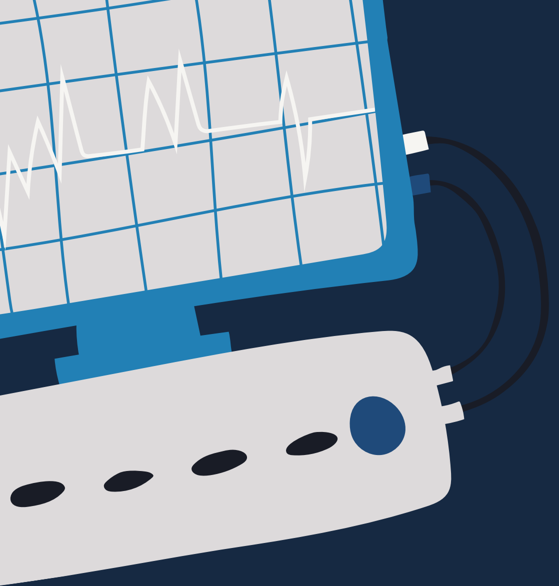
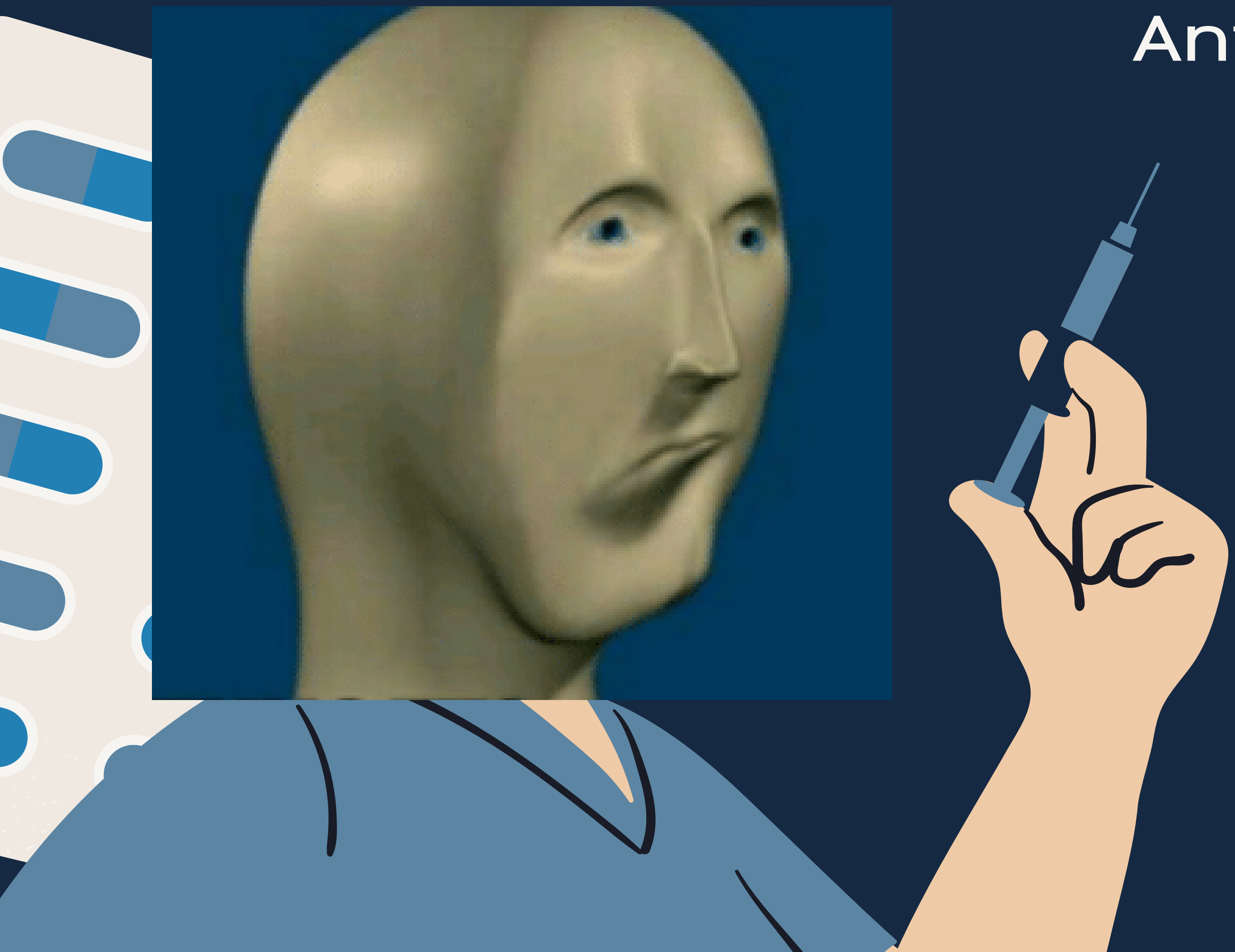


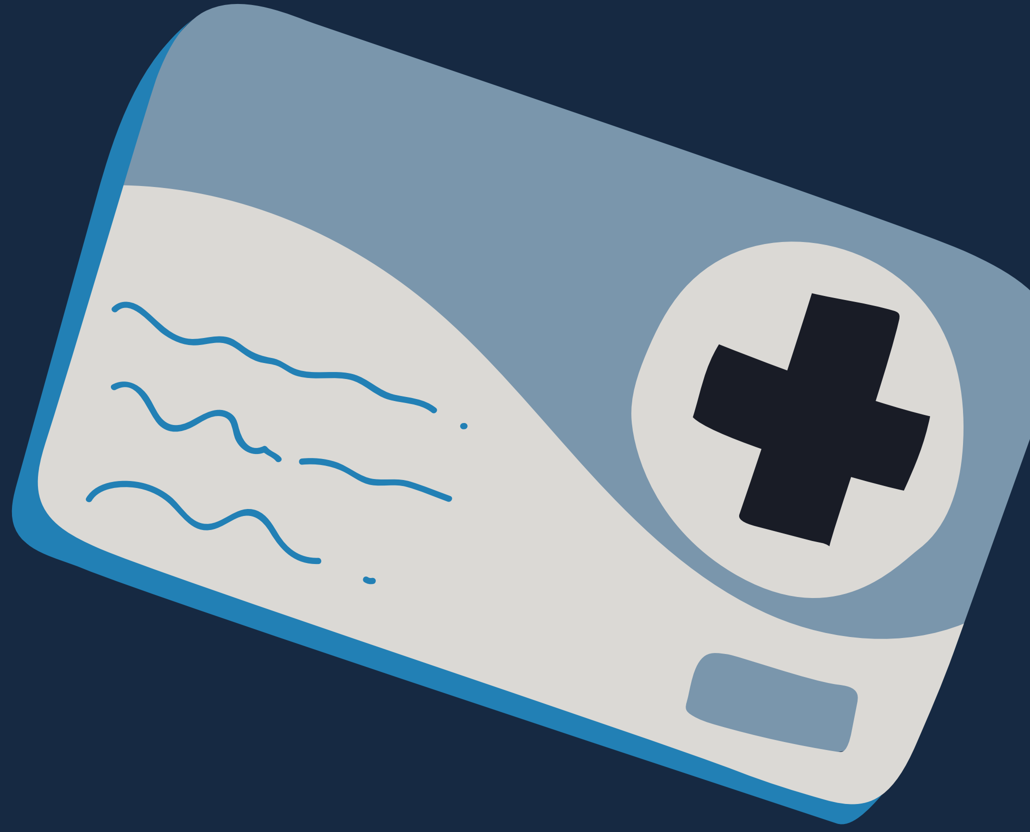
Stroke Prediction Model: Let's keep a straight face!

Anticipate stroke before it happens!



Presented by BCF 1 Group 7:
Si Ming Zhou (U2120609K)
Jeremy Lim Yih Shih (U2122106C)
Siah Wee hung (U2121064J)

Problem Definition



Dataset & Motivation

We used the Stroke Prediction Dataset from Kaggle with the aimed to accurately classify if an individual would have a stroke based on common factors among stroke patients

Real-World Problem

Meaningful and Impactful in a real life context

Exciting challenge but not overwhelming - familiar factors

Apply our knowledge and test our understandings to real-world problems

Project Pipeline

01 Import Kaggle dataset

02 Exploratory Data Analysis (EDA)

03 Data Preprocessing

04 Modelling

05 Recommendations

EDA: Understanding the dataset

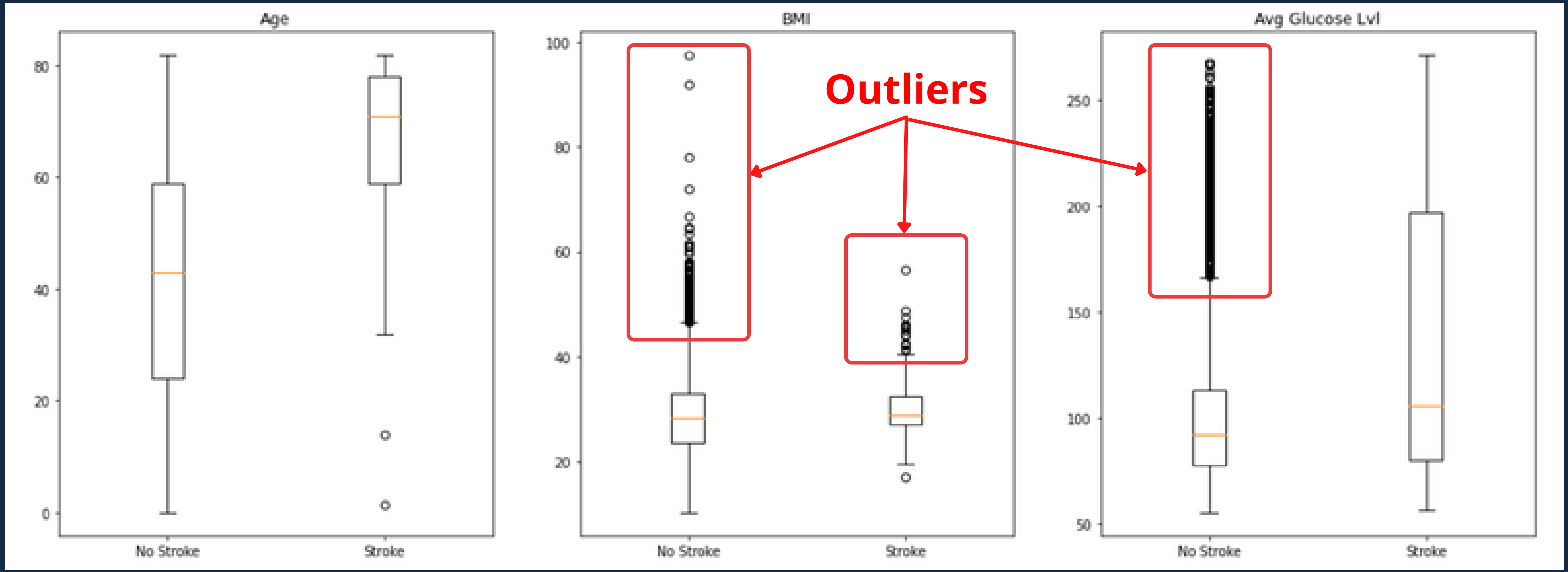
```
[ ] df.head()
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Numeric Variables	**Categorical variables**
age	ever_married
avg_glucose_level	work_type
bmi	Residence_type
	smoking_status

EDA: Numerical Variables

Box plots for age, bmi, avg_glucose_level:

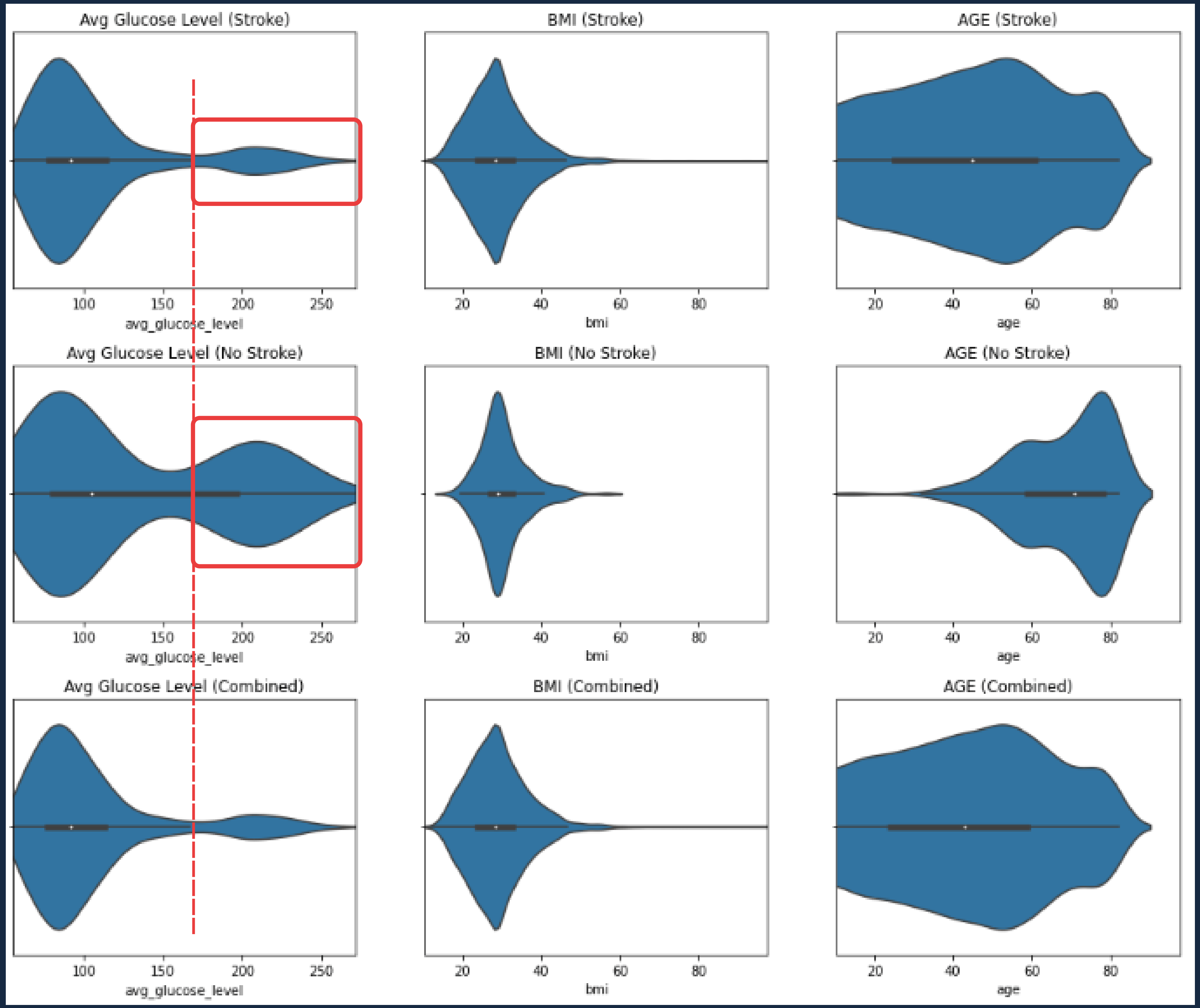


Observation:

For bmi and avg_glucose_level, there is a considerable number of anomalies.

EDA: Numerical Variables

Violin plots for age, bmi, avg_glucose_level:

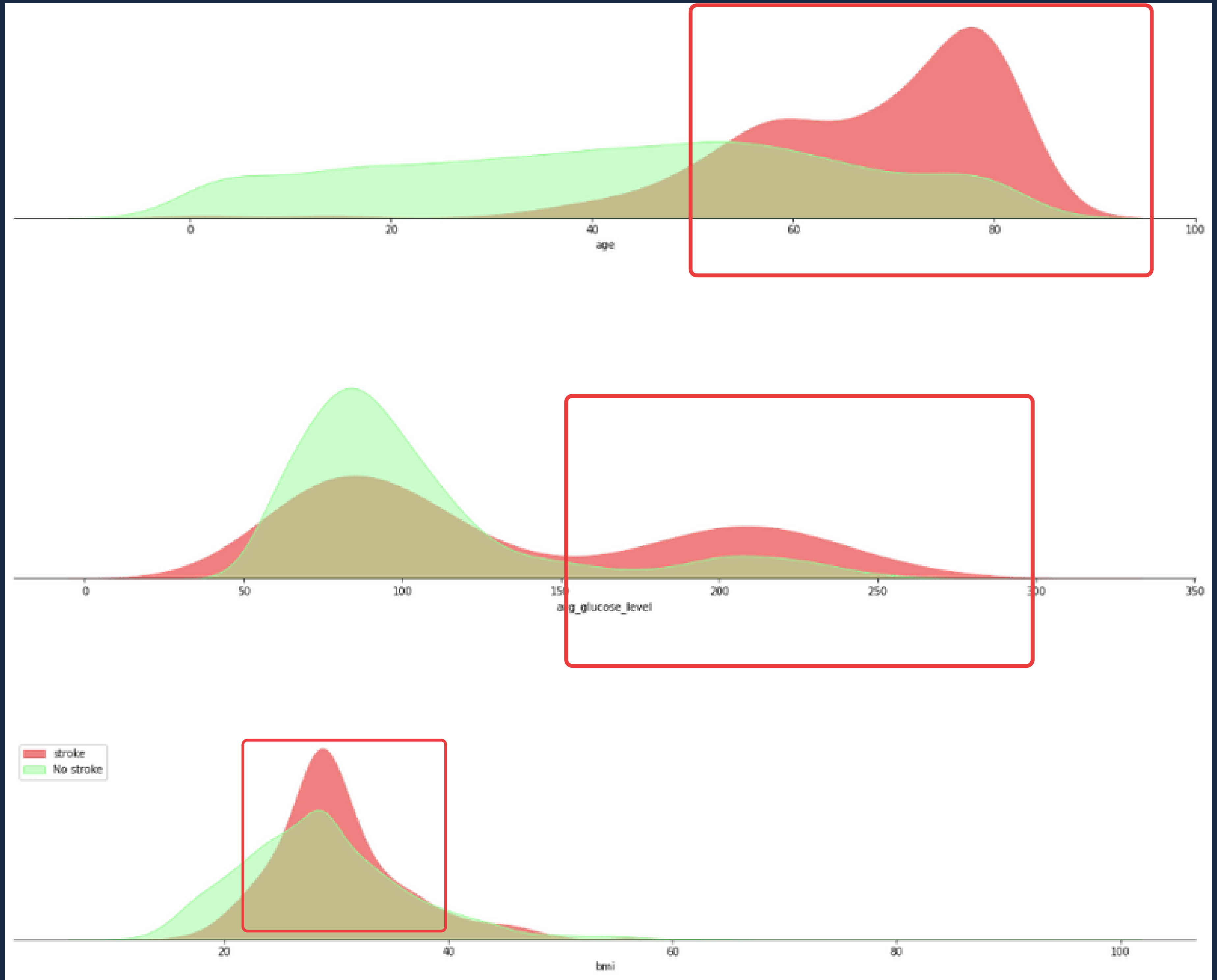


Observation:

- For avg_glucose_level, a large portion of people with stroke have higher average glucose level (larger lump on the right)
- Should not remove anomalies as it will remove significant portion of data points of people with stroke

EDA: Numerical Variables

KDE plots for age, bmi, avg_glucose_level:



Observation:

Age:
Red graph (Stroke) is right skewed compared to green graph (No stroke)
⮡ Older people are more likely to suffer from stroke

Avg_glucose_level:
Bump on red graph (stroke) peaks higher from 150-300 mg/dL
⮡ People with higher glucose level more likely to suffer from stroke

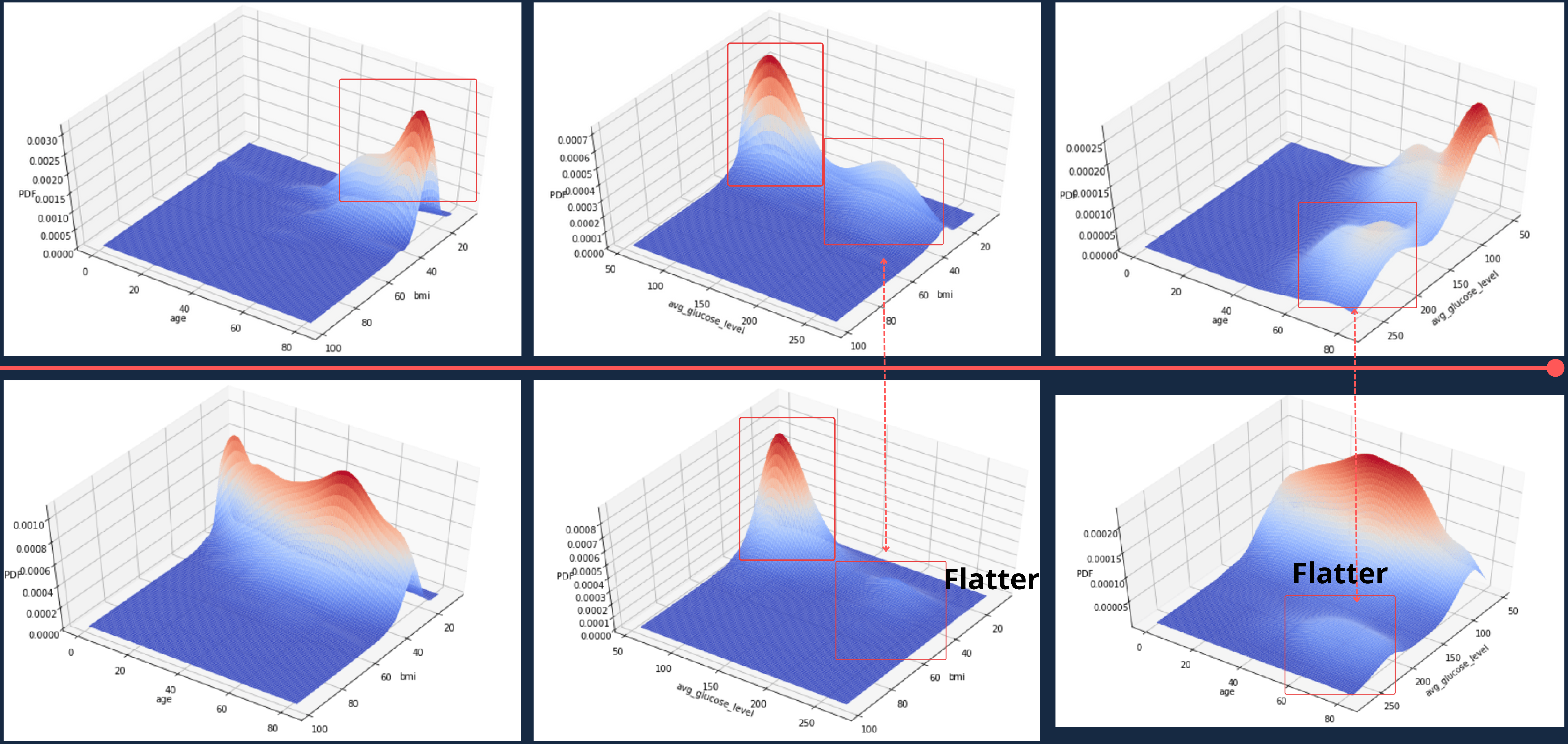
BMI:
Red graph (stroke) is slightly more right-skewed
⮡ People with stroke have a slightly higher bmi

EDA: Numerical Variables

2D-KDE plots for age, bmi, avg_glucose_level:

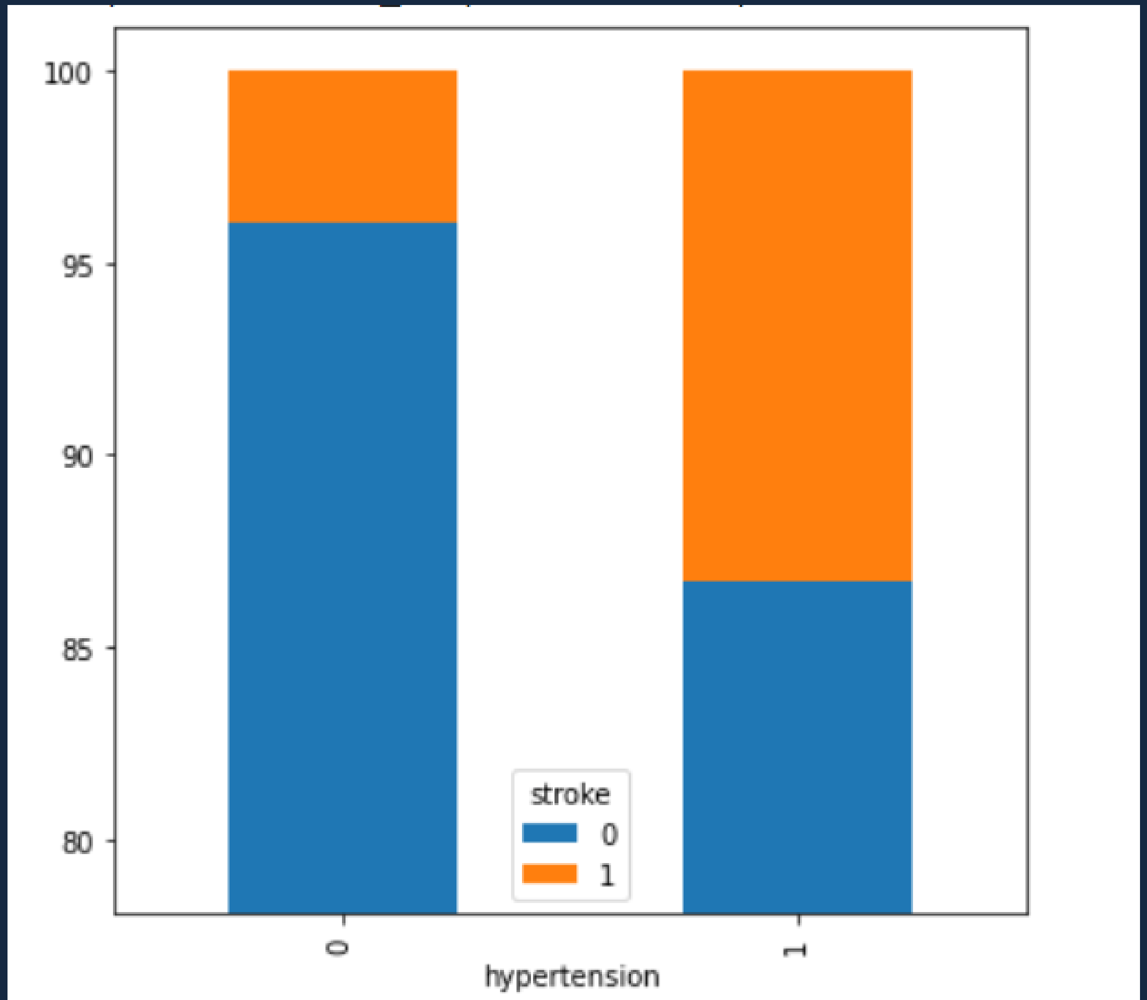
stroke

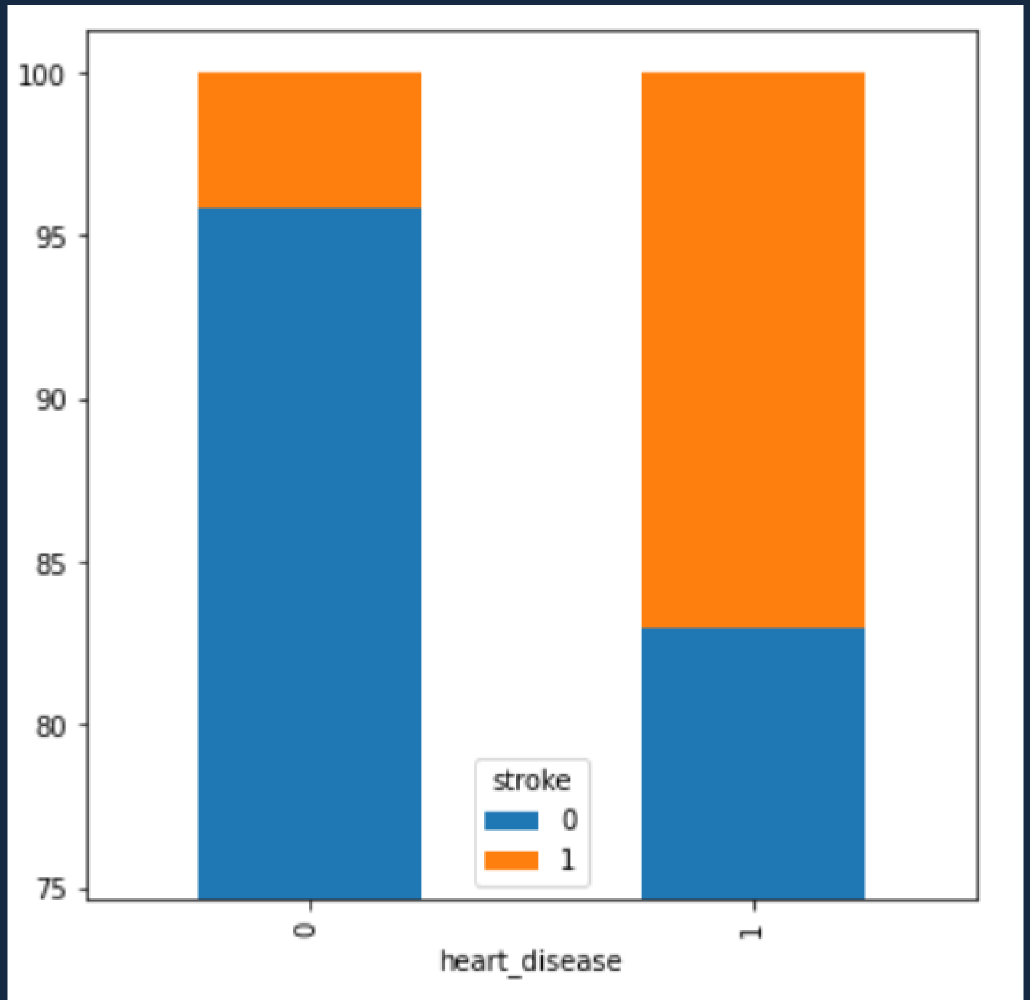
no stroke



EDA: Categorical Variables

Bar chart for hypertension and heart disease:





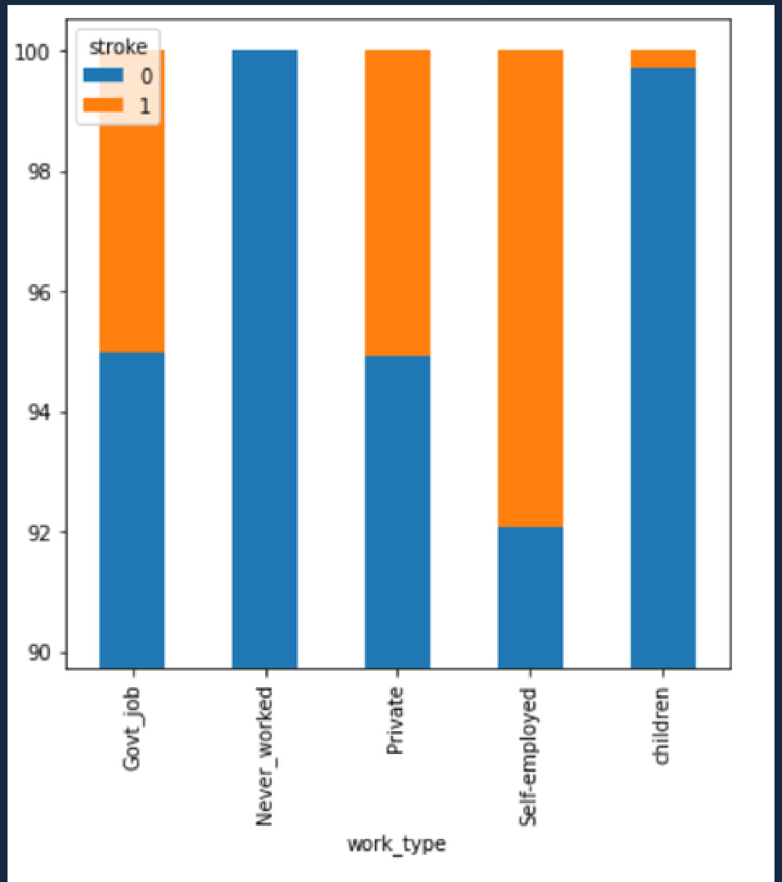
Observation:

The orangs portion (stroke) when they suffer from hypertension and heart disease (==1) is larger

⮑ people with hypertesnion and heart_disease are more likely to have stroke

EDA: Categorical Variables

Bar chart for work type:



	age	
	count	mean
work_type		
Govt_job	657	50.879756
Never_worked	22	16.181818
Private	2925	45.503932
Self-employed	819	60.201465
children	687	6.841339

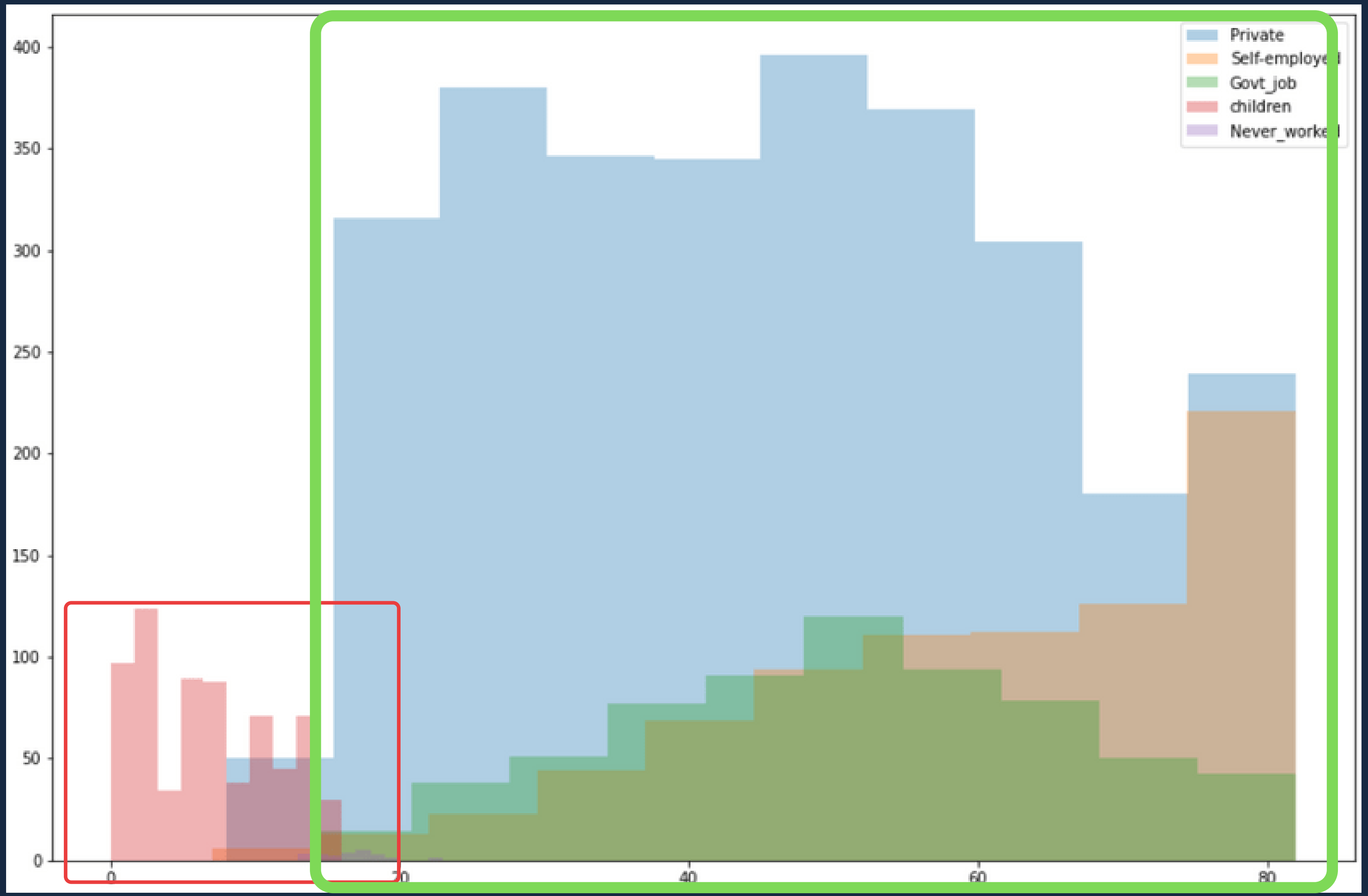
Observation:

For work_type,

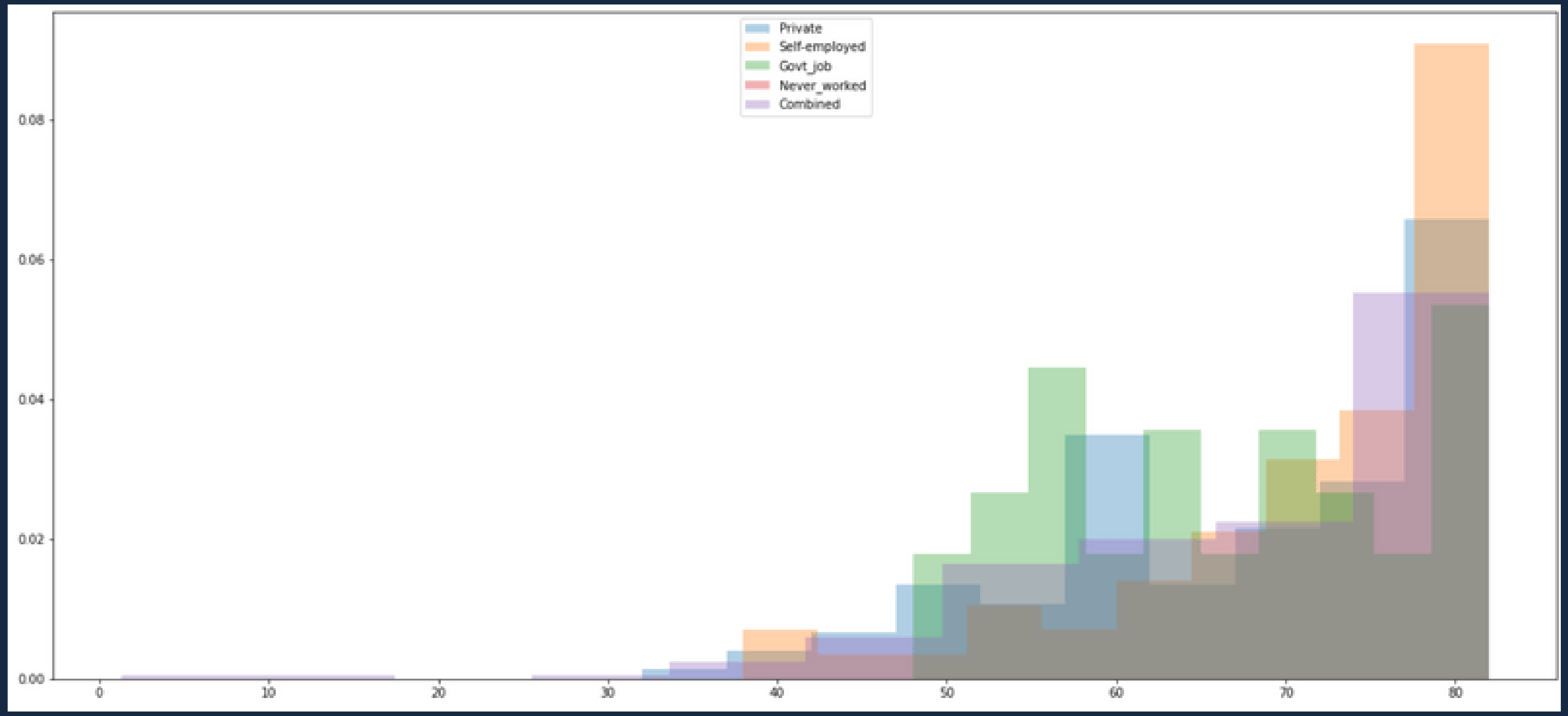
- govt workers, private industry workers and self-employed people are more likely to get a stroke, compared to children and those who never worked
- People who work tend to be of higher age

EDA: Categorical Variables

Age vs work type:



Age distribution for work_type

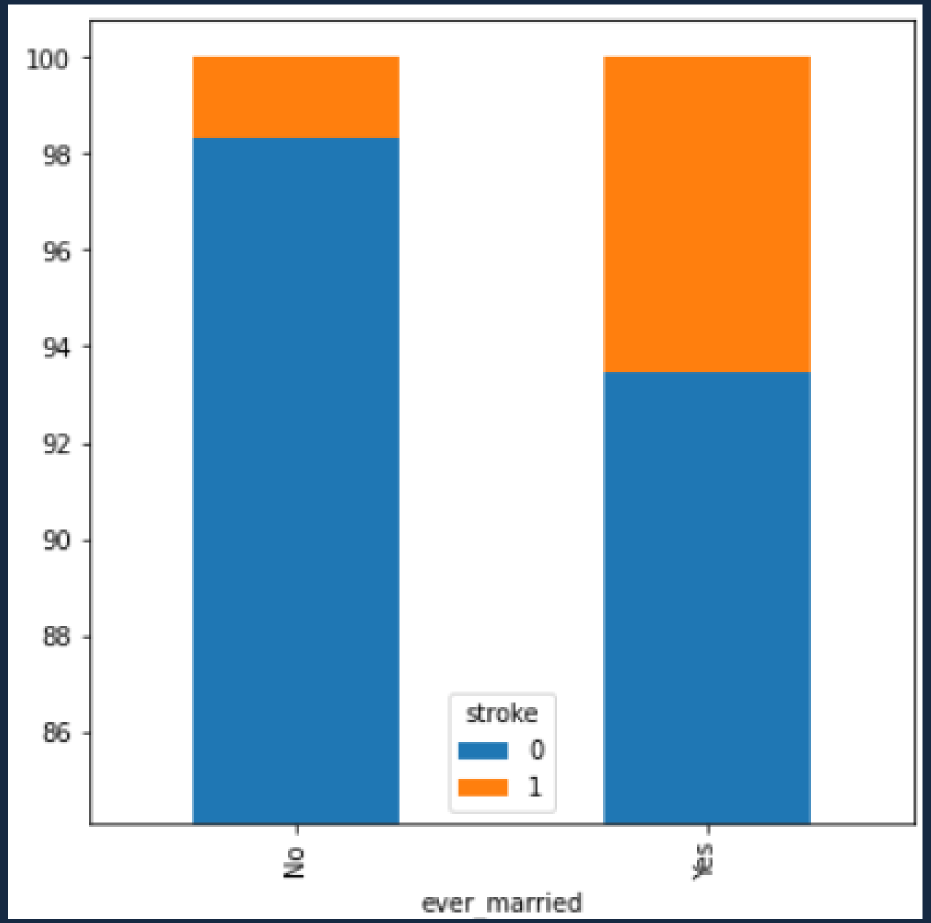


Age distribution for work_type

Generally, for all work_type, chances of getting a stroke increases exponentially with age

EDA: Categorical Variables

Bar chart for ever-married:



	age	
	count	mean
ever_married		
No	1757	22.014229
Yes	3353	54.342082

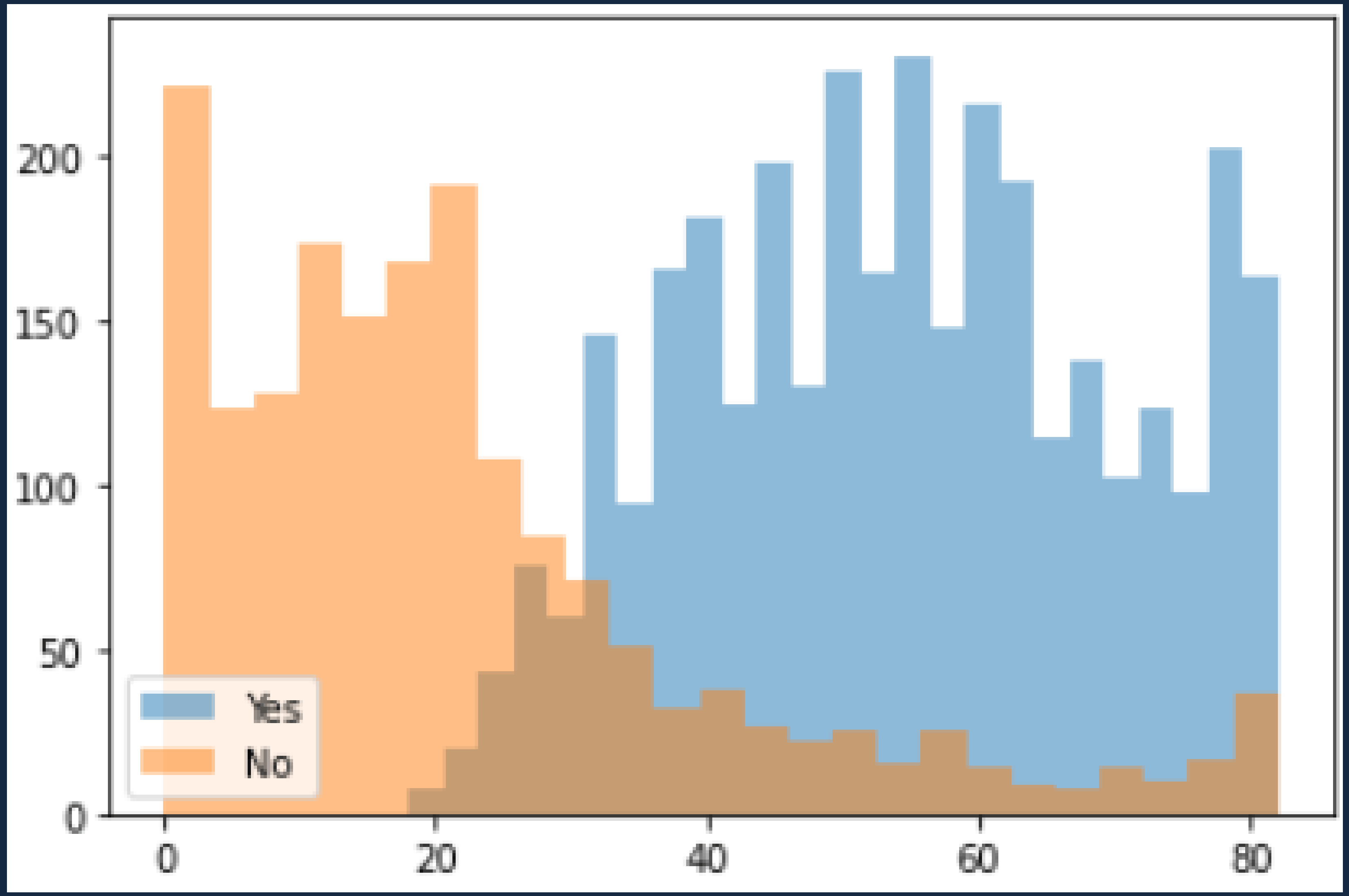
Observation:

For marriage status,

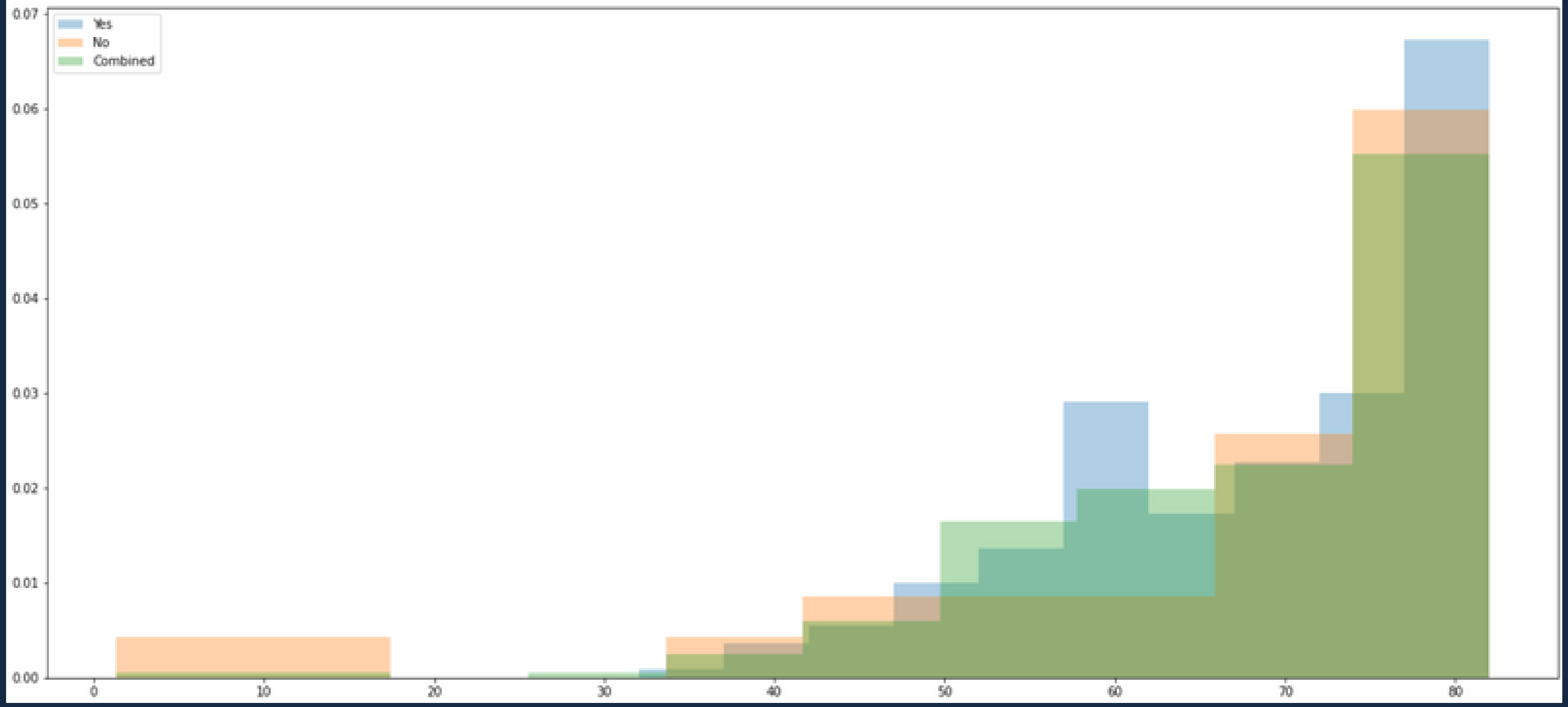
- People who are married before or currently are more likely to suffer from stroke
- They tend to have higher age

EDA: Categorical Variables

Age vs ever-married:



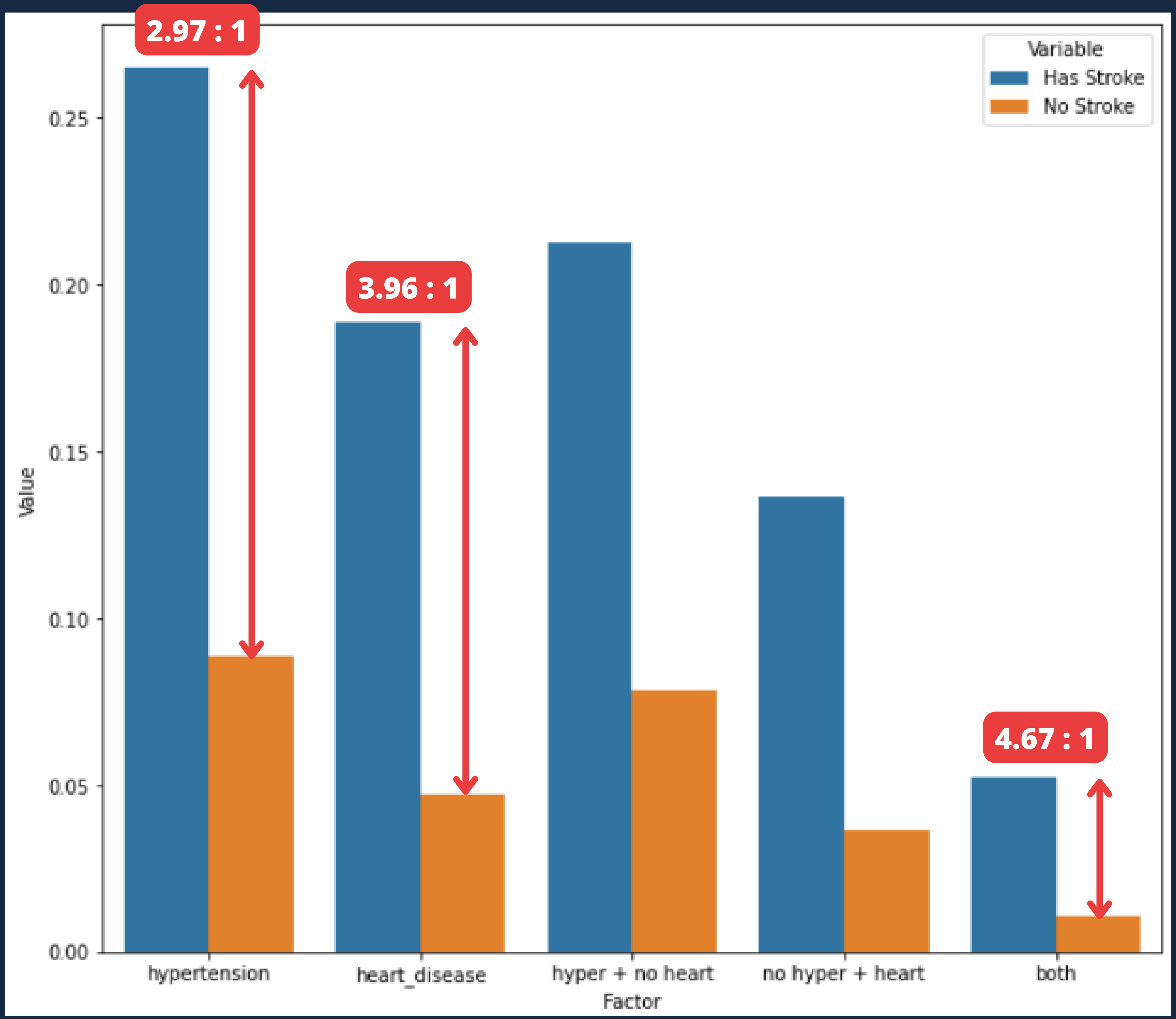
Age distribution for ever_married



Age distribution of stroke patients for ever_married

Regardless of whether person is married or not, chances of getting a stroke increases exponentially with age

EDA: Understanding the dataset

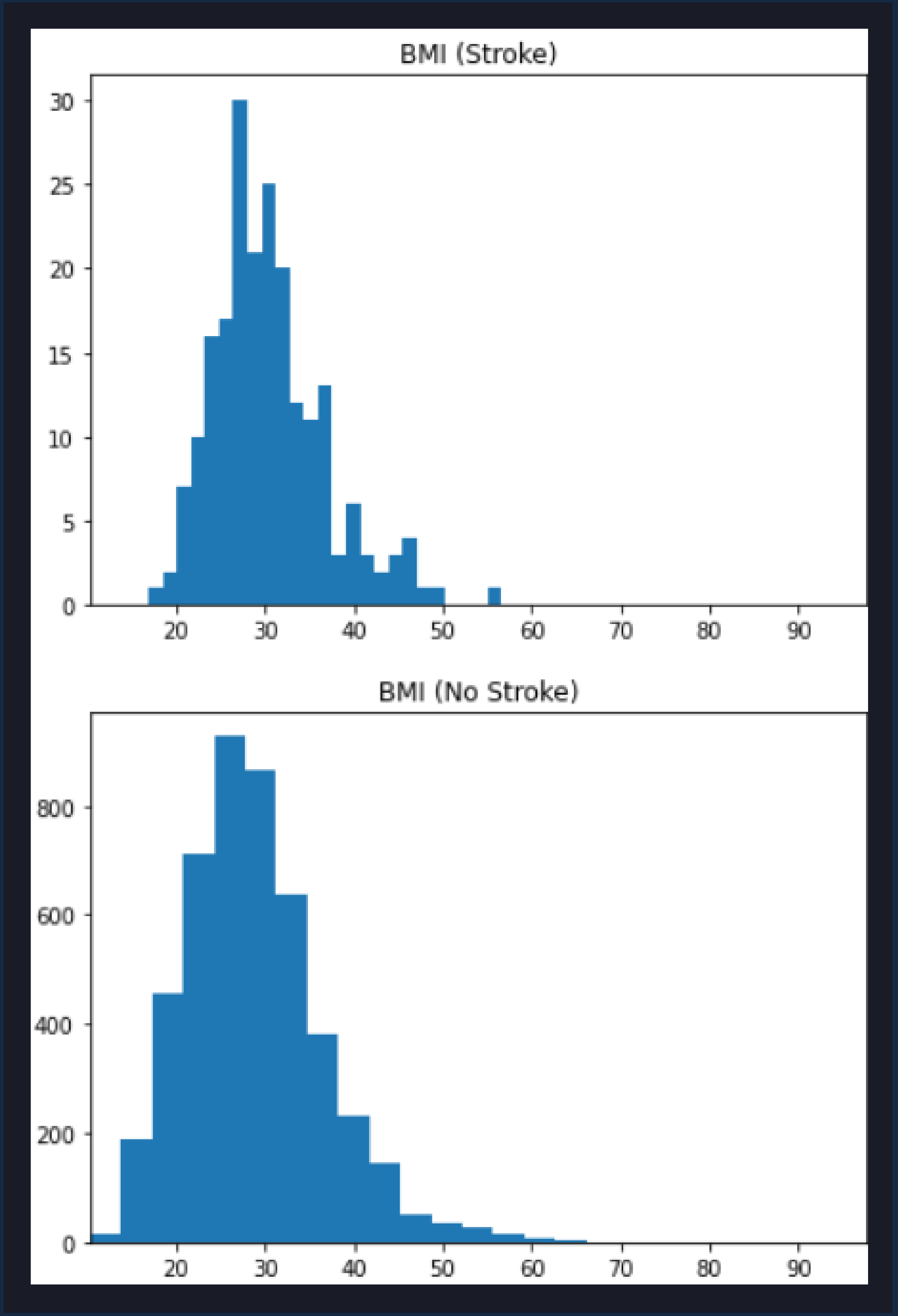


Hypertension + Heart Disease

DATA PREPROCESSING

#	Column	Non-Null Count	Dtype
0	id	5110 non-null	int64
1	gender	5110 non-null	object
2	age	5110 non-null	float64
3	hypertension	5110 non-null	int64
4	heart_disease	5110 non-null	int64
5	ever_married	5110 non-null	object
6	work_type	5110 non-null	object
7	Residence_type	5110 non-null	object
8	avg_glucose_level	5110 non-null	float64
9	bmi	4909 non-null	float64
10	smoking_status	5110 non-null	object
11	stroke	5110 non-null	int64

201 missing values

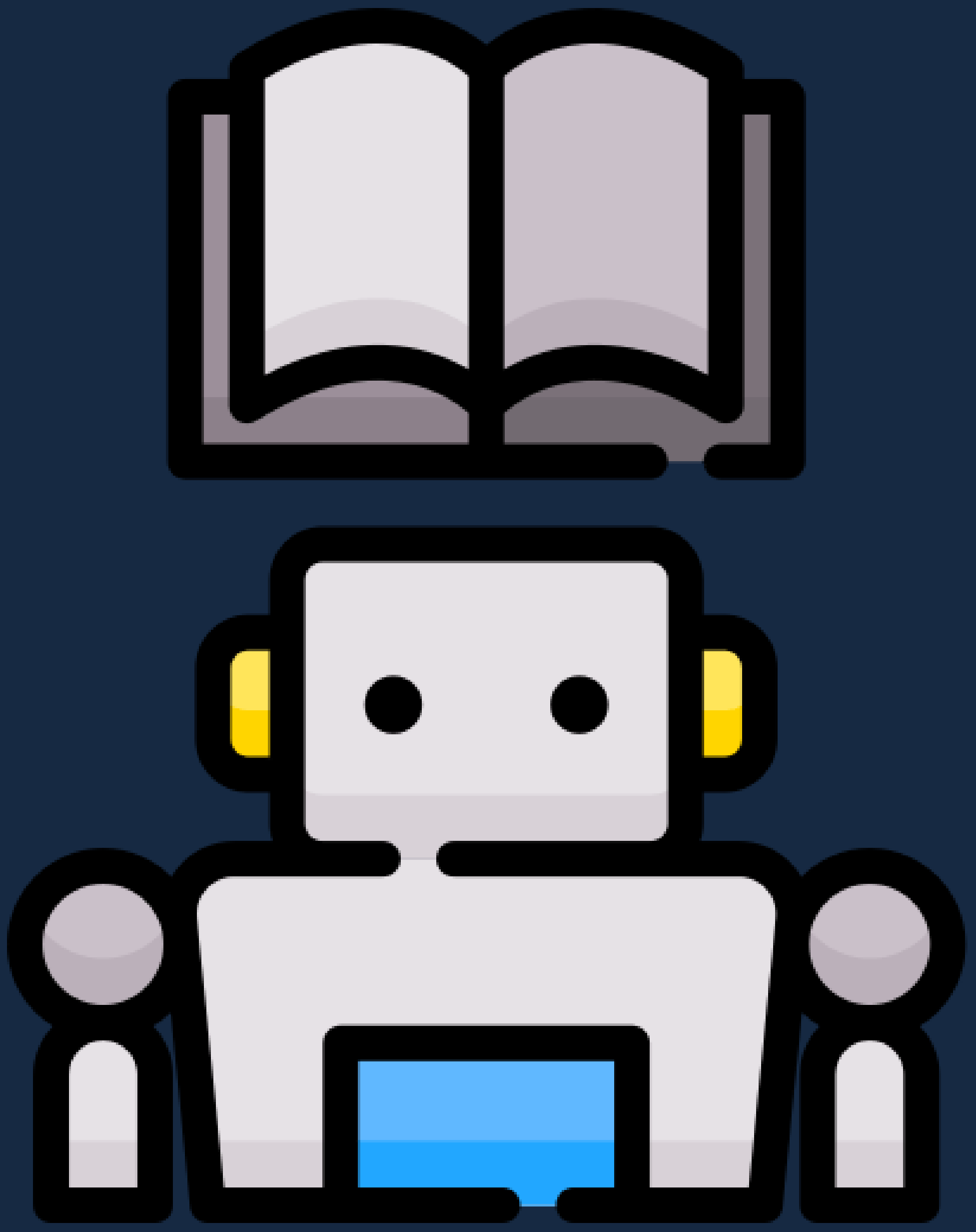


DATA PREPROCESSING

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

smoking_status_formerly smoked	smoking_status_never smoked	smoking_status_smokes	smoking_status_Unknown
1	0	0	0
0	1	0	0
0	1	0	0
0	0	1	0

ONE-HOT ENCODING



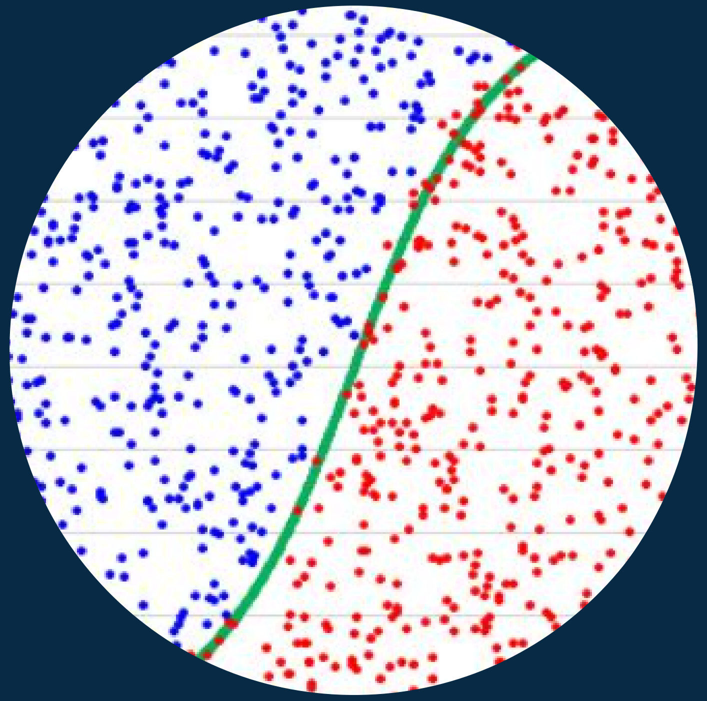
Machine Learning Models

<u>Classification Machine Learning Models</u>

- predicting whether subsequent data would fall into pre-determined categories - Stroke vs No Stroke

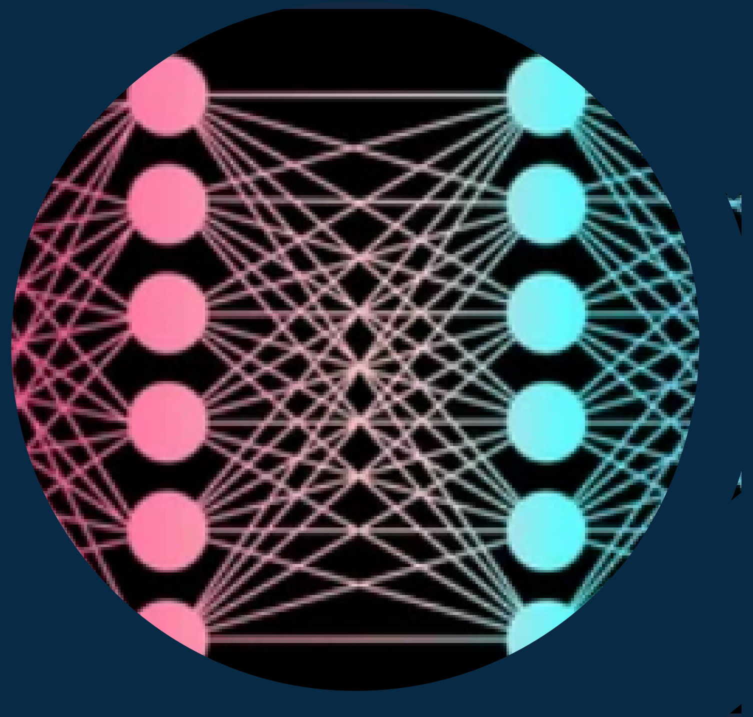
<u>Implemented 4 Models</u>

- Artificial Neural Networks
- XGBoost
- Random Forest
- Logistics Regression



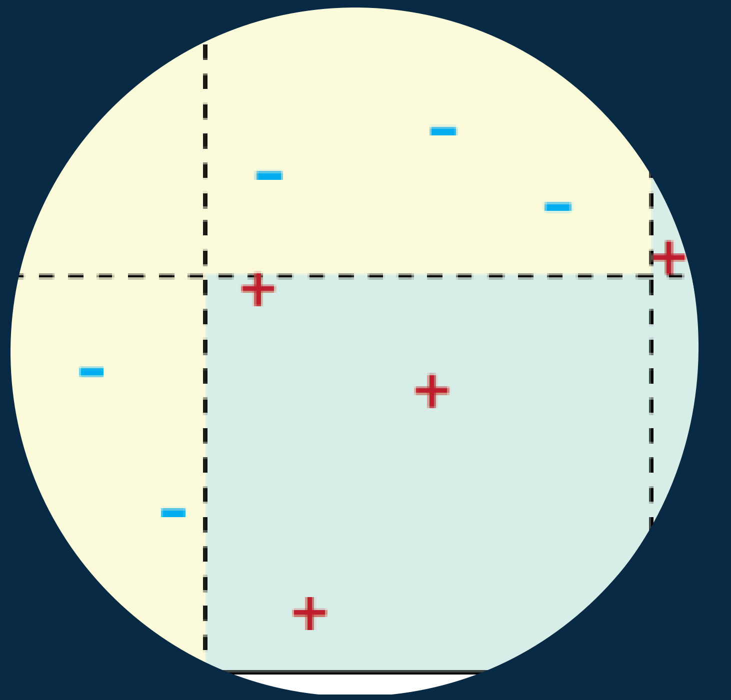
Logistic Regression

- Regression model
- Outputs probability of getting stroke

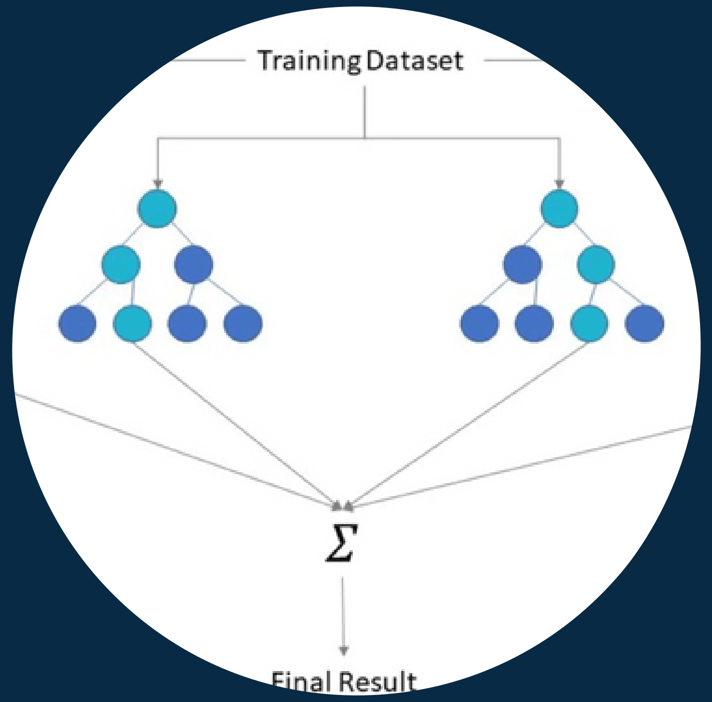


Artificial Neural Network

- Versatile
- Outputs probability of getting stroke



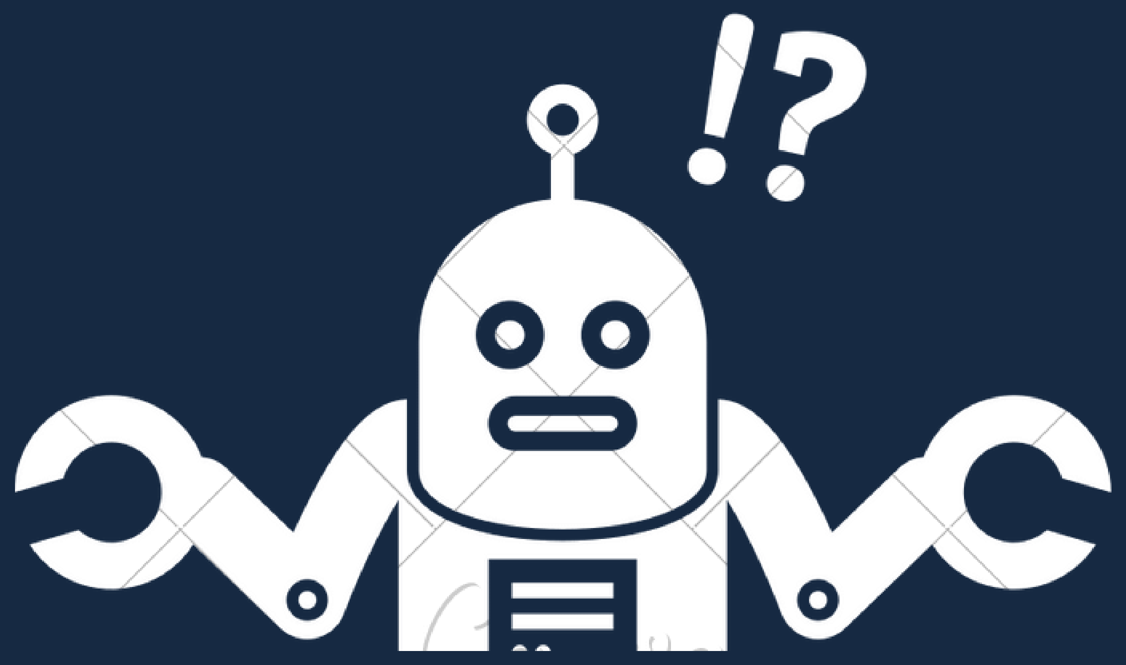
XGBoost



Random Forest

- Classification models
- Decision Trees

Machine Learning Models



Why 4 models?

- Each model utilised different algorithms and concepts to classify
- Allows for comparison and determine which model would best fit our dataset.

Model Evaluation

Why is using only accuracy as our metric not desirable?

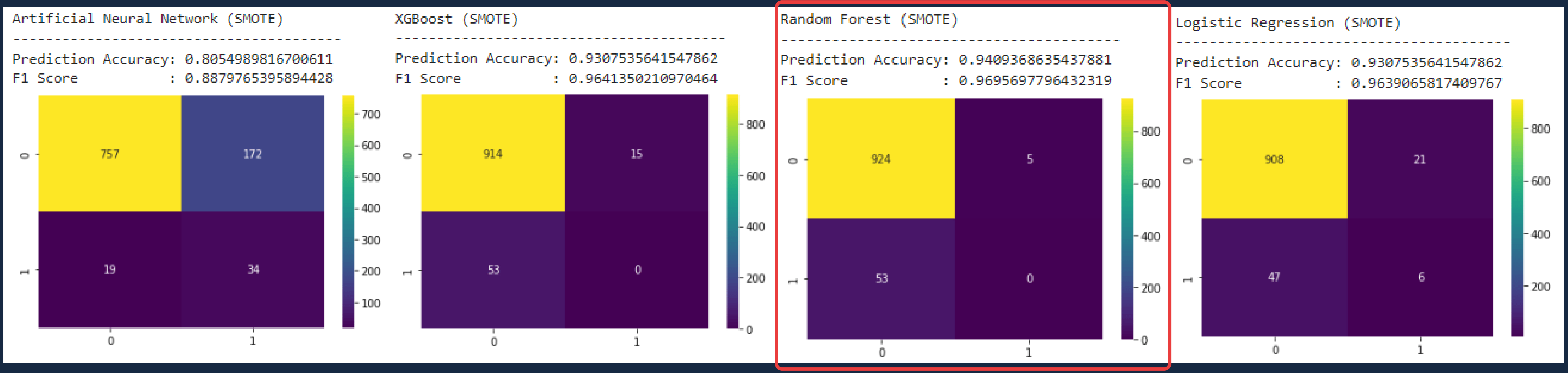
- Highly Imbalanced Dataset - 95% No Stroke, 5% Stroke
- Misleading high accuracy if model predicts all no stroke

- Overcome by applying Synthetic Minority Oversampling Technique (SMOTE)
- Place greater weight onto smaller class (Stroke)

Metrics for Model Evaluation

- Accuracy
- F1 Score

Models Evaluation



Best Model:
Random Forest
- Best accuracy and F1 score

Project Outcome

Able to attain a relatively high classification accuracy and F1 score, achieving our original aim of stroke prediction

Recommendations

Insight: High accuracy despite using unconventional data
Suggestion: Explore other indirect variables to use in conjunction with traditional medical data

Insight: Missing some crucial data (family history of stroke) or incomplete data (null values for BMI and smoking_status)
Suggestion: Seek alternative data to form complete dataset that can improve accuracy

Insight: A binary classifcation (stroke or no stroke) may not be useful
Suggestion: Determine probability of having a stroke instead of a black-or-white classification of having a stroke or not

THANK YOU!