

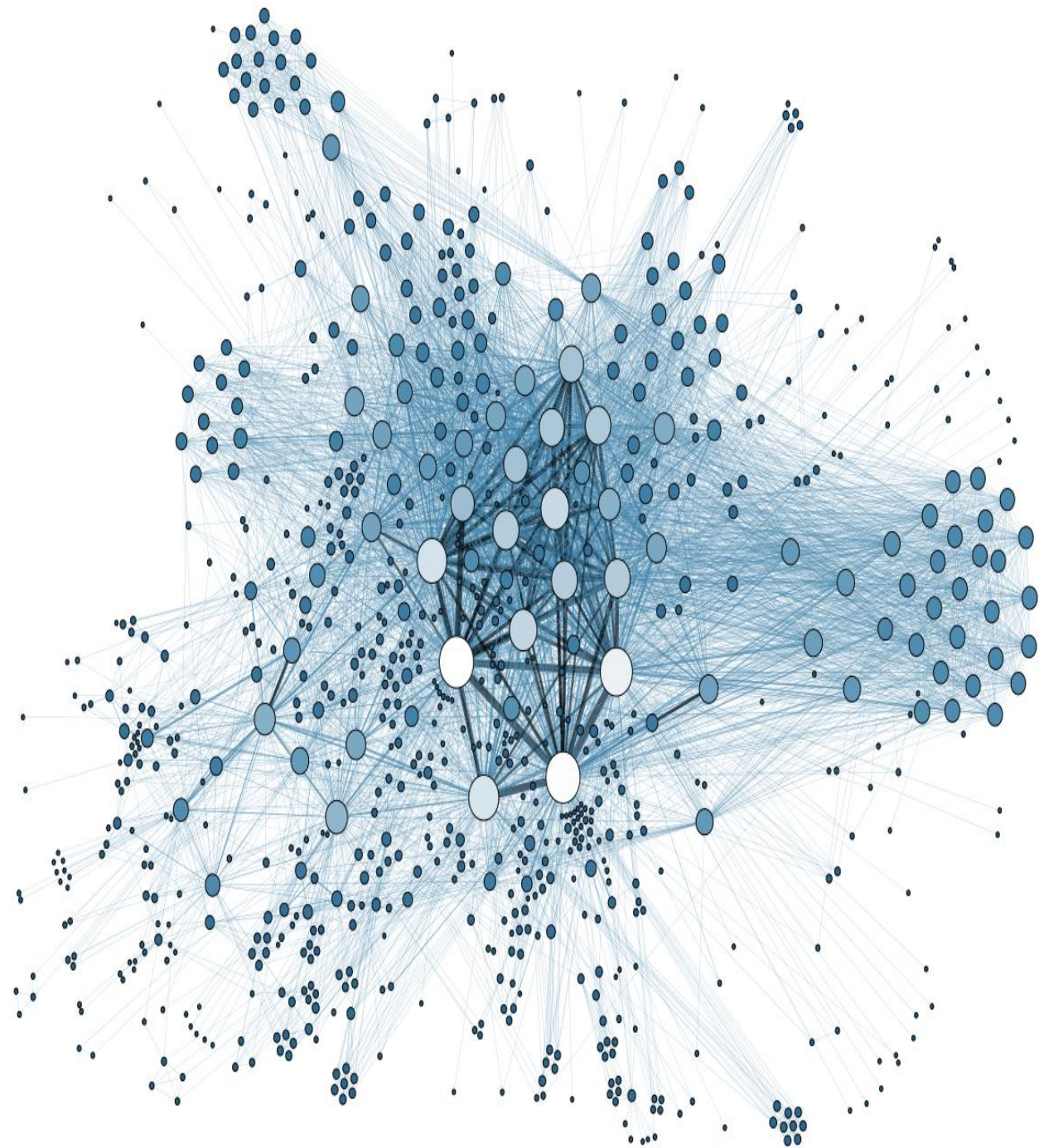
# FINAL PROJECT STROKE PREDICTION

IBM: SUPERVISED ML CLASSIFICATION MODEL

*By Joshua Kabwanga*

# CONTENTS

- Dataset Description
- Exploratory Data Analysis
- Various Classification Models
  - ML analysis and findings
- Recommendations





# OBJECTIVES

In this project I will demonstrate Machine learning algorithm to predict whether a patient is likely to get stroke or not based on the features parameters in the dataset .

- **Data Collection:** Gather comprehensive data from kaggle
- **Data Preprocessing:** Clean and preprocess the data to handle missing values and duplicates
- **Data Visualization:** Exploratory data analysis to reveal the hidden trend in the dataset
- **Model Training:** Train ML models such as logistic regression, Decision Tree, XBOOST and SMOTE
- **Model Evaluation:** Evaluate the model's performance using metrics like accuracy
- **Prediction and Diagnosis:** Use the trained models to predict the probability of stroke case for new patients



# Context

Now according to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.



# What is stroke

## **Type of strokes**

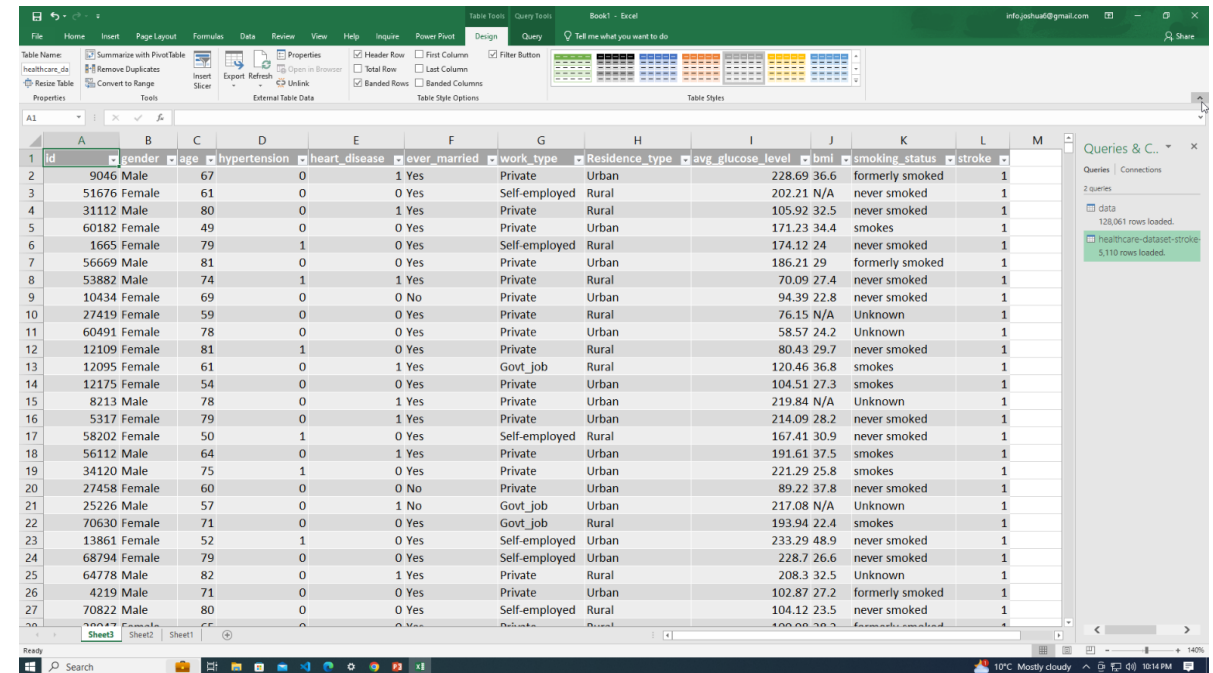
- An ischemic stroke occurs when the blood supply to part of the brain is blocked or reduced. This prevents brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes.
- Another type of stroke is a hemorrhagic stroke. It occurs when a blood vessel in the brain leaks or bursts and causes bleeding in the brain. The blood increases pressure on brain cells and damages them. [click here..](#)



# Attribute Information

- **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

Note: "Unknown" in smoking status means that the information is unavailable for this patient



id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	0	1 Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_jov	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
25226	Male	57	0	1	No	Govt_jov	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_jov	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1
68794	Female	79	0	0	Yes	Self-employed	Urban	228.7	26.6	never smoked	1
64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
4219	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly smoked	1
70822	Male	80	0	0	Yes	Self-employed	Rural	104.12	23.5	never smoked	1

# Dataset imbalance

## What is an imbalance dataset

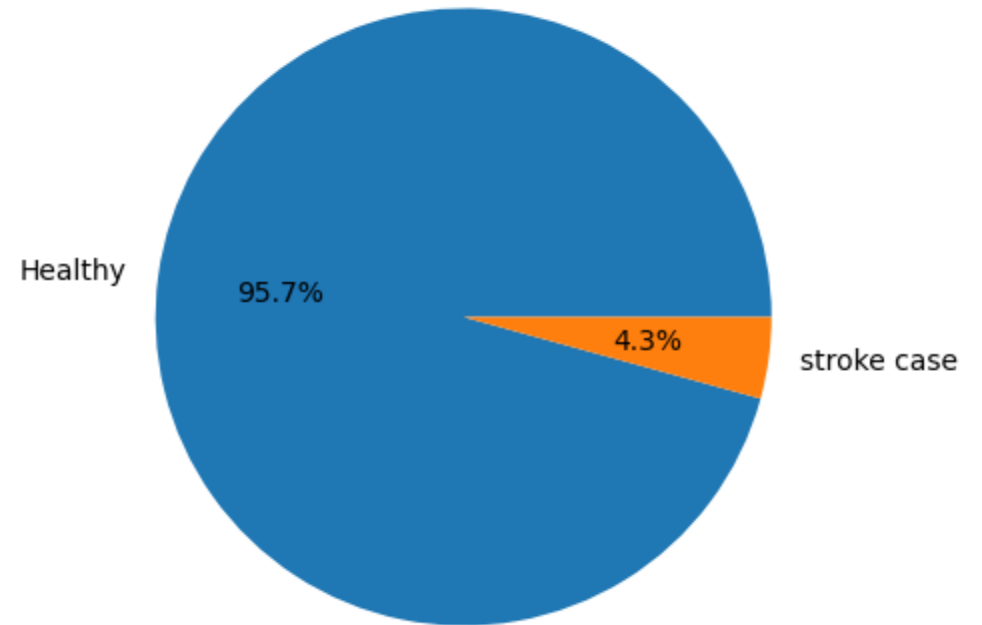
A balanced dataset, the number of Positive and Negative labels is about equal.

However, if one label is more common than the other label, then the dataset is imbalanced. The predominant label in an imbalanced dataset is called the majority class; the less common label is called the minority class.

In this case the majority class is 95.74%

While the minority class is 4.26%

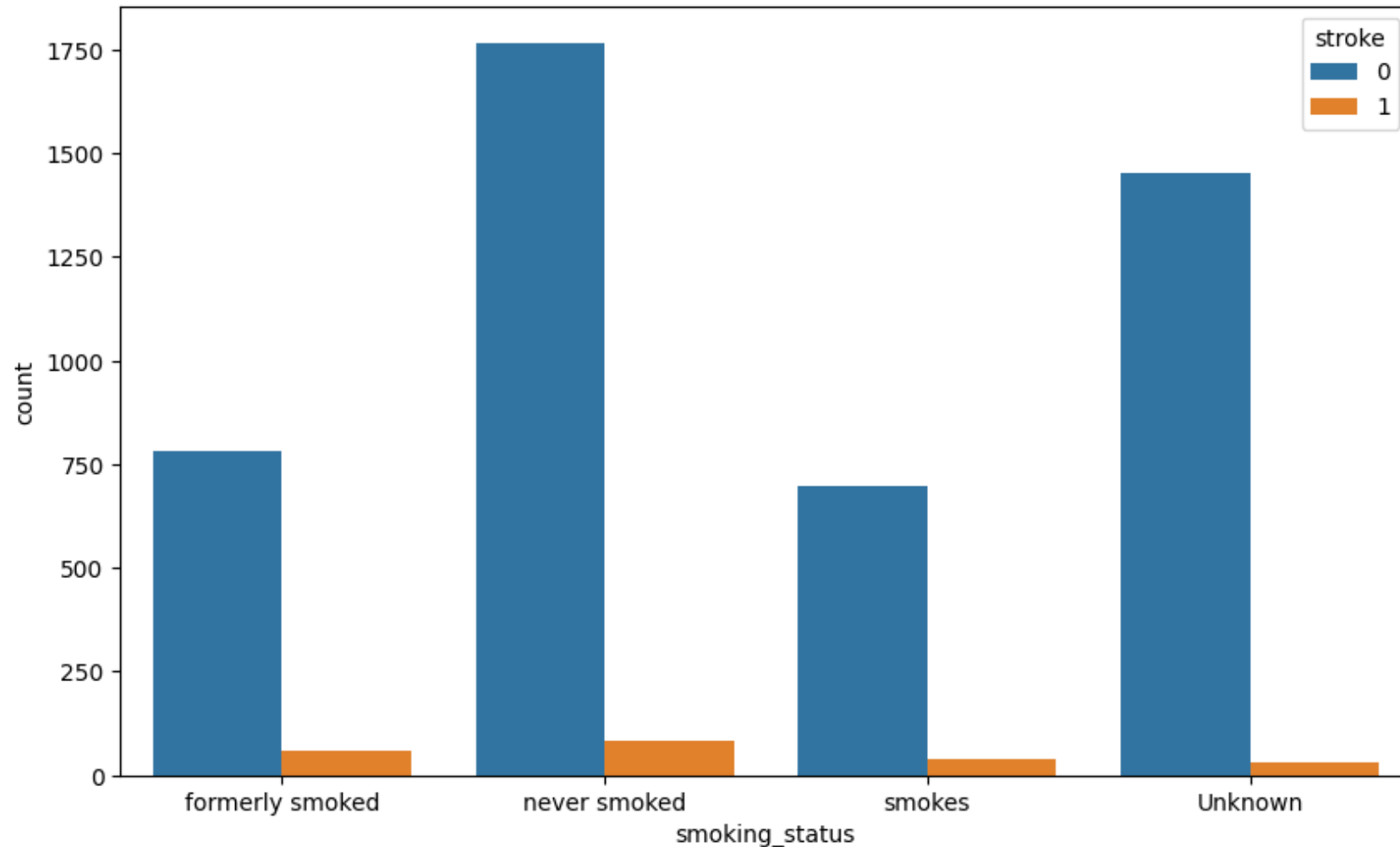
stroke case patients vs hearhy patients



# Categorical features

With the above insight, most of the customers never smoked follow by the those who didn't specify their the smoking history, than we have the formerly smoker. and finally by those that are actually smoking.

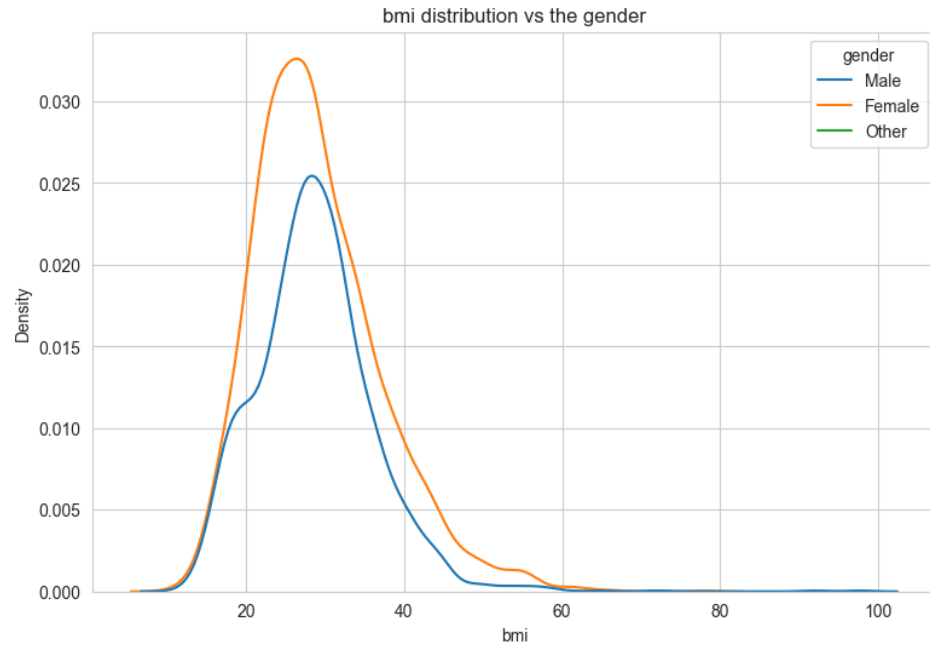
On the other hand most patients with stroke case are those who never smoke.



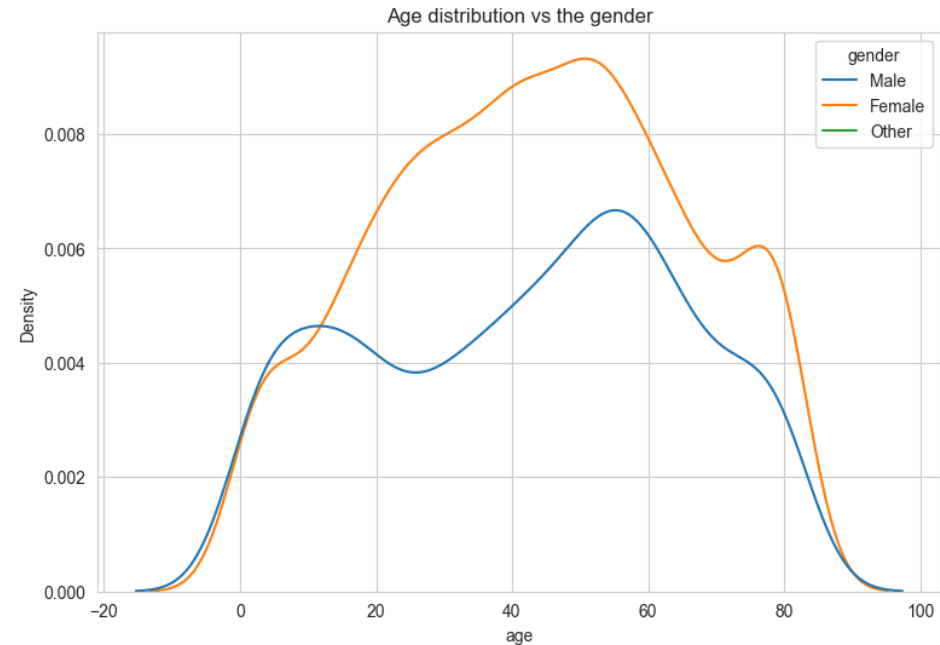


# Gender distribution vs age, and Body mass index

○ Body mass distribution per Gender



● Age distribution Per Gender



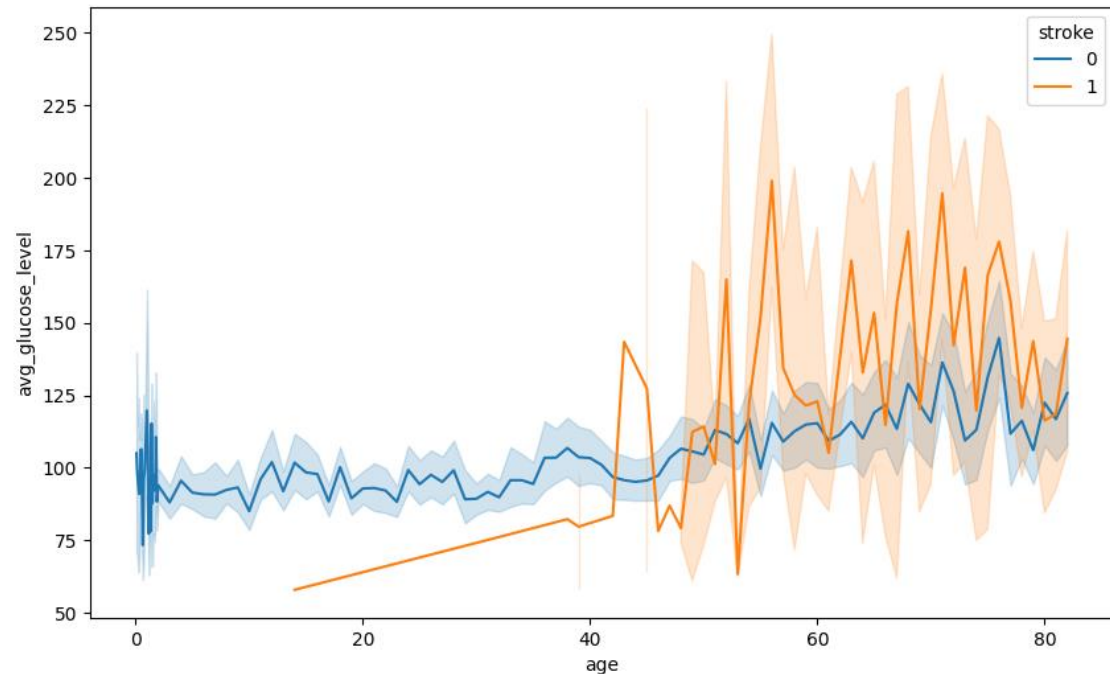
## Insight:

- The density of body mass index distribution showed that mostly female are more weight than male
- The density of age distribution showed that there is more female than male

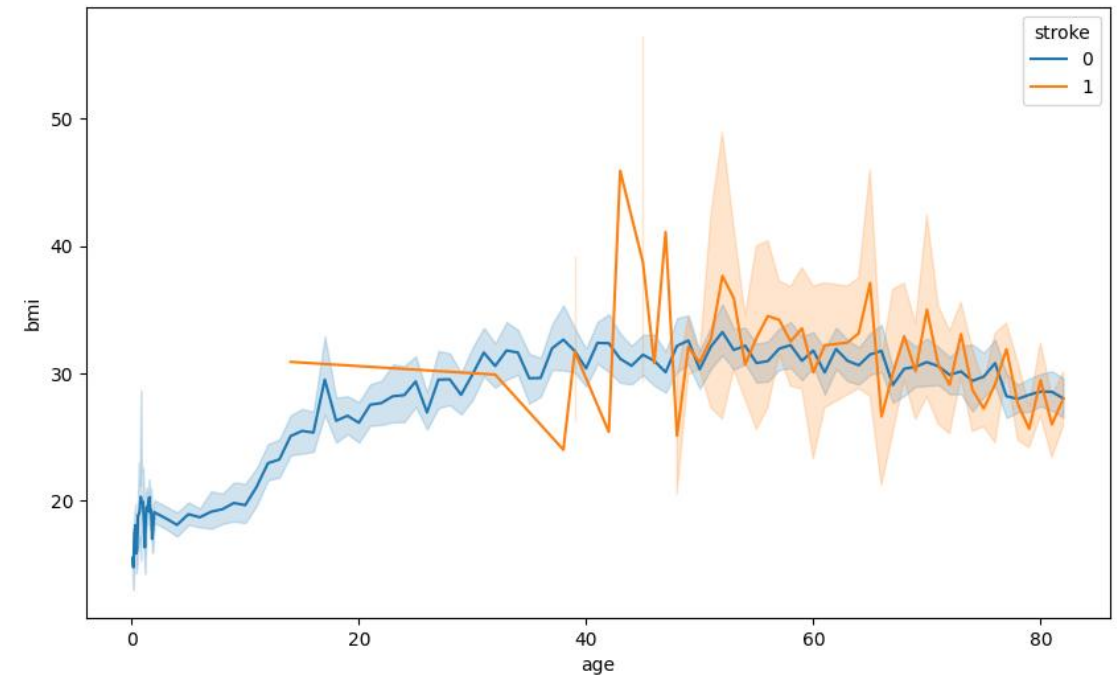
## Patient with stroke cases vs Healthy Patient with no stroke cases history

The above distribution number 1 represent patients with stroke while 0 represent healthy patients

### ● Age and glucose level vs stroke



### ● Age and BMI vs stroke

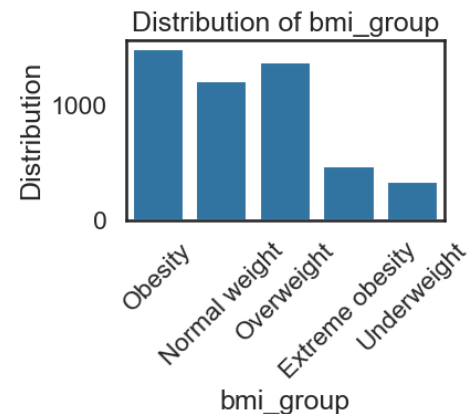
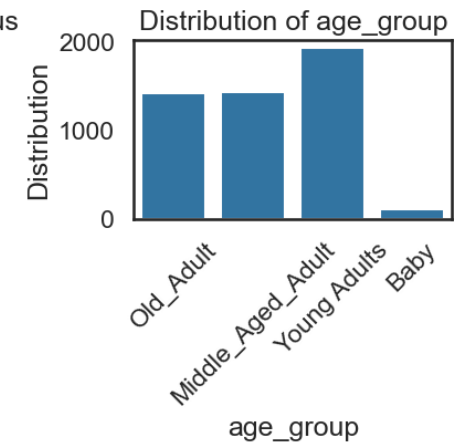
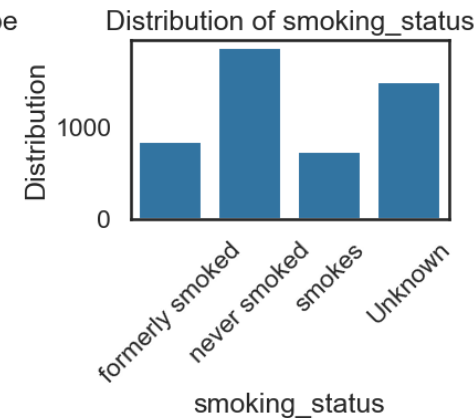
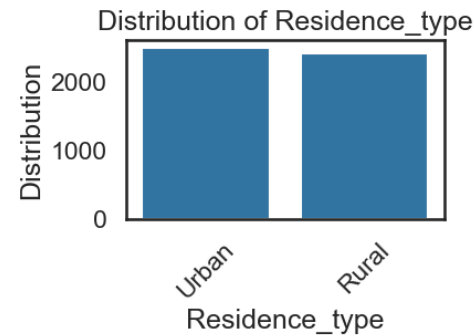
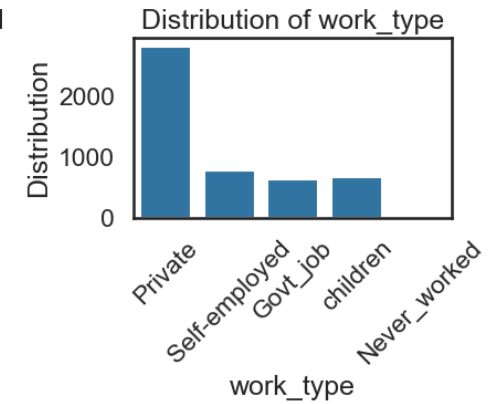
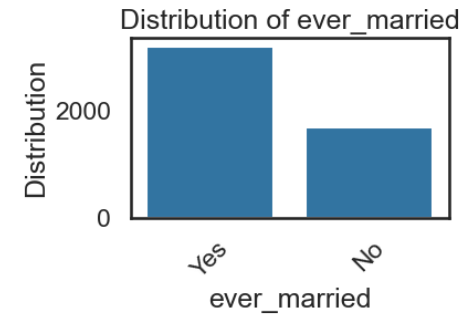
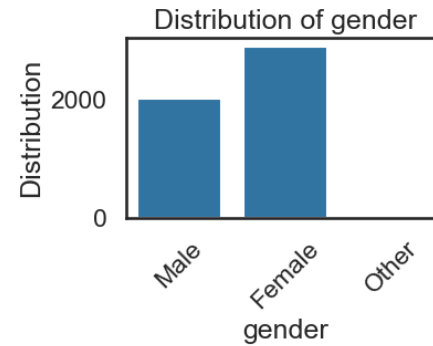


- The above insight reveal that most patient with stroke cases start from above 40 of age with a high percentage of glucose in the body.
- Secondly the more bmi weight with more age is also the highest chance of getting stroke

# Features distributions

The above features revealed the distribution of the following:

1. Gender
2. Ever Married
3. Work type
4. Residence\_type
5. Smoking status
6. Age group
7. Body mass index

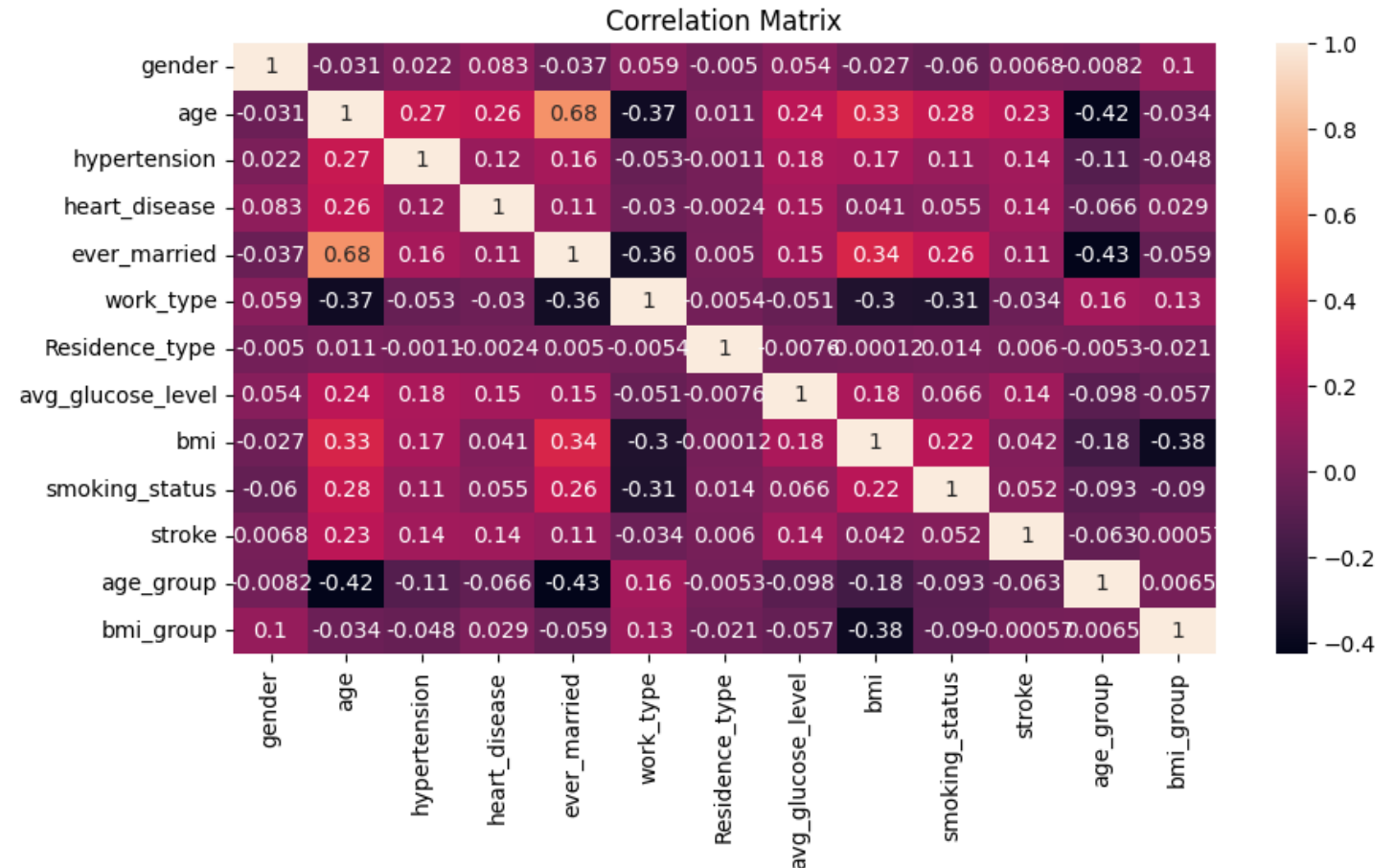




# Correlation between features and target

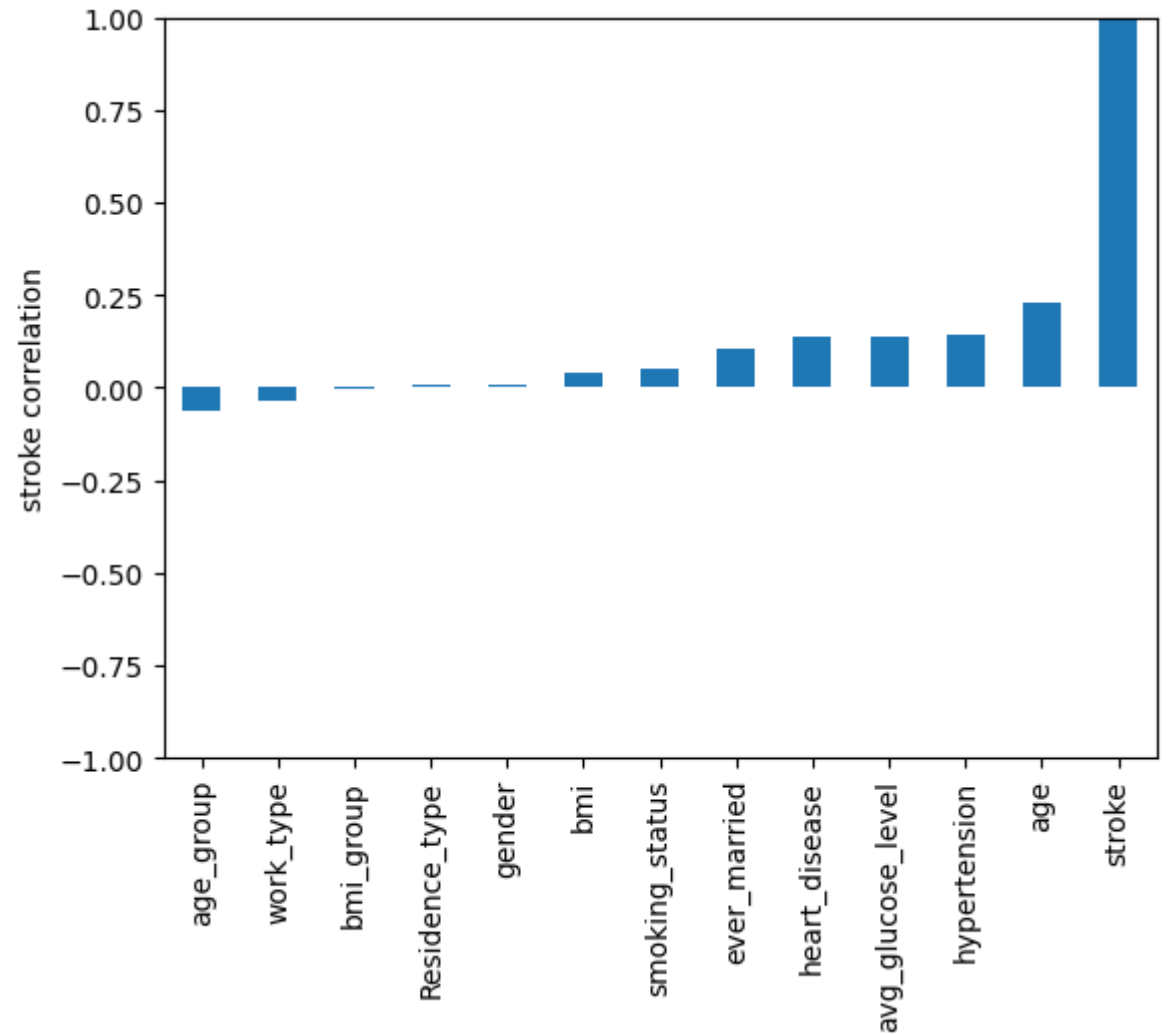
The following heatmap shows the correlation between the features themselves and between the features and the target.

- Age and ever married has 68%
- Eve married with bmi 34%
- Bmi and age 33%
- Age and smoking status 28%
- smoking status with ever married 26%
- Age with stroke 23%



# Features Correlation

The above result. as we can see the positive correlated features with stroke positive patient is Age follow by hypertension, glucose level, heart disease, ever married, smoking status, Body mass index.



# Machine Learning Models

The reason I used `StratifiedShuffleSplit` technique to split the dataset is because the target column is imbalanced

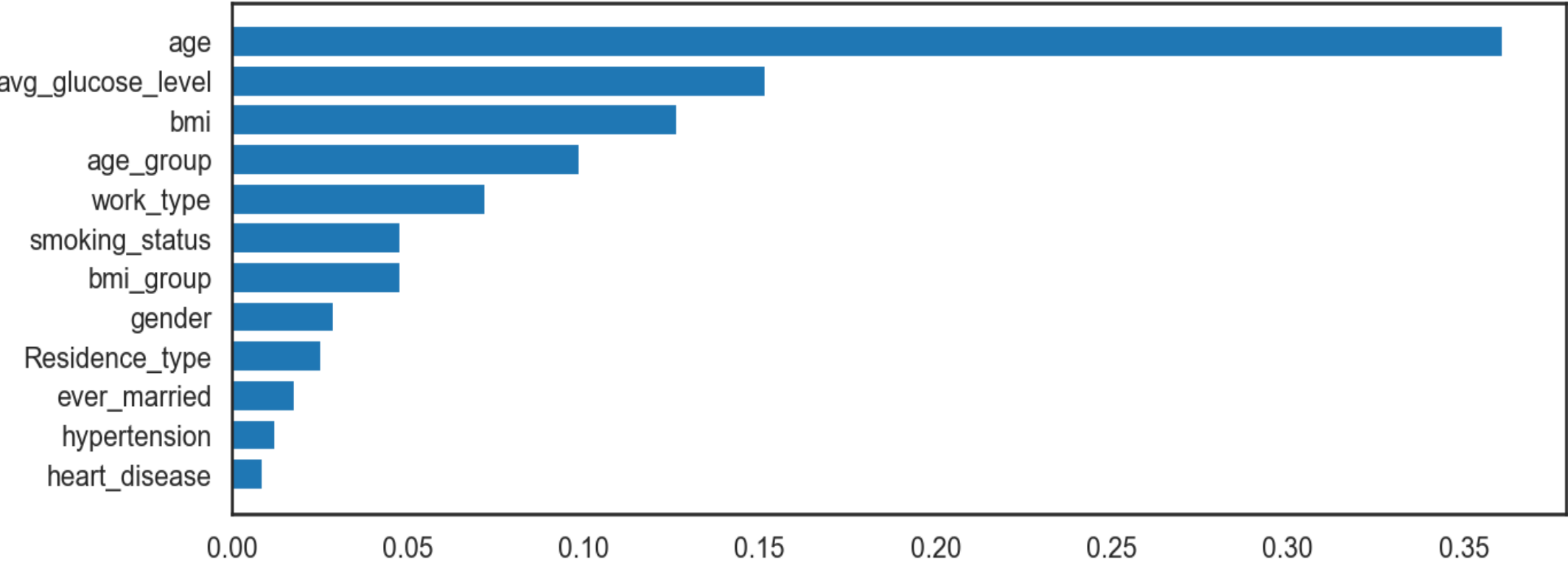
- **LogisticRegression**
- **XGBoostClassifier**
- **DecisionTreeClassifier**



# The best ML Model

Decision Tree

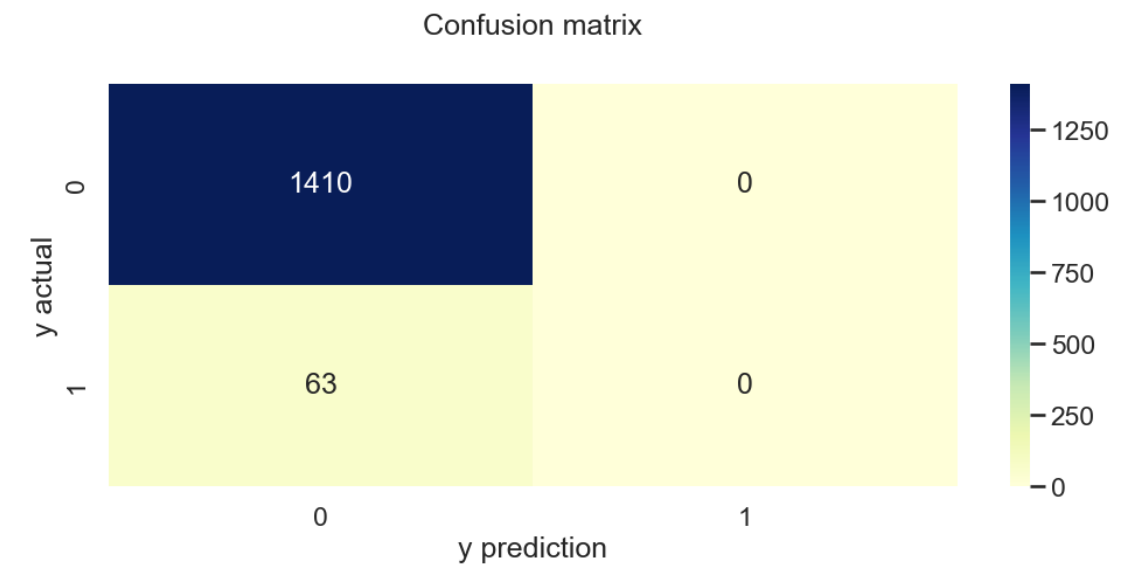
# feature\_importances\_



## Confusion Matrix

Logistic Regression ACC SCORE 95%

	precision	recall	f1-score	support
0	0.96	1.00	0.98	1410
1	0.00	0.00	0.00	63
accuracy			0.96	1473
macro avg	0.48	0.50	0.49	1473
weighted avg	0.92	0.96	0.94	1473

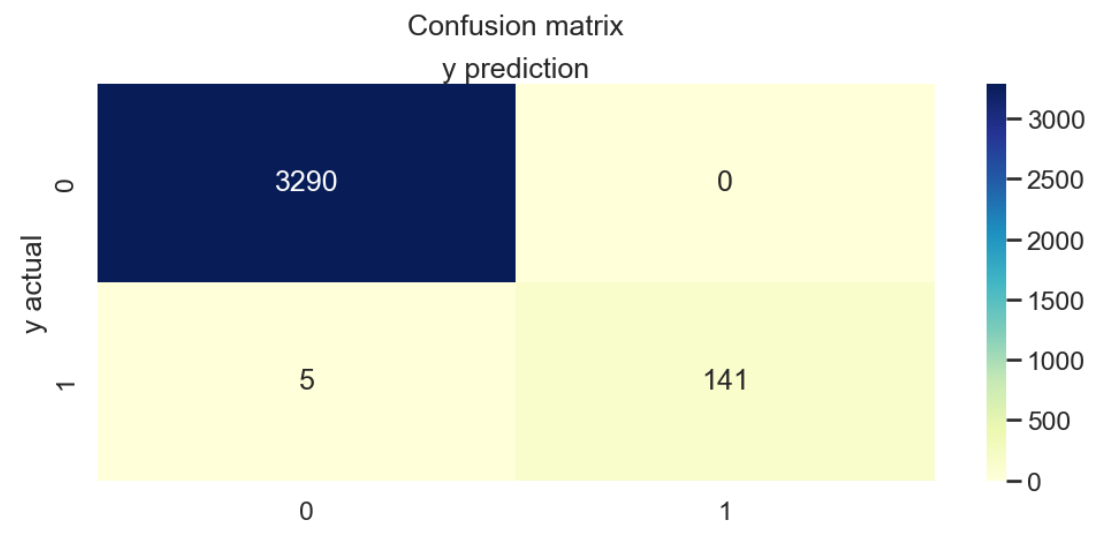




# Confusion Matrix

XBOOST ACC SCORE 99%

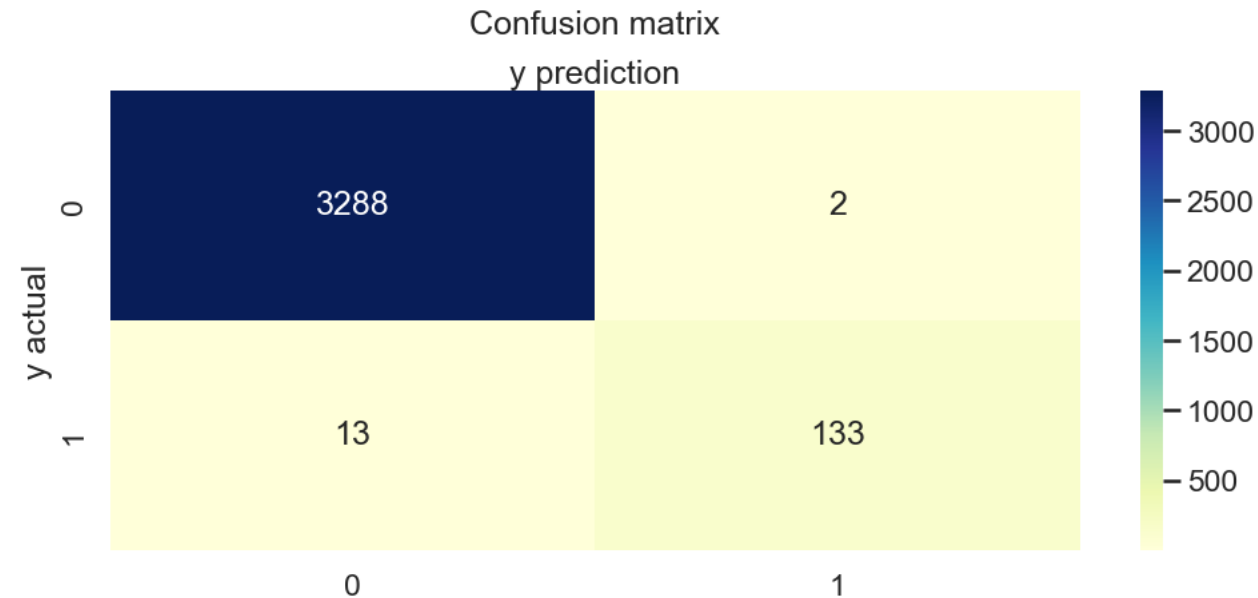
	precision	recall	f1-score	support	
	0	1.00	1.00	1.00	3290
	1	1.00	0.97	0.98	146
accuracy				1.00	3436
macro avg		1.00	0.98	0.99	3436
weighted avg		1.00	1.00	1.00	3436



Confusion Matrix

Decision Tree ACC SCORE 99%

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3290
1	0.99	0.91	0.95	146
accuracy			1.00	3436
macro avg	0.99	0.96	0.97	3436
weighted avg	1.00	1.00	1.00	3436



# Key Insights and Recommendation

Based on the performance above, my personal recommendation will be Decision tree. It provides the highest accuracy 99% and f1-score 94% while using less computational resources than XBOOST

XBOOST accuracy and performance, the best model with accuracy score 99% while f1\_score 0,02%

Decision Tree accuracy and performance, the best model with accuracy score 99% while f1\_score 94%

The 3 top feature \_importance. Age, Glucose, bmi.

This is recommendation is true and inline with medical research.

Joshua Kabwanga  
info.joshua6@gmail.com

