Final-Report

Alex Shen, Steven Yuan, Conner Byrd

12/13/2021

I. Introduction

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. CVDs include a range of conditions that include blood vessel disease, such as coronary artery disease; heart rhythm problems (arrhythmias); heart defects at birth (congenital heart defects); heart valve disease; disease of the heart muscle; heart infections, and many more. Although many forms of CVD can be prevented or treated with healthy lifestyle choices, some can not. CVDs are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 32% of all global deaths. Over 85% of deaths from CVD were due to heart attack and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Cardiovascular diseases (CVDs) are the number one cause of death globally. ADD CITATION

Currently, there are several different ways for physicians to diagnose patients that they believe to be at risk for Cardiovascular Diseases (CVDs). The practices vary by country, but often include the physician checking the patient's blood pressure, cholesterol level, and conducting further tests such as exercise stress tests, X-rays, etc. Currently, there are many issues with the current diagnostic methods. A [study] (https://www.sciencedaily.com/releases/2009/11/091116103435.htm) of 500 patients found a false positive reading between 77 and 82 percent in patients in patients at risk of CVD screened by ECG, and a false negative reading between 6 to 7 percent in the same patient population. People with CVDs or who are at high risk of CVDs need early detection and management wherein a machine learning model can be of great help.

Using our Cardiovascular Heart Disease data, we have two main goals. Our first goal is to create models for the purpose of prediction. These can be used to assess the likelihood of a heart disease diagnostic for potential at-risk patients based on a number of factors such as age of the and sex of the patient, blood pressure, cholesterol, heart rate, and the presence of chest pain.

Our second goal is to create models for the purpose of interpretation, which can be used to provide a greater understanding of signs that at-risk patients can analyze to check their risk for CVDs.

We chose to fit 3 different models to classify whether a patient has heart disease or not. The first model is a logistic regression model with variable selection performed by a backwards selection using AIC. We decided to use this model because of its interpretability so that we can examine the relationship our predictors have with the probability of a patient having heart disease. We also decided to use a random forest and a SVM because of their ability to perform classification and due to their predictive power despite their lack of interpretability. We will use a 10-fold CV to determine the best predictive model based on the classification error of each model. Due to the ability of ensemble models to reduce prediction error, we will also create an ensemble model where the prediction is the most common result of the 3 individual models (https://www.sciencedirect.com/topics/computer-science/ensemble-modeling). Finally, we compare the 10-fold CV errors for each individual model as well as the ensemble to find the one with lowest classification error and therefore highest predictive accuracy.

II. Data

We obtained our data from Kaggle (https://www.kaggle.com/fedesoriano/heart-failure-prediction). The dataset was originally provided by Dr. David Aha, a researcher at the US Naval Research Laboratory. It was

created by combining five heart datasets from Cleveland (303 observations), Hungary (294 observations), Switzerland (123 observations), Long Beach, Virginia (200 observations), and Stalog (270 observations), for 918 unique observations. This makes it one of the largest available heart datasets with multinational data.

The dataset has 11 predictor variables, which are listed below along with their descriptions. Out of the 11 predictor variables, five of them are categorical variables (`Sex`, `ChestPainType`, `RestingECG`, `ExerciseAngina`, and `ST_Slope`), while the rest are numeric. The categorical variables have been factored for this report. For our purposes, the dataset has a single response variable, `HeartDisease`, described further below, which is whether or not the patient has been diagnosed with heart disease.

Variables	Description	Value
Age	Age of the patient in years	Numeric Value
Sex	Sex of the patient	[M/F]
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, LVH: showing probable or definite left ventricular hypertrophy]
MaxHR	Maximum heart rate achieved	[Numeric value between 60 and 202]
ExerciseAngina	Exercise-induced angina	[Y: Yes, N: No]
ST_Slope	the slope of the peak exercise ST segment	[Up: upsloping, Flat: flat, Down: downsloping]
Oldpeak	The level of exercise relative to rest	Numeric value
HeartDisease	Output class denoting if patient has Heart Disease	[1: heart disease, 0: Normal]

Missing Data

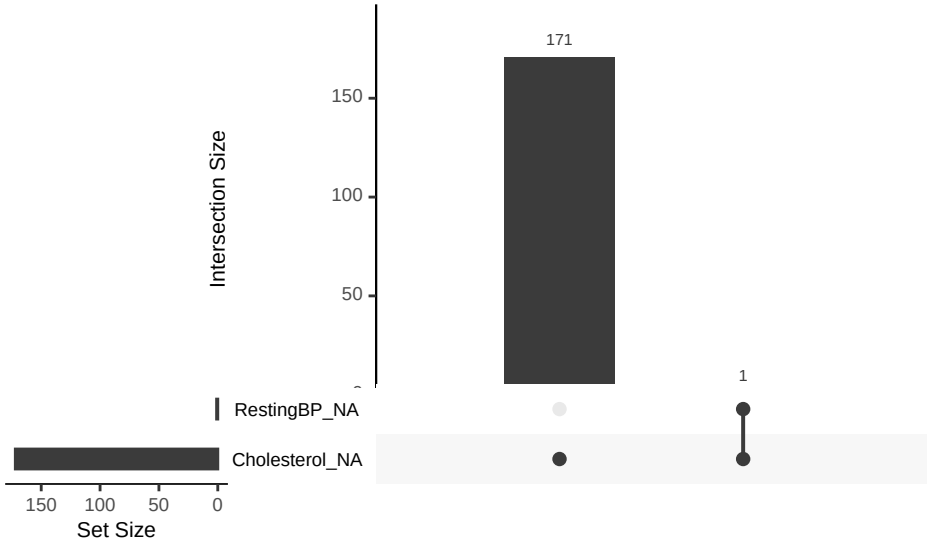


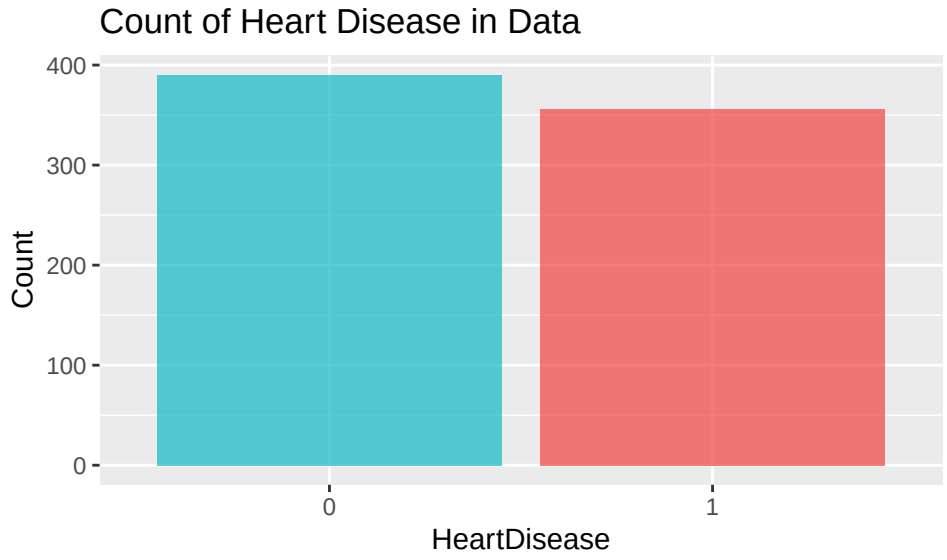
Figure 1: Missing Data Visualization

There were a total of 172 observations with missing values for `Cholesterol` and `RestingBP`. (Cholesterol

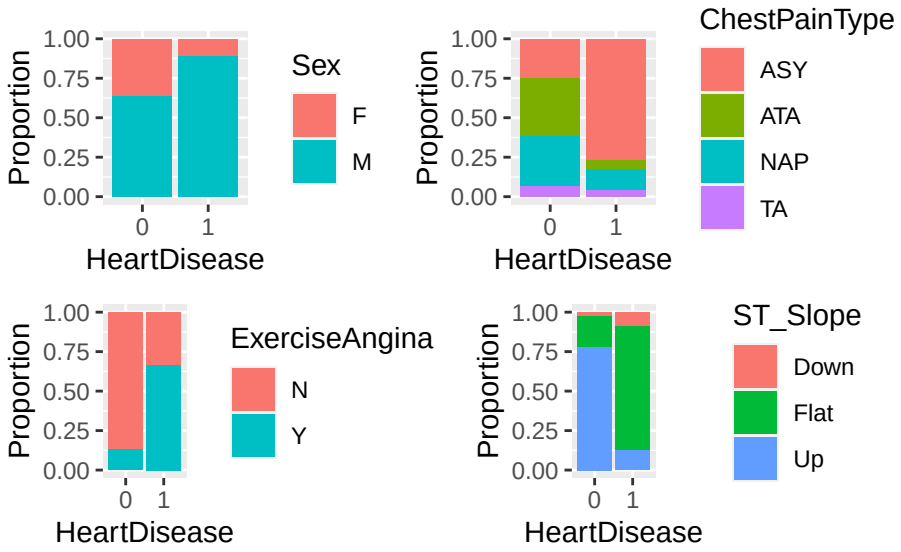
value of 0 and Resting Blood Pressure (BP) value of 0 are considered missing since these are impossible to reach in real life). The visualizations of the missing values in [Figure 1] shows that there are 171 observations with missing Cholesterol value and 1 observation with missing both Cholesterol and Resting BP value. For the one observation with missing Resting BP, it also had missing Cholesterol value. This observation was disregarded for analysis under Missing Completely at Random (MCAR) assumption. For the 171 observations with missing Cholesterol values, imputations were conducted based on Missing at Random (MAR) assumption. Multiple Imputation using Chained Equation (MICE) in R was used to impute the values. Five chains of imputations were conducted for Cholesterol with each chain using the default method of predictive mean matching (ppm) method since Cholesterol was a numerical variable. ADD CITATION

Exploratory Data Analysis

The graph below shows that in the dataset, the frequency of a positive diagnosis for heart disease is roughly equal to the frequency of a negative diagnosis of heart disease. Thus, the data is likely not based on the general population, where the frequency of heart disease is much lower. As a result, the inference and predictions from our models do not apply to the general population, but only to the population this dataset was drawn from, which is the population of patients who are at risk of heart disease and were checked for heart disease.

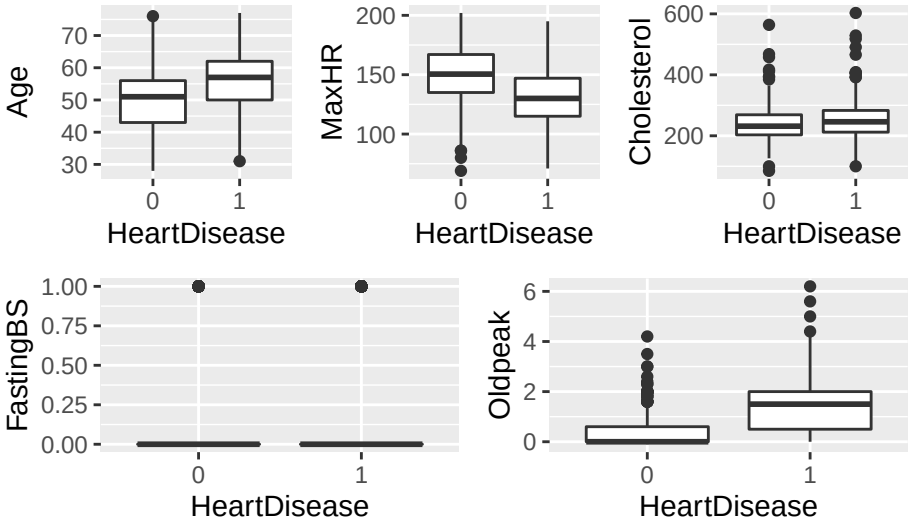


Apart from the spread of `HeartDisease`, it was important to visualize the other predictors and their relation to `HeartDisease` in the dataset. Below are plots four categorical variables (`Sex`, `ChestPainType`, `ExerciseAngina`, and `ST_Slope`)



The above 4 categorical variables, sex, chest pain types, exercise-induced angina, and ST slope seem to have some relationship with heart disease incidence, so these variables will be good to look out for in our final model. From the exploratory data analysis, it appears that a positive diagnosis for CVD tends to occur with the sex of a patient being male, asymptomatic chest pain, the presence of exercise-induced angina, and a flat slope of the peak exercise ST segment.

Next, we plotted the relationship between the incidence of heart disease and five numeric variables (`Age`, `MaxHR`, `Cholesterol`, `FastingBS`, and `Oldpeak`).



From the exploratory data analysis, it appears that the incidence of heart disease tends to occur with higher age, a lower maximum heart rate, slightly higher cholesterol, and a higher level of exercise relative to rest. These variables may be worth exploring in our models later.

III. Methodology

```
##
##  iter imp variable
##    1   1
```

```
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
##   5   4
##   5   5
```

Logistic Regression

Our first model is a logistic regression model. The logistic regression model is formulated by the equation:

$$\log(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Using logistic regression lends itself well to inference and classification goals. Our original plan was to use a lasso logistic regression to perform variable selection for all main effects plus all pairwise interactions due to the ability of lasso to shrink coefficients to 0. Then, we would add the selected variables plus any associated main effects we might need to satisfy the hierarchy principle into a normal logistic regression. However, after observing the output of the normal logistic regression, we realized that many of the p values were very large, with very few p values below an alpha significance threshold of 0.05. Since our goal with the logistic model is to be able to interpret and find relationships between predictor variables and the probability of having heart disease, we decided against using this because having very few significant terms hinders our goals.

Therefore, we decided to use a logistic regression with backwards stepwise selection using AIC. Our starting full model included all main effects and all pairwise interactions. After backwards selection, we made sure that all active interaction effects had their associated main effects included in the selected model, so that the model adheres to the hierarchy principle which aids in the interpretation of interaction effects.

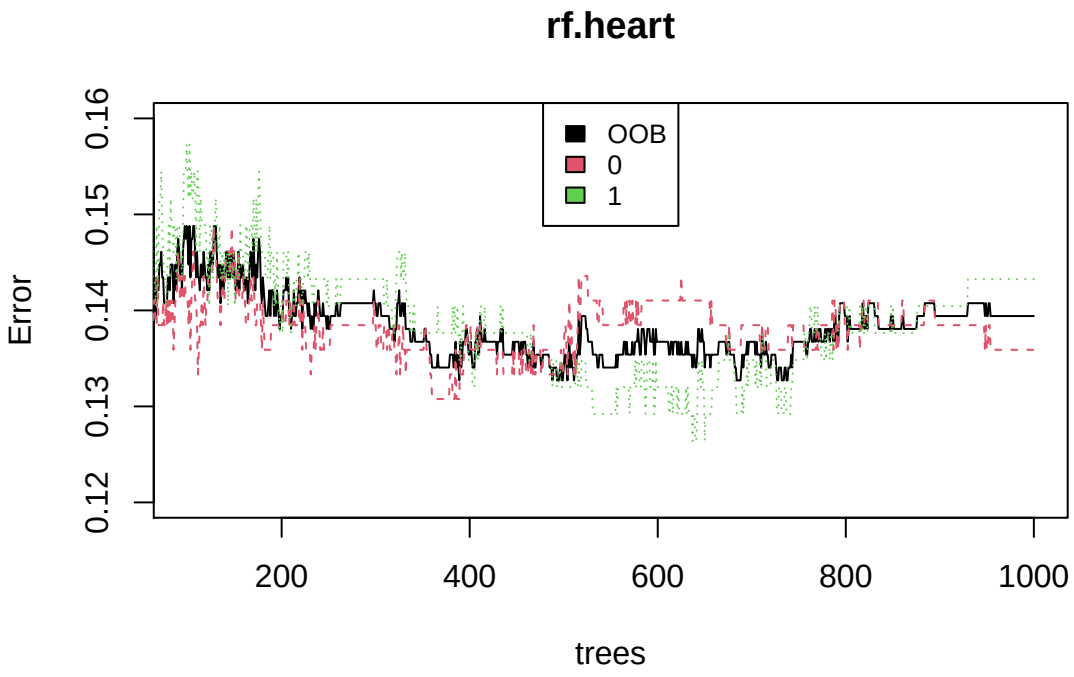
```
##
## Call:
## glm(formula = HeartDisease ~ Age + Sex + ChestPainType + RestingBP +
##     Cholesterol + FastingBS + RestingECG + MaxHR + ExerciseAngina +
##     Oldpeak + ST_Slope + Age:ST_Slope + Sex:FastingBS + Sex:MaxHR +
##     Sex:ExerciseAngina + ChestPainType:Cholesterol + ChestPainType:FastingBS +
##     ChestPainType:RestingECG + ChestPainType:ST_Slope + RestingBP:Oldpeak +
##     Cholesterol:ST_Slope + FastingBS:ST_Slope + RestingECG:ExerciseAngina +
```

```
##     MaxHR:ST_Slope, family = "binomial", data = heart)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.73396  -0.24296  -0.01559   0.31450   2.86486
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       7.569e+02  4.149e+04   0.018  0.98545
## Age                              -3.079e+00  2.190e+02  -0.014  0.98878
## SexM                              6.850e+00  2.496e+00   2.744  0.00607 **
## ChestPainTypeATA                 -1.240e+02  1.078e+04  -0.011  0.99083
## ChestPainTypeNAP                 -2.381e+02  1.269e+04  -0.019  0.98502
## ChestPainTypeTA                  -1.661e+02  7.981e+03  -0.021  0.98340
## RestingBP                         3.366e-03  1.163e-02   0.289  0.77231
## Cholesterol                      -1.288e+00  6.813e+01  -0.019  0.98492
## FastingBS                        -2.694e+01  2.457e+03  -0.011  0.99125
## RestingECGNormal                 -9.199e-01  5.275e-01  -1.744  0.08114 .
## RestingECGST                     -3.444e+00  1.093e+00  -3.152  0.00162 **
## MaxHR                            -1.051e+00  1.266e+02  -0.008  0.99337
## ExerciseAnginaY                   1.435e+00  7.662e-01   1.873  0.06113 .
## Oldpeak                          -2.188e+00  1.388e+00  -1.577  0.11476
## ST_SlopeFlat                     -7.574e+02  4.149e+04  -0.018  0.98544
## ST_SlopeUp                       -7.727e+02  4.149e+04  -0.019  0.98514
## Age:ST_SlopeFlat                  3.065e+00  2.190e+02   0.014  0.98883
## Age:ST_SlopeUp                    3.181e+00  2.190e+02   0.015  0.98841
## SexM:FastingBS                   -4.101e+00  1.362e+00  -3.011  0.00260 **
## SexM:MaxHR                       -2.668e-02  1.655e-02  -1.613  0.10684
## SexM:ExerciseAnginaY             -1.456e+00  7.182e-01  -2.027  0.04262 *
## ChestPainTypeATA:Cholesterol      2.378e-02  1.001e-02   2.375  0.01756 *
## ChestPainTypeNAP:Cholesterol     -9.446e-04  5.868e-03  -0.161  0.87212
## ChestPainTypeTA:Cholesterol       1.564e-02  1.210e-02   1.292  0.19637
## ChestPainTypeATA:FastingBS       -2.235e+00  1.550e+00  -1.442  0.14933
## ChestPainTypeNAP:FastingBS       -3.226e+00  1.357e+00  -2.378  0.01739 *
## ChestPainTypeTA:FastingBS        -3.098e+00  1.470e+00  -2.107  0.03515 *
## ChestPainTypeATA:RestingECGNormal -9.779e-01  9.677e-01  -1.011  0.31223
## ChestPainTypeNAP:RestingECGNormal  1.407e+00  7.984e-01   1.762  0.07813 .
## ChestPainTypeTA:RestingECGNormal   2.261e+00  1.157e+00   1.954  0.05074 .
## ChestPainTypeATA:RestingECGST      2.669e+00  1.958e+00   1.363  0.17291
## ChestPainTypeNAP:RestingECGST      4.768e+00  1.471e+00   3.240  0.00119 **
## ChestPainTypeTA:RestingECGST       4.925e+00  2.273e+00   2.167  0.03025 *
## ChestPainTypeATA:ST_SlopeFlat      1.175e+02  1.078e+04   0.011  0.99131
## ChestPainTypeNAP:ST_SlopeFlat      2.364e+02  1.269e+04   0.019  0.98513
## ChestPainTypeTA:ST_SlopeFlat       1.599e+02  7.981e+03   0.020  0.98402
## ChestPainTypeATA:ST_SlopeUp        1.156e+02  1.078e+04   0.011  0.99145
## ChestPainTypeNAP:ST_SlopeUp        2.350e+02  1.269e+04   0.019  0.98522
## ChestPainTypeTA:ST_SlopeUp         1.599e+02  7.981e+03   0.020  0.98402
## RestingBP:Oldpeak                  2.018e-02  1.057e-02   1.908  0.05636 .
## Cholesterol:ST_SlopeFlat           1.288e+00  6.813e+01   0.019  0.98491
## Cholesterol:ST_SlopeUp             1.290e+00  6.813e+01   0.019  0.98489
## FastingBS:ST_SlopeFlat             3.381e+01  2.457e+03   0.014  0.98902
## FastingBS:ST_SlopeUp               3.105e+01  2.457e+03   0.013  0.98992
## RestingECGNormal:ExerciseAnginaY   3.998e-01  6.897e-01   0.580  0.56210
## RestingECGST:ExerciseAnginaY       2.384e+00  1.210e+00   1.970  0.04879 *
```

```
## MaxHR:ST_SlopeFlat                 1.051e+00  1.266e+02   0.008  0.99338
## MaxHR:ST_SlopeUp                   1.097e+00  1.266e+02   0.009  0.99309
## ---
## 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:  373.28  on 698  degrees of freedom
## AIC: 469.28
##
## Number of Fisher Scoring iterations: 20
```

Random Forest

Our second model uses a random forest. With random forests, we fit a large number of binary decision trees, each on a bagged set of data and each with a random subset of predictors, then average the trees to get a final prediction. By using random subsets of predictors, we decorrelate the trees and reduce the variance compared to other methods like bagging. The random forest method yields itself well to the project goals at hand because of the categorical and binary nature of the outcomes we're trying to predict. We built a random forest of 650 trees built off of bootstrapped heart data. We specified number of predictors sampled for spitting at each node as sqrt(11) ~ 3. We chose sqrt(11) (square root of the number of predictors) since this is the industry-standard for classification. We chose 650 trees in order to lower our false negative rate, as we can see in our error plot.




```
##
## Call:
##  randomForest(formula = HeartDisease ~ ., data = heart, mtry = sqrt(11),      ntree = 650)
##                Type of random forest: classification
##                      Number of trees: 650
## No. of variables tried at each split: 3
```

```
##
##         OOB estimate of  error rate: 13.4%
## Confusion matrix:
##     0   1 class.error
## 0 335  55   0.1410256
## 1  45 311   0.1264045
```

The error rate from this model is 13.4%. At the end, the error rate of this model will be compared with the error rate of our other two models to decide if this is the model we will end up using to predict heart disease in at-risk patients.

SVM

Finally, our third model uses a Support Vector Machine (SVM). An SVM uses kernels to provide a flexible classification model. It separates the domain of the data with the goal maximizing the margin distance while allowing for a small number of misclassified training points. Although SVMs are not very interpretable and therefore wouldn't fit our inference goals, they tend to yield high prediction accuracy and work well with categorical response variables, so we fit an SVM for prediction purposes to compare with our other models. The three types of kernels for SVM that we considered were linear, polynomial, and radial. For each of the kernels, we considered a range of the penalty coefficient C. For polynomial kernel, the degree was set to 2. The best cost tuning parameters for each of our SVM models were determined using the 'tune' function provided by the 'e1071' package.

Then, using the chosen hyperparameters for each models, we ran 5-fold cross validation to determine the estimated test error for each of the models. In each fold, we trained three svm models using 80% of the data, each for the three types of kernels with the chosen parameters. Then, on the other 20% of the data, we computed the misclassification error rate at each fold. Then, we compared the mean of the misclassification rates to select the best model.

```
# Dividing the data into 5 folds
set.seed(1)
shuffled_heart <- heart[sample(nrow(heart)),]
folds <- cut(seq(1,nrow(shuffled_heart)),breaks=5,labels=FALSE)
```

```
## [1] 0.1367248
```

```
## [1] 0.1380492
```

```
## [1] 0.1380582
```

Linear SVM model performed best among the three types of kernels. The best hyperparameter chosen for linear SVM is $c = 1$. In our best SVM model, there are 249 support vectors.

```
## [1] 1
```

```
##
## Call:
## svm(formula = HeartDisease ~ ., data = heart, kernel = "linear",
##     cost = tune.out.linear.cost)
##
##
## Parameters:
```

```
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  252
##
##  ( 125 127 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

Comparision of the models through CV

On top of this, we also considered a combination of the models chosen from each algorithm: linear SVM, random forest, and logistic regression. The combination model predicts the value that appears most among the three models' prediction. This model was also considered to examine how effective the emsembled model of the algorithms would perform. To compare these models, cross-validation was five folds was used. At each fold, each of the models were trained with the hyperparameters that were selected from the above. The misclassification rate on each fold was calculated for each model, and the mean of the misclassification rates were computed in the end. Below is the code for the cross validation, and the results will be analyzed in the following section.

```
## [1] 0.1367248
```

```
## [1] 0.1354004
```

```
## [1] 0.1487875
```

```
## [1] 0.1501387
```

IV. Results

ADD COMMENTARY HERE ABOUT WHICH MODEL WAS THE BEST AND THE CORRESPONDING ERROR RATES

Note: The following analysis of the logistic regression and random forest results are based on fitting each model to the first of 5 chains of imputed results. ### Logistic Regression:

We would like to use the results of the logistic regression in order to interpret the relationships between different variables and the response variable, despite it not being the best model for prediction.

We will be concentrating on interpreting some of the significant terms. One of the categorical terms that is significant is Sex, with an associated coefficient of 6.850. This means that being male, compared to the baseline of female, leads to an expected 6.850 increase in the log odds of having heart disease. We can also say that being male, compared to the baseline of female, leads to the odds of having heart disease to multiply by a factor of e^(6.850). We can also see that having a RestingECG level of ST leads to a decrease in the odds of getting heart disease when compared to the baseline of LVH. All other main effects are not considered significant at alpha = 0.05.

There are a handful of significant interaction effects, most of which are interactions between 2 categorical variables. One example of this is between Sex and FastingBS. When FastingBS takes the baseline value of

0, a male compared to the baseline of female will result in an expected 6.850 increase in the log odds of having heart disease. When FastingBS takes the value 1, a male compared to the baseline of female will now result in an expected 6.850 - 4.101 = 2.749 increase in the log odds of having heart disease. One example of an interaction between a categorical and continuous variable is ChestPainType and Cholesterol. We should keep in mind that this interpretation may not be appropriate because the main effect for Cholesterol is not considered significant. For a baseline ChestPainType of ASY, an increase of 1 unit in Cholesterol will result in an expected decrease of 1.288 in the log odds of having heart disease. For a ChestPainType of ATA, a 1 unit increase in Cholesterol now results in an expected decrease of 1.288 - 0.023 = 1.265 in the log odds of having heart disease.

Random Forest:

Our confusion matrix tells us that there is a 0.141 false positive rate and a 0.126 false negative rate. This means that 14.1% of the time, our model predicts someone who doesn't have heart disease to have heart disease. At the same time, 12.6% of the time, our model predicts someone who has heart disease to not have heart disease. It is better in this scenario that our false positive rate is higher than the false negative rate because it is better that we over predict people having heart disease than to under predict. If we over predict, people who didn't have heart disease but thought they did would get a second opinion and eventually realize that they don't actually have the disease. If we under predicted, then people who have actually heart disease wouldn't get the necessary treatment and may have worsened affects. Despite the false negative rate being quite high, we decided to keep the default cutoff value of 0.5 because without having done a lot more in-depth research on this medical topic, it would be fairly arbitrary of us to choose another cut off point. Intuitively, it would make sense for us to choose a higher cutoff value because there is a heavier social penalty for a high false negative value as compared to a false positive value. One limitation of our random forest is that we can't tell from this model which factors are important in predicting heart disease. Another limitation is that we can't predict heart disease as accurately with this model compared to the SVM model.

SVM:

Compared to the logistic regression and random forest model, the SVM performs better for prediction, as we expected. Therefore, if we wanted a model that would assess whether a patient at risk for heart disease most likely has or does not have heart disease based on their personal and health factors, this is the model we would choose. However, this model does not tell us much about what factors are or aren't important in predicting heart disease, and it cannot tell us actually likelihoods of having heart disease or not.

V. Conclusion

As mentioned earlier, one of the biggest limitations of our findings is the dataset itself. Although the dataset is one of the largest available datasets on heart disease, the dataset's collection methodology limits the usefulness of our models. As mentioned in the Section II of this report, the data is only collected from at-risk patients who received CVD diagnostic check-ups at hospitals. Thus, the data may be weighted disproportionately towards people who would received a positive diagnosis for CVDs, as the sample population in the data were not just at-risk, but also either self-selected for or were chosen to be checked for CVDs. Furthermore, the data was collected from only five hospitals in the United States and Europe. As a result, the data is far from representative of the entire global population that is at-risk for CVDs, and our models and results may not be accurate for predicting the likelihood of CVDs for all at-risk populations in the world, and may provide limited understanding of signs that at-risk populations who are not represented by the data can analyze to check their risk for CVDs.