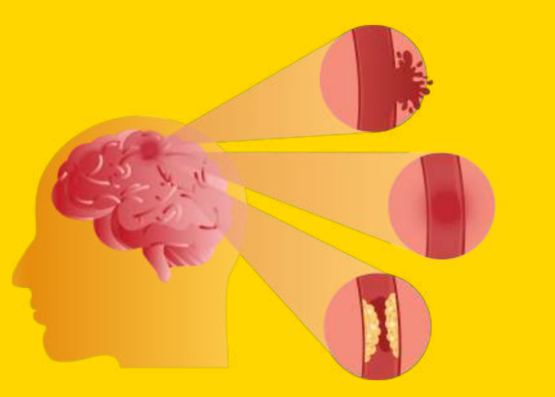
EZRA EUCLID TAMPOLI - 2602150573 - Data Science - ezra.tampoli@binus.ac.id - BINUS University
JEFFREY WIJAYA - 2602158784 - Data Science - jeffrey.wijaya@binus.ac.id - BINUS University
MARCEL KURNIAWAN - 2501977090 - Computer Science - marcel.kurniawan001@binus.ac.id - BINUS University
PATRICIA LAURENCIA WOGE - 2502042095 - Computer Science - patricia.woge@binus.ac.id - BINUS University

Silent Stroke Classification



Problem Statement

According to the **World Health Organization** (WHO) stroke is the **2nd leading cause of death globally**, responsible for approximately **11% of total deaths, incidence of stroke is increasing with age, mainly due to lifestyle changes**. Many people may think that a **stroke must show symptoms**. Whereas there are also **strokes that do not have obvious symptoms** or commonly **called silent strokes**. This condition is even more common. **Silent stroke is a type of stroke that usually does not show common symptoms**. In many cases, sufferers do not even know they have the disease. Therefore, our group wanted to **identify these factors such as age, gender, environment, status, lifestyle and medical history by creating a prediction model**.

Objective

This analysis can **help** people predict **silent stroke**, which is a stroke that **does not cause obvious symptoms**, which can be influenced by various factors and can occur in individuals of any **age, gender, environment, status, lifestyle and medical history**.

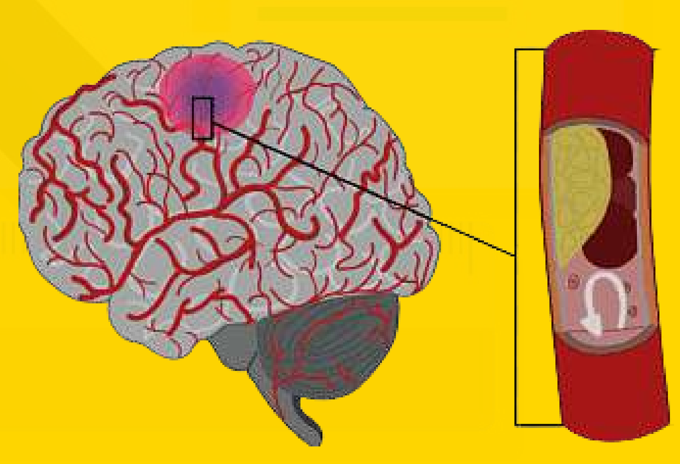
Data Description

This dataset is used to **predict whether a patient is likely to get stroke** based on the input parameters like **gender, age, various diseases, and smoking status**. Each row in the data provides relavant information about the patient.

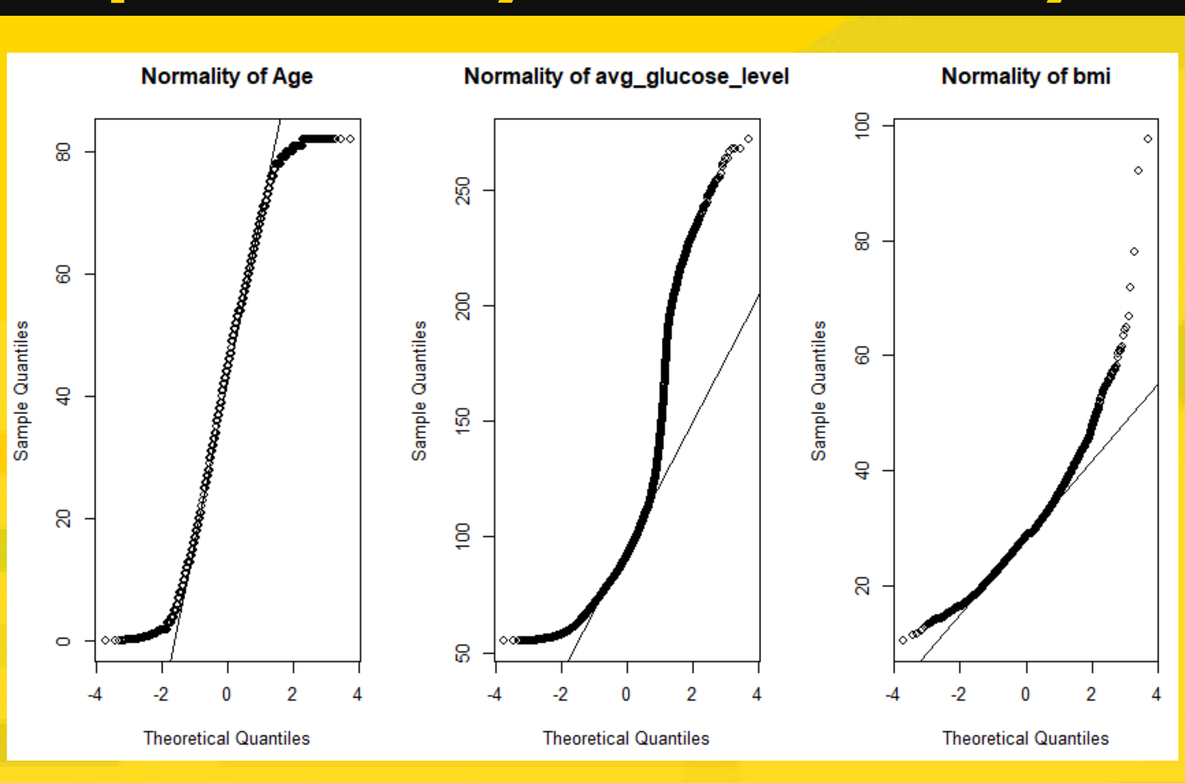
Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download

This dataset consists of 5110 observations and 12 columns

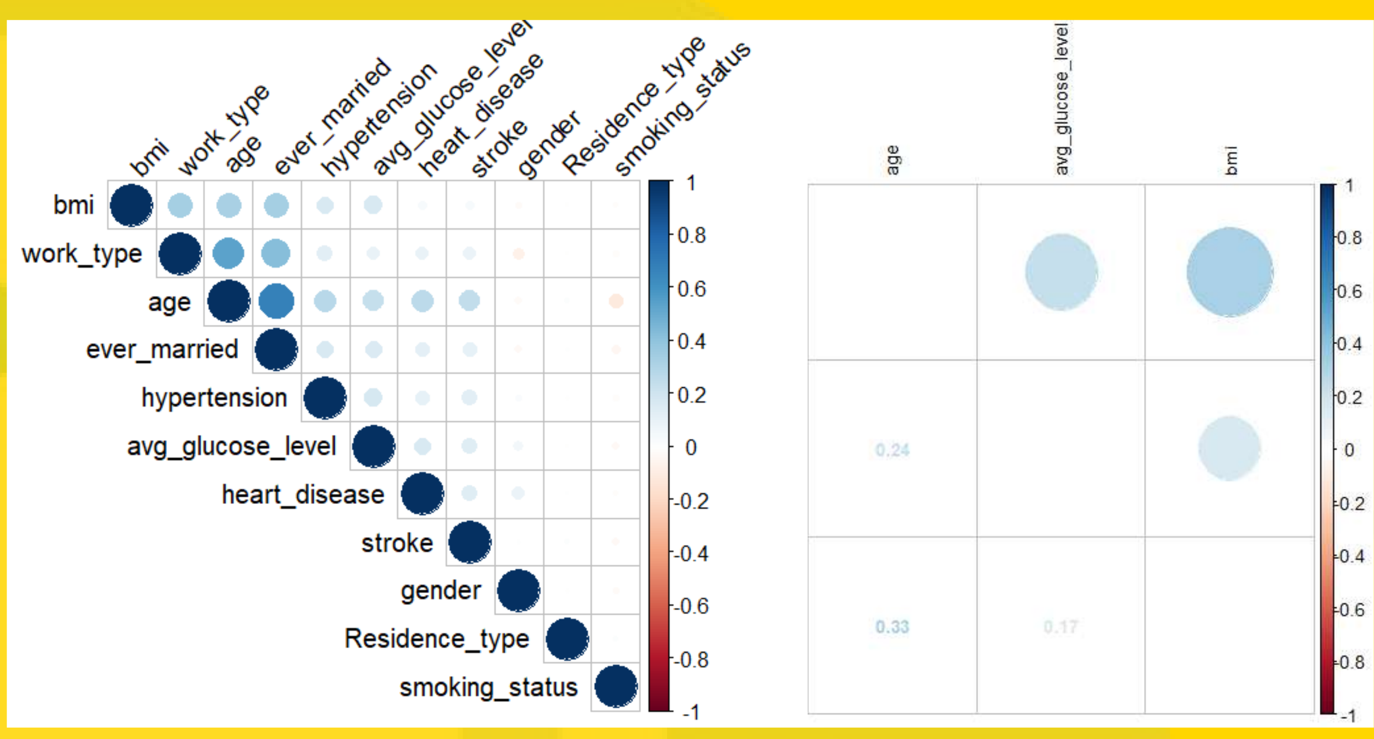
1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 doesn't have hypertension, 1 has hypertension
5) heart_disease: 0 doesn't have heart diseases, 1 has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "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"*
12) stroke: 1 if the patient had a stroke or 0 if not (Note: "Unknown" in smoking_status means that the information is unavailable for this patient)



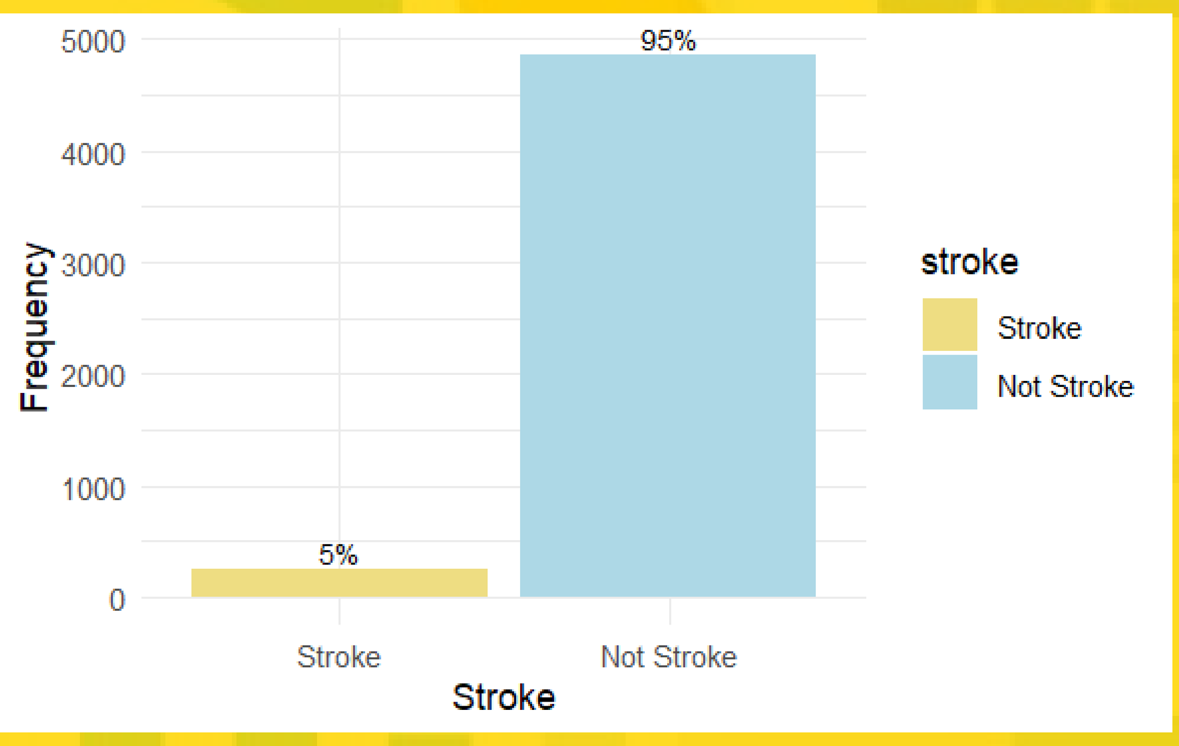
Exploratory Data Analysis



In numerical variables, age and bmi are normally distributed, although for bmi it is not too perfect because there are outliers in it. Meanwhile, avg_glucose_level tends not to be normally distributed and is more directed to the right-skewed distribution.



All correlations are positive, but there are no strong associations because the values are below 0.5 except age and ever_married. The continuous variables do not exhibit strong correlations, because the Pearson's correlation is less than 0.4 for all of them.



In this dataset, it can be seen that the unbalanced dataset on the target variable, stroke, has a highly disproportionate distribution. About 95% of the observations did not have a stroke, while the remaining 5% had a stroke.

- Higher age, may likely to suffer from stroke.
- Higher glucose_level the higher people can get in a stroke!
- Most people's BMI levels are around 20 to 30 and higher does not mean they are more likely to have a stroke. But it can be seen that higher bmi may easier to suffer from strokes.

Predictive Model & Discussion

Logistic Regression

```
"binomial", data = strokeTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1531  -0.3083  -0.1578  -0.0879   3.5133

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -6.463e+00  7.995e-01  -8.084 6.27e-16 ***
genderMale                -5.868e-02  1.613e-01  -0.364  0.71604
age                        7.491e-02  6.704e-03  11.174  < 2e-16 ***
hypertension1              5.526e-01  1.812e-01   3.050  0.00229 **
heart_disease1             2.375e-01  2.163e-01   1.098  0.27240
ever_marriedYes           -1.195e-01  2.551e-01  -0.468  0.63953
work_typeGovt_job         -1.450e+00  8.730e-01  -1.661  0.09675 .
work_typeNever_worked     -1.060e+01  3.326e+02  -0.032  0.97456
work_typePrivate          -1.206e+00  8.491e-01  -1.420  0.15552
work_typeSelf-employed    -1.643e+00  8.767e-01  -1.874  0.06099 .
Residence_typeUrban        3.588e-02  1.568e-01   0.229  0.81898
avg_glucose_level          4.346e-03  1.344e-03   3.233  0.00123 **
bmi                        1.742e-04  1.299e-02   0.013  0.98930
smoking_statusnever smoked -1.518e-01  1.844e-01  -0.823  0.41041
smoking_statussmokes       1.222e-01  2.455e-01   0.498  0.61860
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1555.1  on 4086  degrees of freedom
Residual deviance: 1233.2  on 4072  degrees of freedom
AIC: 1263.2

Number of Fisher Scoring iterations: 14
```

Random Forest	MeanDecreaseGini
gender	10.298477
age	78.039809
hypertension	9.266833
heart_disease	8.206333
ever_married	6.886169
work_type	16.357563
Residence_type	10.558507
avg_glucose_level	97.449988
bmi	78.173896
smoking_status	17.713473

age	avg_glucose_level	bmi	work_type	ever_married
34.287882	15.733506	8.995658	4.599397	2.598070
smoking_status	gender	heart_disease	Residence_type	
2.488491	2.366200	1.841972	1.171199	

Decision Tree

After experimenting with three models, namely **random forest**, **decision tree** and **logistic regression** with each model **using all available variables**, the reason **we used all the variables** in the dataset is because **it goes back to our objective** where **we want to predict silent strokes that can occur regardless of gender, environment, status, lifestyle and medical history that can cause silent strokes**. Therefore, we compared several **different models** for this case.

Based on the performance between the three results, the **random forest model was found to give better results than decision tree or logistic regression and decision tree gave the lowest accuracy**. Although in terms of accuracy the logistic regression model provides a slightly higher accuracy of **0.946 or 0.002 greater than the random forest model**, the **random forest model can better handle non-linear cases, insignificant variables and redundant features**. Therefore, **we chose random forest** as our model for classification in this case.

Conclusion

	0	1
0	956	56
1	1	0

For the **final model**, we used a **random forest** model **using all variables** because it goes **back to the objective**, which is to **predict silent strokes** that can occur regardless of gender, lifestyle, status, environment, and medical history **that can cause silent strokes**. For the **evaluation** of this model, we used a **confusion matrix**, results of which were the number of data correctly predicted as "have stroke" by random forest was 0 (TP), The number of data that should have been classified as "not have stroke" but were incorrectly predicted as "have stroke" by the random forest model was 1 (FN), The number of data that should have been classified as "have stroke" but were incorrectly predicted as "not have stroke" by the random forest model was 56 (FP), The number of data correctly predicted as "not have stroke" by the random forest model was 956 (TN). The **overall accuracy** of this random forest model model is **94.4% with 99% sensitivity**, which means it **can accurately predict TP (True Positive) with a very minimum error rate**. True positives are **very important for medical data** as TP can **lead people to live a much healthier life even if the prediction is wrong**.