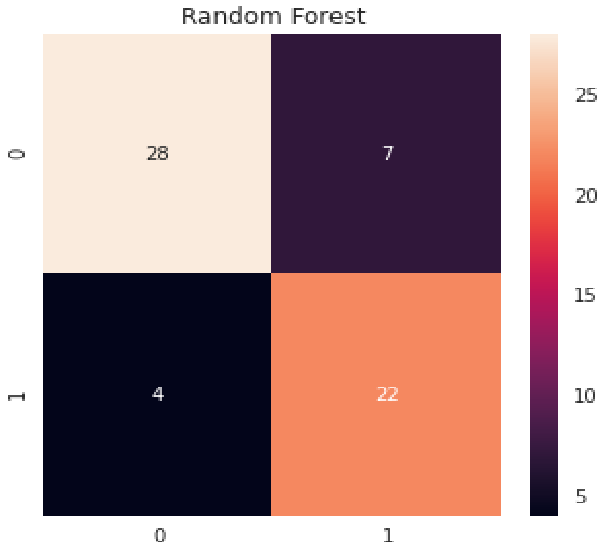

Figure 9: Confusion matrix for random forest classifier

Accuracy: We employed the grid search cross validation with 3 fold cross validation and used the best parameters received for the fitting the model. The accuracy we received was **81.97%**

- **Gradient Boosting:**

  Confusion matrix:

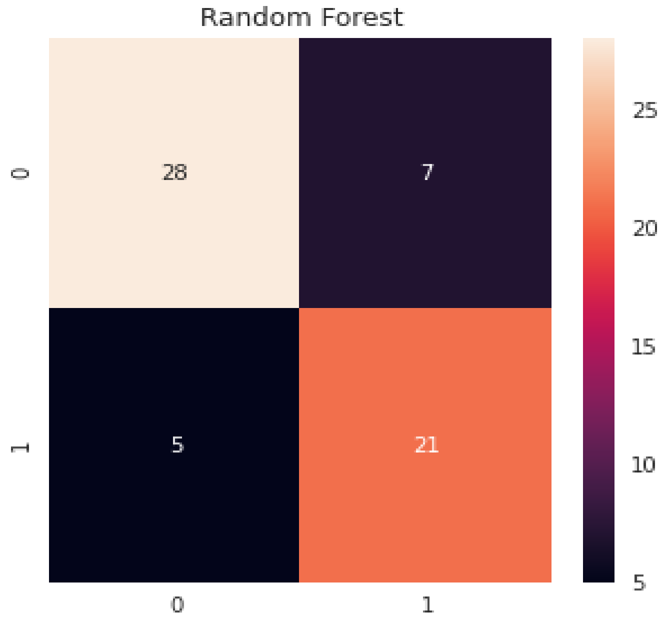

Figure 10: Confusion matrix for gradient boosting

Accuracy: Based on the above confusion matrix, we calculated the accuracy for gradient boosting to be around **78.68%.**

V. Analysis and explanation

Nystroem SVM:

We first transform the dataset to the new feature space using Nystroem method. This is followed by fitting of this dataset using Linear SVM in the primal space.

As expected, the RBF kernel outperforms the linear kernel. In the following figure, we notice a linear trend in the increase of accuracy of the Nystroem method. As more and more samples are selected the approximation gets more precise and hence the result.

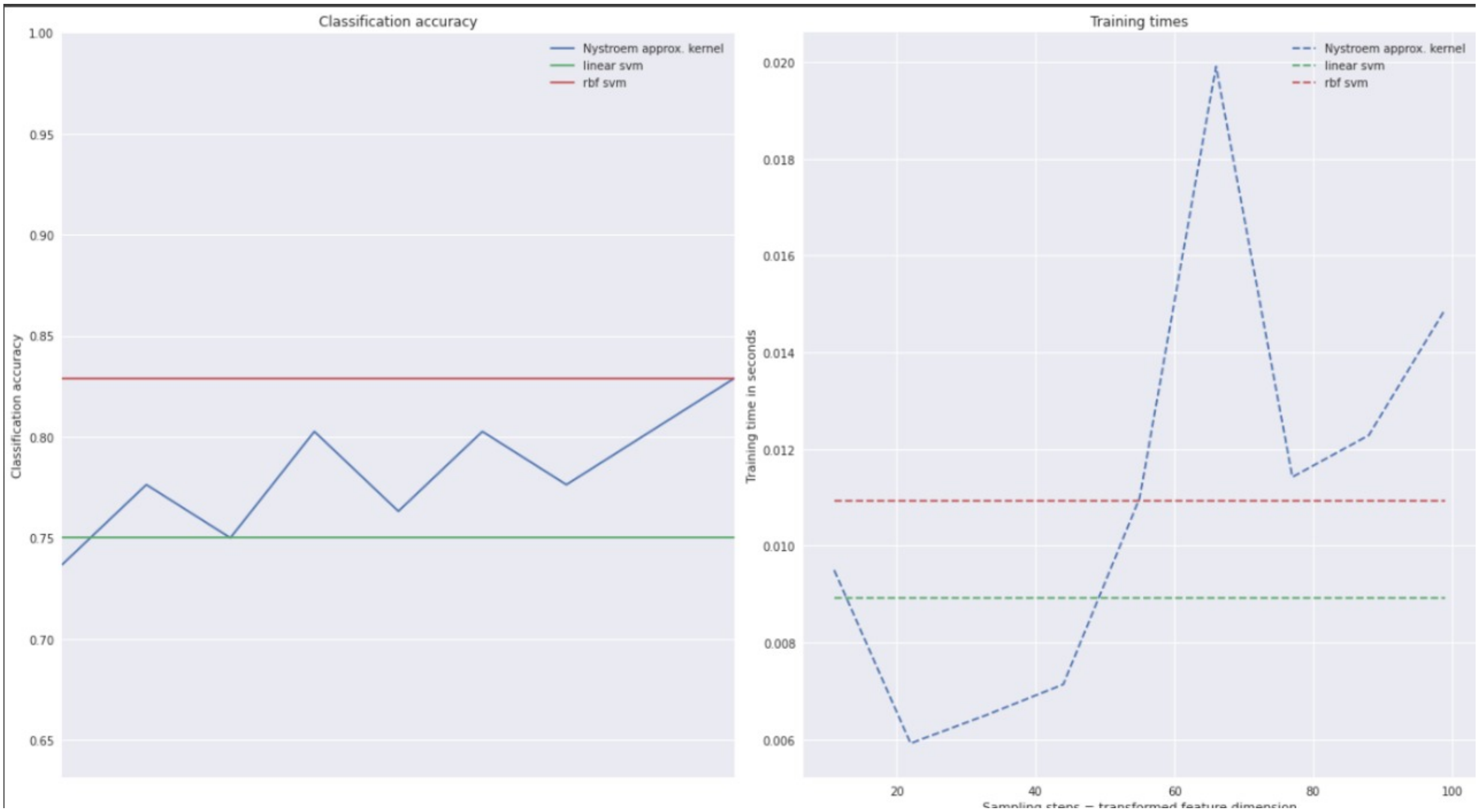

Figure 11: (Left to right) Accuracy of Linear, RBF, Nystroem kernels and their training times respectively. The upper bound is for the RBF while the lower is for linear. Notice The increasing accuracy with increase in sample size.

Decision Tree and Gradient boosting:

The random forest classifier performs better than the gdbt. Linear svm has the worst performance. Random forest is not able to outperform kernel and nystroem SVM.

Explainability:

PDP : We can see that as the number of major blood vessels *increases*, the probability of heart disease *decreases*. It means more blood can get to the heart so the heart is healthier.

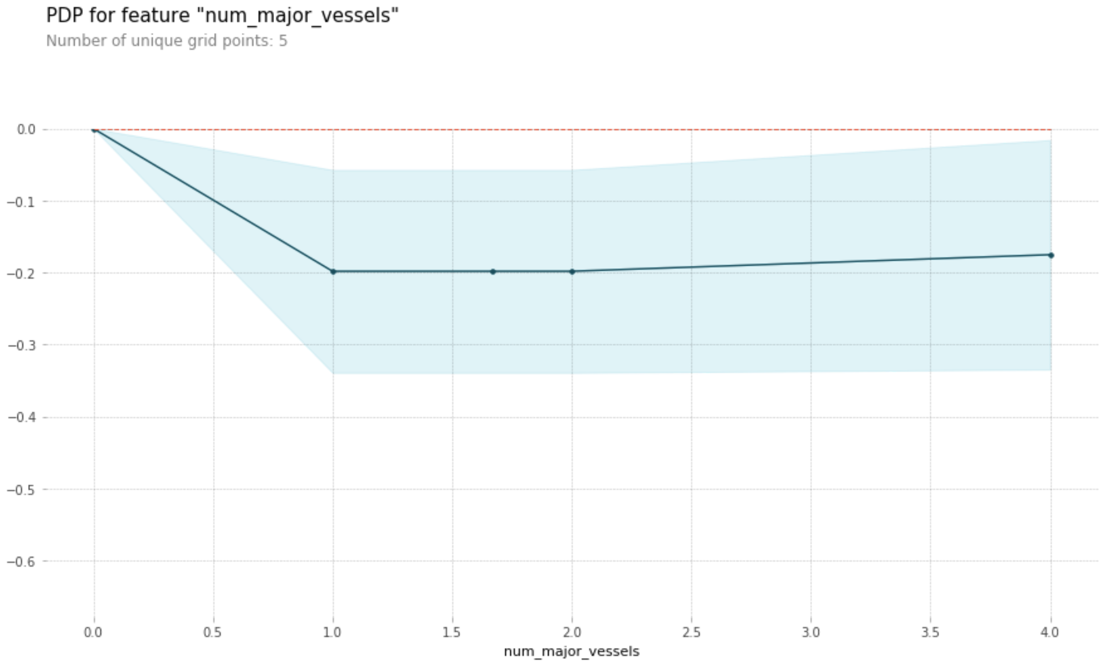

Figure 12: PDF of ‘num_major_vessels’

SHAP: (SHapley Additive exPlanation)

The number of divisions of major vessels is pretty clear, and it says that low values are bad. The thalassemia 'reversible defect' division is very clear. We can see some clear separation in many of the other variables. Exercise induced angina has a clear separation, although not as expected, as 'no' *increases* the probability. Another clear one is the st_slope. It looks like when it's flat, that's a bad sign. It's also odd that the men have a *reduced* chance of heart disease in this model, despite Domain knowledge telling us that men have a greater chance.

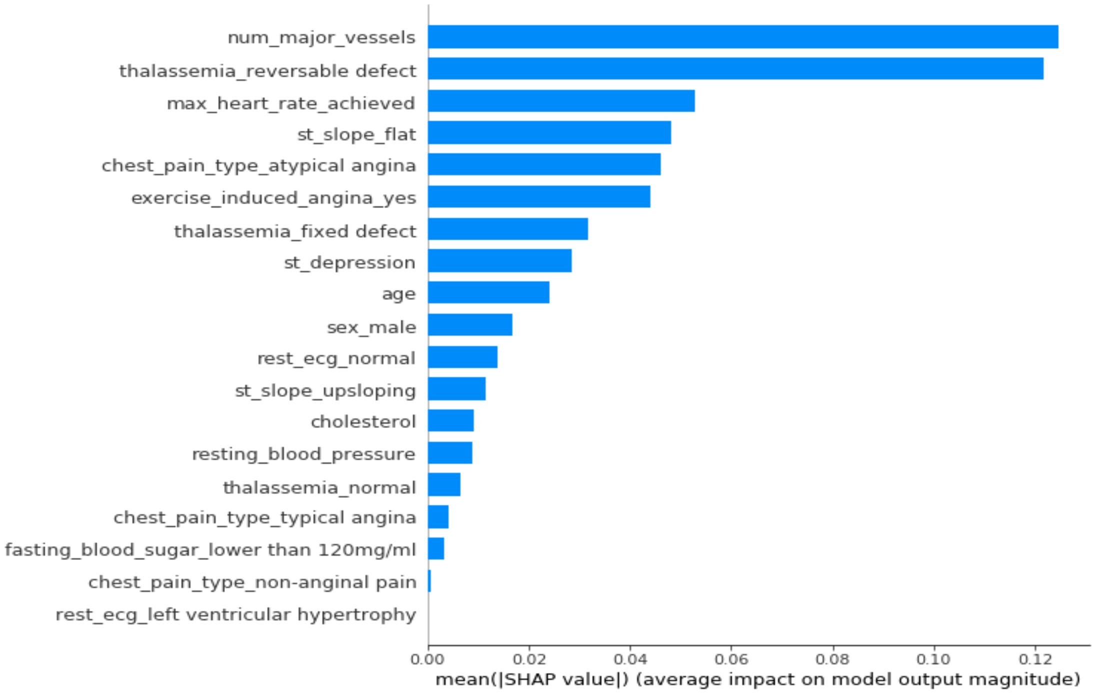

Figure 13: Mean shap values of the features

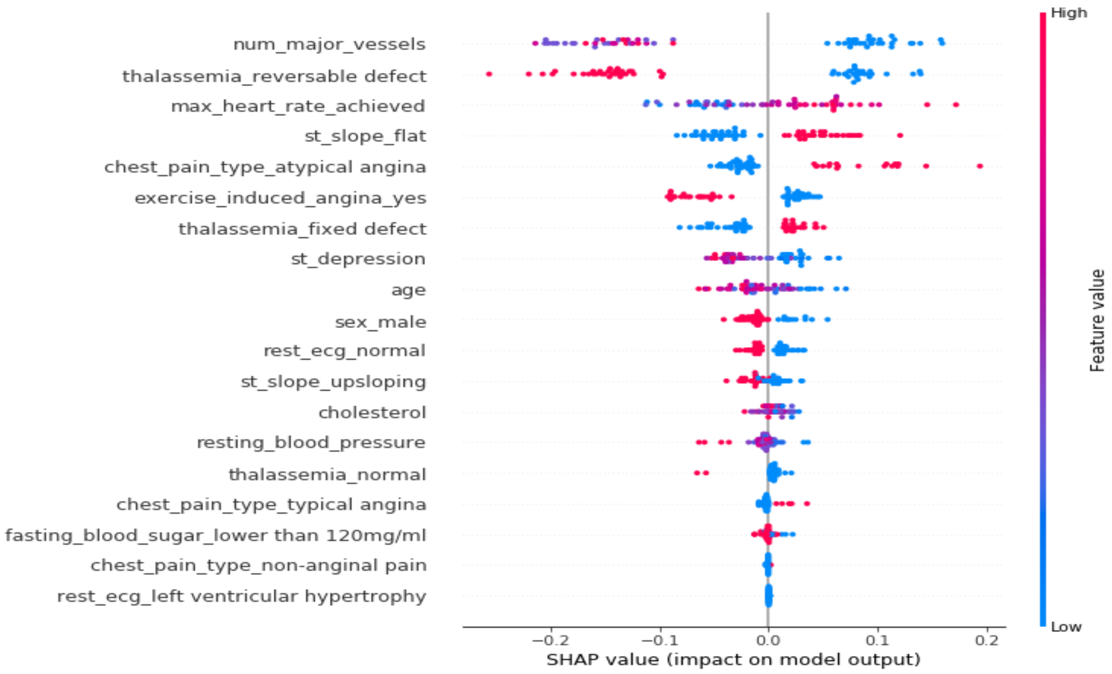

Figure 14: Shap value impact on model

VI. Conclusion and Future work:

In this project we use symptoms and features to predict heart disease. Our best model achieves 83% accuracy. Kernel SVM outperforms all the other algorithms. We see that ‘num major vessels’, ‘thalassemia’ and ‘max heart rate’ are important factors in determining heart disease. We also see that contrary to the domain knowledge, men have less chances of getting heart disease than females.

In the future we would like to explore deep learning and data augmentation. We would also like to see how covid19 affects the chances of heart disease.

VII. Bibliography:

- Williams, C.K.I. and Seeger, M. “Using the Nystroem method to speed up kernel machines”, Advances in neural information processing systems 2001
- T. Yang, Y. Li, M. Mahdavi, R. Jin and Z. Zhou “Nystroem Method vs Random Fourier Features: A Theoretical and Empirical Comparison”, Advances in Neural Information Processing Systems 2012
- Ed Burns. ‘The ST Segment’. Life in the fastlane. 2020.
- Mayo Clinic. ‘Heart disease’ Mayo Clinic. 2020
- Heart Foundation. ‘Medical tests for heart disease’. Heart foundation org. 2020
- BHF. ‘Risk factors for heart and circulatory diseases.’ BHF 2020.
- Heart. ‘Understand Your Risks to Prevent a Heart Attack’ Heart org. 2020
- Caner Dabakoglu. ‘Heart Disease - Classifications (Machine Learning)’ Kaggle. 2019
- Rob Harrand ‘‘What Causes Heart Disease? ’ Kaggle. 2018
- Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. ‘Scikit-learn: Machine Learning in Python’ Journal of Machine Learning Research. 2011
- Mayo clinic. ‘Heart disease symptoms’. Mayo clinic. 2020

VIII. Individual Contribution

- Nanubala Gnana Sai
  - Explicit feature mapping, idea proposal and implementation.
  - Fit and plot Linear and Kernel SVM.
  - Compare performance of Nystroem with standard methods.
  - Brief report of the theory behind it.

- Palaash Agarwal
    - Idea proposal and implementation
    - Random forest and its optimization
    - Hyperparameter tuning
    - Report preparation

- Pranjal Sahu
    - Decision Tree
    - Data Analysis
    - Report
    - GDBT

- Parth Patwa
    - Explainability
    - PDP, SHAP
    - Analysis
    - Conclusion
    - Writing

IX. Code Link: Please visit this link to view the code.