Final Report_Diagnosis of Heart Attack Risk

Group 21: Haoyang Zhang

2024-05-10

1 Introduction

Heart attack represents a significant global public health concern, accounting for the leading cause of mortality globally, with an estimated 17.9 million deaths annually, which comprises approximately 32% of all global deaths. Myocardial infarction, or heart attack, is a significant contributor to this toll, emphasizing the urgent need for effective preventive measures and early diagnosis. The impact of heart attacks extends beyond mortality rates, affecting the quality of life and economic cost. Individuals who survive a heart attack frequently encounter a range of complications that can significantly impair their daily functioning and long-term health. From an economic perspective, the treatment of heart attack patients encompasses acute care, ongoing medication, and potentially lengthy physical rehabilitation. This can result in considerable healthcare spending, with estimates indicating that this cost may be hundreds of billions globally.

The application of machine learning offers the potential for a transformative approach to addressing this global health challenge. Practices on large datasets of patient records, genetic information, and lifestyle factors enables the prediction of the risk of heart attacks with greater accuracy and earlier than is possible with traditional methods. The objective of this project is to develop a predictive classifier that will assist healthcare providers in identifying individuals at risk of a critical event before it occurs.

2 Related Work

Machine learning methods are widely used in the diagnosis of cardiovascular disease in academia and the medical field. The most popular approach is supervised learning with response variable, i.e. the probability of disease occurrence, or the presence or absence of the risk. It often requires the training data contains sufficient information of a patient including baseline, life style, physical and biochemical tests results, living area, socioeconomic factors, medication use and past medical history to ensure the model predicting performance.

As machine learning methodologies continue to be developed and implemented, the predictive accuracy of classifiers utilized to assist in the diagnosis of cardiovascular disease continues to improve. In a previous relevant study (Reddy et al, 2020) multiple classifiers are built on a dataset and gained 85.8% accuracy with Logistic Regression with PCA. The article also mentioned a classifying strategy called ensemble classifier that integrates different classification models in a weighted combination to improve the predictive performance which sounds valuable more exploration, however in this article this approach did not perform better than the single classifiers in term of accuracy. Approaches adopted by this article is instructive in the choice of methodologies for this analysis.

In addition, using well-organized training datasets, researchers proposed two deep neural networks for the effective expectation of coronary heart disease risk using heartsound features (Arslan et al, 2022). Prediction procedures are unable to learn from irregular data in most real-world datasets. Rather than relying on entire or randomly chosen training datasets, they advocated constructing training datasets by distinguishing regular from a highly biased subset. Two processes are involved in the preparation of the training data: to improve the highly biased set, variable autoencoders separate the original training datasets into two sets:

widely dispersed and highly skewed. The last step is to train two separate deep neural network classifiers. According to the suggested approach, the AUC was 0.882 and accuracy was 0.892, which outperformed more common methods in terms of specificity, accuracy, exactness, recall, and the f-measure (0.915).

3 Methods

The main objective of this project is to build a reliable machine learning classifier to predict whether there is the risk of cardiovascular disease based on personal data of individuals. Based on the dataset, detailed personal data are included. Hence, models constructed based on multiple features would be evaluated and utilized separately. To give the best classification outcome and to consolidate our understanding of machine learning in the meantime, five classification models, namely Logistic Regression, Discriminant Analysis, KNN, Random Forest algorithm, and LASSO, would be launched and compared in this project.

Logistic Regression is the statistical model that mostly used when we deal with a classification problem, which models the probability that an observed outcome belongs to a particular category. In a dichotomous classifier, it can be considered as the linear model of the log-odds or logit of the response. This also implies that the Logistic Regression requires the assumption of a linear relationship between the log-odds of the outcome and each predictor variable, indicating it probably will not perform well if there are complex, non-linear relationships in the data. Nevertheless, due to the reduced computational intensity and the favorable interpretability, we have selected Logistic Regression as the initial option.

Discriminant analysis is another statistical technique employed in the field of machine learning with the objective of classifying a set of observations into predefined classes. The method projects features onto a lower-dimensional space with the objective of maximizing the separation between multiple classes. There are two primary types of discriminant analysis: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

The K-Nearest-Neighbor (KNN) classification is a simple, versatile, and easy-to-implement supervised machine learning algorithm used for both classification and regression tasks, but it's particularly popular in classification problems. The principle behind KNN is straightforward: it classifies a new data point based on the majority label of the 'K' closest points in the dataset. KNN is non-parametric, meaning it makes no assumptions about the underlying data distribution. This flexibility allows it to be used in real-world scenarios where the data may not conform to theoretical assumptions. The classification depends on the distance metric used to find the closest neighbors. The choice of distance metric can significantly influence the performance of the classifier. In addition, in this analysis, due to our dataset contains 26 different features, should we cautious about the "curse of dimensionality", because the performance of KNN model can degrade with high-dimensional data.

Random Forest classification is another powerful and versatile machine learning algorithm that belongs to the ensemble learning family. Random Forest combines multiple decision trees to improve the accuracy and stability of the model. The ensemble approach helps in reducing the variance of individual trees, leading to a more robust overall prediction. It employs the bootstrap aggregating (bagging) technique where each tree in the forest is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. However, since our dataset has a large size (over 8,000 observations), we are facing the heavy computational burden when using the Random Forest model.

LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It is originally developed for regression tasks, but LASSO can also be adapted for use in classification through methods such as logistic regression with a LASSO penalty.

4 Data & Experiment Setup

The data we plan to use is Heart Attack Risk Prediction Dataset on Kaggle, which includes 24 features excepting the target variable Heart.Attack.Risk. In the preprocess stage, in order to make the data more easy to be tackled with, variable Blood.Pressure is separated into two variables denoting Systolic and Diastolic, representing the Systolic and Diastolic blood pressure respectively. Thus eventually the dataset contains 25 variables. Dataset records 8,763 individual's data, and for each participant multiple aspects are measured. Detailed composition of the dataset is shown in the following table.

Table 1. Attributes Of Heart Attack Risk Dataset

NO.	Attribute	Type	Description
1	Age	Numeric	Age measured in years (18 to 90)
2	Sex	Binary	Gender (1 = Male; 2 = Female)
3	Cholesterol	Numeric	Cholesterol level measured in mg/dL
4	Systolic	Numeric	Systolic pressure measured in mmHg
5	Diastolic	Numeric	Diastolic pressure measured in mmHg
6	Heart.Rate	Numeric	Heart rate measured in beats per minute
7	Diabetes	Categorical	Presence of diabetes (0 = No; 1 = Yes)
8	Family.History	Binary	Presence of family history of cardiovascular disease (0 = No; 1 = Yes)
9	Smoking	Binary	Presence of smoking behavior (0 = No; 1 = Yes)
10	Obesity	Binary	Presence of obesity (0 = No; 1 = Yes)
11	Alcohol.Consumption	Binary	Presence of alcohol consuming (0 = No; 1 = Yes)
12	Exercise.Hours.Per.Week	Numeric	Time spend on physical exercising per week measured in hours
13	Diet	Categorical	Factor demonstrated dietary status (-1 = Unhealthy; 0 = Average; 1 = Healthy)
14	Previous.Heart.Problems	Binary	Presence of previous heart problem (0 = No; 1 = Yes)
15	Medication.Use	Binary	Presence of any medication use (0 = No; 1 = Yes)
16	Stress.Level	Categorical	Factor demonstrate the level of stress endured (from 1 to 10)
17	Sedentary.Hours.Per.Day	Numeric	Time with sitting per week measured in hours
18	Income	Numeric	Personal annually income
19	BMI	Numeric	Body Mass Index
20	Triglycerides	Numeric	Triglycerides level measured in mg/dL
21	Physical.Activity.Days.Per.Week	Categorical	Factor demonstrates days with physical activity in a week (from 0 to 7)

NO.	Attribute	Type	Description
22	`Sleep.Hours.Per.Day`	Categorical	Time spend on sleeping per week measured in hours
23	`Country`	Categorical	Living area with 20 unique value
24	`Continent`	Categorical	Living area with 6 unique value
25	`Hemisphere`	Binary	Living area with 2 unique value
26	`Heart.Attack.Risk` (target)	Binary	Presence of heart attack risk (0 = No; 1 = Yes)

The target variable `Heart.Attack.Risk` is a binary data and there are 3139 (35.8%) observations marked as exist risk of heart attack, the rest are not.

For these 8763 observations, 70% of them are randomly selected as training set for model building and the rest 30% of data are included in the testing set as our initial settings. In each set, the proportion and distribution of risk observation shall be the same. Training set is used to construct each model and testing set is used to evaluate the model. To avoid overfitting, each training error and testing error are both important measuring standards to choose the best model.

Since the KNN is sensitive to the distance between data points and LASSO requires the regularization of the dataset, before the modeling the dataset has been properly scaled.

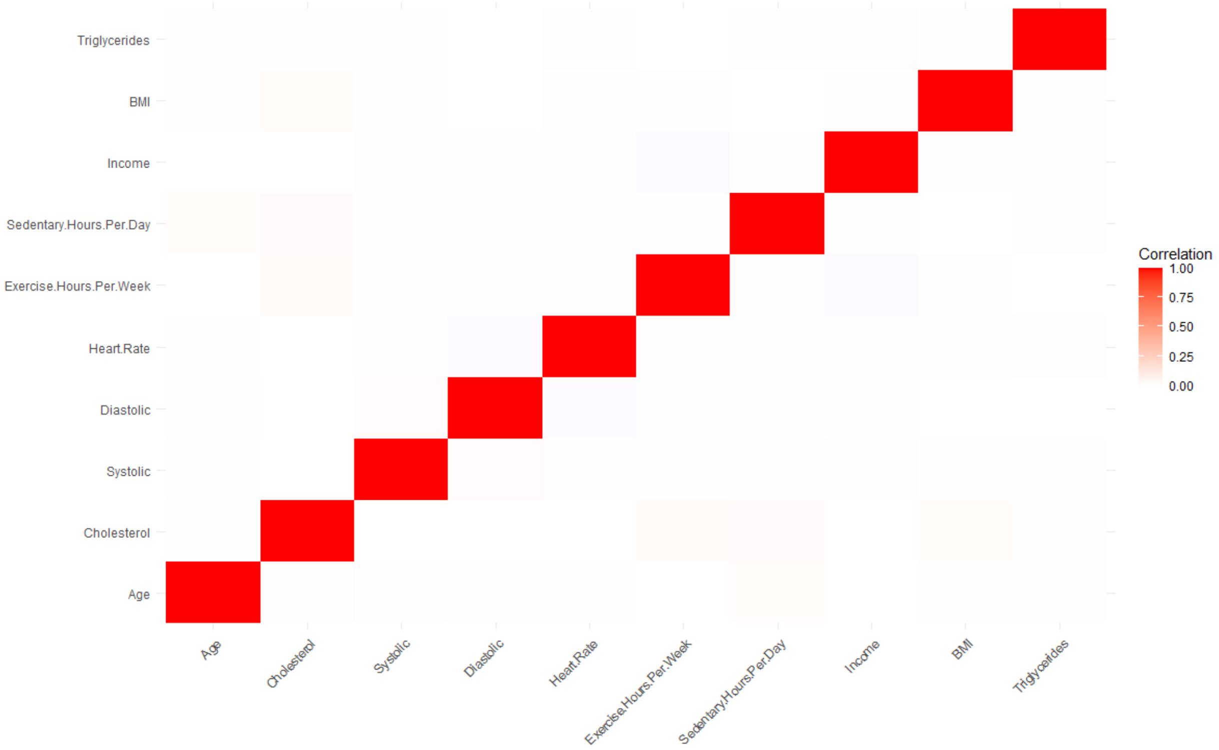


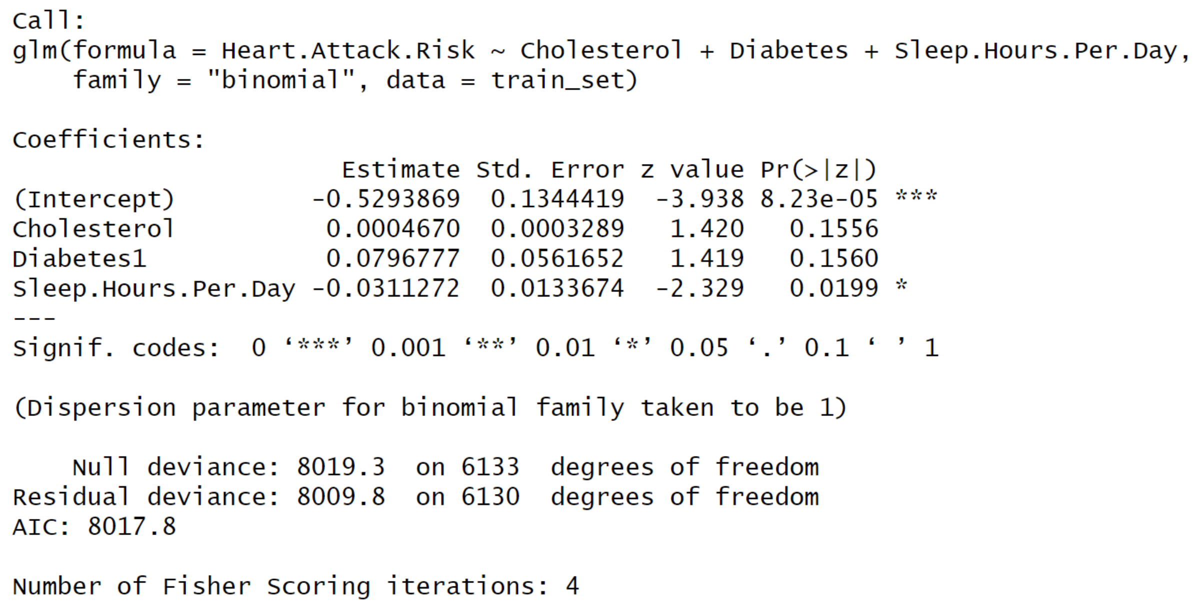
Figure 1: Correlation Matrix of Numeric Variables

The figure above illustrate the correlation heatmap of all continuous variables in the dataset. Colors are extremely faded in the triangle area which implies little correlation between these features.

5 Results

Logistic Regression

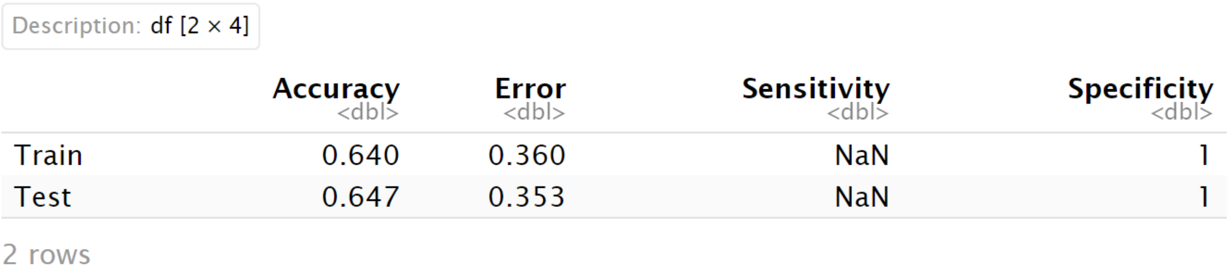
The idea of model selection is we first fit the full model containing all variables and then conduct stepwise selection from both sides using AIC. Should the selected model consisted by the best variable combination subset.



```
Call:
glm(formula = Heart.Attack.Risk ~ Cholesterol + Diabetes + Sleep.Hours.Per.Day,
    family = "binomial", data = train_set)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -0.5293869  0.1344419  -3.938 8.23e-05 ***
Cholesterol          0.0004670  0.0003289   1.420   0.1556    
Diabetes1            0.0796777  0.0561652   1.419   0.1560    
Sleep.Hours.Per.Day -0.0311272  0.0133674  -2.329   0.0199 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8019.3  on 6133  degrees of freedom
Residual deviance: 8009.8  on 6130  degrees of freedom
AIC: 8017.8

Number of Fisher Scoring iterations: 4
```

Figure 2: Summary of Selected Logistic Model

The deviance and AIC looks bad and none of variables selected shown significance.



Description: df [2 × 4]

	Accuracy	Error	Sensitivity	Specificity
	<dbl>	<dbl>	<dbl>	<dbl>
Train	0.640	0.360	NaN	1
Test	0.647	0.353	NaN	1

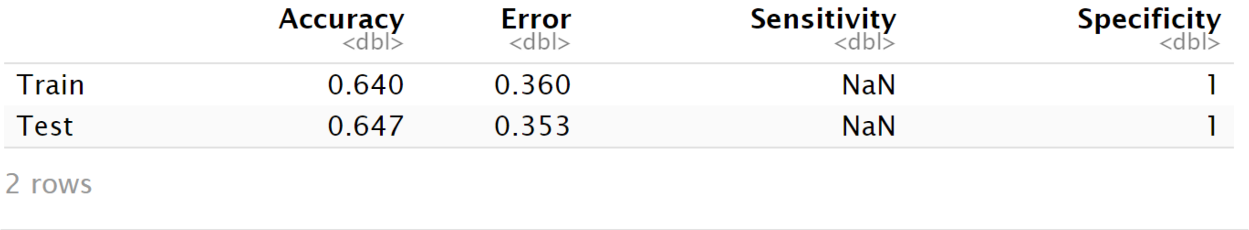
2 rows

Figure 3: Classifying Performance of Selected Logistic Model

Evaluation results shows that the classifier failed to obtain sensitivity and produce 1 specificity, which indicates that the model do not assign any individuals ro positive side, which means that the model is probably doing the random guessing. Besides, an `8-folds` cross validation is also conducted and produce CV error as 0.358, which is also unacceptable for a classifier, but shows consistency with the training and testing evaluation.

Discriminant Analysis

We mainly focus on Linear Discriminant Analysis (LDA) in this study. In order to simplify modeling process and make sure that summary output is not cluttered and redundant, we here directly use the features combination selected by AIC stepwise selection in previous part. Moreover it actually performs similar compared with full model.



	Accuracy	Error	Sensitivity	Specificity
	<dbl>	<dbl>	<dbl>	<dbl>
Train	0.640	0.360	NaN	1
Test	0.647	0.353	NaN	1

2 rows

Figure 4: Classifying Performance of LDA Model

The evaluation results by splitted dataset validation is similar with that of Logistic. Training error got 0.360 and testing error is 0.353. The classifier again failed to assign any individuals to positive class, preducing spectacle sensitivity and specificity.

KNN

Standardization of dataset is done before KNN is conducted. For KNN method previous best subset might not be statistically reliable, thus all variables are involved in this method. We first use `8-folds` cross validation to select value of K, then using the splitted data to evaluate the model with optimal K.

Figure 5 shown that the curve reach the smallest error at $K = 100$. We choose 100-nearest-neighbors to construct the optimal KNN model. However, at the local minimum, the cross validation error is still pretty large, which is 0.358, means the overall performance of the model is not satisfactory neither.

Figure 6 shows overall performance of selected KNN model.

Random Forest

Figure 7 shows mean decrease accuracy on the left and the mean decrease Gini on the right of all variables in the full random forest model. Variables at the top of the plot are the most important, indicating that they have the most impact on the output of the model, based on the chosen importance measure. For some variables have very low importance scores, they may be candidates for removal in simplified models, which could potentially lead to faster and more efficient models without significantly reducing accuracy, and this provide guidence for obtaining the optimal random forest model.

Eventually 5 variables that appears in the top important 10 variables in both rankings, namely `Counrty + Income + Triglycerides + Physical.Activity.Days.Per.Week + Systolic` are selected to constructed the optimal random forest model, and Figure 8 shows the performance of this classifier.

The evaluation result in Figure 8 shows training set presents sign of overfitting. Potential causes could be that those trees generated for this model are too deep, enable them individually learn to perfectly classify all training points, capturing noise and anomalies in the training data that do not generalize to new data. But for the testing set, this time our accuracy reaches 60%.

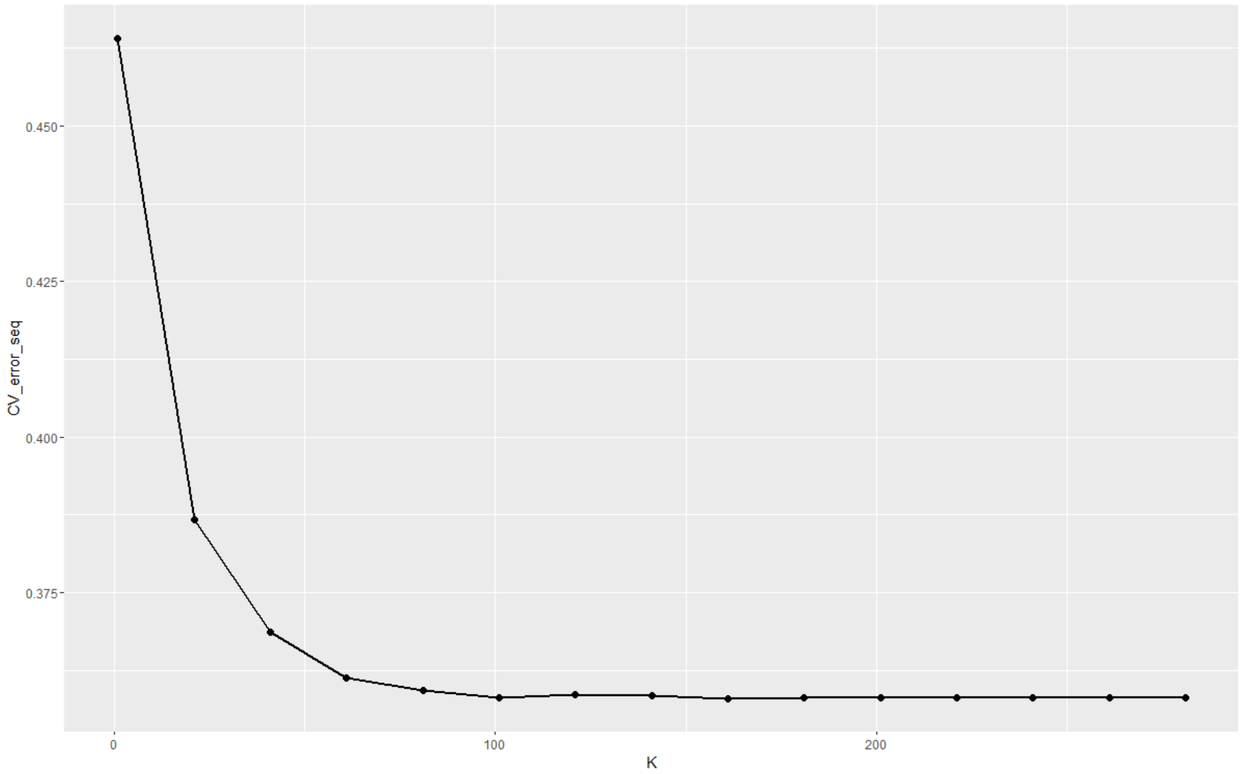
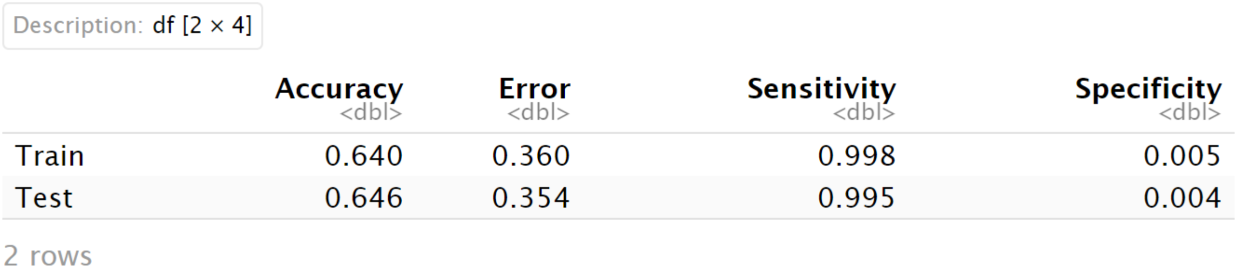


Figure 5: Cross Validation Errors v.s. K



Description: df [2 × 4]

	Accuracy <dbl>	Error <dbl>	Sensitivity <dbl>	Specificity <dbl>
Train	0.640	0.360	0.998	0.005
Test	0.646	0.354	0.995	0.004

2 rows

Figure 6: Classifying Performance of Optimal KNN Model

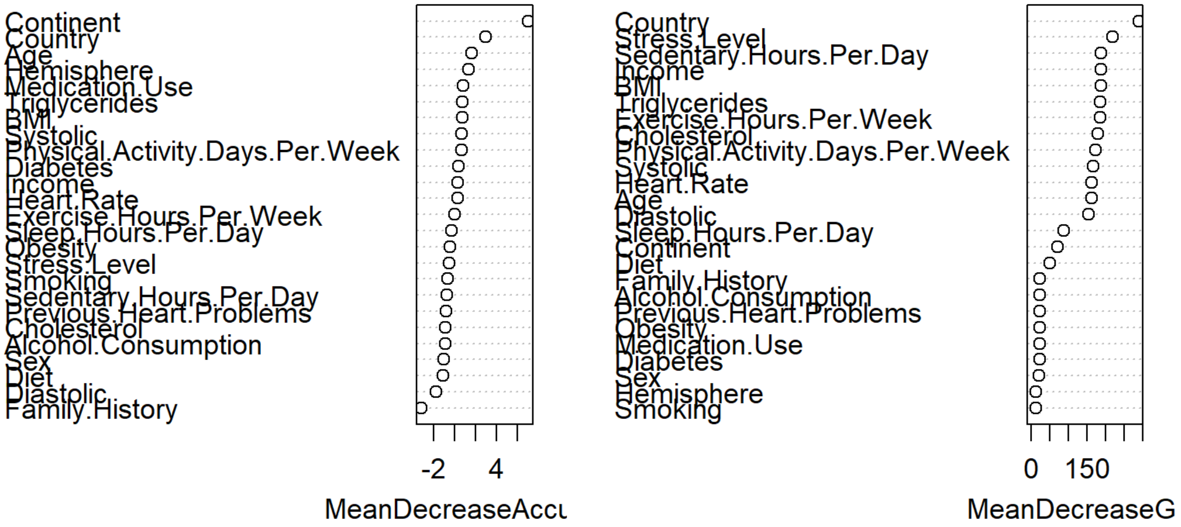
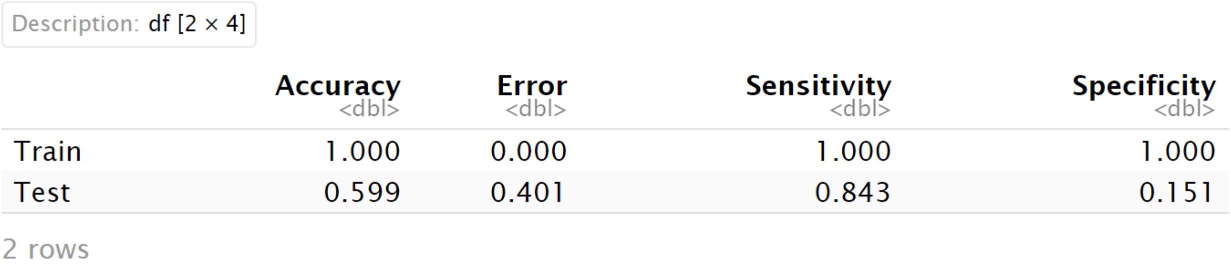


Figure 7: Variables' Importance Plot



Description: df [2 × 4]

	Accuracy <dbl>	Error <dbl>	Sensitivity <dbl>	Specificity <dbl>
Train	1.000	0.000	1.000	1.000
Test	0.599	0.401	0.843	0.151

2 rows

Figure 8: Classifying Performance of Random Forest Model

LASSO

For LASSO regression, our strategy will be similar with that of KNN modeling: using cross validation to find optimal λ, then evaluate the selected model by splitted dataset.

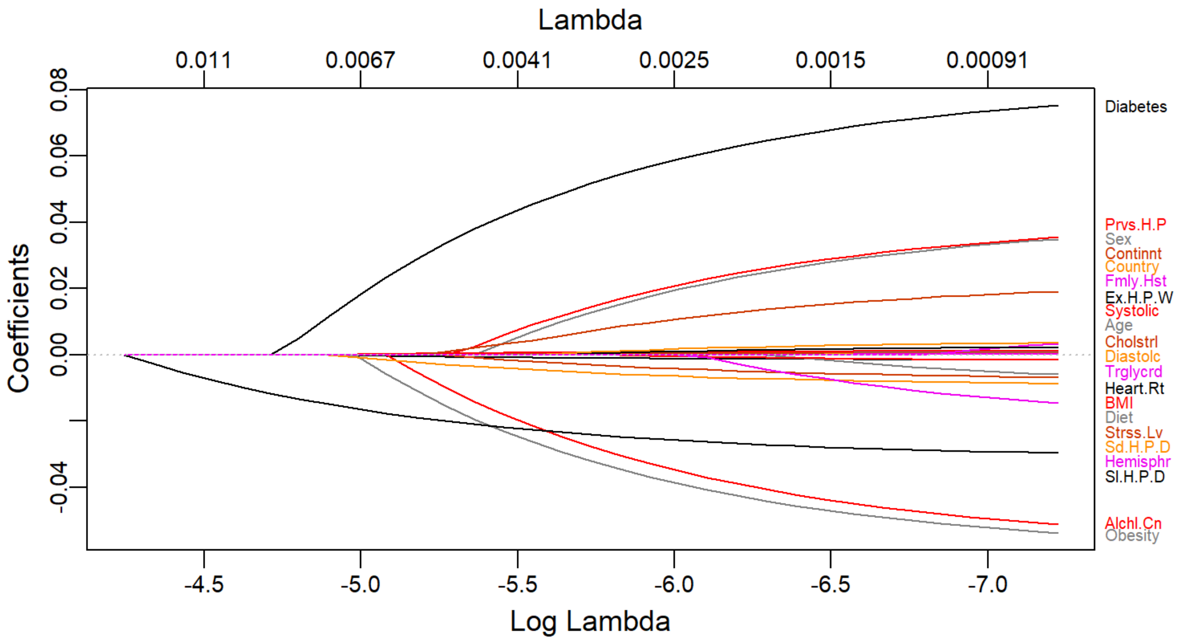


Figure 9: Coefficients of Variables under Different lambda

In Figure 10, the binomial deviance reaches minimum when λ takes value of 0.00927.

Figure shows classifying performance of selected LASSO model. On the contrary of Logistic Regression, LASSO produce 100% sensitivity but failed in specificity, which means LASSO classifier do not assign any individual to negative side.

ROC curve

Figure 12 illustrates ROC curves for 5 models previously built. This kind of ROC literally means that the model attempted perform no better than random guessing.

6 Discussion

This task eventually ended up with failure in the task of constructing a predicting model. This dataset is highly suspected that was intentionally fabricated, rather than collected from real world. Several clues can justify this hypothesis. Firstly, the dataset do not originally contains any missing value, for a dataset with more than 8,000 observations, it is nearly impossible if it is collected from real world. Secondly, not a single pair of variables seem to be correlated with each other. Thirdly, all continuous features tends to be uniformly distributed, and lastly, all attempted models are literally doing random guessing. In the source wbsite of the dataset, one comment suggests that the dataset is probabily generated by AI.

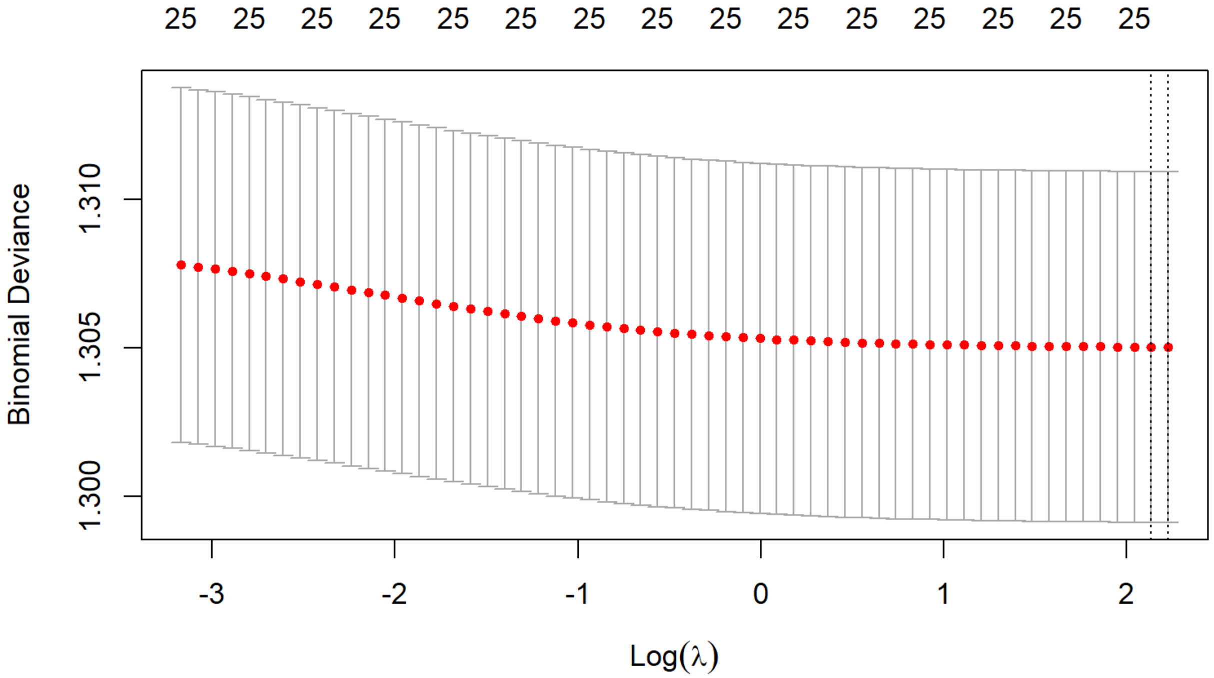
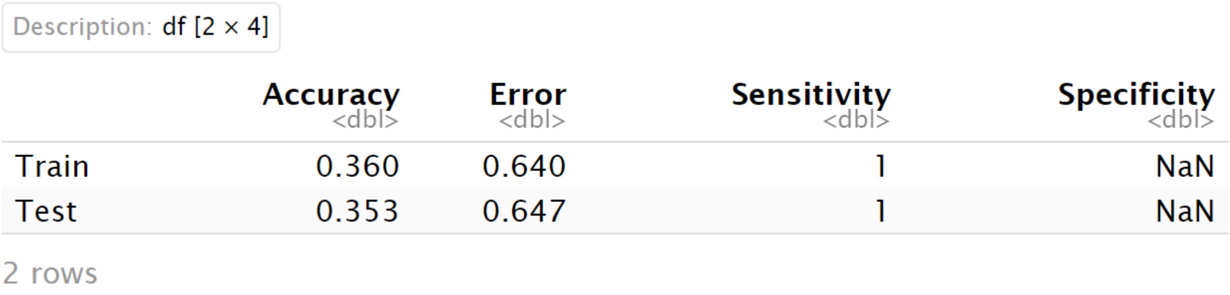


Figure 10: Deviance v.s. log-lambda



Description: df [2 × 4]

	Accuracy	Error	Sensitivity	Specificity
	<dbl>	<dbl>	<dbl>	<dbl>
Train	0.360	0.640	1	NaN
Test	0.353	0.647	1	NaN

2 rows

Figure 11: Classifying Performance of LASSO

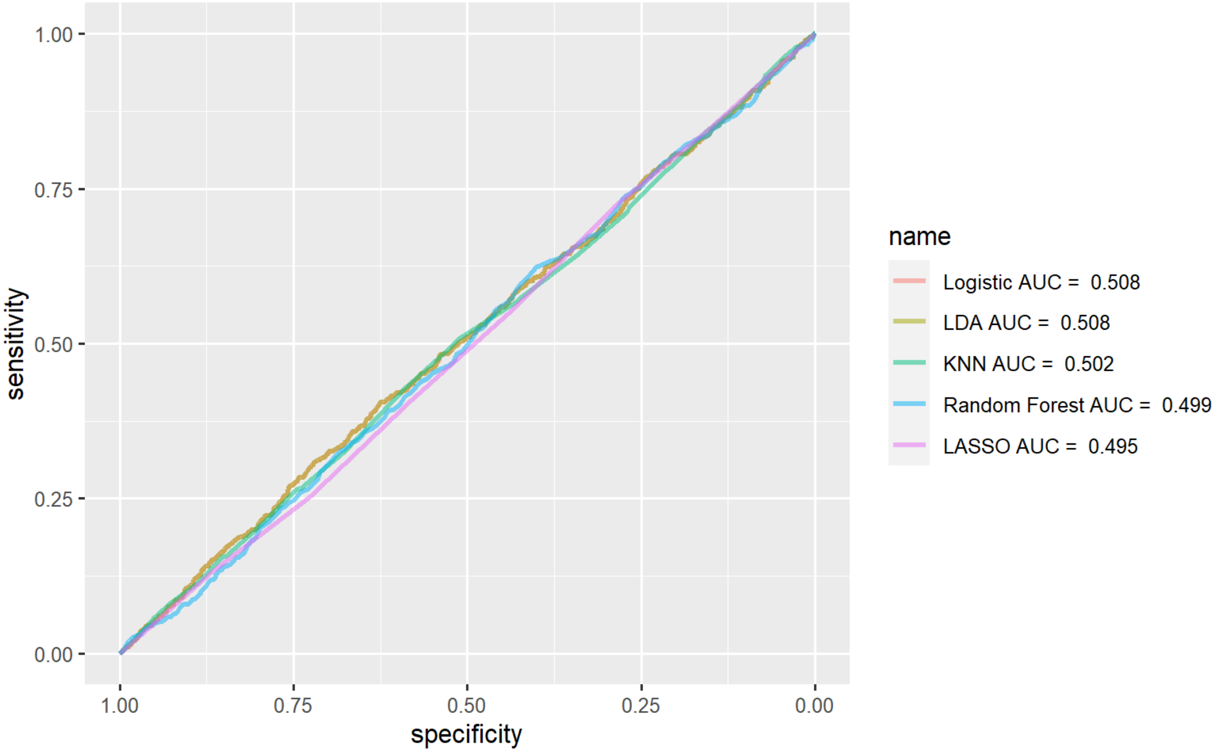


Figure 12: ROC curves

Nevertheless, if the dataset is reliable and collected from the real world, this study still can be improved. For example the consideration of interaction terms, and more modeling methods such as neural network.

References

K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam and H. N. Chua, “Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis,” 2020 8th International Conference on Intelligent and Advanced Systems (ICIAS), Kuching, Malaysia, 2021, pp. 1-6, doi: 10.1109/ICIAS49414.2021.9642676.

Arslan, Ö., & Karhan, M. (2022). Effect of Hilbert-Huang transform on classification of PCG signals using machine learning. Journal of King Saud University-Computer and Information Sciences, 34(10), 9915-9925.

Appendices

The dataset can be access by Heart.Attack.Risk.Predict - Kaggle