### **1. Objective Description:**

### **Predictive Modeling:** Develop a predictive model to identify individuals at high risk of experiencing a stroke based on demographic, lifestyle, and health-related factors.

### **2. Dataset Selection:**

### This [Kaggle dataset](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

### **3. Literature Review:**

Each year, 15 million individuals globally experience a stroke. Among them, 5 million die, while an additional 5 million are left with enduring disabilities, imposing significant challenges on their families. [(World Health Organization)](https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html).

Controllable factors such as high blood pressure, obesity, physical inactivity, poor diet, and smoking contribute to approximately 82-90% of all strokes. However, other uncontrollable factors also influence the likelihood of experiencing a stroke. Strokes can occur at any age, but they are more common in infants under the age of 1 and in adults as they age. Environmental factors like residing in areas with high air pollution, overworking, and prolonged periods of isolation from family and friends can elevate the risk of stroke. Additionally, demographic factors play a role: men have a higher likelihood of experiencing a stroke at younger ages, whereas women, who generally live longer, have a higher lifetime risk. Women who use birth control pills or undergo hormone replacement therapy face an increased risk [(National Heart, Lung, and Blood Institute)](https://www.nhlbi.nih.gov/health/stroke/causes).

Considering this information, the most significant variables from the dataset that can help with predicting stroke are likely hypertension (high blood pressure), bmi, and smoking\_status, as they are supposed to account for 82-90% of strokes.

### **4. Dependent Variable:**

Dependent variable is **“stroke”** — whether a person had a stroke (1) or not (0).

**5. Independent Variables:**

All variables except id (1) will be used to see how they affect having a stroke.

1) id: unique identifier  
2) gender: "Male", "Female" or "Other"  
3) age: age of the patient  
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension  
5) heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease  
6) ever\_married: "No" or "Yes"  
7) work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"  
8) Residence\_type: "Rural" or "Urban"  
9) avg\_glucose\_level: average glucose level in blood  
10) bmi: body mass index  
11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*

### **6. Exploratory Data Analysis (EDA):**

* Clean the data

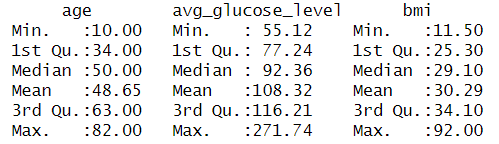
In the “bmi” column there are “N/A” values, which were removed. The number of observations decreased from 5110 to 4909.

In the “smoking\_status” column were unknown values. The number of observations decreased from 4909 to 3426. It is quite a significant loss of data, but in my opinion it is still the best option. I tried fitting models without removing “Unknown” values and even though the fit is better, the “Unknown” values affect the dependent variable significantly, which is illogical.

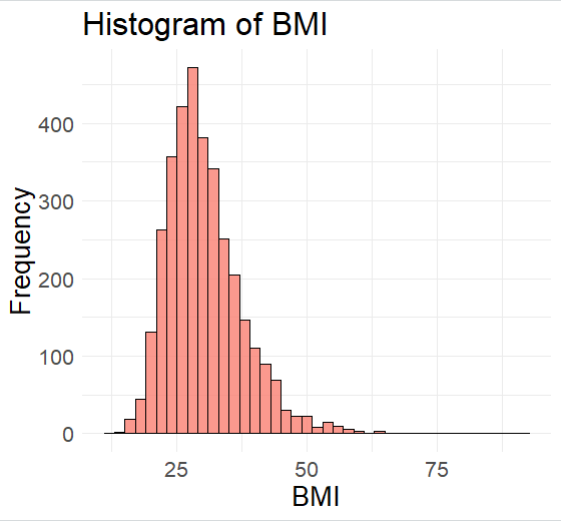
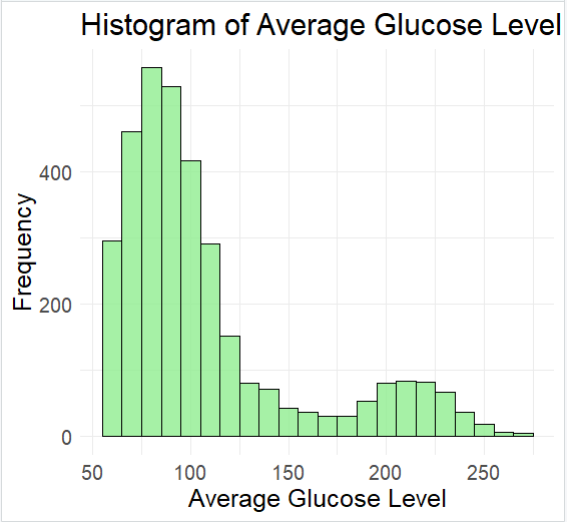
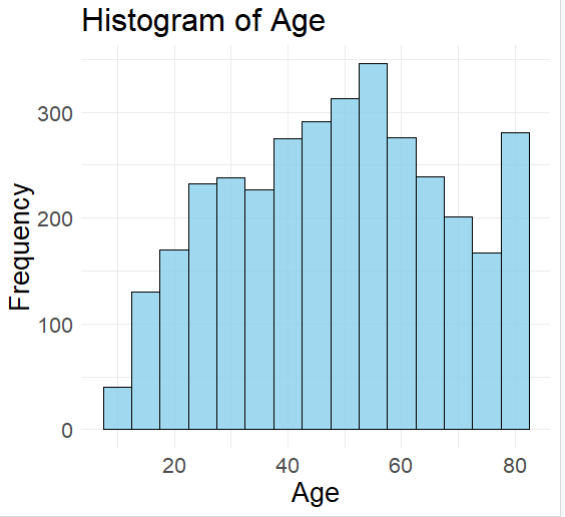
In the “age” column there are float numbers, they were rounded to receive integers.

Values from categorical columns were encoded.

* Descriptive statistics of numeric variables (“age”, “avg\_glucose\_level”, “bmi”)



For “age” and “bmi” columns, mean and median are close, so the data is likely normally distributed. For the “avg\_glucose\_level” column, median is lower than mean, so the data is likely skewed to the right. We will build histograms to check this.



**Age:**

* The cleaned dataset ranges from 10 to 82 years old.
* The interquartile range for age is 29 years, so the middle 50% of the data falls within this range.
* The median age is 50, so half of the individuals in the dataset are younger than 50 and half are older.

**Average Glucose Level:**

* The average glucose level ranges from 55.12 to 271.74 mg/dL.
* The interquartile range for average glucose level is 38.97 mg/dL, representing the middle 50% of the data.
* The median glucose level is 92.36 mg/dL, indicating that half of the individuals have glucose levels below this value and half have levels above.

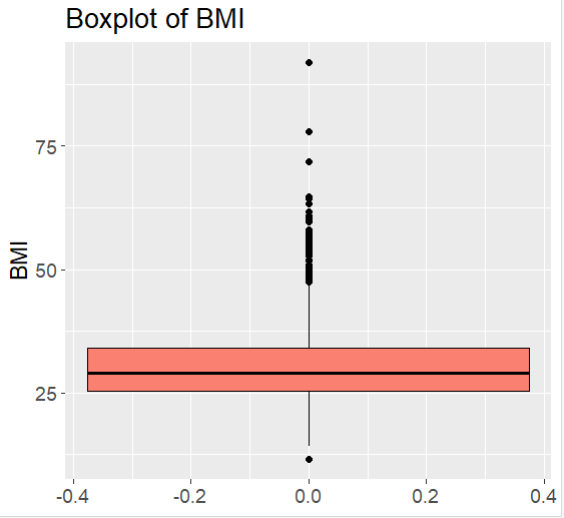
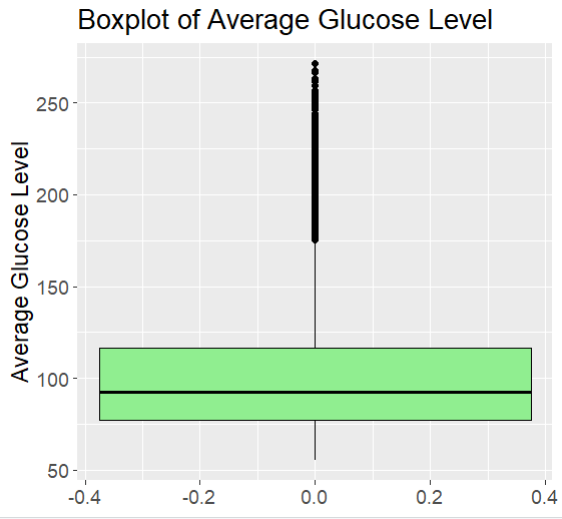
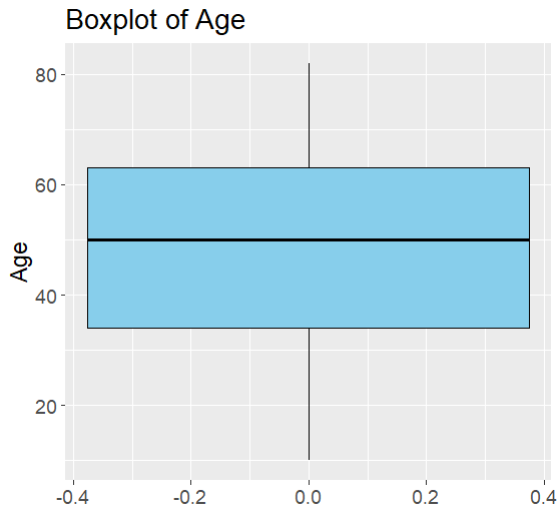
Normal level of glucose depends on the method and conditions of taking the blood test. It is unknown how exactly these tests were taken. Let’s consider that it was a [random blood sugar test](https://www.cdc.gov/diabetes/basics/getting-tested.html). Then, all results above 200 mg/dL are higher than normal and suggest that the patient has diabetes. I calculated that around 10% of people observed have very high glucose levels and likely diabetes. Low glucose level is less than 70 mg/dL. In our case it is around 15% of people observed.

**BMI (Body Mass Index):**

* BMI values range from 11.50 to 92.00.
* The interquartile range for BMI is 8.8, showing the spread of BMI values for the middle 50% of the dataset.
* The median BMI is 29.10, indicating that half of the individuals have a BMI below this value and half have a BMI above.

According to the US [Center for Disease Control and Prevention](https://www.cdc.gov/healthyweight/assessing/index.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obese%20range.) , normal BMI falls within the range 18.5 - 24.9. Median of 29.10 suggests that more than half of the people observed are overweight or obese. I calculated that 76.7% of the observations in the cleaned dataset are higher than 24.9 bmi. So, there are many overweight and obese people in the dataset, but it is a close to normal representation of for example US population where 69% of people are overweight or obese.

Presence of outliers was checked with boxplots.



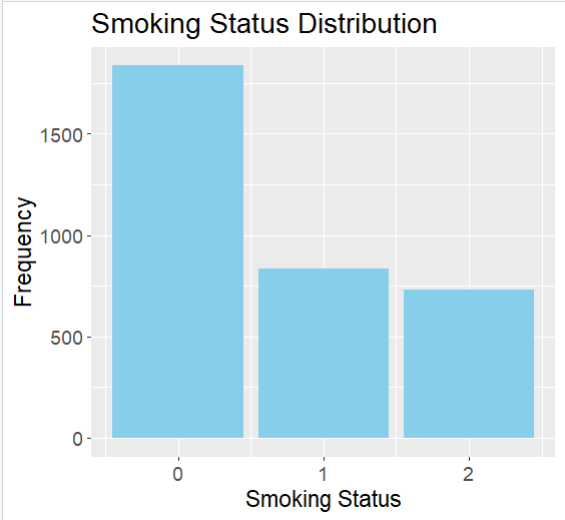
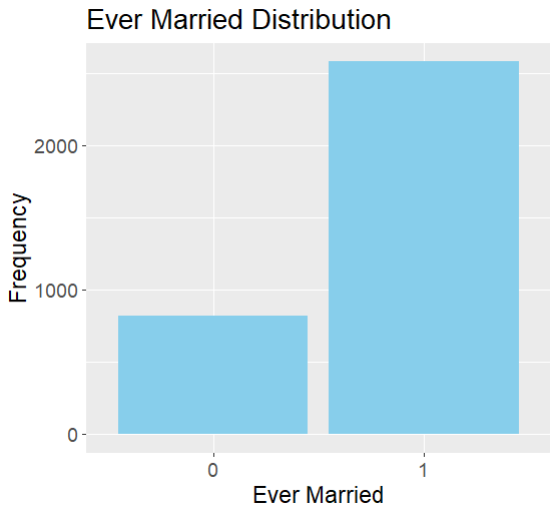
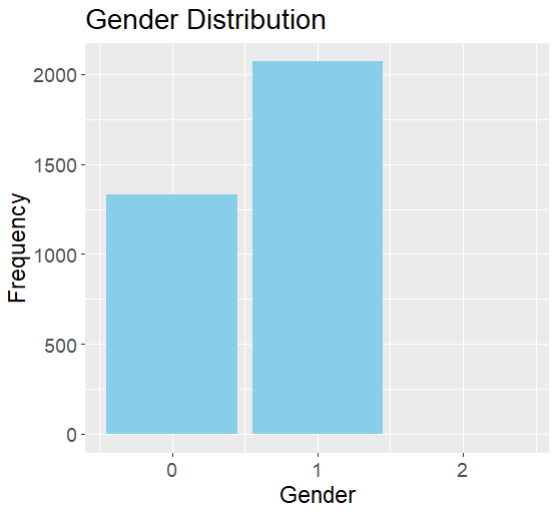
Age is normally distributed, there are no outliers. Both average glucose level and BMI are skewed to the right and have outliers.

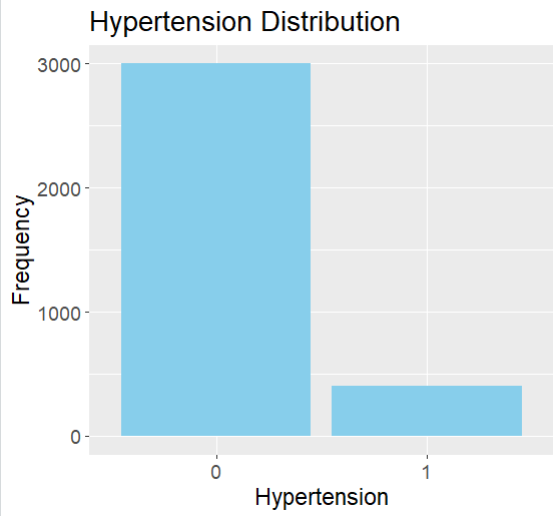
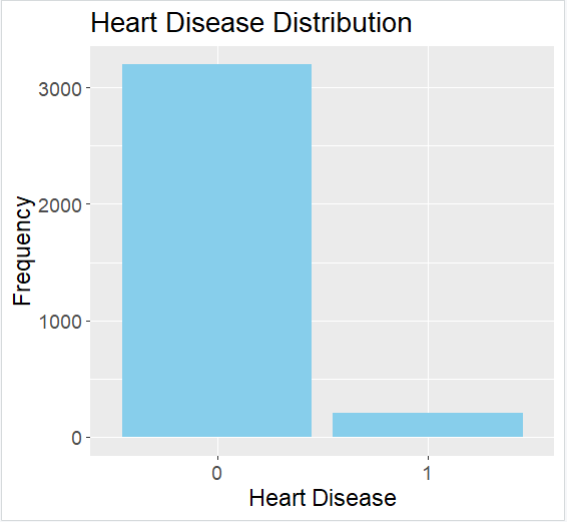
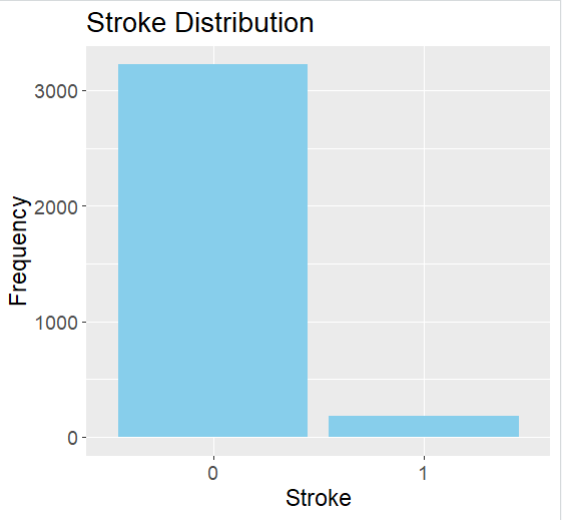
There are questions about the BMI outliers. Having BMI bigger than 50 means being morbidly obese. And having a BMI higher than 70 or 80 is very uncommon. People with high BMI are likely to have diabetes (high glucose level), heart disease, and are more likely to have strokes. In this dataset, however, almost all people with extremely high BMI do not have these problems and almost none of them had a stroke. So, the observations with extremely high BMI could be data entry errors and considering they contradict expectations based on existing knowledge, they can hinder the effectiveness of the model. It is still possible that not all of these outliers are data entry errors, so I believe the best solution is to remove values above 55 BMI leaving some outliers intact to represent a possible group of people with morbid obesity.

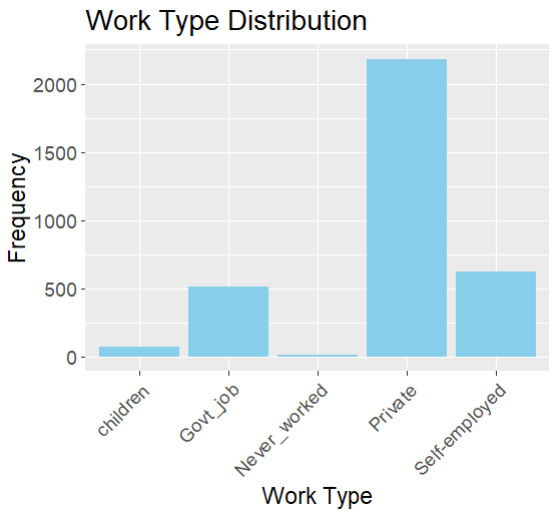
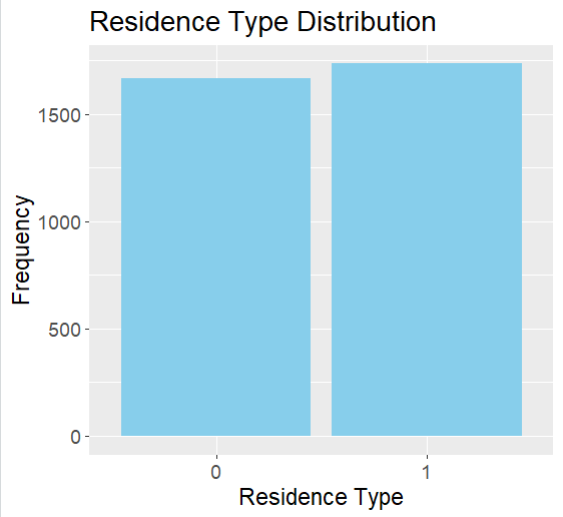
Outliers of the average glucose column can represent people with diabetes, so we will leave them intact.

In general, the numeric varables seem to be a good representation of various groups of people with different characteristics.

Now, categorical variables.







The biggest problem is imbalance in 3 very important columns: “stroke”, “heart\_disease”, and “hypertension”. The data will be balanced with the use of the ROSE library by oversampling and undersampling present data. I will fit the model on both original data and on the balanced data to see which one works best.

### 

### 

### **7. Model Selection:**

**GLM with binomial family** will be used because the dependent variable (stroke) is binary where each observation can fall into one of two categories. I will fit 3 models. One full model for imbalanced data, one full and one reduced (with the best predictors) for balanced data. Models will be compared with ANCOVA and the best one will be selected.

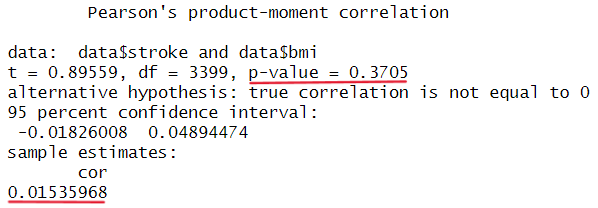
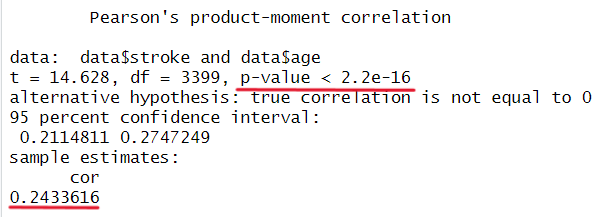
Also, a **Random Forest model** will be used because it can capture complex relationships and is robust against overfitting, so can potentially help with imbalance data.

### **8. Predictor Selection:**

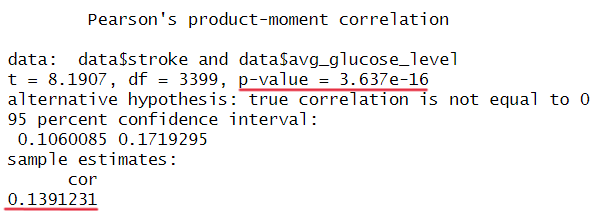
For the reduced model best predictors will be chosen based on analyses and on the result of the full model (p values).

To compare correlation of numeric variables and dependent categorical variable stroke, Point-Biserial Correlation test will be used because it is good for measuring the strength and direction of the relationship between one continuous variable and one dichotomous variable.

Age&Stroke BMI&Stroke



Average\_glucose&Stroke



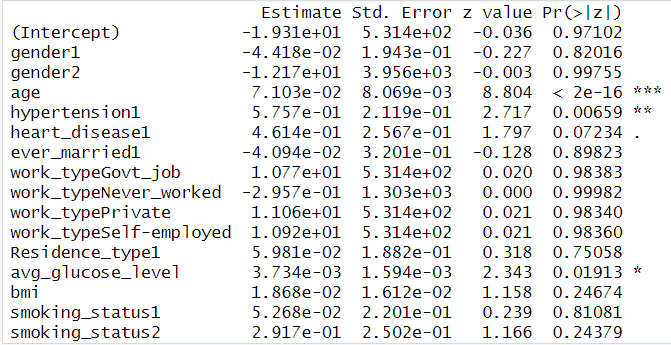
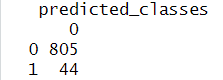
All numeric values have positive relationships with the stroke value (0.24, 0.015, 0.13), but only **age** and **average glucose** have statistically significant correlation (p-value lower than 0.05). So, those 2 values can be used for the reduced model.

For checking correlation of categorical variables and the dependent categorical variable, Chi-square test will be used. The following values have a p-value less than 0.05, are statistically significant and should be used in a reduced model: **hypertension** (p-value < 2.2e-16), **heart disease** (p = 2.056e-15), **ever married** (p = 4.444e-05), **work type** (p = 0.01173), and **smoking status** (p < 2.2e-16).

So, before fitting the full model those are variables that show the most correlation with stroke: **age, average glucose level, hypertension, heart disease, ever married, work type, smoking status.**

**9. Model Fitting and Interpretation:**

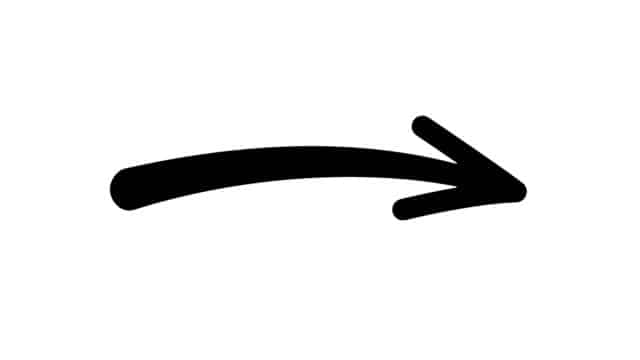
First, a full model was fit on the original data.

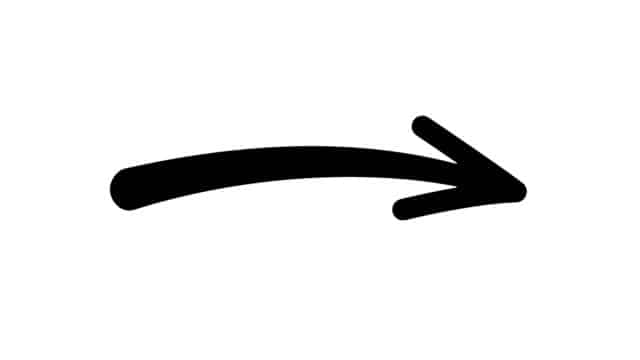
****

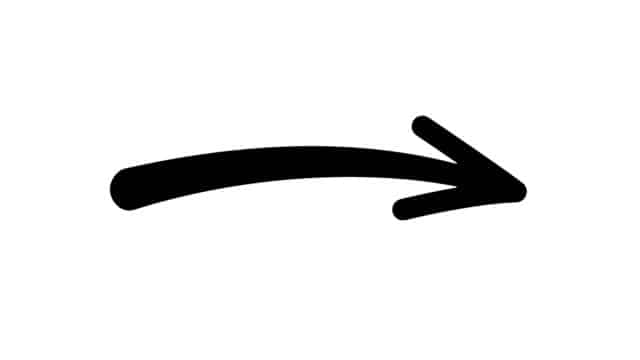


The accuracy of the model is suspiciously high (~94%), when we look at the confusion matrix, we see that the model predicted all the values as 0. Considering an imbalance in the model (most stroke observations are 0), high accuracy is understandable. We will need to balance the data to try fixing this problem.

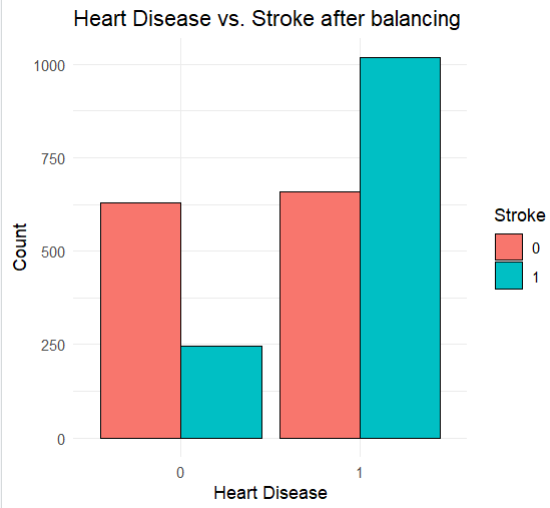
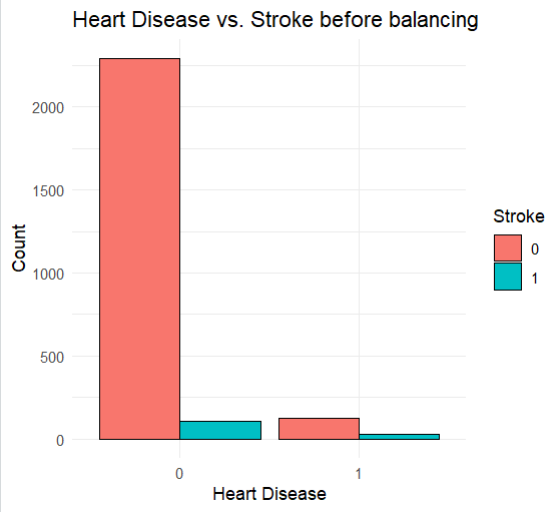
The data was balanced with ROSE.

 Stroke

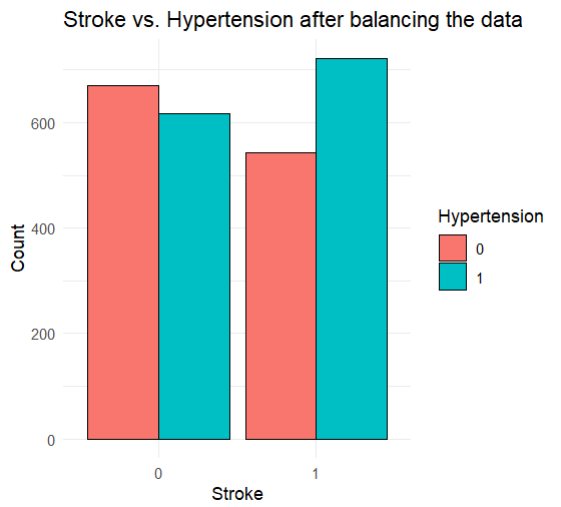
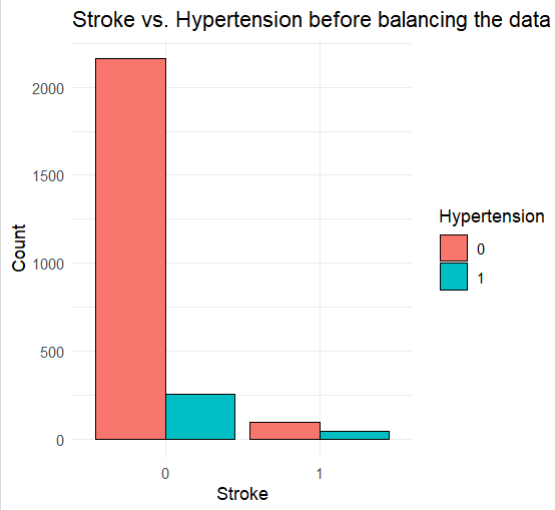
 Heart\_disease

 Hypertension

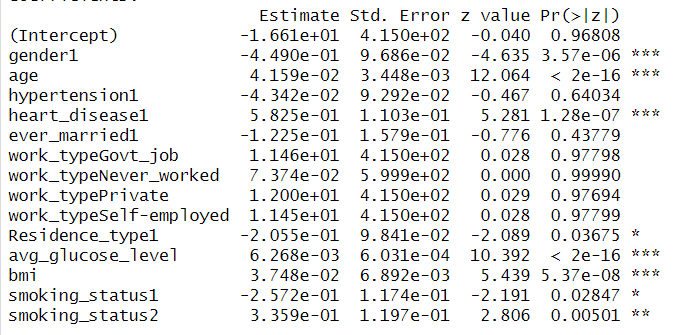
Here are plots of **stroke & heart disease** before and after balancing the data.



Here are plots of **stroke & hypertension** before and after balancing the data.



A model was fit on the balanced data.

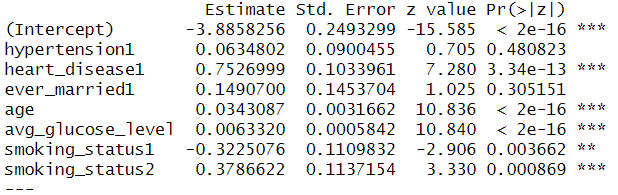


* 1. Precision value: 0.19
  + Only 19% of the positive predictions made by the model are correct.
* 2. Accuracy: 0.89
  + 89% of the predictions made by the model are correct.
* 3. Recall: 0.08
  + The model correctly identifies only 8% of all actual positive instances.
* 4. False Positive rate: 0.08
  + 8% of the negative instances were misclassified as positive.
* 5. False Negative rate: 0.68
  + 68% of the positive instances were misclassified as negative.
* 6. F1 score: 0.11
  + Model's performance is relatively low in terms of both precision and recall.

All in all, it means that the model is far from ideal. Even though the accuracy is high, the model still predicts only negative values well and has problems predicting positive values.

Let’s try fitting a reduced model based on the best predictors to see if there is any difference.

Based on the results of the full model, we can see that the most essential predictors are gender, age, heart\_disease, avg\_glucose level, bmi, and smoking status. After trying different variables taking in consideration previous covariance tests and the results of the full model, the combination of the following variables gives the best result in terms of evaluation metrics: **hypertension, heart\_disease, ever\_married, age, avg\_glucose\_level, smoking\_status.**



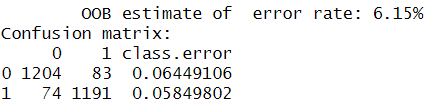
* 1. Precision value: 0.25
  + Only 25% of the positive predictions made by the model are correct.
* 2. Accuracy: 0.91
  + 91% of the predictions made by the model are correct.
* 3. Recall: 0.05
  + The model correctly identifies only 5% of all actual positive instances.
* 4. False Positive rate: 0.05
  + 5% of the negative instances were misclassified as positive.
* 5. False Negative rate: 0.66
  + 66% of the positive instances were misclassified as negative.
* 6. F1 score: 0.07
  + Model's performance is relatively low in terms of both precision and recall.

We will make an ANCOVA test to see whether the difference between models is significant.

### 

### Model 2 has a deviance that is lower than Model 1 by 85.926. And the p-value is very small (8.473e-16), suggesting strong evidence that Model 2 is a better fit than Model 1. But we remember that both models have problems predicting positive outcomes, which is more important in this case because it is better to tell a patient that they have a risk of having a stroke (positive outcome) than to say otherwise and make a mistake. So, neither of the models can be used in such a prediction.

Let’s try a random forest model. It can capture complex relationships and is robust against overfitting, so can potentially help with imbalance data.



### 1204 correct predictions for class 0 and 1191 correct predictions for class 1.

### 83 incorrect predictions for class 0 and 74 incorrect predictions for class 1.

### The class error rate is approximately 6.45% for class 0 and 5.85% for class 1.

### Overall, the Random Forest model seems to perform reasonably well, with an OOB error rate of 6.15%. But let’s try it on unseen test data.

* 1. Precision value: 0.15
  + Only 15% of the positive predictions made by the model are correct.
* 2. Accuracy: 0.91
  + 91% of the predictions made by the model are correct.
* 3. Recall: 0.05
  + The model correctly identifies only 5% of all actual positive instances.
* 4. False Positive rate: 0.05
  + 5% of the negative instances were misclassified as positive.
* 5. False Negative rate: 0.84
  + 84% of the positive instances were misclassified as negative.
* 6. F1 score: 0.07
  + Model's performance is relatively low in terms of both precision and recall.

On test data, the random forest model also performs poorly in terms of predicting positive outcomes.

**10. Conclusion**

We can conclude that neither of the models fitted can be used in such a serious task related to health as predicting strokes. Making a mistake in predicting positive values is especially significant because a person who is in the risk zone of having strokes won’t know about that. It would be reasonable to try other balancing techniques and more complicated models. But the best idea is to gather more data for both classes.