Heart Disease Project

Data Description & Project Objective

This data set dates from 1988 of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It originally contained 14 columns. For the sake of this project, I removed 4 columns of categorical data. After that, this database has columns of age, sex (1 = male; 0 = female), trestbps (resting blood pressure), chol (serum cholestoral), fbs (fasting blood sugar), thalach (maximum heart rate achieved), exang (exercise induced angina), oldpeak (ST depression induced by exercise relative to test), slope (the slope of the peak exercise ST segment), target (valued 0 = no disease and 1 = disease). Since heart disease is one of the toughest diseases in human history, I want to find out if we can prevent heart disease. Through this project, I want to explore whether we could predict the probability of getting heart disease in the future by observing our other body data such as blood sugar, heart rate, blood pressure and so on. Since the major goal is trying to predict the probability, I will apply classification models in this project to see if these attributes can be helpful in predicting.

Literature Review

Heart Disease is the leading cause of death in the United State. That's about 610,000 people who die from the condition each year. People often thought heart disease is one certain illness. Actually it can be divided into a range of heart conditions which can be caused by other diseases of our body such as coronary artery disease (CAD) and peripheral artery disease (PAD). There are a lot of other factors that can cause heart disease, so it is important for us to know how they are related to our heart if we want to predict the probability of getting heart disease.

In the database of this project, the first column is the sex. Actually, there is a strong relation between sex and the heart disease. Women often experience different signs and

symptoms of heart disease than men. Moreover, it is easier for women to confuse the heart disease with other conditions, such as depression, menopause, and anxiety, so it is more important for women to prevent the heart disease than men. Also research found that diabetes, high blood pressure, stress and anxiety can all cause an abnormal heart rhythm.

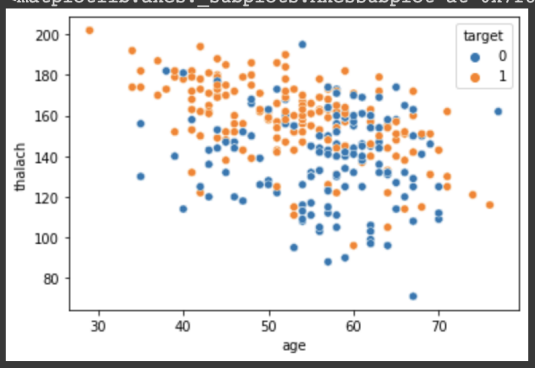
Diabetes also has a strong relationship with heart disease. Cardiovascular disease (CVD) or coronary disease, which can lead to heart attacks and strokes, means that people are more at risk of heart disease when they have diabetes. Cardiovascular disease affects human body circulation, which can make other diabetes complications worse. High blood sugar levels can lead to damage of the blood vessels and serious heart complications. Once our body cannot handle all of this blood sugar properly, they will stick to red blood cells and build up in the blood. Finally, it will prevent the vessels from carrying blood to and from our hearts, starving the heart of oxygen and nutrients.

Besides the CVD, Hypertensive heart disease is the NO.1 cause of death associated with high blood pressure. It includes heart failure, ischemic heart disease, and left ventricular hypertrophy. All of these are related to the heart's pumping chambers, and making it harder for your heart to deliver oxygen, blood and nutrients to our body. The database in this project also includes the data of the ST segment and its slope. According to one research and experiment conducted by the National Library of Medicine, asymptomatic ST-segment depression was a very strong predictor of sudden cardiac death in men with any conventional risk factor but not previously diagnosed CHD. Based on this research, the data of ST-segment is also a valuable source of information.

Exploratory Data Analytics

```
[17] sns.scatterplot(x = df['age'], y = df['thalach'], hue = df['target'])
```

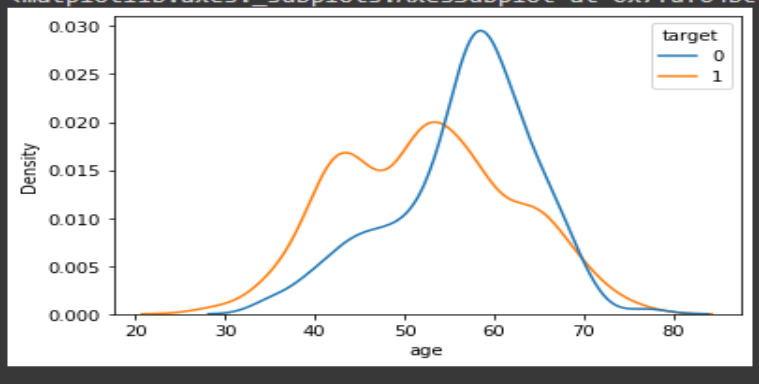
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8fb1a203d0>
```



The most obvious factor that we can observe by ourselves is the heart rate when it comes to heart disease. The above scatter plot indicates different maximum heart rates under different ages. The y-axis thalach stands for maximum heart rate achieved, orange dot represents those people with heart disease and blue dot for no disease. There is a clear trend that the maximum heart rate will decrease with age. However, for those people at the same age, people with higher maximum heart rate are more likely to get heart disease. Therefore, we can conclude that heart rate has a strong relationship with heart disease.

```
sns.kdeplot(x = df['age'], hue = df['target'])
```

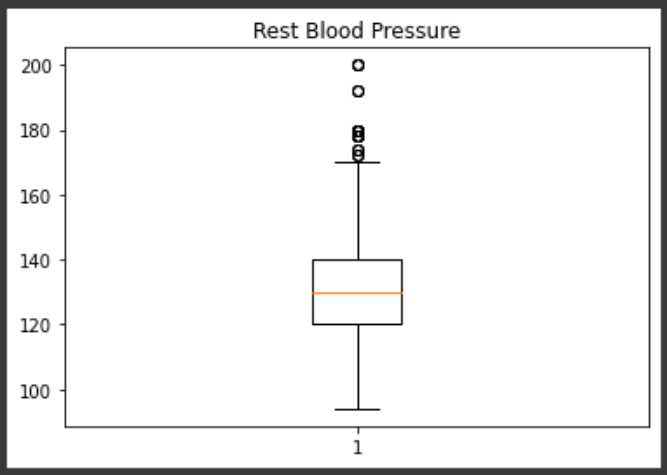
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdf64bc4ad
```



From the above plot, we could easily see the age distribution of the whole dataset. One of the most interesting things is that the amount of people who had heart disease is higher than

those healthy people until mid age. Most people keep recognizing that heart disease is more dangerous for elder group. However, the fact is the rate of younger people getting heart disease is rising.

```
[10] boxplot(df['trestbps'])
     title('Rest Blood Pressure')
     show()
```



This boxplot showed us the data of the resting blood pressure in this dataset. The normal resting blood pressure is between 80 and 100, and 140 could be considered as a high blood pressure. Most samples in this dataset are around 50 years old, it could be more easily to observe the high blood pressure. However, it is clear that high blood pressure has a strong relation with heart disease.

The table below showed the average, Q1, medium, Q3, and maximum number of resting blood pressure, serum cholesterol, and maximum heart rate achieved. Both resting blood pressure and serum cholesterol are higher than the normal range, especially serum cholesterol. Healthy serum cholesterol is usually under 200mg/dL, but the mean serum cholesterol here has reached 246mg/dL.

	age	trestbps	chol	thalach
count	1025.000000	1025.000000	1025.00000	1025.000000
mean	54.434146	131.611707	246.00000	149.114146
std	9.072290	17.516718	51.59251	23.005724
min	29.000000	94.000000	126.00000	71.000000
25%	48.000000	120.000000	211.00000	132.000000
50%	56.000000	130.000000	240.00000	152.000000
75%	61.000000	140.000000	275.00000	166.000000
max	77.000000	200.000000	564.00000	202.000000

Analysis & Discussion

For this dataset, I have applied logistic regression and decision tree classification to analyze the data. From the results, it could be concluded that the decision tree model has a better accuracy than the logistic regression model. The decision tree model returns a higher true positive and true negative. Other data like f1-score and precision are also higher in the evaluation of the decision tree model.

Compared to other similar studies using the same dataset, they have applied more effective graphs and diagrams in the EDA phrase to better interpret the relationship between each variable. Basically, most studies applied the same logistic regression with me, but some of them have visualized their evaluation results by Receiver Operating Characteristic Curve to indicate the performance of the model. Also, a study that used KNN models has better accuracy than mine.

Conclusion

The aim of this study is going to reveal the importance of preventing heart disease. The study revealed the relationship between heart disease and blood sugar, blood pressure and serum cholesterol. I applied logistic regression models and decision tree classifiers to see if we can predict the probability of getting heart disease accurately. There is one study I referenced that applied both KNN and SVM models. The accuracy of these two models are much higher than I thought, especially the KNN model. I have considered using the KNN model instead of the decision tree initially, however, I think the decision tree could have a better performance. After this project, I have learned that it is necessary to conduct as many models as possible to get the most effective one in the future.

Works Cited

bayomars12. "Starter: Heart Disease Dataset 408F5662-7." *Kaggle*, Kaggle, 27 Sept. 2019, https://www.kaggle.com/code/bayomars12/starter-heart-disease-dataset-408f5662-7

Beckerman, James. "High Blood Pressure and Hypertensive Heart Disease." *WebMD*, WebMD, https://www.webmd.com/hypertension-high-blood-pressure/guide/hypertensive-heart-disease

"Coronary Heart Disease - What Is Coronary Heart Disease?" *National Heart Lung and Blood Institute*, U.S. Department of Health and Human Services, https://www.nhlbi.nih.gov/health/coronary-heart-disease

"Cardiovascular Disease: Heart Disease Causes and Symptoms." *Cleveland Clinic*, https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease

Donovan, Robin. "Heart Disease: Risk Factors, Prevention, and More." *Healthline*, Healthline Media, 27 Feb. 2020, https://www.healthline.com/health/heart-disease#causes

EmmaHook. "Diabetes and Heart Disease." *Diabetes UK*, Diabetes UK, https://www.diabetes.org.uk/guide-to-diabetes/complications/cardiovascular_disease

harmeetsingh07. "Detailed Notebook on Logistic Regression + E.D.A." *Kaggle*, Kaggle, 16 Apr. 2022, https://www.kaggle.com/code/harmeetsingh07/detailed-notebook-on-logistic-regression-e-d-a.

Laukkanen, Jari A, et al. “Asymptomatic St-Segment Depression during Exercise Testing and the Risk of Sudden Cardiac Death in Middle-Aged Men: A Population-Based Follow-up Study.” *European Heart Journal*, Oxford University Press, Mar. 2009, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2721711/

Mchtklb. “Kalebasi - Heart Disease ML Modeling.” *Kaggle*, Kaggle, 29 Sept. 2020, https://www.kaggle.com/code/mchtklb/kalebasi-heart-disease-ml-modeling.

NHS Choices, NHS, https://www.nhs.uk/conditions/coronary-heart-disease/causes/

Richard N. Fogoros, MD. “Learn Which Type of Heart Attack Is the Most Serious.” *Verywell Health*, Verywell Health, 2 Mar. 2022, https://www.verywellhealth.com/stemi-st-segment-elevation-myocardial-infarction-1746032

Welch, Ashley, et al. “What Is Heart Disease? Symptoms, Causes, Diagnosis, Treatment, and Prevention.” *EverydayHealth.com*, https://www.everydayhealth.com/heart-disease/

“What Is Cardiovascular Disease?” *Www.heart.org*, 22 July 2021, https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease