# Detecting risk of Stroke through Supervised Learning Analysis

Group 11: Ordinary Leading Students

6/02/2021

**Abstract**

In the following paper we will present a statistical analysis of stroke related data. Our goal will be to predict whether a patient is likely to get a stroke based on multiple input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

After some data cleaning, we plotted some interesting variables of our dataset to get some insight on the content and structure of our data. We then proceeded to the analysis.

## Data

We will now load and present our dataset (that can be retrieved at the following link).

1. id: unique identifier
2. gender: [Male, Female, Other]
3. age: age of the patient
4. hypertension $= \begin{cases} 1 & \text{if the patient has hypertension} \\ 0 & \text{otherwise} \end{cases}$
5. heart_disease $= \begin{cases} 1 & \text{if the patient has a heart disease} \\ 0 & \text{otherwise} \end{cases}$
6. ever_married = [Yes, No]
7. work_type = [children, Govt_jov, Never_worked, Private, Self-employed]
8. Residence_type = [Rural, Urban]
9. avg_glucose_level: average glucose level in blood
10. bmi: body mass index
11. smoking_status = [formerly smoked, never smoked, smokes, Unknown[1]]
12. stroke $= \begin{cases} 1 & \text{if the patient had a stroke} \\ 0 & \text{otherwise} \end{cases}$

[1] The information is unavailable for this patient

```
## 'data.frame':    5110 obs. of  12 variables:
##  $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
##  $ gender           : chr  "Male" "Female" "Male" "Female" ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
##  $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
##  $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
##  $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

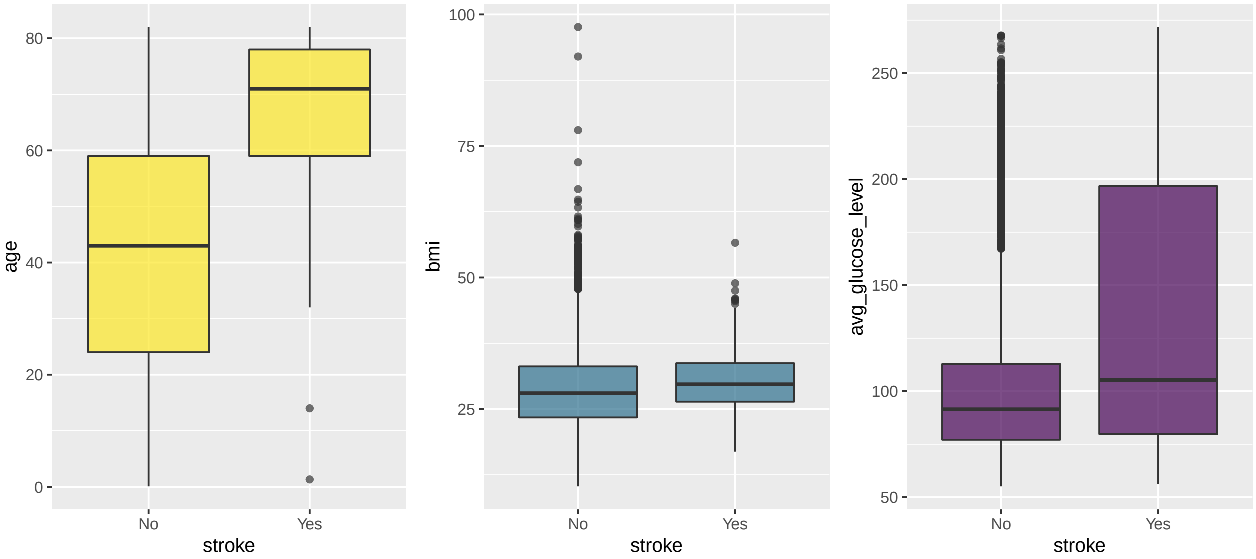
## Data Cleaning

We convert our variables to the right data type and we show the updated summary.

```
## 'data.frame':    5110 obs. of  11 variables:
##  $ gender           : Factor w/ 3 levels "Female","Male",..: 2 1 2 1 1 2 2 1 1 1 ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 1 1 1 ...
##  $ heart_disease    : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 1 ...
##  $ ever_married     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 1 2 2 ...
##  $ work_type        : Factor w/ 5 levels "children","Govt_job",..: 4 5 4 4 5 4 4 4 4 4 ...
##  $ avg_glucose_level: num  229 202 106 171 174 ...
##  $ bmi              : num  36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...
##  $ smoking_status   : Factor w/ 4 levels "formerly smoked",..: 1 2 2 3 2 1 2 2 4 4 ...
##  $ stroke           : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ residence_type   : Factor w/ 2 levels "Rural","Urban": 2 1 1 2 1 2 1 2 1 2 ...
```

## Data Visualization

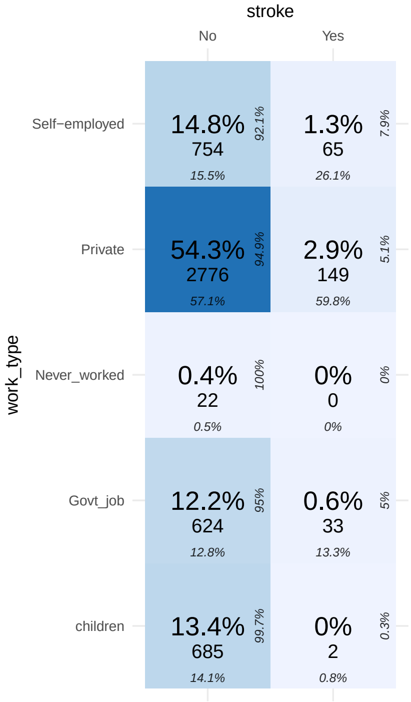
We can use different boxplots to visualize possible relations between our quantitative variables (**age**, **bmi** and **average glucose level**) and stroke status.

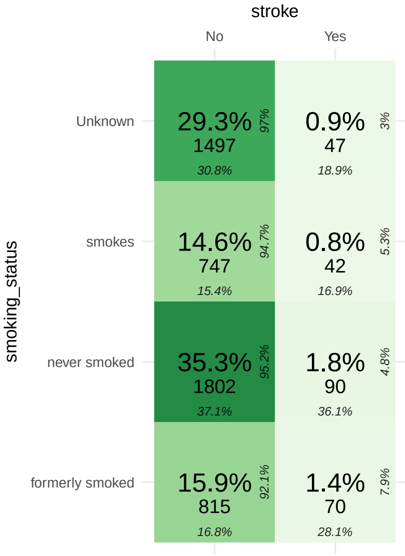


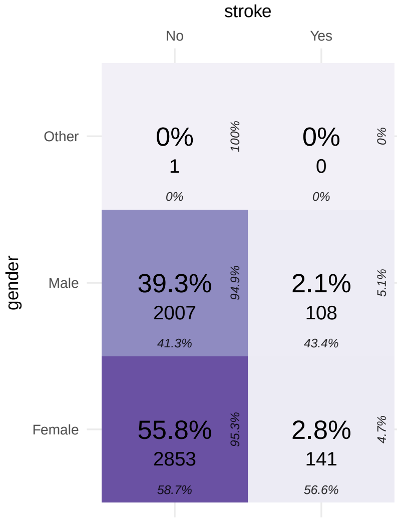
From the plots above we can infer the following:

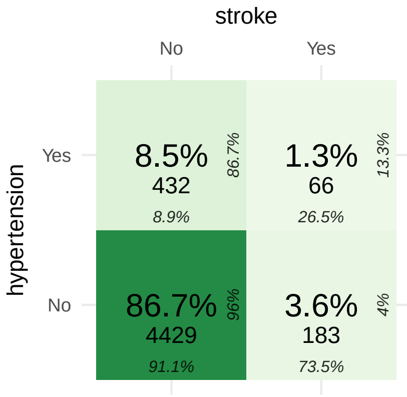
- **Age**: older people are more likely to have a stroke.
- **Bmi**: there is no evident relation between stroke and bmi.
- **Average glucose level**: the higher the level of glucose, the higher the relation with stroke

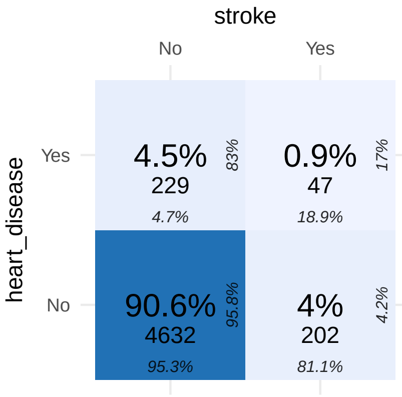
Below we plot a few matrices to visualize possible relations between our qualitative variables and stroke status.

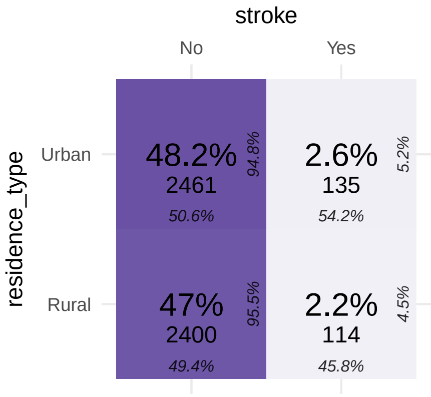


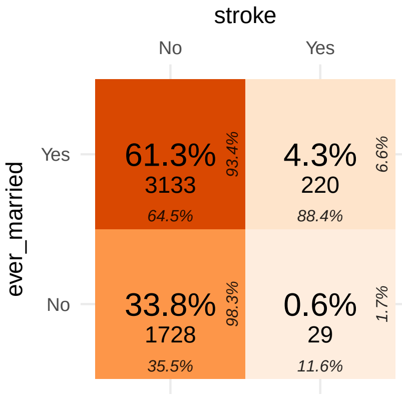








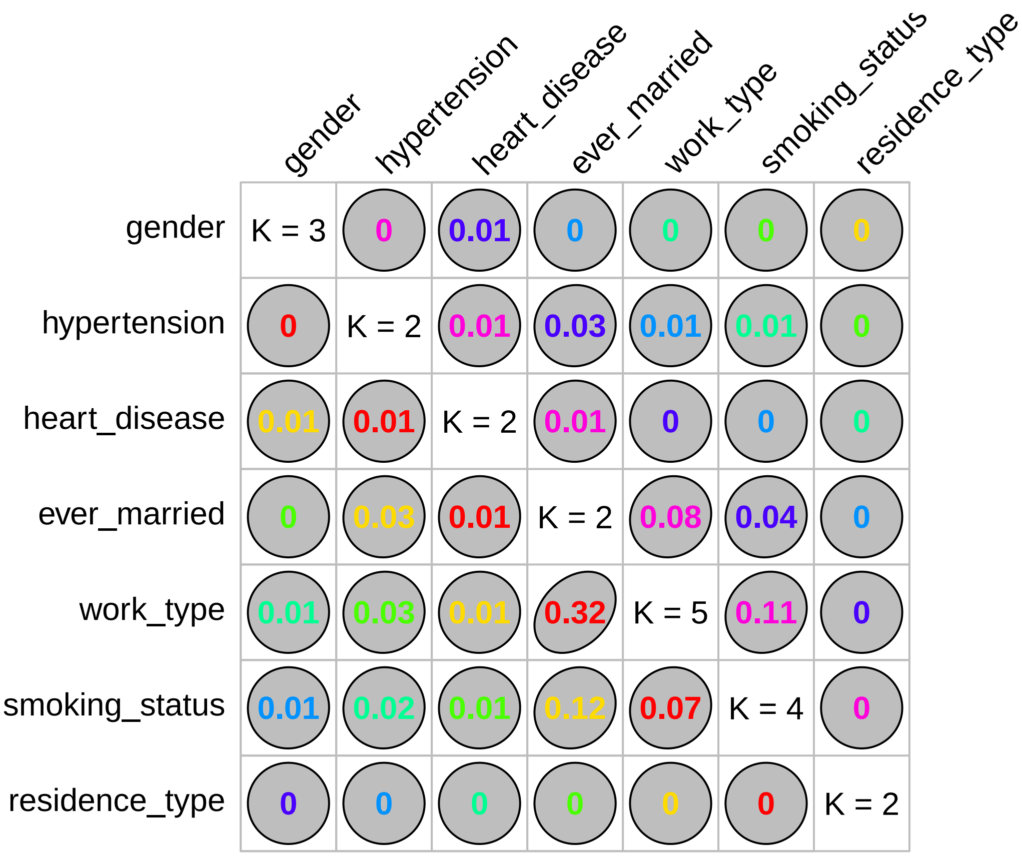




From these comparison matrices we can deduce the following:

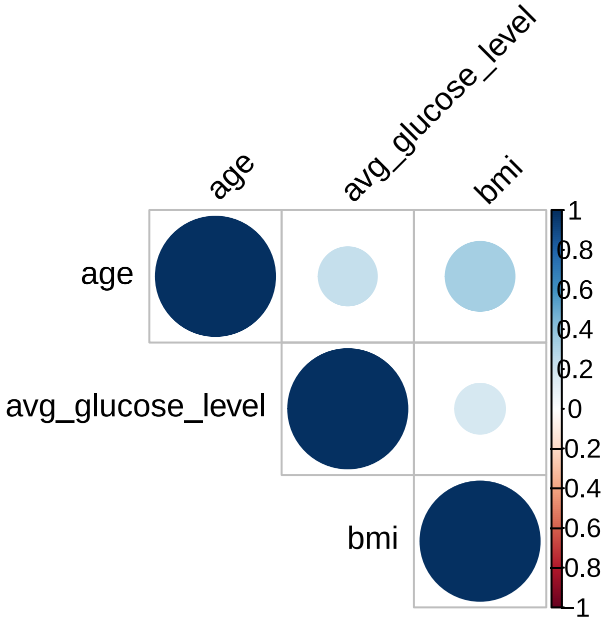
- A lot of our subjects work in the private sector (57.24%) and do not smoke (37%), we can also notice a high predominance of subjects without hypertension (90.25%) and heart diseases (94.6%). These data should not worry us because they do not indicate unbalance in our dataset but just reflects the actual distribution of global population[2].
- We miss a lot of information on the smoking status of our participants (30.2% unknown), this could evolve in a modeling problem if we do not address it properly.
- A lot of our subjects have not suffered a stroke (95.1%), this could cause us some problems but we will address it using the SMOTE technique.

Now we can finally show **qualitative** and **quantitative** correlations.



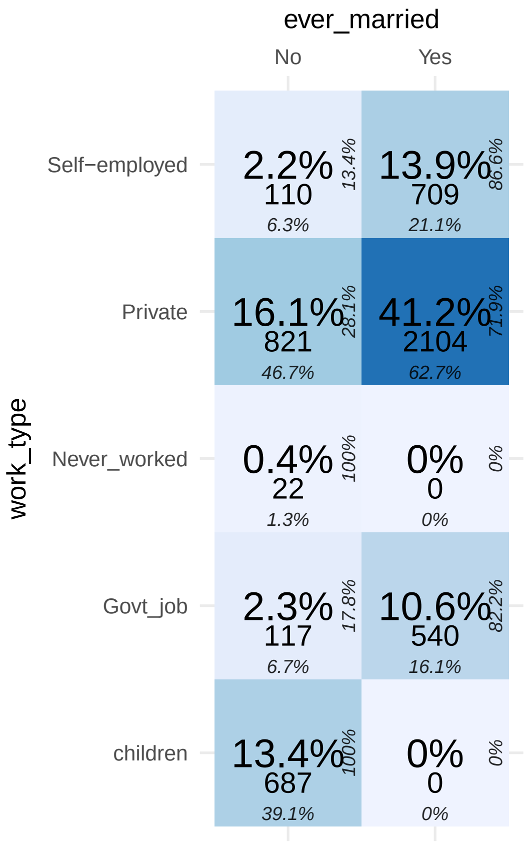
| | gender | hypertension | heart_disease | ever_married | work_type | smoking_status | residence_type |
|---|---|---|---|---|---|---|---|
| gender | K = 3 | 0 | 0.01 | 0 | 0 | 0 | 0 |
| hypertension | 0 | K = 2 | 0.01 | 0.03 | 0.01 | 0.01 | 0 |
| heart_disease | 0.01 | 0.01 | K = 2 | 0.01 | 0 | 0 | 0 |
| ever_married | 0 | 0.03 | 0.01 | K = 2 | 0.08 | 0.04 | 0 |
| work_type | 0.01 | 0.03 | 0.01 | 0.32 | K = 5 | 0.11 | 0 |
| smoking_status | 0.01 | 0.02 | 0.01 | 0.12 | 0.07 | K = 4 | 0 |
| residence_type | 0 | 0 | 0 | 0 | 0 | 0 | K = 2 |

[2] Further information on this subject can be found at the following links:
– Work type
– Smoking status
– Hypertension
– Heart diseases

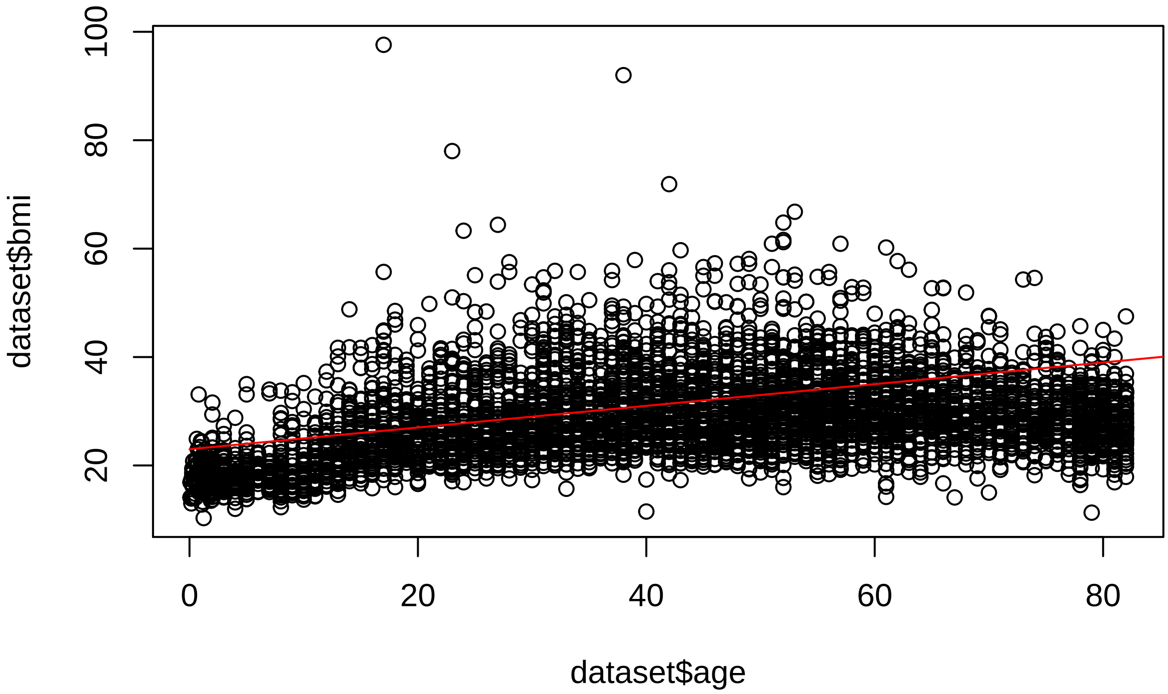


From the results above we are lead to believe that there is no multicollinearity in our instrumental variables. We will provide more evidence for this later, when we perform the regression.

The only two correlations that stand out are the one between work_type and ever_married and the one between age and bmi, as expected. These two correlations are understandable given the nature of the data. Let us inspect this even further with some plots.



This plot shows why there is correlation between the variables work_type and ever_married: the levels Never_worked and children in the work_type variable both have 0% of observations who have ever been married. This is unserstandable since children can't be married and people who have never worked are young are really likely to have never been married.



```
## integer(0)
```

With this plot a slight correlation between the variables bmi and age is noticeable. However, this correlation is not strong enough to influence our model later on.

# Data Manipulation

In order to facilitate our work, we removed the level **other** from the variable gender and the level **never__worked** from the variable work_type, since they are not relevant for the analysis.

## NA Values

We now look for NA values in our dataset:

```
## [1] "bmi ->  201  NA"
```

We notice that we have 201 NA just in the **bmi** variable. In order to avoid losing these rows, given that we already have a small dataset, we will predict their values using a Decision Tree to approximate its value.

# Statistical Model

We will now begin splitting our dataset in train/test.

```
##
##   No  Yes
## 2903  150
```

```
##
##   No  Yes
## 1935   99
```

## SMOTE algorithm

Since our dataset has a problem of unbalance, we will use the SMOTE algorithm to create synthetic new data.

```
## [1] "This was the regular training set balance between the response classes:"
```

```
##
##   No  Yes
## 2903  150
```

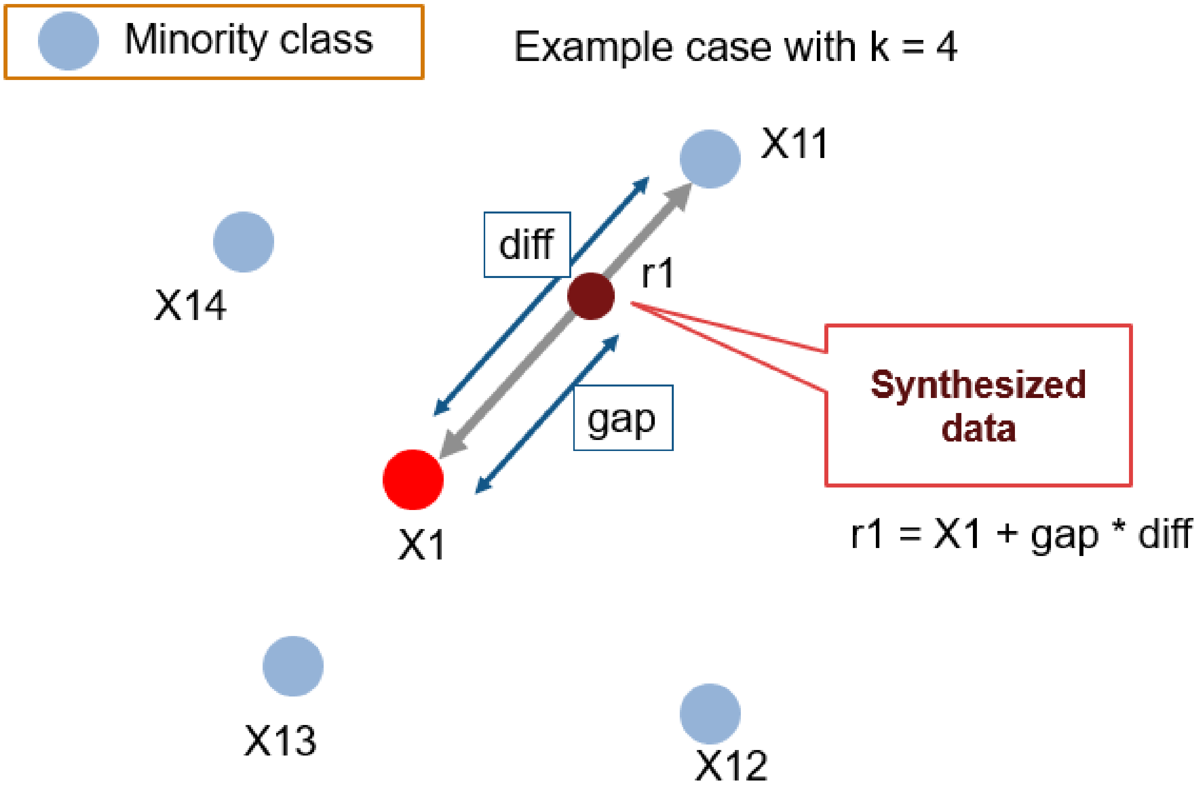
This is the oversampled one. Notice that we could have oversampled since reaching the perfect balance. This is, altough, not advisable. We can obtain better performances helping the algorithms understand this still is a minority class and indeed, we do.

```
table(train_smoted$stroke)
```

```
##
##   No  Yes
## 2903 1452
```

Notice that the SMOTE algorithm was only applied to the train set, this is done on purpose to avoid the evaluation of our final model on synthetic data.

The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors.



Take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features.

### Logistic Regression

We perform a logistic regression using **stroke** as the dependent variable and the rest of the data as independent variables in order to see if we find some significant relationships.

```
##
## Call:
## glm(formula = stroke ~ ., family = binomial(link = "logit"),
##     data = train_smoted)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2672  -0.6358  -0.1857   0.7041   3.3126
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -7.1858514  1.0234287  -7.021 2.20e-12 ***
## genderMale                   0.1440997  0.0855990   1.683 0.092292 .
## age                          0.0990762  0.0036705  26.992  < 2e-16 ***
## hypertensionYes             -0.0105937  0.1119036  -0.095 0.924578
## heart_diseaseYes            -0.3477259  0.1399267  -2.485 0.012953 *
## ever_marriedYes            -0.2294383  0.1353478  -1.695 0.090042 .
## work_typeGovt_job          -0.7793077  1.0421037  -0.748 0.454568
## work_typePrivate            0.5788146  1.0312462   0.561 0.574609
## work_typeSelf-employed     -0.1125092  1.0398675  -0.108 0.913840
## avg_glucose_level           0.0036977  0.0007985   4.631 3.65e-06 ***
## bmi                         0.0058711  0.0071929   0.816 0.414363
## smoking_statusnever smoked -0.1947215  0.1092909  -1.782 0.074801 .
## smoking_statussmokes        0.4408541  0.1294917   3.404 0.000663 ***
## smoking_statusUnknown      -0.0792764  0.1286776  -0.616 0.537838
## residence_typeUrban        -0.1232177  0.0833555  -1.478 0.139349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5544.5  on 4354  degrees of freedom
## Residual deviance: 3572.9  on 4340  degrees of freedom
## AIC: 3602.9
##
## Number of Fisher Scoring iterations: 8
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## gender            1.060133  1        1.029628
## age               1.379844  1        1.174668
## hypertension      1.127855  1        1.062005
## heart_disease     1.053944  1        1.026618
## ever_married      1.078156  1        1.038343
## work_type         1.266750  3        1.040196
## avg_glucose_level 1.154471  1        1.074463
## bmi               1.131193  1        1.063576
## smoking_status    1.220469  3        1.033763
## residence_type    1.017945  1        1.008933
```

The significant variables are `age, heart disease, average glucose level and smoking status`, which have a positive impact on the increase of the logit probability. Note `gender Male` and `Married`.

Now we can also state with certainty that there is no multicollinearity in our data, due to the results of the variance inflation factor test. In fact, all the values of the VIF are very close to 1, which indicates an almost total absence of multicollineairty.

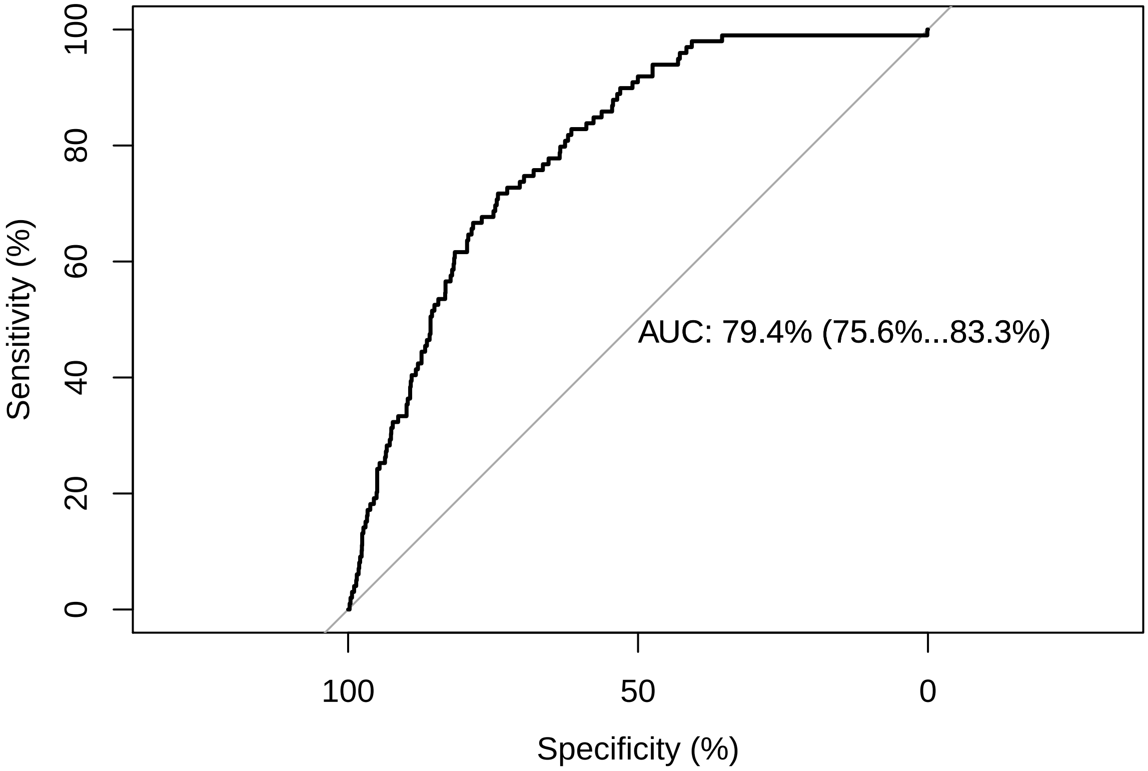
## Confusion Matrix and Statistics

We will now evaluate our model over the test set.

```
## Confusion Matrix and Statistics
##
##          Actual
## Predicted  Yes   No
##       Yes   53  313
##       No    46 1622
##
##                Accuracy : 0.8235
##                  95% CI : (0.8062, 0.8398)
##     No Information Rate : 0.9513
##     P-Value [Acc > NIR] : 1
##
```

```
##                   Kappa : 0.1639
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.53535
##             Specificity : 0.83824
##          Pos Pred Value : 0.14481
##          Neg Pred Value : 0.97242
##              Prevalence : 0.04867
##          Detection Rate : 0.02606
##    Detection Prevalence : 0.17994
##       Balanced Accuracy : 0.68680
##
##        'Positive' Class : Yes
##
```

```
## [1] "F1: 0.227956989247312"
```



```
## Area under the curve: 79.42%
```

As we can see from the outputs above, our model does not perform very well. it correctly predicts just bit more than 50% of the positive results (in our case, people who suffered a stroke), at the expense of a lot of false positives. Because this is such a sensitive subject, we would definitely prefer having more false positives than false negatives.

The problem of this underperforming model could be addressed in two different ways:

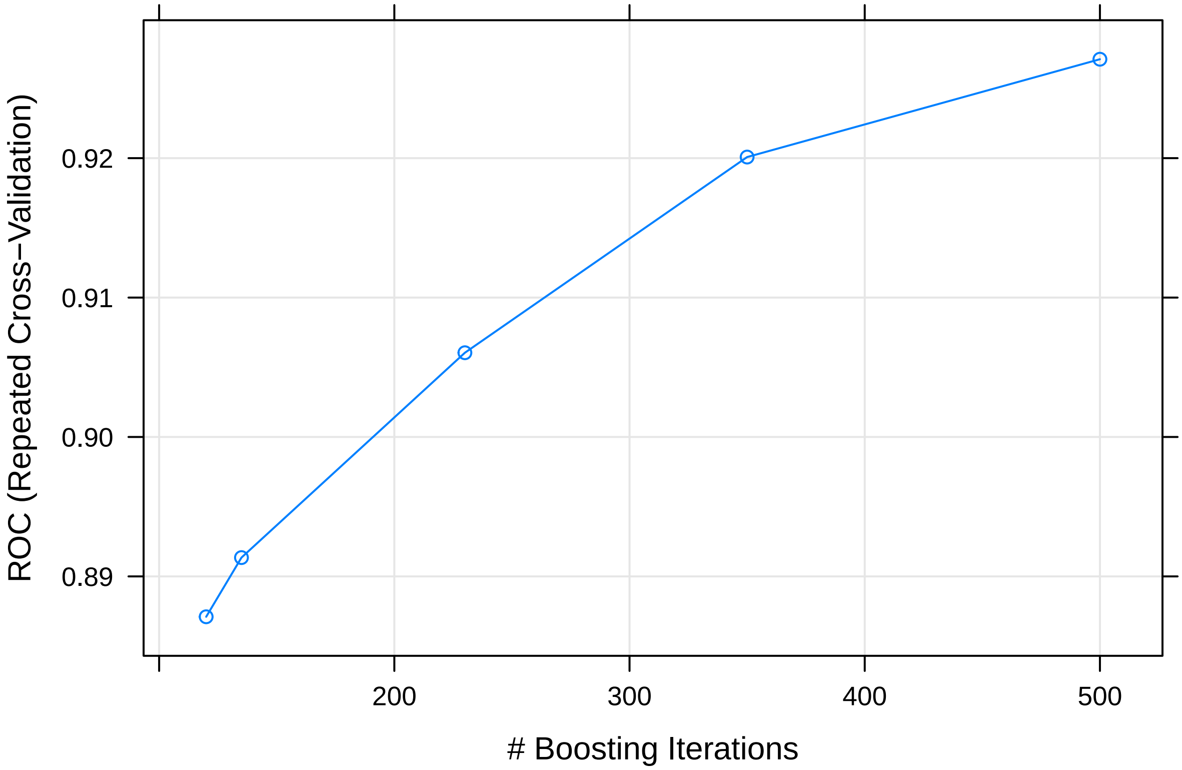
1. We could speak with the stakeholders to understand how to best tune our threshold. In a real-world scenario this would translate into a discussion regarding the number of people we can take care of, even if they are false positives, in order to avoid rejecting people who actually need our help.
2. We could try and increase the accuracy of our model, and we will do that below using a Boosting

technique.

## Boosting

Boosting is a technique used that convert weak learners to strong ones, and it works by sequentially learning weak classifiers with respect to a distribution and adding them to a final strong classifier. Notice that, when they are added to the model, all the learners are weighted with relation to their prediction accuracy.

```
## Boosted Logistic Regression
##
## 4355 samples
##   10 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 3918, 3919, 3920, 3920, 3920, 3920, ...
## Resampling results across tuning parameters:
##
##   nIter  ROC        Sens       Spec
##   120    0.8871032  0.9275128  0.7901012
##   135    0.8913483  0.8704010  0.7287152
##   230    0.9060425  0.9335105  0.8246046
##   350    0.9200796  0.9442832  0.8338712
##   500    0.9270957  0.9529053  0.8334289
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was nIter = 500.
```



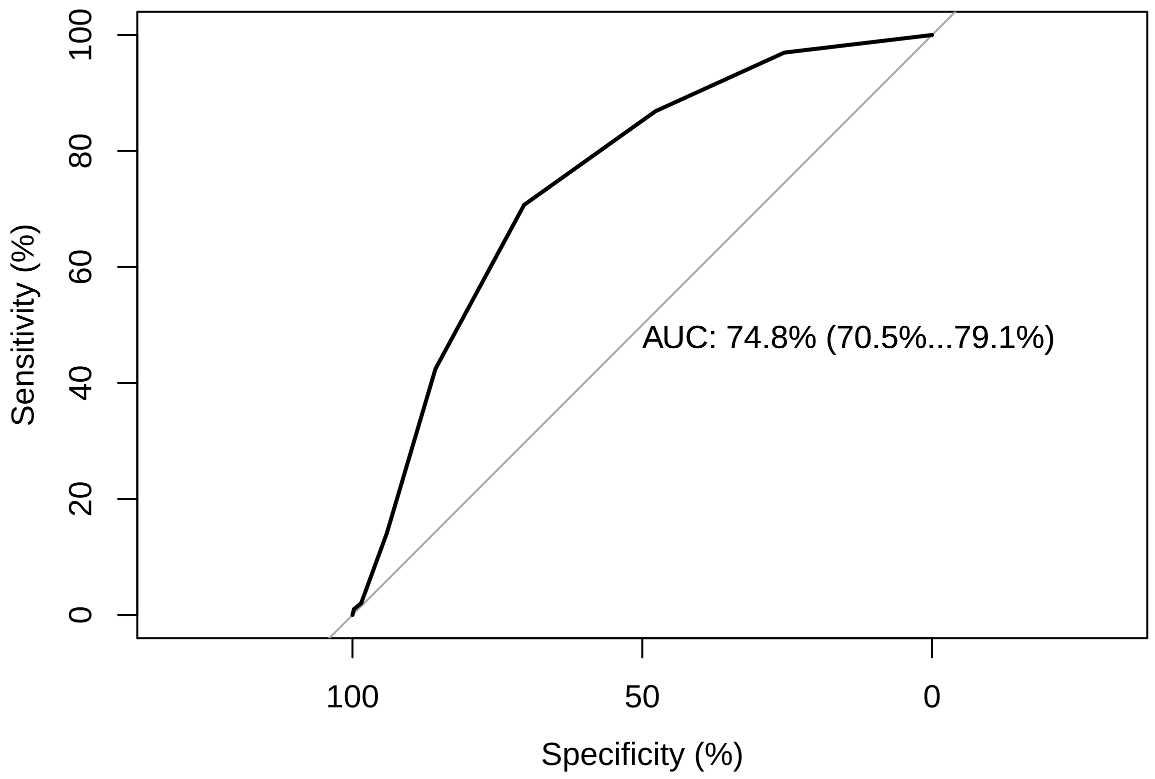
```
## Confusion Matrix and Statistics
```

```
##
##          Actual
## Predicted  Yes   No
##       Yes   14  115
##       No    85 1820
##
##                Accuracy : 0.9017
##                  95% CI : (0.8879, 0.9143)
##     No Information Rate : 0.9513
##     P-Value [Acc > NIR] : 1.0000
##
##                   Kappa : 0.0717
##
##  Mcnemar's Test P-Value : 0.0403
##
##             Sensitivity : 0.141414
##             Specificity : 0.940568
##          Pos Pred Value : 0.108527
##          Neg Pred Value : 0.955381
##              Prevalence : 0.048673
##          Detection Rate : 0.006883
##    Detection Prevalence : 0.063422
##       Balanced Accuracy : 0.540991
##
##        'Positive' Class : Yes
##
```

```
## [1] "F1: 0.12"
```

```
## Setting levels: control = 1, case = 2
```

```
## Setting direction: controls < cases
```



We can notice in the output above that boosted logistic our model improves its accuracy at a constant threshold of 0.5 (from 82% to 90%) when 5000 iterations are computed but with a worst ROC AUC (74%).

It would probably be better to stand with a lower threshold and let space for less accuracy and more Sensitivity.

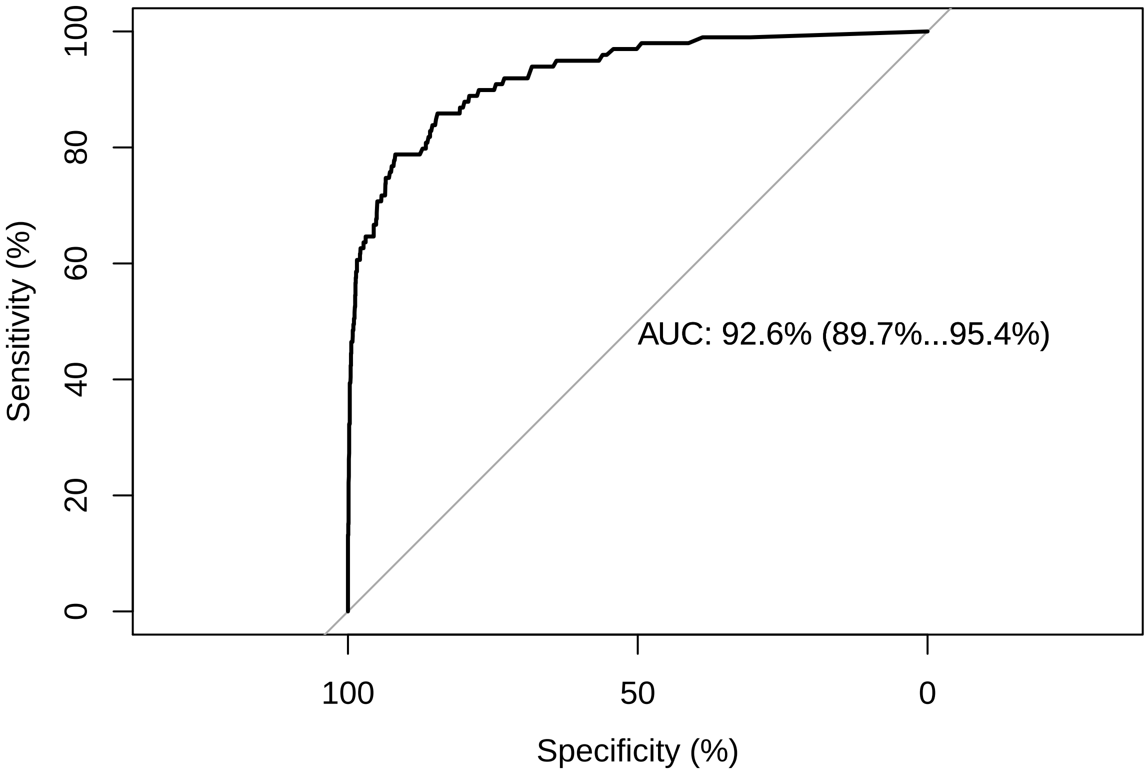
## Random Forest

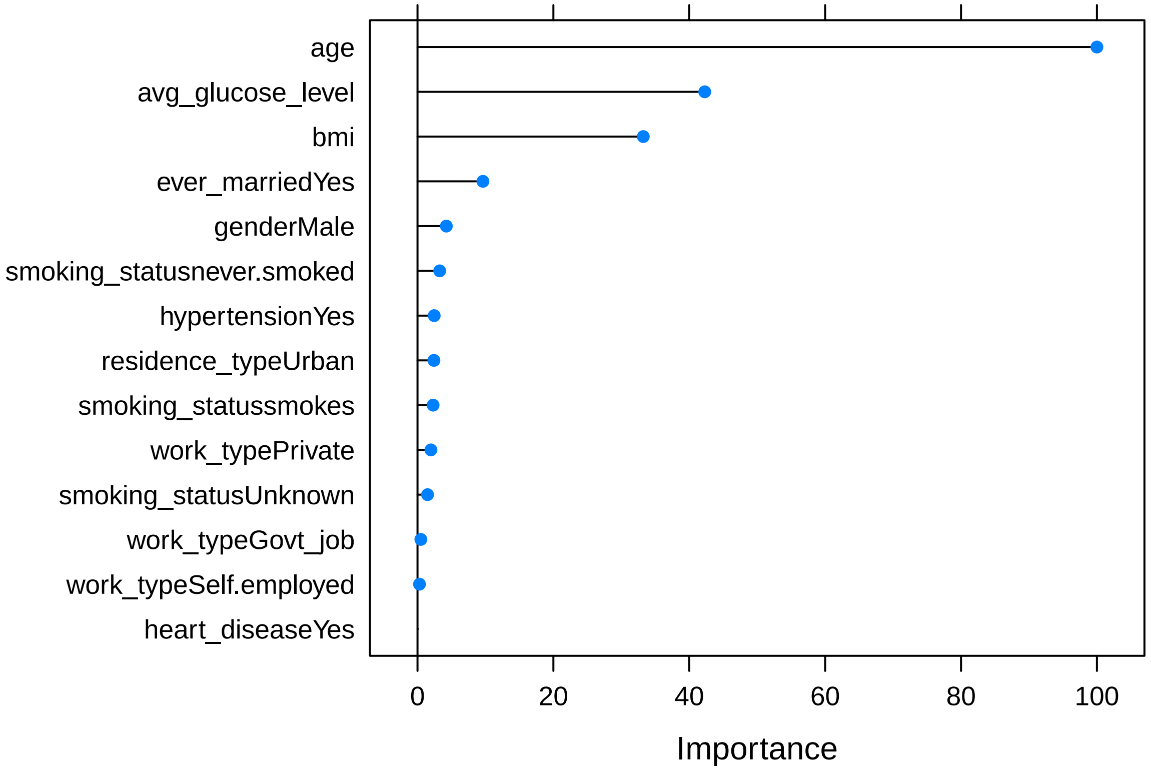
We further proceed with our last model. We're going to train a random forest with Repeated Cross Validation. for what it concerns the number of var at every tree, we're going to use the squared root of nVar + 6 since the number of significant variables is pretty tight

```
## Random Forest
##
## 5806 samples
##   10 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (3 fold, repeated 5 times)
## Summary of sample sizes: 3871, 3871, 3870, 3871, 3870, 3871, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.9179121  0.8358245
##
## Tuning parameter 'mtry' was held constant at a value of 9.316625
```

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicted  Yes   No
##       Yes   64   63
##       No    35 1872
##
##                Accuracy : 0.9518
##                  95% CI : (0.9416, 0.9607)
##     No Information Rate : 0.9513
##     P-Value [Acc > NIR] : 0.485631
##
##                   Kappa : 0.5413
##
##  Mcnemar's Test P-Value : 0.006383
##
##             Sensitivity : 0.64646
##             Specificity : 0.96744
##          Pos Pred Value : 0.50394
##          Neg Pred Value : 0.98165
##              Prevalence : 0.04867
##          Detection Rate : 0.03147
##    Detection Prevalence : 0.06244
##       Balanced Accuracy : 0.80695
##
##        'Positive' Class : Yes
##
```

```
## [1] "F1: 0.57"
```




age
avg_glucose_level
bmi
ever_marriedYes
genderMale
smoking_statusnever.smoked
hypertensionYes
residence_typeUrban
smoking_statussmokes
work_typePrivate
smoking_statusUnknown
work_typeGovt_job
work_typeSelf.employed
heart_diseaseYes
0
20
40
60
80
100
Importance

# Conclusion

In conclusion, we were able to construct a model that is able to predict with an 95% level of accuracy and with an outstanding 92% of ROC_AUC if an individual is more likely to suffer from a stroke than others based on his clinical records. Furthermore, we can state that:

- **Age**: Age is surely the most important variable for us to exclude whether or not a person is likely to suffer a stroke or not.
- **Average Glucose Level**: in terms of the logistic regression a one unit increase in average glucose level is associated with an increase in the log odds of having a stroke by 0.003 units
- **BMI**: one Notable difference between the Random Forest and the Logistic Regression are BMI and Smoking Status. While Logit model stated the latter as significant for our analysis and the first as not, the random forest attributes the third spot of the podium to the BMI and only a minor contribute to the smoking Status.