Assignment 12.2

Final Project

Sucharitha Puppala

Data Science, Bellevue University.

DSC550-T301 Data Mining (2231-1)

Professor Brett Werner.

November 10, 2022.

Introduction

Introduce the problem

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you're born with (congenital heart defects), among others.

The term “heart disease” is often used interchangeably with the term “cardiovascular disease”. Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

Justify why it is important/useful to solve this problem

According to a news article, heart disease proves to be the leading cause of death for both women and men. About 610,000 people die of heart disease in the United States every year–that's 1 in every 4 deaths. Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men. Coronary Heart Disease (CHD) is the most common type of heart disease, killing over 370,000 people annually. This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease. With this analysis we can know about the important features that are dominant in a person having Heart Disease.

How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?

Machine learning in healthcare is used to draw insights from large medical data sets to enhance clinicians' decision-making, improve patient outcomes, automate healthcare professionals' daily workflows, accelerate medical research, and enhance operational efficiency. If a person identified with any of the features identified can take preventive measures and help prevent from having heart disease.

Explain where you obtained your data

The Heart Failure Prediction dataset is collected from Kaggle.com and the link to the dataset is given below.

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

This dataset contains 11 features that can be used to predict a possible heart disease.

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]

Organized and detailed summary of Milestones 1-3

Milestone – 1 (EDA; include any visuals you think are important to your project)

The Milestone -1 mainly focused on better understanding of the dataset. The distribution of all the features and the target variables in the dataset, to know the counts of the categorical features of the dataset, identify any duplicated, missing values and outliers in the data set. The following are the points summarizing the dataset.

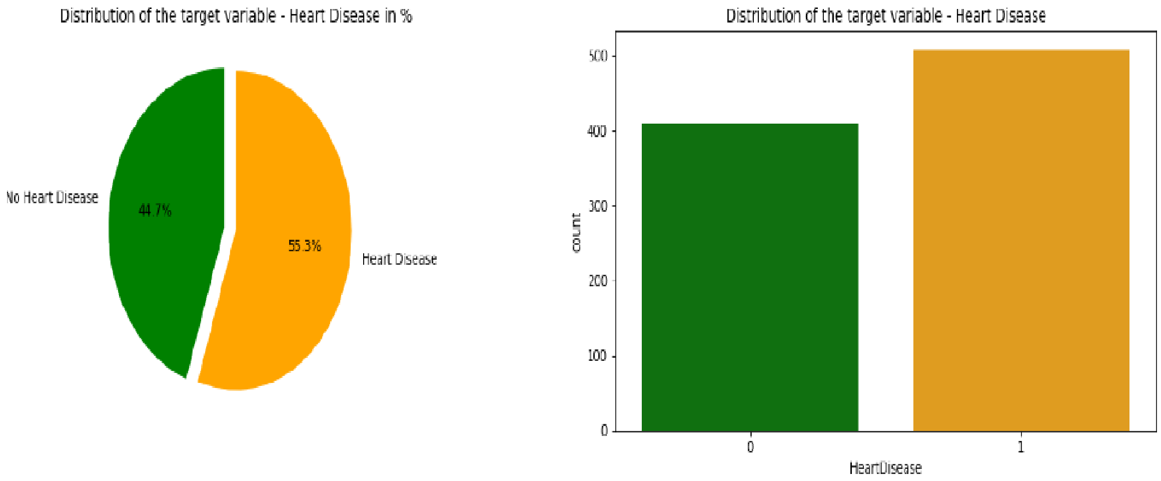
- The dataset has 918 rows and 12 columns.
- The target variable "Heart Disease" is well distributed.
- There are no missing values and duplicated data in the dataset.
- The observations from the plots of patients having positive Heart Disease with different variables in the dataset says that the males are more in number, ChestPainType ASY is more dominant, FastingBS is less, RestingECG is low, ST_Slope is flat.

- The people of age above 50 years, having RestingBP between 95 and 170, having Cholesterol values between 160 and 340 are more prone to Heart Disease.
- We can see some features having a good correlation with the Heart Disease like Age, Sex, but the correlation strength is very low.

Visualizations:

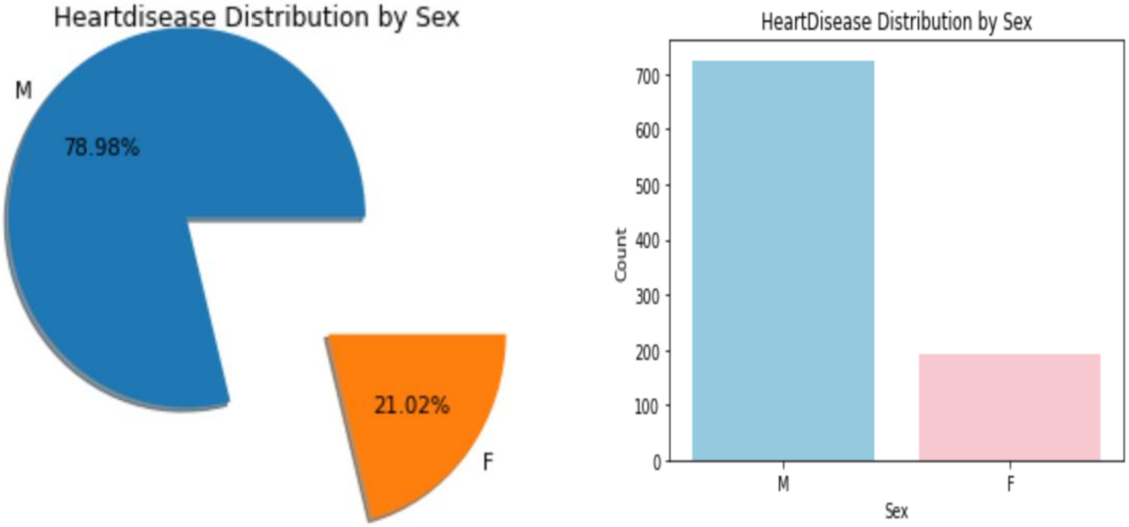
- The following visuals help for better understanding the distribution of the target variable and the distribution of the numerical and categorical features with the target variable from the dataset. I have used count plots, pie charts, histograms, heat map, pair plot and box plots for the visual analysis of the features and the target variables in the dataset. The count plots, histograms help to understand the distribution of each feature and the target variables. Pie charts help understand the percent value of the distribution of each feature.
- From the below visualizations we can see that the dataset is evenly balanced. We can see from the pie plot that 55.3% are having Heart Disease and 44.7% are not having any Heart Disease.

Figure 1: Distribution of the target variable – "Heart Disease"



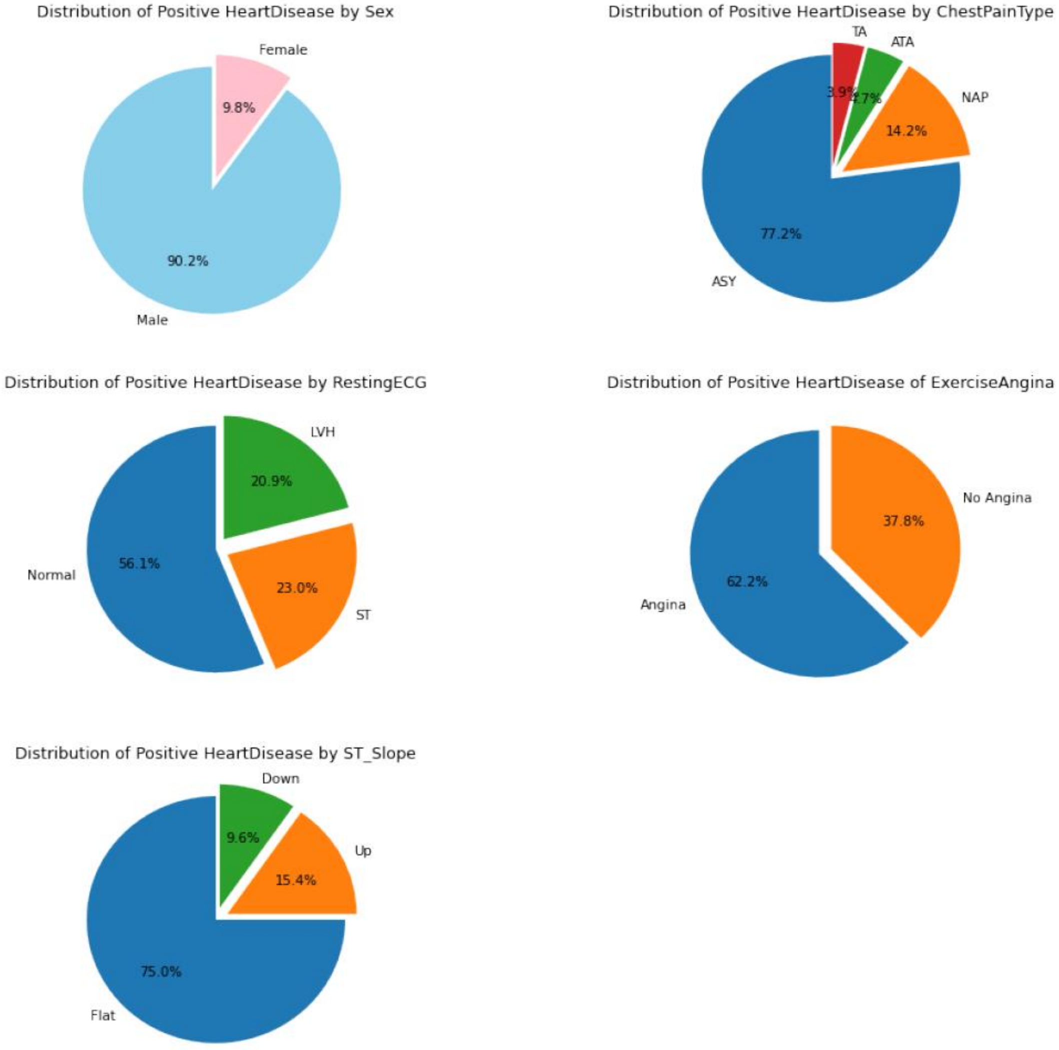
- From the below pie plot we see that about 78.98% of males have heart disease and 21.02% of females have Heart Disease. We can say that Males are approximately 3 times more likely to have Heart Disease than females.

Figure 2: Distribution of "Heart Disease" by sex



- From the pie plot of 'Sex vs. Positive Heart Disease' we see that about 90% of the patients are male, 'ChestPainType vs. Positive Heart Disease' we see that about 77.2% of the patient with Heart Disease are having ASY type of Chest Pain, 'RestingECG vs. Positive Heart Disease' we see that the patients having Heart Disease have RestingECG normal level which is 56.1%,'Exercise Angina' we see that the patients having Heart Disease have the problem of Exercise Induced Angina of about 62.2%, 'ST_Slope vs. Positive Heart Disease' we see that about 75% of patients having Heart Disease have a flat ST_Slope.

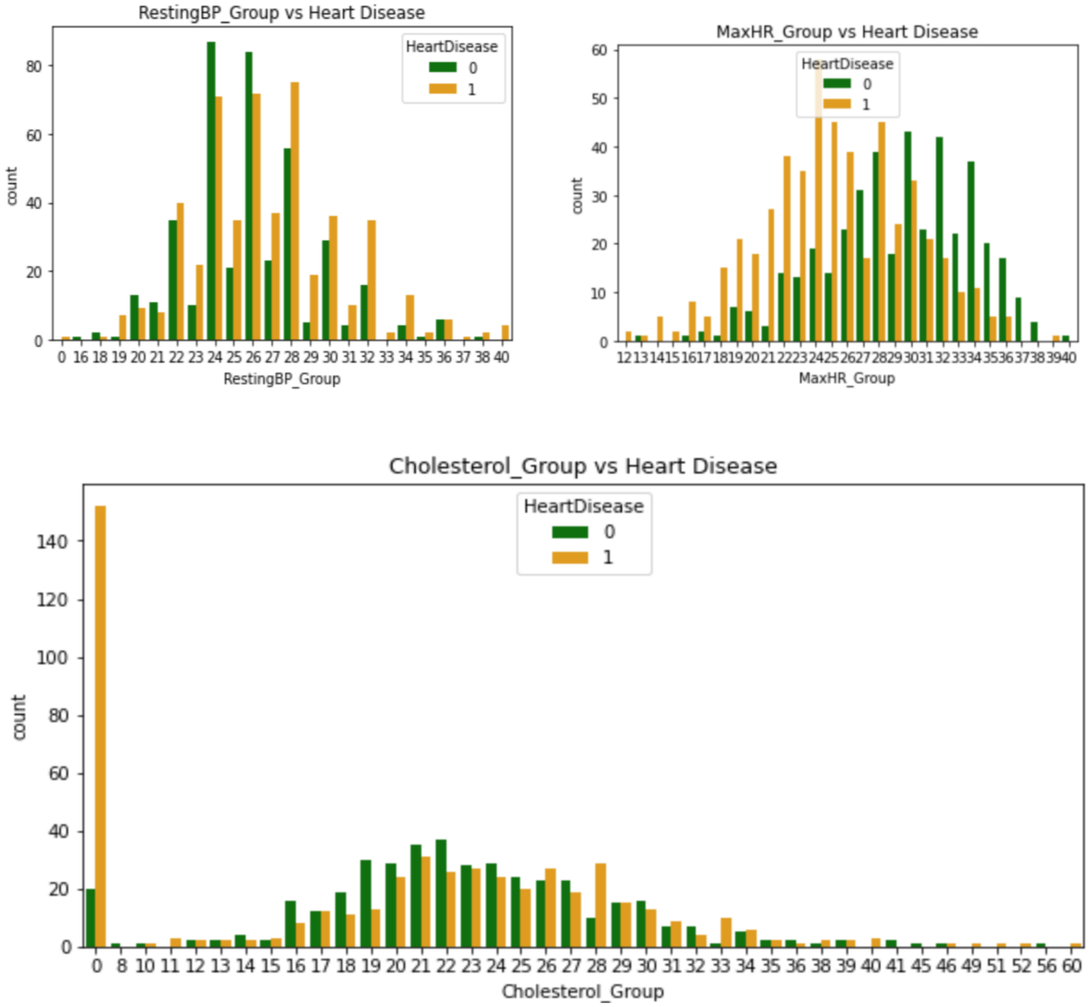
Figure 3: Distribution of the positive Heart Disease in different categorical features

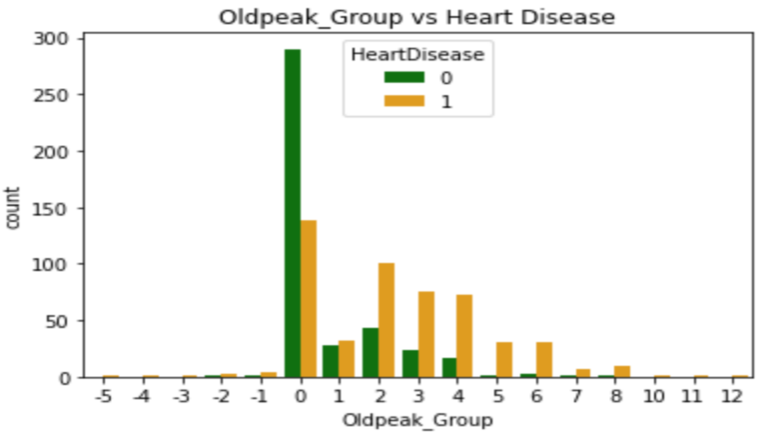


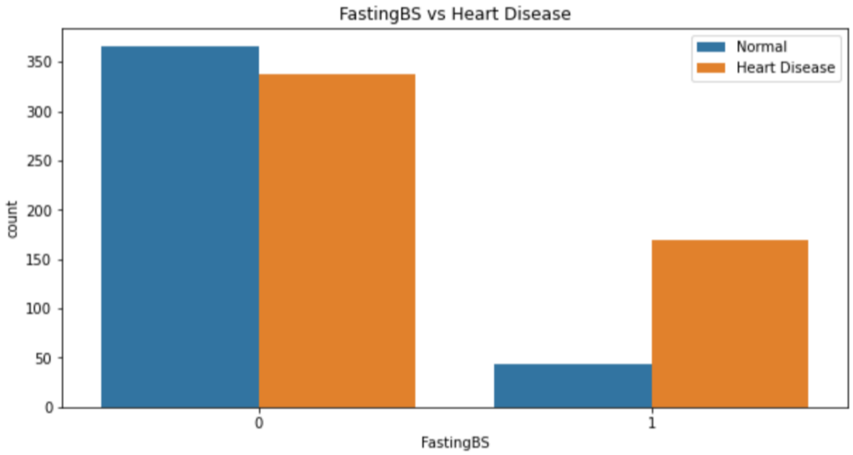
- From the Figure 4, we see the distribution of the numerical variables in the dataset. When observed the plot of RestingBP, we can say that the values 95 to 170 ((19*5)- (34*5))are mostly prone to have Heart Disease. From the graph of scaled Cholesterol, we can say that the values 160 to 340 ((16*10)- (34*10))are prone to have Heart Disease. From the graph of scaled MaxHR, we can say that the values 70 to 180 ((14*5) - (36*5)) are mostly prone to have Heart Disease.

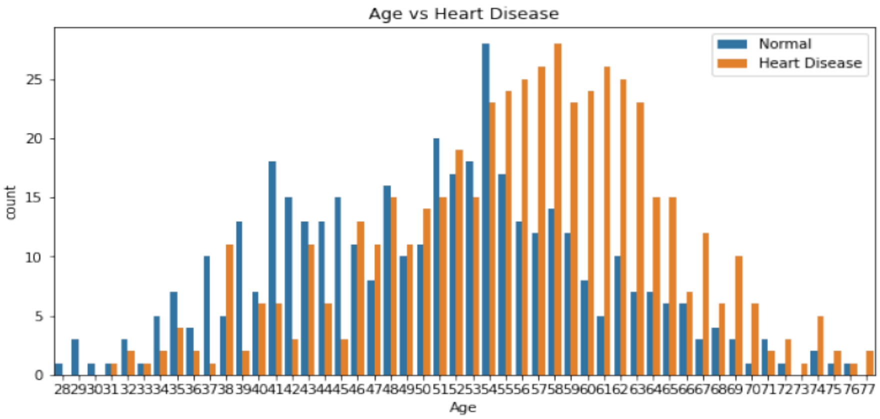
From the graph of scaled Oldpeak, we can say that the values 0 to 8 i.e. 0 to 4((0*5/10) - (8*5/10)) are mostly prone to have Heart Disease. From the count plot of Age vs. Heart Disease most of the heart disease Patients have age between 55 and 65.From the above count plot of Fasting Blood Sugar vs. Heart Disease, we see that the persons having FastingBS and positive Heart Disease counts up to 150.

Figure 4: Distribution of the target variable with the numerical features.



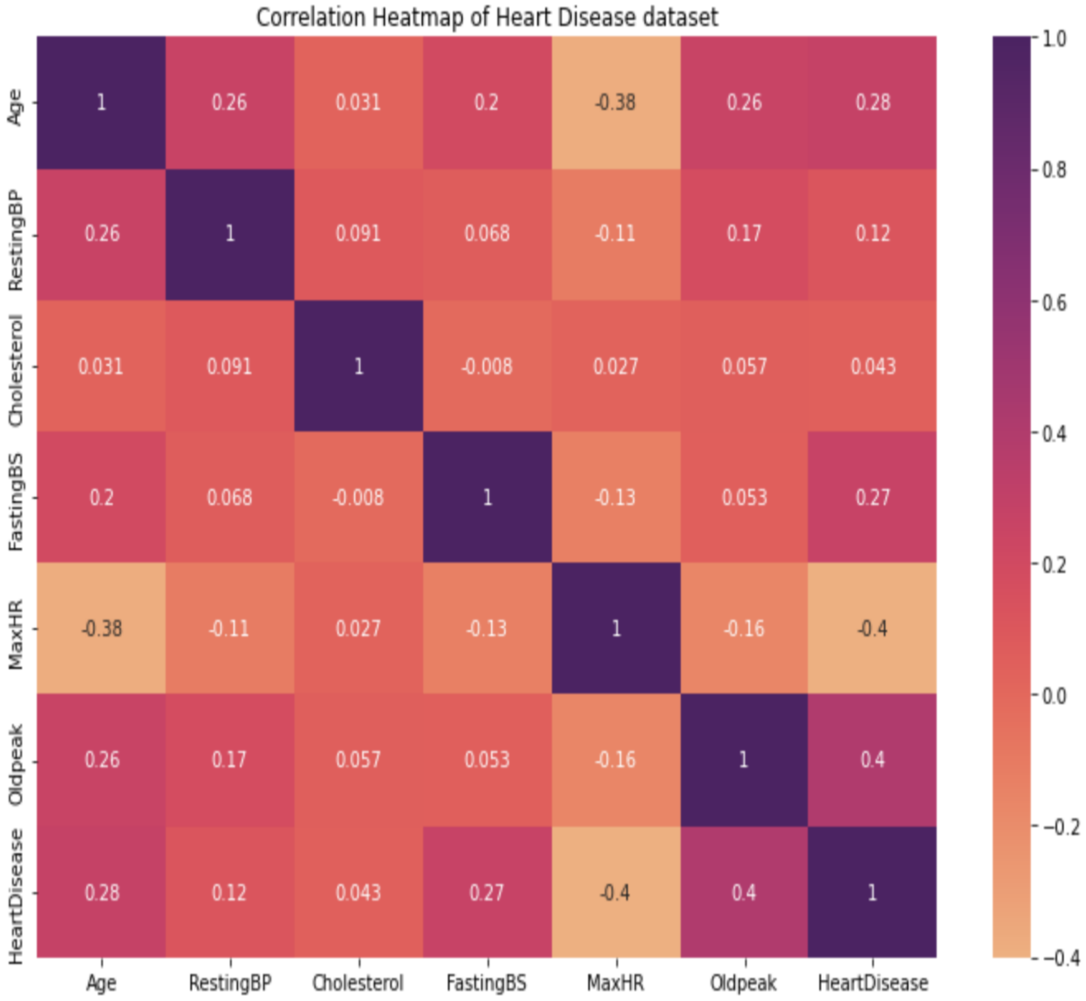






- Heat map helps in understanding the correlation of the numerical features with the target variables. From the below correlation heat map we see some features are positively correlated with Heart Disease and some features are negatively correlated, but the correlation seems to be very low. As per the correlation heat map we can say that Age, RestingBP are having good correlation with the Heart Disease.

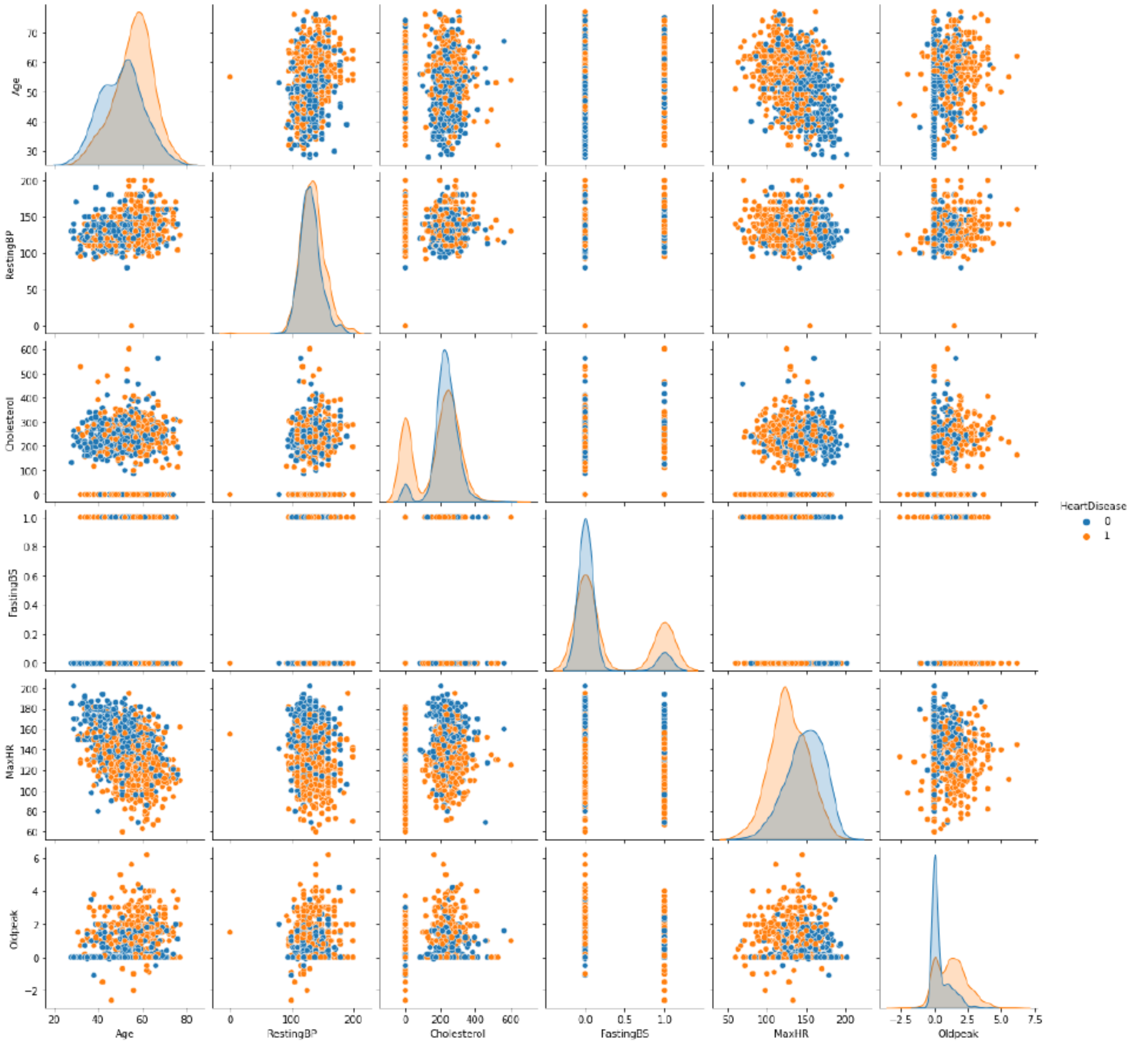
Figure 5: Correlation heat map of Heart Disease dataset.



- From the pair plot below we can see that some features are correlated with the target variable.

  In individuals who are older and having high RestingBP are more prone to Heart Disease.

  Figure 6: Pair Plot of the variables in the dataset.

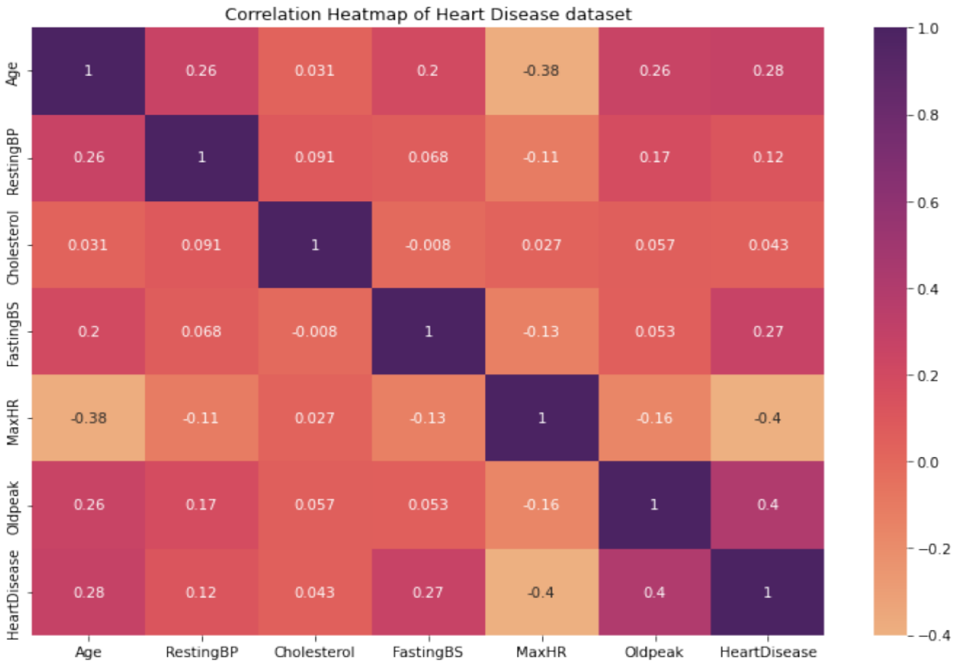


Milestone- 2 (Data preparation)

Milestone 2 focuses on data preparation, which includes handling of outliers, creating dummy variables for the categorical features, elimination of the non-essential features if any, and finally splitting the data into training (80%) and test (20%) data sets. The following are the points from the data preparation.

- Removed outliers identified in the dataset using the interquartile range method.
- The columns 'Cholesterol' and 'RestingBP' are having zeros; the zeros are replaced with the median value of the respective columns.
- Correlation heat map is plotted, and identified the features that are having good correlation to the target variable. From the below correlation map, we can see the correlation between Heart Disease and the remaining features has been improved after the removal of outliers. We can see that Age, RestingBP, Old peak are having high correlation with the target "Heart Disease."

Figure 5: Correlation heat map of Heart Disease dataset after removing the outliers.

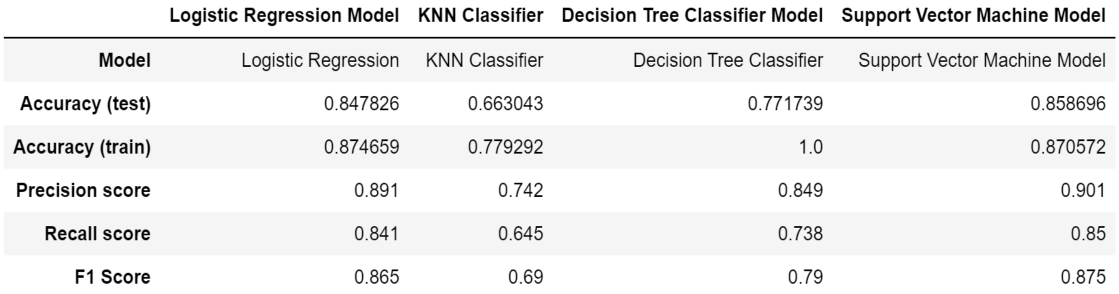


- Dummy variables are created for the categorical columns and the data set is split into 80% test and 20% train datasets.
- Principal Component Analysis (PCA) is applied and second set of train and test data sets are created, with this the features have been reduced from 15 to 11.

Milestone -3 (Model building and evaluation)

Milestone 3 focuses on building four models with the test and train datasets and the PCA applied test and train datasets. For this Heart Disease prediction dataset, I have built 4 models; they are Logistic Regression model, K- Nearest Neighbor Classifier, Decision Tree Classifier, Support Vector Machine (SVM).All the models are created, trained and fit with the test and train datasets and PCA applied test and train datasets respectively. For the selection of the best model that fits our data, evaluation metrics Accuracy, Precision, Recall and F1 Score are calculated with the test data and the PCA applied test data. Confusion Matrices that summarizes the performance of the model build are plotted for each model respectively. After observing the evaluation metrics calculated for each model with the test dataset and the PCA applied test and train datasets, the best model that fits the dataset, is selected. Following points are observed in model building and evaluation stage.

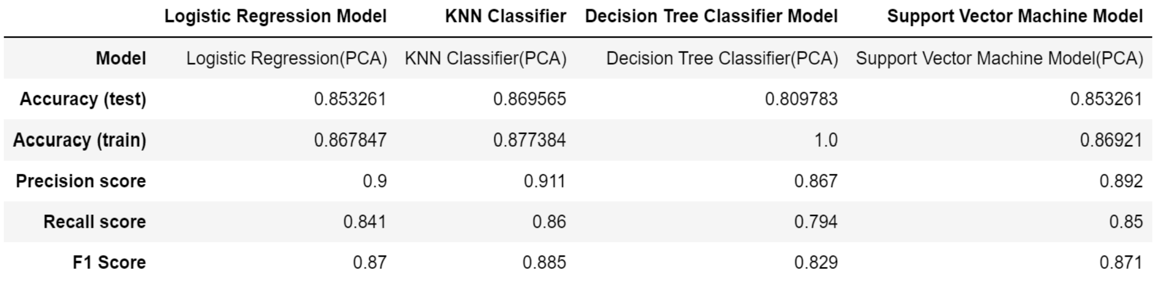
- The summary of the metrics calculated for the models are summarized below.



	Logistic Regression Model	KNN Classifier	Decision Tree Classifier Model	Support Vector Machine Model
Model	Logistic Regression	KNN Classifier	Decision Tree Classifier	Support Vector Machine Model
Accuracy (test)	0.847826	0.663043	0.771739	0.858696
Accuracy (train)	0.874659	0.779292	1.0	0.870572
Precision score	0.891	0.742	0.849	0.901
Recall score	0.841	0.645	0.738	0.85
F1 Score	0.865	0.69	0.79	0.875

- When observed the summary metrics calculated using the actual train and test datasets, the Support Vector Machine Classifier Model performed best. The accuracy score obtained is 0.858(~86%).
- The second best model is the Logistic Regression Model with accuracy score of 0.847(~85%).

- The KNN classifier performed with low accuracy score and the Decision Tree classifier accuracy score is 0.77 but the accuracy of the trained dataset is 1.0 indicating there is some over fit with the trained dataset.
- Summary of the evaluation metrics of the models using PCA trained and test datasets.



	Logistic Regression Model	KNN Classifier	Decision Tree Classifier Model	Support Vector Machine Model
Model	Logistic Regression(PCA)	KNN Classifier(PCA)	Decision Tree Classifier(PCA)	Support Vector Machine Model(PCA)
Accuracy (test)	0.853261	0.869565	0.809783	0.853261
Accuracy (train)	0.867847	0.877384	1.0	0.86921
Precision score	0.9	0.911	0.867	0.892
Recall score	0.841	0.86	0.794	0.85
F1 Score	0.87	0.885	0.829	0.871

- When used the PCA applied training and test datasets the KNN classifier is best performed model, which improved the performance with accuracy score 0.869(~87%) when compared to the performance of the actual trained and test datasets.
- When observed the summary metrics calculated using the PCA applied train and test datasets, the Logistic Regression Model with accuracy score 0.853(85%) and Support Vector Machine Model with accuracy score 0.853(85%) performed second best.
- The Decision Tree classifier model performance didn't change even after using the PCA applied trained dataset.
- Both the KNN Classifier, Logistic Regression and Support Vector Machine models performed good with good accuracy scores when the features are reduced from the 15 to 11 with the PCA application.

Conclusion

What does the analysis/model building tell you?

In the analysis/ model building we see that the models Logistic Regression and the Support Vector Machine models performed best with the normal training and test datasets and the PCA test and train datasets. When PCA applied KNN classifier model performed best with good accuracy. The Decision Tree classifier model accuracy was good with the train data set and PCA applied train dataset, but the accuracy score with the test and PCA applied test dataset is dropped with huge difference indicating that the model is not generalizing and is being over fit with the train dataset. Overall we can say that the Support Vector Machine model performed well with the selected dataset with and without PCA applied train and test datasets.

Is this model ready to be deployed?

Though the Support Vector Machine model performed well with the features available in the dataset, the model is not recommended for deployment, as there are a lot of other features that are to be added to the dataset in prediction of Heart Disease like family history, diet, daily active hours, habit of smoking, habit of alcohol consumption etc.also play an important role in the prediction of Heart Disease. However the model can be tested with the real life data collected from the hospitals and helps understand the features that are having more influence in a person having Heart Disease.

What are your recommendations?

For future recommendations for this project I would like to perform the Hyper parameter tuning on the models and check the model performance, and for evaluation Cross Validation can be included for better model evaluation.

What are some of the potential challenges or additional opportunities that still need to be explored?

Additionally I would like to use the functions and Pipelines for transforming the data and to improve the code readability. There are few features that can be added to the dataset, and more samples can be added to the dataset which helps in better prediction of the Heart Disease.

References

Shubhankar Rawat. "HeartDisease Prediction." Medium,Towards Data Science, 10 Aug 2019, https://medium.com/towards-data-science/heart-disease-prediction-73468d630cfc

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Google, itransition Machine Learning, Aug 24 2022, https://www.itransition.com/machine-learning/healthcare