Heart Disease Prediction

Team 5: Fun Times

Tanush Sharanarthi, Hsueh-Yao Lu, Shreyas Iyer, Aishah Matsatsa, Titus Yong

Johns Hopkins Carey Business School

BU.330.780.T1.SP22 Data Science and Business Intelligence

19 May 2022

Business Understanding

Cardiovascular diseases (CVDs) are the global leading cause of death, attributing to 31% of all deaths and claiming 17.9 million lives each year (CDC, 2022). Eighty percent of these deaths have been attributed to heart attacks and strokes and one-third of these occur in individuals under 70 years of age (CDC, 2022). In the United States, it is estimated that someone has a heart attack every 40 seconds and about 805,000 Americans have a heart attack every year (Thomas, 2020). Due to the large number of individuals impacted, individuals with CVDs or who carry a higher risk of CVDs would greatly benefit from early detection and management through the use of a machine learning model. We want to know what factors significantly contribute if a person has heart disease and if we can predict it using a machine learning model.

The use of electronic health records (EHR) has dramatically increased the quantity of clinical data allowing for easier analysis to gain better healthcare insights (Dash et al., 2019). Therefore, EHR systems can collect and store meaningful data including a number of characteristics that contribute directly to whether a person has heart disease. As a result, there are unprecedented opportunities to use big data and machine learning to reduce the costs of health care in the United States. Through data mining, we can set our target variable as having heart disease and analyze which characteristics contribute most to it (Liu et al., 2017). Policymakers and healthcare providers can implement policies on healthcare reform, improved quality care, educate the most high-risk patients, and save healthcare costs. We can use decision tree, random forest, and logistic regression models to determine which features are the most important indicators to influence heart disease and also determine which model is most accurate and reliable in predicting heart disease.

To combat heart disease, most healthcare providers previously focused on encouraging a healthy lifestyle and preventative education such as reducing alcohol and

tobacco use, eating healthy, and exercising. But this is not enough. If a person has an annual physical, it is easier to catch risk factors and provide early treatment. It is difficult to identify high-risk patients because of the multi-factorial nature of several contributory risk factors such as chest pain type, old peak, high blood pressure, high cholesterol, and more (Bates et al., 2014). This is where machine learning and data mining become useful.

Data Understanding and Visualization

The dataset we used for our model is the Heart Failure Prediction Dataset from Kaggle. It contains 918 observations and 12 attributes with the target variable being HeartDisease where output 1 refers to the patient having heart disease and 0 refers to a patient not having heart disease. The data dictionary is **Table 1** in the appendix.

Heart failure is commonly associated with CVDs, therefore, using this dataset that contains the above attributes a machine learning model that could be used to predict heart disease can be built. Since early detection of CVDs is vital to saving a patient's life and only through early detection doctors can take the necessary measures to help the patients we can analyze the risk factors such as diabetes and hypertension to see which attributes are highly likely to lead to heart failure.

The modeling was analyzed on RStudio. The dataset is a CSV file containing a combination of 5 heart datasets. In RStudio the *tidyverse* library was used to make it easy to load all the *tidyverse* packages which can be used for visualization using *ggplot*.

Before performing Exploratory Data Analysis (EDA) the data was cleaned to determine if there were any missing values or duplicates. Upon checking the data no missing values or duplicates were found. The data was also described to show the counts of the categorical variables and the distribution of the continuous variables. Since there are several issues with the data we made some modifications. Since RestingBP and Cholesterol cannot

be 0 we changed the zero values to NA. We also re-coded, the Y/N values in ExerciseAngina to 0/1 values to make the analysis more manageable and we factorized the HeartDisease column.

With this, we could start visualizations to draw some insights. We used both R and Tableau to perform visualizations. By plotting a countplot of the sex feature we could see that the data had a majority of male patients and it is clear from the plot that males were far more susceptible to heart diseases than females were. However, as shown in **Figure 1**, it seems that there are many more observations of males than females so we may still need to obtain more data for a better understanding of the data.

Next, we plotted a histogram, **Figure 2**, of the age and observed that the patients in the age group between 55 and 65 years old were most vulnerable to heart diseases. This indicated that age could be a key predictor in our analysis of heart diseases.

We then plotted a countplot of the ChestPainType, as seen in **Figure 3**. Here we found that patients with ASY or asymptomatic chest pain were the most susceptible to heart diseases. This may have been because an asymptomatic patient would have not known anything was wrong since there was no physical indicator that they were suffering from heart disease. This could also explain why such a patient would go under the radar when a doctor examined them.

Next, we plotted the RestingBP, **Figure 4**. The data points were primarily concentrated within 100 and 120. Since the graph did not give us many details, it's possible that Resting BP may not be a good indicator of heart disease. Finally, we plotted the Cholesterol level, **Figure 5**. The data points were concentrated around 200 to 300 and resembled a normal distribution for both patients with heart disease and those without.

From our visualizations, we were able to observe several relationships between the different features and the target variable. Based on this preliminary analysis we can build models and see whether our assumptions were in line with the results.

Modeling

Since our goal is to predict whether a person is vulnerable to heart diseases, the dependent variable is a categorical variable indicating whether a person has heart disease or not. Therefore, we decided to train our data set with three different classification models: decision tree, random forest, and logistic regression, and compare the results to find out the most optimal model for our data set. We will first build our models with all of our data to have a generic understanding of them. Then, we will use the hold-out sample method to check how each model performs on validation sets.

Decision Tree

First, we built a decision tree model. The advantages of using a decision tree are that it requires less data pre-processing compared to other data mining methods. It also does not require normalization or scaling of the data and is not affected by missing values. The most important attribute is that a decision tree is easy to interpret visually and easy to explain to other people how each variable contributes to the final outcome. There are also some downsides with a decision tree. First, it is not as accurate as other kinds of models, and it can only be used on classification outcomes. Also, it is sensitive to outliers, and adding a single data can change the result dramatically. Last but not least, it takes a long time to train decision trees and it is computationally more expensive than other models.

Looking at the decision tree (see **Figure 6**), we can see that the most critical variables that affect the outcome are ST_Slope, ChestPainType, and Oldpeak. We limited the depth of the tree to be 4 to prevent the problem of overfitting.

Random Forest

Next, we built a random forest model. Random forest models normally have higher accuracy than decision tree models, and they have a balanced bias-variance trade-off. Like the decision tree models, random forest models do not require scaling or transformation of the variables. On the other hand, unlike the decision tree models, random forest models can work with both classification and regression tasks, and do not get influenced by outliers. The downsides of random forest models are that they are not as interpretable as decision tree models, and they cannot handle missing data. Also, they are computationally intensive and we have limited control over what we can do with the model.

We built the random forest model (see **Figure 7)** with the variable importance plot (see **Figure 8**), and we found that ST_Slope is the most important variable, along with Oldpeak and ChestPainType

Logistic Regression

Lastly, we built a logistic regression model with all of the independent variables. The advantages of logistic regression include its good interpretability and easy implementation. It works particularly well with linearly separable data and is able to work with big data with efficient computation. It is also unlikely for a logistic regression model to overfit a set of training data. Some of the disadvantages are the inability to handle missing data, assumption of linearity between dependent and independent variables and it is more sensitive to outliers.

The results of the logistic regression model (see **Figure 9**) shows us that Age, Sex, ChestPainType, ExerciseAngina, Oldpeak, and ST_Slope are the most important variables.

Combining the results of all three models, we can conclude that ST_Slope, Oldpeak, and ChestPainType are the three most important variables when predicting whether a person has heart disease.

Next, we use the hold-out sample method, which is to separate our data into training and testing data sets, and train the three models using the training data set. For the decision tree model, we pruned it by setting the complexity parameter to be the ones with the lowest X-value relative error. For the random forest model, we tuned it by using the entry with the lowest OOB error. For the logistic regression model, we used forward and backward selection and use the set of variables with the lowest AIC for the new model. After that, we use the testing data set to test our model accuracy, and build a confusion matrix and ROC curve, which would all be evaluated in the next section.

Evaluation

In our Heart failure prediction project, we have used three machine learning classifiers. The classifiers used are decision trees, random forests, and logistic regression. Firstly, we will compare the accuracies of the three models used on the whole dataset (see **Table 2**). As we can see, if we take the accuracy as a metric, the logistic regression model is the most accurate followed by the random forest model and then the decision tree model. Next, we tune and validate the models further to get more accurate metrics.

Decision Tree

Firstly in the decision tree model, we will create a holdout sample ie. split the dataset into training and testing, prune the decision tree and test it on the testing sample. After this, we get the three most important features as ST_Slope, ChestPainType, and OldPeak. Based on the resulting confusion matrix (see **Table 3**), the accuracy of the test dataset is (56+99)/183 = 0.847.

Random Forest

Next we will evaluate the random forest model. In this model, we will again create a holdout sample and perform hyperparameter tuning. The OOB estimate of error is 13.23% (see **Figure 10**). Based on the resulting confusion matrix (see **Table 4**), the accuracy of the test dataset is (68+62)/ 149 = 0.872.

Logistic Regression

In this model, we will again create a holdout sample, and perform stepwise regression. Based on the resulting confusion matrix (see **Table 5**), the accuracy of the test dataset is (61+66)/149 = 0.852. After this, we get the most important variables for the logistic regression model as Age, Sex, ChestPainType, ExerciseAngina, Oldpeak, and ST_Slope. So we can conclude that when we use accuracy as our metric, the random forest is the best model, followed by logistic regression, and the worst is the decision tree.

Performance Visualisation with ROC

Looking at the visualization (see **Figure 11**) and by taking AUC as our metric we can conclude that the logistic regression is the best model followed by random forest, and the worst is the decision tree.

This model can be used to accurately predict if a person has heart disease or not based on the given features. Since we will want to predict this target variable accurately, accuracy is an important metric that we should consider while evaluating the model. Simultaneously we must also look at the AUC since it is important that the model can clearly distinguish between the positive and negative classes. This is because we want to minimize any false positives or false negatives that might cause any unnecessary risks to the person. If we take into account both these metrics we can see that the random forest is the best model since it has a good AUC and accuracy.

The results from this model can be evaluated using real-life examples in hospitals and clinics where real-time tests are run on patients along with the model to compare and analyze the results to check the reliability of the model. The business case of utilizing this model as a standalone solution in the future in hospitals can be possible but since in the medical field, reliability is a concern the model should be coupled with real-time testing until the confidence for its accuracy is really high and has proven accurate results in testing.

As time goes we should be able to get more data to train the model and make it more and more reliable and can be used as a standalone solution. The entire process of evaluation and improvement should not be too difficult and if faced with any setbacks, alternative solutions such as using more sets of people ie data points and identifying other features that can impact the final target variable should be discovered. Limiting our model to the new strongest impact features and coupled with the features we have already identified to be strong predictors will help improve our model further and improve the accuracy and its ability to distinguish the two classes of having heart disease and not having disease clearly.

Deployment

In deploying the results from this study, we will be able to determine if a patient is likely to fall into the category of being more vulnerable to heart disease based on ST_Slope, ChestPainType, and OldPeak. Those individuals with asymptomatic chest pain, ST_SlopeFlat, and a higher OldPeak should be more closely assessed for the possibility of heart diseases. We suggest for these individuals to undergo a more detailed review for potential heart diseases.

There were some issues that if better addressed could help us to provide a more accurate analysis. For a start, we were not certain about the specifics behind how the data was collected. There were more males than females in the dataset and attributes that we had

assessed to be significant by visually interpreting the graphs (age and asymptomatic chest pain) were different from the models that we ran (ST_Slope, ChestPainType, and OldPeak).

Having an early detection of heart disease could help in saving lives. However, there could be some ethical issues that need to be considered as well. The firm should ensure that they do not jump to preemptive conclusions, patient's consent is obtained prior to storing the data, and that the data is anonymized and protected. We have provided a flow chart to depict the entire process (See **Figure 12**).

The firm should be careful and do closer tests to determine if the patient has heart disease as compared to preemptively determining that the patient has heart disease just based on the three points of data of asymptomatic chest pain, ST_SlopeFlat, and higher OldPeak. Preemptively notifying patients that they have heart disease without proper medical assessment could result in undue stress for the patient as well as possible lifestyle changes and even affect the mental wellbeing of the patient. There would be costs incurred for the firm should the patient decide to sue the firm for making an inaccurate assessment.

Prior to collecting the data, the firm should get the patient’s consent so as to respect their privacy, especially with the prevalence of privacy laws. Following this, the firm needs to also be able to protect its data well to ensure that there is no data leak or breach. The data that has been collected from patients in forming the model should be anonymized so that in the event of a data breach, patients’ privacy would not be compromised.

By taking the above steps, we believe that more lives can be saved as heart diseases can be detected in their earlier stages and the additional data that is collected can make the model even more accurate, thereby saving even more lives.

References

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 33(7), 1123-1131. https://doi.org/10.1377/hlthaff.2014.0041

CDC. (2022, February 7). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm#:~:text=Heart%20Disease%20in%20the%20United%20States&text=About%20659%2C000%20people%20in%20the,1%20in%20every%204%20deaths.&text=Heart%20disease%20costs%20the%20United,year%20from%202016%20to%202017.&text=This%20includes%20the%20cost%20of,lost%20productivity%20due%20to%20death

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis, and future prospects. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0217-0

Holypython. (2021, June 29). Logistic regression pros & cons. HolyPython.com. https://holypython.com/log-reg/logistic-regression-pros-cons/

K, D. (2020, December 26). Top 5 advantages and disadvantages of decision tree algorithm. Medium. https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method. Computational and Mathematical Methods in Medicine, 2017, 1-11. https://doi.org/10.1155/2017/8272091

Singh, J. (2020, December 26). Random forest: Pros and cons. Medium. https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04

Thomas, J. (2020, July 16). Facts and statistics on heart disease. Healthline. https://www.healthline.com/health/heart-disease/statistics#How-much-does-it-cost?

Appendix

Table 1

Age	age of the patient [years]
Sex	sex of the patient [M: Male, F: Female]
ChestPainType	chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
RestingBP	resting blood pressure [mm Hg]
Cholesterol	serum cholesterol [mm/dl]
FastingBS	fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG	resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
MaxHR	maximum heart rate achieved [Numeric value between 60 and 202]
ExerciseAngina	exercise-induced angina [Y: Yes, N: No]
Oldpeak	oldpeak = ST [Numeric value measured in depression]
ST_Slope	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
HeartDisease	output class [1: heart disease, 0: Normal]

Table 2

Model	**Accuracy**
Decision Tree	0.841
Random Forest	0.859
Logistic Regression	0.87

Table 3

	actual	
predicted	No	Yes
No	56	7
Yes	21	99

Table 4

	actual	
predicted	No	Yes
No	62	8
Yes	11	68

Table 5

	actual	
predicted	No	Yes

No	61	10
Yes	12	66

Figure 1

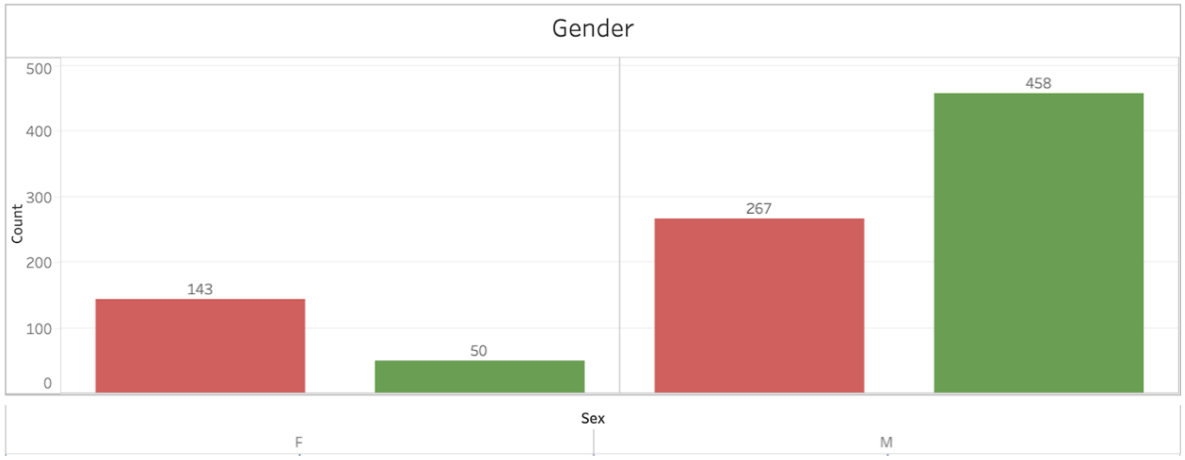

Figure 2

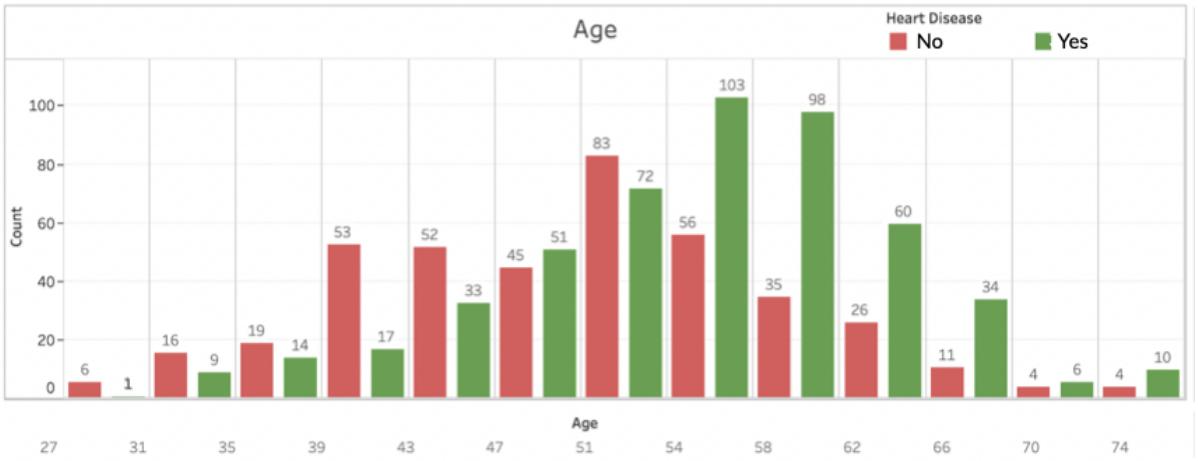

Figure 3

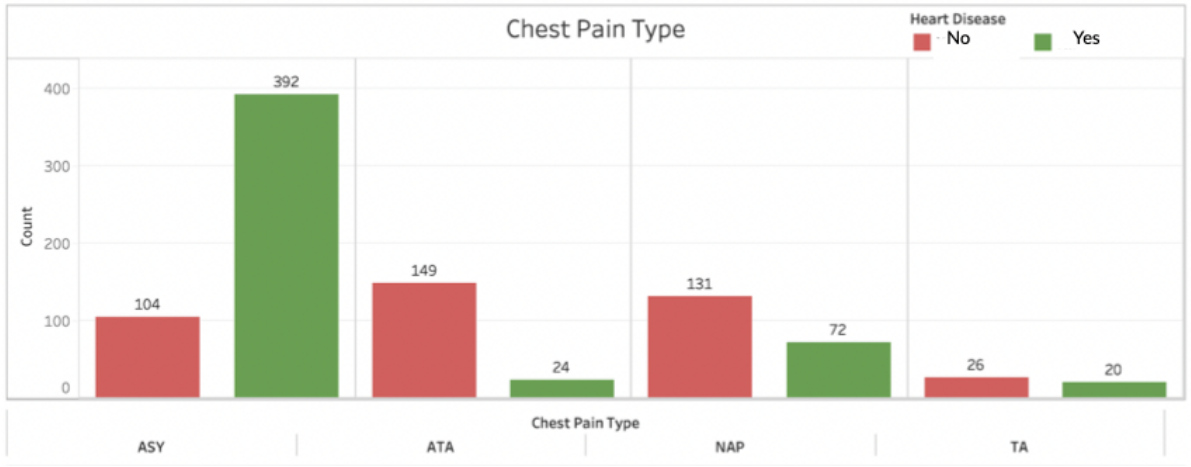

Figure 4

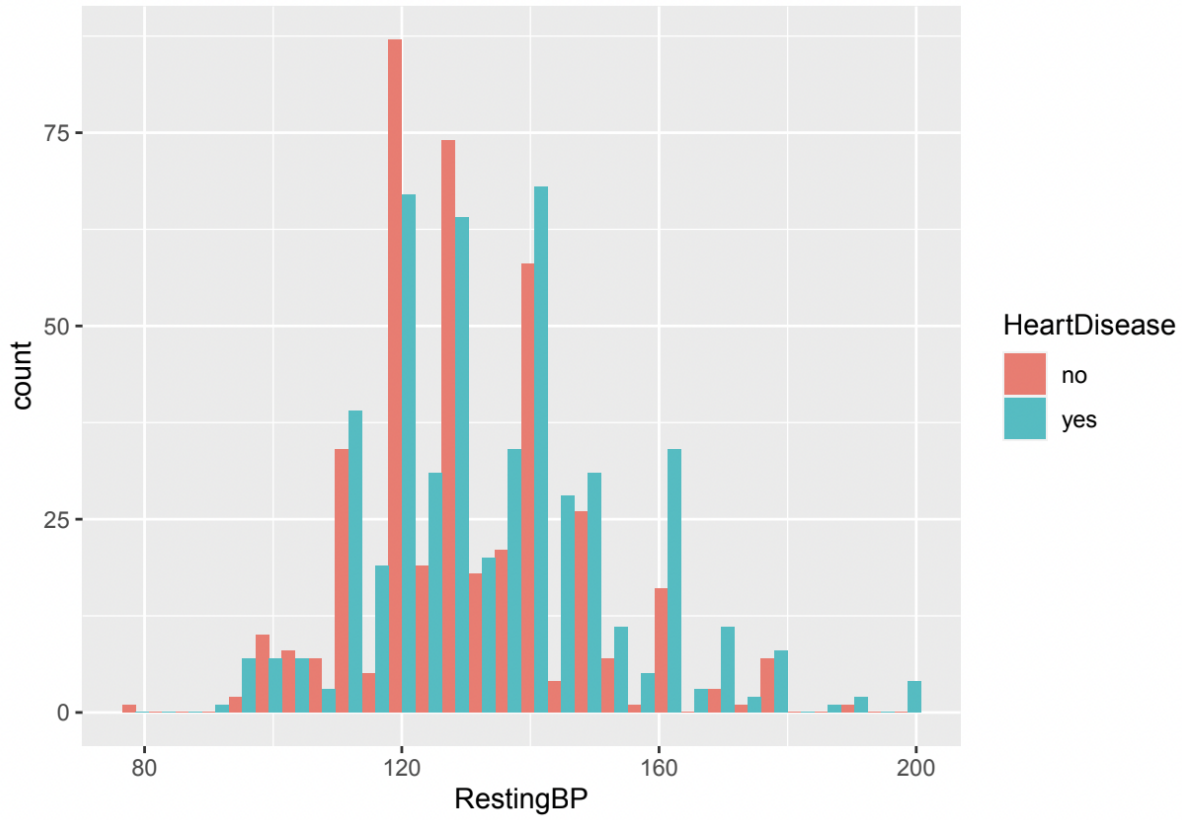

Figure 5

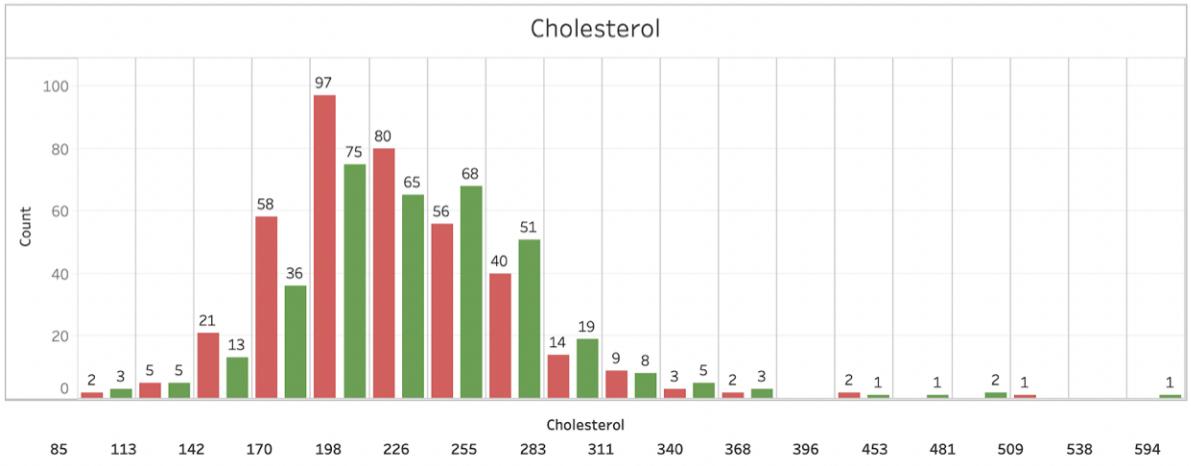

Figure 6

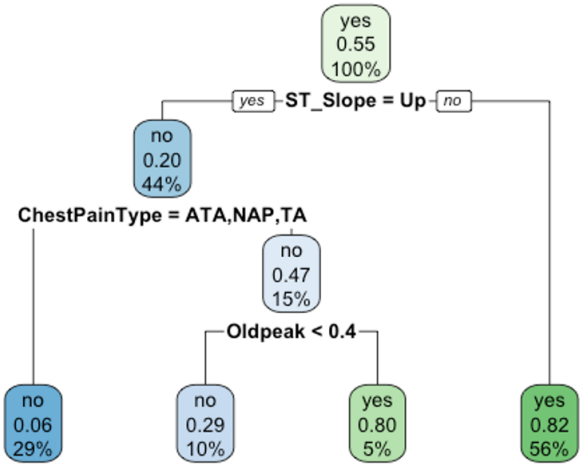

Figure 7

```
Call:
 randomForest(formula = HeartDisease ~ ., data = training, ntree = 500,      cutoff = c(0.5, 0.5), mtry = 2, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 13.23%
Confusion matrix:
     no yes class.error
no  278  39   0.1230284
yes  40 240   0.1428571
```

Figure 8

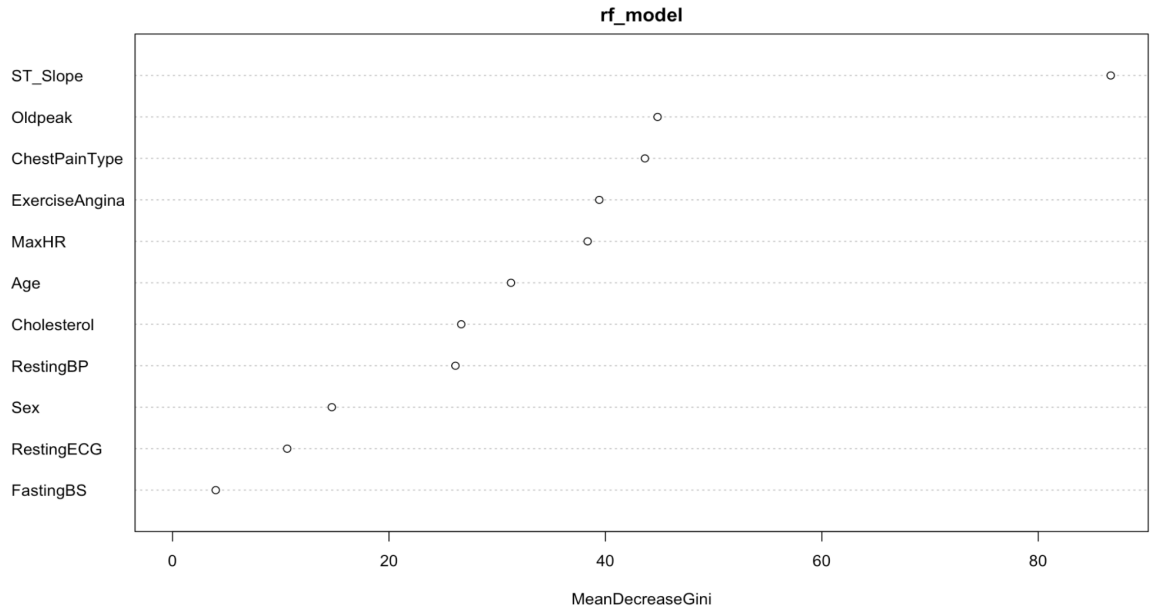

Figure 9

```
Call:
glm(formula = HeartDisease ~ ., family = "binomial", data = heart_disease)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6697  -0.3841  -0.1085   0.4465   2.7371

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.4373046  1.7625169  -3.085 0.002036 **
Age               0.0313784  0.0148105   2.119 0.034119 *
SexM              1.8655490  0.3134065   5.952 2.64e-09 ***
ChestPainTypeATA -1.6731804  0.3544226  -4.721 2.35e-06 ***
ChestPainTypeNAP -1.5730121  0.3029404  -5.192 2.08e-07 ***
ChestPainTypeTA  -1.6332529  0.4838117  -3.376 0.000736 ***
RestingBP         0.0117792  0.0072988   1.614 0.106557
Cholesterol       0.0024955  0.0019773   1.262 0.206928
FastingBS         0.2923999  0.3311265   0.883 0.377212
RestingECGNormal -0.2297888  0.2842091  -0.809 0.418791
RestingECGST     -0.1746017  0.3941671  -0.443 0.657792
MaxHR             0.0005807  0.0057810   0.100 0.919991
ExerciseAngina    0.9073515  0.2671360   3.397 0.000682 ***
Oldpeak           0.4108355  0.1406671   2.921 0.003493 **
ST_SlopeFlat      1.3038217  0.5197574   2.509 0.012124 *
ST_SlopeUp       -1.2100372  0.5655279  -2.140 0.032382 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1032.63  on 745  degrees of freedom
Residual deviance:  483.58  on 730  degrees of freedom
  (172 observations deleted due to missingness)
AIC: 515.58

Number of Fisher Scoring iterations: 6
```

Figure 10

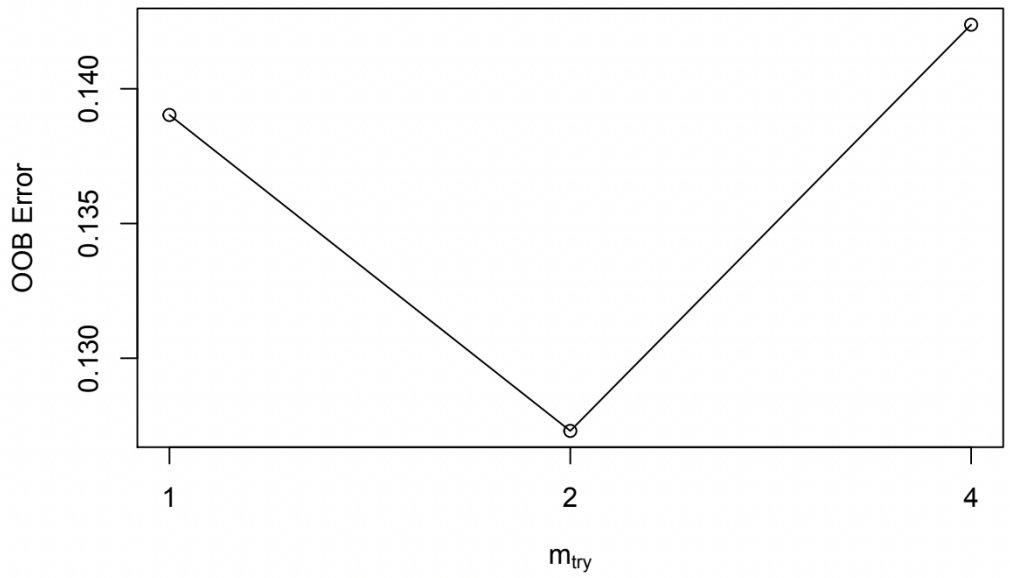

Figure 11

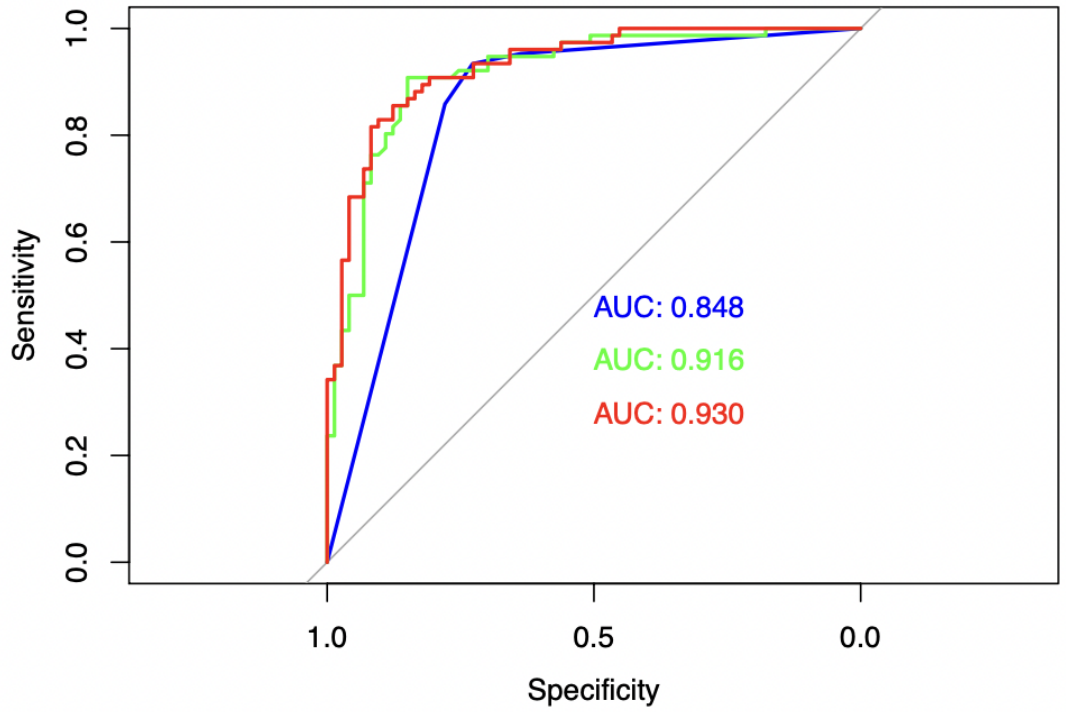

Figure 12