STROKE

AS PROJECT 3 SDAIA DATA SCIENCE BOOTCAMP(T5)

BY:

- AHAD ALMUTAIRI
- RAYAN ALMUWAYNI
- ABDULSALAM ALQARNI

TABLE OF CONTENTS

01 **BACKGROUND**
About stroke and problem statement?

02 **METHODOLOGY**
How the works is done?

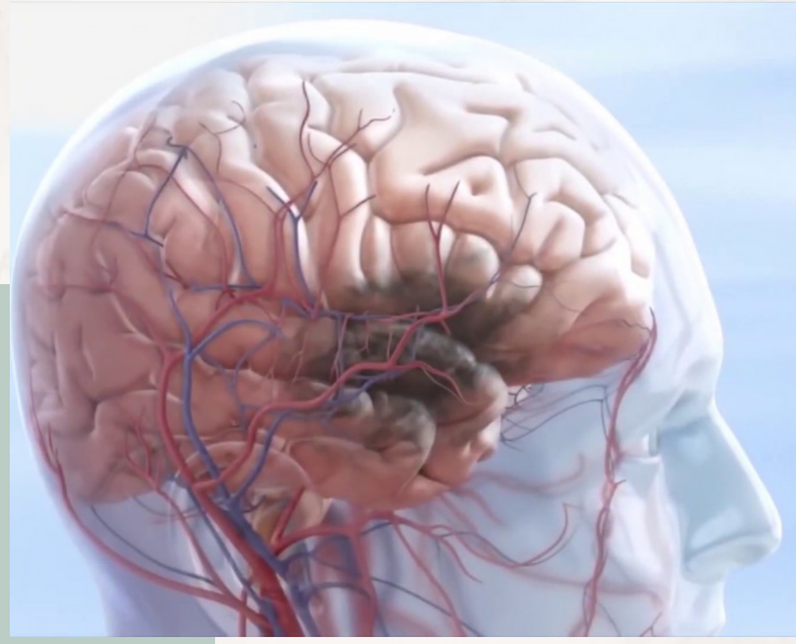
03 **TOOLS**
What are the tools used?

04 **CONCLUSION**
Future work and summery

01

BACKGROUND

About stroke



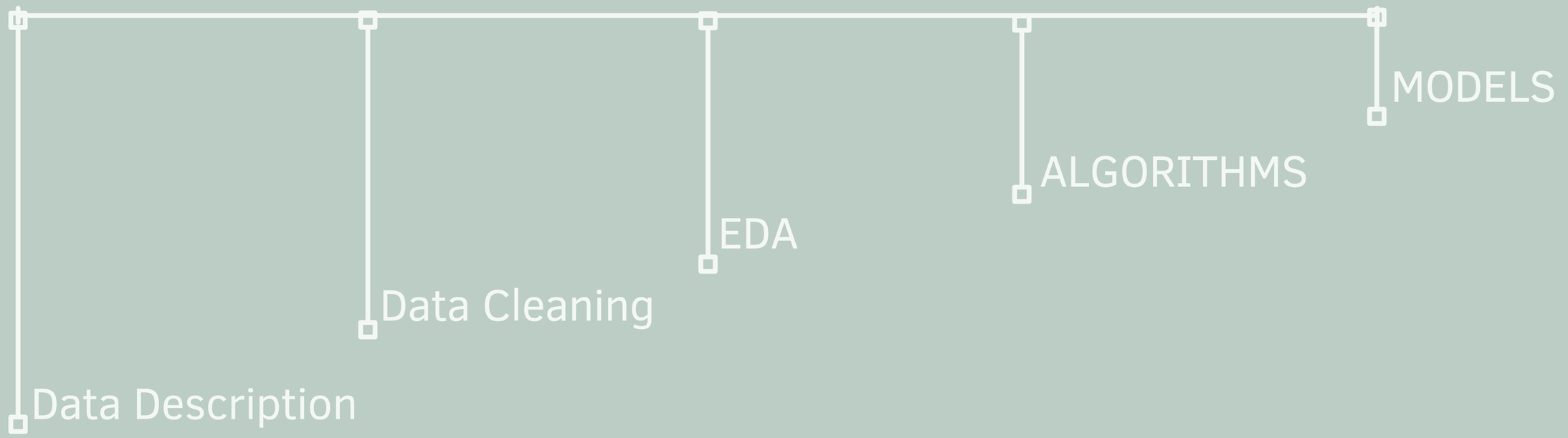
According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking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

Problem statement

In the given study, we have a binary classification problem.
We will make a prediction on the target variable - Strokes
Lastly we will build a variety of Classification models and compare the models giving the best prediction on strokes.

02

METHODOLOGY



1. Data Description

Dataset from KAGGLE.COM and contains 5110 observations × 12 feature

- id: unique identifier
- gender: "Male", "Female" or "Other“
- age: age of the patient
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- ever_married: "No" or "Yes“
- work_type: "children", "Govt_jov", "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

DATA CLEANING

Null values

We have 1 column with null value

01

Drop columns

Drop Id columns, So we have 11 columns

02

03

Outlier

We have two columns with outlier

04

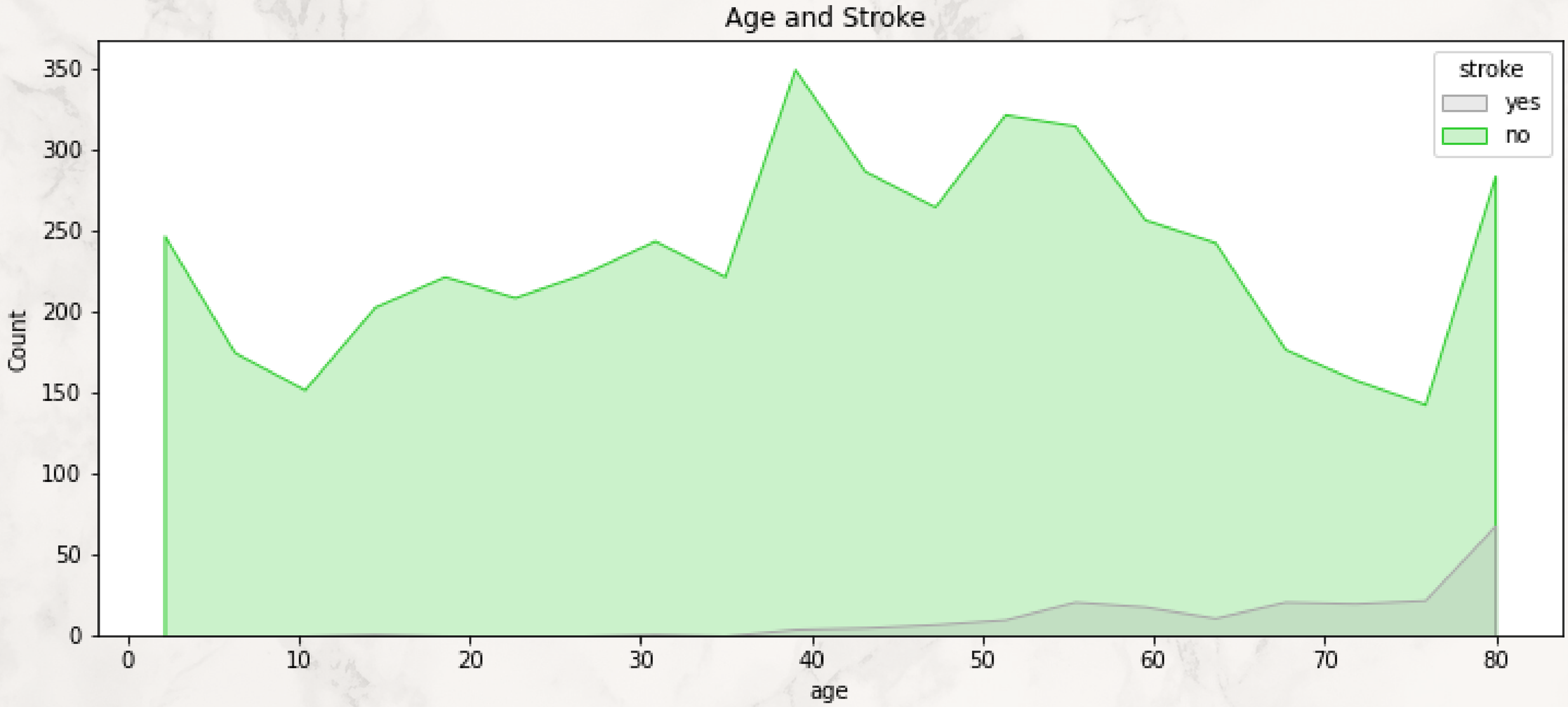
Replace

replace them by mean

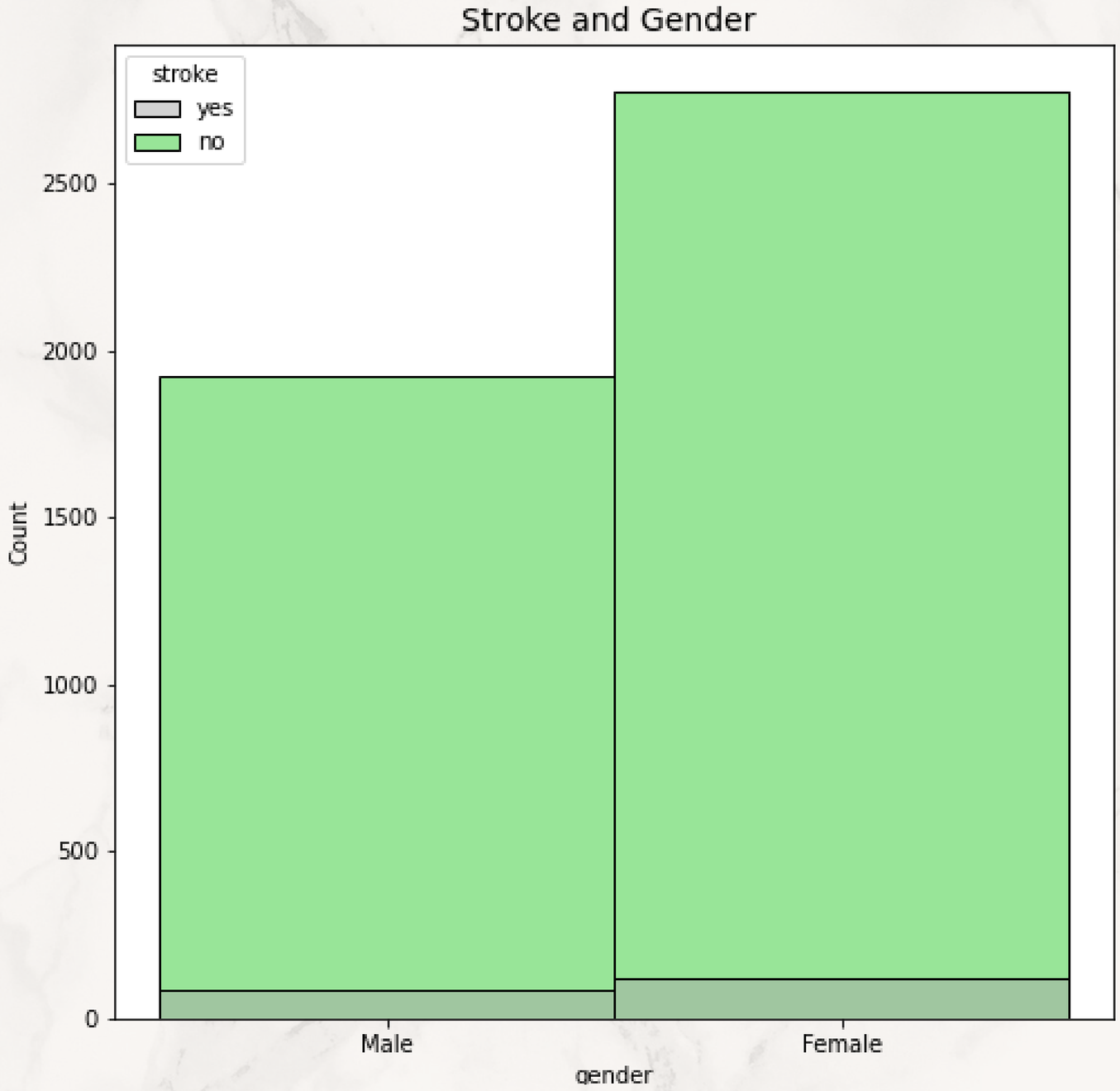
2.EDA

1. Is age a cause of stroke?
2. Who is more likely to have a stroke male or female?
3. Is smoking a cause of stroke?

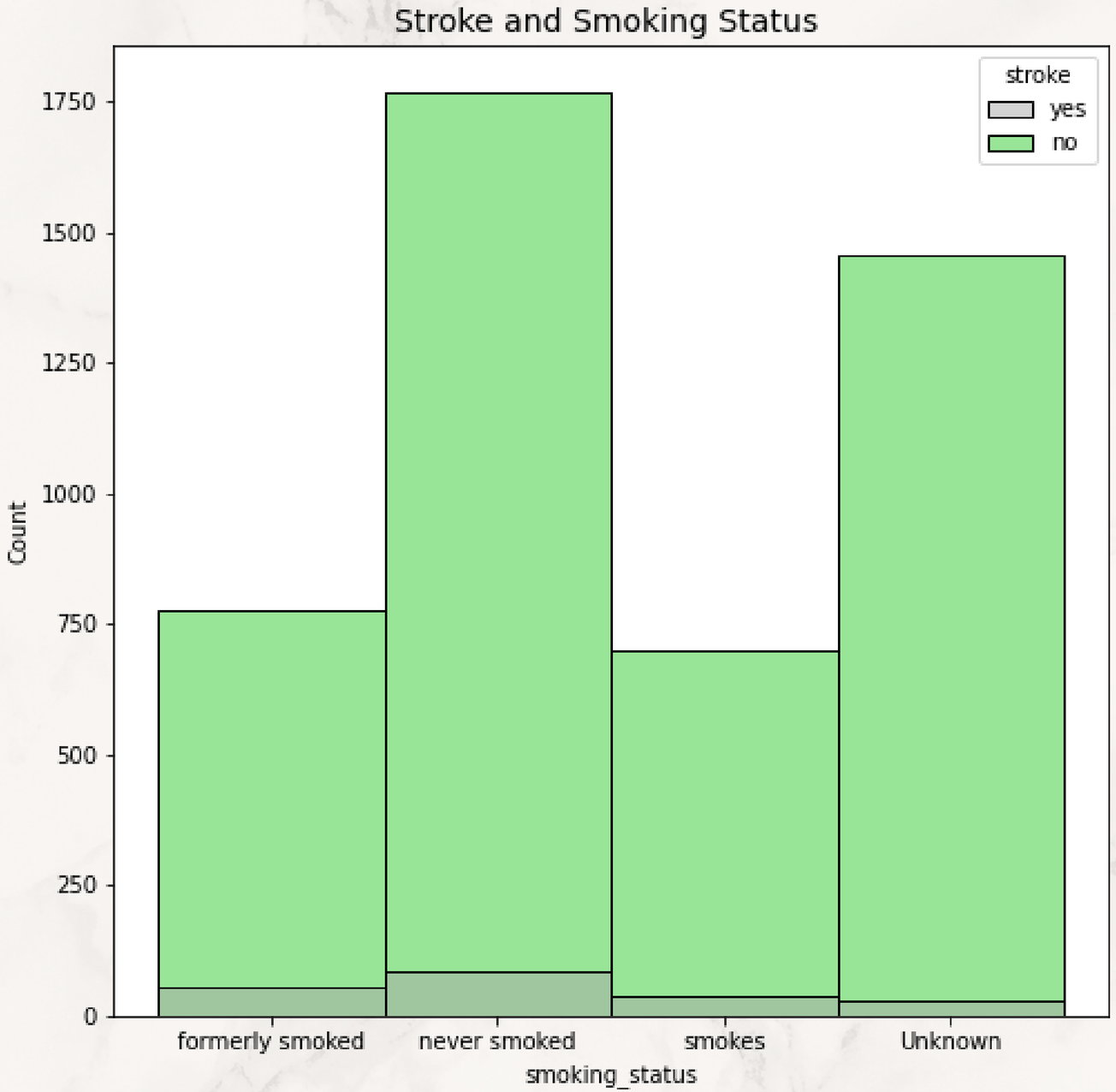
- Is age a cause of stroke?



- Who is more likely to have a stroke male or female?

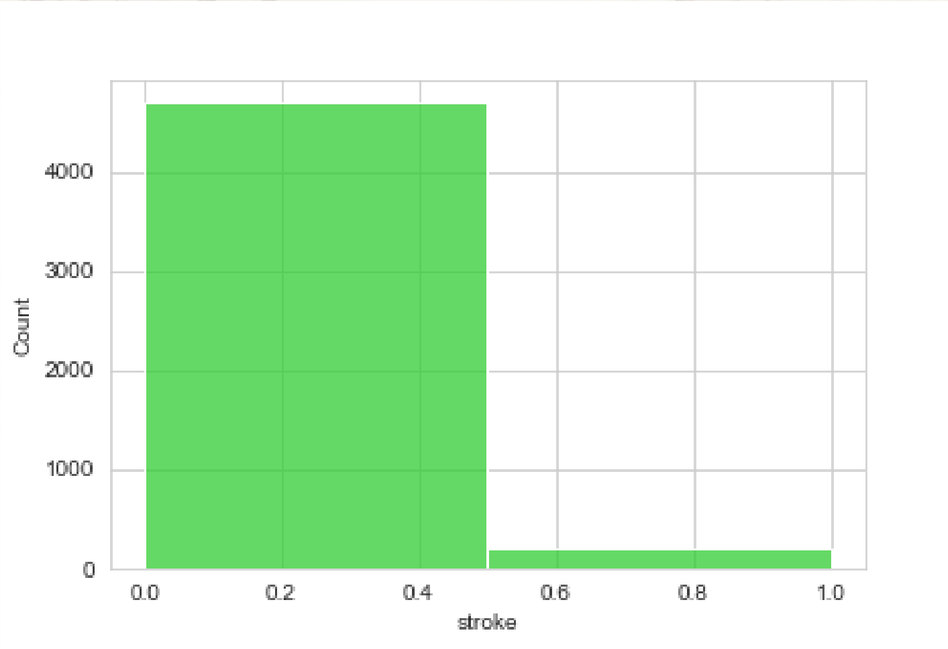


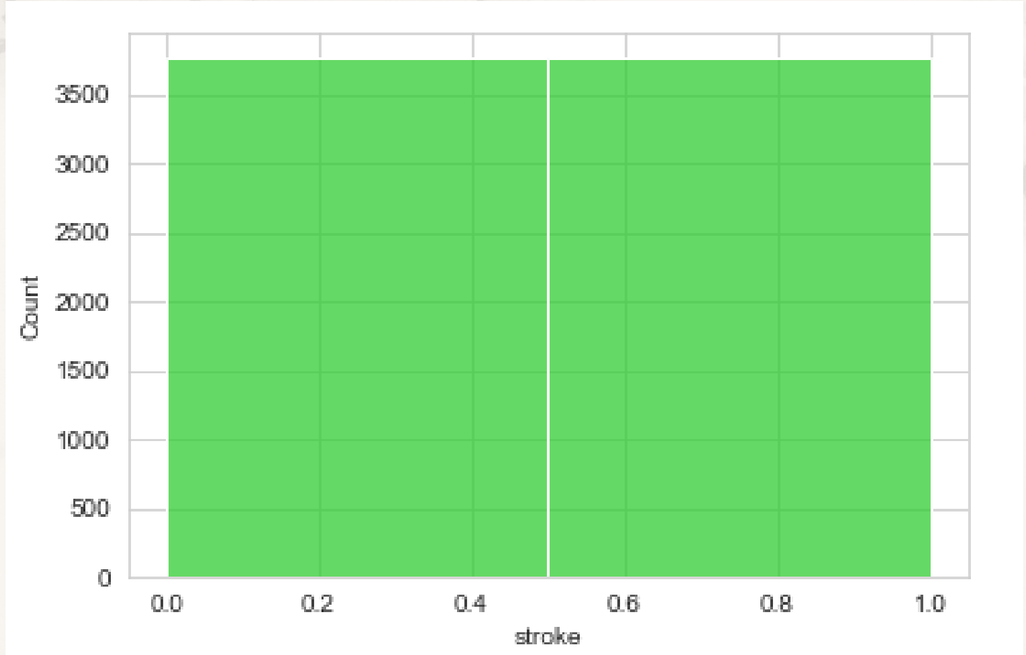
- Is smoking a cause of stroke?



SOLVED IMBLANCE PROBLEM :

RandomOverSampler





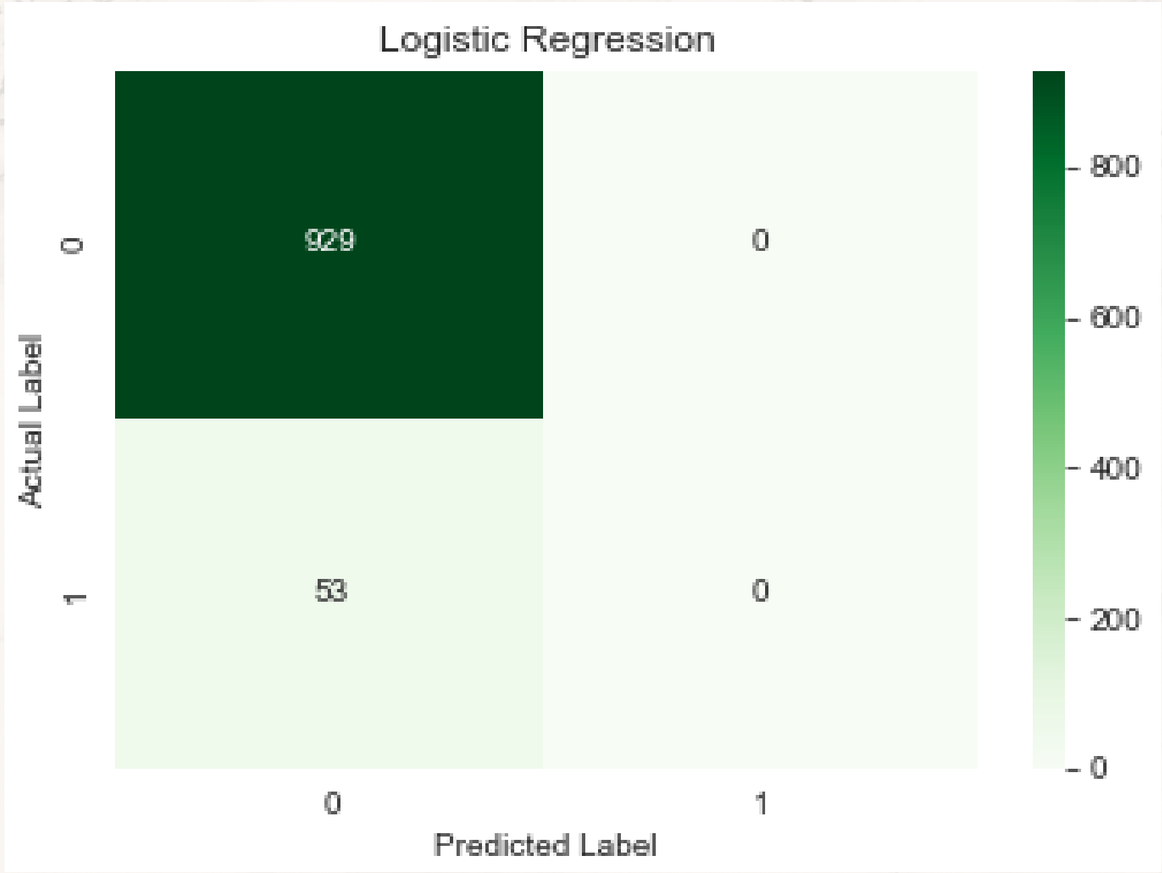
ALGORITHMS

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. K-Neighbors Classifier
5. Support Vector Classifier
6. XGB Classifier
7. Stacking

APPLY MODEL BEFORE SOLVE A IMBLANCE PROBLEM :

- Logistic Regression

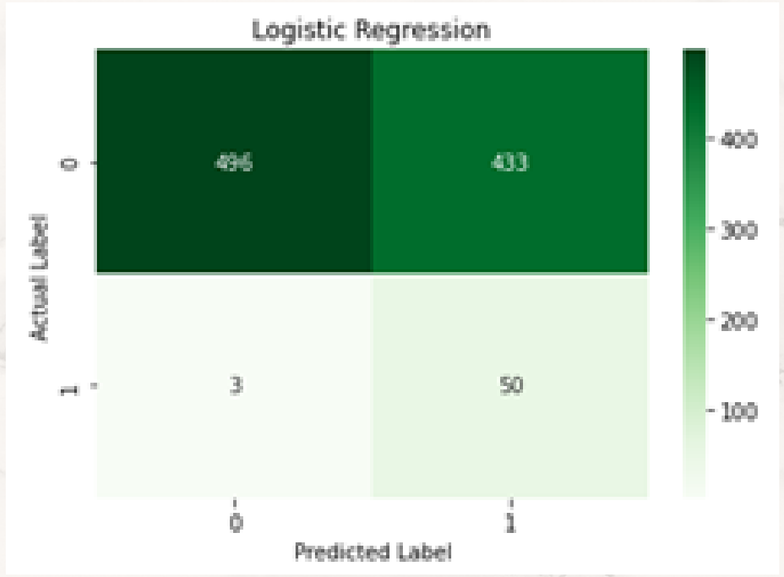
	precision	recall	f1-score
0	0.95	1.00	0.97
1	0.00	0.00	0.00



APPLY MODEL AFTER SOLVE A IMBLANCE PROBLEM :

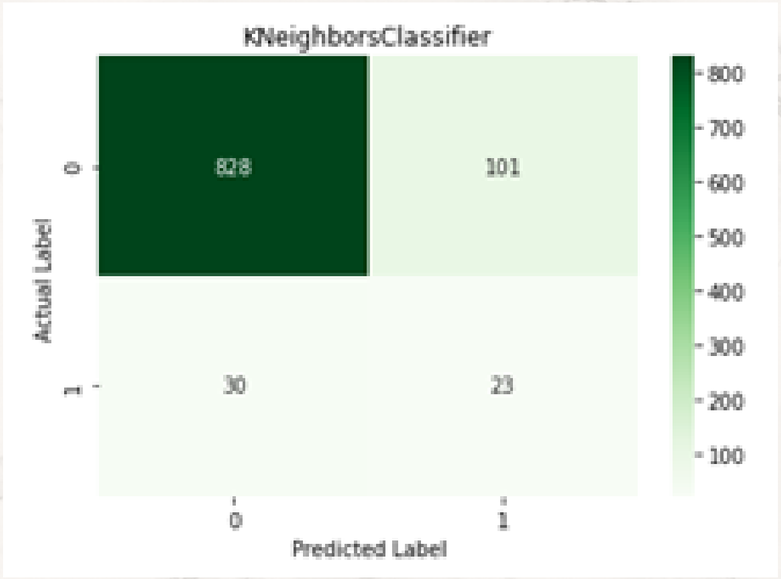
- **Logistic Regression**

	precision	recall	f1-score
0	0.99	0.53	0.69
1	0.10	0.94	0.19



- **K-Neighbors Classifier**

	precision	recall	f1-score
0	0.97	0.89	0.93
1	0.19	0.43	0.26

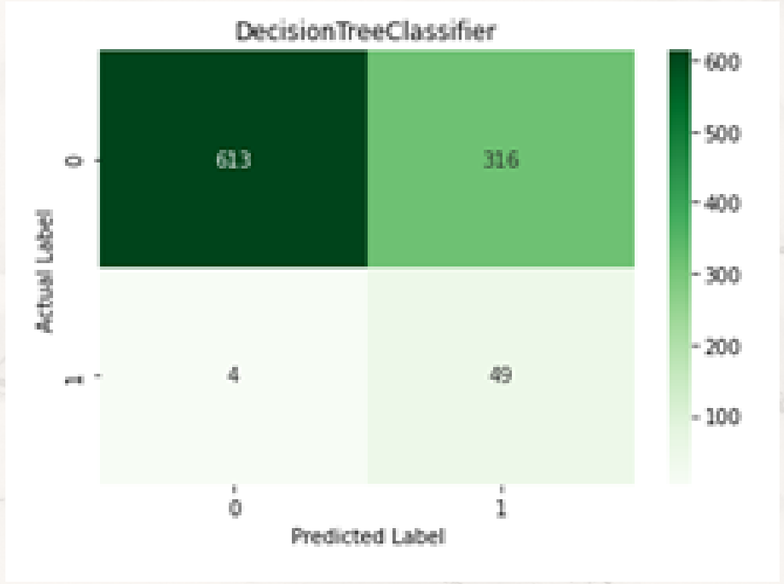


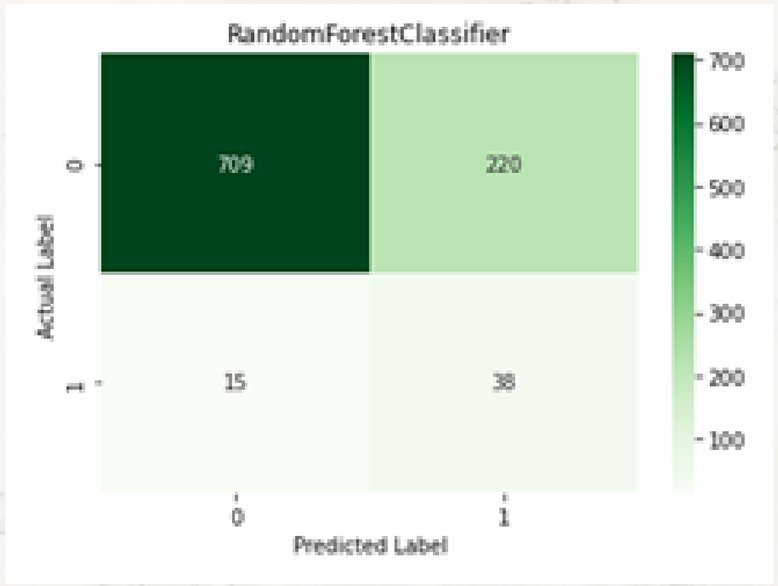
- Decision Tree Classifier

	precision	recall	f1-score
0	0.99	0.66	0.79
1	0.13	0.92	0.23

- Random Forest Classifier

	precision	recall	f1-score
0	0.98	0.76	0.86
1	0.15	0.72	0.24



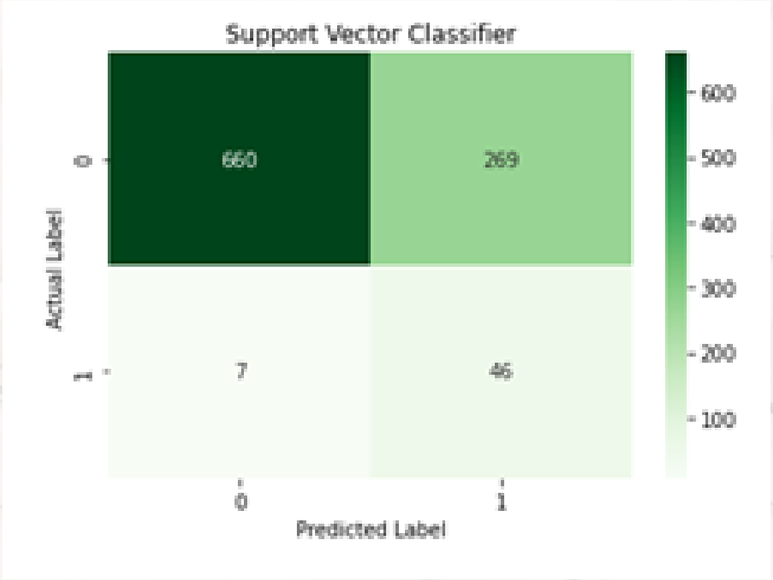


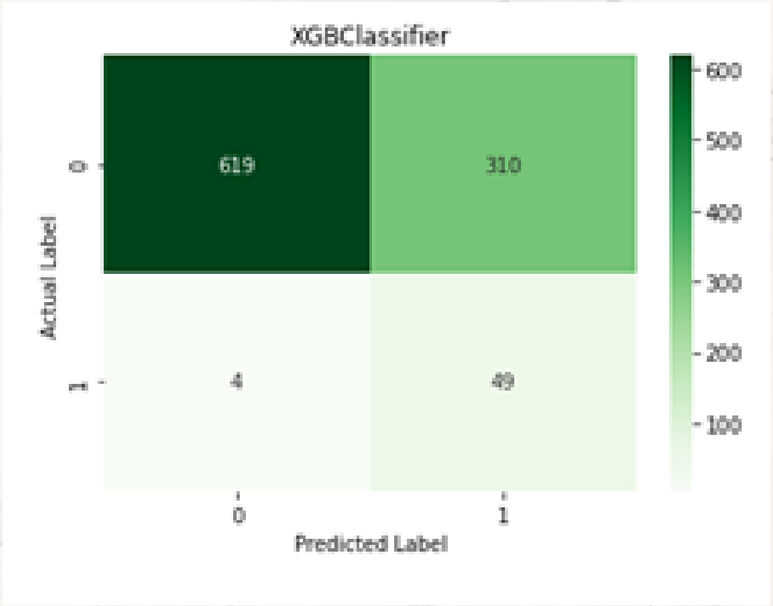
- Support Vector Classifier

	precision	recall	f1-score
0	0.99	0.71	0.83
1	0.15	0.87	0.25

- XGB Classifier

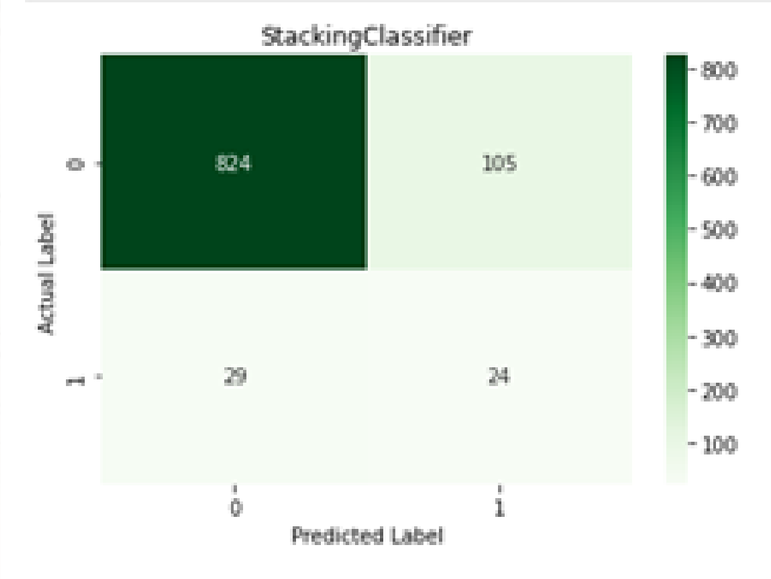
	precision	recall	f1-score
0	0.99	0.67	0.80
1	0.14	0.92	0.24





- Stacking

	precision	recall	f1-score
0	0.97	0.89	0.92
1	0.19	0.45	0.26



THE BEST MODEL

Logistic Regression

03 TOOLS

- Technologies

Jupyter Notebook , Python

- Libraries

Pandas , Numpy , Matplotlibe , Seaborn , Sklearn , patsy

- Canva

Presentation the project

04 CONCLUSION

We have made 6 models that predicted the whether patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. The model will help the hospitals to the Discovering a stroke. and we have achieve a very good accuracy in the best model.

THANKS