Wilfrid Laurier University

ST494/694 Final Project

Department of Mathematics

Application of Statistical Learning on Heart Disease Data

Authors

Nora Fan

Tianliu Fu

Yeemui Lau

Xiaocong Lian

Instructor

Dr. Devan Becker

Apr 18, 2023

Abstract

Heart disease is a significant health concern and remains the leading cause of death globally. Early diagnosis and treatment are essential in preventing the progression of the disease. This report aims to utilize statistical learning techniques and non-invasive test results to predict the presence of heart disease. The data used was collected from patients referred for coronary angiography, and three non-invasive tests were performed. The report focuses on the accuracy of predictions and introduces unsupervised techniques and a new boosting algorithm to improve the model's accuracy. Furthermore, the sensitivity and specificity of the model are reviewed. Data cleansing and pre-processing, statistical learning techniques, and model selection were also been discussed.

The results showed that the Random Forest model combined with PCA achieved the highest accuracy on validation set, and also had the best performance in predicting heart disease. After comparing the results of test and validation set, the conclusion and limitations of the study are discussed in detail. In addition, the trade-off between correctly identifying patients with heart disease and predicting wrongly for some healthy patients should also be considered.

Contents

1 Introduction

Heart disease, also known as cardiovascular disease, is a condition that affects the heart and blood vessels. It is a major public health issue and the leading cause of death worldwide, accounting for approximately 17.9 million deaths annually [1]. Early diagnosis and treatment of heart disease are crucial in preventing its progression and reducing its impact on patients' quality of life.

The target of this project is to apply statistical learning techniques to predict if a patient was suffering from heart disease using non-invasive test results to replace the need to take invasive tests. Non-invasive tests are preferred because they do not require any surgical procedure or incision, which minimize the risk of complications and reduces the recovery time for patients. The data we used was collected at the Cleveland Clinic for patients referred for coronary angiography, an invasive test [5]. All the participating patients did not have a history or evidence of prior heart diseases. The patients took three non-invasive tests, including exercise electrocardiogram, cardiac fluoroscopy and thallium scintigraphy, as part of the research protocol. Our project is to use these test results and other patients' general information to predict heart disease and then compare the actual result from invasive test, angiography.

We will focus on prediction accuracy in selecting the best model fit for the data. Apart from the statistical learning techniques we learnt from the textbook [4], we will explore to combine unsupervised techniques with supervised techniques and introduce new boosting algorithm to improve the model accuracy. On the other hand, sensitivity and specificity will also be reviewed as these measure how good our model is in predicting with and without heart disease.

In this report, Section 2 covers the details of the data and data cleansing/pre-processing performed. Section 3 discusses the statistical learning techniques employed in fitting the data as well as the process to choose best tuned parameters for each model. The best tuned parameters on test set are discussed in Section 4, and then, conclusion and limitation of our study are in Section 5.

2 Data

The data was downloaded from Kaggle [2], where the underlying data source was from UCI Machine Learning Repository [3]. The data was collected at the Cleveland Clinic between year 1981 and 1984 for 303 consecutive patients. The dataset consists of 76 attributes/features, only 13 features are included in our analysis, similar to what other scholars did, as other attributes, such as timing of ECG reading and exercise protocol, expected to have little effect on heart disease.

The patient was considered as suffering from heart disease if the diameter of at least one of the major vessels was narrowed by more than 50% as shown by angiography. The data field *condition* in the dataset stores such information with 0 = no heart disease and 1 = with heart disease. This variable is simplified from the original dataset, in which the number of major vessels with diameter narrowed by more than 50% was recorded.

Figure 1 on next page summarizes the 13 features in more details. Items 1 to 6 are general information collected and item 7 to 11 are the results from electrocardiogram. Items 12 to 13 are the results from cardiac fluoroscopy and thallium scintigraphy, which were done by injecting radioactive material or Thallium into the body, and followed by X-ray or Gamma to observe any abnormality in the blood vessels and heart muscles.

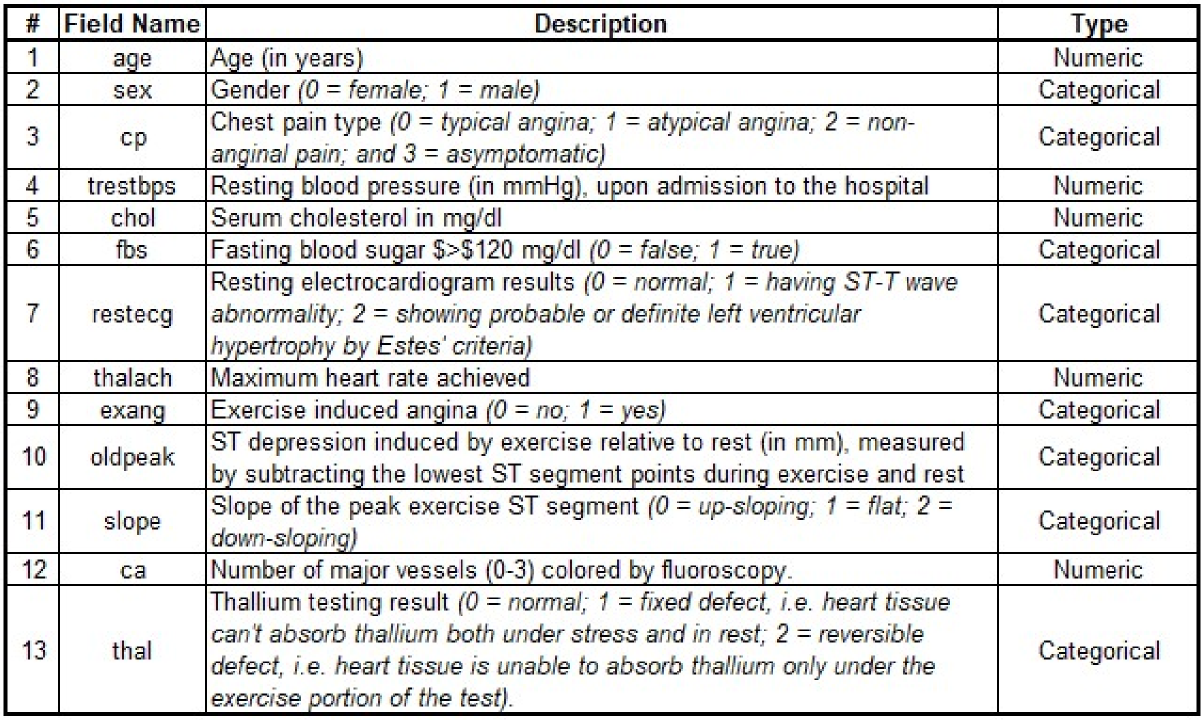
After collecting the data, data cleansing and data grouping are done:

1. Invalid input removal - 6 records with "?" input in fields *ca* and *thal* are removed.

2. Outlier removal - 1 record with *chol* amounted 564mg/dl, while the normal level is 150mg/dl and the next highest record in the data is 417mg/dl. To avoid distortion of results from outlier, such record is removed.

3. Data grouping - To give a more meaningful output from Logistic Regression, sub-category 1 of *restecg* which has 4 records only, is grouped with sub-category 2. 4 sub-categories for *cp* are grouped to become 2.

After data cleansing, there are 296 records remaining, and out of which around 45%



#	Field Name	Description	Type
1	age	Age (in years)	Numeric
2	sex	Gender *(0 = female; 1 = male)*	Categorical
3	cp	Chest pain type *(0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; and 3 = asymptomatic)*	Categorical
4	trestbps	Resting blood pressure (in mmHg), upon admission to the hospital	Numeric
5	chol	Serum cholesterol in mg/dl	Numeric
6	fbs	Fasting blood sugar $>$120 mg/dl *(0 = false; 1 = true)*	Categorical
7	restecg	Resting electrocardiogram results *(0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)*	Categorical
8	thalach	Maximum heart rate achieved	Numeric
9	exang	Exercise induced angina *(0 = no; 1 = yes)*	Categorical
10	oldpeak	ST depression induced by exercise relative to rest (in mm), measured by subtracting the lowest ST segment points during exercise and rest	Categorical
11	slope	Slope of the peak exercise ST segment *(0 = up-sloping; 1 = flat; 2 = down-sloping)*	Categorical
12	ca	Number of major vessels (0-3) colored by fluoroscopy.	Numeric
13	thal	Thallium testing result *(0 = normal; 1 = fixed defect, i.e. heart tissue can't absorb thallium both under stress and in rest; 2 = reversible defect, i.e. heart tissue is unable to absorb thallium only under the exercise portion of the test).*	Categorical

Figure 1: Details of 13 features

of records are with heart disease. This proportion is much higher than general population as the data was collected in clinic for patients that were suspected to have heart disease.

Plots of each features against the presence of heart disease are reviewed and noted that all the features have some associateship with the presence of heart disease, though some results are unexpected, like higher proportion of people without chest pain are having heart disease as compared those with chest pain. Plots among features are checked and no strong multi-collinearity is expected. The plots are included in Appendix 6.1.

We then divide our data into training set and validation set in 80%/20% proportion. Cross validation is used by splitting the training set into 10 folds and repeated 3 times. Standardization of features, using min-max or mean-standard deviation scaling, are done as pre-processing for different models. For categorical features with more than two possible values, one-hot encoding is used to represent each possible value as a separate numerical variable as applicable.

3 Methods

In this section, 8 statistical methods are discussed. We will discuss the rationale in choosing those models and the process to choose best tuned parameters for each model.

3.1 Logistic Regression

Logistic regression is a linear model that uses a sigmoidal function to map the input features to the output probability, that is the logarithm of odds $log(p/(1-p))$ is modeled as linear regression on covariates.

To identify the optimal model, we applied the backward feature selection method, that is to fit the model using all features first and remove the features with insignificant coefficients one-by-one according to p-values. The coefficients of covariates can give us a sense of how relevant the feature is related to the response variable.

3.2 K-Nearest Neighbors

K-Nearest Neighbors(KNN) is a popular statistical learning algorithm which predicts the class of a new data point based on the class of its nearest K neighbors in the feature space. KNN is a non-parametric method, meaning that it does not make any assumptions about the underlying distribution of the data.

We experimented with both unscaled and scaled KNN models by setting parameter $K = seq(1, 35, 1)$, here scaled refers to normalize the data using the Min-Max scaling. We believe scaled KNN is more appropriate as the dataset consists of numerical features with different units, and normalization helps to bring these features to the same scale.

3.3 Naive Bayes

Naive Bayes is a classification method based on the Bayes theorem. Naive Bayes assumes that the features are independent given the response variable. We chose Naive Bayes over Linear Discriminant Aanlysis (LDA) and Quadratic Discriminant Aanlysis

(QDA) as this dataset has likely independent features, while LDA and QDA assume that the features are normally distributed, which may not work well for categorical inputs.

3.4 Neural Network

Neural network is an upgraded version of logistic regression with multiple layers and neurons. To work with binary output as in our project, error calculation method and activation function are chosen as cross-entropy and default logistic. Cross-entropy measures the difference between two probability distributions. It is chosen as it provides a smooth and continuous measure of the error.

To find the optimal Neural Network model, different combinations of number of hidden layers and number of neurons in each layer are tested - mainly 2 hidden layers are considered, with number of neurons in layer 1 is up to 5 and layer 2 up to 2. With our small dataset, more noise is expected if model complexity is further increased.

Figure 2 visualizes the complexity of the neural network model.

3.5 Support Vector Machine

Support Vector Machine (SVM) finds the optimal hyperplane that best separates the data into different classes or predicts a continuous target value. In our data set is the case of classification, SVM will try to maximize the margin between the classes, which is the distance between the separating hyperplane and the closest data points from each class.

We chose SVM in view of its ability to handle high-dimensional data, robustness against over-fitting, and the flexibility to model both linear and non-linear relationships. We examined different combinations for tuning parameters for polynomial and radial kernel including *degree* / *gamma* and *cost* to choose the best model fit for our data.

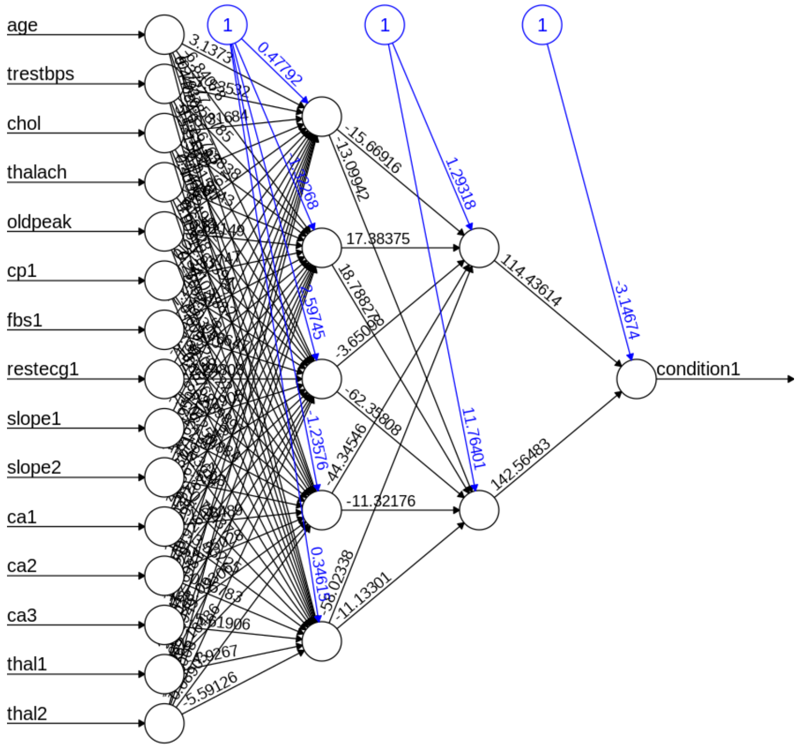


Figure 2: Visualization of Neural Network

3.6 Random Forest

Random forest builds an ensemble of decision trees with the leverage of multiple weak learners (in this case, decision trees) to produce a strong learner. In our study, we use $mtry$, the number of variables considered at each split as the tuning parameters. Generally, a smaller $mtry$ can lead to more diverse trees, which may help reduce overfitting and improve generalization.

3.7 Random Forest with Principal Component Analysis

Building from Random Forest, we tried to test if incorporating unsupervised technique could further improve the model accuracy. Principal Component Analysis (PCA) is applied - PCA is a dimensionality reduction technique that transforms a dataset by projecting it onto a lower-dimensional space.

The reasons of using PCA:

1. **Noise reduction**: PCA can help remove noise from the data, by focusing on the principal components that capture the most variance in the dataset. Reducing noise may improve model performance by reducing the chances of over-fitting.

2. **Dimensionality reduction**: By reducing the number of dimensions using PCA, the model can focus on the most important features, potentially improving the performance by reducing the curse of dimensionality.

3. **Computation efficiency**: Training a Random Forest on reduced number of features is computationally more efficient, which may allow for using more trees or more in-depth trees in the ensemble. This could indirectly contribute to performance improvement.

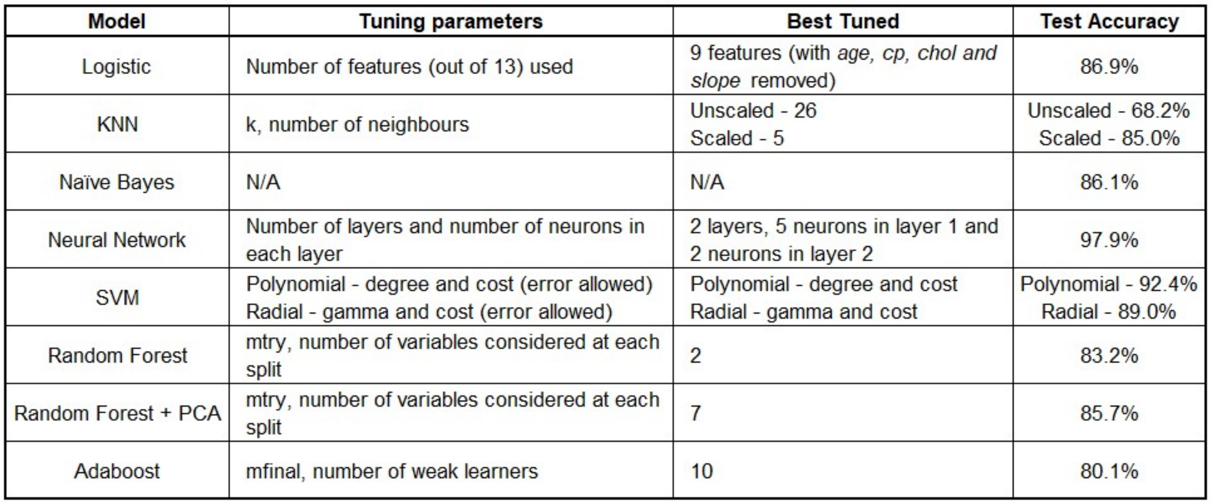
3.8 Adaptive Boosting

Adaptive Boosting (AdaBoost) is a boosting ensemble method that sequentially builds simple Decision Trees. Unlike Random Forest, the trees in AdaBoost consist of only a root and two leaves and are commonly referred to as ”Decision Stumps” [6]. There are 3 main ideas behind Adaboost. Firstly, it combines multiple weak learners into a single strong learner. Secondly, decision stumps have varying contributions to the final predictions. Lastly, each stump is made by taking the previous stump's mistakes into account, which helps improve the model's overall accuracy by focusing on instances that are harder to classify. It is a fast and powerful algorithm that can achieve high accuracy, however, it is sensitive to noisy data and outliers, and not suitable for large datasets.

The data set was prepossessed using Min-Max Scaling and one-hot encoding before the AdaBoost model was trained. One of the pivotal parameters in AdaBoost is $mfinal$, which denotes the number of weak learners that are sequentially combined to create the final ensemble model. Five distinct values of $mfinal$: 10, 30, 50, 70, and 90 were experimented to optimize the model performance.

4 Results

Figure 3 summarizes the tuning parameters and best tuned parameters using cross validation on training set. As the split of with and without heart disease is 55% / 45% respectively in our data, prediction accuracy is used as the selection criterion when identify the optimal parameters in each model. The details of the test set results are included in Appendix 6.2.3.



Model	Tuning parameters	Best Tuned	Test Accuracy
Logistic	Number of features (out of 13) used	9 features (with *age, cp, chol and slope* removed)	86.9%
KNN	k, number of neighbours	Unscaled - 26 Scaled - 5	Unscaled - 68.2% Scaled - 85.0%
Naïve Bayes	N/A	N/A	86.1%
Neural Network	Number of layers and number of neurons in each layer	2 layers, 5 neurons in layer 1 and 2 neurons in layer 2	97.9%
SVM	Polynomial - degree and cost (error allowed) Radial - gamma and cost (error allowed)	Polynomial - degree and cost Radial - gamma and cost	Polynomial - 92.4% Radial - 89.0%
Random Forest	mtry, number of variables considered at each split	2	83.2%
Random Forest + PCA	mtry, number of variables considered at each split	7	85.7%
Adaboost	mfinal, number of weak learners	10	80.1%

Figure 3: Optimal parameters on test set

Logistic: After fitting all the features in logistic regression model not all the coefficients are statistically significant. *age*, *cp*, *chol*, *slope* are removed and the remaining coefficients are significant. The test accuracy is improved from 85.7% to 86.9%.

KNN: As expected, the accuracy for scaled model outperforms un-scaled one as normalization helps to bring numerical features to the same scale that improve the performance. Scaled KNN will be used for final model selection.

SVM: As polynomial basis function kernel outperforms the radial, it is chosen for the final model selection.

Random Forest + PCA: The first 10 principal components, which can explain 90% of data variance, are used in Random Forest. The prediction accuracy is slightly improved from Random Forest alone, as expected.

5 Conclusion

To select the best model to predict heart disease using our dataset, we have prepared the results of each model on the validation set. Figure 4 shows that the validation accuracy for each model is around 90%, suggests that every model performs well. Among them, the Random Forest combined with PCA achieves the highest accuracy, at 96.6%. Additionally, this model also has the best performance in predicting heart disease with sensitivity score 97.3%, and predicting without heart disease, same as specificity as Naive Bayes and Random Forest alone.

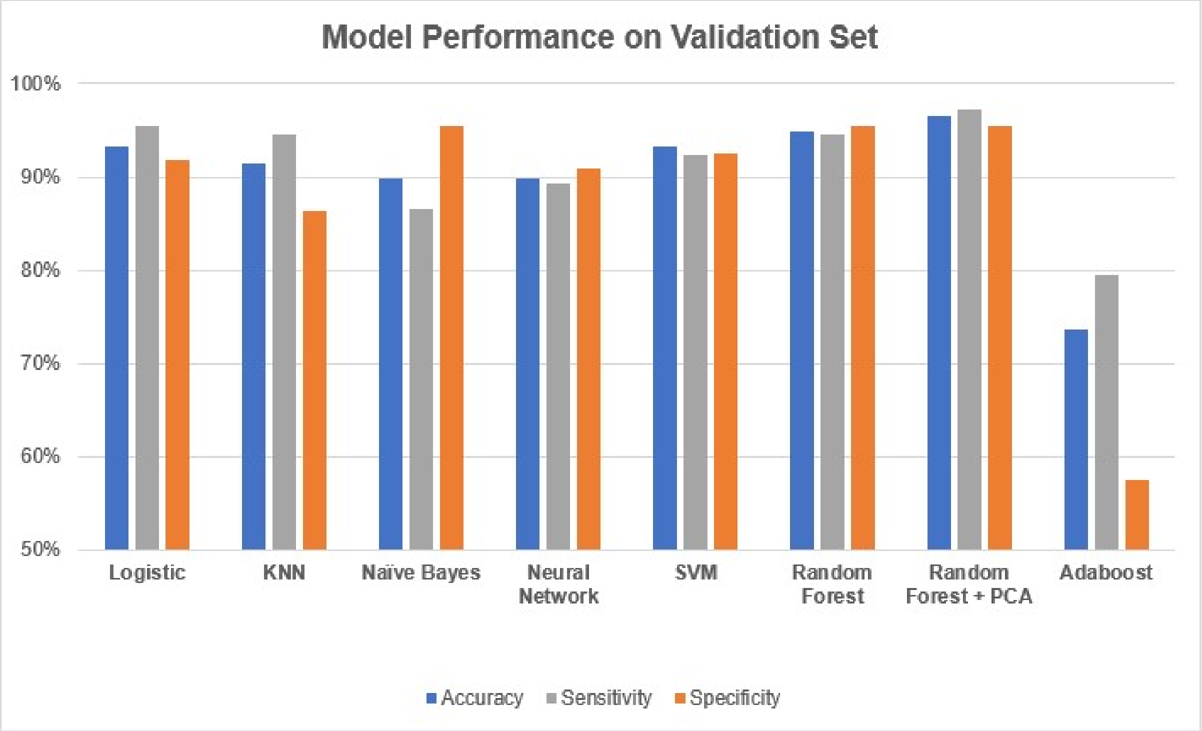


Figure 4: Result Comparison on Validation Set

Another observation is that the validation accuracy is generally higher than the test set. This is not normal, the reasons may be due to the small validation data with approximately 60 records and the input variables in the validation set is less diverse than the training set. The only exception is Adaboost has a lower validation accuracy compared to the test accuracy. This can be attributed to the model's tendency to over fit to the training set, leading to memorization of the training data and resulting in poor performance on new and unseen data.

In this project, we tried to incorporate unsupervised techniques (PCA) to improve supervised technique (Random Forest) and introduce Adaboost, a boosting algorithm. These methods can improve the prediction accuracy, but not the interpretability/explanability of the model. In terms of limitation, our dataset was collected in 1980s that cannot reflect the medical advancement in diagnosing heart diseases. We believe newer data should be collected to make the prediction more useful. After collecting the new data, similar tuning process is recommended.

In addition, prediction accuracy is used as the primary selection criterion in our study, however, as our ultimate goal is to identify patients with heart disease by using non-invasive tests only. Thus, we may consider to change the selection criteria so that the model can correctly identify patients with heart disease with the trade-off to predict wrongly for some of the healthy patients as unhealthy. This trade-off is essential as we do not want to miss any unhealthy people in our prediction.

References

[1] *Cardiovascular diseases (CVDs).* `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`. 2021.

[2] *Heart Disease Cleveland UCI.* `https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci/discussion/245147/`. 2020.

[3] *Heart Disease Data Set.* `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`. 1988.

[4] G. James et al. *An Introduction to Statistical Learning with Applications in R, second edition.* New York: Springer, 2021.

[5] Hamid Reza Marateb and Sobhan Goudarzi. “A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system”. In: *Journal of Research in Medical Sciences* (2015).

[6] Albers Uzila. *Understanding Ensemble Methods: Random Forest, AdaBoost, and Gradient Boosting in 10 Minutes.* `https://medium.com/towards-data-science/understanding-ensemble-methods-random-forest-adaboost-and-gradient-boosting-in-10-minutes-ca5a1e305af2`. 2022.

6 Appendices

6.1 Data and Data Analysis

The following plots examine the relationship between each feature and target variable.

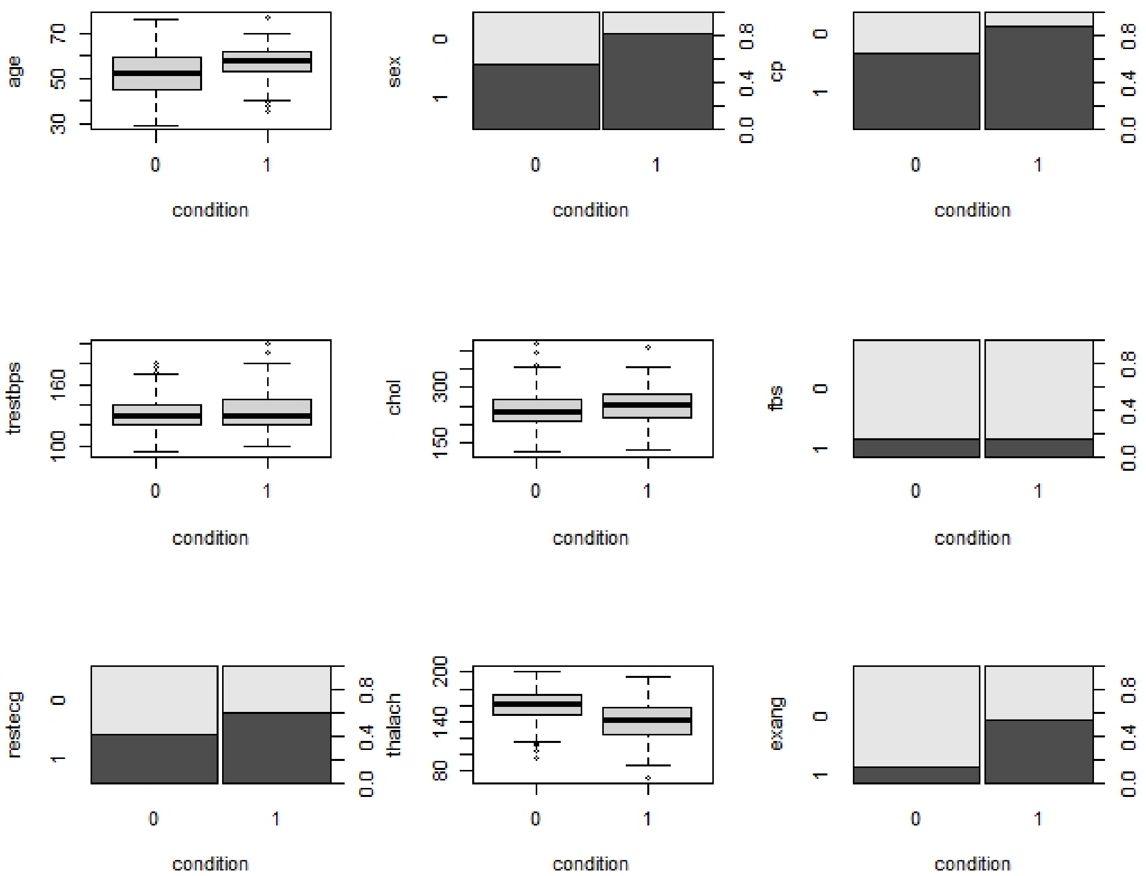


Figure 5: Plots of each features versus response variable

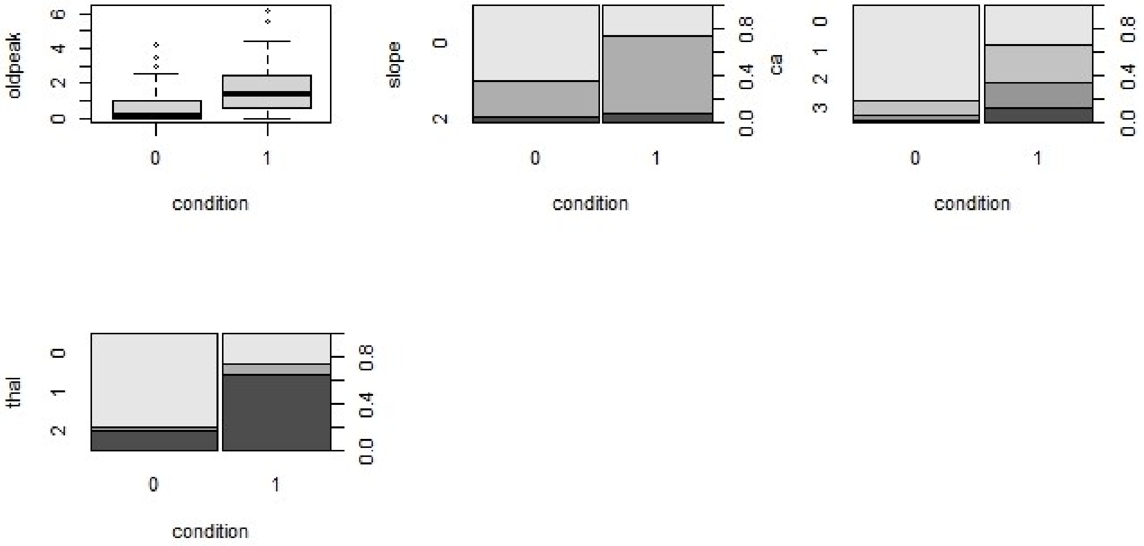


Figure 6: Plots of each features versus response variable (cont'd)

The following plots visualize the relationship among features, and no strong multicollinearity is expected.

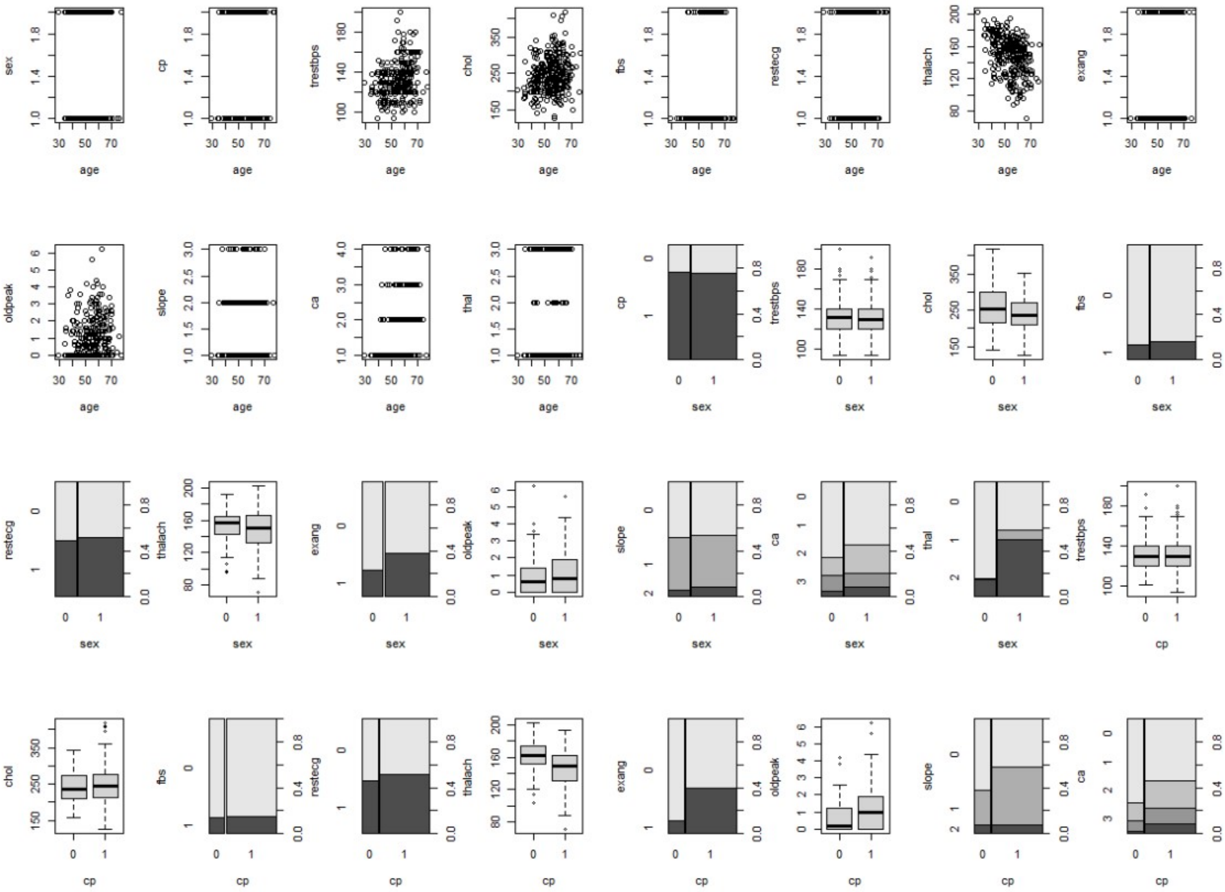


Figure 7: Plots among features

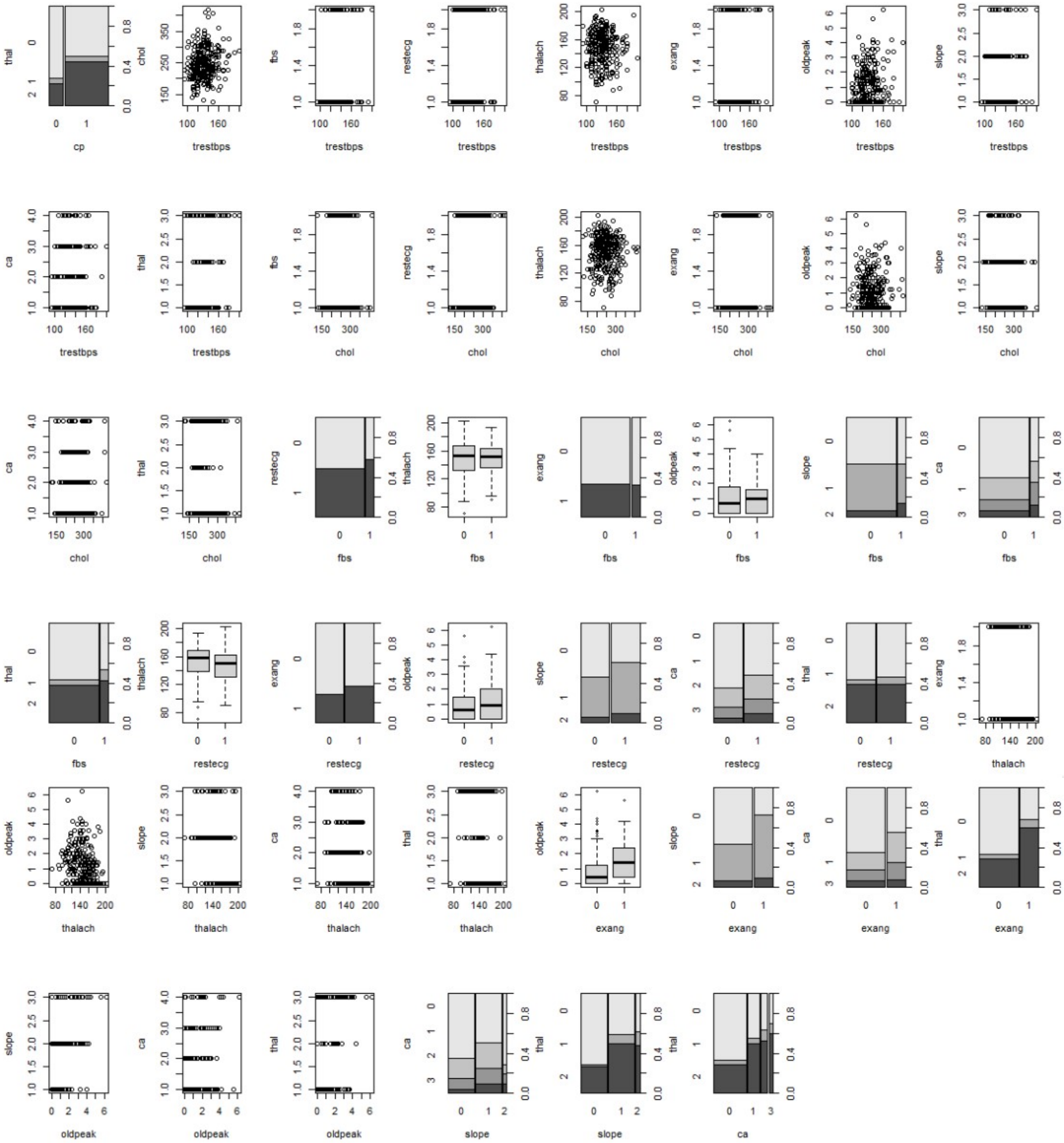
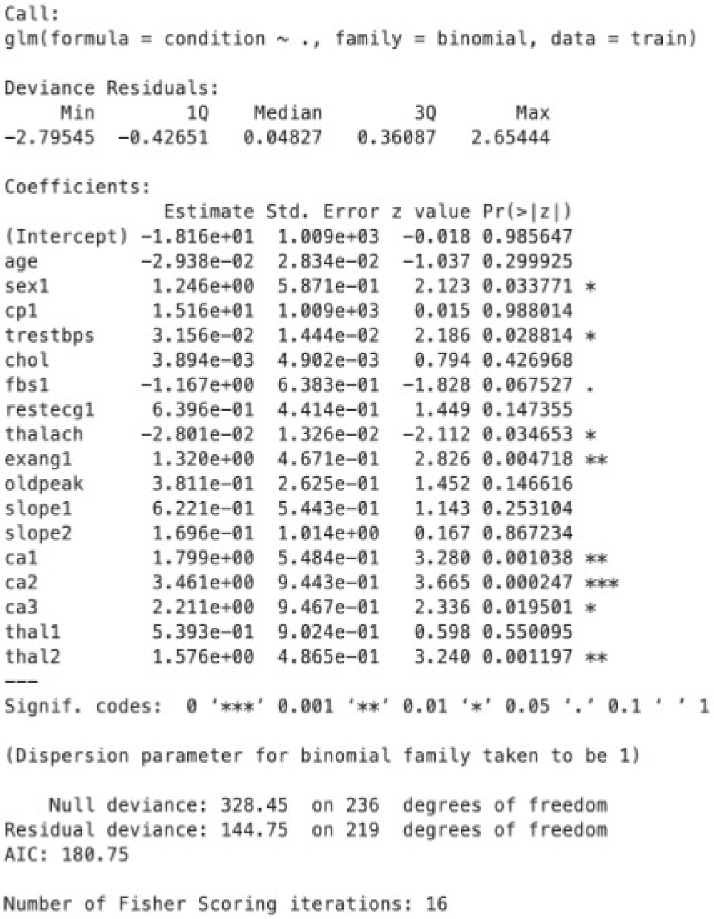


Figure 8: Plots among features (cont'd)

6.2 Parameter tuning for test set

The following figures show the prediction accuracy varied by different tuning parameters. The best tuned parameters are chosen as the one with highest prediction accuracy.

6.2.1 Logistic Regression



```
Call:
glm(formula = condition ~ ., family = binomial, data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.79545  -0.42651   0.04827   0.36087   2.65444

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.816e+01  1.009e+03  -0.018 0.985647
age         -2.938e-02  2.834e-02  -1.037 0.299925
sex1         1.246e+00  5.871e-01   2.123 0.033771 *
cp1          1.516e+01  1.009e+03   0.015 0.988014
trestbps     3.156e-02  1.444e-02   2.186 0.028814 *
chol         3.894e-03  4.902e-03   0.794 0.426968
fbs1        -1.167e+00  6.383e-01  -1.828 0.067527 .
restecg1     6.396e-01  4.414e-01   1.449 0.147355
thalach     -2.801e-02  1.326e-02  -2.112 0.034653 *
exang1       1.320e+00  4.671e-01   2.826 0.004718 **
oldpeak      3.811e-01  2.625e-01   1.452 0.146616
slope1       6.221e-01  5.443e-01   1.143 0.253104
slope2       1.696e-01  1.014e+00   0.167 0.867234
ca1          1.799e+00  5.484e-01   3.280 0.001038 **
ca2          3.461e+00  9.443e-01   3.665 0.000247 ***
ca3          2.211e+00  9.467e-01   2.336 0.019501 *
thal1        5.393e-01  9.024e-01   0.598 0.550095
thal2        1.576e+00  4.865e-01   3.240 0.001197 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 328.45  on 236  degrees of freedom
Residual deviance: 144.75  on 219  degrees of freedom
AIC: 180.75

Number of Fisher Scoring iterations: 16
```

Figure 9: Result summary for logistic regression

6.2.2 KNN

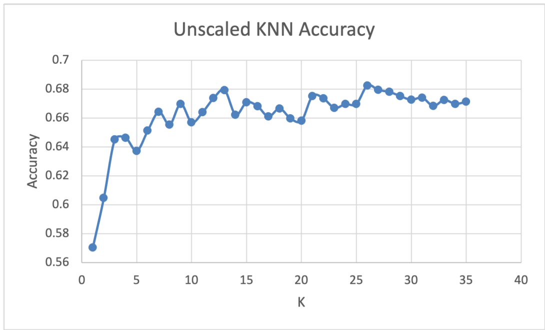


Figure 10: Tuning parameter for unscaled KNN

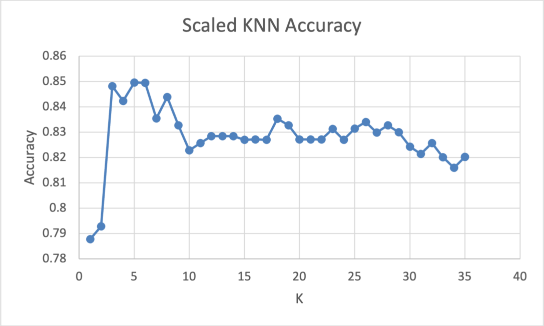


Figure 11: Tuning parameter for scaled KNN

6.2.3 Adaboost

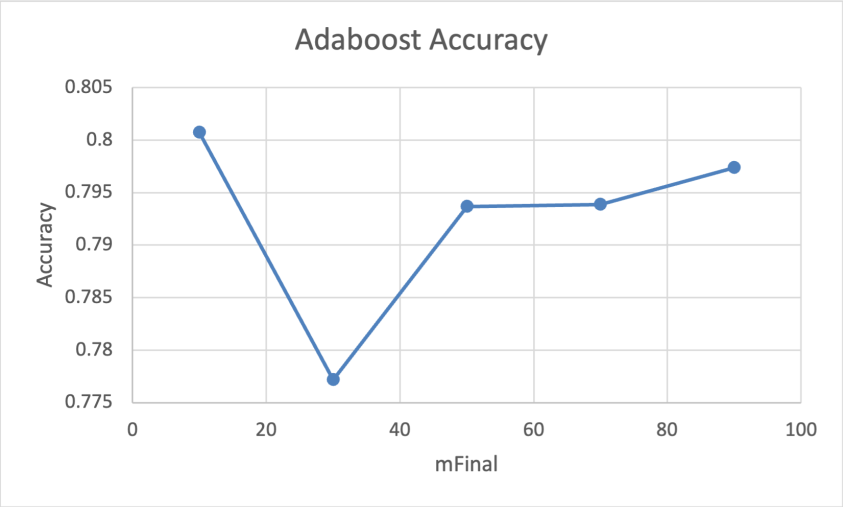
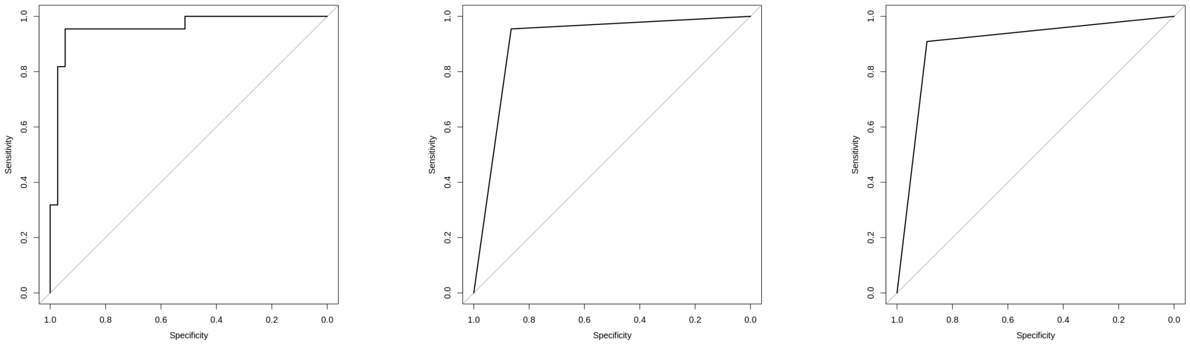


Figure 12: Tuning parameter for Adaboost

6.3 ROC curves

Figures below are ROC curves of logistic regression when choosing the best subset, Naive Bayes, and Neural Network.



(a) Logistic with the best subset

(b) Naive Bayes

(c) Neural Network

Figure 13: ROC curves

6.4 R code

GitHub: https://github.com/MichaelLXC/ST694-Final-Project/tree/main

7 Delineation of Work

Nora Fan - KNN, Adaboost and conclusion

Tianliu Fu - Logistic, Naive Bayes and Neural Network

Yeemui Lau - Data and overall report review

Xiaocong Lian - SVM, Random Forest, Random Forest + PCA