Project Report

Team name: Data Dudes
Richard Lim
Jake Li
Jianhong Li
Dev Patel

Heart Disease Data Analysis Report

We have decided to use data to analyze the risk of developing heart disease because we believe that the information derived from this will prove useful throughout our entire lifetime. It is crucial to understand high risk factors that heavily impact heart disease early on in one's life so that they can take precautionary measures to preserve their long term health. The CDC has claimed that the leading cause of death for a large majority of people of different races in the United States is heart disease. Almost half of the American population (47%) has 1 of 3 key heart disease risk factors. These factors which will be examined include: smoking, high blood pressure, high cholesterol, obesity, lack of physical exercise, diabetic status, overconsumption of alcohol, and many more. We are going to analyze a heart disease dataset that we found on Kaggle that originated from the CDC website. This data was largely collected by the Behavioral Risk Factor Surveillance System which gathers data on the health state of United States citizens by conducting annual phone surveys. This data originally had 279 different variables but by using the Kaggle dataset, we have narrowed it down to what we believe are the 18 most relevant factors. Our main objective is to determine which variables have the largest impact on heart disease probability. Other questions we are interested in answering include: "Which factors affect each gender the most?", "Which gender is the most susceptible to developing a heart disease?", "Which age group is the most likely to develop a heart disease?", "What is the likelihood of getting heart disease when exposed to more than one variable?", and "Are people who do not partake in unhealthy habits still at significant risk for heart disease?".

Prior to implementing predictive models, we determined that it was essential to understand the dataset through the exploratory data analysis process so that we can appropriately analyze key statistical metrics as well as better understand the contents for restructuring purposes. We first determined that the dimensions of the dataset consisted of 319,795 rows and 18 columns and that our variables were formatted in primarily the factor or num datatype due to a large quantity of categorical and quantitative data. In order to test data quality to determine what degree of cleaning was necessary, we tested for null and duplicate values. Although there were no null values in the dataset, there were in fact, 18,078 duplicate values which we removed to reduce redundancies and bias on model results. Upon generating summary statistics for each of our variables, we noticed a few factors that may impact the interpretation of our predictive model results. One such observation according to figure 1 was that our dataset was largely imbalanced as the quantity of individuals that didn't have any heart disease was heavily weighted towards the factor level "No" with 292,422 occurrences accounting for over 91% of all occurrences. Such a sampling of the data could significantly hamper our results and inflate model accuracy since our models would not have enough minority occurrences to best identify correlations between response and explanatory variables. Due to a large degree of imbalance within our data, most models will be likely to predict the majority class and also will portray an inaccurate measure of error as the minority class will contribute little toward it. Hence, in order to address this class problem, we implemented the Synthetic Minority Oversampling Technique (SMOTE) from the Performance Estimation Library to undersample our "No" observations to 272,610 and oversample our "Yes" observations to 299,871 for the Heart Disease variable. In order to successfully implement this procedure, we also decided to create dummy variables for each of the columns so that a new dataset can be replicated to facilitate desired re-ordering of columns and remaining of variable names without altering the original dataset for further analysis.

This, coupled with SMOTE function, created a more balanced distribution that allowed us to remove any potential sampling bias in the survey.

The response variable (Y) we want to predict is qualitative and we are interested in estimating the probabilities that Y belongs to each category. Therefore, we need to use classification data mining methods. In this project, we fitted our data to logistics regression, classification and regression tree (CART), and random forest. First, we split the dataframe into 80% for training for 20% for testing. Next, we fitted all the predictors into the logistics regression model and produced a summary of the model. Based on the coefficient from Figure 2, we found that most predictors are statistically significant except physical health, mental health, age 25 to 29, and skin cancer. Based on the coefficients, it seems like males in general have a higher probability of having heart disease. We also see that as age increases, the coefficient of the variable also increases significantly. In other words, the probability of having heart disease gets higher as people get older, which is intuitive. It was also interesting to see that drinking alcohol lowers the probability of having heart disease but only true for moderate intake. In sum, the variables that lead to a higher probability of heart disease are older age, stroke, and poor general health. Some variables that reduced the risk of heart disease are sleep time and alcohol in moderation. To evaluate model performance, we decided to use a confusion matrix to calculate the accuracy and recall score. The accuracy is how many predictions were correct in total. We are also using recall because it calculates how many times the actual result is "Yes" and we got it right. The recall is important in this case because if a person has heart disease, we don't want to misclassify it. Based on Table 1, we calculated the accuracy by adding true positive and true negative and then dividing it by the total number of observations. The logistic regression yielded an accuracy of 76%. We calculated the recall by dividing true positive by true positive plus false negative. The logistic regression yielded a recall of 81%. This is great because we want higher recall than accuracy. In addition, we performed 5-fold cross-validation on logistics regression and also got an accuracy of 76%. This indicates that the model is not overfitting. Next, we fitted our model to the classification and regression tree (CART) model. According to Figure 3, we pruned the tree to terminal nodes of 7 because that gives the lowest cross-validation error rate. By pruning the tree, we also prevent it from overfitting. According to the pruned tree, Figure 4, we can see the logic the model use to determine if someone has heart disease. Note that the values on Figure 4 is very close to 0 so it's safe to pretend that they are 0. For instance, if one has difficulty walking, then the model classify the person has heart disease. If not, the model will check the next condition until it reaches a classification. It seems like the CART model concludes that using difficulty walking, diabetic, and age are enough to determine if someone has heart disease or not. Based on Table 2, the CART model yielded an accuracy of 72%. It's slightly lower than the logistics regression model. However, it yielded a recall of 85% which is higher than the logistic regression model. Therefore, even though CART has a lower accuracy than logistic regression, we still consider it the better model due to having a higher recall. Last but not least, we fitted our data to a random forest model. Based on Table 3, the random forest model yielded an accuracy of 91% and a recall of 93%. Since random forest got the highest accuracy and recall, it will be our final model.

Random forest adds additional randomness to the model while growing trees. When splitting a node, it searches for the best feature among a random subset of features instead of looking for the most important feature. Thus, it reduces the overfitting problem in decision trees and lessens the variance, improving accuracy. As in Figure 5, it shows the specific values of importance in the model. We can also

visualize in Figure 6, which is a fundamental outcome of the random forest and it shows, for each variable, how important it is in classifying the data. The Mean Decrease Accuracy plot expresses how much accuracy the model losses by excluding each variable. The more the accuracy suffers, the more important the variable is for the successful classification. The variables are presented from descending importance. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable in the model. By looking at Figure 6, we can see that the most important variables in Mean Decrease Accuracy include: Sex_Male, Asthma_Yes, BMI（Body Mass Index), Stroke_Yes and Diabetic_Yes. The most important variables in Mean Decrease Gini are Sleep time, Physical Health, BMI (Body Mass Index), Difficulty to walk, Mental Health and Diabetic.

Our main objective for this project was to determine the most influential factors in predicting the likelihood of someone developing a heart disease. We wanted to remove any potential sampling bias in the survey so we used the SMOTE function and created dummy variables to ensure we had a balanced distribution. We decided to use random forest as our final model because it yielded the highest accuracy and recall percentages out of all the models we decided to test. We concluded that the most influential factors based off the random forest model were: if somebody was male, if they have asthma, if they have a high body mass index, if they've had a stroke, if they're diabetic, the amount of sleep they get, their physical and mental health, and if they have difficulty walking. When referencing the set of questions we sought to answer, we were able to determine that males were more susceptible to developing a heart disease and by using logistic regression we were able to determine that the older somebody gets, the higher chance they have of developing a heart disease. Further research would incorporate figuring out the correct models to answer our other unanswered questions which were: "Which factors affect each gender the most?","What is the likelihood of getting heart disease when exposed to more than one variable?", and "Are people who do not partake in unhealthy habits still at significant risk for heart disease?". The results from this project have given us a good understanding of which factors affect heart disease probability the most and we plan on using this insightful knowledge to encourage and preserve the good health of ourselves and loved ones.

Figure 1. Heart disease frequency plot

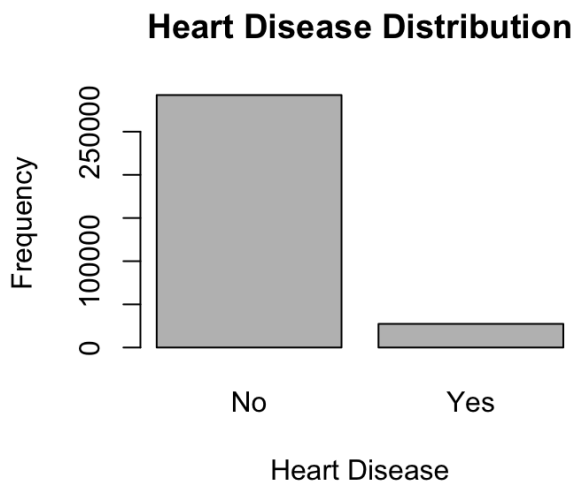

Figure 2. Summary of the logistic regression model

```
Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -4.1339077  0.0533413 -77.499  < 2e-16 ***
BMI                                0.0062284  0.0006617   9.413  < 2e-16 ***
PhysicalHealth                     0.0000987  0.0005389   0.183  0.85467
MentalHealth                      -0.0001505  0.0005185  -0.290  0.77164
SleepTime                         -0.0271844  0.0026370 -10.309  < 2e-16 ***
Smoking_Yes                        0.3726000  0.0075516  49.340  < 2e-16 ***
AlcoholDrinking_Yes               -0.5136372  0.0174204 -29.485  < 2e-16 ***
Stroke_Yes                         1.1066360  0.0164019  67.470  < 2e-16 ***
DiffWalking_Yes                    0.2638820  0.0106043  24.884  < 2e-16 ***
Sex_Male                           0.7386757  0.0076702  96.305  < 2e-16 ***
AgeCategory_TwentyFiveToTwentyNine 0.0604947  0.0464622   1.302  0.19291
AgeCategory_ThirtyToThirtyFour     0.4808004  0.0415566  11.570  < 2e-16 ***
AgeCategory_ThirtyFiveToThirtyNine 0.5433598  0.0400582  13.564  < 2e-16 ***
AgeCategory_FortyToFourtyFour      0.9719838  0.0378390  25.687  < 2e-16 ***
AgeCategory_FourtyFiveToFourtyNine 1.2760847  0.0366125  34.854  < 2e-16 ***
AgeCategory_FiftyToFiftyFour       1.7507043  0.0352312  49.692  < 2e-16 ***
AgeCategory_FiftyFiveToFiftyNine   2.0089835  0.0346127  58.042  < 2e-16 ***
AgeCategory_SixtyToSixtyFour       2.2800475  0.0342384  66.593  < 2e-16 ***
AgeCategory_SixtyFiveToSixtyNine   2.5447759  0.0341554  74.506  < 2e-16 ***
AgeCategory_SeventyToSeventyFour   2.8286749  0.0342487  82.592  < 2e-16 ***
AgeCategory_SeventyFiveToSeventyNine 3.0198762  0.0348031  86.770  < 2e-16 ***
AgeCategory_EightyOrOlder          3.3218114  0.0347329  95.639  < 2e-16 ***
Race_Asian                        -0.4839300  0.0443288 -10.917  < 2e-16 ***
Race_Black                        -0.0905040  0.0332848  -2.719  0.00655 **
Race_Hispanic                      0.0548129  0.0335846   1.632  0.10266
Race_Other                         0.1033148  0.0367350   2.812  0.00492 **
Race_White                         0.1867245  0.0304941   6.123 9.17e-10 ***
Diabetic_No_borderline_diabetes   -0.2156022  0.0244750  -8.809  < 2e-16 ***
Diabetic_Yes                       0.4493035  0.0097093  46.275  < 2e-16 ***
Diabetic_Yes_during_pregnancy     -0.4593714  0.0595375  -7.716 1.20e-14 ***
PhysicalActivity_Yes               0.0858774  0.0088031   9.755  < 2e-16 ***
GenHealth_Fair                     1.6734999  0.0160236 104.440  < 2e-16 ***
GenHealth_Good                     1.0899112  0.0132370  82.338  < 2e-16 ***
GenHealth_Poor                     2.1262961  0.0233057  91.235  < 2e-16 ***
GenHealth_Very_good                0.5114605  0.0132693  38.545  < 2e-16 ***
Asthma_Yes                         0.2094534  0.0107615  19.463  < 2e-16 ***
KidneyDisease_Yes                  0.4205244  0.0164175  25.614  < 2e-16 ***
SkinCancer_Yes                     0.0115401  0.0111754   1.033  0.30178
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3. Error rate graph for the CART model.

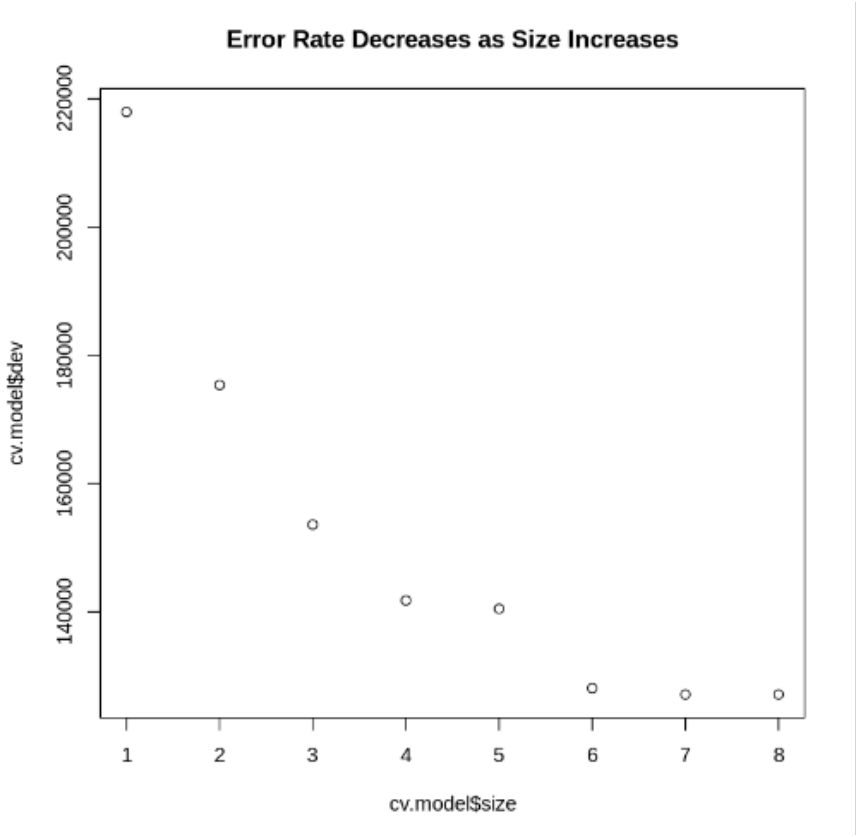

Figure 4. Pruned tree of the classification tree.

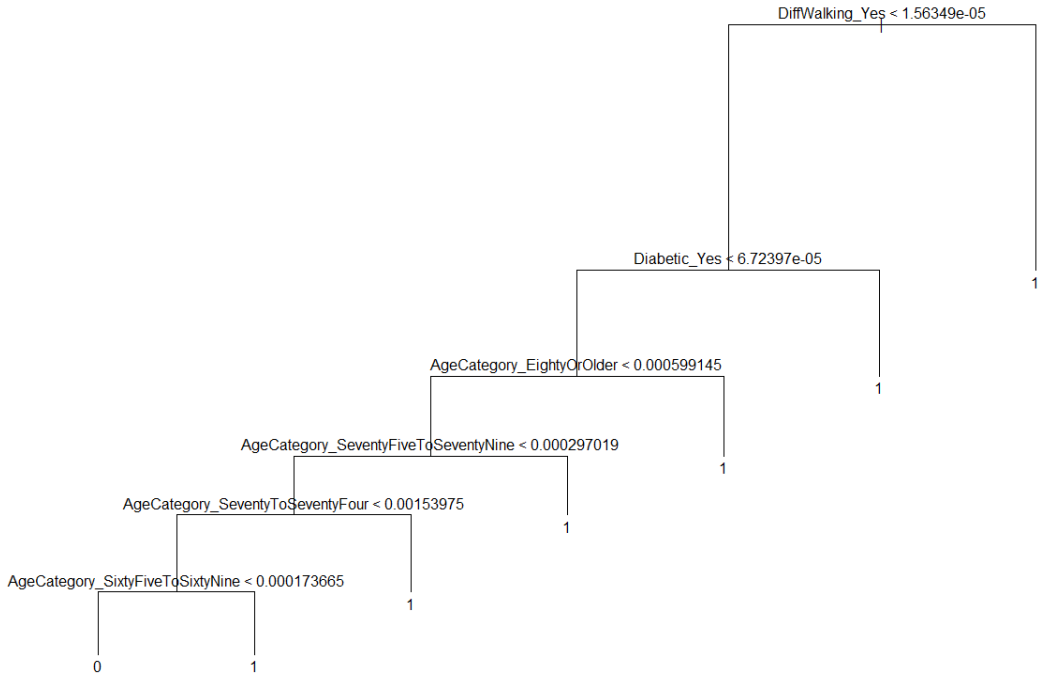

Figure 5. Random Forest - Importance

```
> importance(rf.balanced_df)
                                            0        1 MeanDecreaseAccuracy MeanDecreaseGini
BMI                                152.276940 221.96794            229.87607       13068.3561
PhysicalHealth                      59.270355 124.86836            103.08076       15498.2403
MentalHealth                        63.498272 178.25446            141.95529        9606.8576
SleepTime                           79.101504 154.36012            120.73982       22284.2322
Smoking_Yes                         89.307089 124.04573            118.78547        5551.0837
AlcoholDrinking_Yes                 78.034864 137.24499            133.03490        1866.8572
Stroke_Yes                         147.170942 185.81179            176.99632        6587.6272
DiffWalking_Yes                     74.520194 115.26623            107.37285        9821.4860
Sex_Male                           132.206003 266.26403            238.10140        6876.5864
AgeCategory_TwentyFiveToTwentyNine   6.132190  45.51462             50.81753        1886.1330
AgeCategory_ThirtyToThirtyFour      10.032923  52.79243             58.79458        1728.3092
AgeCategory_ThirtyFiveToThirtyNine  14.655657  64.12823             71.13078        2103.1118
AgeCategory_FortyToFourtyFour       26.445611  72.36854             78.88603        1498.6873
AgeCategory_FourtyFiveToFourtyNine  31.735262  74.28445             79.11857        1241.4853
AgeCategory_FiftyToFiftyFour        27.138704  92.42307             95.89506        1285.2161
AgeCategory_FiftyFiveToFiftyNine     6.475936 123.30368            136.72057        1443.0996
AgeCategory_SixtyToSixtyFour       -15.691595 109.52695            120.50441        2114.6655
AgeCategory_SixtyFiveToSixtyNine    -8.975884 117.09145            112.87371        2899.4068
AgeCategory_SeventyToSeventyFour    36.043299 124.25849            119.72290        4262.5657
AgeCategory_SeventyFiveToSeventyNine 55.027748 128.30767           127.73852        4402.3151
AgeCategory_EightyOrOlder           69.804705 161.77004            149.52205        7305.1193
Race_Asian                          29.277513  75.23017             86.31988         393.6329
Race_Black                          33.467774 100.04122            113.04489        1042.3504
Race_Hispanic                       24.406082  86.93321            100.19319        1012.4056
Race_Other                          27.895078 123.07172            126.73949         673.5958
Race_White                          46.544397 154.60288            134.51316        3184.6004
Diabetic_No_borderline_diabetes     85.089102 136.62565            139.86019         945.2576
Diabetic_Yes                       116.357350 172.36836            160.88511        9250.0756
Diabetic_Yes_during_pregnancy       31.983796  66.31576             70.01417         230.3582
PhysicalActivity_Yes                39.787005  77.00954             68.57026        3142.7815
GenHealth_Fair                      69.592299  64.70080             75.59413        5101.3714
GenHealth_Good                       8.443616 113.07996            121.04252        3994.4406
GenHealth_Poor                      38.243040  59.87681             59.76493        3166.5656
GenHealth_Very_good                -32.519441  65.12507             63.49990        2969.1536
Asthma_Yes                          95.285573 234.82912            230.50269        2480.0002
KidneyDisease_Yes                   42.635382  98.47311             72.66360        2368.1547
SkinCancer_Yes                      52.686709  95.83675             80.02032        2395.2760
> varImpPlot(rf.balanced_df)
```

Figure 6. Random Forest - Plot

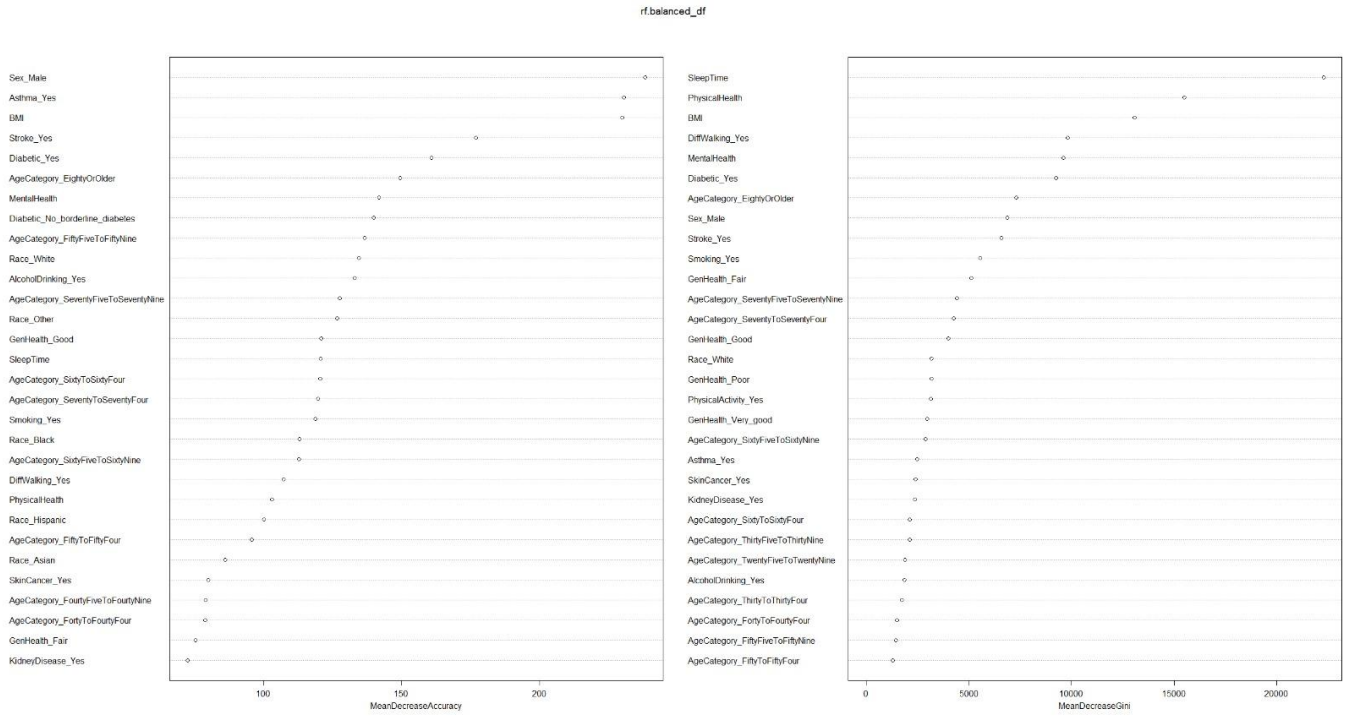

Table 1. Confusion matrix for logistic regression

glm.pred	test.truevalue 0	1
0	39153	11559
1	15500	48285

Table 2. Confusion matrix for CART

prunetree.pred	test.truevalue 0	1
0	31777	9188
1	22324	50872

Table 3. Confusion matrix for random forest

yhat.rf	test.truevalue 0	1
0	48335	4113
1	6318	55731

References

https://www.cdc.gov/heartdisease/risk_factors.htm

https://rikunert.com/smote_explained

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/#h2_6