Stroke Analysis Report

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

2022-12-2

Introduction

With heart disease being the leading cause of death in the United States, we think it's important to examine the factors that may lead to strokes. Examining the factors for strokes helps predict and create an image of strokes and heart disease. Our data is sourced from Kaggle with a confidential source and is designated for educational purposes only.

Dataset Information

Data Variables:

- smoking_status: The smoking status of the observation. [Categorical factor w/3 levels]
- id: The unique identifier for the given observation [numeric]
- gender: The gender of the observation [Categorical Factor w/3 levels]
- age: Age of the observation [numeric]
- hypertension: Whether or not the observation has hypertension [Binary dummy variable]
- heart_disease: Whether or not the observation has heart disease [Binary dummy variable]
- ever_married: Whether or not the observation is married [Binary dummy variable]
- work_type: The type of work the observation participates in [Categorical Factor w/5 levels]
- Residence_type: The location in which the observation is located [Categorical factor w/2 levels]
- avg_glucose_level: The average glucose level [numeric]
- bmi: The body mass index [numeric]
- smoking_status: The smoking status of the observation. [Categorical factor w/3 levels]
- stroke: If the observation has had a stroke [Binary dummy variable]

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.

Main Question

Which factors demonstrate statistical significance in relation to the response variable stroke.

Logistic Regression (Christine Do, Alyssa Guillory, Jimmy Vuong)

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 had been used to predict the probability of whether or not a person will have a stroke instead, 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:

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

This is the logistic regression formula when we consider all of our variables as predictors for the response variable stroke. However, we automatically excluded id as a predictor, as it is a unique number arbitrarily used to identify patients, therefore we know it does not influence our response variable. Using the glm() function, we created our initial logistic model with stroke being the response and all the other variables as predictors.

```
##
## 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 shows a moderate level of significance. Since our p-values for all four of these predictors are less than α=0.1, we believe that these are the most important variables in predicting the probability of having a stoke; 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 on 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.

$$P(stroke = 1) =$$

$$\frac{exp(-7.632810 + 0.067773 * age + 0.568379 * hypertension + 0.453704 * heart\_disease + 0.004701 * avg\_glucose\_level)}{1 + exp(-7.632810 + 0.067773 * age + 0.568379 * hypertension + 0.453704 * heart\_disease + 0.004701 * avg\_glucose\_level)}$$

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
```

```
## Train Error Rate Mean:
train.error.m
```

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

```
## Test Error Rate Mean:
test.error.m
```

```
## [1] 0.05080292
## Test Sensitivity Mean:
test.sensitivity.m
```

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

```
## Test Specificity Mean:
test.specificity.m
```

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

Between the three models, health2.glm and health3.glm provided the closest fit for our data, with health2.glm having a slight advantage with lower residual deviance and AIC values. Each model's R^2 value is extremely low, with values being between 0.1 and 0.2, indicating that these models are not a good fit for our data. Even so, we conclude that the best linear regression model for fitting the data and predicting whether or not a stroke will occur is the model health2.glm (predictors: age, hypertension, heart_disease, avg_glucose_level).

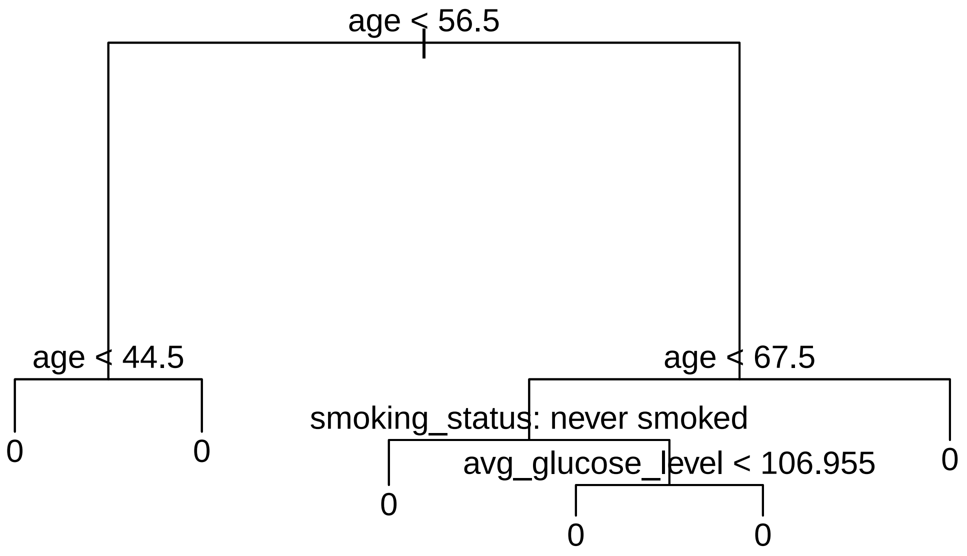
Decision Tree (Johnny Le, Giselle Ruiz, Christopher Turcios)

Base Decision Tree

```
## Test Error:  0.05107998
```

```
## Test Sensitivity:  0.001204819
```

```
## Test Specificity:  0.9999387
```

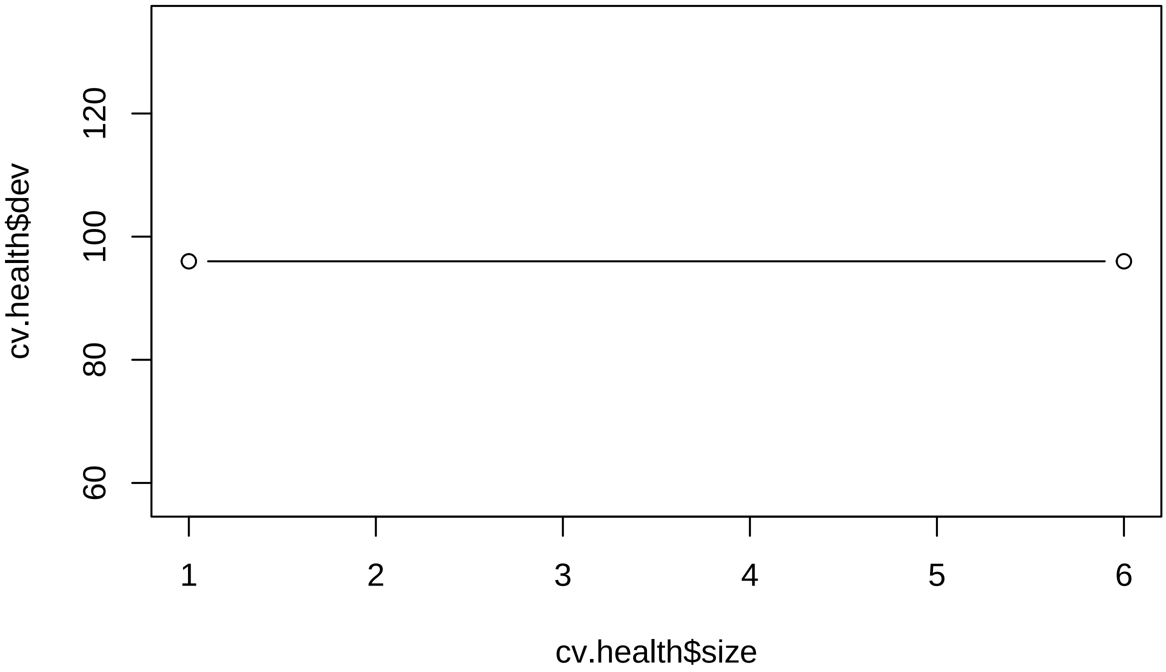


Pruned Decision Tree

```
## Test Error:  0.05107998
```

```
## Test Sensitivity:  0
```

```
## Test Specificity:  1
```

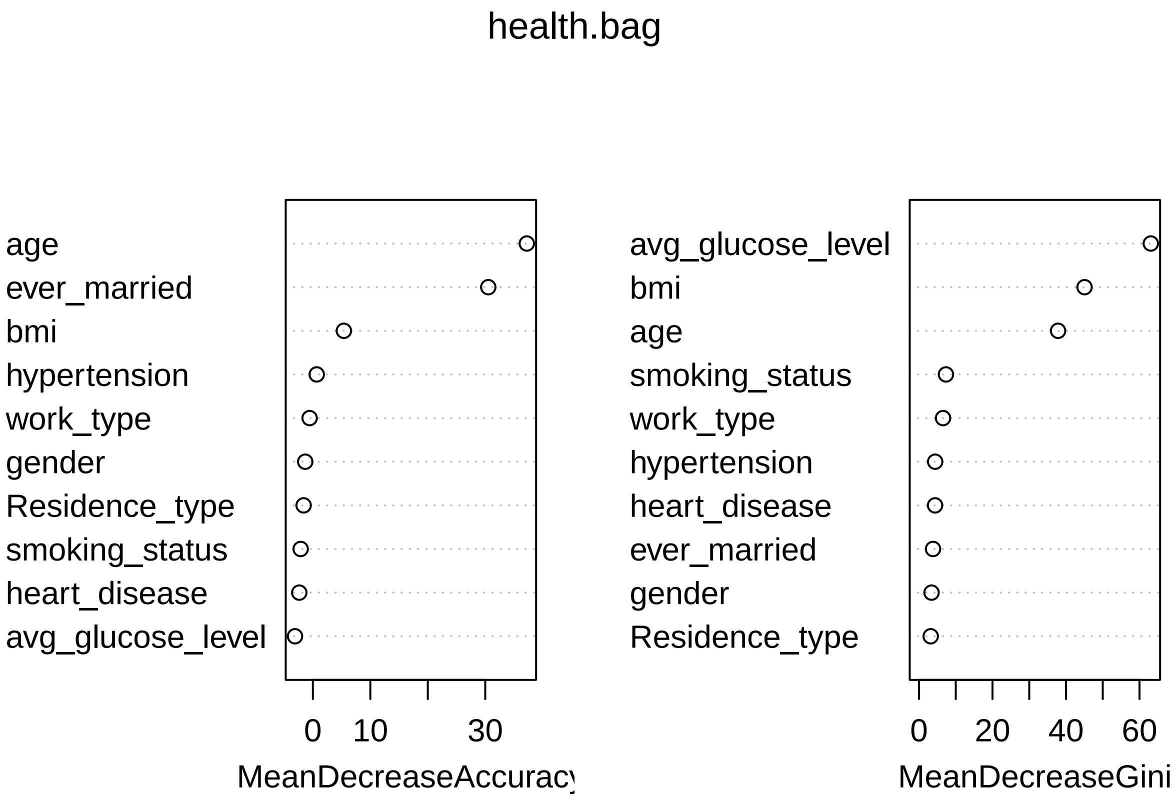


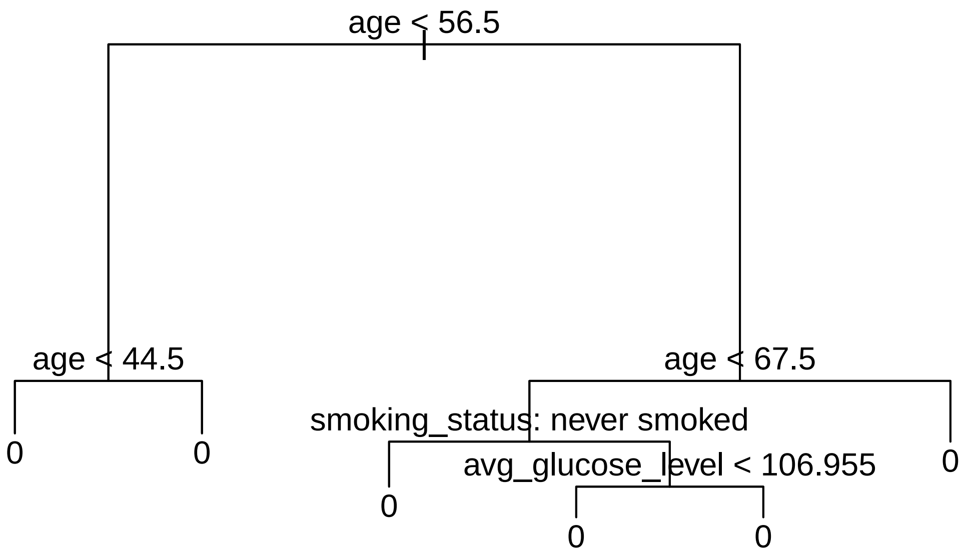
Bagged Decision Tree

```
## Test Error:  0.05452423
```

```
## Test Sensitivity:  0.02140718
```

```
## Test Specificity:  0.9952017
```


health.bag
age
ever_married
bmi
hypertension
work_type
gender
Residence_type
smoking_status
heart_disease
avg_glucose_level
0
10
30
MeanDecreaseAccurac
avg_glucose_level
bmi
age
smoking_status
work_type
hypertension
heart_disease
ever_married
gender
Residence_type
0
20
40
60
MeanDecreaseGini

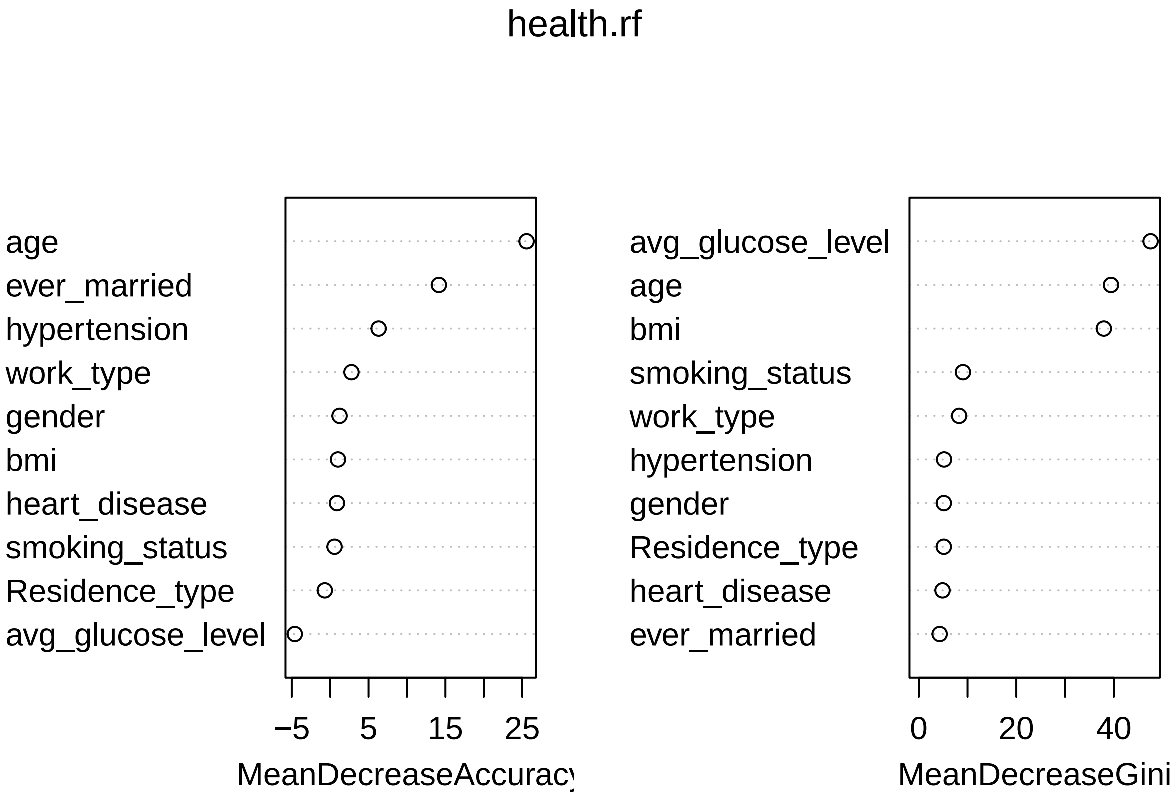


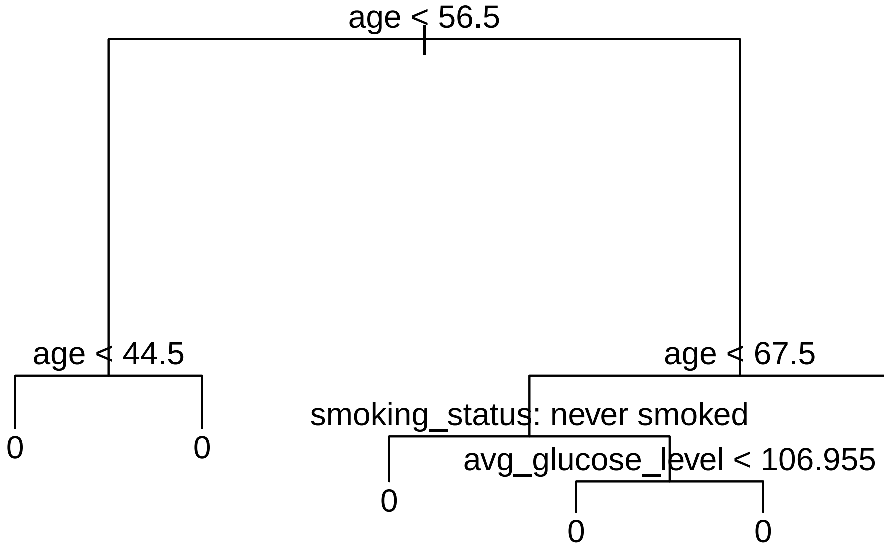
Random Forest

```
## Test Error:  0.05207239
```

```
## Test Sensitivity:  0.007972732
```

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


health.rf
age
ever_married
hypertension
work_type
gender
bmi
heart_disease
smoking_status
Residence_type
avg_glucose_level
−5
5
15
25
MeanDecreaseAccuracy
avg_glucose_level
age
bmi
smoking_status
work_type
hypertension
gender
Residence_type
heart_disease
ever_married
0
20
40
MeanDecreaseGini


age < 56.5
age < 44.5
0
0
age < 67.5
smoking_status: never smoked
0
avg_glucose_level < 106.955
0
0
0

Conclusion