DSC-680 Applied Data Science
Bellevue University
Summer Term- 2022

Heart Failure Prediction

By:
Yograj Karki
07/02/2022

Table of contents

Table of Figures

Introduction

Background

The cardiovascular system is made up of the heart and blood vessels. Cardiovascular disease (CVD) is defined as any serious, abnormal condition of the heart or blood vessels (arteries, veins). Cardiovascular disease includes coronary heart disease (CHD), stroke, peripheral vascular disease, congenital heart disease, endocarditis, and many other conditions. Many cardiovascular diseases are preventable.

The heart failure syndrome has first been described as an emerging epidemic about 25 years ago. Today, because of a growing and ageing population, the total number of heart failure patients still continues to rise. However, the case mix of heart failure seems to be evolving. Incidence has stabilized and may even be decreasing in some populations, but alarming opposite trends have been observed in the relatively young, possibly related to an increase in obesity.

Most of the risk factors for cardiovascular disease and stroke are modifiable or entirely preventable. By modifying risk factors, you decrease the chances of getting diseases. Modifiable risk factors include tobacco use, high blood pressure, physical inactivity, high blood cholesterol, obesity, heavy alcohol consumption, and poor nutrition. Non-modifiable risk factors are age and family history. The more risk factors one has, the higher the risk of developing disease.

Problem statement

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. **More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age**.

The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity.

Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. So, there is an urgent need of a system which will predict accurately the possibility getting heart disease. Predicting a Heart Disease in early stage will save many people's Life.

Scope

The main objective of this project is to design a robust system which works efficiently and will able to predict the possibility of heart failure accurately. This paper uses the dataset originally from the UCI repository and having 11 important attributes but the combined version of dataset available in Kaggle with 918 unique observations was used. Here are the attributes of the dataset explained:

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]

Methodology

Technical approach

This course project on predictive analytics on heart failure will follow CRISP-DM model for data understanding, modeling, and evaluation. Python programming language will be used to load, explore, and model the prediction system along with necessary machine learning libraries for Python.

Data sources

The required dataset for this project will be retrieved from Kaggle website. Here's the link to the data source: https://www.kaggle.com/fedesoriano/heart-failure-prediction

This dataset on Kaggle is also originally sourced from University of California, Irvine, Machine Learning Repository. Here's the link to that: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Data collection

I acquired the dataset from Kaggle website. This dataset was created by combining different datasets already available independently but not combined before. However original datasets were sourced from University of California, Irvine Machine Learning Repository. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- 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

Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository on the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

Data Analysis

Initially, exploratory data analysis was performed to better understand the data and compute some major descriptive statistics. I produced some descriptive visualizations representing the dataset as well. For the prediction model, various machine learning algorithms was employed to see and choose the best resulting model. Particularly, SVM, Naïve Bayes, Logistic Regression, Decision Tree and KNN are the potential algorithms to be used.

Collected dataset is analyzed using Python programming language and Jupyter Notebook is used as an scripting interface. Several python libraries is used for the various purposed of exploratory data analysis, visualization and machine learning model creation. Some of the notable ones are:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn

Initially after loading the dataset into the pandas dataframe, several information about the dataset was generated such as shape of the dataet, data types, missing values, duplicated observations etc. Then Exploratory data analysis was performed using visualization tools. From the insights gained from visualizations, outliers were detected and dealt with accordingly.
Then the important step of the project, modeling was performed. Data scaling, assigning the numerical and categorical data and encoding was accomplished.

As the data processing was complete, data was trained with several classification algorithms. Trained models were attempted to improve by hyper parameter tuning. Data was first trained with Decision tree classifier and then followed by Random Forest classifier, Logistic regression, K-Nearest neighbor and AdaBoost ensemble. Fine tuning of each of those models were performed and evaluated in each step with testing data. For the model evaluation mainly, accuracy, F1 score and whole classification report was used. In the case of KNN classifier, various k values were tested and the best one was selected to train the model.

Results

Exploratory Data Analysis

Dataset was analyzed in Python's Jupyter notebook. Initially dataset was imported as a pandas data frame. Here's how the dataset looked like.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
				140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

ck to scroll output; double click to hide

Figure 1 : Dataset preview

Dataset has 12 columns and 918 rows of observations. As soon as the data was loaded, missing values and duplicated values were checked. It was found that there were neither missing nor duplicates in the dataset.

Count plot of each feature or column was plot in two sections. Here is the diagram of categorical variables.

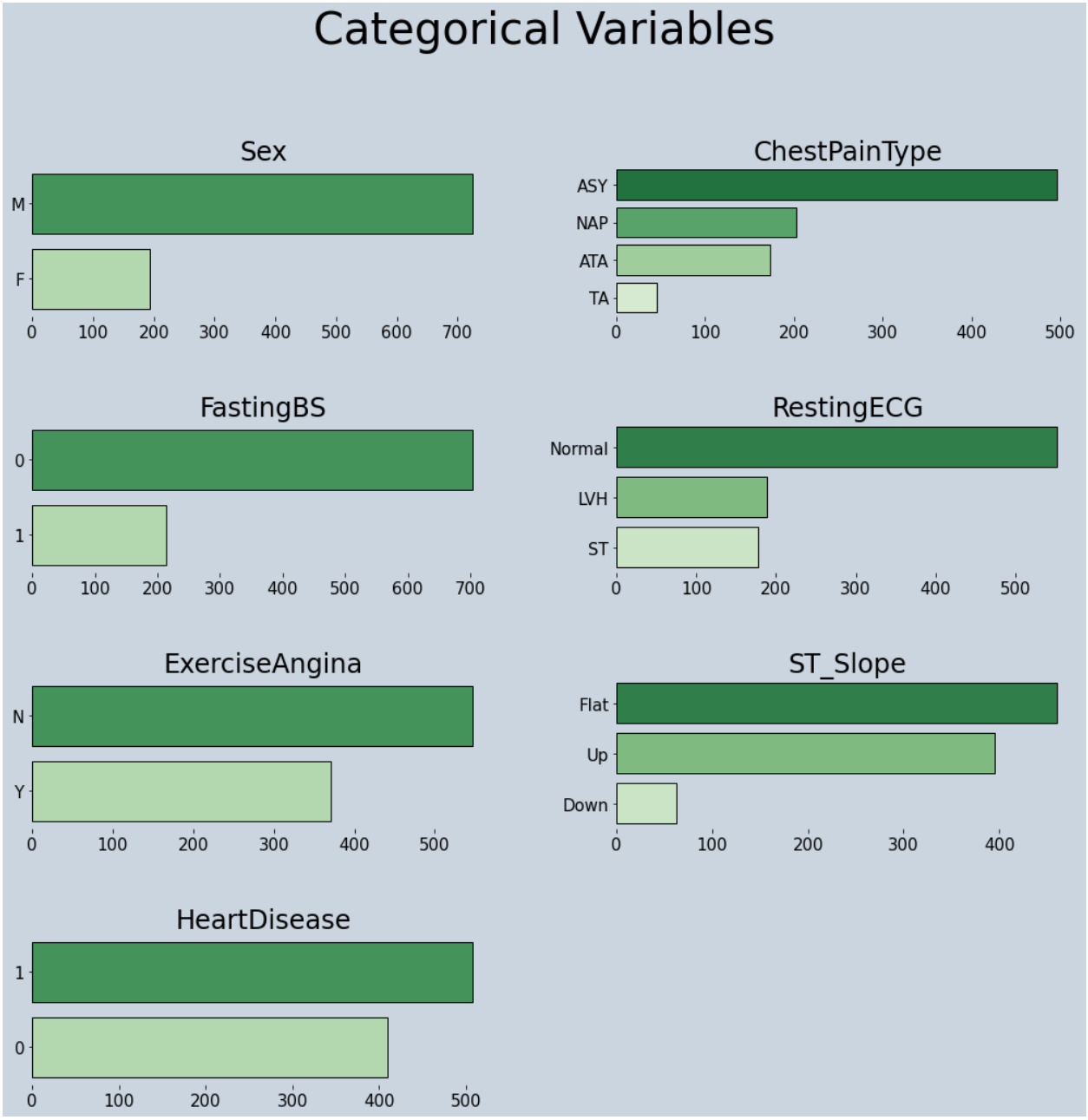

Figure 2 : Count plot of Categorical variables

From the plot above, it can be seen that females were about 25% of the total observations and rest of them were males.

Majority of report about chest pain was asymptomatic followed by non-anginal pain and Atypical Angina. Patients with Typical angina were the lowest in number.

An important observation here is, target variable heart disease status is fairly balanced.

Another chart here is count plot of numerical variables. In the chart below we can see that the variable Age and Max hear rate is fairly normally-distributed.

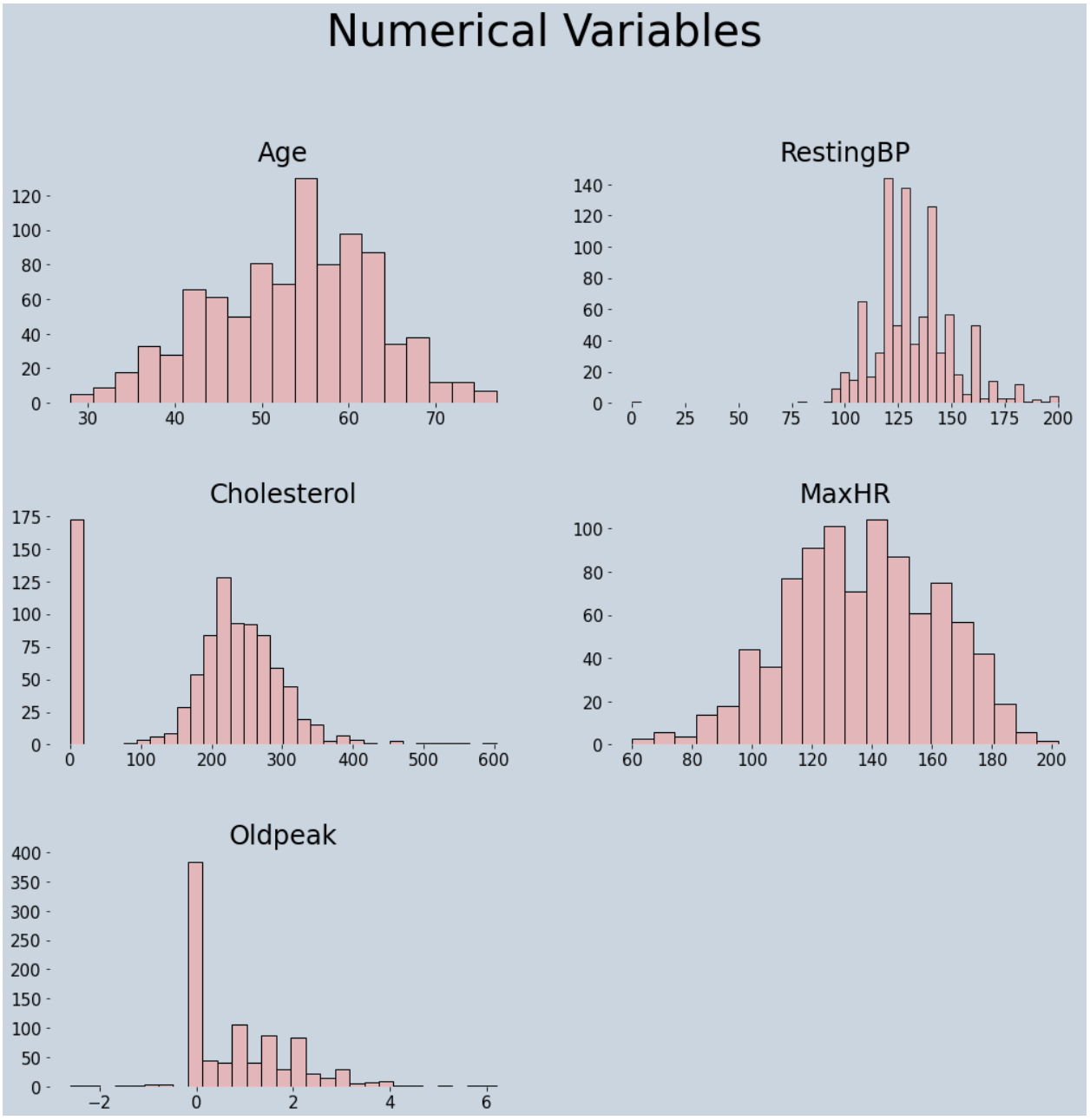

Figure 3 : Histograms of Numerical variables

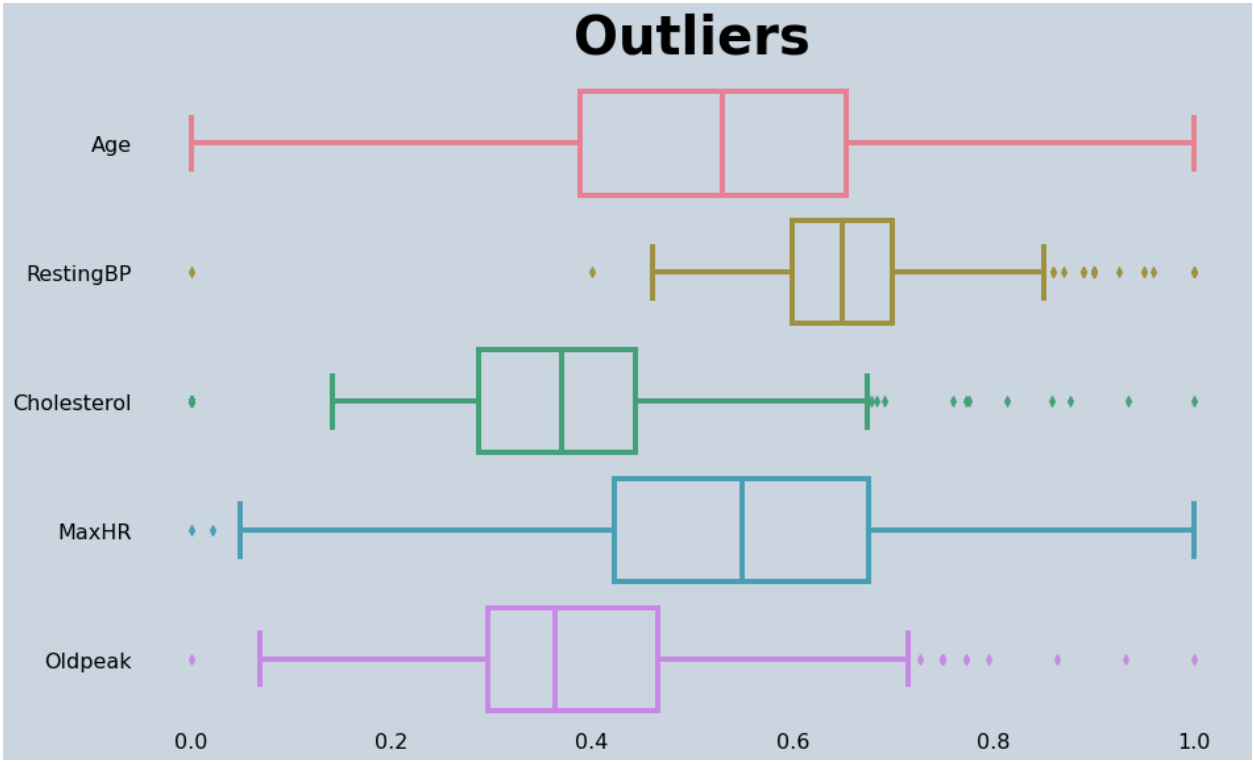

Figure 4 : Outliers in the numerical features

Figure 4 shows the outliers in the numerical variable of the dataset. It is important to note that the data was scaled before plotting this boxplot. Age variable didn't have any outliers where as Resting Blood Pressure had some. Observations which had RestingBP value of 0 were dropped from the dataset since it's impossible to have the Blood pressure of 0 unless the person is dead. Outliers were detected using inter quartile range. Outliers in RestingBP were imputated using median value so was done in Cholesterol variable too. Other outliers were not dealt with because most of them were associated with having heart disease, so it was important to keep them as it is.

Correlation between variables

Next figure here is the heat map of correlations of each variables of the dataset. Figure 5 shows that there is not any strong correlation between the variables. However the most notable one

was with the MaxHR and Age. It has the correlation coefficient of -0.38 which meant that the people of young age tend to higher heart rate values.

Figure 5 : Correlation heatmap

Model selection and evaluation

After the exploratory data analysis, focus shifted onto the model development. First model that was built was using the decision tree classifier. The model had the accuracy score of 0.81, F1 score of 0.81 as well.

Second classifier that was used is Random Forest classifier, it seemed to perform better than the decision tree classifier since it had the F1 score and accuracy score of 0.88. Random forest classifier was again fine-tuned using hyper parameters. Even then it seemed to be improved just slightly.

Another classifier was Logistic regression which is one of the most common classifiers for both classification and regression problems. Standard logistic regression produced the accuracy of 0.88 and F1 score of 0.90. Then it was fine-tuned by setting the L1 and L2 regularization and other parameters. It didn't improve the model by much.

After that, K-nearest neighbor classifier was employed but it had even poorer performance in prediction. It came up with the accuracy score of 0.83. As the value of k matters the most in KNN classification, so I went ahead and experimented with multiple k values.

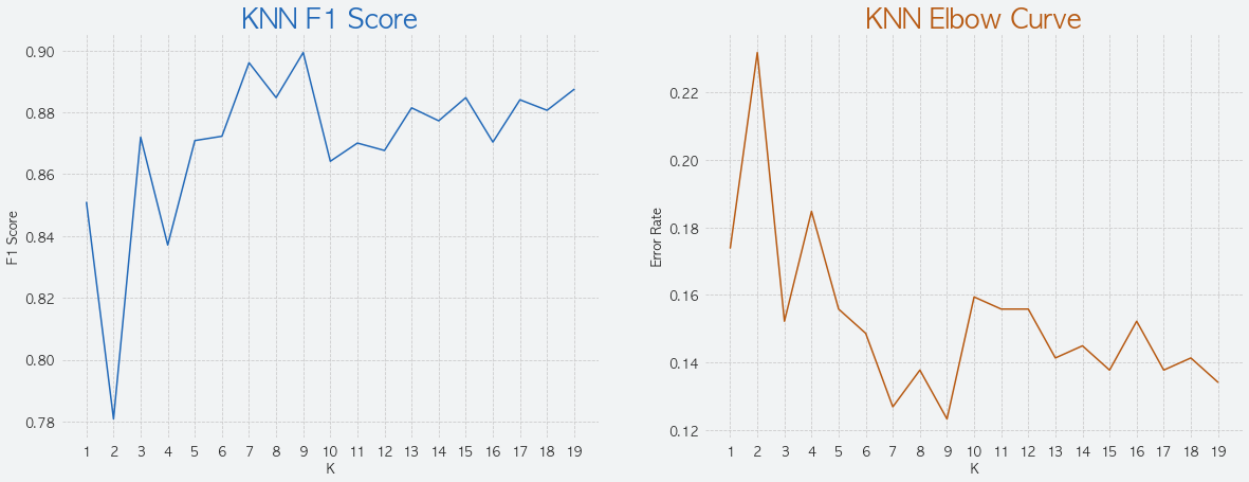

Figure 6 : KNN f1 score and error rate with multiple k values

From figure 6, we can clearly see that the k=9 have highest f1 score and lowest error rate. Then again, KNN classifier with k=9 is used to train the data which produced the accuracy of 0.88. That was quite an improvement to be honest.

Finally, ensemble method was used which is AdaBoost ensemble with decision tree classifier. It showed a promising result out of all the previous models. It had the accuracy of 0.90. AdaBoost ensemble has been selected as the final model for this project.

Discussion

Data understanding

The whole purpose of this project is to create a predictive model and analyze the data of people having heart disease. It is very important to note that this dataset itself is small and with a very few features. But it was something of my interest and the best data I could find out there for free. For a person to develop a disease there are hundreds of factors playing role. By no means, I could be able to create a model which is able to predict the heart disease based on few parameters. It definitely provided some insights about the heart disease itself and the process of developing a predictive model.

I should be thankful that the data I acquired was pretty clean with no missing or duplicate values which saved me a lot of time and effort. I have used mainly the matplotlib and seaborn as the visualization tools. They've produced some good visuals in the notebook.

I chose count plots over other plots for categorical variable because it's easier to see the distribution and symmetry at a quick glance. For the numerical variables though, I generated a set of histograms for each variable.

Data preprocessing

Dealing with outliers was one of the important tasks during this project. Outliers in Resting Blood pressure was quite tricky as some of the data have 0 values in it. I dealt with the outliers in RestingBP and cholesterol by imputing median value and replacing the outlier with it. However, other outliers such as Age, MaxHR and Oldpeak were not imputed and replaced since majority of the observations were associated with the presence of heart disease, so it was important to leave is just like that.

Feature engineering

During the feature engineering, numeric features were scaled using sklearn's MinMaxScaler estimator. It normally scales and translates each feature individually such that it is in the given range between zero and one. Whereas categorical features were encoded using one hot encoding method of Pandas library.

Model development

Model development was the most important and insightful part of the project as I had to go through a bunch of classifiers and experiment with them. Some of the classifiers performed so bad that I did not include them in my notebook.

The first classifier I used was decision tree. The main benefit of this algorithm is its straightforward interpretation and visualization. Impurity refers to the quality of a split in a determined decision. There are several ways to measure the impurity of the decision tree.

Decision Trees don't require any normalization or standardization. This algorithm isn't always the best one but it helps introduce algorithms like the Random Forest. Decision tree classifier model gave the accuracy score of 0.81 upon its evaluation with test data. 0.81 is not a very good score as there are others with higher scores.

Intuitively, I had to go with the Random Forest classifier after the decision tree. Bagging is an ensemble algorithm that fits multiple models on different subsets of a training dataset, then combines the predictions from all models. Random forest is an extension of bagging that also randomly selects subsets of features used in each data sample. Random forest consists of a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The trees protect each other from their individual errors. This classifier performed way better than its predecessor as it had the f1 score of 0.88. I tried to improve it by fine tuning it's parameters, but it only improved by a very small amount. F-1 score went up by 0.01 making it 0.89.

Then, I played with logistic regression in the hope of having a better model. The logistic regression is one of the most popular and straightforward models for classification and regression as well. This model can be used for binary, multinominal or ordinal classification. It can be set a threshold to predict which class an observation relates. The main disadvantage of this model is that the interpretation may result complex. Logistic regression model didn't seem any better than Random forest one. I even attempted to fine tune it using both L1 and L2 regularization. Just like previous model, accuracy score went up just by 0.01 upon fine tuning.

K-Nearest neighbor was another classifier I trained the trained the model with. Without determining the k-value, the standard KNN model produced the accuracy of 0.82 with f1 score of 0.85 which is terrible in a way. I had to go ahead and find the optimal value of k, so did I. Optimal k was found to be 9 and it produced accuracy of 0.88, quite a notable improvement.

At last, the highest accuracy score model seemed to be AdaBoost ensemble combined with Decision Tree classifier. AdaBoost is a boosting ensemble model and works especially well with the decision tree. Boosting model's key is learning from the previous mistakes of misclassification data points. AdaBoost learns from the mistakes by increasing the weight of misclassified data points. It gave out 0.90 accuracy and f1 score. It also had precision of 0.93 to predict the presence of heart disease and 0.86 to predict the absence.

Conclusion

Best predictive model found out to be AdaBoost ensemble classifier which has the highest classification accuracy of 90%. I must admit that predicting a heart disease is a very sensitive undertaking for anyone. I am confident that this model is the best model for the given limited features as in this study. But it is not the best for predicting a heart disease in general because this model doesn't take other important risk factors of heart diseases like smoking, exercise, alcohol consumption, nutrition etc. into account. As this project was an academic project, it gave me practical perspectives of data mining and predictive modeling. Some of the challenging parts of the project were pre-processing the dataset, dealing with outliers, fine tuning the various classification algorithms.

Ethical Considerations

I have explicitly stated the dataset and its source that I used. All 918 observations of the dataset was anonymized by removing the patient's identifiers to protect the privacy rights of patients. The prediction model which will be built by the end of this project will not guarantee the presence/absence of heart disease since there are multitude of factors affecting the development of the disease and all those factors haven't been accounted for in the dataset.

Risks and Assumptions

This study assumes that the dataset is valid and accurate. There is always a chance of losing some data due to its incompleteness and extremeness.

One of the potential risks associated with this study is low number of observations. Also, this dataset does not account all risk factors associated with cardio-vascular diseases for example, physical activity, fruits and vegetables intake, smoking and alcohol consumption etc. so the prediction model cannot be fully accurate.

Recommendation

This model can be used to predict heart disease based on 13 features included in this particular study with 90 percent accuracy, but it can be improved heavily if there is more data and more features including all other important risk factors of cardiovascular diseases. I would recommend any stakeholder to use this model with caution.

References

1. Chris Albon. (2018). *Machine Learning with Python Cookbook Practical Solutions from Preprocessing to Deep Learning*.
2. Fedesoriano. (2021). *Heart Failure Prediction Dataset | Kaggle*. https://www.kaggle.com/fedesoriano/heart-failure-prediction
3. *Introduction Cardiovascular Disease, Introduction - Publication Bureau for Public Health, West Virginia*. (n.d.). Retrieved September 12, 2021, from http://www.wvdhhr.org/bph/cvd/intro.htm
4. Raj, S. (2020). *How to Evaluate the Performance of Your Machine Learning Model - KDnuggets*. https://www.kdnuggets.com/2020/09/performance-machine-learning-model.html
5. scikit-learn. (2021). *API Reference — scikit-learn 1.0.1 documentation*. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics
6. Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. John Wiley & Sons.
7. WHO. (2021). *Cardiovascular diseases*. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
8. Yuan, J. (2018). *HR Predictive Data Analytics in the Era of Big Data*. *67*(Ebmcsr), 388–390. https://doi.org/10.2991/ebmcsr-18.2018.75