Stroke Analysis Report

Christine Do, Alyssa Guillory, Johnny Le, Giselle Ruiz, Christopher Turcios, Jimmy Vuong

2022-11-21

Introduction

Response Variable:

- stroke:

Possible Predictors:

- smoking_status: The smoking status of the observation. Factor variable with 3 levels: "formerly smoked", "never smoked", and "smokes".

Cleaning the data

To use this data set, we first converted variables to their correct representations and then omitted any incomplete observations.

Notable fixes:

- The categorical variable smoking_status was reformatted to remove the level "Unknown", as it was used to represent unavailable data. Any cells that previously had "Unknown" were updated to reflect their unavailable status.
- The quantitative variable bmi was fixed by converting the data type from character to numeric. This change does not apply to non-numeric cells, so the cells containing "N/A" were changed to reflect their unavailable status in the now numeric column.

—**still need to list the rest of the variables, short description of data, and the question we want to answer**

Logistic Regression Model

We chose to use a logistic regression model for our data because our response variable is qualitative with two classes. It was desirable to have a model predict the probability of a person having a stroke using binary classification. If linear regression is used to predict the probability of whether or not a person will have a stroke, then the model may have predicted Y values outside of our intended range of 0-1. The logistic regression model allows us to enforce this restriction of range.

Model Formula

Our basic logistic regression formula with all our predictors would be as follows:

$$P(stroke = 1) = \frac{exp^{(b0 + b1*id + b2*gender + b3*age + b4*hypertension + b5*heart\_disease + b6*ever\_married + b7*work\_type + b8*Residence\_type + b9*avg\_glucose\_level + b10*bmi + b11*smokin + b12*g\_status)}}{1 + exp^{(b0 + b1*id + b2*gender + b3*age + b4*hypertension + b5*heart\_disease + b6*ever\_married + b7*work\_type + b8*Residence\_type + b9*avg\_glucose\_level + b10*bmi + b11*smokin + b12*g\_status)}}$$

This is what our logistic regression formula would look like if we considered all of our predictors and the response variable being stroke. However, we automatically did not consider the unique identifier (id) as this variable is only used to identify patients, which does not have a significant influence on predicting the probability of strokes. Using the glm() function, we created our initial logistic model with stroke being the response and all the other variables as predictors (except id).

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = health)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2163  -0.3357  -0.1914  -0.1051   3.1357
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.911e+01  4.775e+02  -0.040 0.968076
## genderFemale                 7.625e-02  1.681e-01   0.454 0.650091
## genderOther                 -1.234e+01  3.956e+03  -0.003 0.997511
## age                          7.304e-02  6.958e-03  10.497  < 2e-16 ***
## hypertension1                5.694e-01  1.828e-01   3.114 0.001845 **
## heart_disease1               3.906e-01  2.211e-01   1.767 0.077249 .
## ever_marriedYes             -1.795e-01  2.627e-01  -0.683 0.494366
## work_typeGovt_job            1.092e+01  4.775e+02   0.023 0.981747
## work_typeNever_worked       -2.958e-01  1.157e+03   0.000 0.999796
## work_typePrivate             1.117e+01  4.775e+02   0.023 0.981339
## work_typeSelf-employed       1.081e+01  4.775e+02   0.023 0.981936
## Residence_typeUrban         -4.830e-03  1.624e-01  -0.030 0.976276
## avg_glucose_level            4.696e-03  1.377e-03   3.409 0.000651 ***
## bmi                          6.501e-03  1.292e-02   0.503 0.614789
## smoking_statusnever smoked -7.476e-02  1.894e-01  -0.395 0.692999
## smoking_statussmokes         3.151e-01  2.305e-01   1.367 0.171639
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1411.0  on 3425  degrees of freedom
## Residual deviance: 1141.8  on 3410  degrees of freedom
## AIC: 1173.8
```

```
##
## Number of Fisher Scoring iterations: 16
```

As the summary shows, the predictors age, hypertension, and avg_glucose_level show high levels of significance, and the predictor heart_disease show moderate levels of significance. Since our p-values for all four predictors are less than 0.1, we can confidently reject the null hypothesis that states H_0: $B_0 = B_1 = \ldots = B_n$. To confirm that these variables are significant, we used the backwards step() function to see if it would give us the same significant variables, and it did.

```
#Model created from backwards step()
summary(health2.glm)
```

```
##
## Call:
## glm(formula = stroke ~ age + hypertension + heart_disease + avg_glucose_level,
##     family = "binomial", data = health)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1420  -0.3357  -0.1927  -0.1072   3.1976
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.632810   0.439480 -17.368  < 2e-16 ***
## age                0.067773   0.006359  10.659  < 2e-16 ***
## hypertension1      0.568379   0.181386   3.134 0.001727 **
## heart_disease1     0.453704   0.216660   2.094 0.036253 *
## avg_glucose_level  0.004701   0.001334   3.524 0.000426 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1411.0  on 3425  degrees of freedom
## Residual deviance: 1149.9  on 3421  degrees of freedom
## AIC: 1159.9
##
## Number of Fisher Scoring iterations: 7
```

```
#Model created without heart_disease
summary(health3.glm)
```

```
##
## Call:
## glm(formula = stroke ~ age + hypertension + avg_glucose_level,
##     family = "binomial", data = health)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0523  -0.3398  -0.1939  -0.1058   3.2054
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.750124   0.437181 -17.728  < 2e-16 ***
## age                0.070099   0.006262  11.194  < 2e-16 ***
## hypertension1      0.579247   0.180838   3.203 0.001359 **
```

```
## avg_glucose_level  0.004932   0.001326   3.720 0.000199 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1411.0  on 3425  degrees of freedom
## Residual deviance: 1154.1  on 3422  degrees of freedom
## AIC: 1162.1
##
## Number of Fisher Scoring iterations: 7
```

By removing the predictors that had little to no significance in predicting the response variable, we are able to refit our data into a second model (health2.glm) with the four predictors that are significant. The original model with all predictors had an AIC of 1173.8, while the second model with the four significant predictors had an AIC of 1159.9. As there was not much of a difference between the first and second model, we decided to remove the lowest significant predictor (heart_disease) out of the four to perform one last model refitting of our data. This time, our AIC for the third model (health3.glm) came out to be 1162.1, which was a slight increase compared to the second model.

$$P(stroke = 1) = \frac{exp(b_0 + b_3 * age + b_4 * hypertension + b_9 * avg\_glucose\_level)}{1 + exp(b_0 + b_3 * age + b_4 * hypertension + b_9 * avg\_glucose\_level)}$$

Additional information was obtained by using the Goodness of Fit statistical hypothesis test. Given the three models we tested above, we are able to determine each models' null deviance, residual deviance, R^2 value, AIC value, and BIC value. The lower the residual deviance, AIC, and BIC values, the better the model is able to predict the value of the response variable. As for R^2, a value closest to 1 is best.

```
##             Null Deviance Residual Deviance       R^2      AIC      BIC
## health.glm       1411.002          1141.806 0.1907835 1173.806 1272.032
## health2.glm      1411.002          1149.914 0.1850374 1159.914 1190.609
## health3.glm      1411.002          1154.069 0.1820928 1162.069 1186.625
```

Between all three models, better.health and best.health provided the closest fit for our data, with better.health having a slight advantage in lower residual deviance and AIC values. The test results show that each model's R^2 value is extremely low, with values being between 0.1 and 0.2, informing us that only 10%-20% of the variance can be explained by our models. This means that using logistic regression to fit our data may not result in the best fitting model. Even so, we can conclude that the model better.health (with predictors age, hypertension1, heart_disease1, and avg_glucose_level) had the best fit compared to the other models and is somewhat accurate in predicting whether or not a stroke will occur.

```
#Mean of the train error rate
(train.error.mean = mean(train.error))
```

```
## [1] 0.05275447
```

```
#Mean of the test error
(test.error.mean = mean(test.error))
```

```
## [1] 0.05080292
```

```
#Test Sensitivity
test.sensitivity
```

```
##  [1] 0.00000000 0.00000000 0.03225806 0.00000000 0.00000000 0.00000000
##  [7] 0.00000000 0.00000000 0.00000000 0.00000000
```

```
(test.sensitivity.mean = mean(test.sensitivity))
```

```
## [1] 0.003225806
```

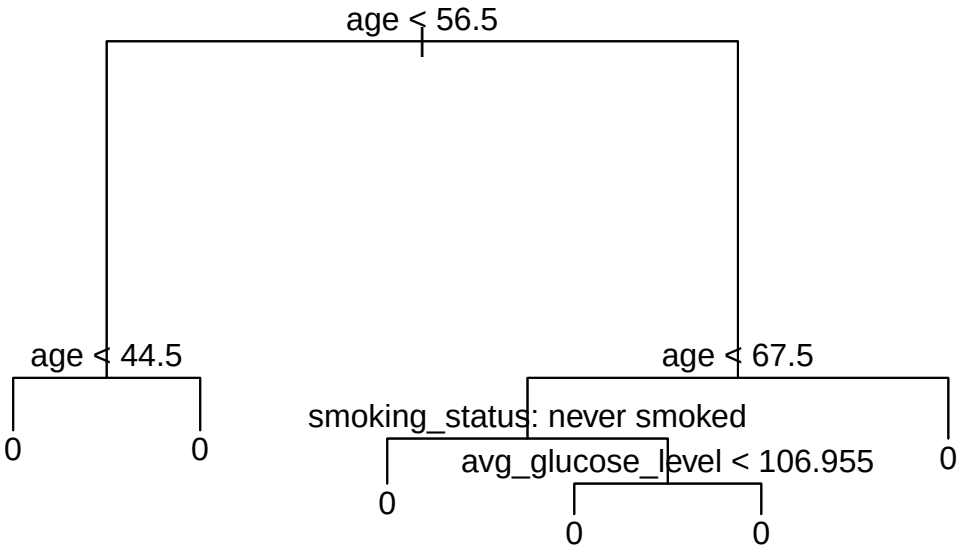
```
#Test Specificity
(test.specificity.mean = mean(test.specificity))
```

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

Decision Tree Model

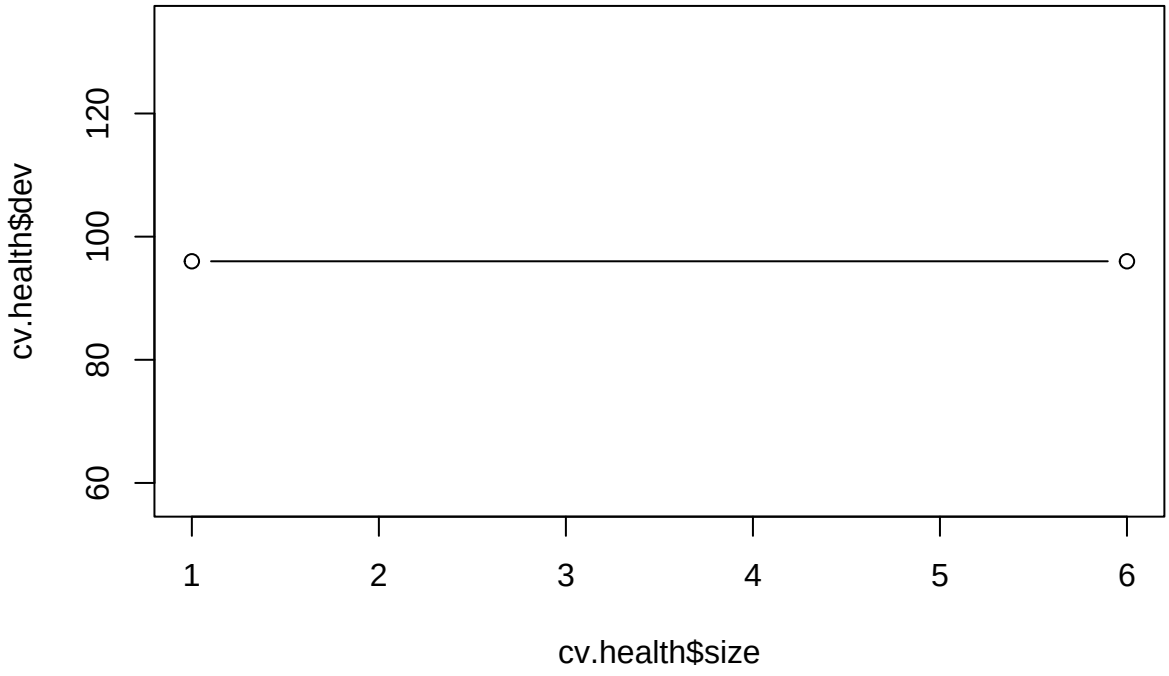
Base Decision Tree

```
## Test Error:  0.05107998
## Test Sensitivity:  0.001204819
## Test Specificity:  0.9999387
```



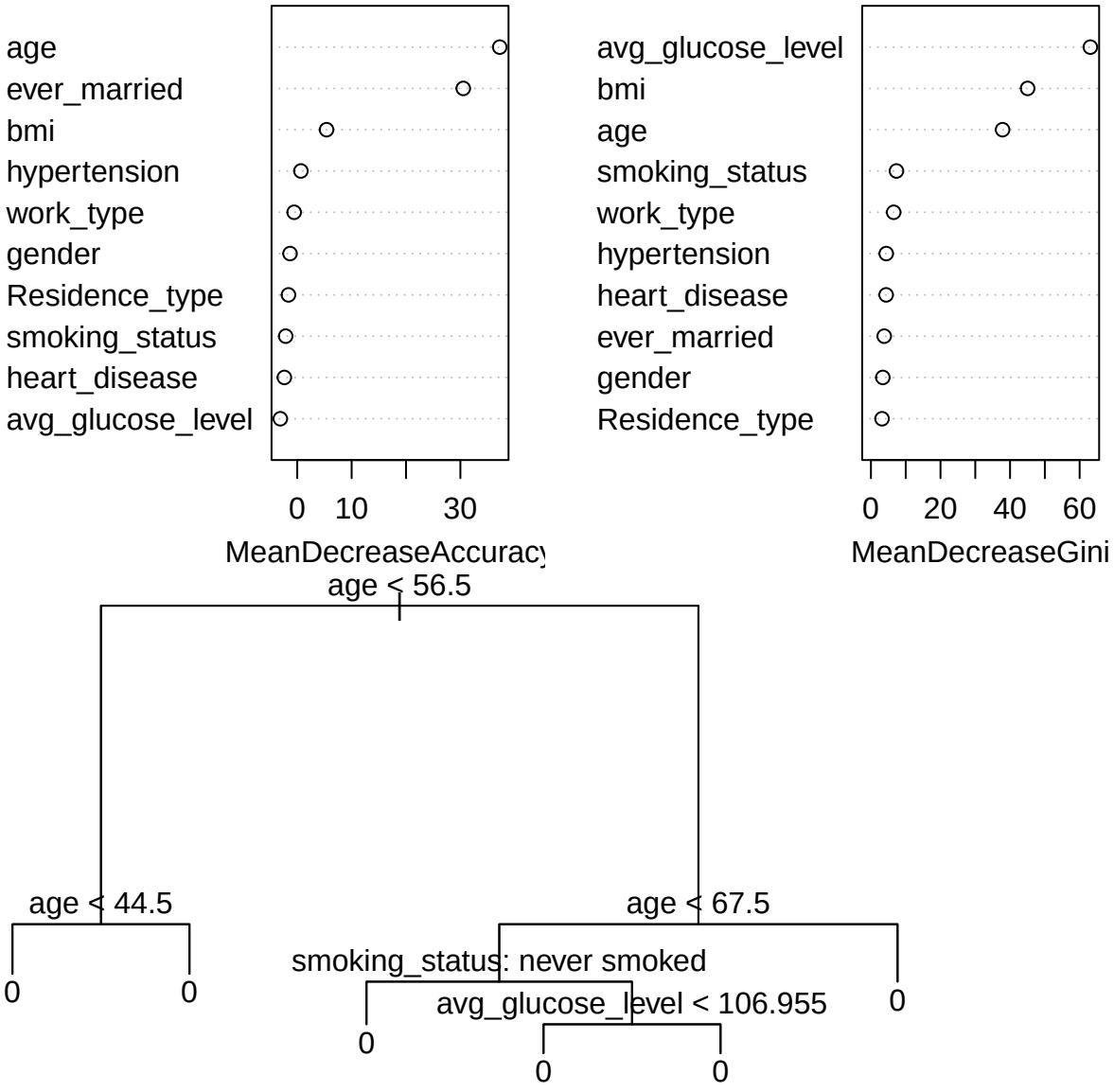
Pruned Decision Tree

```
## Test Error:  0.05107998
## Test Sensitivity:  0
## Test Specificity:  1
```



Bagged Decision Tree

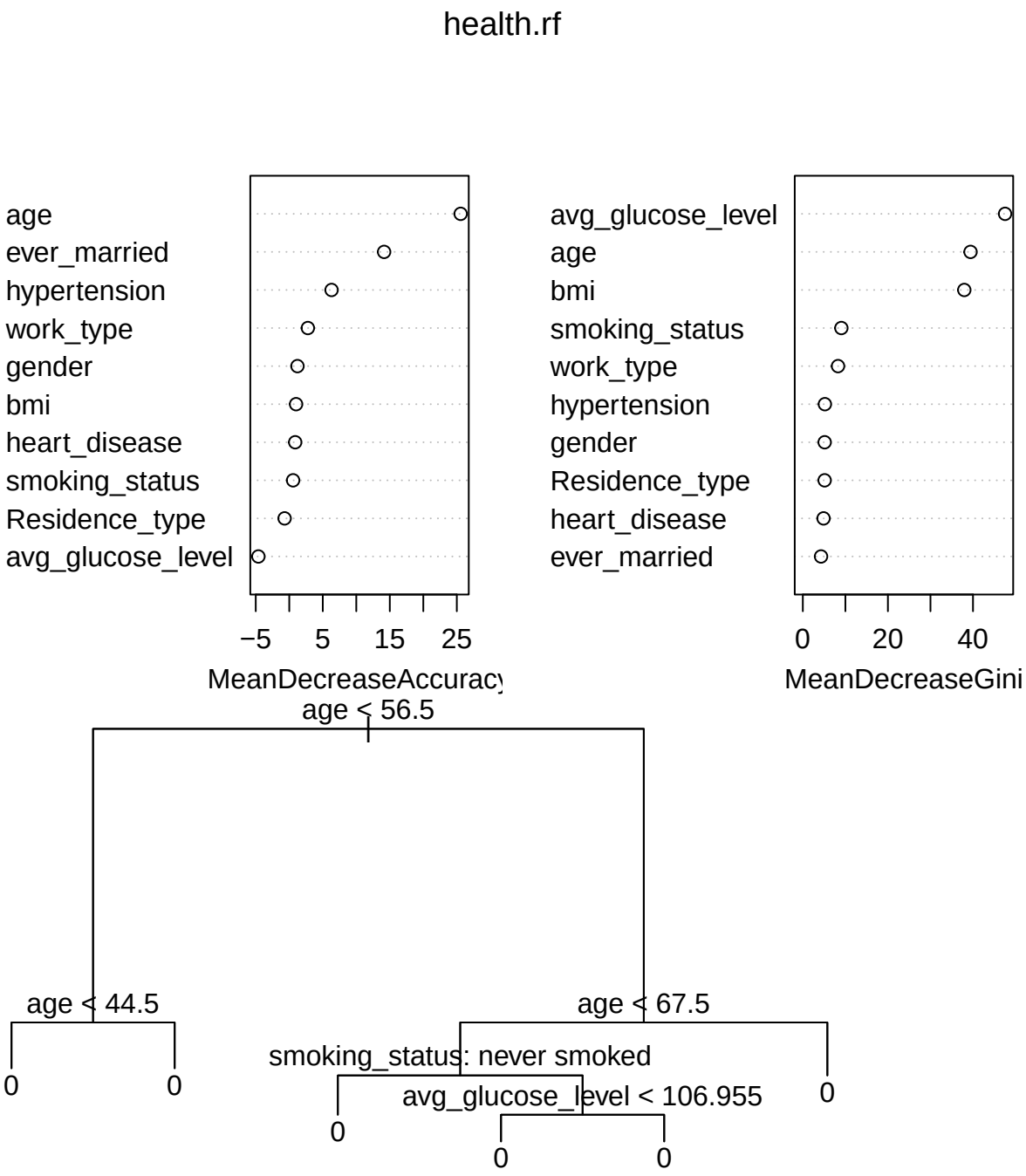
```
## Test Error:  0.05452423
## Test Sensitivity:  0.02140718
## Test Specificity:  0.9952017
```



Random Forest

```
## Test Error:  0.05207239
## Test Sensitivity:  0.007972732
```

```
## Test Specificity:  0.9985244
```



Conclusion