An Investigation into the Prediction of Heart Disease

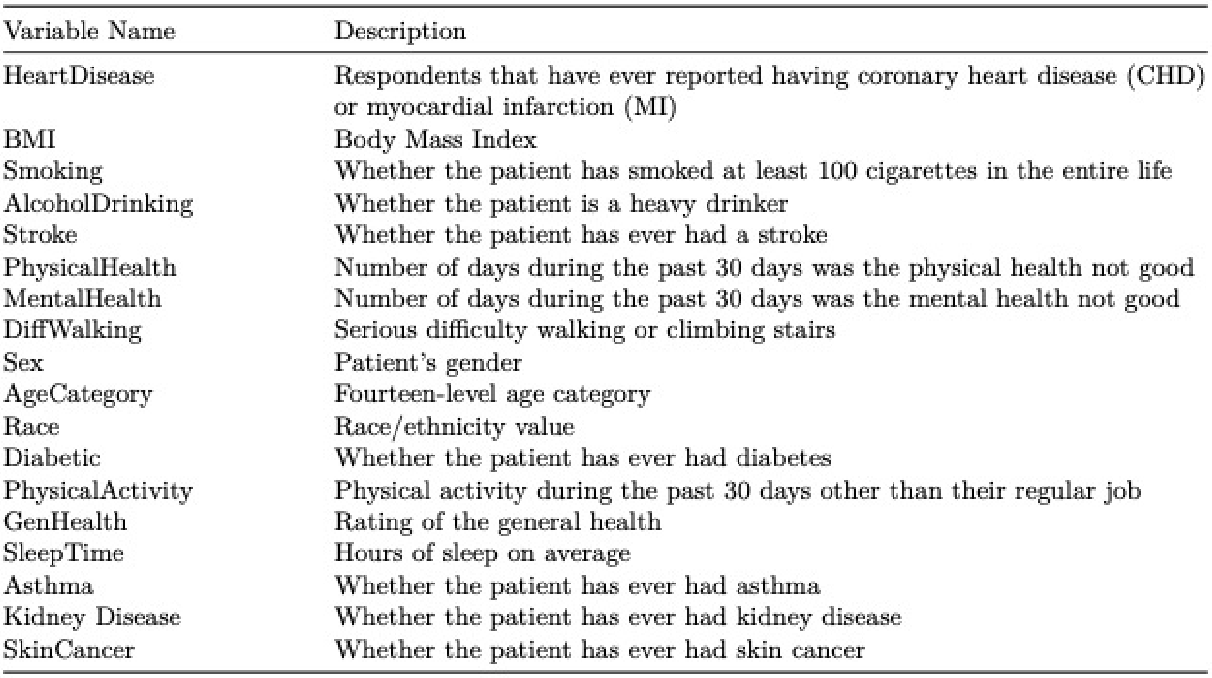
Executive Summary

According to the Centers for Disease Control and Prevention, heart disease is one of the leading causes of death for people of most races in the US. High blood pressure, high cholesterol, and smoking are several key risk factors for heart disease. Other key risk factors include diabetic status, obesity, drinking, not getting enough exercise, etc. This project aims to find out which variables have a significant effect on the likelihood of heart disease.

This project mainly serves as a reference for people in the healthcare industry who are interested in detecting and preventing the factors that have a great impact on heart disease. By leveraging the data tools and power of data mining, we are able to seek the patterns and build models to predict heart disease patients' conditions. The project mainly consists of four parts: data cleaning and feature engineering, explanatory analysis, split training and validation, and model implementation.

Data Cleaning and Feature Engineering

The dataset is directly downloaded from the Kaggle website. There are 17 input features and one target variable.



Variable Name	Description
HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
BMI	Body Mass Index
Smoking	Whether the patient has smoked at least 100 cigarettes in the entire life
AlcoholDrinking	Whether the patient is a heavy drinker
Stroke	Whether the patient has ever had a stroke
PhysicalHealth	Number of days during the past 30 days was the physical health not good
MentalHealth	Number of days during the past 30 days was the mental health not good
DiffWalking	Serious difficulty walking or climbing stairs
Sex	Patient's gender
AgeCategory	Fourteen-level age category
Race	Race/ethnicity value
Diabetic	Whether the patient has ever had diabetes
PhysicalActivity	Physical activity during the past 30 days other than their regular job
GenHealth	Rating of the general health
SleepTime	Hours of sleep on average
Asthma	Whether the patient has ever had asthma
Kidney Disease	Whether the patient has ever had kidney disease
SkinCancer	Whether the patient has ever had skin cancer

This dataset consists of numerical variables such as *BMI* and *SleepTime*, binary categorical variables including *Smoking* (whether the patient smokes) and *Sex*, as well as multi-value categorical variables including *AgeCategory* and *Race*. There is no missing value in this dataset.

In this section, data cleaning and feature engineering will be conducted to transform all variables into numerical ones. Firstly, the binary variables with possible values “Yes” and “No” are transformed into numerical values 1 and 0. Secondly, for the binary variable with possible values other than “Yes” and “No”, such as the *Sex* variable with values “Female” and “Male”, a new variable *Sex_female* will be constructed with numerical values 1 and 0. Thirdly, for the multi-value categorical variables, their possible values will be examined first and transformations will be determined accordingly.

In this dataset, there are three multi-value categorical variables, including *AgeCategory*, *Race,* and *GenHealth* (general health condition). For *AgeCategory*, the possible values consist of “18-24”, “25-29”, “30-34”, “35- 39”, “40-44”, “45-49”, “50-54”, “55-59”, “60-64”, “65-69”, “70-74”, “75-79”, and “80 or older”. The transformation for this variable transforms each age range into its median value (“80 or older” is transformed into 80). For *Race*, a binary variable is constructed for each possible value, such as *Race_Hispanic*, denoting whether the patient belongs to a certain race. For *GenHealth*, the possible values consist of “Poor”, “Fair”, “Good”, “Very Good”, and “Excellent”. Considering these values have order information, they are transformed into numerical values 1~5 with 1=“Poor” and 5=“Excellent”.

It is noted that this dataset is huge (319795 patients in total) and highly imbalanced (27373 patients have heart disease while 292422 patients don’t have heart disease). 50000 patients (25000 with heart disease and 25000 without heart disease) are sampled without replacement to construct the balanced dataset with appropriate size.

Exploratory Data Analysis

For exploratory data analysis, the variables’ distributions by the *HeartDisease* variable are examined. When grouped by the values of *HeartDisease*, variables *PhysicalHealth* and *Age*’s distributions are very different between the two groups, which are shown in Figure 1 and 2.

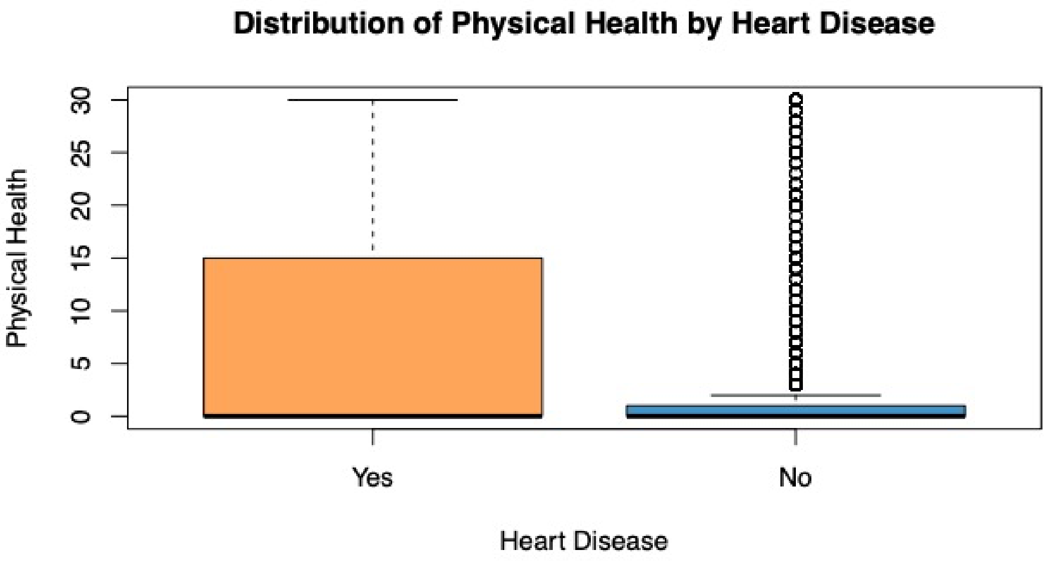


Figure 1: Boxplot of variable *PhysicalHealth*'s distribution by variable *HeartDisease*.

From the distribution of physical health by heart disease shown in Figure 1, patients with heart disease tend to have higher values of *PhysicalHealth*. Recall that this variable denotes the number of days during the past 30 days was the patient's physical health not good, it makes sense for patients with heart disease to have higher values for this variable.

From the distribution of age by heart disease shown in Figure 2, patients with heart disease tend to have higher age than those without heart disease. This brings us insight that age may have influence on the risk of having heart disease.

Besides boxplots, a correlation matrix of this dataset is shown in Figure 3. From this figure, variable *HeartDisease* is more correlated with variables *Age, GenHealth, Diabetic, PhysicalHealth* and *DiffWalking*. Among these variables, *GenHealth* is negatively correlated with *HeartDisease*, while the others are positively correlated with *HeartDisease*.

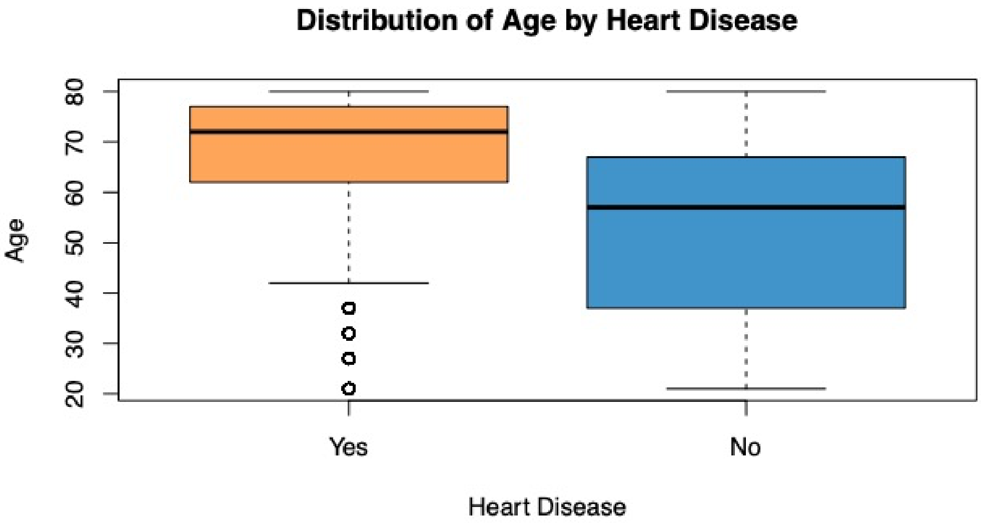


Figure 2:Boxplot of variable Age’s distribution by variable HeartDisease.

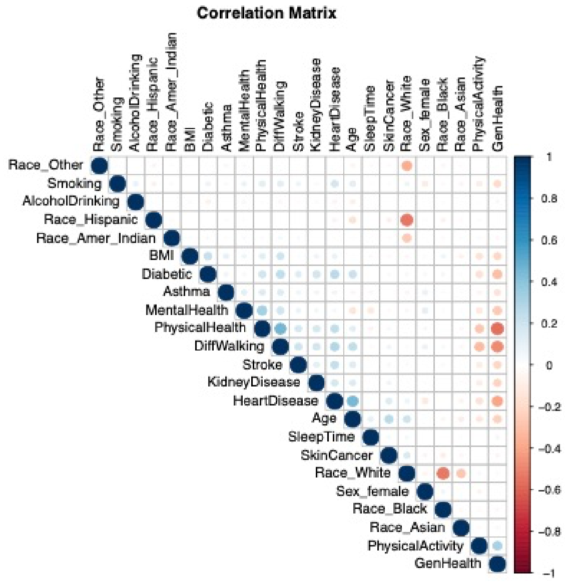


Figure 3: Correlation matrix of the dataset

Methods

Split Heldout Set

The whole dataset is split into a training set and a heldout set with 80% and 20% of the data respectively. The heldout set is never used in the model training process, and the training set is used to tune the parameters of models as well as train the final models. During the parameter tuning process, the training set will be further split into training and validation sets. A random seed of 2022 is set for the reproducibility purpose.

Selection of Models

The selected baseline model for this project is the simple model that predicts all samples to belong to the most common class. Considering this dataset has equal number of patients with or without heart disease, the baseline model is set to predict all patients to have heart disease and achieves a 50% classification accuracy.

Selected models for this classification task consist of Lasso, decision tree and random forest. Lasso is selected as there are many input features in this dataset, some of which may be irrelevant. Therefore, lasso can serve as the feature selection method that can eliminate irrelevant features. Decision tree method is selected for a similar reason, as decision tree chooses the most predictive features at each split and ignores the irrelevant features. Random forest is selected since it is an ensemble method consisting of multiple decision trees and possibly an improvement of the decision tree model.

Evaluation

Two evaluation metrics are considered for this project: (1) classification accuracy and (2) confusion matrix consisting of “True Positive”, “True Negative”, “False Positive” and “False Negative” values, which provides more information about the classification.

Parameter Tuning and Best Model

All models have a parameter tuning step which optimize the model’s parameters. The best model is selected as the model with the optimal parameter obtained from the parameter tuning step.

Lasso

Lasso is a statistical method which performs variable selection and L1 regularization to improve the prediction accuracy and reduce the risk of overfitting. For this model, the parameter lambda is tuned by k-fold (k=10) cross validation on the training set. Figure 4 shows the parameter tuning plot for lasso model.

From lasso’s parameter tuning step, the optimal parameter lambda is selected as the lambda.1se (largest value of lambda such that error is within 1 standard error of the cross-validated errors for lambda.min). With this optimal parameter, the final lasso model is trained on the entire training set and tested on the heldout set (testing set). The overall testing accuracy is 0.7619, and the confusion matrix is shown below:

```
##
##         0    1
##    0 3675 1296
##    1 1113 3916
```

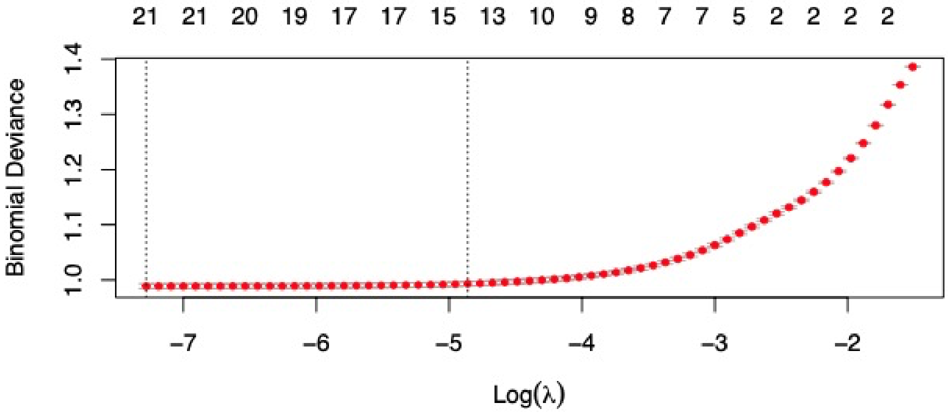


Figure 4: Parameter tuning plot for lasso model.

Decision Tree

Decision tree is a tree-structure model where each internal node represents a "test" on an attribute, each branch denotes the outcome of the "test", and each terminal leaf node denotes a decision class label. For this model, the tuned parameter is the size of the pruned tree. During the parameter tuning process, the entire training set is further split into training set (75%) and validation set (25%). The decision tree is trained and pruned on the training set and validated on the validation set. The parameter tuning plot in shown in Figure 5.

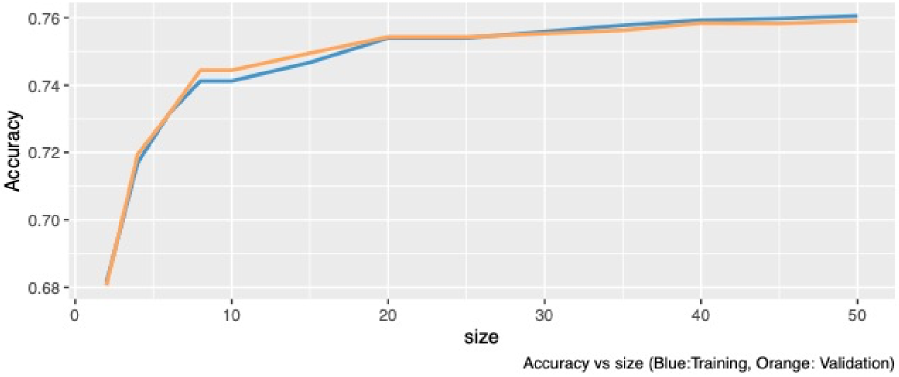
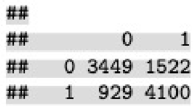


Figure 5: Parameter tuning plot for decision tree model.

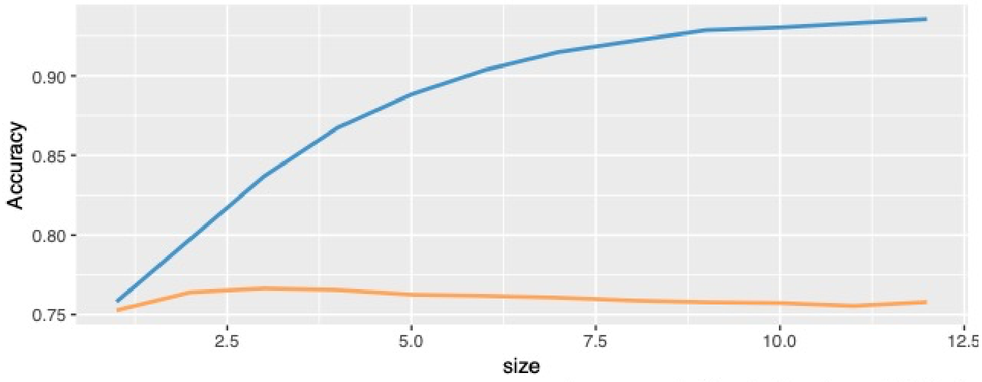
From decision tree's parameter tuning step, the optimal parameter is selected as the smallest size with comparably best validation accuracy: size=35. With this optimal parameter, the final decision tree model is trained and pruned on the entire training set and tested on the heldout set (testing set). The overall testing accuracy is 0.7566, and the confusion matrix is shown below:



```
##
##        0    1
##    0 3449 1522
##    1  929 4100
```

Random Forest

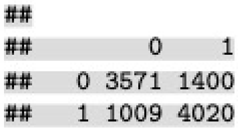
Random forest is an ensemble learning method that constructs multiple decision trees. For this classification task, the output of the random forest is the class predicted by most trees. For this model, the tuned parameter is mtry (the number of variables randomly sampled as candidates at each split). During the parameter tuning process, the entire training set is further split into training set (75%) and validation set (25%). The random forest is trained on the training set and validated on the validation set. The parameter tuning plot in shown in Figure 6.



Accuracy vs size (Blue:Training, Orange: Validation)

Figure 6: Parameter tuning plot for random forest model.

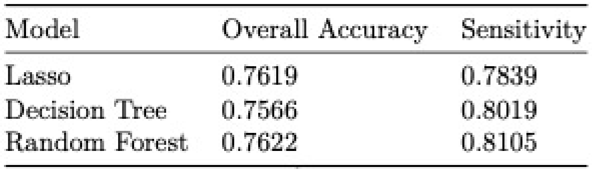
From random forest's parameter tuning step, the optimal parameter is mtry=4. With this optimal parameter, the final random forest model is trained on the entire training set and tested on the heldout set (testing set). The overall testing accuracy is 0.7622, and the confusion matrix is shown below:



```
##
##        0    1
##    0 3571 1400
##    1 1009 4020
```

Conclusion

In this project, three models with optimal parameters are trained for predicting the risk of heart disease. Considering the context of this problem, it is more important to correctly predict patients with heart disease than patients without heart disease. Therefore, the metric of sensitivity (true positive rate) that refers to the probability of a positive test conditioned on truly being positive is important. This metric can be computed from the confusion matrix. Results of overall accuracies and sensitivities are summarised in the table below.



Model	Overall Accuracy	Sensitivity
Lasso	0.7619	0.7839
Decision Tree	0.7566	0.8019
Random Forest	0.7622	0.8105

From this table, random forest has the highest values for both overall accuracy and sensitivity. Therefore, it is safe to conclude that random forest has the best performance for this task.