أكاديمية سدايا
SDAIA Academy

Stroke Prediction

16 December 2021

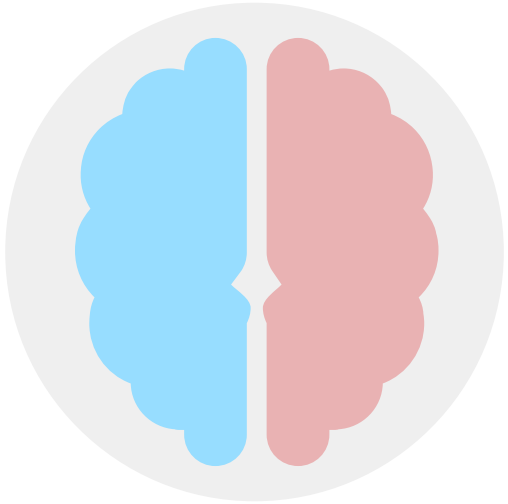
MRS Make robust solution

Rahaf Almayouf , Shatha Alghamdi , Munira Alshahrani

Outline

1. Introduction
2. Workflow
3. Data Overview
4. Exploring Data Analysis (EDA)
5. Algorithm
6. Challenges
7. Tools
8. Conclusion

Introduction

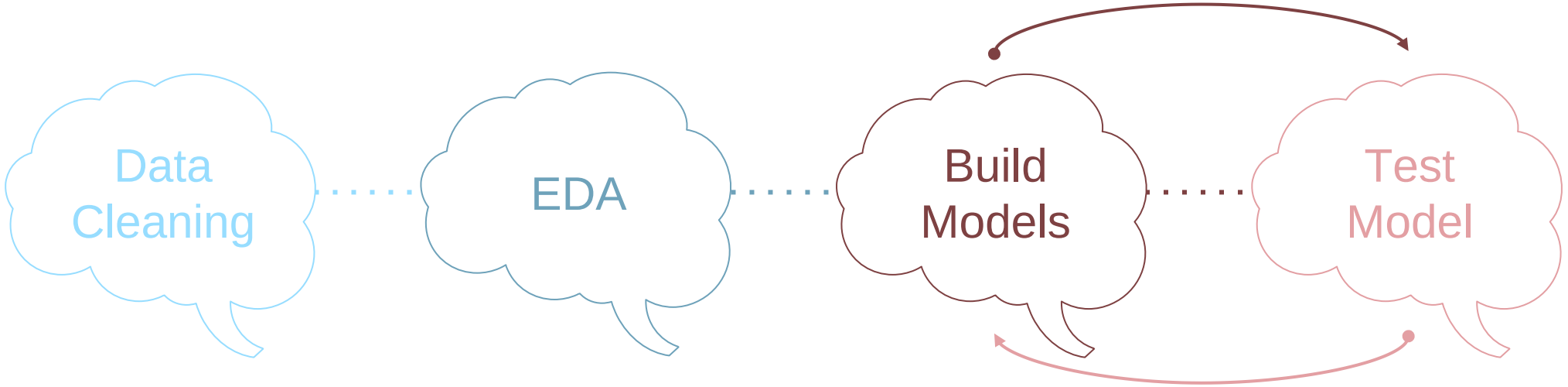


According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Targets

- Awareness of the most common causes of strokes.
- Concluding who are the most susceptible to strokes from people.
- Realizing the difference between smokers and non-smokers in having strokes.
- Knowing the ages most likely to have strokes.

Workflow



Data Overview

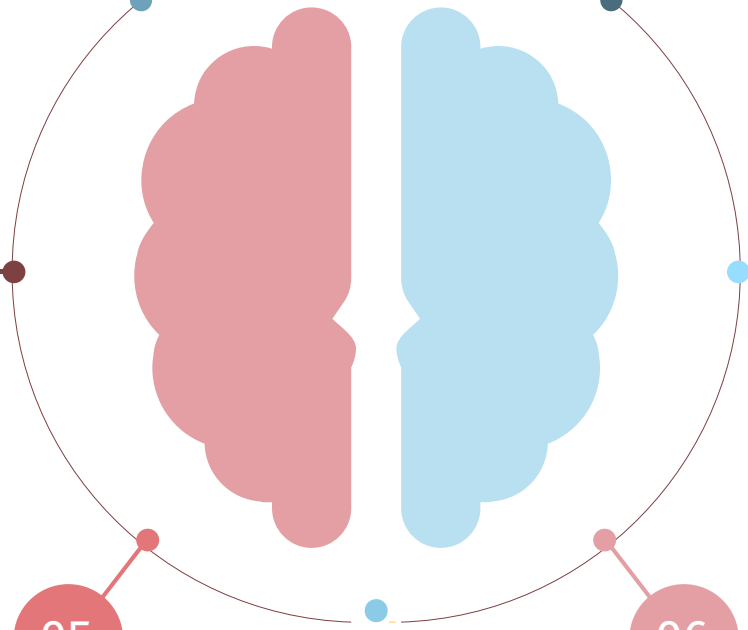
- The dataset consists of 12 features and 5110 observations.
- Source from Kaggle.
- Features description:

id	int
age	float
gender	object
hypertension	int
heart_disease	int
ever_married	object
work_type	object
Residence_type	object
avg_glucose_level	float
bmi	float
smoking_status	object
stroke	int

Exploring Data Analysis (EDA)

01 Does age have impact on strokes?

02 Does body mass index and glucose levels in a person, propel a stroke?



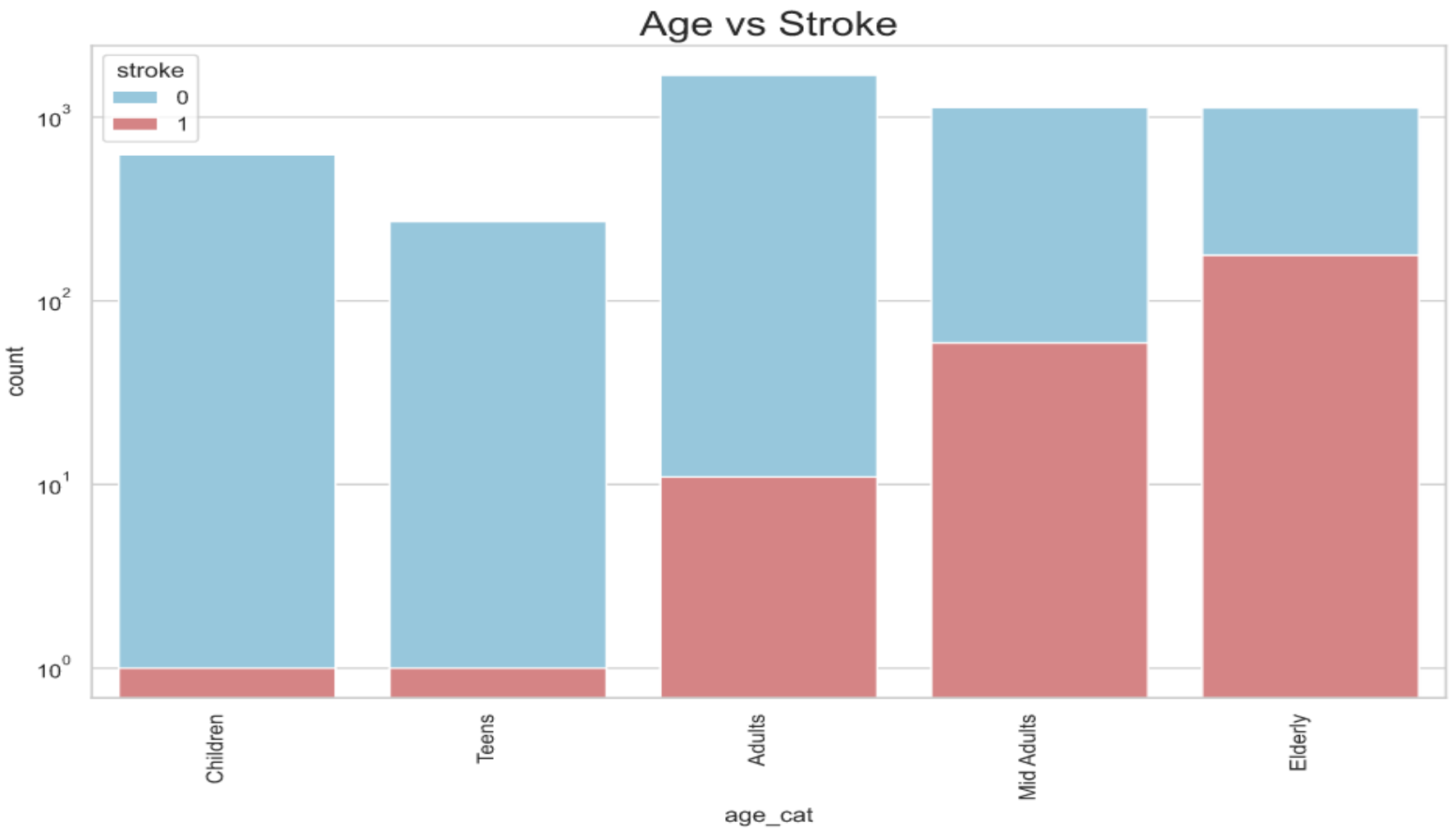
03 Assumption: Smoking can induce Stroke, is it true?

04 Assumption: Heart with a Heart Disease is prone to Stroke, is it true?

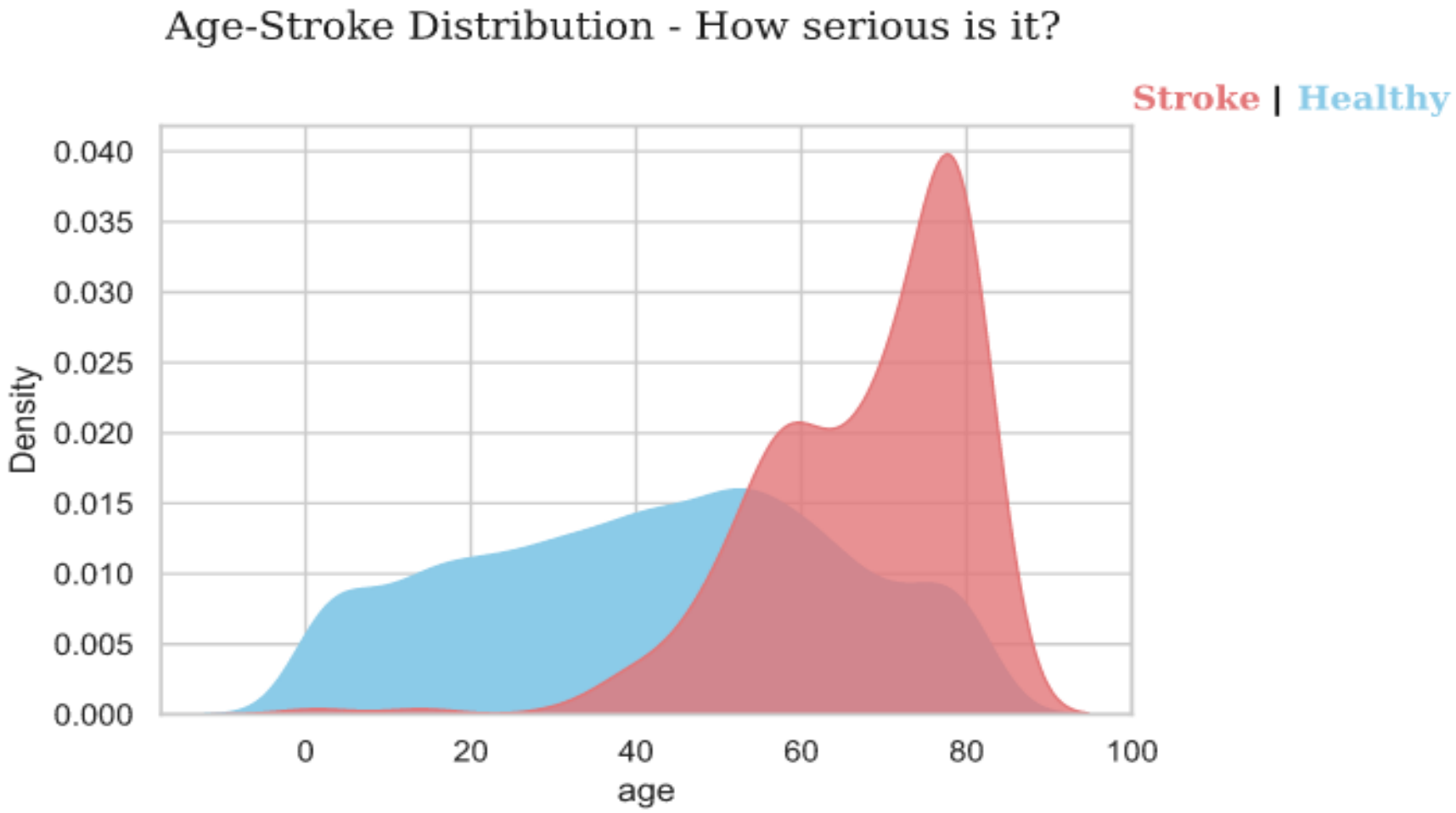
05 Assumption: Workload results in high blood pressure and that could lead to Stroke, is it true?

06 Assumption: Males are most susceptible to strokes due to high work-related stress, is it true?

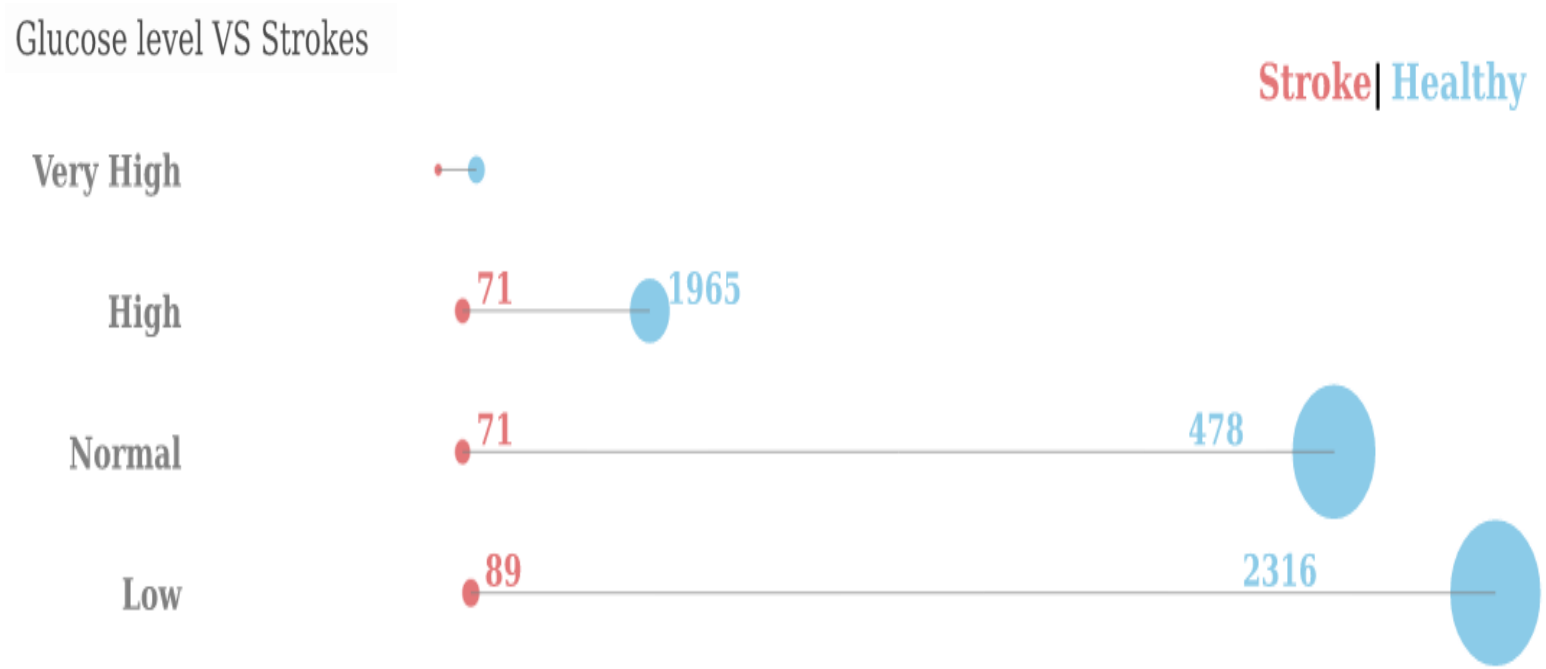
01 Does age have impact on strokes?



01 Does age have impact on strokes?

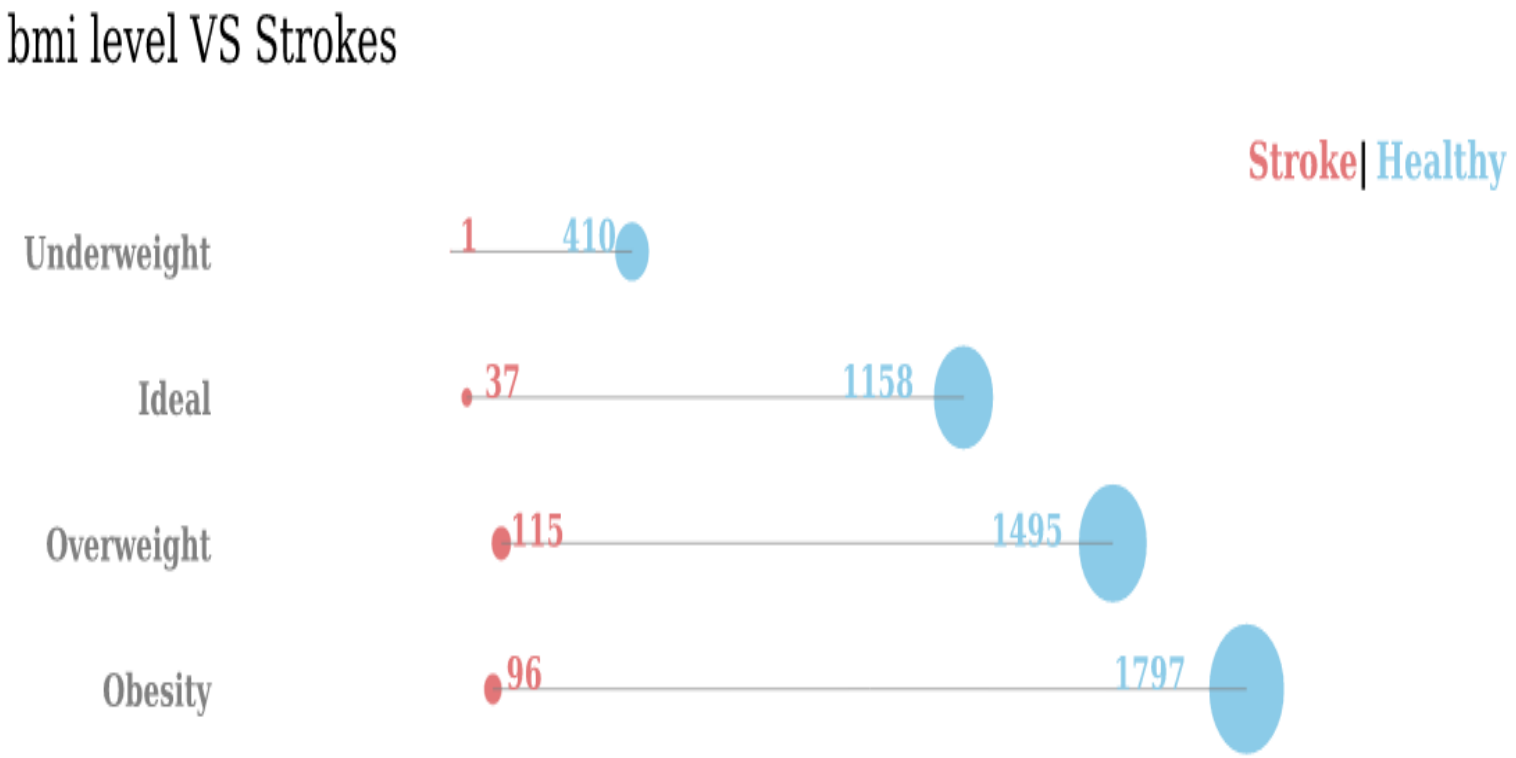


02 Does body mass index and glucose levels in a person, propel a stroke?



02

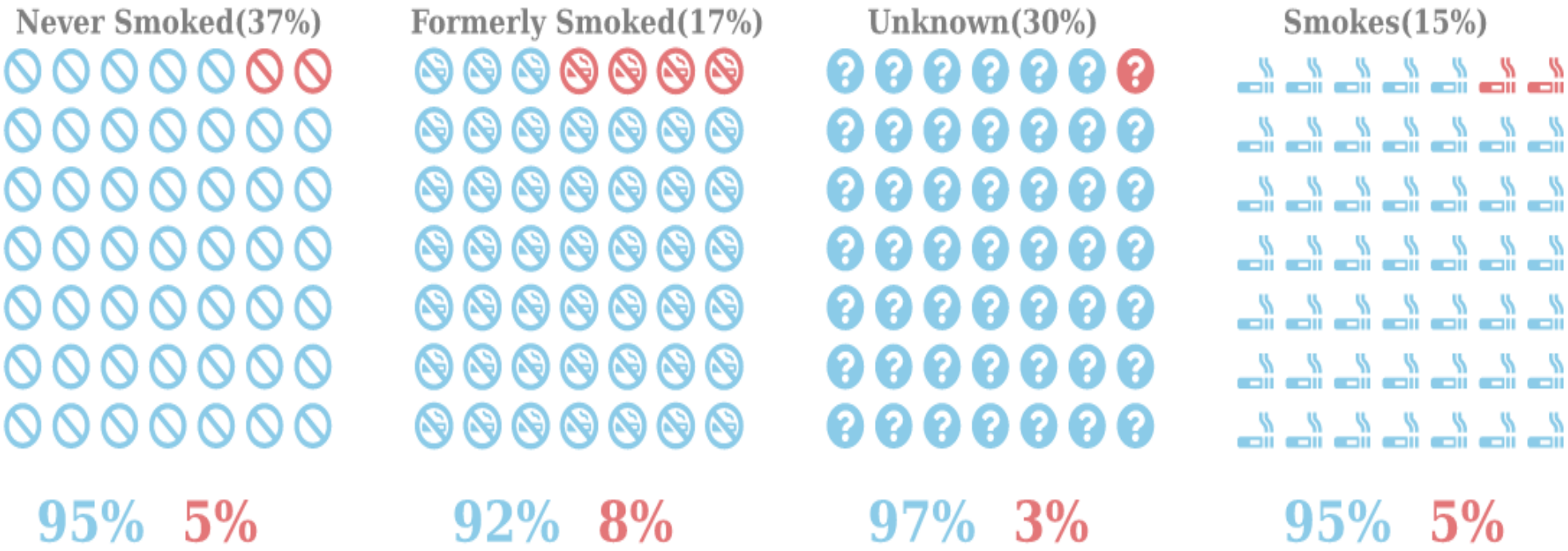
Does body mass index and glucose levels in a person, propel a stroke?



03 Assumption: Smoking can induce Stroke, is it true?

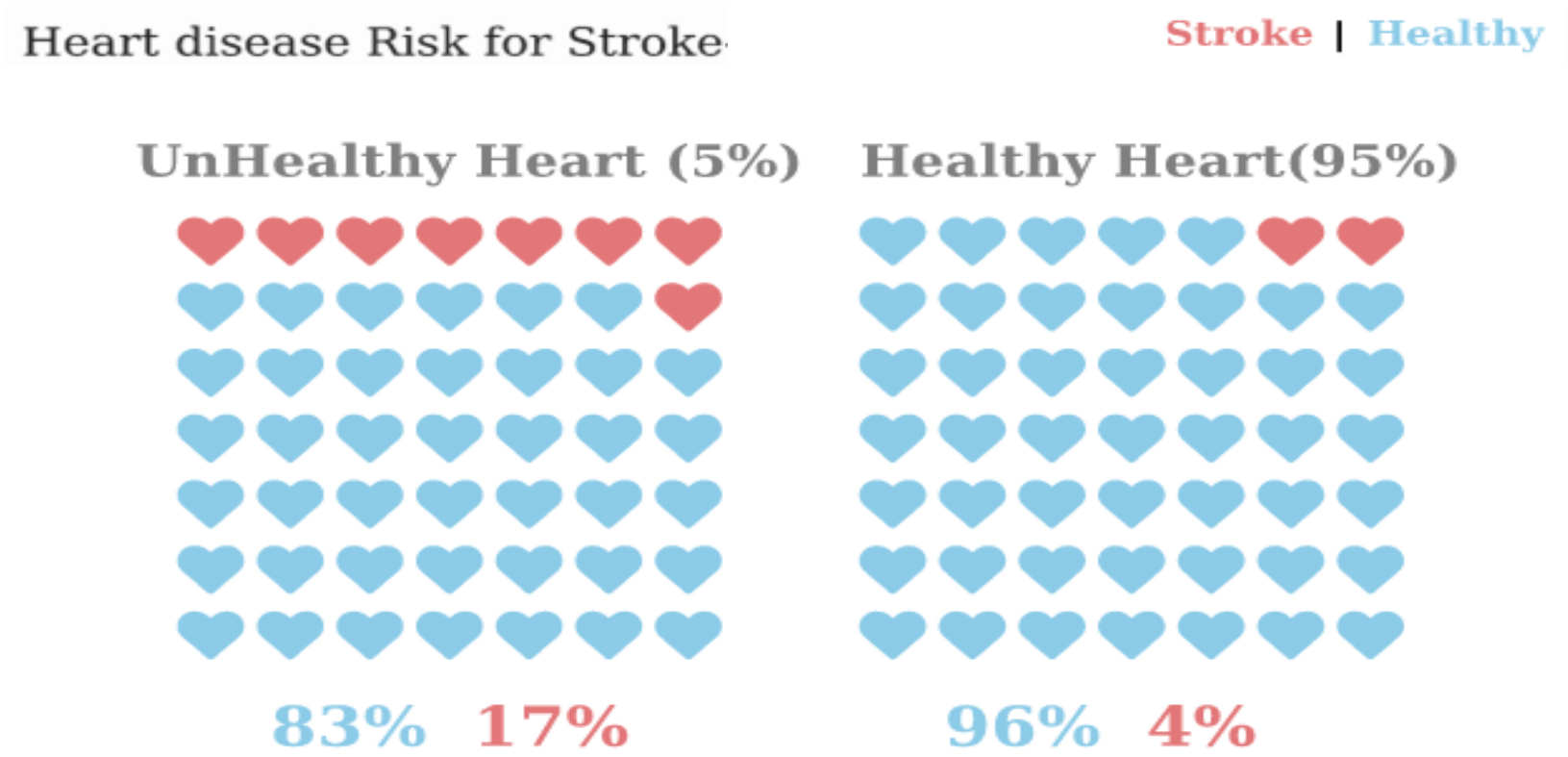
Smoking and Stroke- Does smoking habit could cause Stroke?

Stroke | Healthy



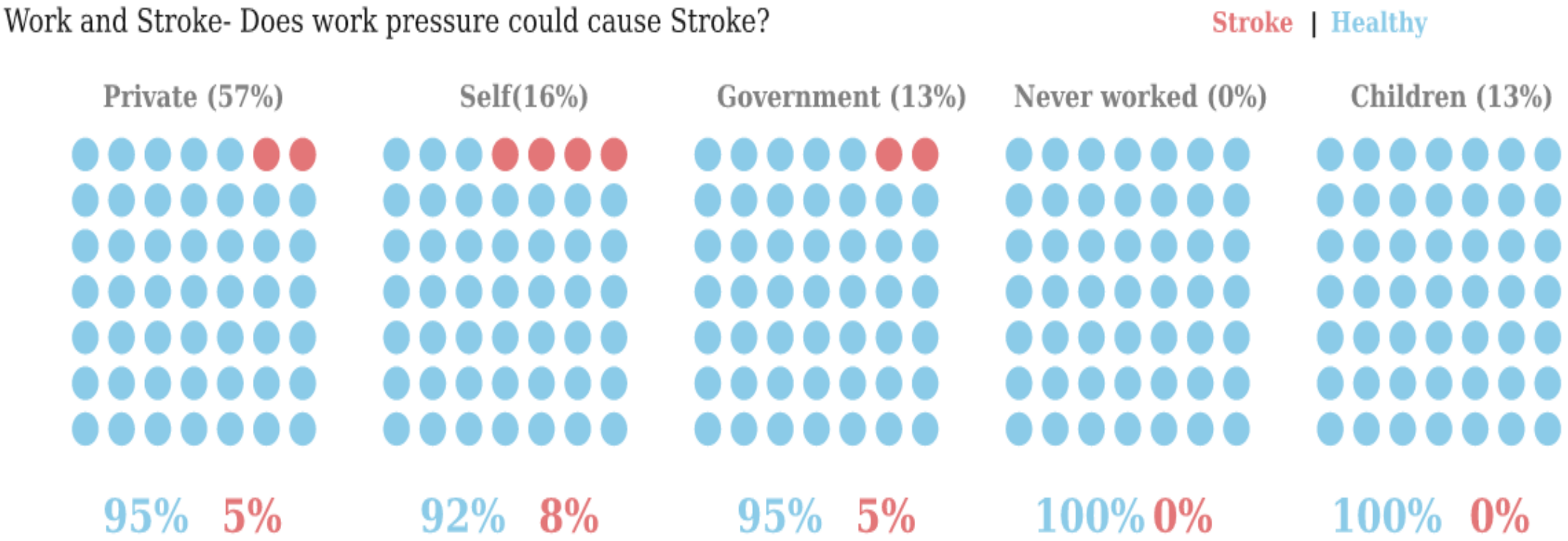
04

Assumption: Heart with a Heart Disease is prone to Stroke, is it true?



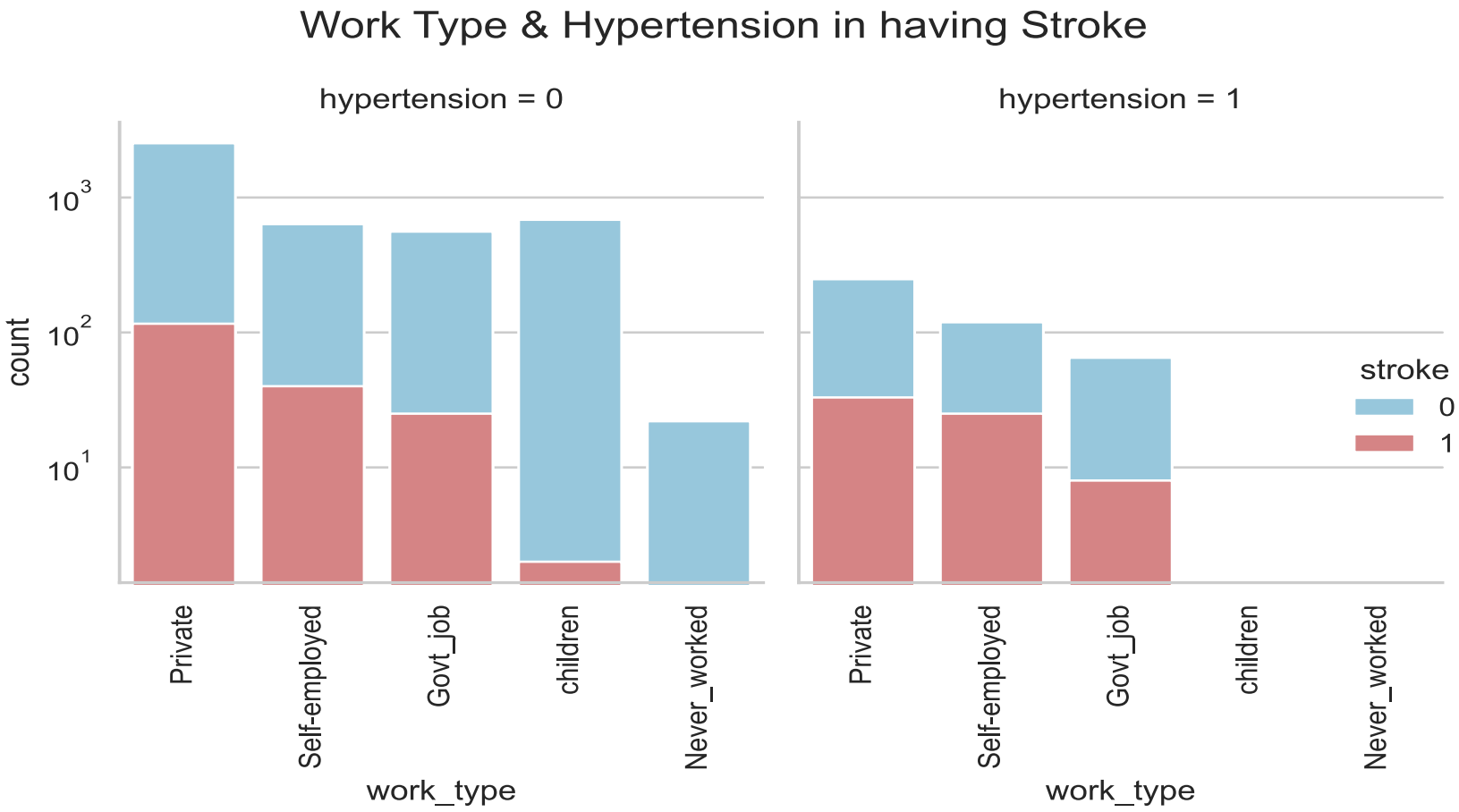
05

Assumption: Workload results in high blood pressure and that could lead to Stroke.



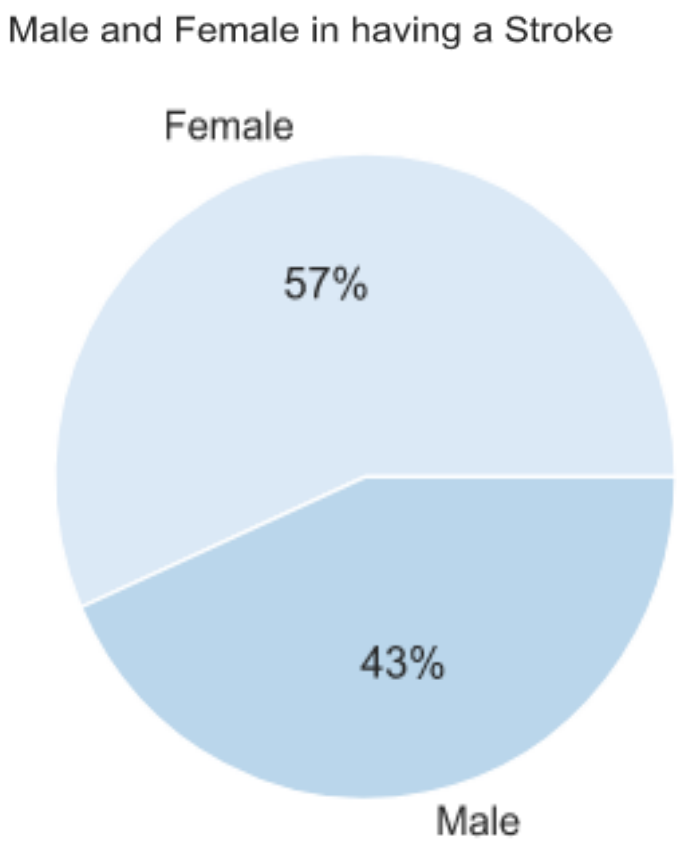
05

Assumption: Workload results in high blood pressure and that could lead to Stroke

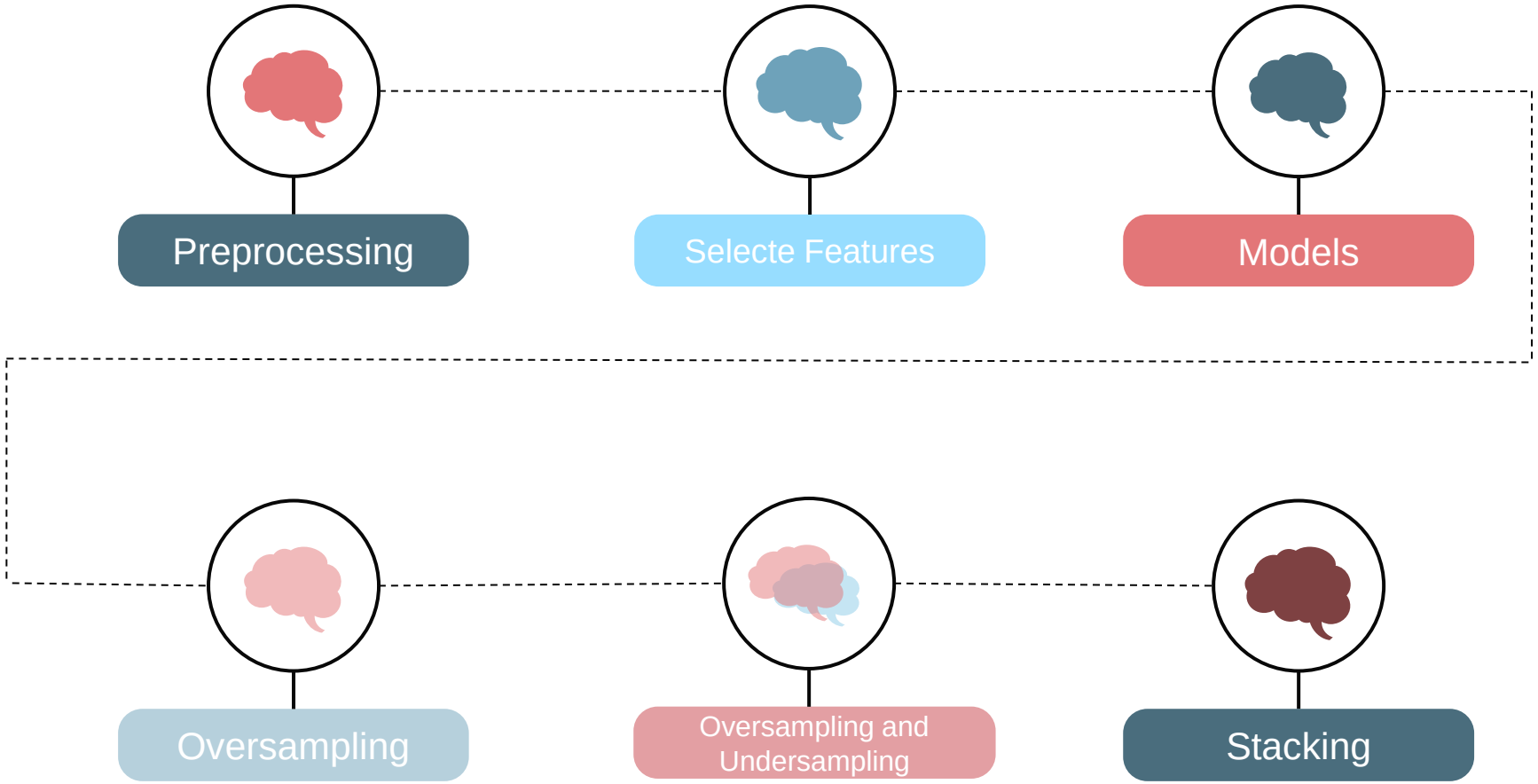


06

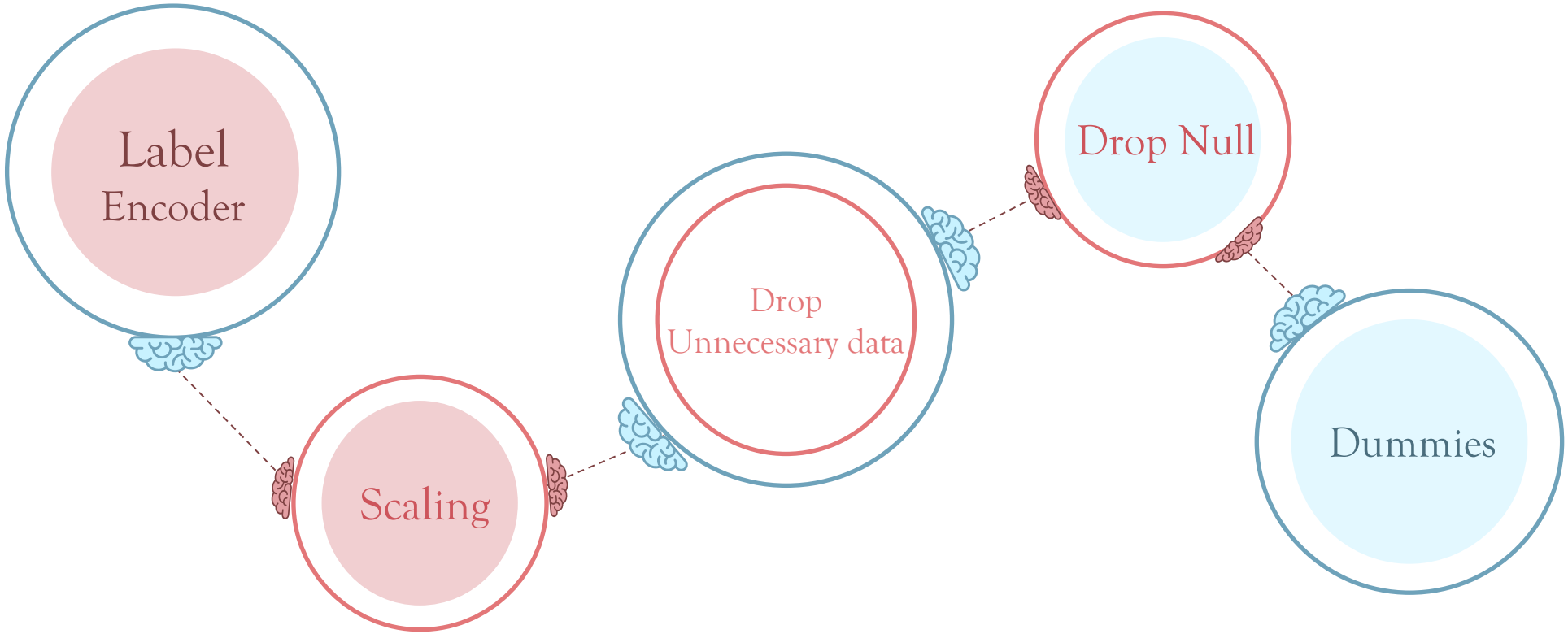
Assumption: Males are most susceptible to strokes due to high work-related stress.



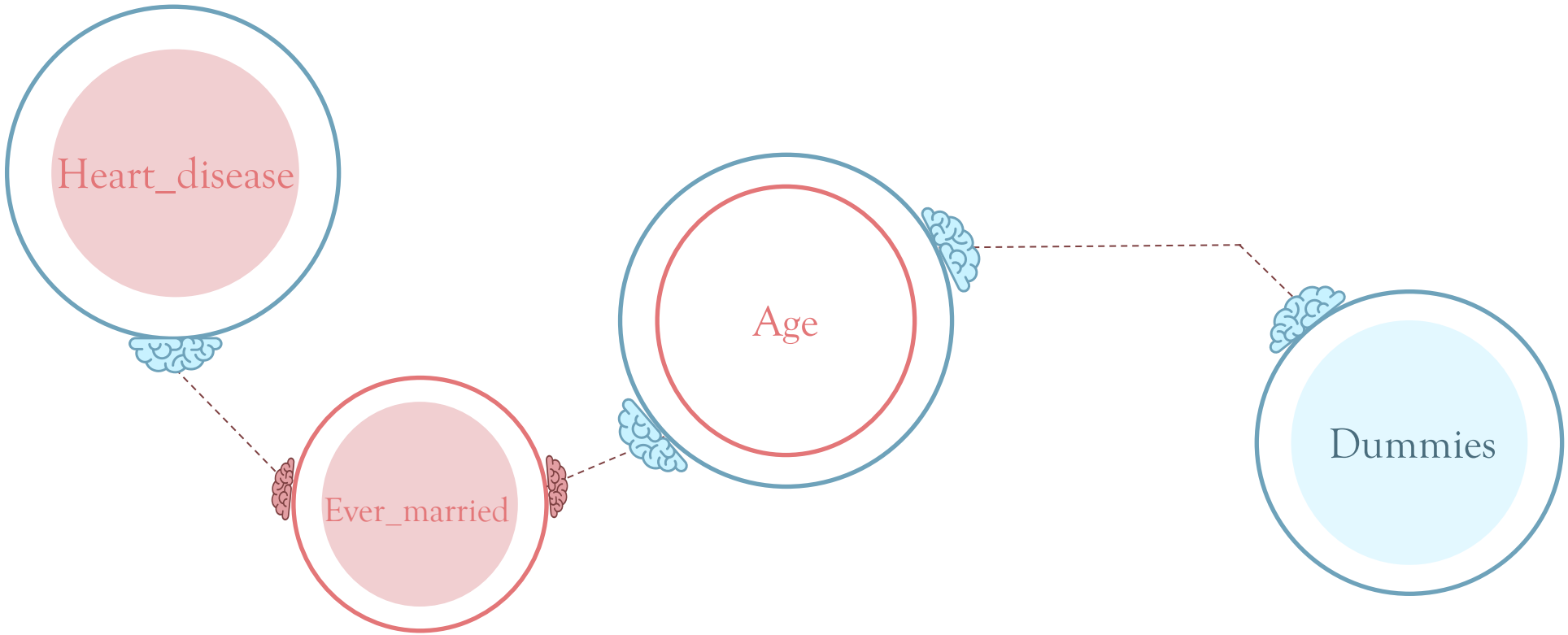
Algorithm


Preprocessing
Selecte Features
Models
Oversampling
Oversampling and Undersampling
Stacking

Preprocessing

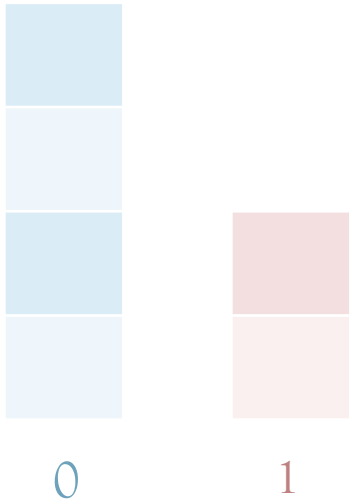


Selected Features



BUT !

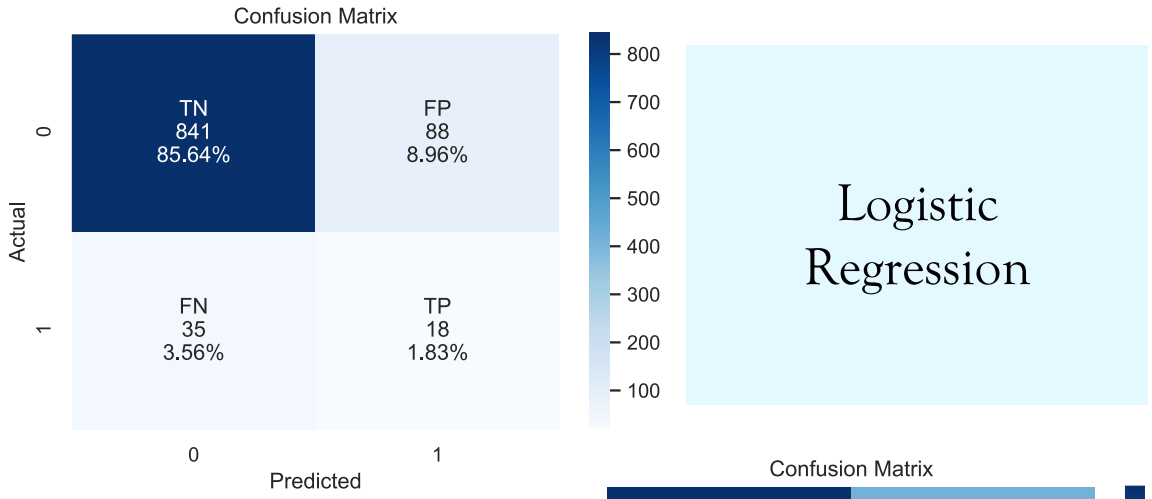
We found that we have a problem of imbalanced data set, namely that target of **1** is less than number of **0**.



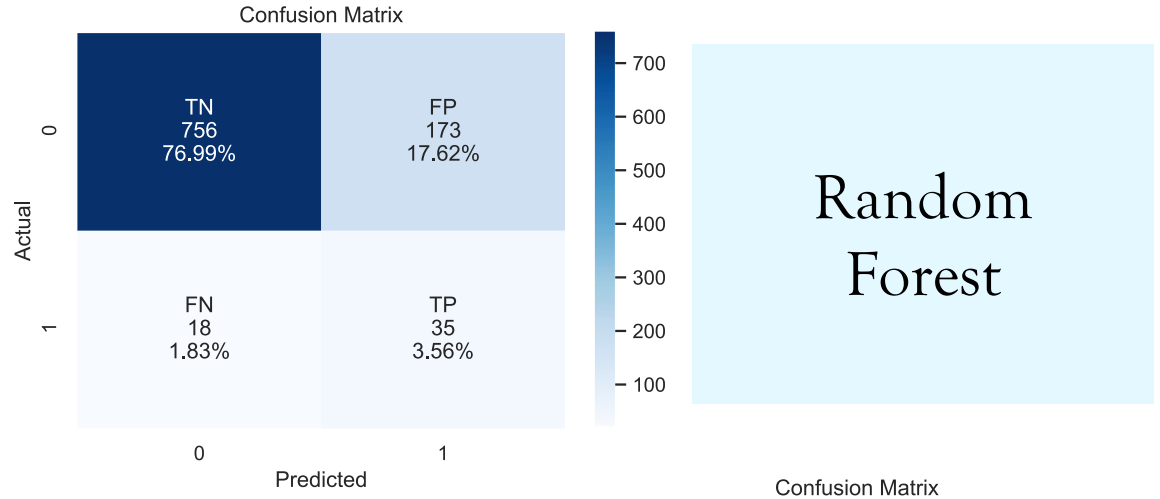
Oversampling

SMOTE

0 → 929 1 → 53

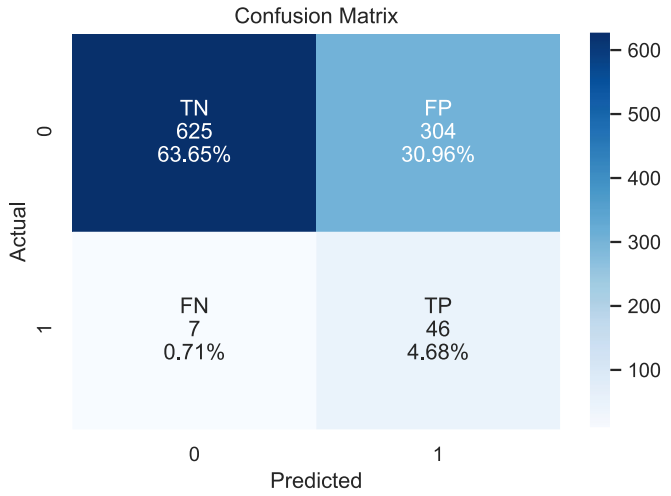


Logistic Regression

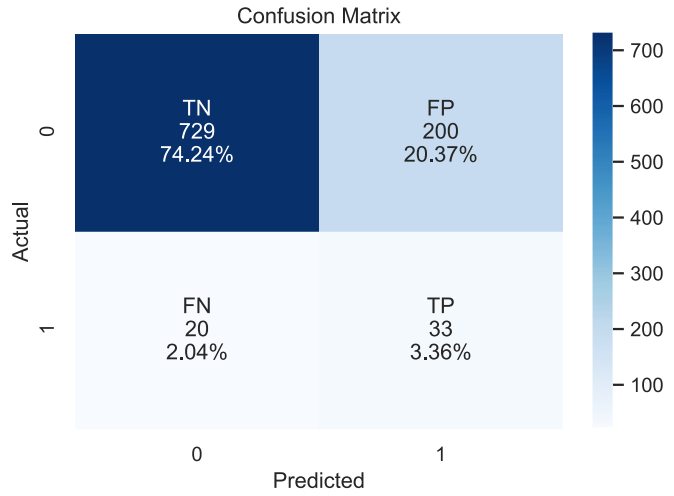


Random Forest

KNN



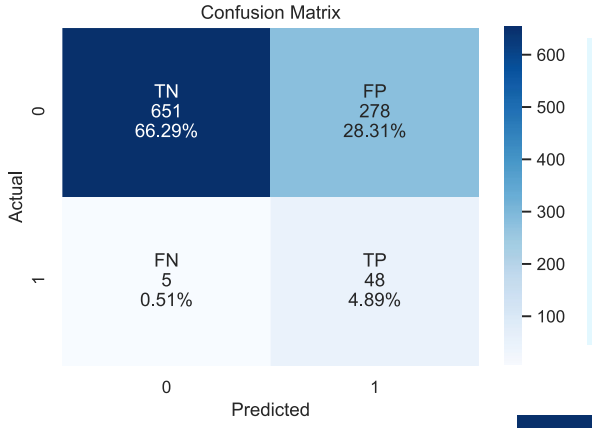
Decision Tree



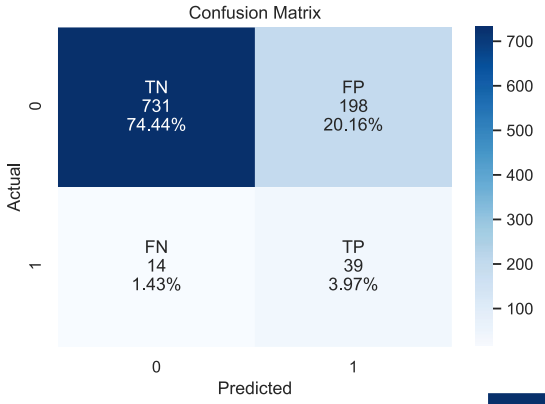
Oversampling

SMOTE

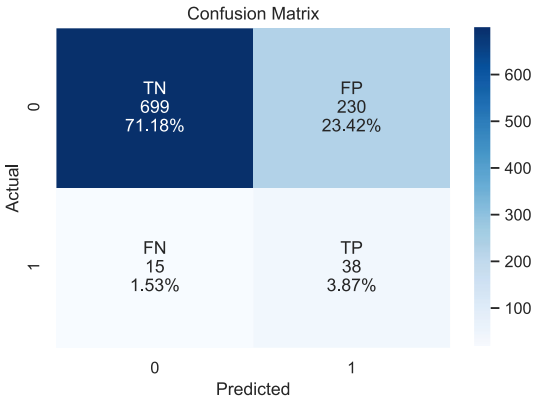
0 → 929 1 → 53



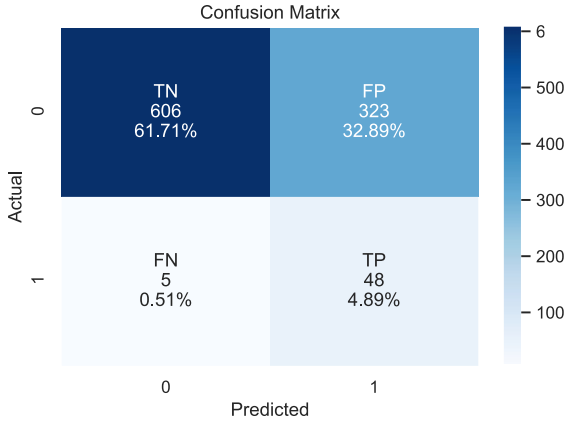
Naive Bayes



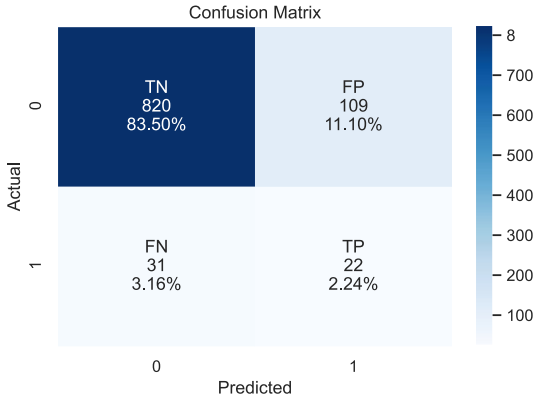
XGBoost



SVM



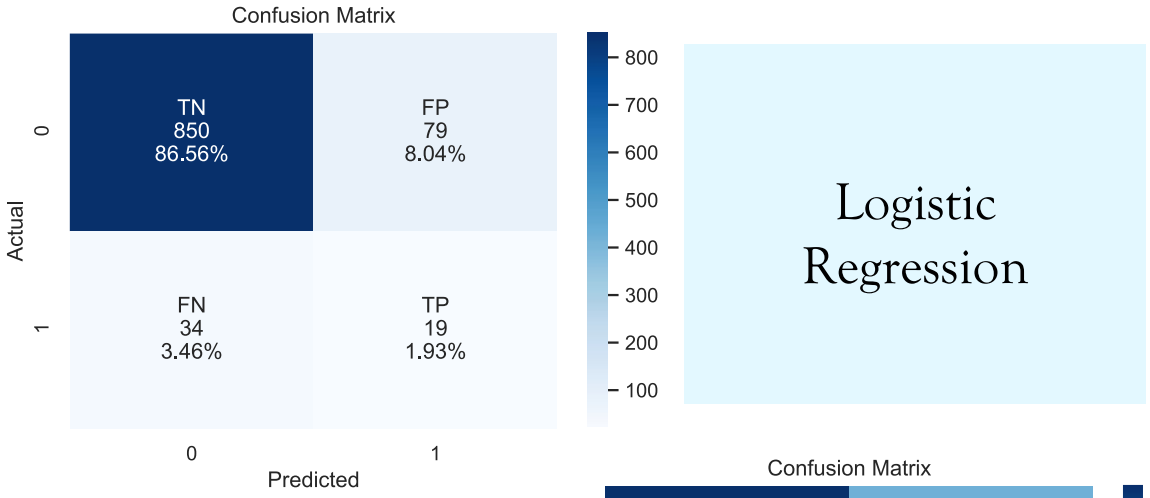
Gradient Boosting



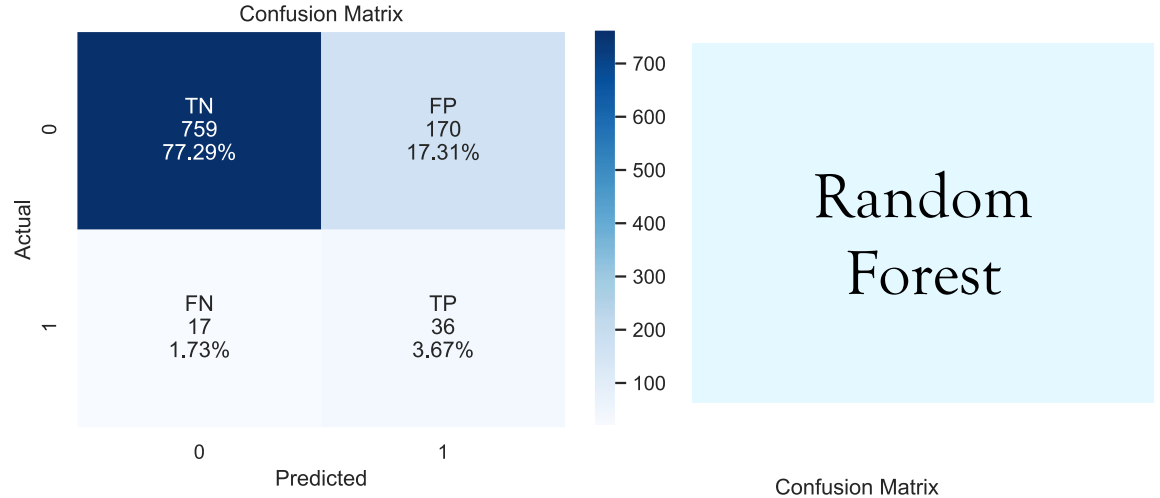
LightGBM

Oversampling &Undersampling

SMOTE + Tomek 0 → 929 1 → 53

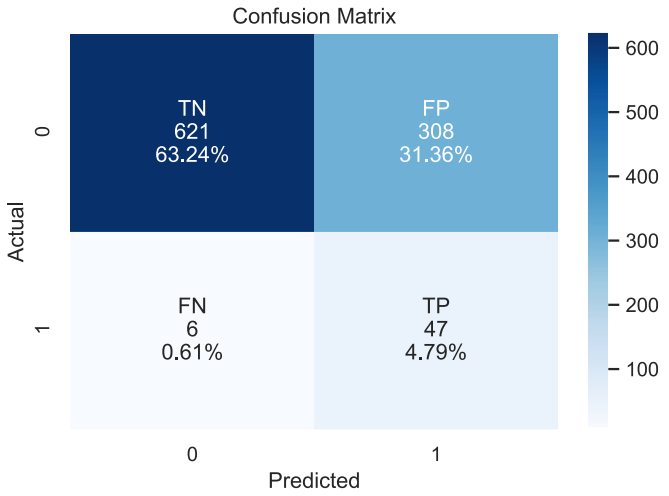


Logistic Regression

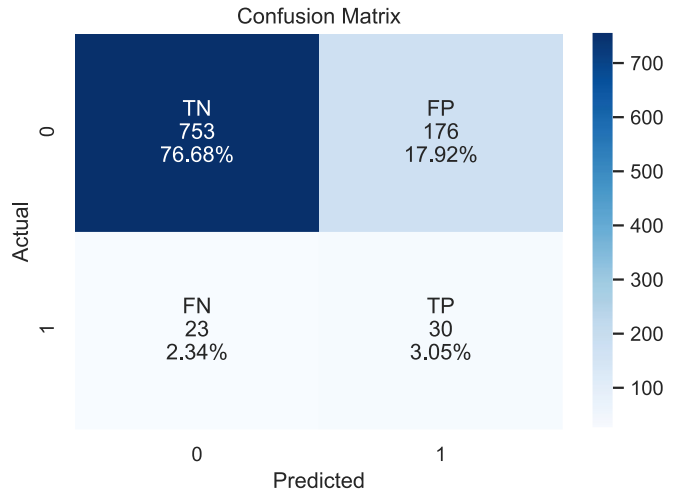


Random Forest

KNN

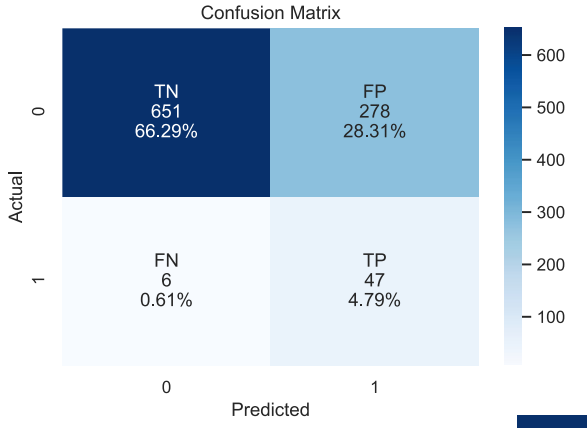


Decision Tree

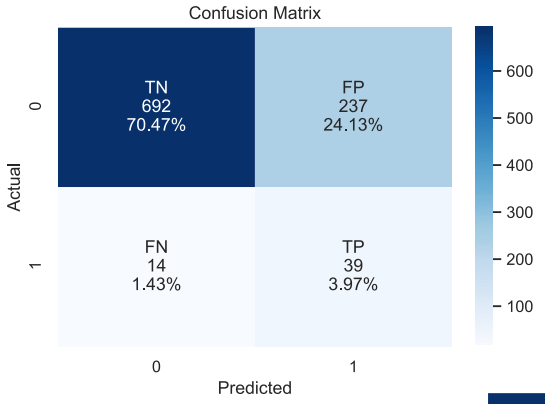


Oversampling & Undersampling

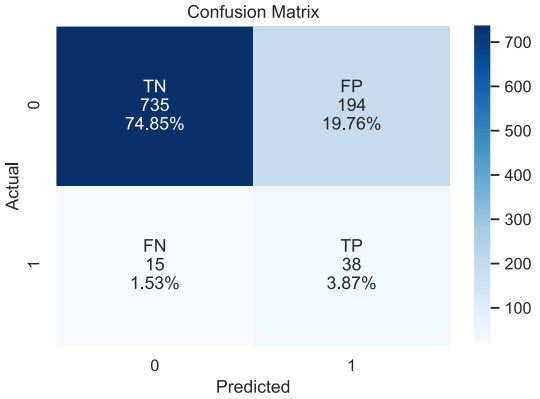
SMOTE + Tomek 0 → 929 1 → 53



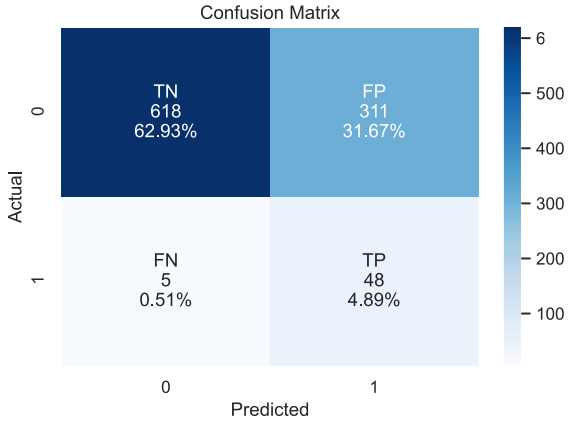
Naive Bayes



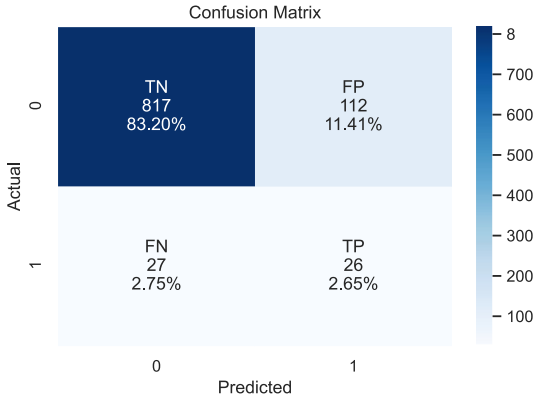
XGBoost



SVM

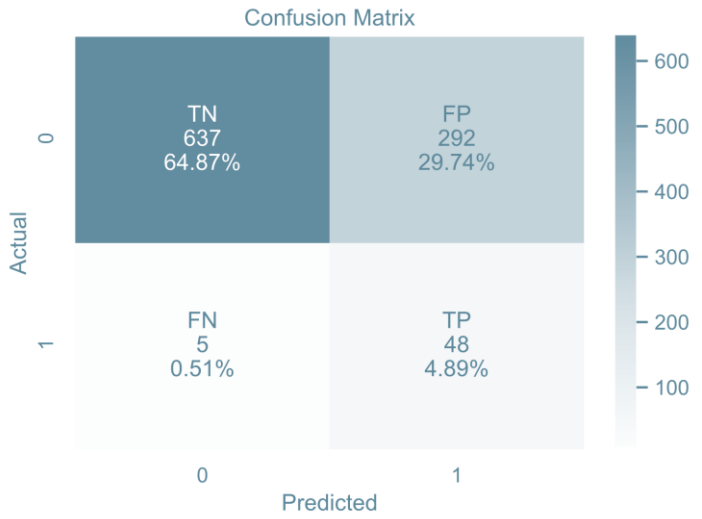


Gradient Boosting

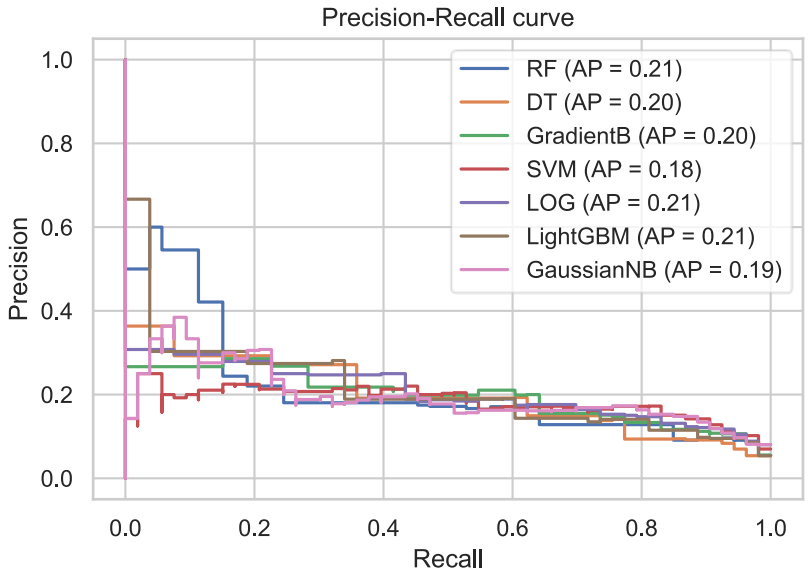


LightGBM

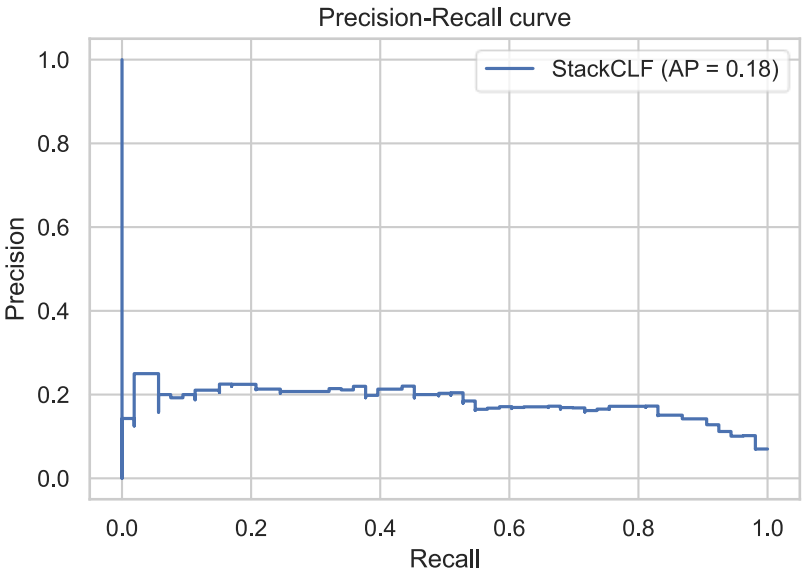
Stacking



Stacking Model



PRC for Models



PRC for Stacking Model

Challanges

01

Handling Imbalance data

02

Find a suitable hyperparameter

03

Long execution time

Tools

Libraries

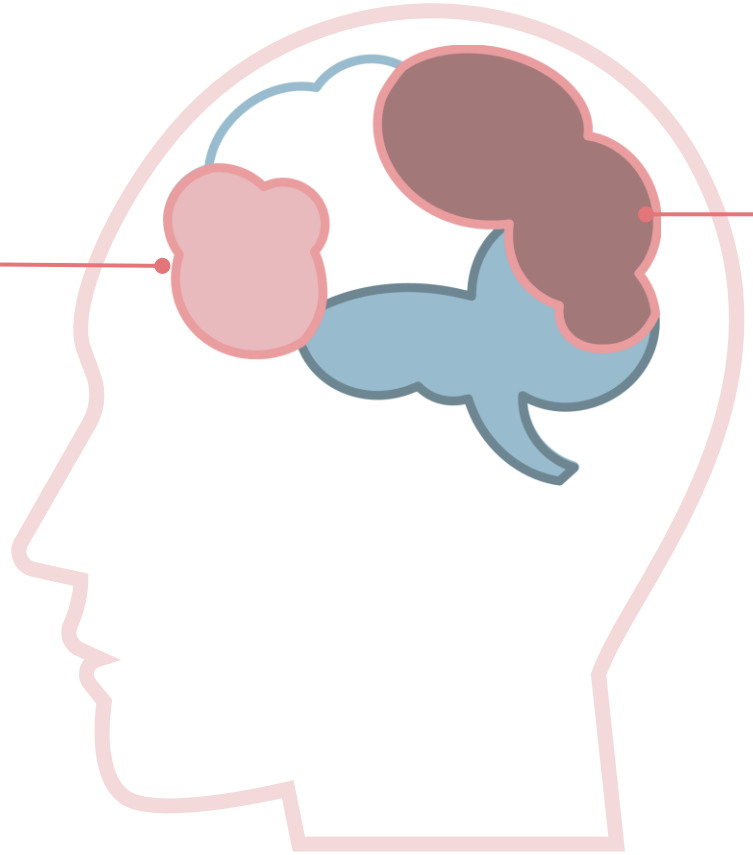
pandas, numpy, matplotlib, seaborn, plotly, sklearn, imblearn, pywaffle.

Software

Trello, GitHub, Jupyter, VSCode, Word & PowerPoint, Zoom.

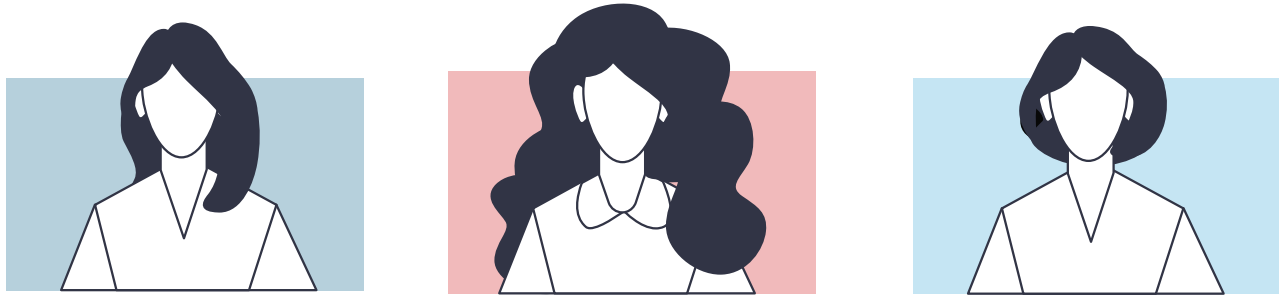
Conclusion

Do an awareness program that explains all the causes for being exposed to get a stroke, teach people how to do the first medical help until the Ambulance.



Build a platform to predict if any patient is exposed to get a stroke at any time, to call the patient and take the first step to prevent any bad situation.

Thank you



Feedback ?