

# Final project output

Hongkai Wang

2022-11-30

## Introduction

Heart disease is the leading cause of death for all demographic groups in the United States. On average, one person would die from some form of cardiovascular conditions every 34 seconds, and heart diseases account for one out of every 5 deaths in the US in 2020. Given the severity and prevalence of heart diseases in the United States, it is important to develop accurate diagnostic approaches and provide strong guidance in the prevention, treatment, and recovery of heart diseases. One of the possible way of improving the diagnostic process and providing accurate prevention measures is by utilizing machine learning algorithms to predict the development of heart diseases given relevant clinical parameters. In this project, I will attempt to create multiple machine learning models that could predict the presence of heart diseases based on several common clinical measurements.

A data set was sourced from a page named Heart Failure Prediction Dataset. The data set contains 11 common clinical features related to cardiovascular health on a total of 918 patients. The patient information were compiled from a collection of heart disease data sets posted on the UCI Machine Learning Repository. This data set contains information sourced from Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog, and could be considered as a comprehensive description of global heart diseases with a fairly small data set. Attempts were made to increase the size of the available data, however, there is limited publications with detailed patients data. The current data set used for the project was the largest data set I was able to acquire.

The data set is composed of 11 clinical parameters and 1 target feature indicating the presence of heart diseases in a particular patient sample. The 11 clinical features include: age, sex, type of chest pain experience (includes Typical Angina, Atypical Angina, Non-Angina pain, no chest pain), resting blood pressure, level of serum cholesterol measured in millimeters of Mercury, level of fasting blood sugar (denoted 1 if fasting blood sugar  $\geq 120$  mg/dl, 0 otherwise), resting electrocardiogram results (categorized by Normal, ST-T wave abnormal, Left Ventricular Hypertrophy), maximum heart rate achieved, presence of exercise-induced angina (reported as yes or no ), numerical value of ST value (denoted as old peak), slope of peak exercise ST segment (defined as upsloping, flat, or downsloping). There is a total of 6 categorical predictor covariates, and 5 continuous numerical predictor covariates. Here is a name of each column defined in the data set, in the order shown above.

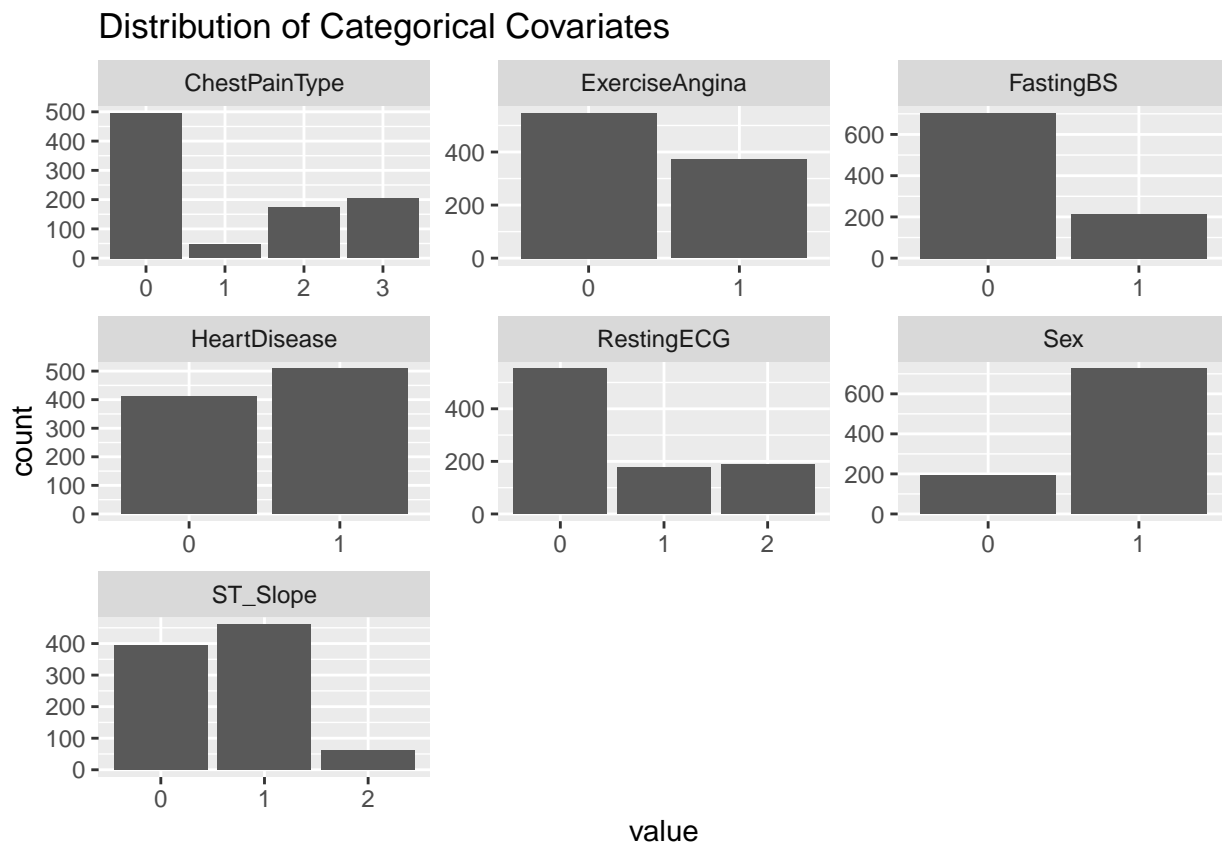
##	[1]	"Age"	"Sex"	"ChestPainType"	"RestingBP"
##	[5]	"Cholesterol"	"FastingBS"	"RestingECG"	"MaxHR"
##	[9]	"ExerciseAngina"	"Oldpeak"	"ST_Slope"	"HeartDisease"

Data transformation was conducted on all categorical covariates into the following numerical coding.

1. sex: female = 0, male = 1
2. ChestPainType: Asymptomatic = 0,
3. RestingECG: Normal = 0, ST = 1, LVH = 2

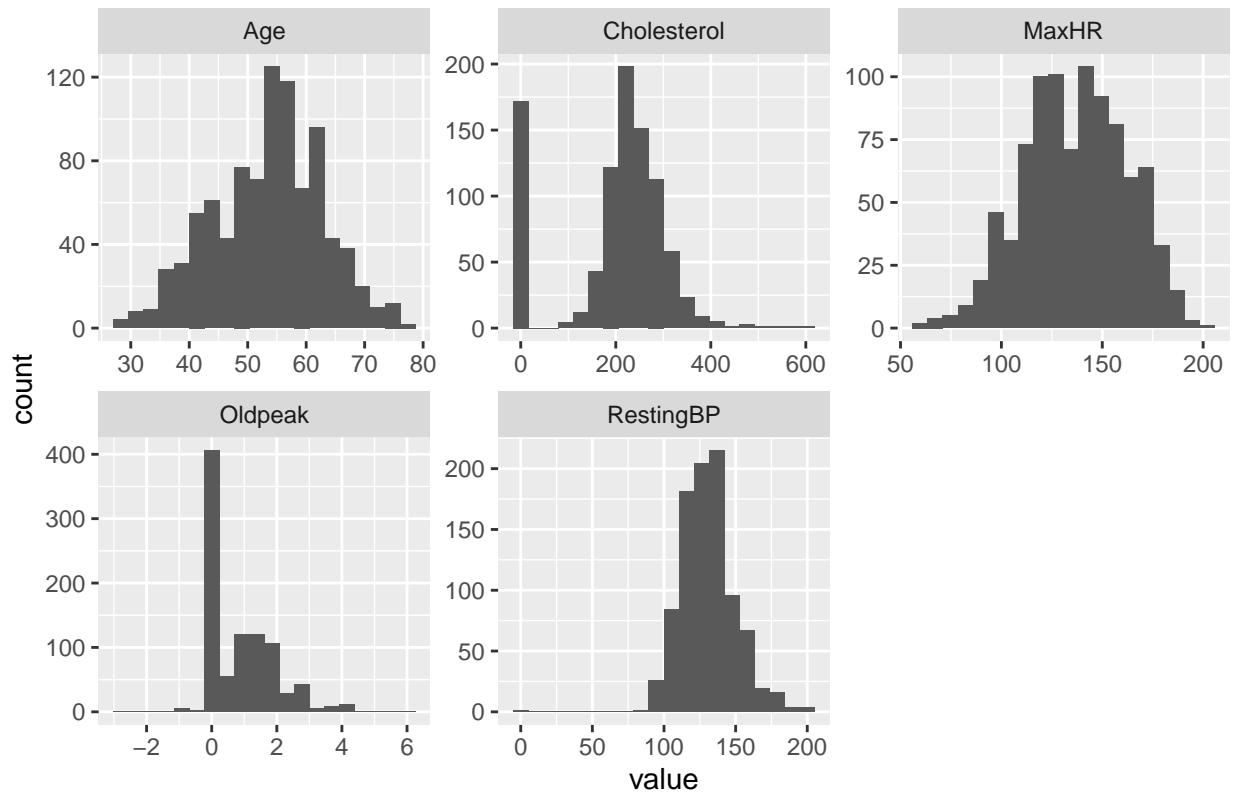
4. ExerciseAngina:  $N = 0$ ,  $Y = 1$
5. ST\_Slope: up = 0, flat = 1, down = 2

Bar graphs were constructed to reveal the distribution of the categorical predictor covariates. The results shows that the binary outcome of heart diseases is balanced, and there is not need for re sampling to balance the data set. There are two categories that is somewhat concerning, we see that ChestPainType and ST slope both have categories that has less than 100 samples in their categories. However, since they are not the outcome classification, I will hold off on changing any of the existing data set.



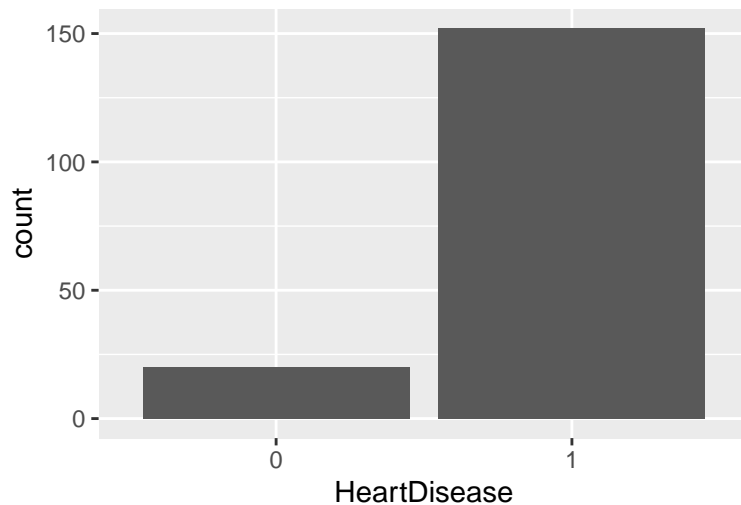
Histograms were constructed to reveal the distribution of the continuous numerical variables. Age, MaxHR, and restingBP all appear to be normally distributed with only a few outlier. There is a significant amount of 0 values in the cholesterol categories, and that was deemed to be missing data for some samples. These missing data will be removed from the machine learning analysis due to a lack of information on possible ways to fill these in. A possible approach is to fill the data with some form of regression analysis. However, that could affect our classification of heart disease, as there is no clear explanation on underlying reason for missing data. At the same time, we see that the old peak covariates is mostly 0, probably due to the fact that it is a data not collected in all of the source studies. Thus, I decided to exclude old peak from further analysis.

## Distribution of Numerical Continuous Covariates



Further analysis of the data shows that most of the missing data are heart disease patients, and data filling techniques could result in bias in our classification model. Thus, data samples with missing cholesterol levels are removed from the data set.

## Classification missing cholesterol levels



Another round of EDA was conducted after removing the missing values, and no significant changes in data distribution were observed. Normalization was then conducted on the continuous variables to achieve better machine learning outcomes. The plots for the final distribution of the data are included in the appendix.

## Methods

This project will utilize several machine learning classification methods to conduct predictions on the presence of heart disease based on the clinical parameters. Cross validations will be conducted before testing of the machine learning algorithms to create test and train data sets. Cross validation is essential in effective machine learning as it can significantly reduce overfitting and gives us a proper approach of evaluating our fitted models. There are several machine learning algorithms that could be useful for this project. A logistic regression model provides a foundation and baseline for our project. Logistic regression is chosen as it is one of the tried and true methods for classifying binary outcomes. The excellent explainability of the model provides important insight into the factors that has a strong influence on heart diseases. K-Nearest Neighbor is another helpful algorithm in data classification through data clustering. This method is used due to ease of fitting a model, and easy hyper parameter tuning with the value of k. Random forest is another helpful ML algorithm for this project. Random forests utilizes bagging in regression trees to reduce the usual issue of high variance in decision trees.

In summary, the three methods that I have chosen for this project, logistic regression, KNN, and random forests, cover three distinct type of classification methods, which are regression analysis, clustering, and decision tree based classification. It is a goal of this project to explore the advantage and disadvantage of utilizing different types of machine learning algorithms for medical data.

## Results

A training-test 8:2 data split was conducted before any model fitting processes. There was no significant deviation from the distribution of the original data set. Visual inspection of the data distribution is included in the code section.

### logisitic regression

A logistic regression model was first fitted onto the training data. Different combination of the covariates were tested for a best fitted model, and several covariates didn't produce a statistically significant fit for our model. Several classification threshold for the probability output is also tested. Interestingly, cholesterol was proven to be a insignificant predictor within the full model, but it is a significant predictor in a reduced model consisted only of itself. This possibility due to a stronger predictive power from other covariates that contains similar information to the cholesterol level. As a result, cholesterol was not included in the final model. The full model summary of the final model is shown below

```
##
## Call:
## glm(formula = HeartDisease ~ Age + Sex + ChestPainType + ExerciseAngina +
##      ST_Slope, family = "binomial", data = training_set)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.5968  -0.4028  -0.1184   0.4375   2.9547
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.9747    0.4033  -7.375 1.64e-13 ***
## Age             0.3602    0.1360   2.649 0.008073 **
## Sex1            1.9953    0.3384   5.895 3.74e-09 ***
## ChestPainType1 -1.1888    0.5113  -2.325 0.020081 *
## ChestPainType2 -1.5716    0.3817  -4.117 3.84e-05 ***
```

```
## ChestPainType3    -1.4555      0.3269   -4.452 8.51e-06 ***
## ExerciseAngina1   1.2052      0.2865    4.207 2.59e-05 ***
## ST_Slope1         2.8793      0.2885    9.980 < 2e-16 ***
## ST_Slope2         2.0729      0.5337    3.884 0.000103 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 826.40  on 596  degrees of freedom
## Residual deviance: 394.48  on 588  degrees of freedom
## AIC: 412.48
##
## Number of Fisher Scoring iterations: 5
```

We can explore the predictability of the model by looking at the confusion matrix comparing the prediction outcomes of the test data set.

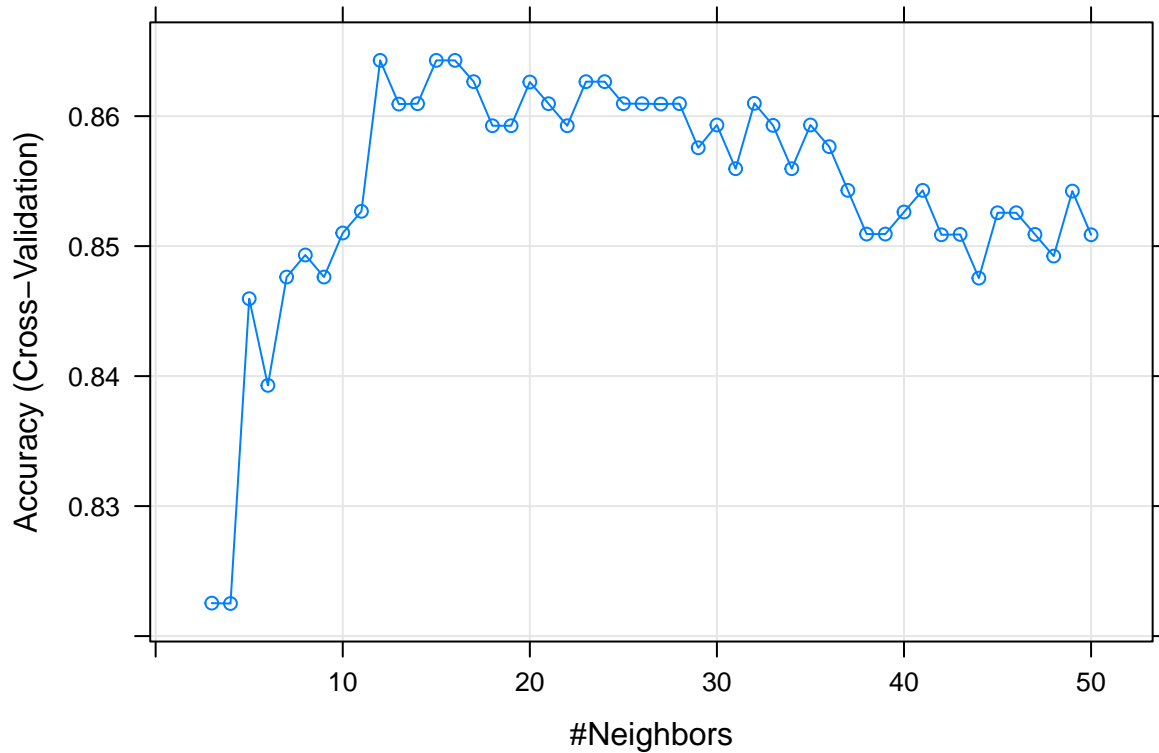
```
##           Reference
## Prediction  0  1
##           0 71 13
##           1  7 58

##           Accuracy           Kappa  AccuracyLower  AccuracyUpper  AccuracyNull
## 8.657718e-01 7.299257e-01 8.003330e-01 9.160427e-01 5.234899e-01
## AccuracyPValue McNemarPValue
## 7.550115e-19 2.635525e-01
```

From the confusion matrix output, we see that logistic regression model achieves an respectable accuracy of 86.5%. However, the model is very likely to make false negative predictions, meaning that we have a fairly low recall score. From a clinical point of view, this is not very helpful as false negative will do significant damage to the diagnostic prediction process. On the other hand, false positive is not as bad, as further diagnosis by a physician could clear up confusion and perhaps serve as precautionary tale for the patients. One way to increase the recall of the model is to lower the classification threshold. However, in this case, even with the classification threshold set to 0.1, we are still getting a false positive rate of 0.13. Thus, the logistic regression might not be the most appropriate model of choice here for predicting the presence of heart disease.

## K-Nearest Neighbor

Next, we are going to see if clustering methods provide a better result when compared to the logistic regression model. The K-Nearest neighbor model was established with an internal cross validation process and the test/train ratio was set to 1:9. All 10 remaining covariates were fed into the model in the hopes of finding new correlations between the clinical parameters and heart diseases. However, comparison of different input parameters shows that it is better to stick with the covariates that was significant in the previous logistic regression model and we can achieve better model accuracy. K values ranging from 3 to 50 were used to create different KNN models.



We can see that the highest accuracy of the model is obtained at  $K = 8$ , and any further clustering would only increase the noise in the model and reduce accuracy. Now let's evaluate the model outcome with the confusion matrix again.

```
##           Reference
## Prediction  0   1
##           0  65   8
##           1  13  63
```

```
##           Accuracy           Kappa  AccuracyLower  AccuracyUpper  AccuracyNull
##  8.590604e-01  7.183872e-01  7.926613e-01  9.105905e-01  5.234899e-01
## AccuracyPValue McNemarPValue
##  4.271252e-18  3.827331e-01
```

We can see that the KNN method has a similar overall accuracy to the logistic regression model. However, we see some improvement in the recall of the model. In a clinical setting, this model outcome is more preferable than the logistic regression model. A lower false negative rate would result in a lower number of missed heart diseases.

## Random Forest

```
## Random Forest
##
## 597 samples
## 10 predictor
```

```

## 2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 478, 477, 477, 478, 478
## Resampling results across tuning parameters:
##
## mtry Accuracy Kappa
## 1 0.8425490 0.6836438
## 5 0.8592997 0.7181363
## 10 0.8492297 0.6981143
## 25 0.8342017 0.6678724
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 5.

## Reference
## Prediction 0 1
## 0 66 13
## 1 12 58

## Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull
## 8.322148e-01 6.634746e-01 7.623590e-01 8.883770e-01 5.234899e-01
## AccuracyPValue McNemarPValue
## 2.599543e-15 1.000000e+00

```