# Heart Stroke Prediction

ADVANCED STATISTICAL METHODS PROJECT REPORT


Submitted by:

**Ayesha Shariff (21BDA18)**

**Aaran D'Lima (21BDA23)**

**Shivanshi Maheshwari (21BDA33)**

Under the supervision of

Asst. Prof. (Ms.) Jayati Kaushik

**DEPARTMENT OF ADVANCED COMPUTING**

ST. JOSEPH'S COLLEGE (AUTONOMOUS)

36, Lalbagh Road, Bengaluru-560027


**JULY 2022**

## ACKNOWLEDGEMENT

The successful completion of any task is incomplete and meaningless without giving any due credit to the people who made it possible without which the project would not have been successful and would have existed in theory.

First and foremost, we are grateful to **Dr. Jayati Bhadra**, HOD, Department of Advanced Computing, St. Joseph's College, for giving us this opportunity. We owe a lot of thanks to our supervisor, **Ms. Jayati Kaushik**, Asst. Professor, Department of Advanced Computing, St. Joseph's College, for igniting and constantly motivating us and guiding us in the idea of a creatively and amazingly performed Major Project, in undertaking this endeavor and challenge and also for being there whenever we needed her guidance or assistance.

We would also like to take this moment to show our thanks and gratitude to one and all, who indirectly or directly have given us their hand in this challenging task. We feel happy and joyful and content in expressing our vote of thanks to all those who have helped us and guided us in presenting this project work for our project. Last, but never least, we thank our well-wishers and parents for always being with us, in every sense and constantly supporting us in every possible sense whenever possible.

| Ayesha Shariff | Aaran D'lima | Shivanshi Maheshwari |
|---|---|---|
| (21BDA18) | (21BDA23) | (21BDA33) |

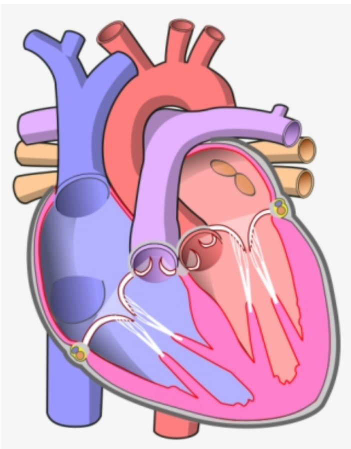
# Contents

# Chapter 1

# Introduction

## 1.1 Overview and Background

In today's world, Stroke is a critical health problem. It severely affects human health and lives. It is the second most deadly disease since 20th century. Stroke is caused as a result of blockage or bleeding of blood vessels which reduces the flow of blood to the brain. Due to this brain does not receives sufficient oxygen or nutrients and brain cells start to die.

Some of the symptoms of stroke are age, being overweight, high blood pressure, diabetes, high cholesterol, heart disease, etc. In present scenario, we come across many people who die of heart stroke and we feel that this study may help to some extent to provide insights to some of the reasons for heart stroke.

## 1.2 Definition

In 1970, the World Health Organization defined stroke as ’rapidly developed clinical signs of focal (or global) disturbance of cerebral function, lasting more than 24 hours or leading to death, with no apparent cause other than of vascular origin’.

## 1.3 Need for Study

In present scenario we come across many people die to heart stroke and we feel that this study may help to some extent to provide the some of the reasons for heart stroke.

## 1.4 Objectives

As a data analytic students we want to identify the risk factors for stroke. Main objective of our project is to predict whether people will have a stroke and reasons for stroke based on historical data. We are also interested in finding the stroke outcome for potential patients.

- We are trying to find out whether input variables or features affect the stroke outcome.
- We are predicting stroke outcome.

# Chapter 2

# Dataset

## 2.1 Information Regarding Dataset

We got the dataset of kaggle.
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv
Our dataset showcases person's body features and their stroke status. It contains 5110 rows with 12 columns. Each row gives the required information about the person. Overall we are finding whether a person is likely to get stroke based on the input parameters such as BMI, hypertension, work type, etc.
Below is the attribute information.

- id: unique
- gender: "Male", "Female" or "Other"
- age: age of the person
- hypertension: 0 if the person does not have hypertension, 1 if the person has hypertension
- heart_disease: 0 if the patient does not have any heart diseases, 1 if the patient has a heart disease
- ever_married: "No" or "Yes"
- work_type: "Children", "Govt_job", "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

## 2.2 Summary of the dataset

- There are 9 input variables and 1 outcome(stroke) in the dataset. For our analysis we don’t need ID.
- We can see that some of the columns values are in character, therefore we must change it into factor or number.
- The interesting fact that we must observe is then mean of stroke is 0.04, which means only 4% of the patients have stroke.

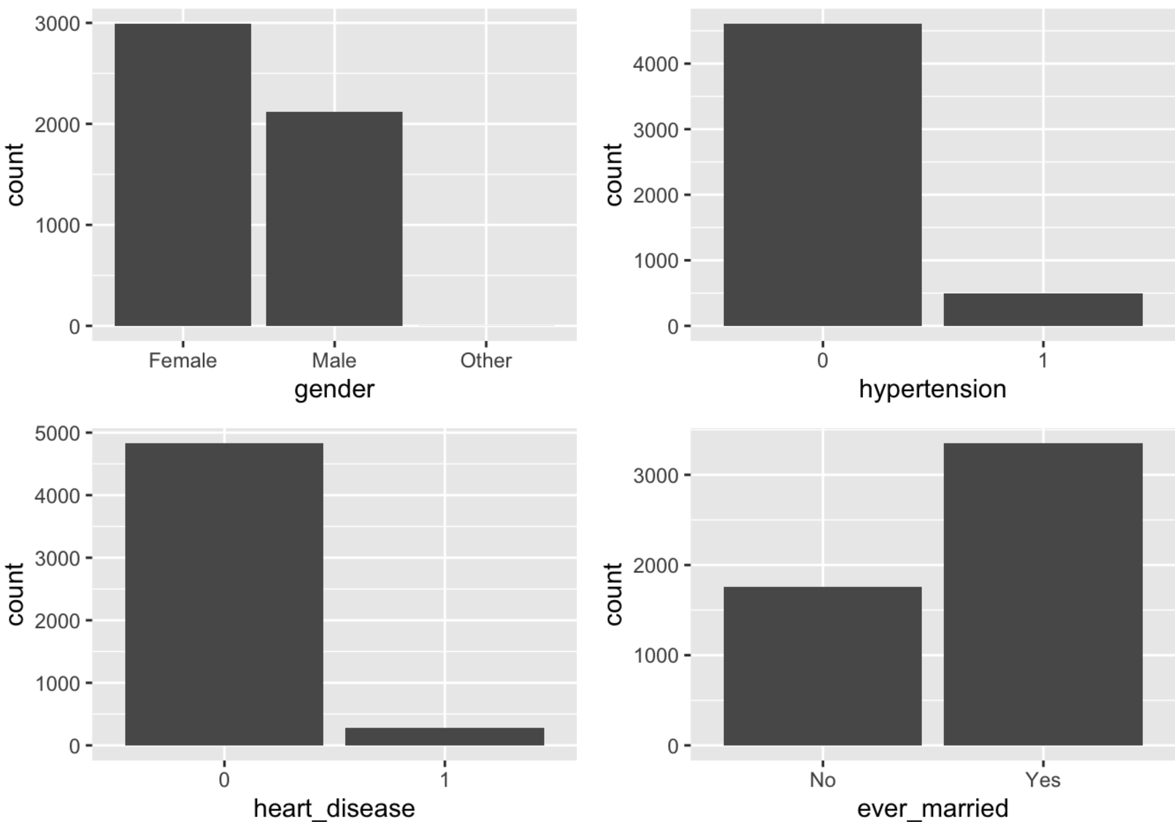
## 2.3 Data Cleaning

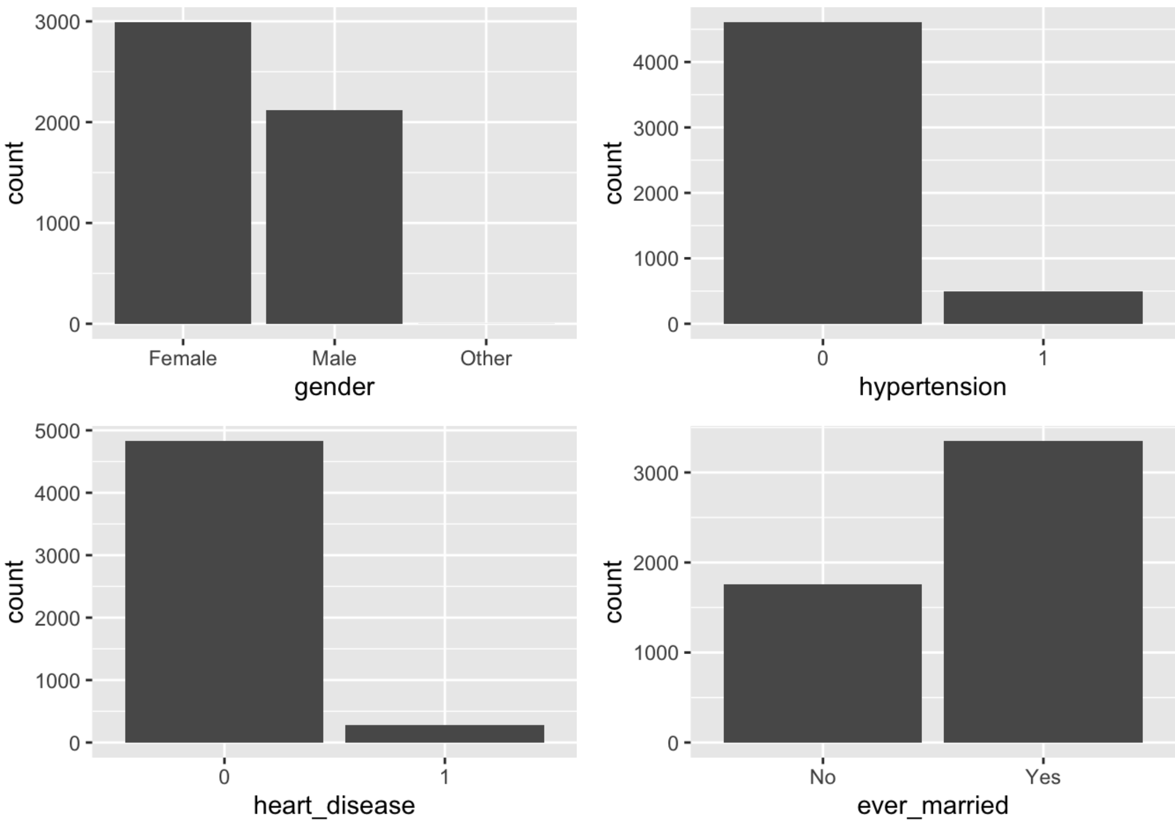
- Firstly, we have dropped the column ID.
- Secondly, we have found that there are many null values in the column BMI. Therefore we have replaced null values with median of the BMI.
- In the smoking_status column, we have found that there are large number of unknowns. Therefore we have replaced unknown with most frequent category ‘Never smoked’.
- Similarly, we examined the columns such as work_type, residence_type and ever_married.
- In order to perform Exploratory Data Analysis, we have transformed the columns that are categorical, into binary variables.

# Chapter 3

# Exploratory Data Analysis

## 3.1 In our dataset we have discrete variables. We have used barplot to show their distribution.

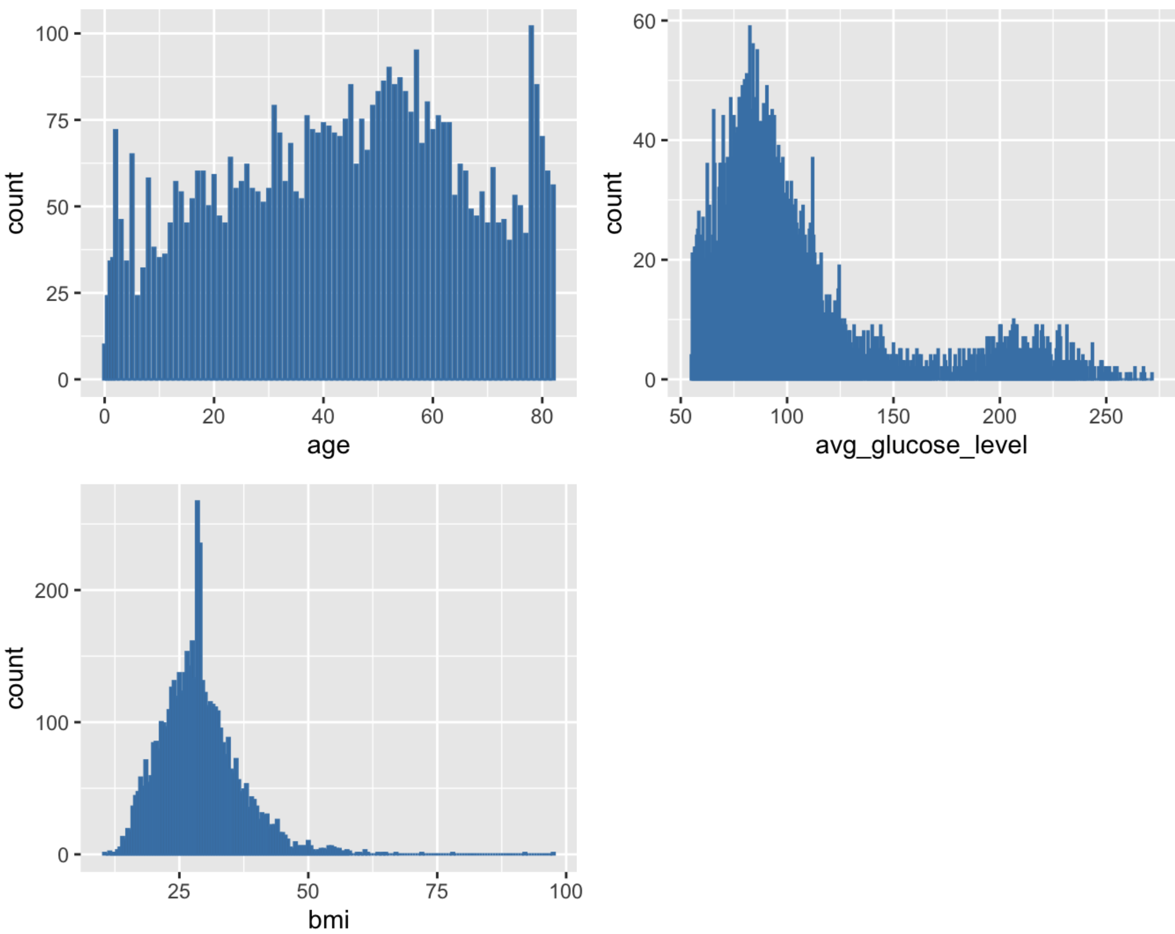





From the above barplots, we can conclude:

- Female and married people are the majority.
- Most of the people do not have heart disease and hypertension.
- We can also conclude that private workers and non-smoker are the majority.

## 3.2 For continuous variables, we have made use of the histogram.

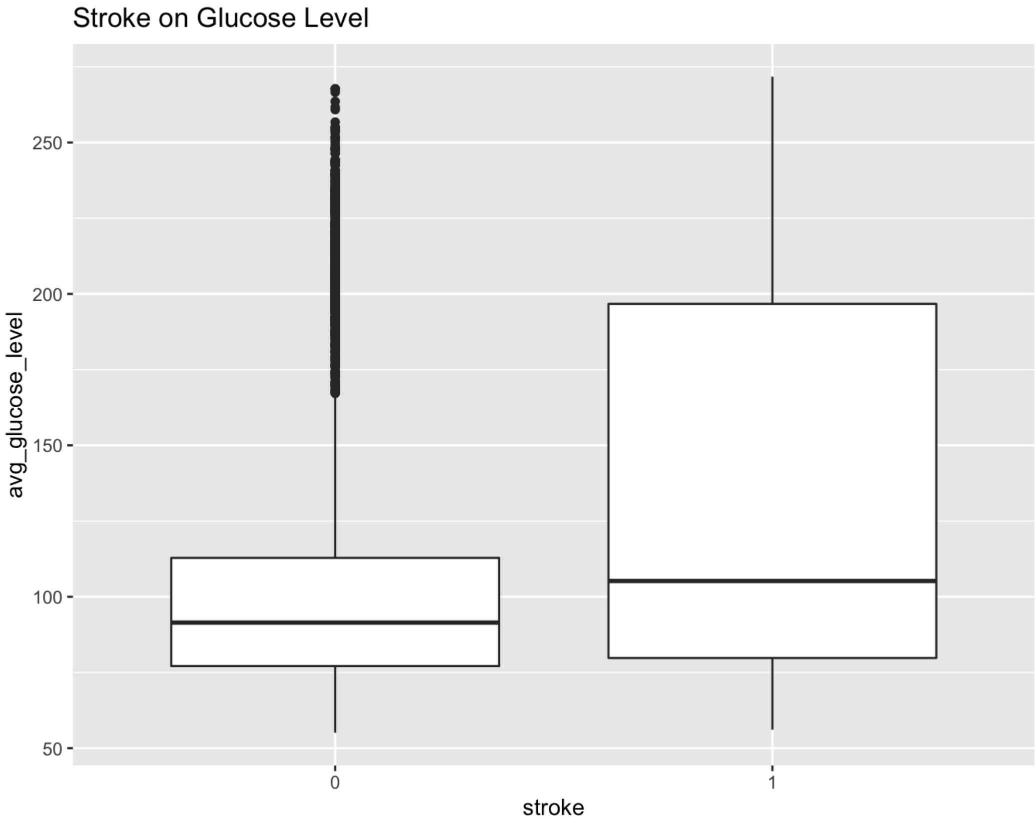



From the above plots, we can conclude:

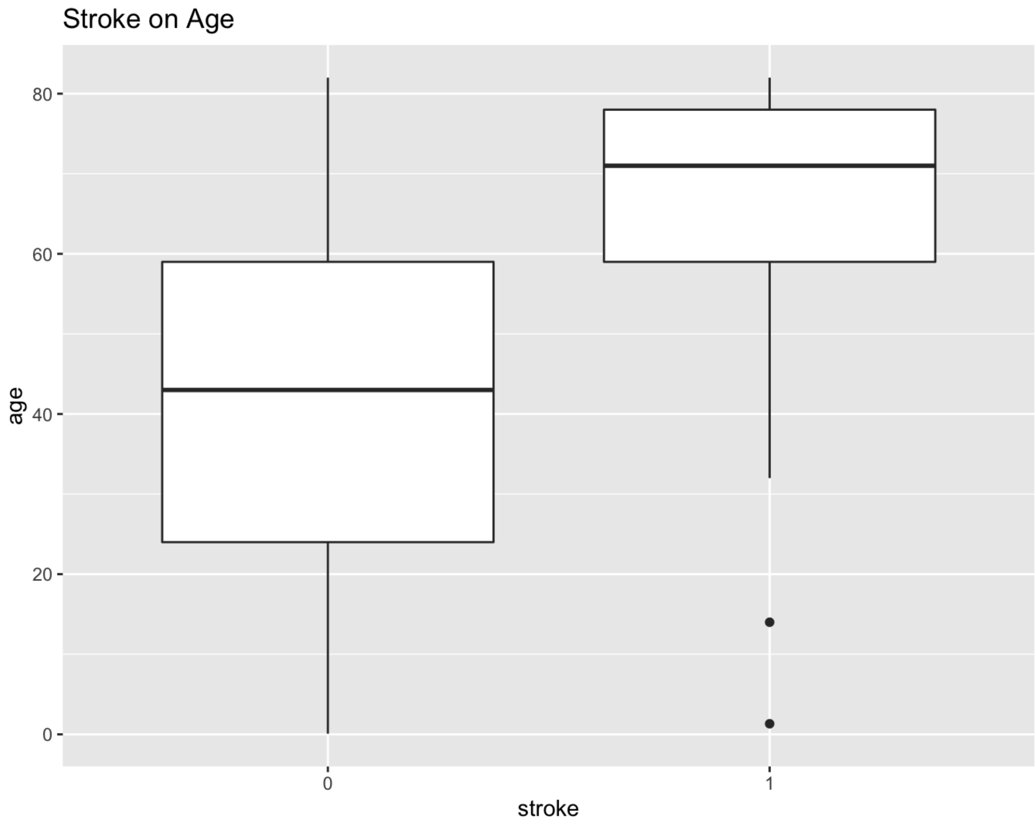
- age is slightly left skewed and glucose level and BMI are right skewed.
- One interesting fact we must observe is that the spike in the BMI plot which is the result of replacing null values with median, so we can get an important conclusion that median is the best estimator than mean as BMI is right skewed.

**We have successfully done EDA for individual features. Let us find the association between different variables.Our dataset's outcome variable is stroke, so let us find the association between stroke and some of the input features. Some of the plots which we have plotted is boxplot, mosaic plot, heat map.**

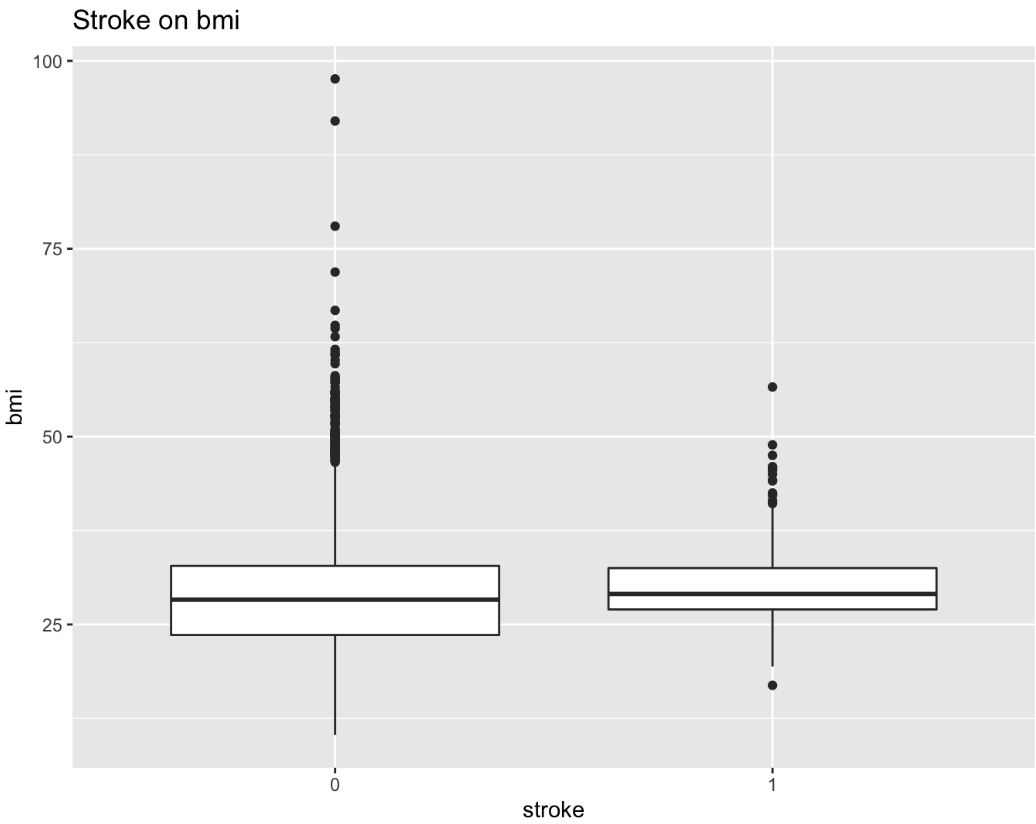
## 3.3 Boxplot



Stroke on Glucose Level
avg_glucose_level
250
200
150
100
50
0
1
stroke

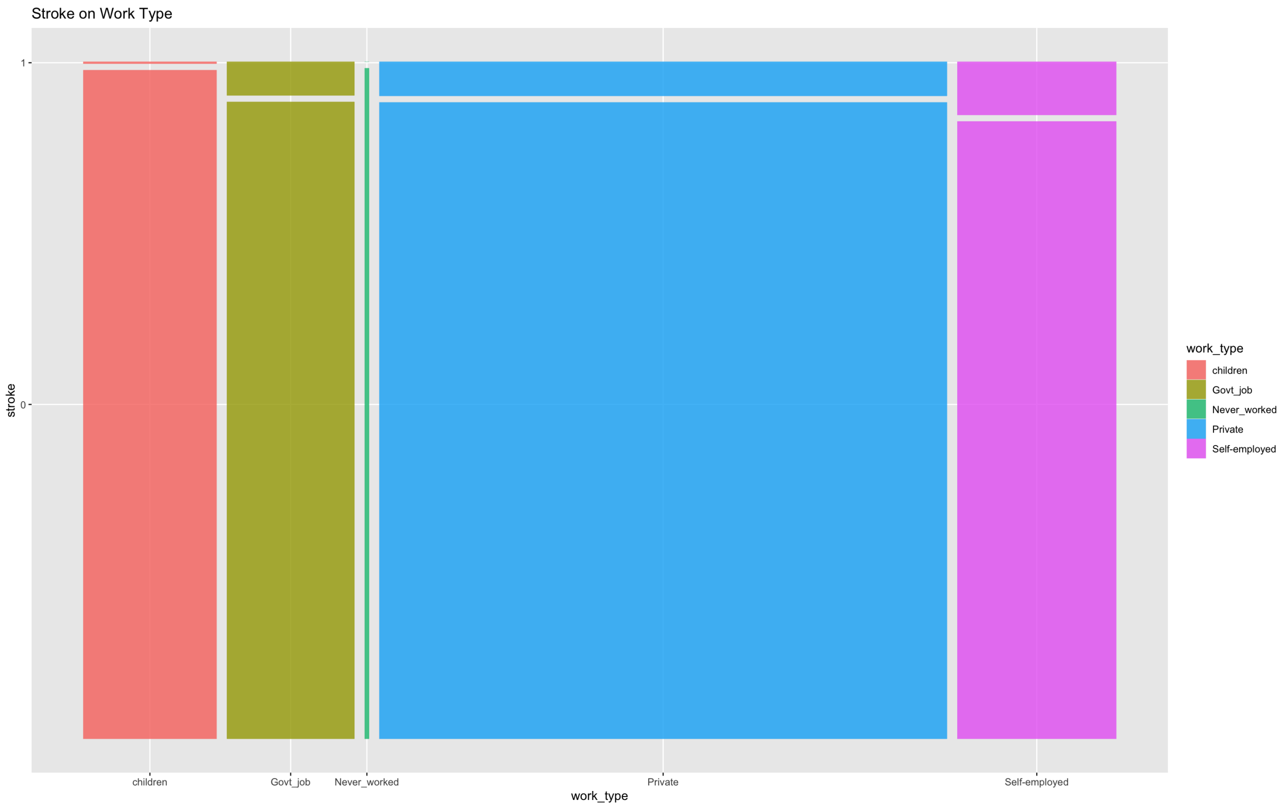
Stroke on Age
age
80
60
40
20
0
0
1
stroke

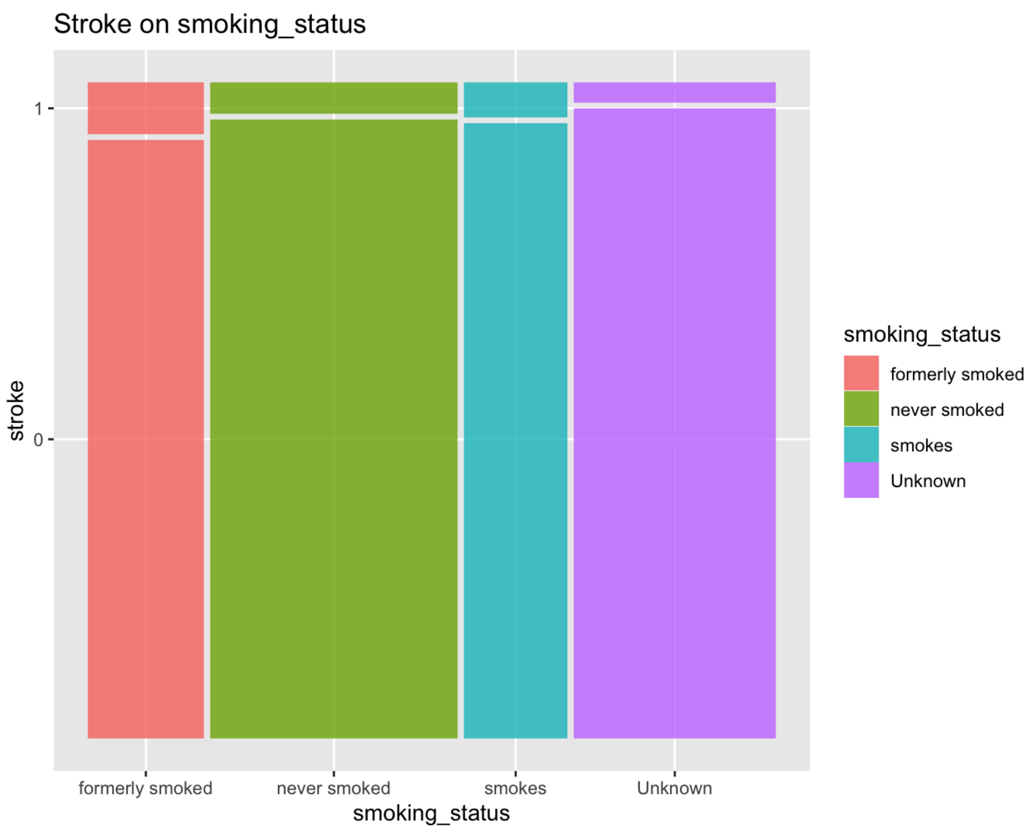


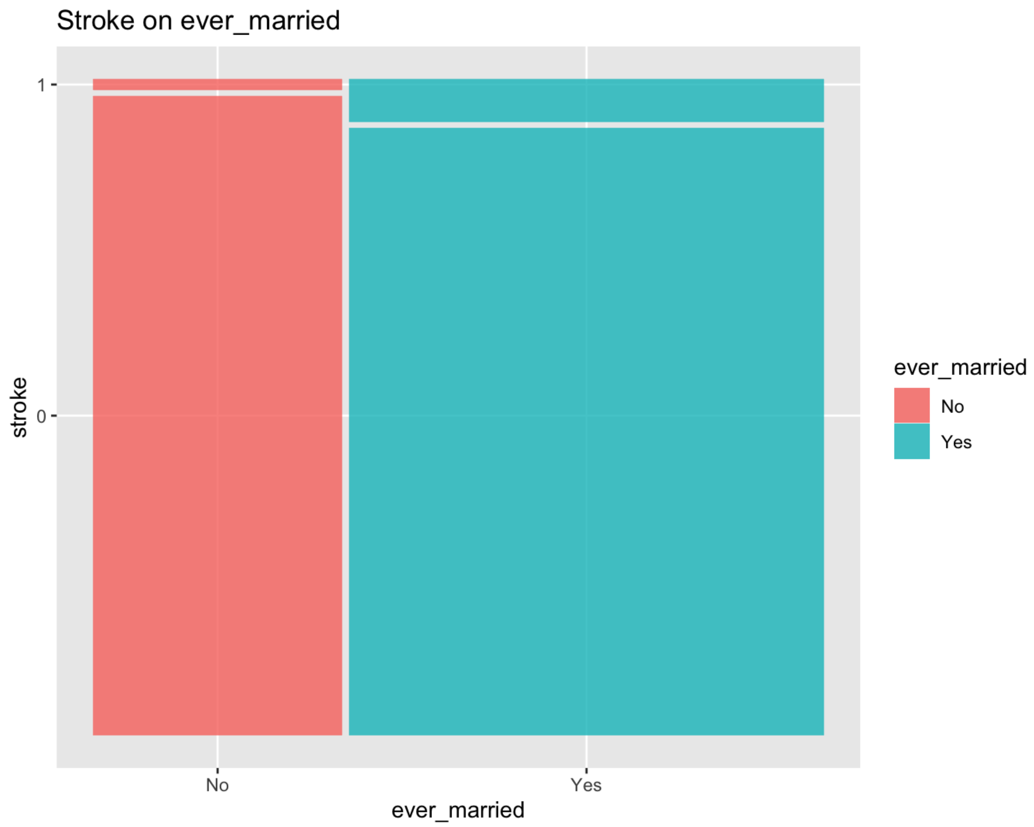

From boxplot,we can say:

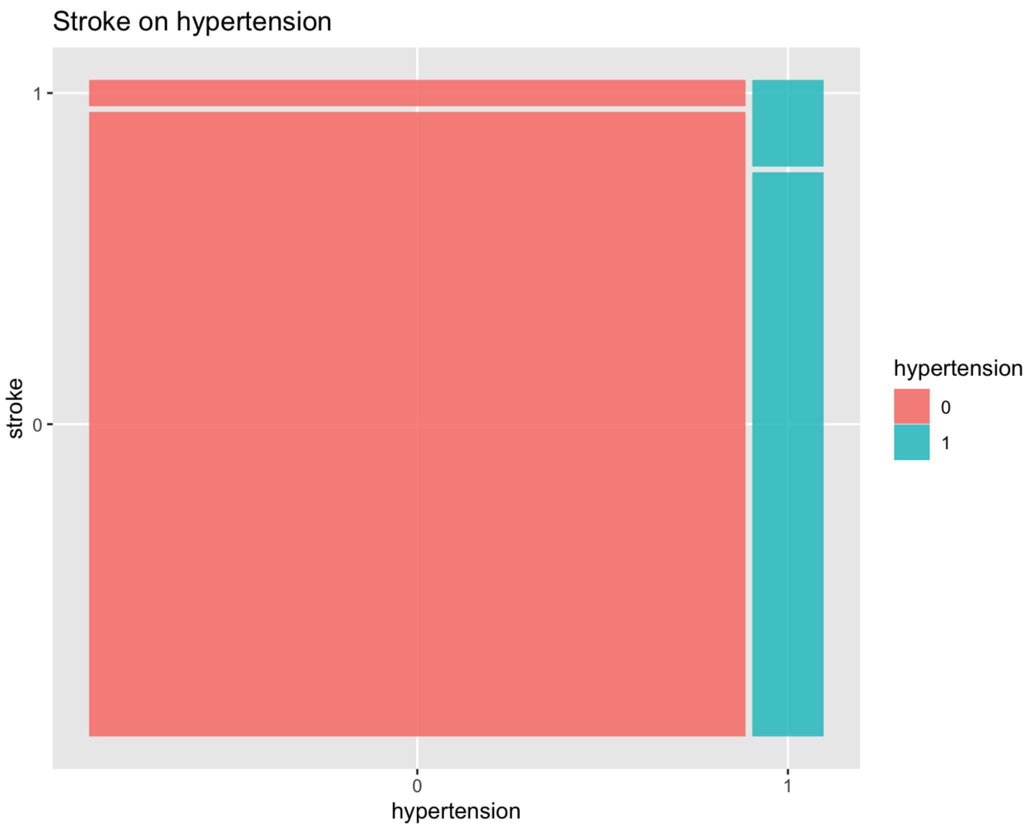
- Older people are more likely to get stroke.
- Similarly those who had stroke have higher glucose level and BMI, but it's not that significant.

## 3.4 Mosaic Plot





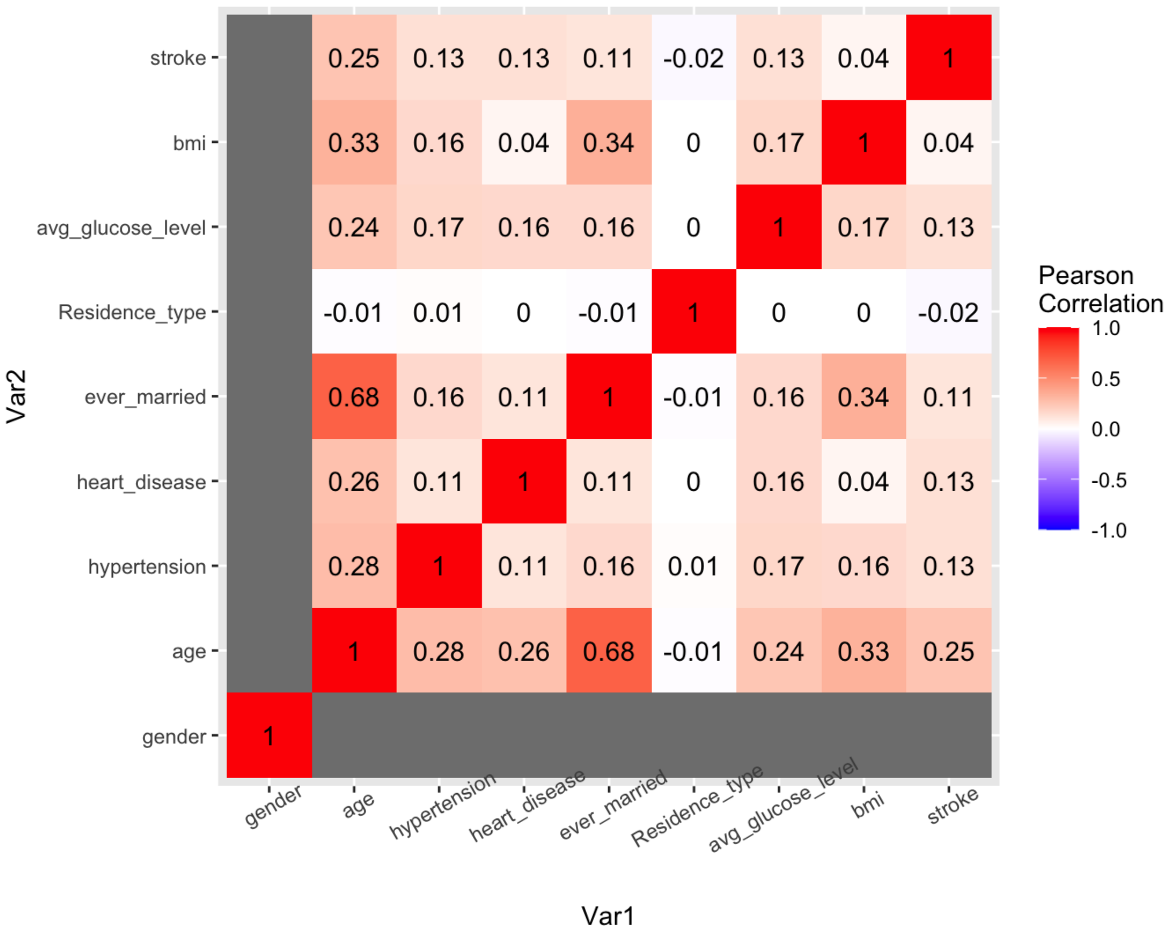




**We have used Mosaic plot to find the association between stroke and some discrete input features.**

Here we can conclude:

- Self-employed workers, those who have hypertension and those who are married are more likely to get stroke.
- Smoke seems to have little effect on the stroke.

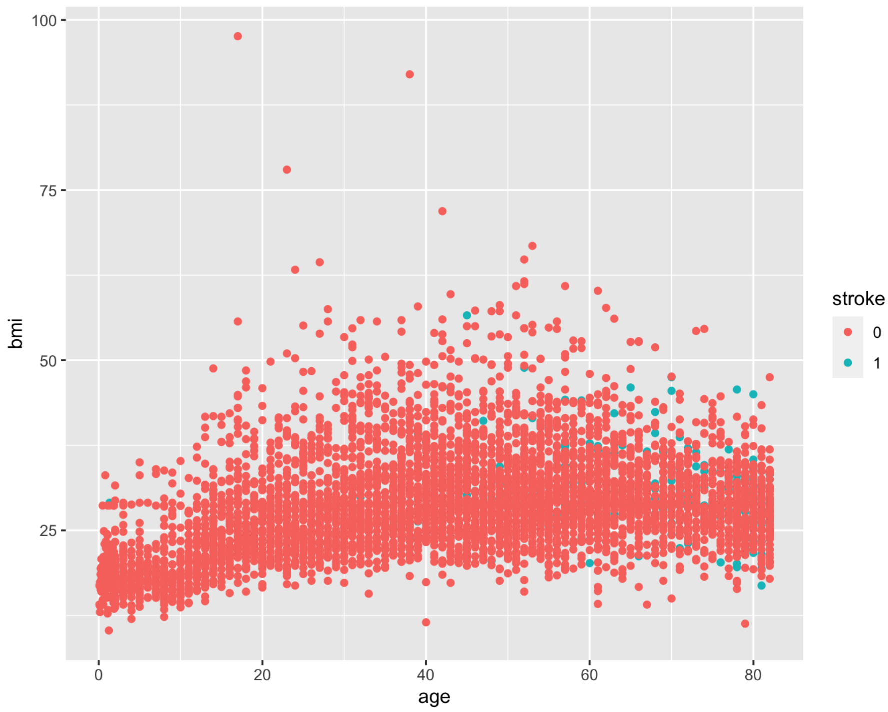
## 3.5 Heatmap

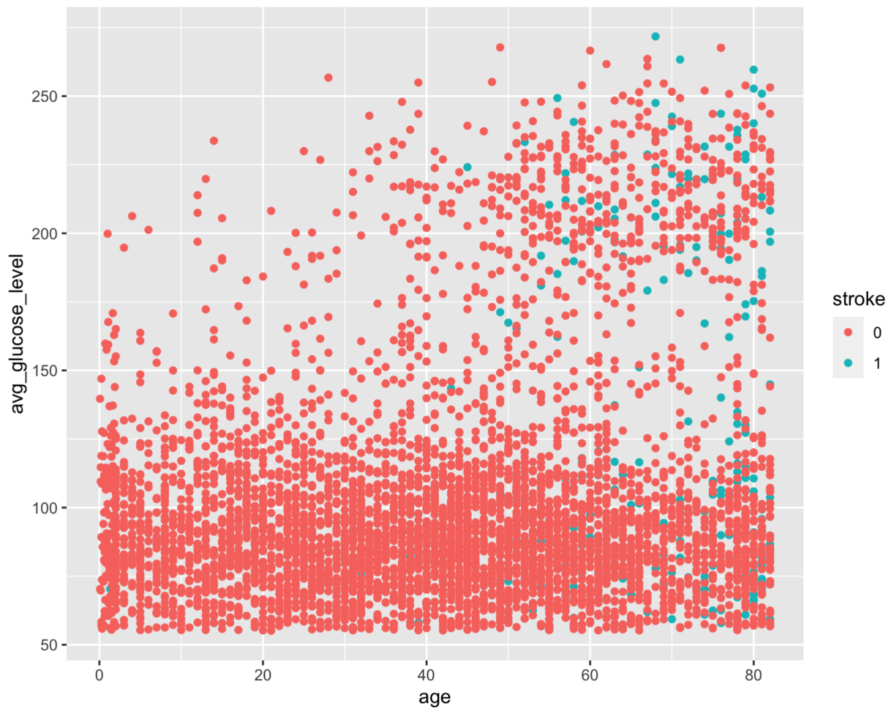



**We have made use of the heatmap to know the correlation among the input features.**

- In order to get the heatmap, we have converted qualitative variables into quantitative variables.
- The important factor that we must observe is that the correlation between ever married and age. Here age has highest correlation with stroke.

## 3.6 Scatter Plot




**From the Scatter plot, we can clearly say:**

- People with high BMI and older people are likely to get stroke.
- Also, it is clearly visible that the data is highly imabalanced.

# Chapter 4

# Model

## 4.1 Logistic Regression for Prediction

From the heatmap, we have seen that there is no certain feature that has a strong correlation with stroke. To determine how a certain input feature affect the outcome , we go for a regression model. Since stroke is a binary variable, we will use logistic regression.

### 4.1.1 Data Building and Model Training

As we have considered logistic regression, we need to build the training dataset.
To test the regression result:

- We split dataset into training set(70%) and testing set(30%).
- We have used stroke_dummy for regression.
- Next, we carry out the regression using Generalized linear model (GLM) and set the link to be logit. The model is trained on the training set.

```
glm(formula = stroke ~ ., family = binomial(link = "logit"),
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1231  -0.3076  -0.1542  -0.0899   3.4319

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.498e+00  8.179e-01  -7.945 1.94e-15 ***
age                          7.766e-02  7.066e-03  10.991  < 2e-16 ***
hypertension1                3.594e-01  2.015e-01   1.784 0.074411 .
heart_disease1               2.498e-01  2.287e-01   1.092 0.274774
ever_marriedYes             -2.732e-01  2.673e-01  -1.022 0.306806
Residence_typeUrban          6.757e-02  1.660e-01   0.407 0.684043
avg_glucose_level            5.001e-03  1.440e-03   3.472 0.000516 ***
bmi                          5.308e-03  1.353e-02   0.392 0.694887
gender_Male                 -2.946e-01  1.746e-01  -1.687 0.091522 .
gender_Other                -1.150e+01  2.400e+03  -0.005 0.996175
work_type_Govt_job          -1.492e+00  8.888e-01  -1.679 0.093248 .
work_type_Never_worked      -1.177e+01  5.212e+02  -0.023 0.981984
work_type_Private           -1.420e+00  8.673e-01  -1.637 0.101588
`work_type_Self-employed`   -1.639e+00  8.917e-01  -1.838 0.066071 .
`smoking_status_never smoked` -1.364e-01  2.145e-01  -0.636 0.524699
smoking_status_smokes        2.963e-01  2.574e-01   1.151 0.249606
smoking_status_Unknown      -6.431e-02  2.599e-01  -0.247 0.804578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1397.5  on 3577  degrees of freedom
Residual deviance: 1093.7  on 3561  degrees of freedom
AIC: 1127.7
```

From the regression result, we were able to conclude that:

1. The input features like age, hypertension and work_type self-employed are statistically significant in the regression results, with p-value smaller than 0.005. Among these age has the lowest p-value.

2. We have noticed in our EDA that age and hypertension are postively correlated with stroke and here we see the confirmation of it.

3. Self-employed is negatively correlated with stroke. So we can say that self-employed would reduce the risk of getting stroke. Probably self-employed people would better enjoy the life and have a healthy lifestyle.

### 4.1.2 Anova

Now we can run the anova() function on the model to see the deviance of the regression model.

ANOVA shows how features lower the original deviance to residual deviance.

```
> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: stroke

Terms added sequentially (first to last)


                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                           3577     1397.5
age                            1  272.385      3576     1125.1 < 2.2e-16 ***
hypertension                   1    5.987      3575     1119.2 0.0144103 *
heart_disease                  1    2.598      3574     1116.5 0.1069670
ever_married                   1    0.977      3573     1115.6 0.3229362
Residence_type                 1    0.382      3572     1115.2 0.5364661
avg_glucose_level              1   12.669      3571     1102.5 0.0003718 ***
bmi                            1    0.047      3570     1102.5 0.8284755
gender_Male                    1    2.346      3569     1100.1 0.1256050
gender_Other                   1    0.015      3568     1100.1 0.9011491
work_type_Govt_job             1    0.002      3567     1100.1 0.9628779
work_type_Never_worked         1    0.124      3566     1100.0 0.7242357
work_type_Private              1    0.735      3565     1099.2 0.3913461
`work_type_Self-employed`      1    2.360      3564     1096.9 0.1245109
`smoking_status_never smoked`  1    1.312      3563     1095.6 0.2520375
smoking_status_smokes          1    1.848      3562     1093.7 0.1740227
smoking_status_Unknown         1    0.061      3561     1093.7 0.8043320
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Besides three significant features shown in the regression result, avg_glucose_level also significantly lowers the deviance. Those features with a large p-value shows that even without these features, the model would explain more or less of the same of total variation.

### 4.1.3 Prediction on the testing dataset

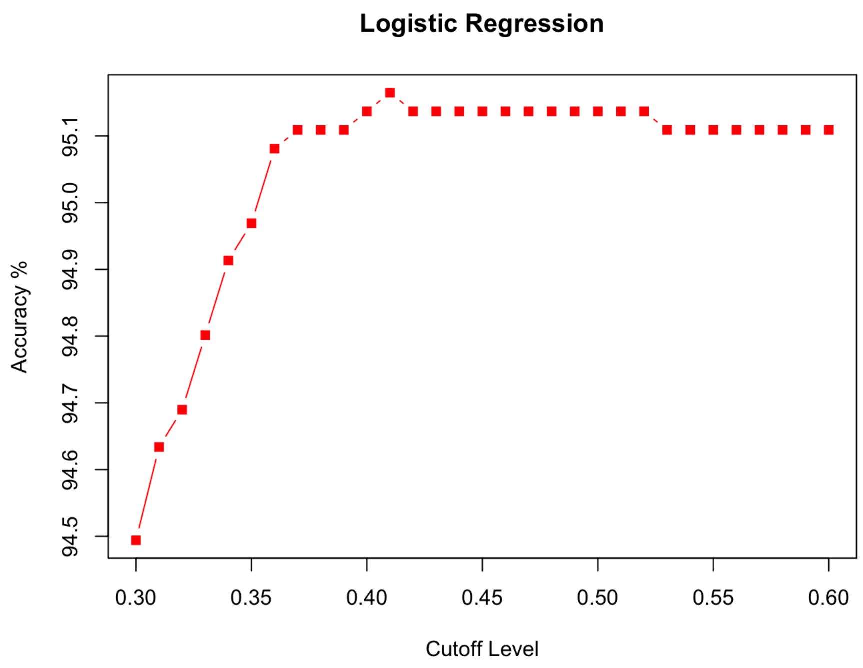
To visualize the prediction result, we could use the confusion Matrix function from package caret. We set the prediction result to be 0.5, indicating that if the predicted stroke is larger than 0.5, we believe that this person is likely to get stroke.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1458   74
         1    0    0

               Accuracy : 0.9517
                 95% CI : (0.9397, 0.9619)
    No Information Rate : 0.9517
    P-Value [Acc > NIR] : 0.5309

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.9517
         Neg Pred Value :    NaN
             Prevalence : 0.9517
         Detection Rate : 0.9517
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

### 4.1.4 Calculating the prediction accuracy

- From the confusion matrix report, we can see that the accuracy is 95%, which is relatively high.
- If we have a close look we can see that most of the prediction would just predict the outcome to be 0.
- Therefore we can come to most important conclusion that 'stroke' is imbalanced that almost all the outcome to be 0.
- Hence, it is a difficult task to differentiate between those who have stroke and those who don't.

One possible way to get better result is to change the threshold. Therefore we have plotted a relationship between threshold and accuracy.




We get a result which is not so pleasing to us. Therefore we can conclude that even by raise the accuracy by changing the cutoff threshold. So we have made use of random forest classifier to test the outcome.

## 4.2 Random Forest Classifier

```
 randomForest(formula = stroke ~ ., data = data, importance = TRUE,       proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 4.93%
Confusion matrix:
     0 1 class.error
0 4856 5 0.001028595
1  247 2 0.991967871
```

Even in the random forest classifier we get the accuracy of 95.07% and we can see that confusion matrix is similar to logistic regression.

**Therefore we can say that imbalance of the dataset would greatly affect the prediction result.**

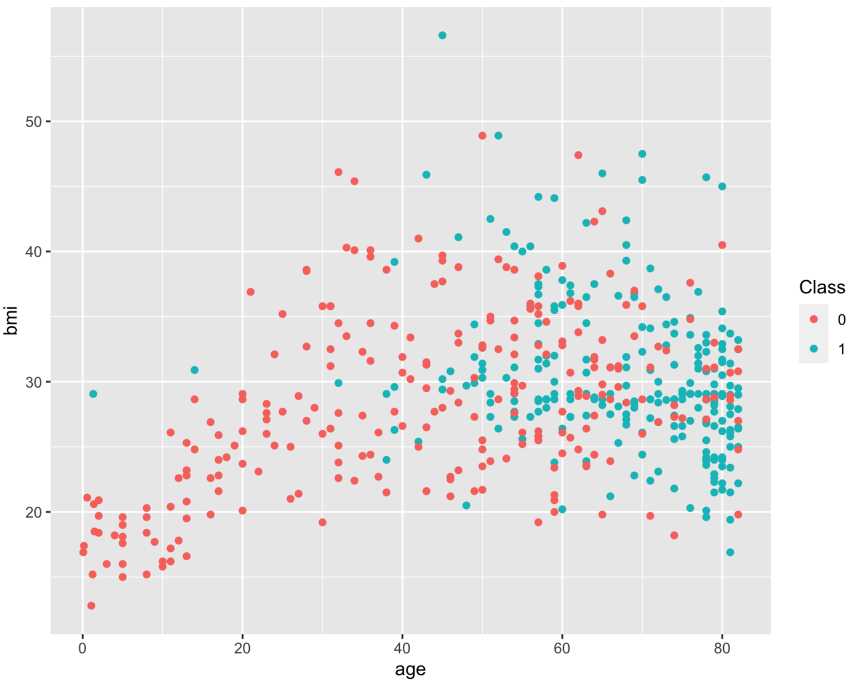
# Chapter 5

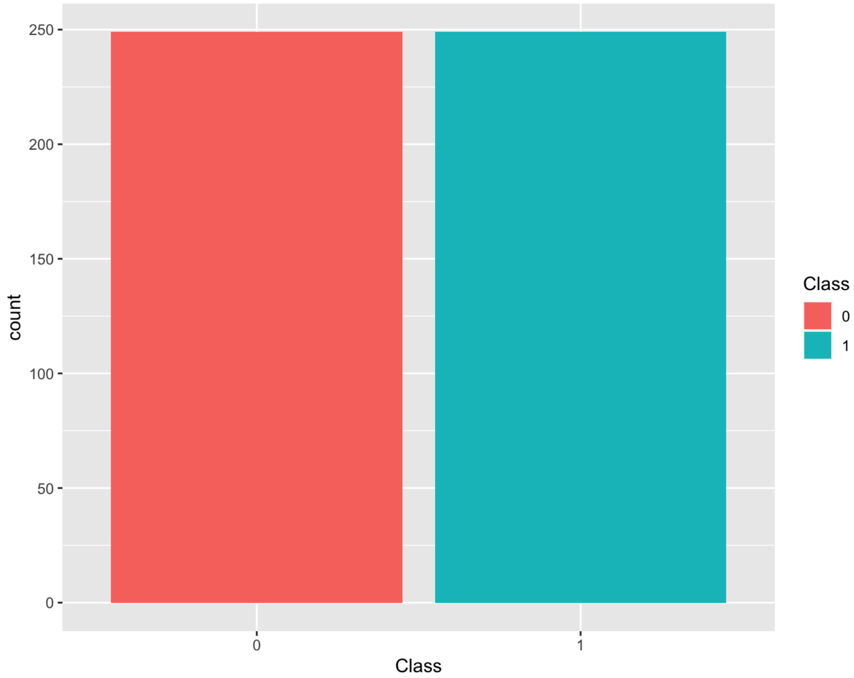
# Dealing with Imbalanced Data

Since our data is imbalanced, we use the Data Resampling techniques, to deal with it.

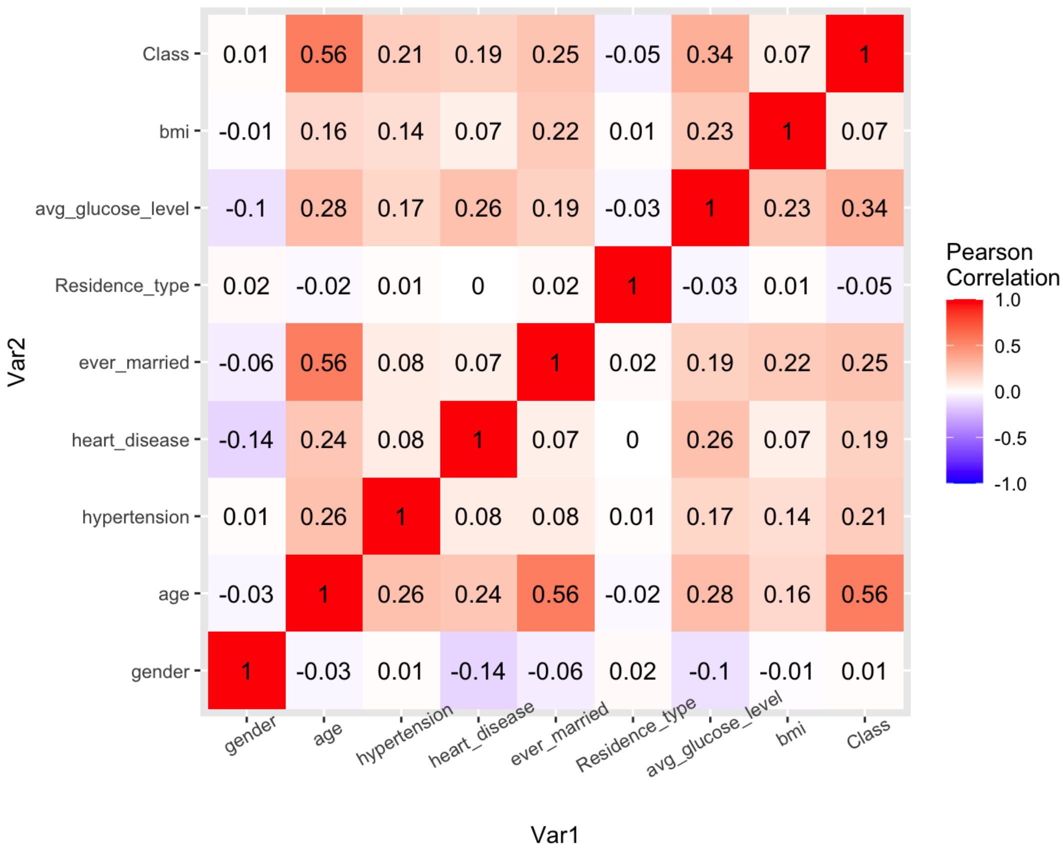
## 5.1 Under Sampling Technique

Under Sampling is the technique where some observations of the Majority Class are removed and then the observations are made to be equal.






We can see from the plots above that the Data is balanced now.




- It is visible that age has a high correlation with stroke.
- None of the features have a very strong correlation.

### 5.1.1 Fitting the Logistic Regression Model

```
Call:
glm(formula = Class ~ ., family = binomial(link = "logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4154  -0.7103   0.1213   0.7203   2.4613

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -4.520202   1.006242  -4.492 7.05e-06 ***
age                            0.084279   0.011739   7.179 7.00e-13 ***
hypertension1                  0.696541   0.383966   1.814   0.0697 .
heart_disease1                 0.259665   0.492721   0.527   0.5982
ever_marriedYes               -0.446955   0.498256  -0.897   0.3697
Residence_typeUrban           -0.114045   0.278539  -0.409   0.6822
avg_glucose_level              0.003185   0.002827   1.127   0.2599
bmi                            0.019524   0.024069   0.811   0.4173
gender_Male                    0.004941   0.290340   0.017   0.9864
work_type_Govt_job            -1.186160   1.101319  -1.077   0.2815
work_type_Never_worked       -12.659627 882.743712  -0.014   0.9886
work_type_Private             -1.082780   1.067923  -1.014   0.3106
`work_type_Self-employed`     -1.147267   1.146702  -1.000   0.3171
`smoking_status_never smoked` -0.406020   0.349542  -1.162   0.2454
smoking_status_smokes          0.530878   0.425975   1.246   0.2127
smoking_status_Unknown         0.751987   0.448806   1.676   0.0938 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 485.20  on 349  degrees of freedom
Residual deviance: 328.37  on 334  degrees of freedom
AIC: 360.37

Number of Fisher Scoring iterations: 13
```

From the regression result, we were able to conclude that:

1. The input features like age and hypertension are statistically significant in the regression results, with p-value smaller than 0.005. Among these age has the lowest p-value.

2. We have noticed in our EDA that age and hypertension are postively correlated with stroke and here we see the confirmation of it.

### 5.1.2 Anova

ANOVA shows how features lower the original deviance to residual deviance.

```
> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Class

Terms added sequentially (first to last)


                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                            349     485.20              
age                            1  130.329       348     354.87 < 2.2e-16 ***
hypertension                   1    0.510       347     354.36   0.47493    
heart_disease                  1    2.336       346     352.03   0.12639    
ever_married                   1    0.056       345     351.97   0.81221    
Residence_type                 1    0.084       344     351.89   0.77248    
avg_glucose_level              1   18.024       343     333.86 2.182e-05 ***
bmi                            1    0.879       342     332.98   0.34856    
gender_Male                    1    0.705       341     332.28   0.40102    
work_type_Govt_job             1    2.899       340     329.38   0.08864 .  
work_type_Never_worked         0    0.000       340     329.38              
work_type_Private              1    1.262       339     328.12   0.26124    
`work_type_Self-employed`      1    2.572       338     325.55   0.10880    
`smoking_status_never smoked`  1    0.128       337     325.42   0.72017    
smoking_status_smokes          1    0.709       336     324.71   0.39991    
smoking_status_Unknown         1    2.519       335     322.19   0.11251    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Besides two significant features shown in the regression result, avg_glucose_level also significantly lowers the deviance.

### 5.1.3 Prediction on the testing dataset

We used the confusion Matrix to visualize the Prediction Result. We set the prediction result to be 0.5, indicating that if the predicted stroke is larger than 0.5, we believe that this person is likely to get stroke.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 57 22
         1 17 52

               Accuracy : 0.7365
                 95% CI : (0.6578, 0.8054)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 3.732e-09

                  Kappa : 0.473

 Mcnemar's Test P-Value : 0.5218

            Sensitivity : 0.7703
            Specificity : 0.7027
         Pos Pred Value : 0.7215
         Neg Pred Value : 0.7536
             Prevalence : 0.5000
         Detection Rate : 0.3851
   Detection Prevalence : 0.5338
      Balanced Accuracy : 0.7365

       'Positive' Class : 0
```

### 5.1.4 Calculating the prediction accuracy

- From the confusion matrix report, we can see that the accuracy is 73.65%, which is okay.

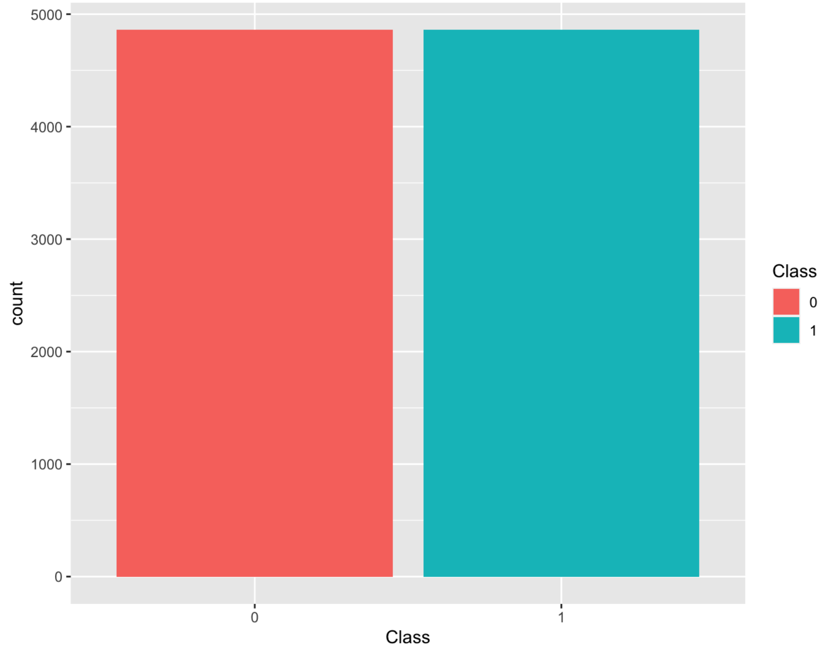
### 5.1.5 Random Forest Classifier

```
Call:
 randomForest(formula = Class ~ ., data = underData, importance = TRUE,      proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 25.3%
Confusion matrix:
    0   1 class.error
0 177  72   0.2891566
1  54 195   0.2168675
> 
```

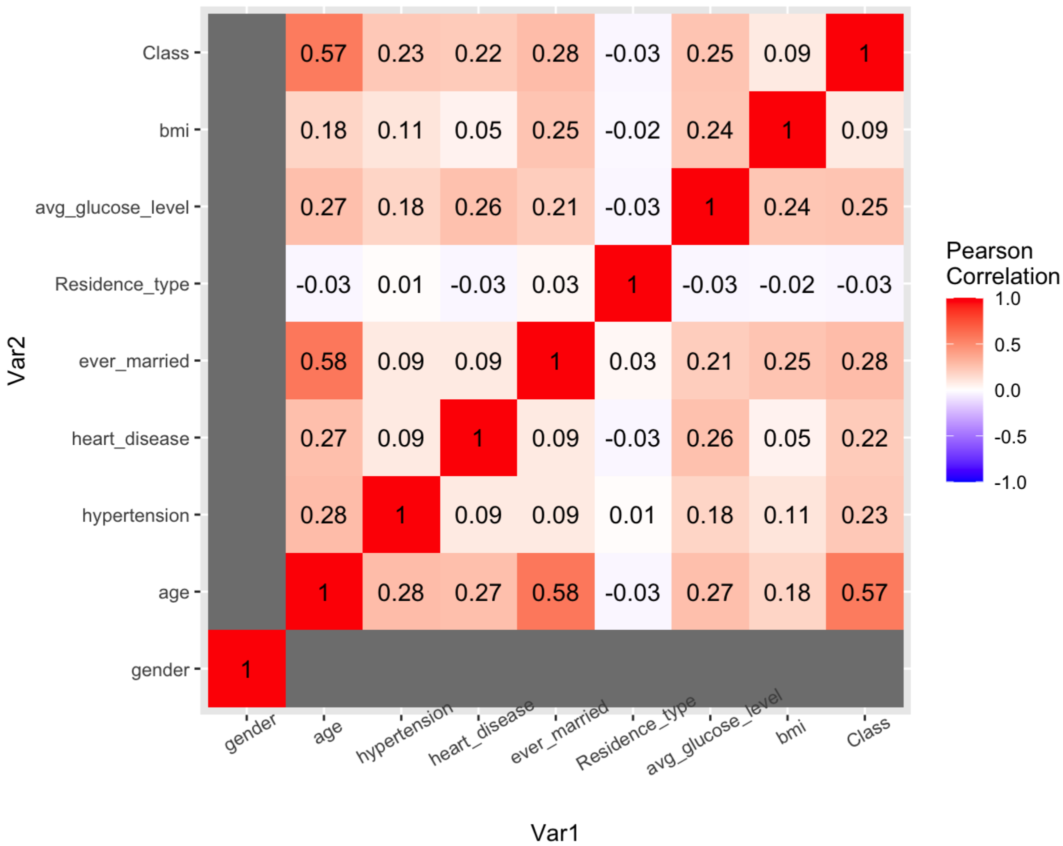
Even in the random forest classifier we get the accuracy of 74.7% and we can see that confusion matrix is similar to logistic regression.

## 5.2 Over Sampling Technique

Over Sampling is the technique where the number of observations of the Minority Class is increased so as to make them equal to that of the Majority Class.




We can see from the plot above that the Data is balanced now.



- It is visible that age has a high correlation with stroke.

- None of the features have a very strong correlation.

### 5.2.1 Fitting the Logistic Regression Model

```
Call:
glm(formula = Class ~ ., family = binomial(link = "logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4951  -0.6772   0.1301   0.7093   2.4369

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.476e+00  2.257e-01 -15.402  < 2e-16 ***
age                            8.161e-02  2.496e-03  32.699  < 2e-16 ***
hypertension1                  4.899e-01  8.338e-02   5.875 4.23e-09 ***
heart_disease1                 1.267e-01  1.024e-01   1.238   0.2159
ever_marriedYes               -9.162e-02  1.011e-01  -0.906   0.3648
Residence_typeUrban            5.596e-02  6.181e-02   0.905   0.3653
avg_glucose_level              4.424e-03  6.068e-04   7.291 3.09e-13 ***
bmi                            5.859e-03  5.091e-03   1.151   0.2498
gender_Male                   -1.246e-01  6.378e-02  -1.954   0.0507 .
gender_Other                          NA         NA      NA       NA
work_type_Govt_job            -1.993e+00  2.425e-01  -8.219  < 2e-16 ***
work_type_Never_worked        -1.171e+01  1.475e+02  -0.079   0.9368
work_type_Private             -1.757e+00  2.317e-01  -7.583 3.37e-14 ***
`work_type_Self-employed`     -1.871e+00  2.469e-01  -7.578 3.51e-14 ***
`smoking_status_never smoked` -4.101e-01  8.321e-02  -4.928 8.29e-07 ***
smoking_status_smokes          2.506e-01  9.797e-02   2.558   0.0105 *
smoking_status_Unknown        -8.559e-02  9.519e-02  -0.899   0.3686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9435.1  on 6805  degrees of freedom
Residual deviance: 6480.5  on 6790  degrees of freedom
AIC: 6512.5

Number of Fisher Scoring iterations: 12
```

From the regression result, we were able to conclude that:

1. The input features like age, hypertension, avg_glucose_level and some other factors are statistically significant in the regression results, with p-value smaller than 0.005.

2. We have noticed in our EDA that age and hypertension are postively correlated with stroke and here we see the confirmation of it.

### 5.2.2 Anova

ANOVA shows how features lower the original deviance to residual deviance.

```
> anova(model, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Class

Terms added sequentially (first to last)


                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                           6805     9435.1
age                            1  2712.35      6804     6722.8 < 2.2e-16 ***
hypertension                   1    39.29      6803     6683.5 3.661e-10 ***
heart_disease                  1    15.65      6802     6667.8 7.629e-05 ***
ever_married                   1     1.03      6801     6666.8 0.3106572
Residence_type                 1     0.98      6800     6665.8 0.3212047
avg_glucose_level              1    63.20      6799     6602.6 1.867e-15 ***
bmi                            1     0.18      6798     6602.4 0.6742989
gender_Male                    1     0.28      6797     6602.2 0.5955169
gender_Other                   0     0.00      6797     6602.2
work_type_Govt_job             1     8.83      6796     6593.3 0.0029648 **
work_type_Never_worked         1     1.37      6795     6592.0 0.2422014
work_type_Private              1     0.15      6794     6591.8 0.6982888
`work_type_Self-employed`      1    52.54      6793     6539.3 4.208e-13 ***
`smoking_status_never smoked`  1    46.91      6792     6492.4 7.415e-12 ***
smoking_status_smokes          1    11.03      6791     6481.3 0.0008951 ***
smoking_status_Unknown         1     0.81      6790     6480.5 0.3686756
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Anova provides information about variability within the Regression Model, hence proving the significance.

### 5.2.3 Prediction on the testing dataset

Using the confusion Matrix to visualize the Prediction Result again.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1070  294
         1  388 1164

               Accuracy : 0.7661
                 95% CI : (0.7503, 0.7814)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5322

 Mcnemar's Test P-Value : 0.0003692

            Sensitivity : 0.7339
            Specificity : 0.7984
         Pos Pred Value : 0.7845
         Neg Pred Value : 0.7500
             Prevalence : 0.5000
         Detection Rate : 0.3669
   Detection Prevalence : 0.4678
      Balanced Accuracy : 0.7661

       'Positive' Class : 0
```

### 5.2.4 Calculating the prediction accuracy

From the confusion matrix report, we can see that the accuracy is 76.6%, which is okay.

### 5.2.5 Random Forest Classifier

```
Call:
 randomForest(formula = Class ~ ., data = overData, importance = TRUE,      proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.89%
Confusion matrix:
     0    1 class.error
0 4774   87  0.01789755
1    0 4861  0.00000000
```

In the random forest classifier we get the accuracy of 99.11% and we can see that confusion matrix is not similar to logistic regression here.

# Chapter 6

# Conclusion

- In this project, we used logistic regression to discover the relationship between stroke and other input features.
- We get the conclusion that age, hypertension and work_type_self-employed would affect the possibility of getting stroke.
- We also use logistic regression and random forest to build a prediction model for stroke. Both model reach the accuracy of 95%, but the imbalance of dataset has limited the accuracy level.
- Using the Under Sampling method, we saw that the accuracy reduced to 75%.
- Using the Over Sampling method, we saw that the accuracy increase to 99.08% using Random Forest Classifier.
- It appears that for this particular Dataset, Over Sampling and Random Forest are among the best of the options that we have tried here

# Chapter 7

# Bibliography

Following links were used for reference purpose:


- Dealing with Imbalanced Data
- https://towardsdatascience.com/understanding-random-forest-58381e0602d2
- https://www.statology.org/logistic-regression-in-r/