Heart Disease Predication with Supervised Machine Learning Algorithm

Hua Mo

Table of Content

Introduction

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 37 seconds in the United States from cardiovascular disease. About 647,000 Americans die from heart disease each year, which is 1 in every 4 deaths. Heart disease costs the United States about $219 billion each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death.

There are different root causes of the heart disease. Therefore, a few lifestyle choices can increase the risk of heart disease. These include high blood pressure and cholesterol, smoking, overweight and obesity, diabetes, family history, diet of junk food, age, a history of preeclampsia during pregnancy, staying in a stationary position for extended periods of time. Having any of these risk factors greatly increases the risk of heart disease. There are also a lot of symptoms during the heart disease, such as chest pain, etc.. However due to the complex of human body, an unfavorable lifestyle will not always 100% to have heart disease. The same thing also happens to the symptoms. Not all the people have unfavorable symptom will have heart disease. However, if there are more symptoms, there will be a higher possibility to have heart disease. The detail is illustrated in the Fig. 1.

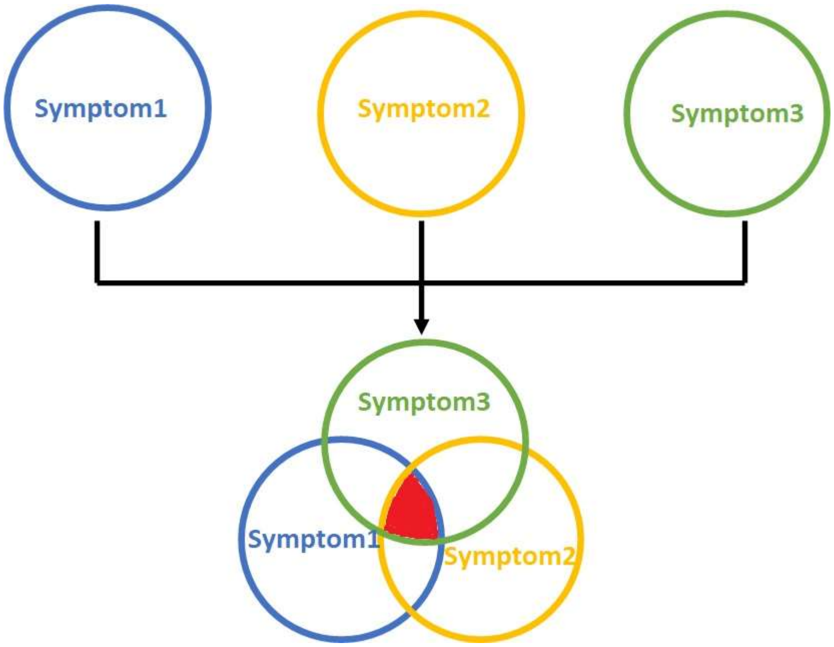


Fig. 1. Symptoms for Heart Disease

From Fig.1, a person with has only symptom1, or symptom2, or symptom3 doesn’t mean that he/she has a heart disease. However, when a person found these three symptoms, there is a high possibility that he/she is having heart disease, which is marked in the red area.

Fig. 1 indicated that there are correlations between several symptoms and potential to have heart diseases, not necessary for only one symptom. To find out the correlations between these parameters and potential to have heart diseases will be a challenge for the health workers. The best way to do it is to establish models.

Machine Learning, especially supervised learning, has been used in different areas. In this project, we will use machine learning technique to setup models. These models will be used to get the correlation between the symptoms and potential to have heart diseases. We will also tune the parameters and compare the advantage and disadvantage of these models.

Data Wrangling

The first step to process data is data wrangling. Data wrangling involves taking raw data and preparing it for processing and analysis. It will have

- Data Collection / Acquistion
- Data Organization
- Data Definition
- Data Cleaning

I. Data Collection / Acquisition

The heart data is extracted from heart disease in Kaggle (https://www.kaggle.com/ronitf/heart-disease-uci) . Based on the information the original database contains 76 attributes. But all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The original data was extracted and found some errors. After search for the discussion board, an updated data version was obtained.

II. Data Organization

The raw data was organized with Python. Pandas read_csv was used to input the data from the excel spreadsheet. To test whether the data input was correct, the first five lines of data format was displayed and compared with the original excel spreadsheet. The results showed that there were no differences between input data and original excel spreadsheet, which indicated that the data input was successful. The first five lines of data is shown in Table 1 and Table 2. The Table I and Table 2 showed a well-organized data format. Therefore, there would be no need to organize the data.

Table 1. The first five lines of data (first eight columns)

	age	**sex**	**cp**	**trestbps**	**chol**	**fbs**	**restecg**
0	69	1	0	160	234	1	2
1	69	0	0	140	239	0	0
2	66	0	0	150	226	0	0
3	65	1	0	148	282	1	2
4	64	1	0	110	211	0	2

Table 2. The first five lines of data (last seven columns)

thalach	**exang**	**oldpeak**	**slope**	**ca**	**thal**	**condition**
131	0	0.1	1	1	0	0
151	0	1.8	0	2	0	0
114	0	2.8	2	0	0	0
174	0	1.4	1	1	0	1
144	1	1.8	1	0	0	0

III. Data Definition

There were 297 rows and 14 columns, exclude the index column. The column heads represented the meaning of the data. The rows are the data under a certain criterion. Since some column heads were either abbreviations, or jargons, or something unknown, the column heads need to be explained to help understand the meaning of data. The explanation of individual column head is shown in Table 3. The data type information and calculation of potential null is shown in Table 4. There are 13 int64 data and 1 float64.

Table 3. Explanation of Column Head

Column #	Column Head	Explanation	Note
1	age	Age of the people to take the test	
2	sex	Sex of the people to take the test	Value 0: female Value 1: male
3	cp	Chest Pain Type (4 values) The severe of chest pain increase when value increase	Value 0: asymptomatic Value 1: atypical angina Value 2: non-anginal pain Value 3: typical angina
4	trestbps	Resting blood pressure	
5	chol	serum cholestoral in mg/dl	
6	fbs	Fasting blood sugar > 120 mg/dl	Value 0: No Value 1: Yes
7	restecg	Resting electrocardiographic results (three values)	Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria Value 1: normal Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
8	thalach	Maximum heart rate achieved	
9	exnag	Exercise induced angina	
10	oldpeak	ST depression induced by exercise relative to rest	
11	slope	The slope of the peak exercise ST segment (three values)	Value 0: downsloping Value 1: flat Value 2: upsloping
12	ca	Number of major vessels (0-3) colored by flourosopy	
13	thal	Unknown parameter	Value 0: normal Value 1: fixed defect Value 2: reversable defect
14	condition	Potential for heart disease (two values)	Value 0 well people Value 1 heart disease

Table 4. Data Information

#	Column	Non-Null Count	Dtype
0	age	297 non-null	int64
1	sex	297 non-null	int64
2	cp	297 non-null	int64
3	trestbps	297 non-null	int64
4	chol	297 non-null	int64
5	fbs	297 non-null	int64
6	restecg	297 non-null	int64
7	thalach	297 non-null	int64
8	exang	297 non-null	int64
9	oldpeak	297 non-null	float64
10	slope	297 non-null	int64
11	ca	297 non-null	int64
12	thal	297 non-null	int64
13	condition	297 non-null	int64

IV. Data Cleaning

Since there are not any categorical data and null data, the data were clean and ready to be used.

Data Exploration

Data exploration is used to explore the relations among different features. We will investigate the relationship among different features by heatmap and individual feature vs. condition

I. Heat Map

A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. The color on the heatmap proportional to the relationship between different features. The heatmap with annotation is especially useful because it can provide a relative quantitative relationship among different features. The positive number in the heatmap indicates both features increase at the same direction. The negative number in the heatmap indicates the trend of features has the opposite directions. For examples, if a feature vs. condition is positive, when the value of feature increase, the condition will also increase. If a feature vs. condition is negative, when the value of feature increase, the condition will decrease.

The heatmap of different features is shown in Fig. 2.

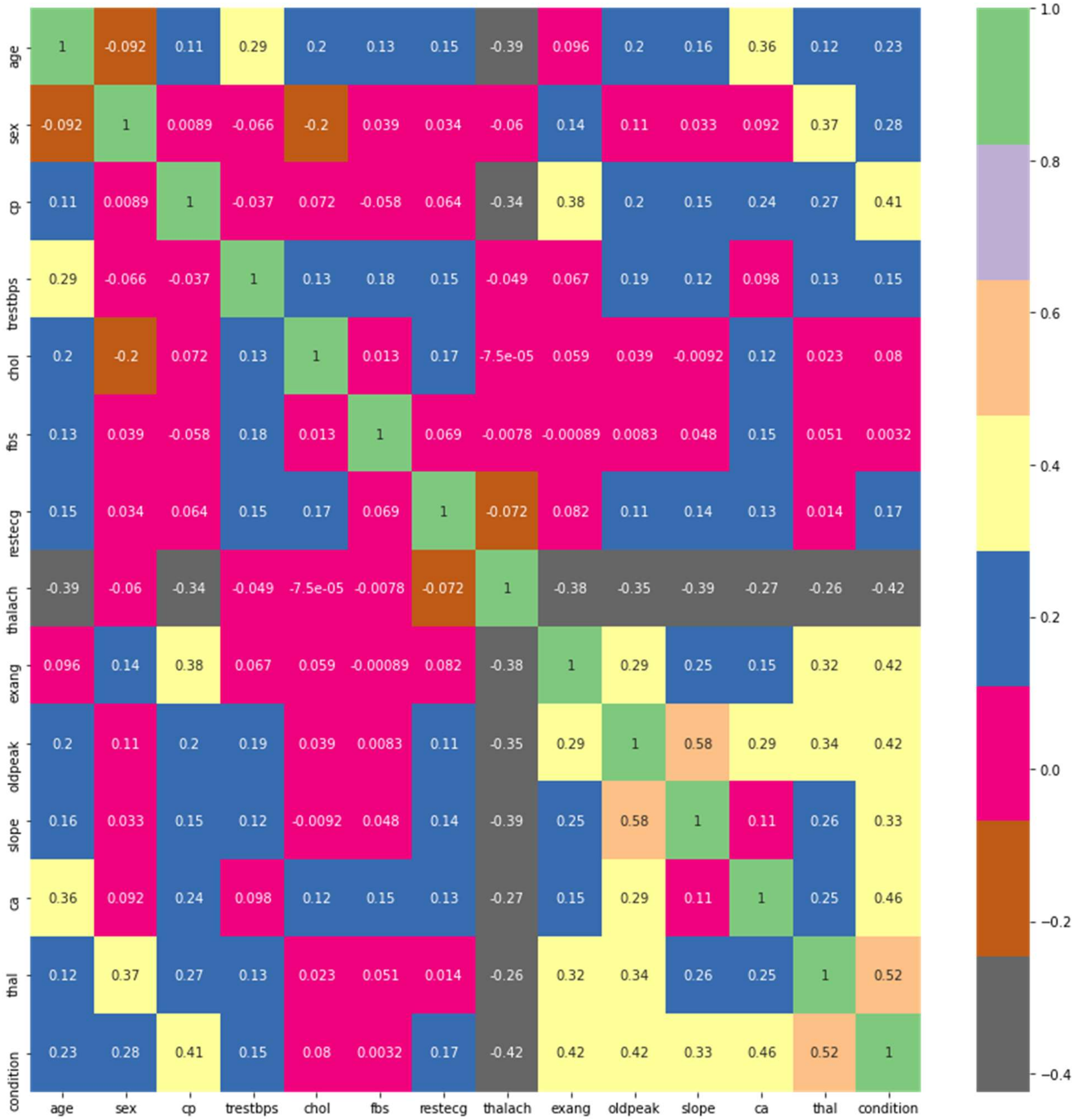


Fig.2. The Heatmap of Different Features

Based on the heatmap, we could see

- Age
  - relative small relationship with sex, exnag(exercise induced angina), which was expected. The older people normally does not exercise a lot.
  - positive relation with cp(chest pain), trestbps(resting blood pressure), chlo(serum chllestoral), fbs(Fasting blood sugar>120 mg/dl), restecg(Resting

electrocardiographic results), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy), thal (unknown parameter), condition (potential for heart disease), which was reasonable.
- negative relation with thalach (Maximum heart rate achieved), which is also reasonable. The old people does not have a strong heart compared with young people

- Sex
  - relative small relationship with age, cp(chest pain), trestbps(resting blood pressure), fbs(Fasting blood sugar>120 mg/dl), restecg(Resting electrocardiographic results), thalach (Maximum heart rate achieved), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy),which was expected.
  - positive relation with exnag(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest), thal (unknown parameter),condition (potential for heart disease).
  - negative relation with chlo(serum chllestoral). The results indicated that female seems to have lower serum chllestoral

- CP (chest pain)
  - relative small relationship with sex, trestbps(resting blood pressure), chlo(serum chllestoral), fbs(Fasting blood sugar>120 mg/dl), restecg(Resting electrocardiographic results).
  - positive relation with age, exnag(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy), thal (unknown parameter),condition (potential for heart disease).
  - negative relation with thalach (Maximum heart rate achieved)

- Trestbps (Resting Blood Pressure)
  - relative small relationship with sex, cp(chest pain), thalach (Maximum heart rate achieved),exnag(exercise induced angina), ca(Number of major vessels colored by fluoroscopy),
  - positive relation with age, chlo(serum chllestoral), fbs(Fasting blood sugar>120 mg/dl), restecg(Resting electrocardiographic results), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), thal (unknown parameter),condition (potential for heart disease).

- Chol(Serum cholestoral in mg/dl)

- relative small relationship with cp (chest pain), fbs(Fasting blood sugar>120 mg/dl), thalach (Maximum heart rate achieved), exnag(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), thal (unknown parameter),condition (potential for heart disease).
- positive relation with age, trestbps(resting blood pressure), restecg(Resting electrocardiographic results), ca(Number of major vessels colored by fluoroscopy)
- negative relation with sex

- Fbs (Fast blood sugar > 120 mg / dl)
  - relative small relationship with sex, cp(chest pain), chlo(serum chllestoral), restecg(Resting electrocardiographic results), thalach (Maximum heart rate achieved), exnag(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), thal (unknown parameter),condition (potential for heart disease)
  - positive relation with age, trestbps(resting blood pressure), ca(Number of major vessels colored by fluoroscopy)

- Restecg (Resting electrocardiographic results)
  - relative small relationship with sex, cp(chest pain),fbs(Fasting blood sugar>120 mg/dl), thalach (Maximum heart rate achieved), exnag(exercise induced angina), thal (unknown parameter)
  - positive relation with age, trestbps(resting blood pressure), chlo(serum chllestoral), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy), condition (potential for heart disease).

- Thalach(Maximum heart rate achieved)
  - relative small relationship with sex, trestbps(resting blood pressure), chlo(serum chllestoral), fbs(Fasting blood sugar>120 mg/dl), restecg(Resting electrocardiographic results)
  - negative relation with age, cp(chest pain),exnag(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy), thal (unknown parameter),condition (potential for heart disease).

- Exnag (Exercise induced angina)
  - relative small relationship with age, trestbps(resting blood pressure), chlo(serum chllestoral), fbs(Fasting blood sugar>120 mg/dl), restecg(Resting electrocardiographic results)

- positive relation with sex, cp(chest pain), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy), thal (unknown parameter),condition (potential for heart disease).
- negative relation with thalach (Maximum heart rate achieved)

- Oldpeak (ST depression induced by exercise relative to rest)
  - relative small relationship with chlo (serum chllestoral), fbs(Fasting blood sugar>120 mg/dl),
  - positive relation with age, sex, cp(chest pain), trestbps(resting blood pressure), restecg(Resting electrocardiographic results), exnag(exercise induced angina), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy), thal (unknown parameter),condition (potential for heart disease).
  - negative relation with thalach (Maximum heart rate achieved)

- Slope (The slope of the peak exercise ST segment)
  - relative small relationship with sex, chlo(serum chllestoral), fbs(Fasting blood sugar>120 mg/dl)
  - positive relation with age, cp(chest pain), trestbps(resting blood pressure), exnag(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy), thal (unknown parameter),condition (potential for heart disease).
  - negative relation with thalach (Maximum heart rate achieved)

- Ca (Number of major vessels by fluoroscopy)
  - relative small relationship with sex, trestbps(resting blood pressure)
  - positive relation with age, cp(chest pain), chlo(serum chllestoral), , fbs(Fasting blood sugar>120 mg/dl), restecg(Resting electrocardiographic results), exnag(exercise induced angina), oldpeak(ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST segment), ca(Number of major vessels colored by fluoroscopy), thal (unknown parameter),condition (potential for heart disease).
  - negative relation with thalach (Maximum heart rate achieved)

- Thal (unknown parameter)
  Since this parameter is unknown, we will not discuss about it.

II. The potential heart disease affected by individual feature

There were 297 people. 137 people had heart disease. 160 people were well people, who do not have heart disease. The samples size for the heart disease vs. well people were similar. In this project, we will compare the features between people with heart disease and well people.

1. Age

The age comparison of heart diseases vs. well people is shown in Fig.3

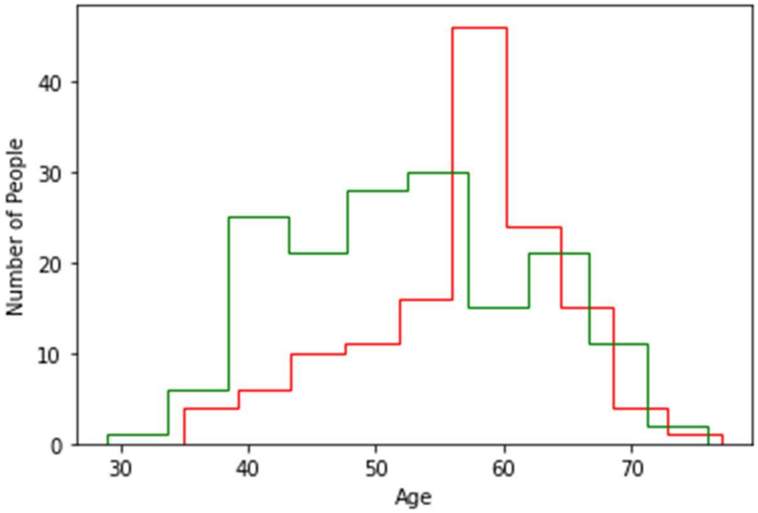


Fig. 3. Age Comparison between Heat Disease (red) People vs. Well People (green)

The green line is the age distribution of well people. The red line is the distribution of the heart disease people. The result showed

- the age for well people follow a normal distribution of age group. The age of heart disease people showed the distorted normal distribution. The distribution age of heart diseases is narrower.
- The maximum age distribution shifted from 55 (well people) to ~ 60 (heart disease people)
- The data indicated that people at ~ 60 would have a higher potential for heart disease

To confirm the hypothesis, Cumulative Distribution Functions (CDF) was used to test the hypothesis. If the two data set are the same, the CDF curve should be similar. The CDF curve is shown in Fig. 4.

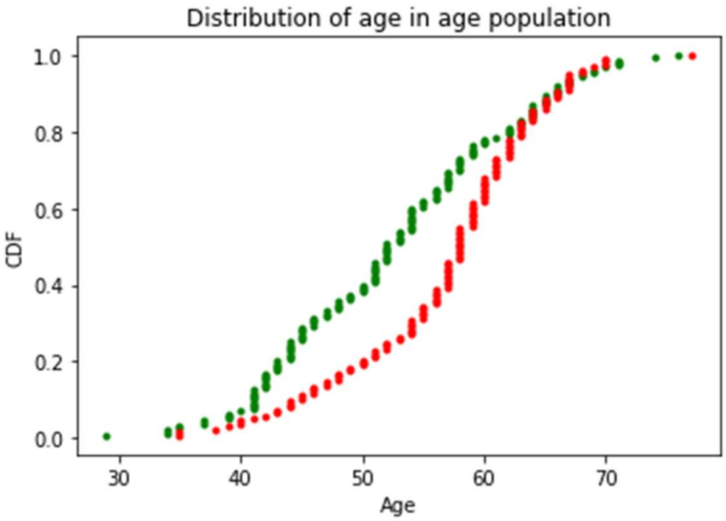


Fig. 4. The CDF of Age Distribution of Heart Disease People (red) vs. Well People (gree n)

The results confirmed that there is a distribution difference in the center of the age. The 0.5 on the CDF was 55 for well people (green line). The 0.5 on the CDF was 60 for heart disease people. The CDF confirmed hypothesis

Boxplot was also used to test the spreading and mean of the data. The result is shown i n Fig. 5.

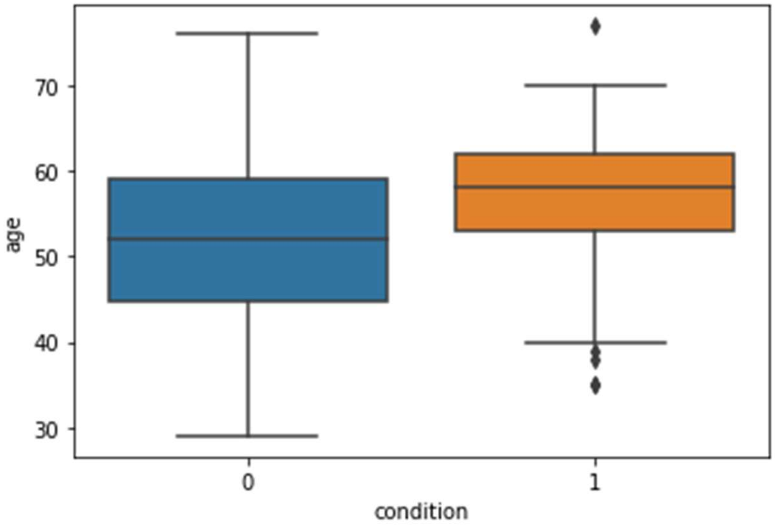


Fig. 4. The Boxplot of Age Distribution of Heart Disease People (orange) vs. Well Peopl e (blue)

The results indicated that all the data for well people (blue) were within the outlier. The mean was around 52. However, some data for heart disease people (orange) were out of the outlier. That indicated that the heat disease peoples in this test had more data spr eading. The mean age to have the heart diseases was around 59.

Based on the data analysis, the potential age to have heart disease will around 59-60.

2. Sex (Gender)
The gender comparison of heart diseases vs. well people is shown in Fig.5 and Fig. 6

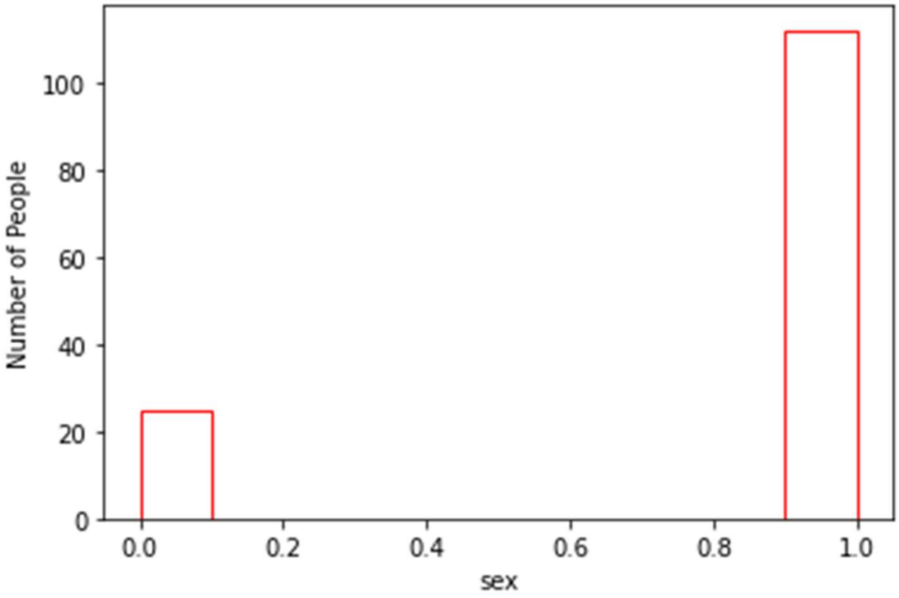


Fig. 5. Gender Distribution of Heat Disease People (red)

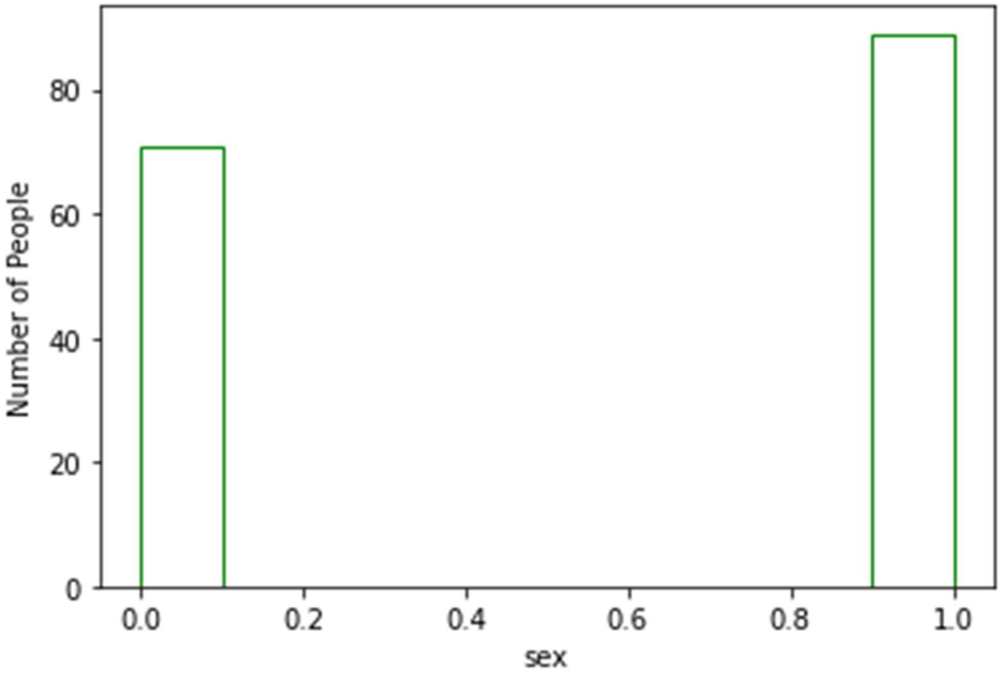


Fig.6. Gender Distribution of Well People (green)

In Fig.5 and Fig. 6, Value 0 represents female. Value 1 represents male. The result sho wed that in the well people, female and male were almost equally distributed. However more male than females in the heart diseases. Therefore, assume there is no difference between the test people, the male would likely to have higher potential for heart diseas e.

3. Chest Pain (cp)

Chest pain is one of symptoms for heart disease. There are four levels of chest pain – asymptomatic (0), atypical angina (1), non-anginal pain (2) and typical angina(3). The comparison of CP is shown in Fig. 7 and Fig. 8.

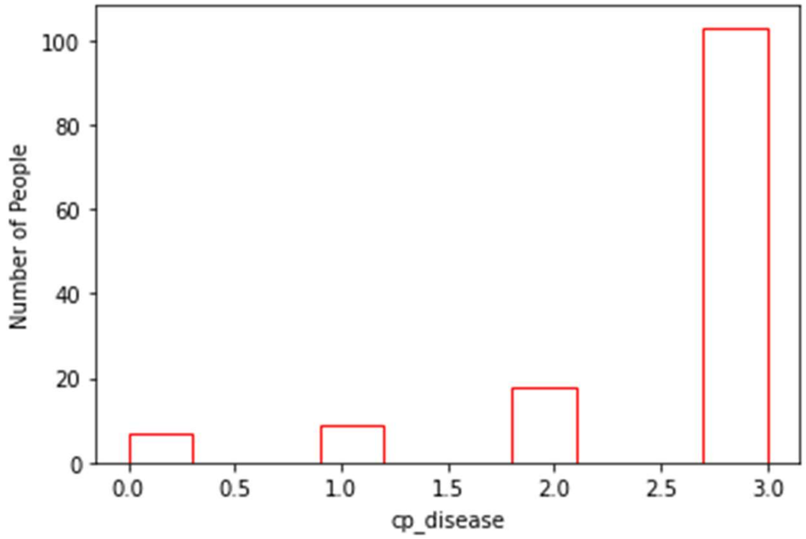


Fig. 7. Chest Pain of Heart Disease

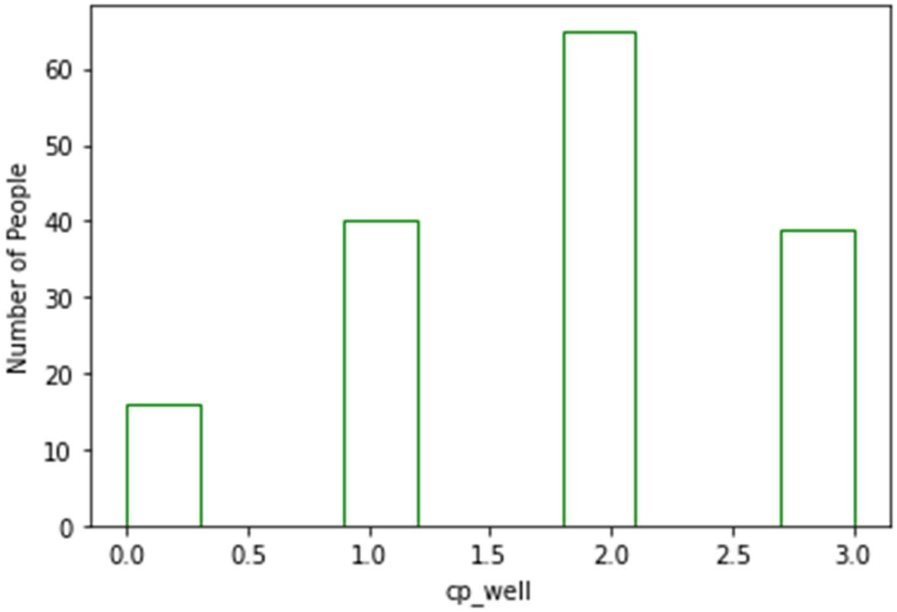


Fig. 8. Chest Pain of Well People

The result showed that the heart disease people had a higher portion of typical angina (value =3).

In another word, when a people have a typical angina (value=3), he / she likely to have a heart disease.

4. Trestbps (Resting Blood Pressure)

The resting blood pressure (Tresbps) comparison of heart diseases vs. well people is shown in Fig.9

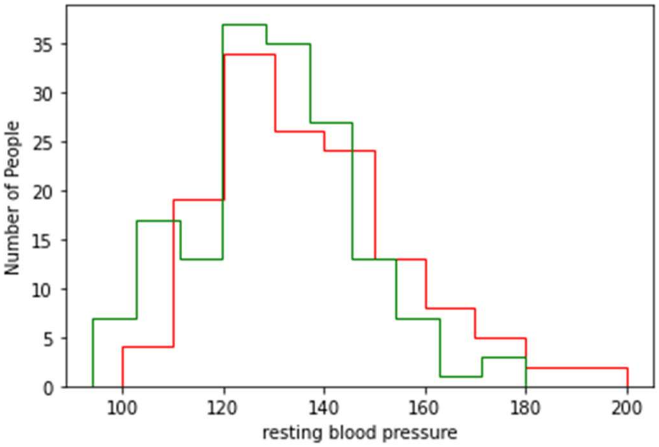


Fig. 9. Resting Blood Pressure Comparison between Heat Disease (red) People vs. Well People (green)

The green line is the resting blood pressure distribution of well people. The red line is the distribution of the heart disease people. The result showed that there was a slight shift of resting blood pressure to higher distribution level.

To confirm the hypothesis, Cumulative Distribution Functions (CDF) was used to test the hypothesis. If the two data set are the same, the CDF curve should be similar. The CDF curve is shown in Fig. 10.

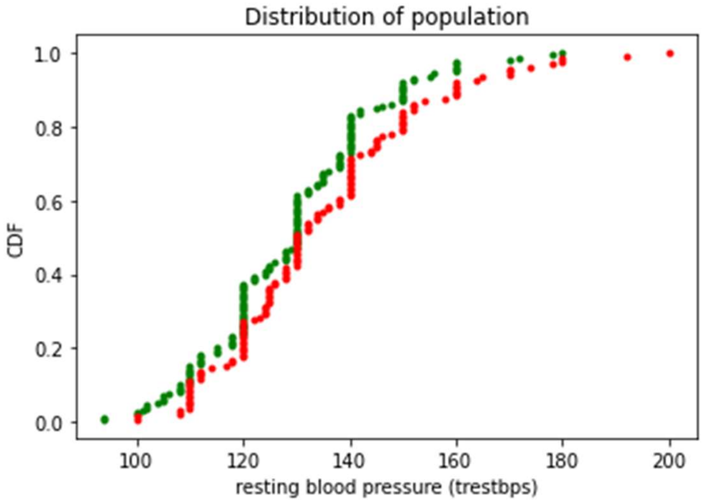


Fig. 10. The CDF of Resting Blood Pressure Distribution of Heart Disease People (red) vs. Well People (green)

The results confirmed that two data sets were similar with a shift of distribution. The highest resting blood pressure of people was around 180. A small portion people with heart diseases could go to 200.

Boxplot was also used to test the spreading and mean of the data. The result is shown in Fig. 11.

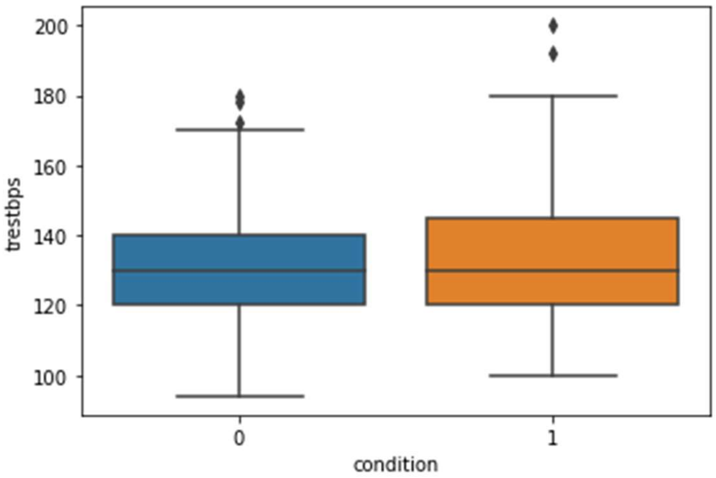


Fig. 11. The Boxplot of Resting Blood Pressure Distribution of Heart Disease People (orange) vs. Well People (blue)

The results indicated that some data for well people (blue) and people with heart diseases(orange) were laying outside of the boundary. There was a slight difference of the mean.

Based on the data analysis, the heart diseases people appeared to have a higher resting blood pressure.

5. Serum Cholesterol Level (chol)

The serum cholesterol (chol) comparison of heart diseases vs. well people is shown in Fig.12

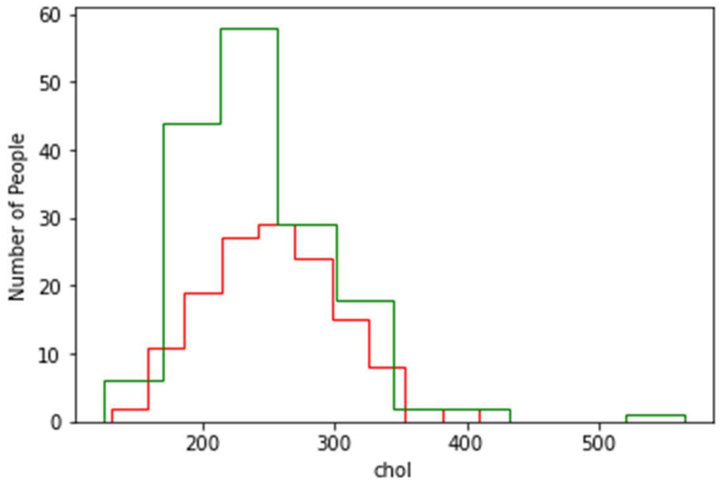


Fig. 12. Serum Cholesterol (chol) Comparison between Heat Disease (red) People vs. Well People (green)

The green line is the serum cholesterol distribution of well people. The red line is the distribution of the heart disease people. The result showed that the serum cholesterol did not seems to have a huge difference. The most serum cholesterol for well people was arount 250 mg /dl.

To confirm the hypothesis, Cumulative Distribution Functions (CDF) was used to test the hypothesis. If the two data set are the same, the CDF curve should be similar. The CDF curve is shown in Fig. 13.

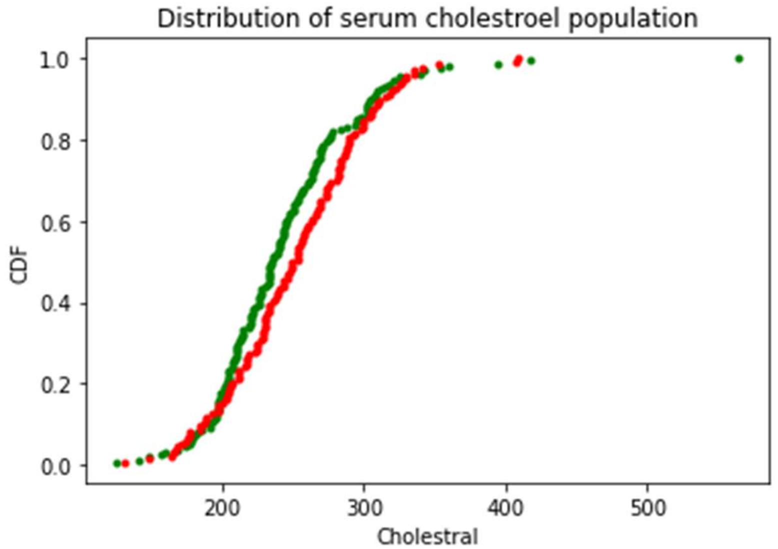


Fig. 13. The CDF of Serum Cholesterol (chol) Distribution of Heart Disease People (red) vs. Well People (green)

The results confirmed that two data sets were similar. The result confirmed that serum cholesterol (chol) did not show a significant difference between people with heart disease and well people.

Boxplot was also used to test the spreading and mean of the data. The result is shown in Fig. 14.

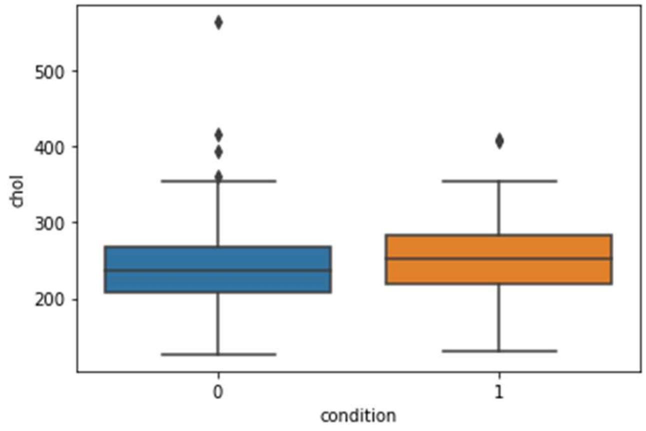


Fig. 14. The Boxplot of Serum Cholesterol (chol) Distribution of Heart Disease People (orange) vs. Well People (blue)

The results indicated that some data for well people (blue) and people with heart diseases (orange) were laying outside of the boundary. There was a slight difference of the mean between well people and people with heart disease

Based on the data analysis, the serum cholesterol may not be an indication of heart disease people by a single test.

6. Fasting Blood Sugar > 120 mg / dl (fbs)

The fasting blood sugar has two value. If the level is above 120 mg /dl, the value is 1. Otherwise, the value is 0. The comparison of fasting blood sugar of heart disease people vs. that of well people are shown in Fig. 15 and Fig. 16.

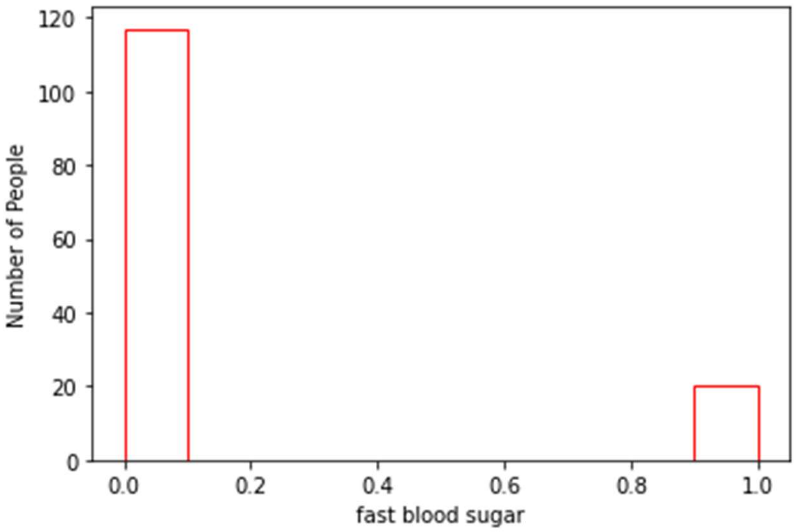


Fig. 15. Fast Blood Sugar > 120 mg /dl for Heart Disease People

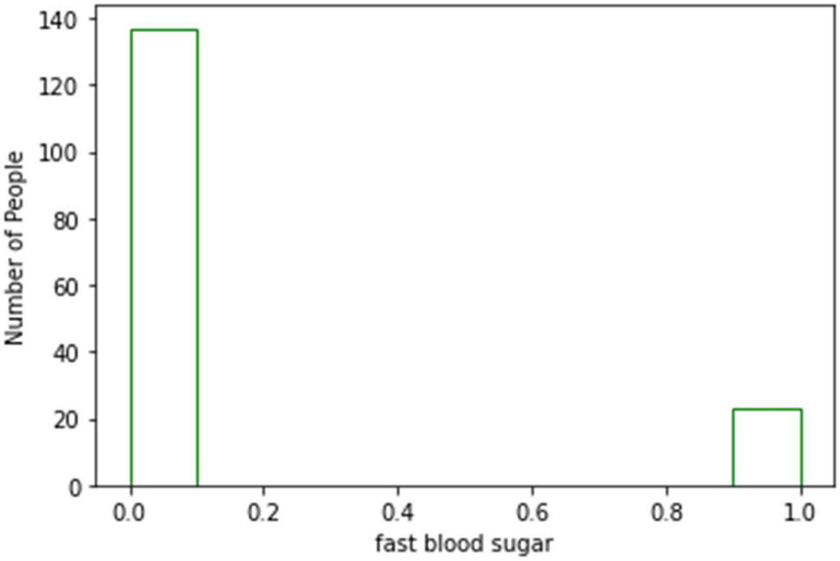


Fig. 16. Fast Blood Sugar > 120 mg / dl for Well People

The result showed that the distribution fast blood sugar level was almost the same for both heart disease peoples and well people.

7. Resting Electrocardiographic Results (restecg)

The resting electrocardiogram is a test that measures the electrical activity of the heart. There were three values. If the value was 0, it showed probable or definite left ventricular hypertrophy. If the value was 1, it was normal. If the value was 2, it showed that ST-T wave abnormal. The comparison of resting electrocardiographic are shown in Fig. 17 and Fig. 18.

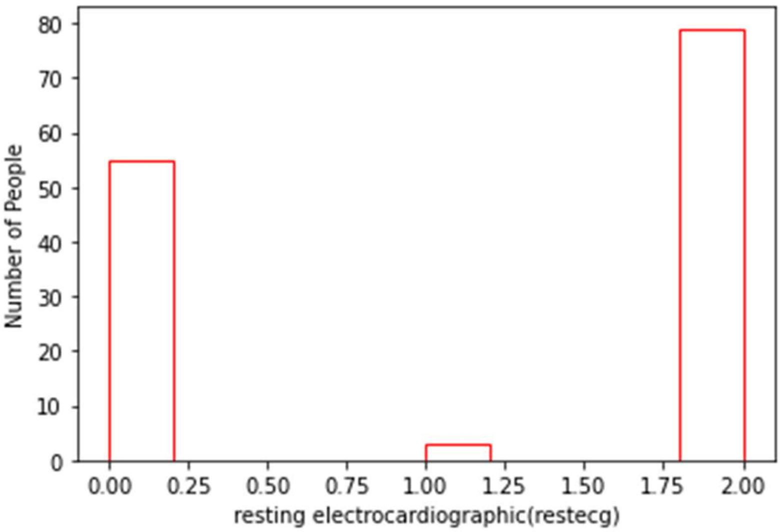


Fig. 17. Resting Electrocardiographic for Heart Disease People

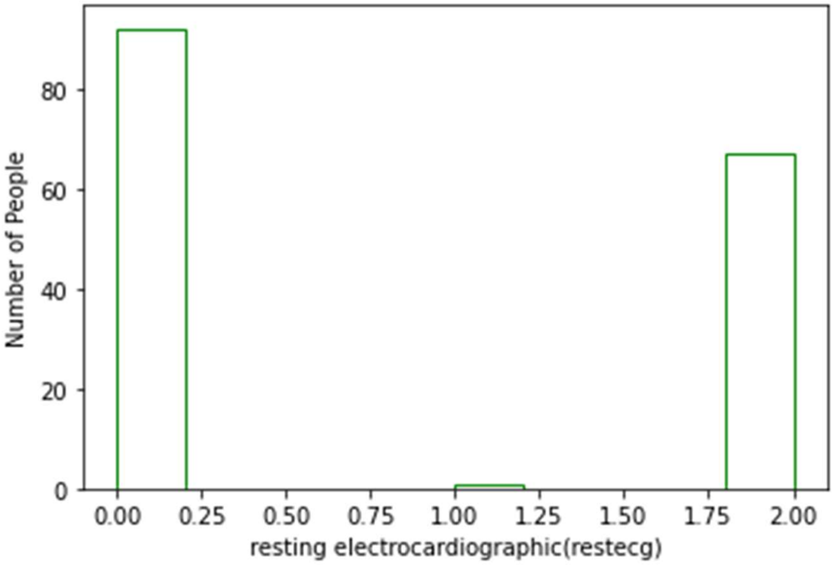


Fig. 18. Resting Electrocardiographic for Well People

The result did not seem to have a significant shift of resting electrocardiographic from heart disease people to well people

8. Maximum Heart Rate Achieved (Thalach)

The maximum heart rate comparison between heart disease people and well people is shown in Fig. 19.

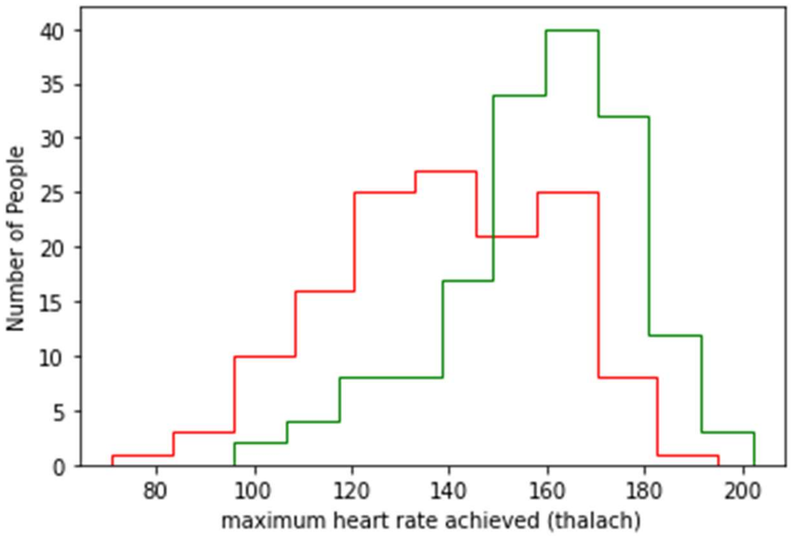


Fig. 19. Maximum Heart Rate Achieved (thalachl) Comparison between Heat Disease (red) People vs. Well People (green)

The green line is the maximum heart rate achieved distribution of well people. The red line is the distribution of the heart disease people. The result showed that the shift of distribution. The maximum heart rate achieved for well people was around 170. The maximum heart rate achieved for heart disease people was around 130.

To confirm the hypothesis, Cumulative Distribution Functions (CDF) was used to test the hypothesis. If the two data set are the same, the CDF curve should be similar. The CDF curve is shown in Fig. 20.

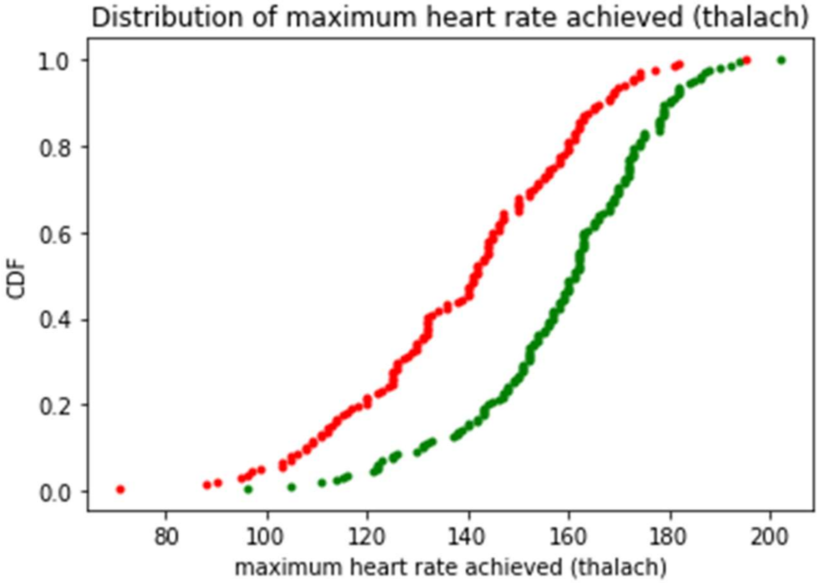


Fig. 20. The CDF of Maximum Heart Rate Achieved Distribution of Heart Disease People (red) vs. Well People (green)

The result confirmed that the maximum heart rate achieved data did not show a significant difference between people with heart disease and well people. The people with heart diseases tends to have a lower maximum heart rate.

Boxplot was also used to test the spreading and mean of the data. The result is shown in Fig. 21.

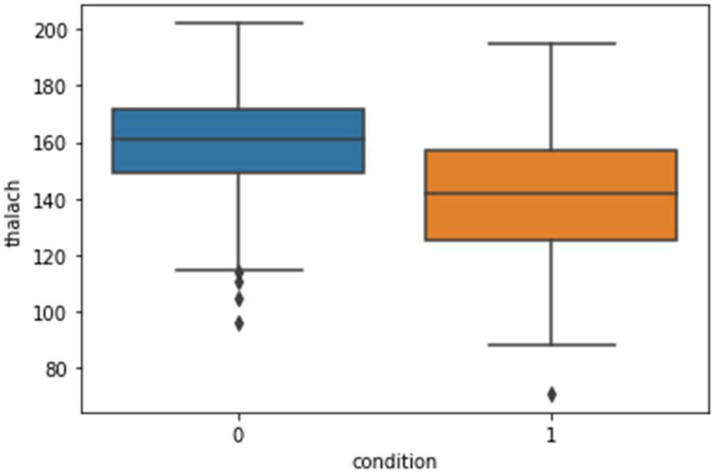


Fig. 21. The Boxplot of Maximum Heart Rate Achieved (thalach) Distribution of Heart Disease People (orange) vs. Well People (blue)

The results indicated that some data for well people (blue) and people with heart diseases (orange) were laying outside of the boundary. There was a significant difference of the mean between well people and people with heart disease. The mean for the well people was above 160. The people with heart disease was around 140.

Based on the data analysis, we conclude that heart disease people tend to have a less achieved maximum heart rate.

9. Exercise Induced Angina (exnag)

The Exercise Induced Angina (exnag) has two value. If it happened, , the value is 1. Otherwise, the value is 0. The comparison of Exercise Induced Angina of heart disease people vs. that of well people are shown in Fig. 22 and Fig. 23.

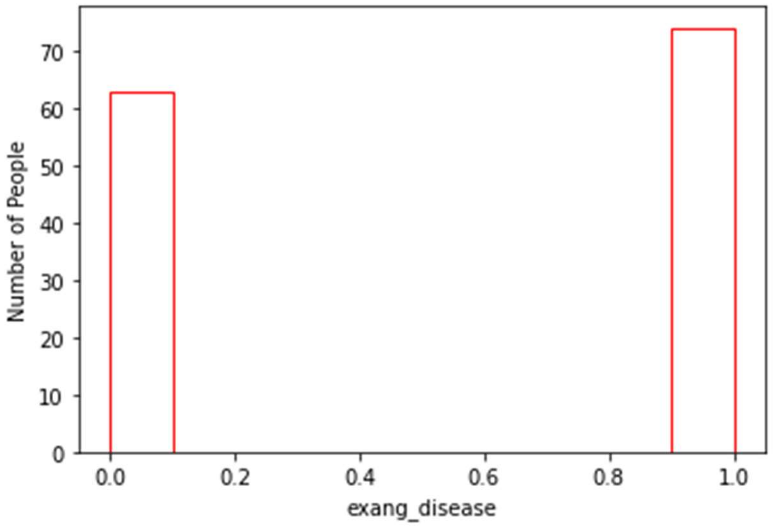


Fig. 22. Exercise Induced Angina (exnag) for Heart Disease People

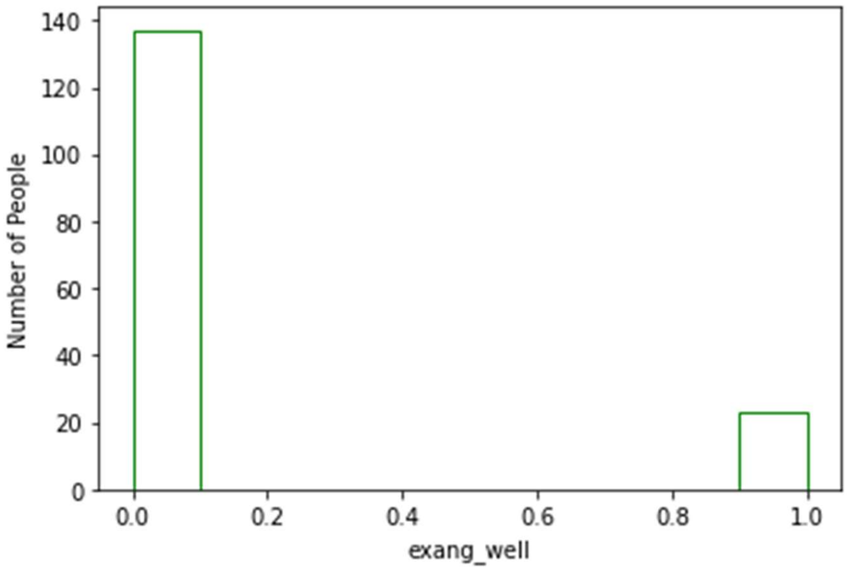


Fig. 23. Exercise Induced Angina (exnag) for Well People

The result showed that the exercise induced angina would likely happen for the heart di sease people.

10. ST Depression Induced by Exercise Relative to Rest (Oldpeak)

The ST depression induced by exercise relative to rest between heart disease people a nd well people is shown in Fig. 24.

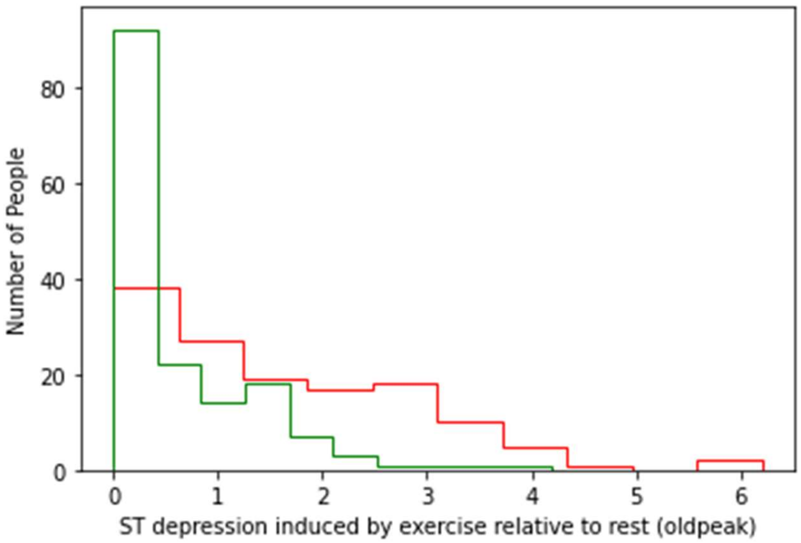


Fig. 24. ST Depression Induced by Exercise Relative to Rest (Oldpeak) Comparison between Heat Disease (red) People vs. Well People (green)

The green line is the ST depression induced by exercise relative to rest distribution of well people. The red line is the distribution of the heart disease people. The result showed that the shift of distribution. Most well people were 0. When people has heart disease, the data will spread out to high value, which indicated that people with heart disease would have higher potential to have ST depression induced by exercise relative to rest.

To confirm the hypothesis, Cumulative Distribution Functions (CDF) was used to test the hypothesis. If the two data set are the same, the CDF curve should be similar. The CDF curve is shown in Fig. 25.

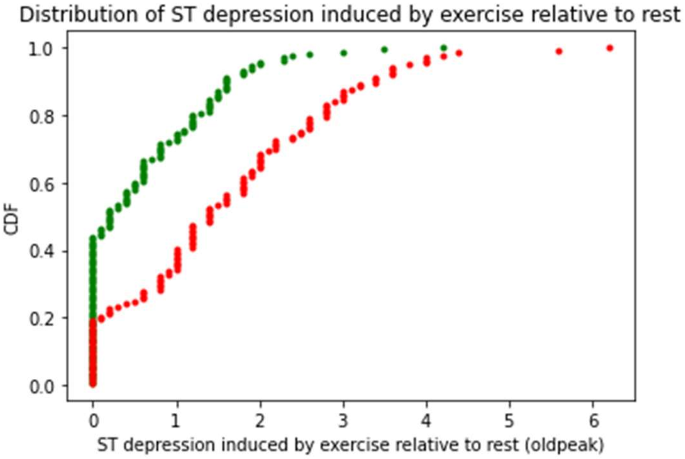


Fig. 25. The CDF of ST Depression Induced by Exercise Relative to Rest (Oldpeak) Distribution of Heart Disease People (red) vs. Well People (green)

The result confirmed that the ST depression induced by exercise relative to rest show a significant difference between people with heart disease and well people. The people wi th heart diseases tends to have a higher potential to have ST depression induced by exercise relative to rest.

Boxplot was also used to test the spreading and mean of the data. The result is shown in Fig. 26.

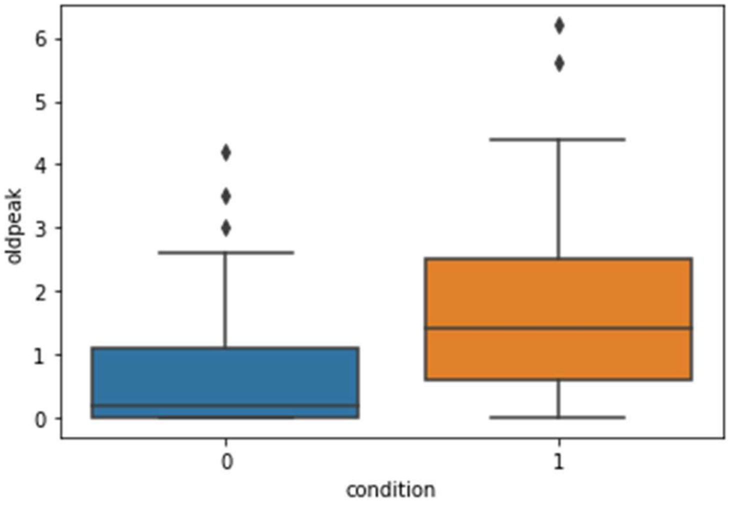


Fig. 26. The Boxplot of ST Depression Induced by Exercise Relative to Rest (Oldpeak) Distribution of Heart Disease People (orange) vs. Well People (blue)

The results indicated that some data for well people (blue) and people with heart diseases (orange) were laying outside of the boundary. There was a significant difference of the mean between well people and people with heart disease. The people with heart disease tend to have a higher potential of ST depression induced by exercise relative to rest.

Based on the data analysis, we conclude that heart disease people tend to have a higher potential for ST depression induced by exercise relative to rest.

11. The Slope of the Peak Exercise ST Segment (Slope)
The slope of the peak exercise ST segment has three value. If it was downsloping, , the value was 0. If it was flat, the value was 0, if it was upsloping, the value was 2. The comparison of the slope of the peak exercise ST segment of heart disease people vs. that of well people are shown in Fig. 27 and Fig. 28.

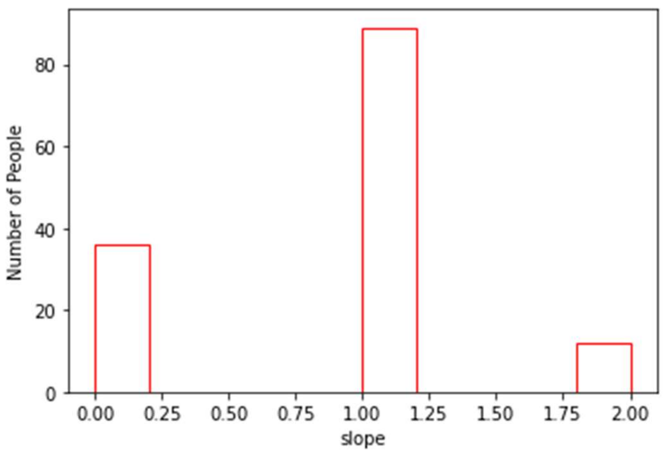


Fig. 27. The Slope of the Peak Exercise ST Segment (Oldpeaks) for Heart Disease People

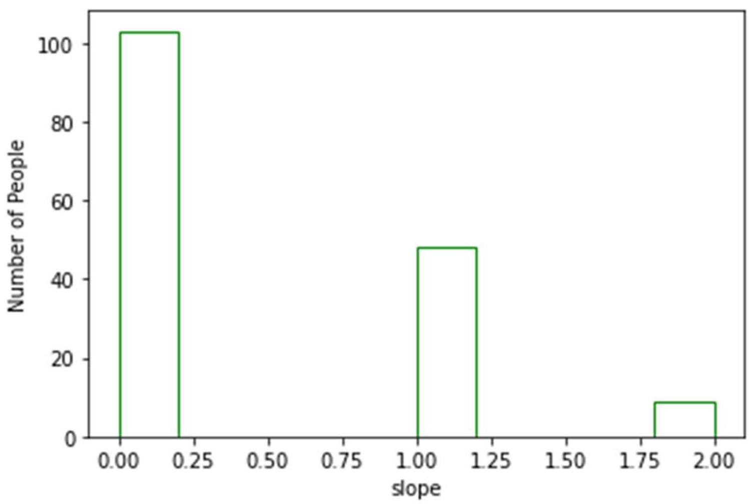


Fig. 28. The Slope of the Peak Exercise ST Segment (Oldpeaks) for Well People

The result showed that the distribution of slope of the peak exercise ST segment were flat for the people with heart diseases.

12. Number of Major Vessels Colored by Fluoroscopy (ca)

The comparison of number of major vessels colored by fluoroscopy of heart disease people vs. that of well people are shown in Fig. 29 and Fig. 30.

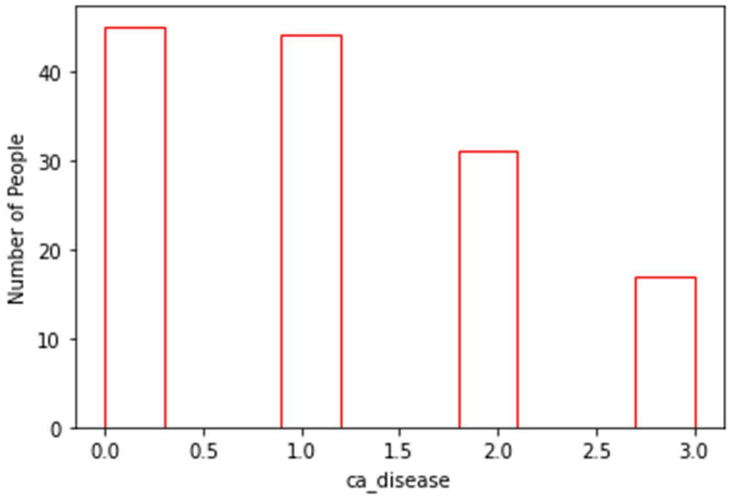


Fig. 29. Number of Major Vessels Colored by Fluoroscopy (ca)
for Heart Disease People

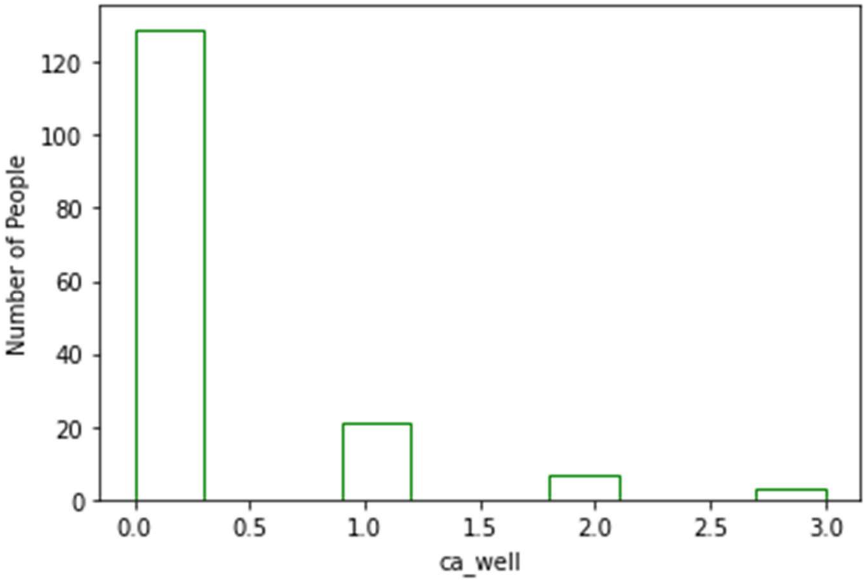


Fig. 30. Number of Major Vessels Colored by Fluoroscopy (ca)
for Well People

The result showed that the distribution of number of major vessels colored by fluoroscop y changed for the people with heart diseases. The number of major vessels colored by fl uoroscopy increased for the people with heart diseases.

13. Unknown Parameter (Thal)
The comparison of thal of heart disease people vs. that of well people are shown in Fig. 31 and Fig. 32.

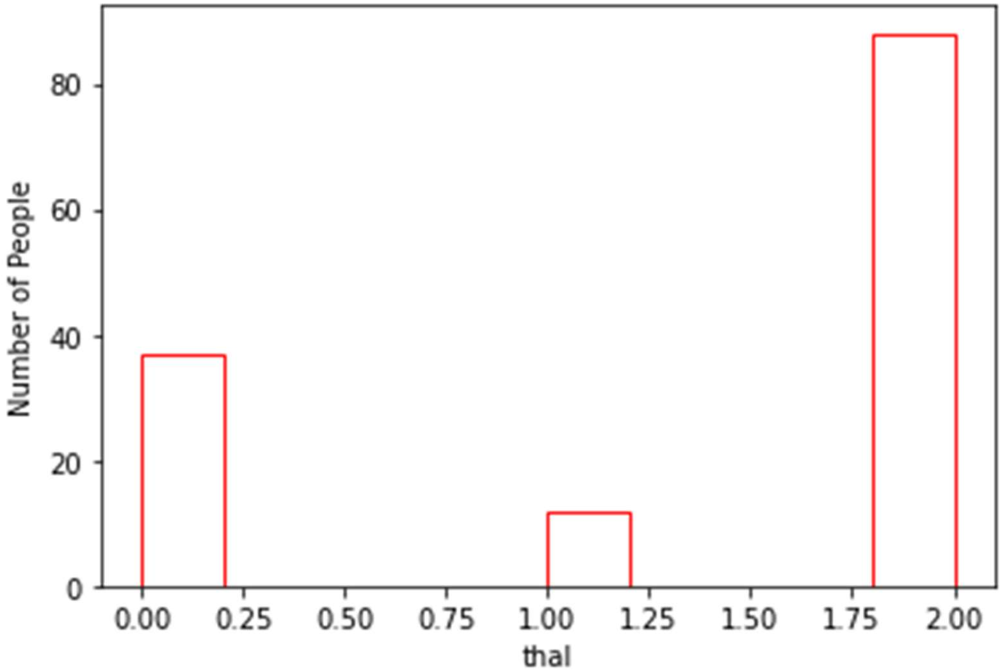


Fig. 31. Thal for Heart Disease People

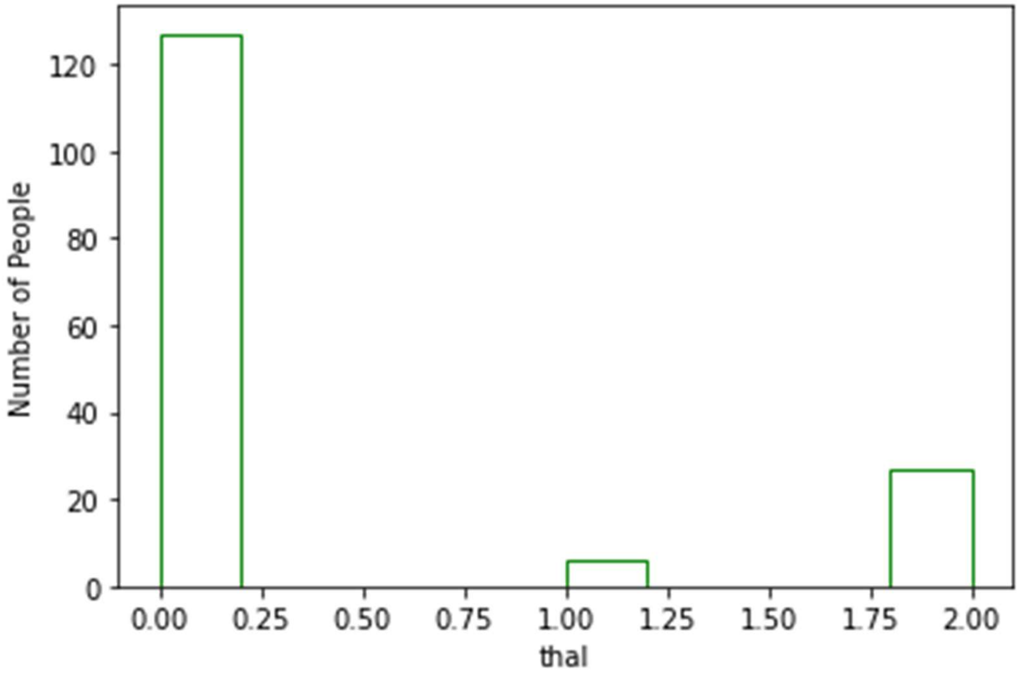


Fig. 32. Thal for Well People

The result showed that the distribution of thal changed for the people with heart diseases. Thal increased for the people with heart diseases.

14. Conclusion

Based on the data, we concluded the relationship of heart disease people with individual feature:

- Age related to people with heart disease. The potential age to have heart disease will around 59-60.
- Male would likely to have higher potential for heart disease.
- when a people have a typical angina (value=3), he / she likely to have a heart disease.
- Heart diseases people appeared to have a higher resting blood pressure.
- the serum cholesterol may not be an indication of heart disease people by a single test.
- Fast blood sugar level was almost the same for both heart disease peoples and well people
- Resting electrocardiographic from heart disease people vs that of well people did not seem to significantly shift
- Based on the data analysis, we conclude that heart disease people tend to have a less achieved maximum heart rate.
- The exercise induced angina would likely happen for the heart disease people.
- Heart disease people tend to have a higher potential for ST depression induced by exercise relative to rest
- The distribution of slope of the peak exercise ST segment were flat for the people with heart diseases
- The number of major vessels colored by fluoroscopy increased for the people with heart diseases
- The distribution of thal changed for the people with heart diseases. Thal increased for the people with heart diseases

Data Modeling

The data exploration explained the relationship of individual feature with heart diseases. However, as we discussed above, the heart disease is a combination of different features. The potential to have a heart disease will be high if there are more symptoms appears. To predict the potential to have a heart disease, models will be needed to combine different features, which is shown in Fig. 33. In the Fig. 33, the features were input through a Machine Learning Model (MLM). After processing through the model, the output will be either people with heart disease or well people. In general, there are two ways to set up a model. One is through supervised learning. The other is through unsupervised learning. The supervised learning is ‘train’ model with known data sets. Then the ‘trained’ model is used to predict the test data. The unsupervised learning is directly predicting the data without any training. In this project, we will use supervised learning. We will use three models (Logistic Regression, Decision Tree and RandomForest) to predict the potential of heart disease with different features. We will also evaluate the different model and see whether there will be any difference between these models. The models in this project were classification models.

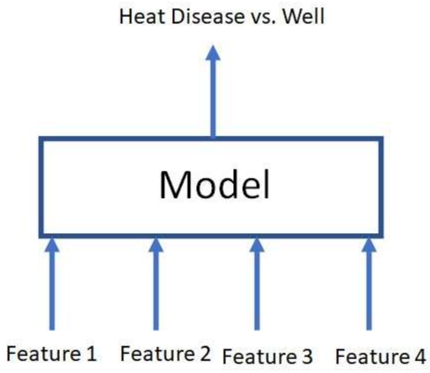


Fig. 33. Model

I. Data Preprocessing
The data was split as training data (80%) and test data (20%). The training data were used to 'train' model and tune the parameters within the model. The test data were used to validate the performance of model. All the data have been scaled to the same level for model prediction.

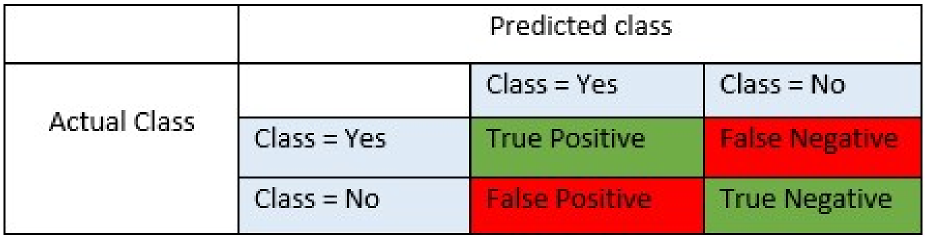
II. Logistic Modeling
Data Modeling is to train data with "training' data. Then cross validate the model with 'test' data set. The statistical analysis was used to see the effectiveness of the model. In this project, we used the following types of statistical method:

- accuracy score
- f-1 score
- confusion matrix

We also use the parameter to evaluate the importance of feature on the output.

A confusion matrix is a table. It is often used to describe the performance of a classification model on a set of test data, which the true values are known. The table is shown in Table 5

Table 5. Confusion Matrix



	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

True positive and true negatives are the observations, which are correctly predicted and therefore shown in green. The false positive and false negatives are the observations, which are not correctly predicted and therefore shown in red. In the modeling, we want to minimize false positives and false negatives.

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. In this project, both actual class and predict class indicates that there is a heart disease.
True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. In this project, both actual class and predict class indicates there is no heart disease.
False Positives (FP) – These are the incorrectly predict the positive value. When actual class is no and predicted class is yes. In this project, if actual class says there is no heart disease. But predicted class says there is heart disease.
False Negatives (FN) – These are the incorrectly predict the negative value. When actual class is yes but predicted class in no. In this project, if actual class value indicates that there is heart disease and predicted class says that there is no heart disease.
The accuracy, accuracy score and F1-score were calculated from the confusion matrix

Accuracy - Accuracy is the most intuitive performance measure. It measures the distance between a data and 'true' value. From the confusion matrix, it is simply a ratio of correctly predicted observation to the total observations. The calculation is shown below

Accuracy = (TP+TN) / (TP+FP+FN+TN)

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision usually relates to the low false positive rate. The calculation is shown below:

Precision = TP / (TP+FP)

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The value for accepted model is above 0.5. The calculation is shown below:

Recall = TP / (TP+FN)

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. if there is an uneven class distribution, F1 is usually more useful than accuracy. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are different, it's better to look at both Precision and Recall. The calculation of F1 score is shown below:

F1 Score = 2(Recall * Precision) / (Recall + Precision)*

1. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable(0 and 1). ‘1’ means pass. ‘0’ means fail. As the simplest regression analysis model, logistic regression normally used to estimate the parameters, which will affect the output variables.

After the model was trained, the parameters are shown in Table 6 and Table 7

Table 6. Parameters of Different Features

Age	Sex	cp	trestbps	chol	fbs	restecg
0.0373	0.506	0.710	0.357	0.154	-0.272	0.130

Table 7. Parameters of Different Features

thalach	Exang	oldpeak	slope	ca	thal
-0.250	0.372	0.259	0.373	1.088	0.720

The higher parameter in the table, the more important contribution for the feature to the output, which was the potential to have heart disease. Based on Table 5 and Table 6. The importance of features was shown as following:

ca > thal > cp > sex > slope > exang > trestbps > fbs> oldpeak > thalach > chol > restecg > age

The accuracy and F-1 score of the model is shown in Table 8

Table 8. Accuracy and F-1 score

Accuracy	F-1 Score
0.867	0.867

The normalized confusion matrix is shown in Fig. 33

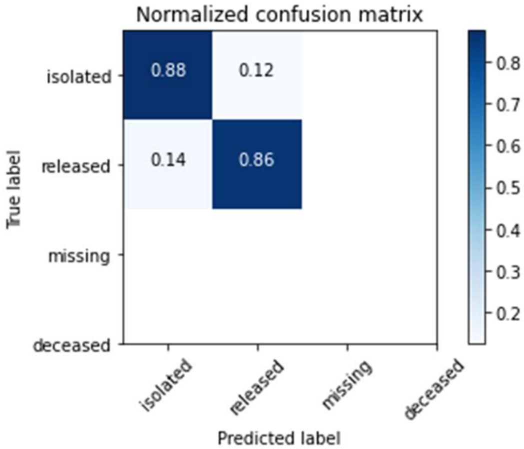


Fig. 33. Normalized Confusion Matrix

2. Decision Tree

A decision tree is a flowchart-like structure (see Fig. 34) in which each internal node represents a "test" on an attribute. The ‘test’ is like a coin flipping- head vs. tail. Each branch represents the outcome of the test. and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

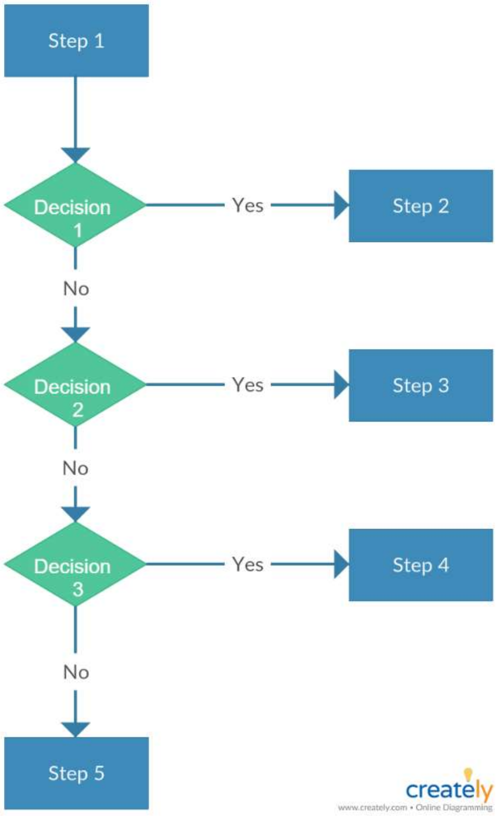


Fig. 34. Decision Tree Flow Chart

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes:

1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

Decision trees are commonly used in operations research and operations management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

The detail parameter for the decision tree in this project is shown below:

tree.DecisionTreeClassifier(criterion='entropy',max_depth=3,random_state=0)

The decision tree used 'entropy' model. There were 3 layers in the decision tree.

After the decision tree has been trained, the test data were input. The accuracy and F-1 score is shown in Table 8.

Table 8. Accuracy and F-1 Score

Accuracy	F-1 Score
0.833	0.833

The normalized confusion matrix is shown in Fig. 34

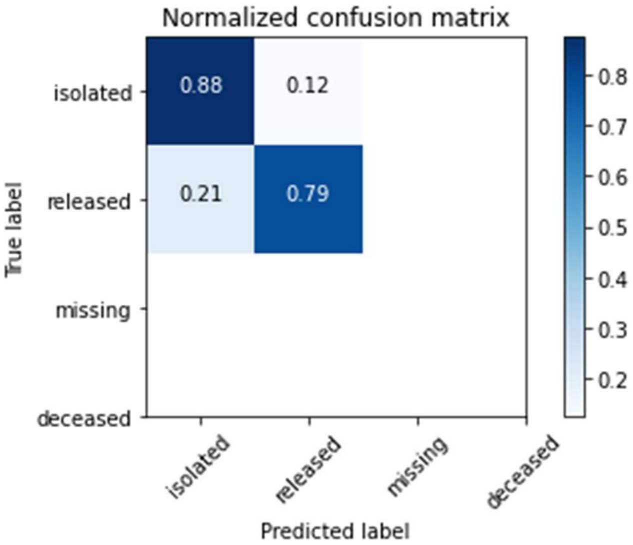


Fig. 34. Normalized Confusion Matrix

Based on the confusion matrix, the feature importance for the output is shown in Fig. 35

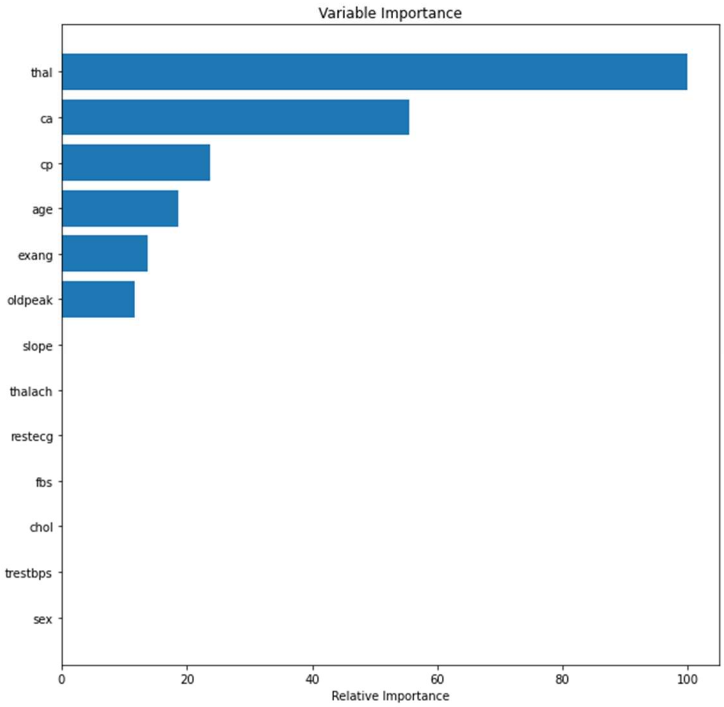


Fig. 34. The Feature Importance for Heat Disease

The results showed that thal > ca > cp > age > exang > oldpeak. The results was differed from Logistic Regression Model.

3. Random Forest Model

A decision tree is built on an entire dataset, using all the features/variables of interest, whereas a random forest randomly selects observations/rows and specific features/variables to build multiple decision trees from and then averages the results. After many trees are built using this method, each tree "votes" or chooses the class, and the class receiving the most votes by a simple majority is the "winner" or predicted class.
When using a decision tree model on a given training dataset the accuracy keeps improving with more and more splits. people can easily overfit the data and doesn't know when they have crossed the line unless using cross validation (on training data set). The advantage of a simple decision tree is model is easy to interpretA random forest is like a black box. It's a forest people can build and control. Although number of trees in the forest(n_estimators) and max num of features can be specified in each tree, the randomness of the forest cannot be controlled. For examples, which feature is part of which tree in the forest, which data point is part of which tree, etc.. Accuracy keeps increasing as the number of trees increase. However, it will become constant at certain point. Unlike decision tree, it won't create highly biased model and reduces the variance.

After the Random Forest Model (119 trees, each tree is ‘gini’) has been trained, the test data were input. The accuracy and F-1 score is shown in Table 9.

Table 9. Accuracy and F-1 Score

Accuracy	F-1 Score
0.833	0.833

The normalized confusion matrix is shown in Fig. 35

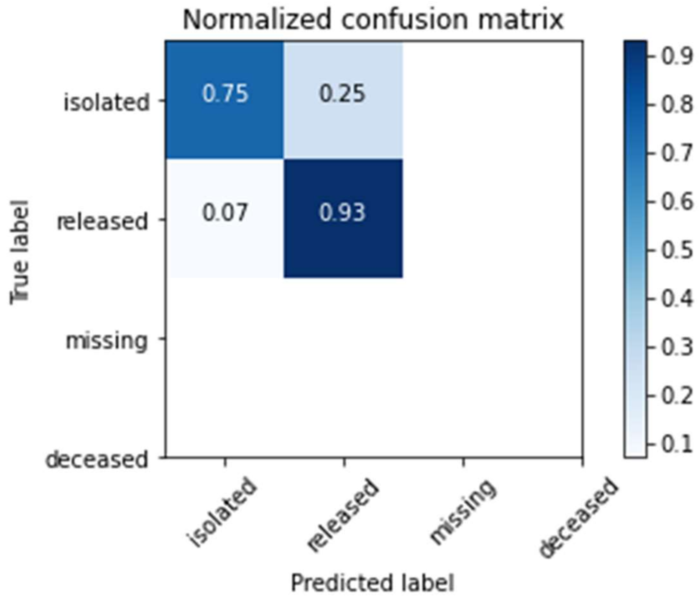


Fig. 34. Normalized Confusion Matrix

Based on the confusion matrix, the feature importance for the output is shown in Fig. 36

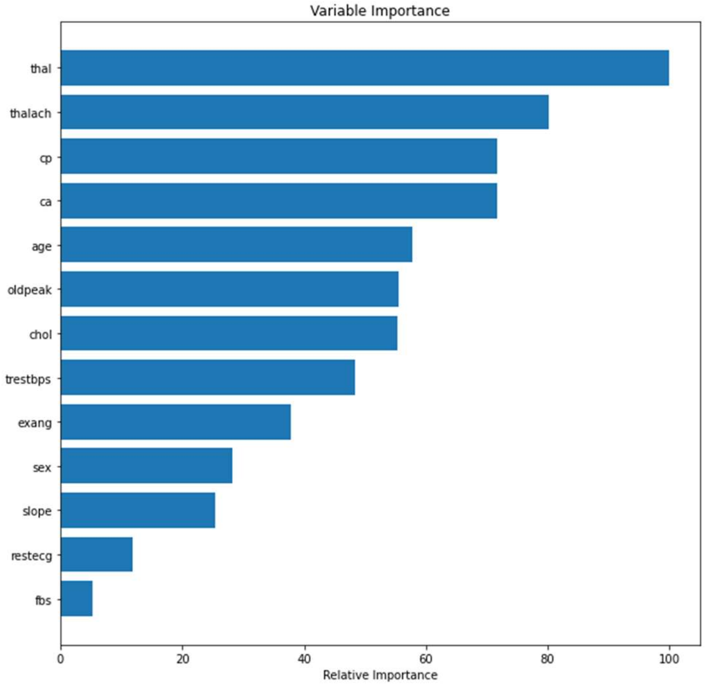


Fig. 36. The Feature Importance

The importance of features is shown as following:

RM model: thal > thalach > cp > ca > age > chol > oldpeak > tresbps > exang > sex > slope > restecg > fbs

III. Model Comparison

1. Accuracy and F-1 Score

We used three models to predict the potential of heart disease. The accuracy and F-1 score in different model are shown in Table 10

Table 10. Accuracy and F-1 Score of Different Model
Table 9. Accuracy and F-1 Score

Model	Accuracy	F-1 Score
Logistic Regression	0.867	0.867
Decision Tree	0.833	0.833
Random Forest	0.833	0.833

The results showed that Logistic Regression was a better model. Normally decision tree and random forest were used for a more complex data set. However since there was not a significant difference for 0.867 and 0.833, Random Forest Model will be recommended in case there will be more data added.

2. Feature importance

Different model provides a different feature importance. The feature importance for individual model is shown below:

Logistic Regression: ca > thal > cp > sex > slope > exang > trestbps > fbs> oldpeak > thalach > chol > restecg > age

Decision Tree: thal > ca > cp > age > exang > oldpeak

Random Forest: thal > thalach > cp > ca > age > chol > oldpeak > tresbps > exang > sex > slope > restecg > fbs

Conclusion

In this project, we used three models to predict the potential for the heart disease. The result showed that there was no significant difference between these three models. However, we would like to recommend Random Forest Model.
The results also showed that the feature importance also depend on the model.