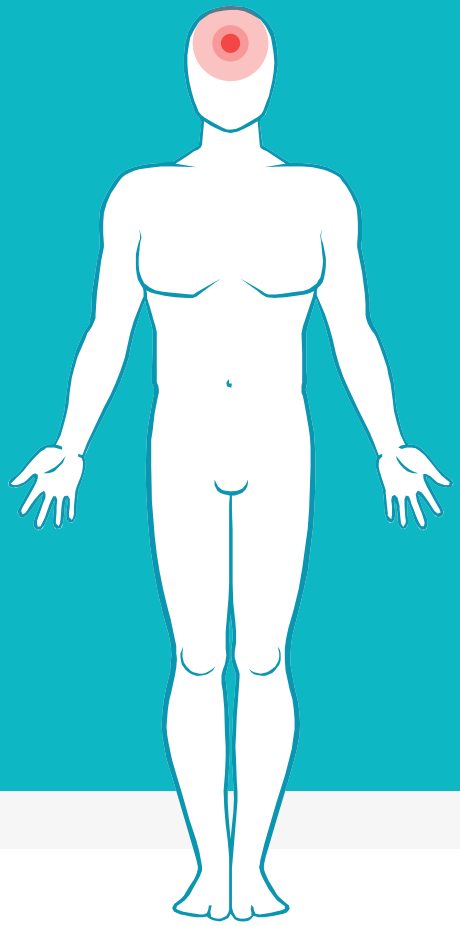
# Intro to DS Project: **Stroke Prediction**



Stefan Roata, Zheng Zhiwei (Mike)

# 1.
# Introduction 

# Project Introduction

**Context:**

According to the World Health Organization, 15 million people suffer stroke worldwide each year. Of these, 5 million die and another 5 million are permanently disabled. High blood pressure contributes to more than 12.7 million strokes worldwide.

Given the percentage of stroke skyrocketed with a significant 11% each year, our team decided to further investigate the stroke population and establish a model for the purpose of prediction and prevention.

**Data Set Selection:**

The dataset (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset) we have chosen is found on Kaggle and contains information about age, gender, BMI (and 7 other parameters) along with whether or not the patient has had a stroke. Each row in the dataset provides relevant information about each of the 5110 patients.


**Source:** http://www.strokecenter.org/patients/about-stroke/stroke-statistics/

The location where these data were collected, as well as the period of collection, remain undisclosed. The dataset is under a license called “Data files © Original Authors”, meaning that the dataset does not belong to the original uploader. The original source, as stated on Kaggle, is confidential and the use of this dataset is restricted to educational purposes only.

# Data Structure Overview: Attribute Information

- **id**: unique identifier
- **gender**: "Male", "Female" or "Other"
- **age**: age of the patient
- **hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart_disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- **ever_married**: "No" or "Yes"
- **work_type**: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- **Residence_type**: "Rural" or "Urban"
- **avg_glucose_level:** average glucose level in blood
- **bmi**: body mass index
- **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"*
- **stroke:** 1 if the patient had a stroke or 0 if not

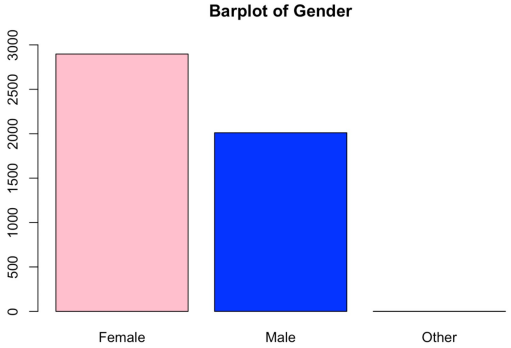
## Research Question:

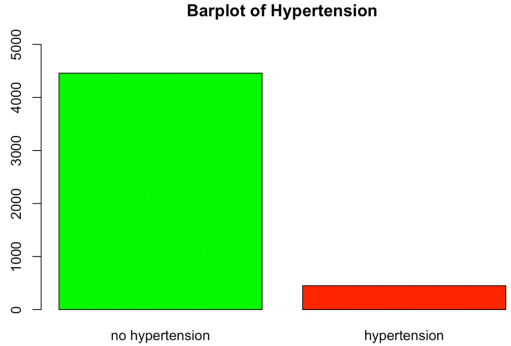
What are the variables that best allow us to predict stroke for an individual? Is it possible to predict stroke using these variables?

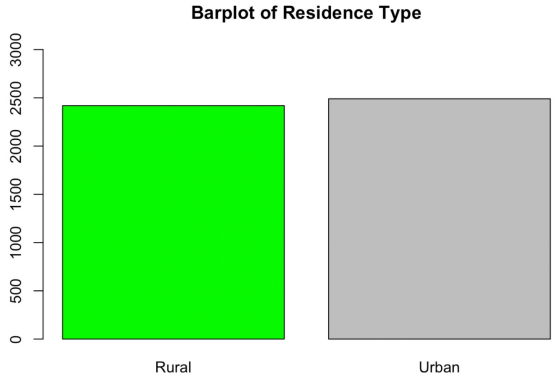
”

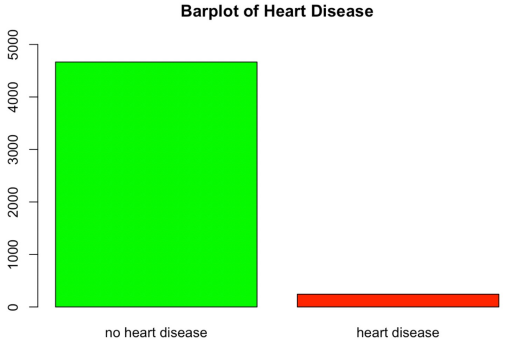
# 2.
# EDA Analysis 

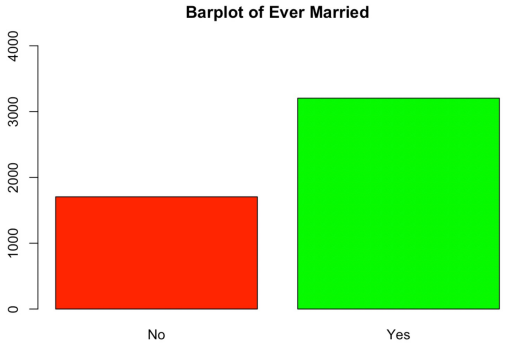
# The dataset contains mostly binary variables. Is the dataset balanced with respect to these binary variables?

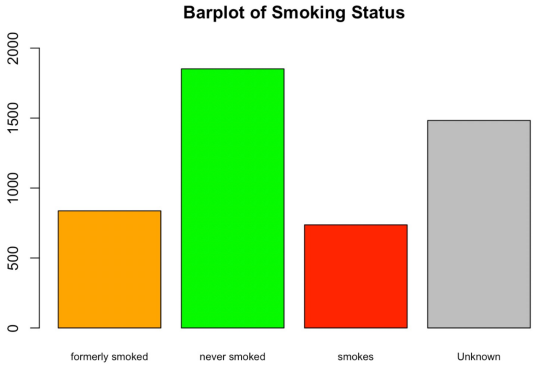


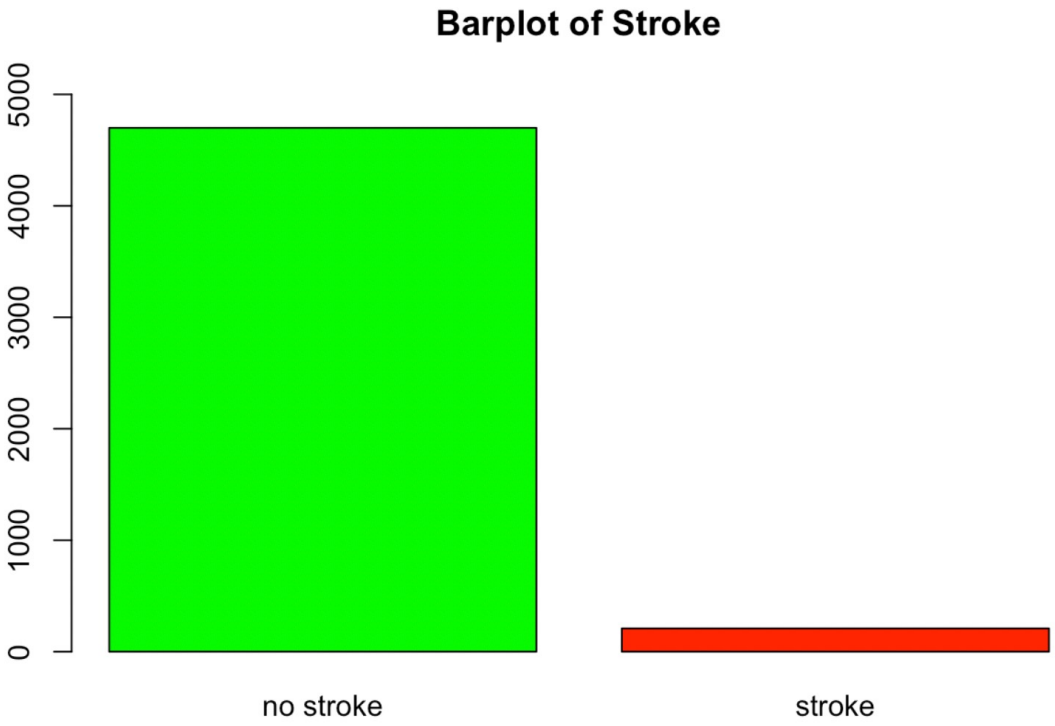



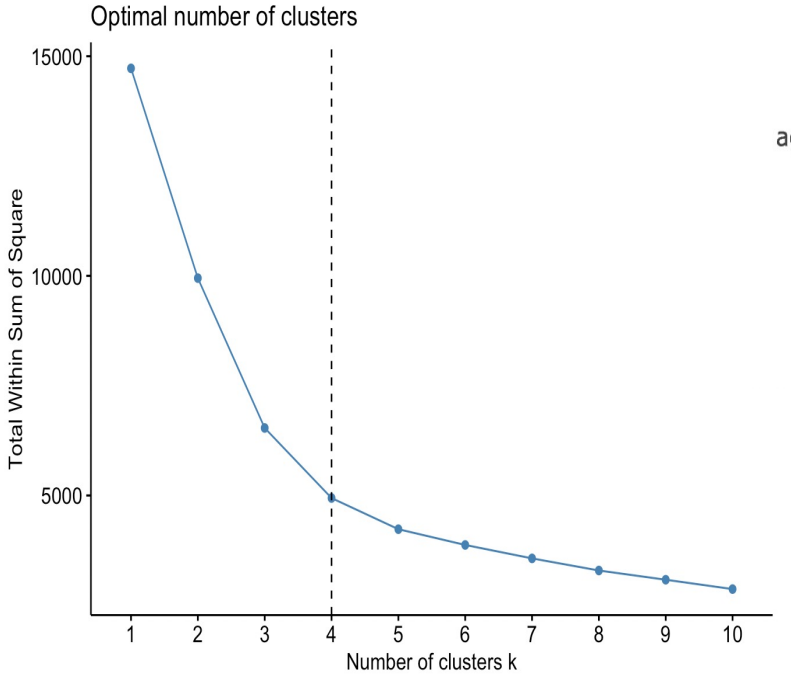



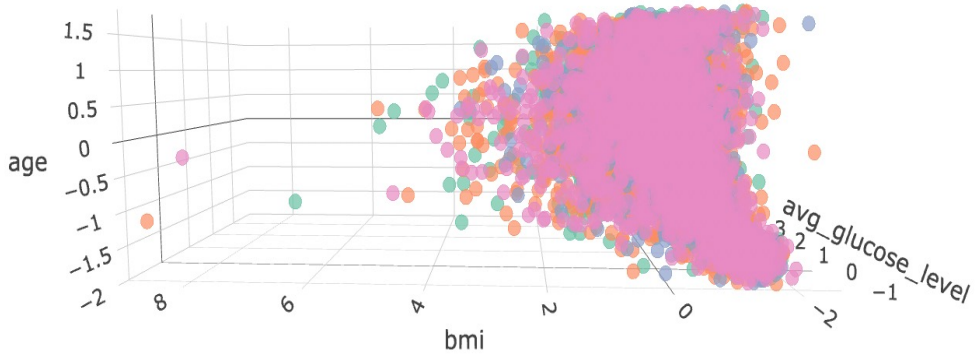



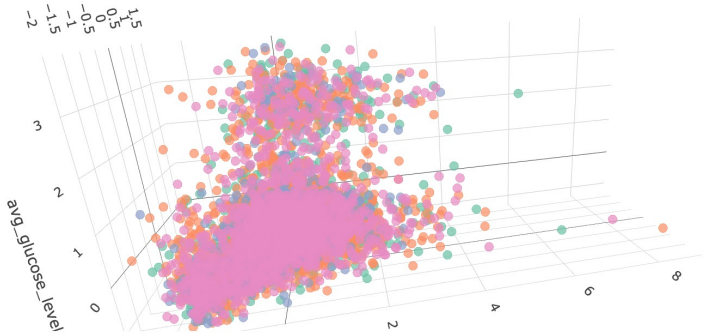




Barplot of Stroke
0
1000
2000
3000
4000
5000
no stroke
stroke

# Is there any natural segmentation between people who have had a stroke or not based on their lifestyle habits (e.g. Bmi, glucose level etc)?

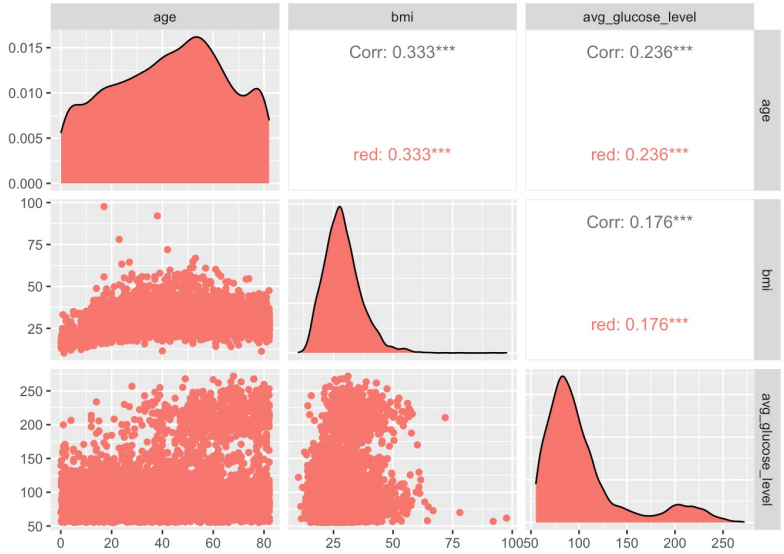


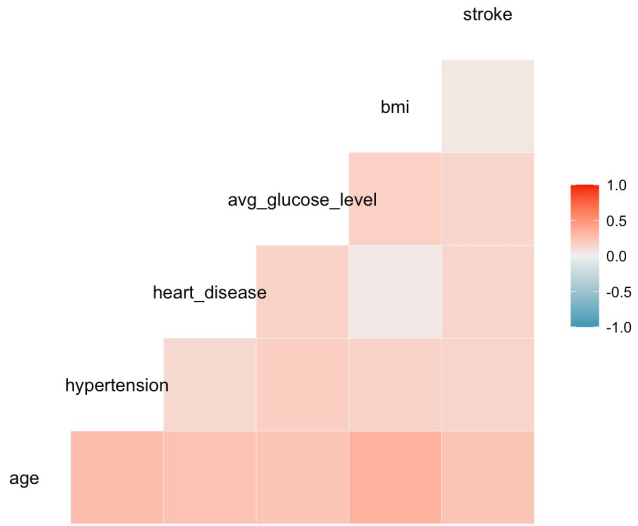



There don’t seem to be any prominent clusters of data, so there is no natural clustering between the participants of this study.

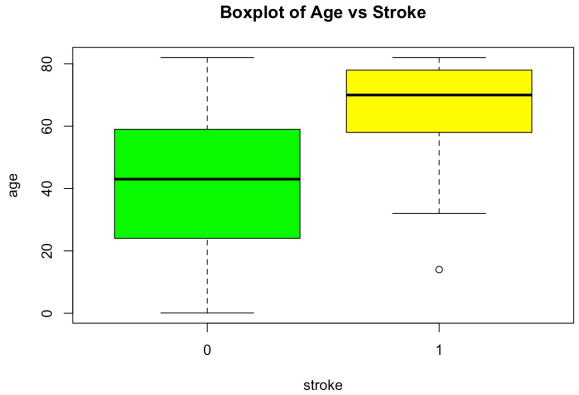
# Is there any significant linear relationship between the variables in this dataset (age, bmi, avg_glucose level)?

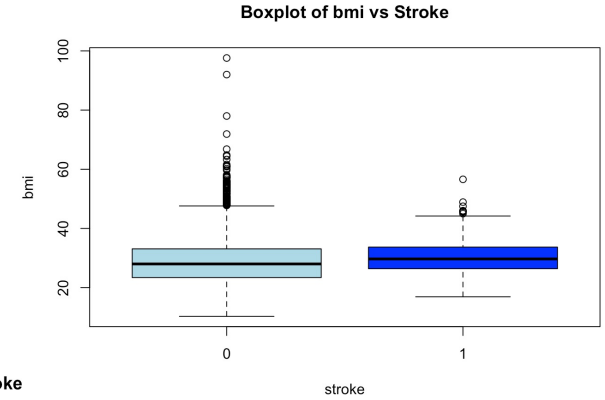


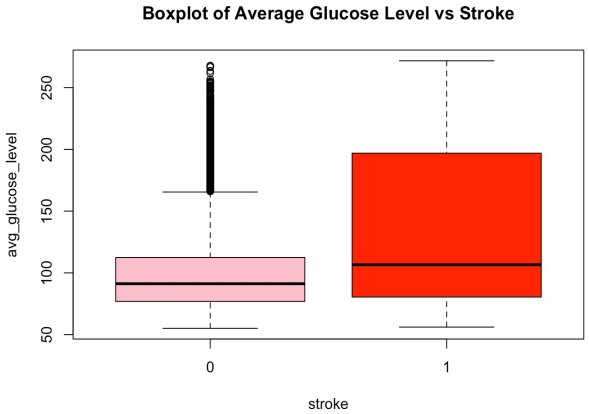



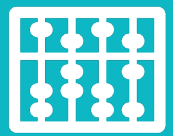

- There are no obvious strongly linear relationships between the continuous variables bmi, average glucose level and age. Thus, when we compute the model with stroke as dependent variable, we will not have to worry about **multicollinearity**.
- The strongest correlated variables seem to be bmi and age, and then age and hypertension, but even these correlations are quite low.

# Are older people/people with higher bmis/people with higher average glucose levels more likely to suffer from stroke?



# 3.
# Feature Selection 

Since our main research question is

"**What are the variables that best allow us to predict stroke for an individual?**",

we will try to do model selection using three methods:

- Lasso Regression with the most appropriate value for lambda
- Stepwise regression based on p-values
  - Backward Elimination
  - Forward Selection

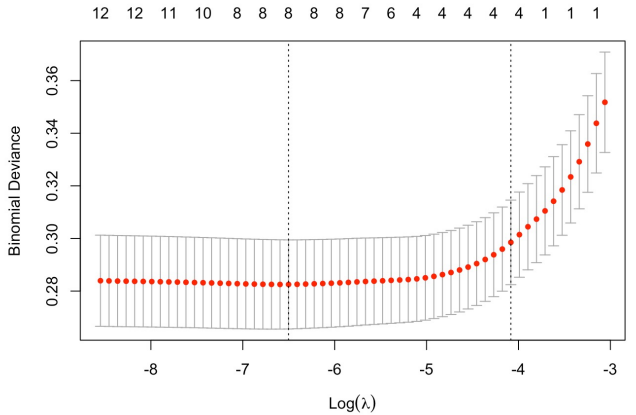
# Lasso Regression for Model Selection

Lasso Regression has discovered the following representative attributes for predicting stroke:

- age
- hypertension
- heart_disease
- work_type
- avg_glucose_level
- smoking_status

```
lasso_fit <- glmnet(x = X, y = stroke$stroke, lambda = 0.001500674, family = "binomial")
coef(lasso_fit)
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                                      s0
## (Intercept)                -7.522473918
## genderMale                  .
## genderOther                 .
## age                         0.066255878
## hypertension                0.504055849
## heart_disease               0.351618296
## ever_marriedYes             .
## work_typeGovt_job           .
## work_typeNever_worked       .
## work_typePrivate            0.084822677
## work_typeSelf-employed     -0.199362826
## Residence_typeUrban         .
## avg_glucose_level           0.004446061
## bmi                         .
## smoking_statusnever smoked  .
## smoking_statussmokes        0.245736020
## smoking_statusUnknown      -0.110532397
```



# To consolidate and examine the lasso regression, we used stepwise regression for model selection as well.

**P-value Backward elimination**

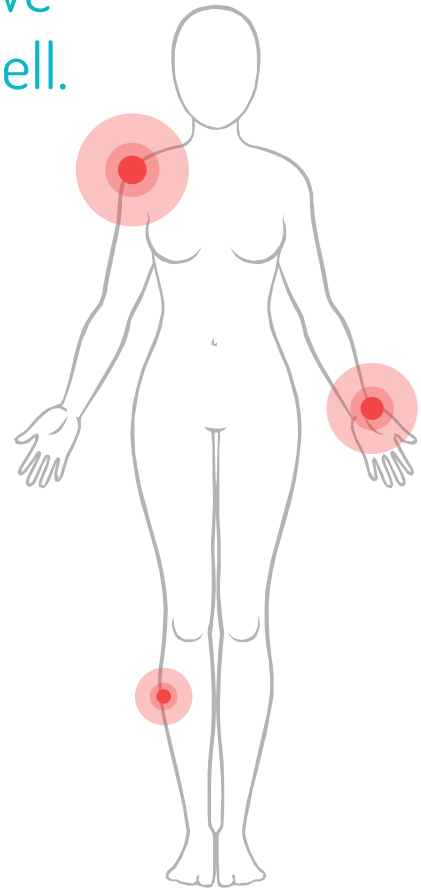
The model discovered by backward elimination based on p-values contains:

- age
- hypertension
- heart_disease
- work_type
- avg_glucose_level
- ever_married

**P-value Forward selection**

Final model by forward selection based on p-values contains the following attributes:

- age
- hypertension
- avg_glucose_level
- heart_disease



Since lasso, backward elimination and forward selection all yield different models, for the classification part we will stick to the results obtained via

- Lasso regression (**more reliable**)
- Forward selection based on p-values (**yields a minimal model**)

”

## 4.Prediction:

Use decision tree and naive bayes to construct prediction model respectively using parameters obtained via lasso regression and forward selection

# Decision Tree (with parameters from lasso regression)

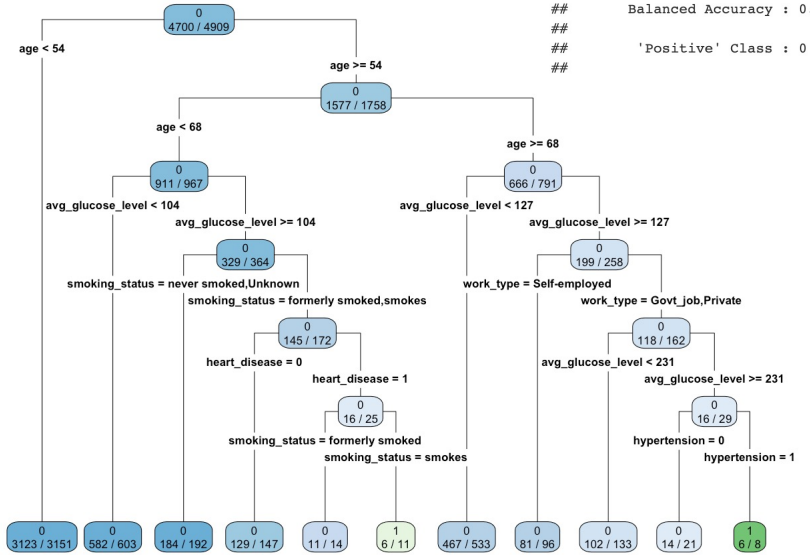
- For this decision tree, the number of false positives is very low at 7. On the other hand, the number of true positives is relatively good at 12. This seems like a **moderately good model** in terms of performance, because the number of false negatives is a bit high at 197.
- In order to be classified as a positive, a person must be either:

❑ older than 54 years, but younger than 65, with an average glucose level of 104 or more, currently a smoker, having heart disease or

❑ older than 68, with an average glucose level of 231 or more, working for the government or at a private company, and having hypertension

- PPV = 95.9% (quite high)
- Detection Rate = 95.6%.

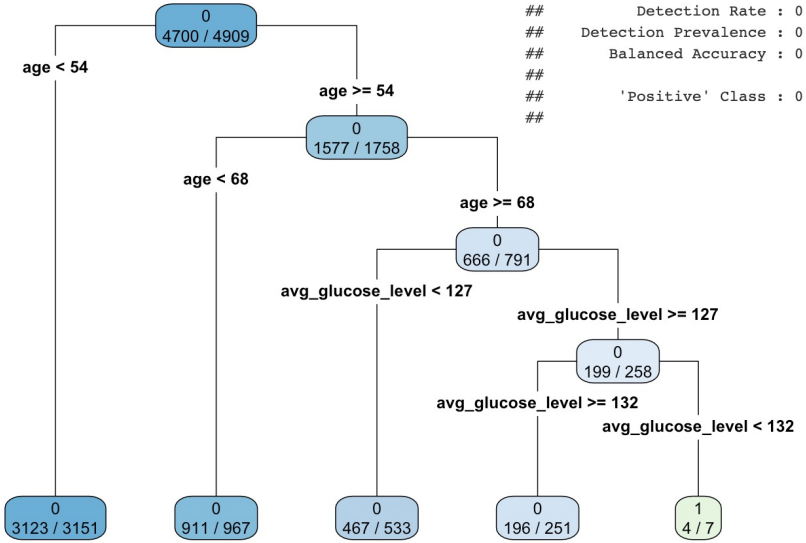
```
## Confusion Matrix and Statistics
##
##
## pred    0    1
##    0 4693  197
##    1    7   12
##
##                Accuracy : 0.9584
##                  95% CI : (0.9525, 0.9639)
##     No Information Rate : 0.9574
##     P-Value [Acc > NIR] : 0.3789
##
##                   Kappa : 0.0989
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.99851
##             Specificity : 0.05742
##          Pos Pred Value : 0.95971
##          Neg Pred Value : 0.63158
##              Prevalence : 0.95743
##          Detection Rate : 0.95600
##    Detection Prevalence : 0.99613
##       Balanced Accuracy : 0.52796
##
##        'Positive' Class : 0
##
```



# Decision Tree (with parameters from forward selection)

- For the second decision tree, the situation is similar, in that the number of false positives is 3 and the number of true positives is 4. This seems to be an only **moderately good model,** unfortunately, in terms of performance, because, again, the number of false negatives is high at 205.
- In order to be classified as a positive, a person must be older than 68, with an average glucose level between 127 and 132.
- PPV = 95.8% (quite high)
- Detection Rate = 95.6%.

```
## Confusion Matrix and Statistics
##
##
##          pred    0    1
##             0 4697  205
##             1    3    4
##
##                Accuracy : 0.9576
##                  95% CI : (0.9516, 0.9631)
##     No Information Rate : 0.9574
##     P-Value [Acc > NIR] : 0.4902
##
##                   Kappa : 0.0344
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.99936
##             Specificity : 0.01914
##          Pos Pred Value : 0.95818
##          Neg Pred Value : 0.57143
##              Prevalence : 0.95743
##          Detection Rate : 0.95681
##    Detection Prevalence : 0.99857
##       Balanced Accuracy : 0.50925
##
##        'Positive' Class : 0
##
```



# Naïve Bayes Prediction (with parameters from Lasso Regression)

- Our second option was to do a Naive Bayes classifier, in the hope that we will obtain more reliable results:
- The results from the first Naive Bayes classifier are indeed a lot more reliable. The model did classify some instances as 1, with 121 false negatives and 88 true positives. The accuracy is high at 87.6%, as well as the positive predictive value at 97%. The detection rate is 85%. All in all, this is a better classifier than decision tree.

```
## Confusion Matrix and Statistics
##
##
##  pred    0    1
##     0 4214  121
##     1  486   88
##
##                Accuracy : 0.8763
##                  95% CI : (0.8668, 0.8854)
##     No Information Rate : 0.9574
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1732
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8966
##             Specificity : 0.4211
##          Pos Pred Value : 0.9721
##          Neg Pred Value : 0.1533
##              Prevalence : 0.9574
##          Detection Rate : 0.8584
##    Detection Prevalence : 0.8831
##       Balanced Accuracy : 0.6588
##
##        'Positive' Class : 0
##
```

# Naïve Bayes Prediction (with parameters from Forward Selection)

- We also implemented another Naive Bayes classifier based on the minimal model discovered through forward selection based on p-values:
- This model is comparable to the previous one. It yielded 124 false negatives and 85 true positives. The accuracy is 88.4%, and the positive predicted value is 97%. The detection rate is also higher at 86.68%.

```
## Confusion Matrix and Statistics
##
##
##  pred    0    1
##     0 4255  124
##     1  445   85
##
##                Accuracy : 0.8841
##                  95% CI : (0.8748, 0.8929)
##     No Information Rate : 0.9574
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.18
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9053
##             Specificity : 0.4067
##          Pos Pred Value : 0.9717
##          Neg Pred Value : 0.1604
##              Prevalence : 0.9574
##          Detection Rate : 0.8668
##    Detection Prevalence : 0.8920
##       Balanced Accuracy : 0.6560
##
##        'Positive' Class : 0
##
```

# 5.
# Conclusion 

Since the Naive Bayes model with lasso regression predictors yield a higher number of true positives, along with more reliable accuracy and prediction outcomes, our team decides to choose this model as our optimal option.

**Findings:**

- Curiously enough, bmi was not included in any classification model after feature selection, even though, from a practical perspective, we would expect it to be somewhat important.

- Naive Bayes Classifier and Decision Tree seems to be relatively robust to imbalance in the dataset, and are better predictive models for stroke in this context.

**Future Directions:**

- In order to build even better models in the future, we could try to counter the imbalance in the dataset by oversampling the minority class (~ 200 people, i.e. the people who had a stroke) or downsampling the majority class (~ 4700 people, i.e. the people who have not had a stroke).

# THANK YOU!