Heart Disease

Member 1 name Junxi Chen
Member 2 name Zhongwen Liang
Member 3 name Yizhang Huang

Member 1 Contribution:

-use the logistic regression method to test the data set for the important variables, to see the coefficient relationship.
-Design the structure and content for methodology. Provide some ideas for the result and interpretation.
-communicate and brainstorm.

Member 2 Contribution:

-Write the introduction and methodology
-Interpreting/analyze graphs with words
-Find ways to visualize the structure and format of the PDF
-Communicating, gathering all the code and combining into one file
-Markdown

Member 3 Contribution:

-Use python to show correlation matrix.
-Use cross validation to find the best parameter for k nearest neighbors.
-fit knn and logistic lasso regression.
-Calculate the confusion matrix and accuracy for knn and logistic lasso algorithms.
-Writing results and interpretation

Introduction

Our data:

For this project, we will be using the heart disease data set from Kaggle https://www.kaggle.com/johnsmith88/heart-disease-dataset. This data set includes age, sex, chest pain type, resting blood pressure, serum cholesterol, and so on with a total of 14 columns. The first 13 columns are our experiment variables consisting of binary and numeric variables. The last column named "target" is our response variable y which is a binary variable with 1 meaning the individual has heart disease and 0 meaning the individual does not have heart disease.

Our Intention:

Cardiovascular disease has been the leading cause of death worldwide for the past 20 years, with heart disease accounting for nearly 9 million deaths, or 16 percent of all deaths, through 2019. We want to use the data set from UCI to do a simple heart disease prediction model by using statistical learning methods such as classification and regression. As individuals who are able to develop heart disease in the future, we want to find out which variable or variables have greater effects on developing heart disease so that we could try to prevent it from happening and minimize the chance of it occurring to us. After our research on this topic, we can also apply our learning to some elderly people for the same reason as well.

Methodology

For this project, we use a dataset that includes information about heart disease. This data all uses numerical variables to show the information. This includes age, sex, ST depression induced by exercise relative to rest, and so on. We will use the data set from UCI to perform a simple heart disease prediction algorithm by using statistical learning methods including K-nn classification, cross validation, and logistic regression. Moreover, we want to determine the importance of each variable and the relationship between each variable with heart disease. To be more specific, we first find the correlation table for each variable. Then, as we can see from the graph, we primarily target a few variables. Then we will use K-nn and logistic lasso to make predictions. Furthermore, we use the confusion matrix to try to determine the correct rate of prediction. And we use the original table to compare with the table that only includes the important variables. Our null hypothesis is that the variables, sex, age, cp, and thalach have an influence on whether the target has heart disease or not.

Implementation

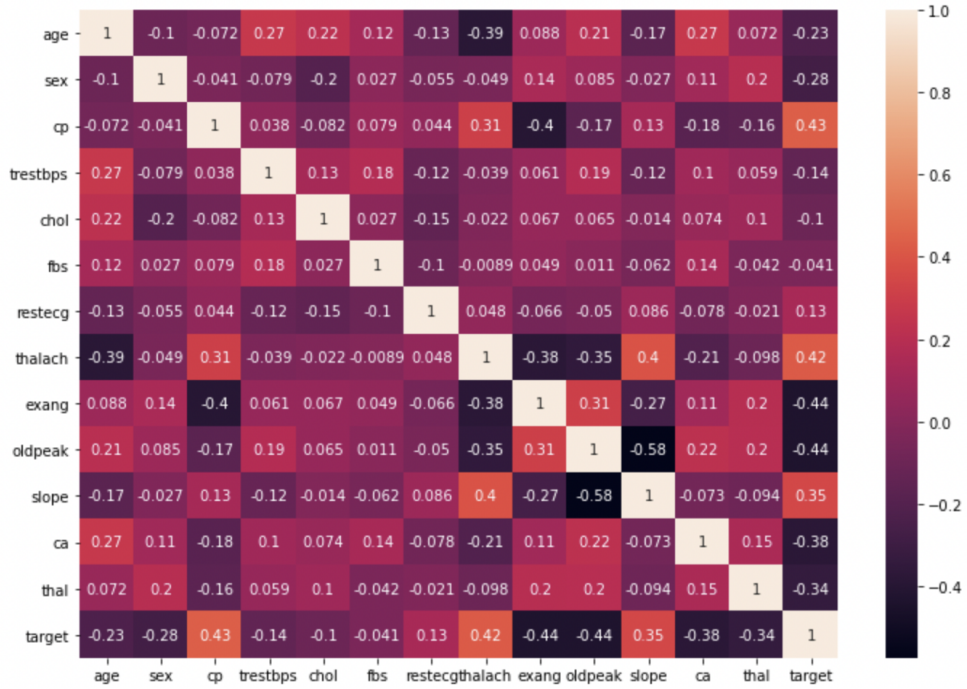
This is our heart disease data

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

1025 rows × 14 columns

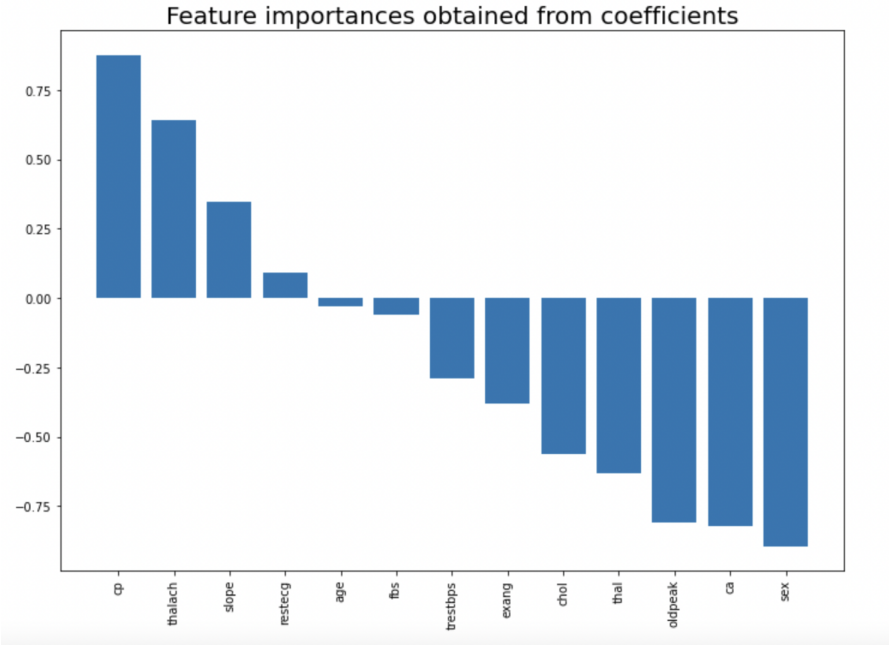
The first 13 columns are the experiment variables and the last column “target” being our response variable. None of our columns contains categorical variables and they are all numeric variables.

This is our correlation matrix



As we can see from the correlation matrix, for the target variable, the exang, oldpeak, thalach, and cp are highly correlated. Then we use the logistic regression method to test it.

This graph displays the importance of the experiment variables



Looking at the graph, we can see that the variables cp, thalach, sex, oldpeak, and ca all have importance on our response variable target. All five of them have more than 50% influence either positive or negative. We believe that they must have a strong influence on the result variable. Next, we decided to keep only these 5 variables and drop the rest to determine the influence on the target.

Data after dropping non significant variables

	sex	cp	thalach	oldpeak	ca	target
0	1	0	168	1.0	2	0
1	1	0	155	3.1	0	0
2	1	0	125	2.6	0	0
3	1	0	161	0.0	1	0
4	0	0	106	1.9	3	0
...	...	...	...	...	...	...
1020	1	1	164	0.0	0	1
1021	1	0	141	2.8	1	0
1022	1	0	118	1.0	1	0
1023	0	0	159	0.0	0	1
1024	1	0	113	1.4	1	0

1025 rows × 6 columns

Confusion matrix for knn method:

```
print(confusion)
print(kacc)
```

```
[[270 137]
 [ 78 335]]
0.7378048780487805
```

```
print(confusion_r)
print(kacc_r)
```

```
[[310  97]
 [ 69 344]]
0.7975609756097561
```

Confusion matrix for logistic lasso method:

```
print(confusion2)
print(lacc)
```

```
[[305 102]
 [ 57 356]]
0.8060975609756098
```

```
print(confusion2_r)
print(laccr)
```

```
[[320  87]
 [ 60 353]]
0.8207317073170731
```

The images on the left are the confusion matrix and prediction accuracy before dropping non-significant variables and on the right is after dropping variables. As we can see, the former with knn method gives us a 73.78% accuracy rate and the latter gives us a better performance with a 79.76% accuracy rate. And the former with the logistic lasso method gives us an 80.60% accuracy rate and the latter gives us a better performance with an 82.07% accuracy rate.

Results and Interpretation

Looking at our finding by correlation table and logistic regression, we can see that for this heart disease dataset, heart disease exist related to sex, cp, thalach, oldpeak, and ca. To be more specific, those variables mean sex, chest pain type, maximum heart rate achieved, ST depression induced by exercise relative to rest, and the number of major vessels (0-3) colored by fluoroscopy. In our predicting algorithms, both K-nn and logistic lasso give us better predictions after dropping non-significant variables. One interesting fact about these two methods is that different data splitting would affect the results of the Knn method largely but not very impactful with the lasso method. When we split the data with 50% training and 50% testing(the above uses a 20% training 80% testing split), Knn would have a prediction accuracy of 92% after dropping variables since it would have more neighbors for prediction. However, since the lasso method has a penalizing term to reduce error, thus different data splitting wouldn't affect the algorithm as much. With these findings, we can conclude that our null hypothesis is that sex, age, cp, and thalach do have an impact on whether an individual has heart disease or not.