Data Science Bootcamp

Project presentation

11 clinical features for

Predicting Stroke Events

Done by: Eman Ahmed Alzhrani

supervised by: Ms. Mariam Elmasry

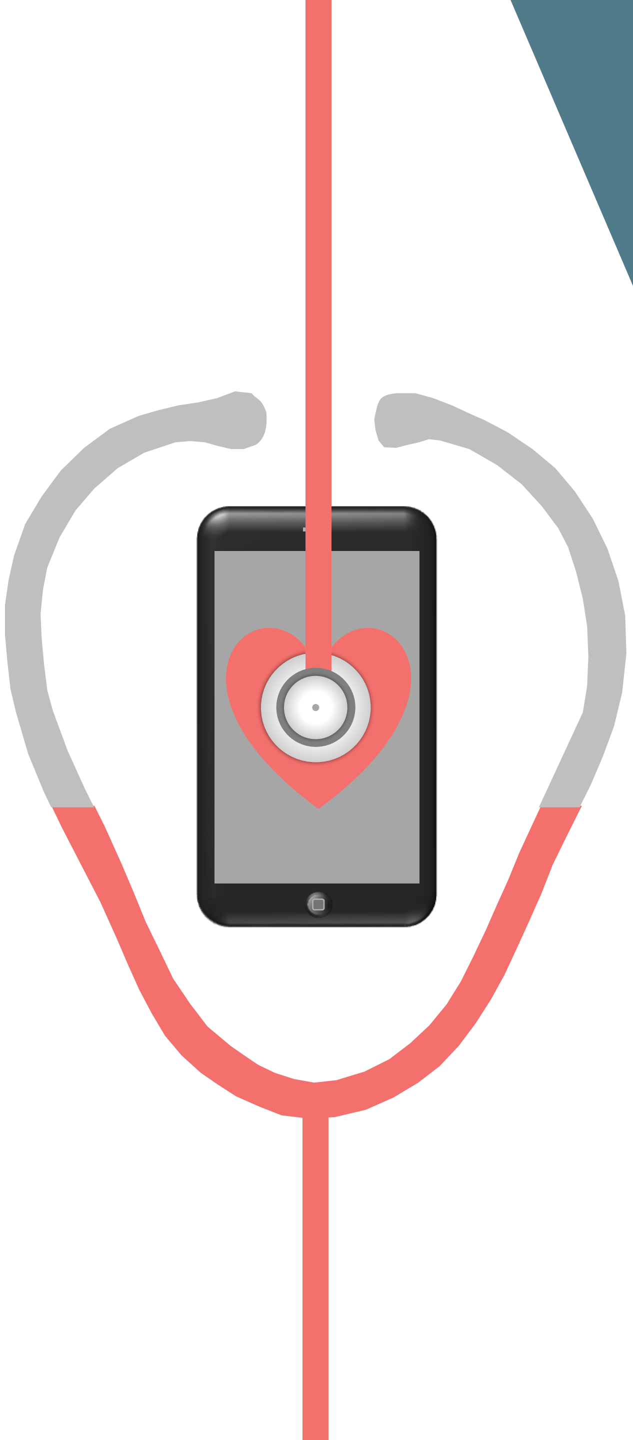
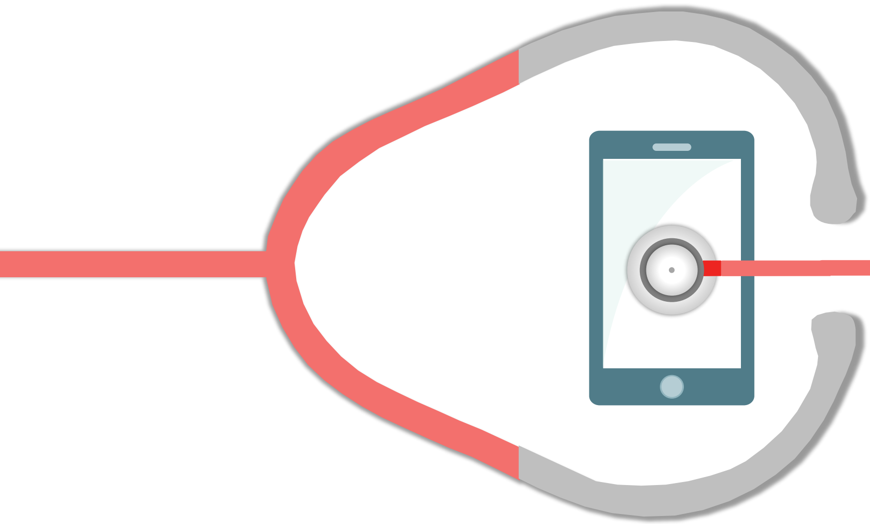


Table of contents

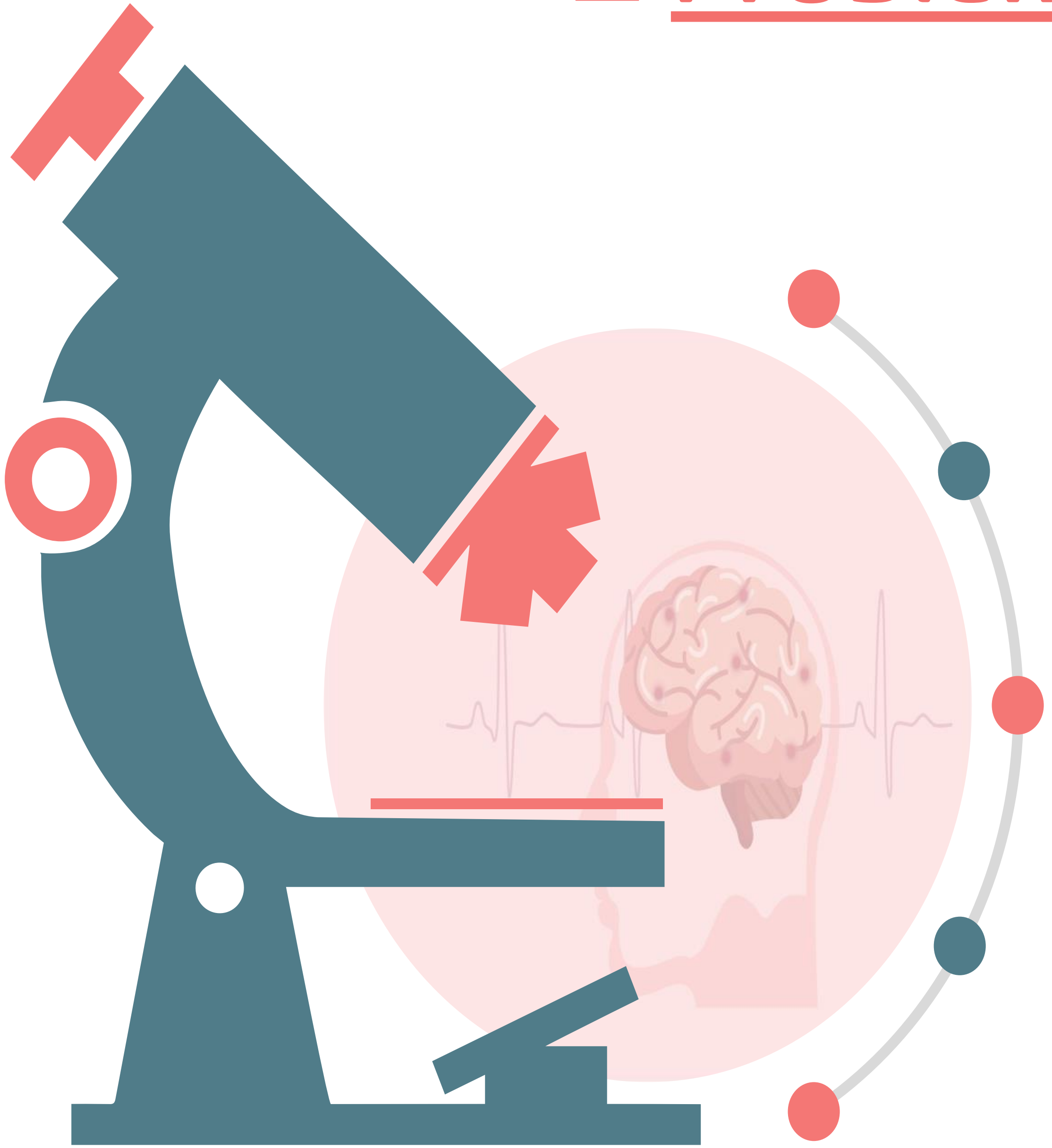
Stroke DS project

Stroke DS project



Problem and Purpose

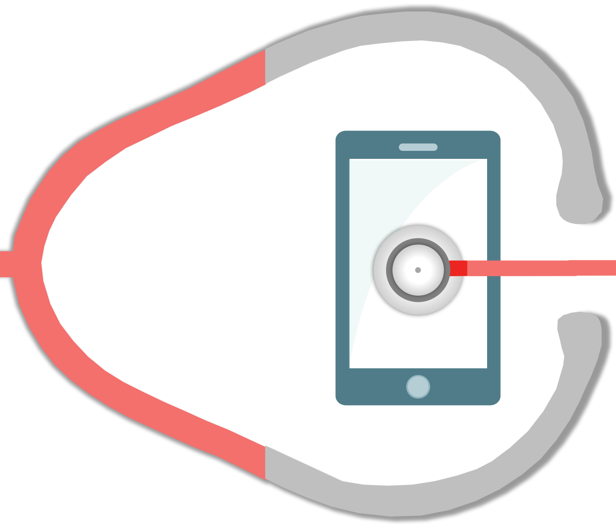
❑ Problem and purpose



•A stroke is a medical emergency called "brain attack”. occurs when part of the brain loses its blood supply and stops working. This causes the part of the body that the injured brain controls to stop working.

•Early action can reduce brain damage and other complications, the purpose of this project is to build a classification model that helps 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.

Stroke DS project



Used Tools

Stroke DS project

Tools

Data Processing

Pandas, Numpy

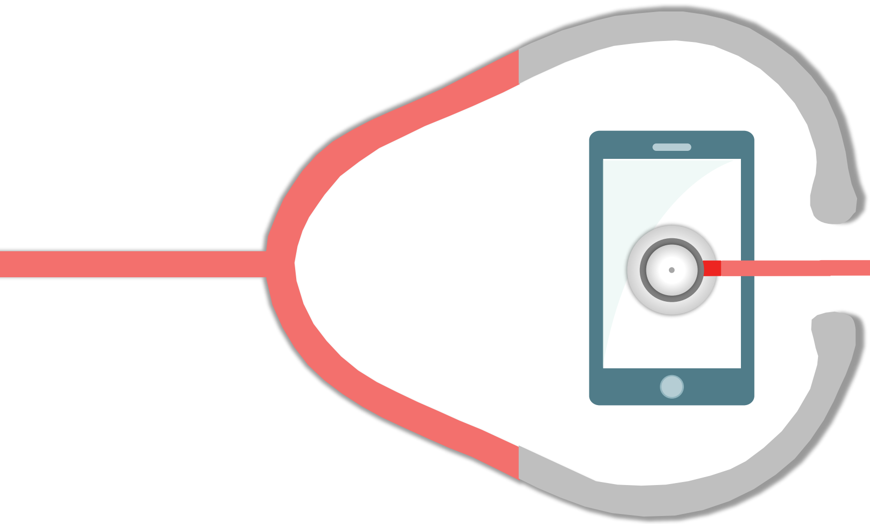
Modelling

scikit-learn , Imbalanced-learn

Visualization

Matplotlib, Seaborn, Plotly, cufflinks

Stroke DS project



Dataset & data cleaning

Stroke Prediction Dataset

11 clinical features for predicting stroke events

Data source "Kaggle":

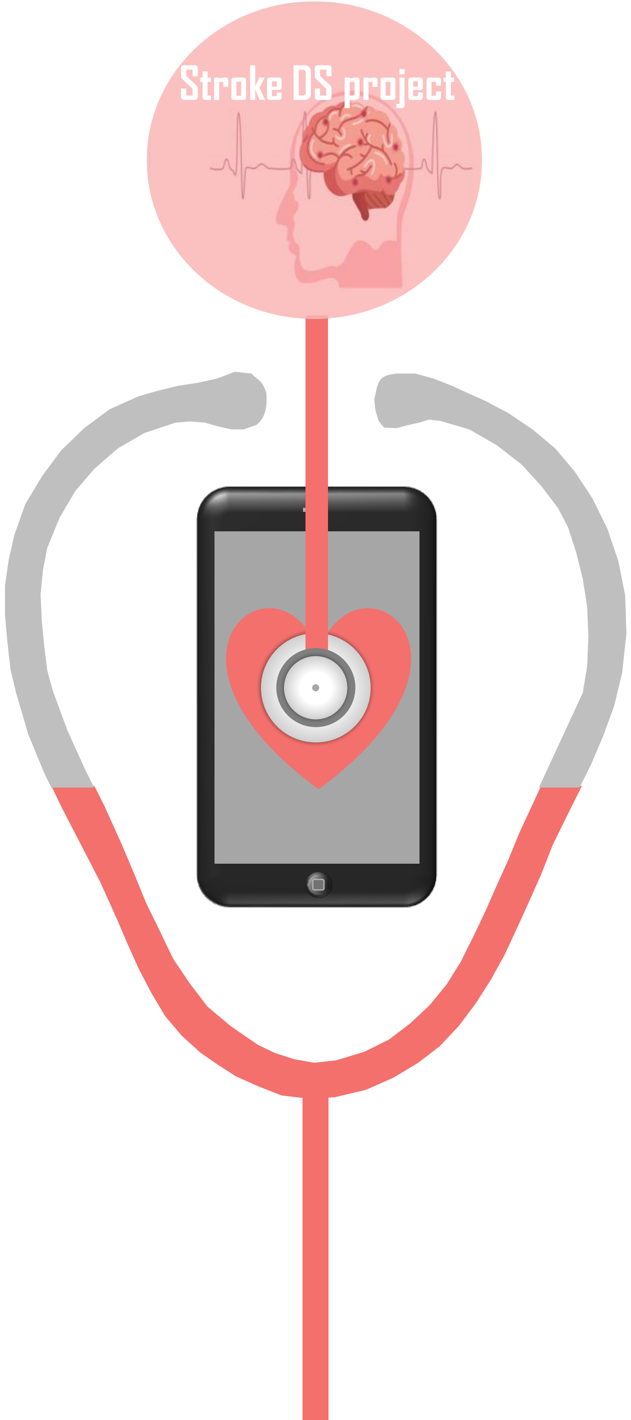
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Target

```
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

Data has **5,110** instances and **11** attributes. "id is excluded"



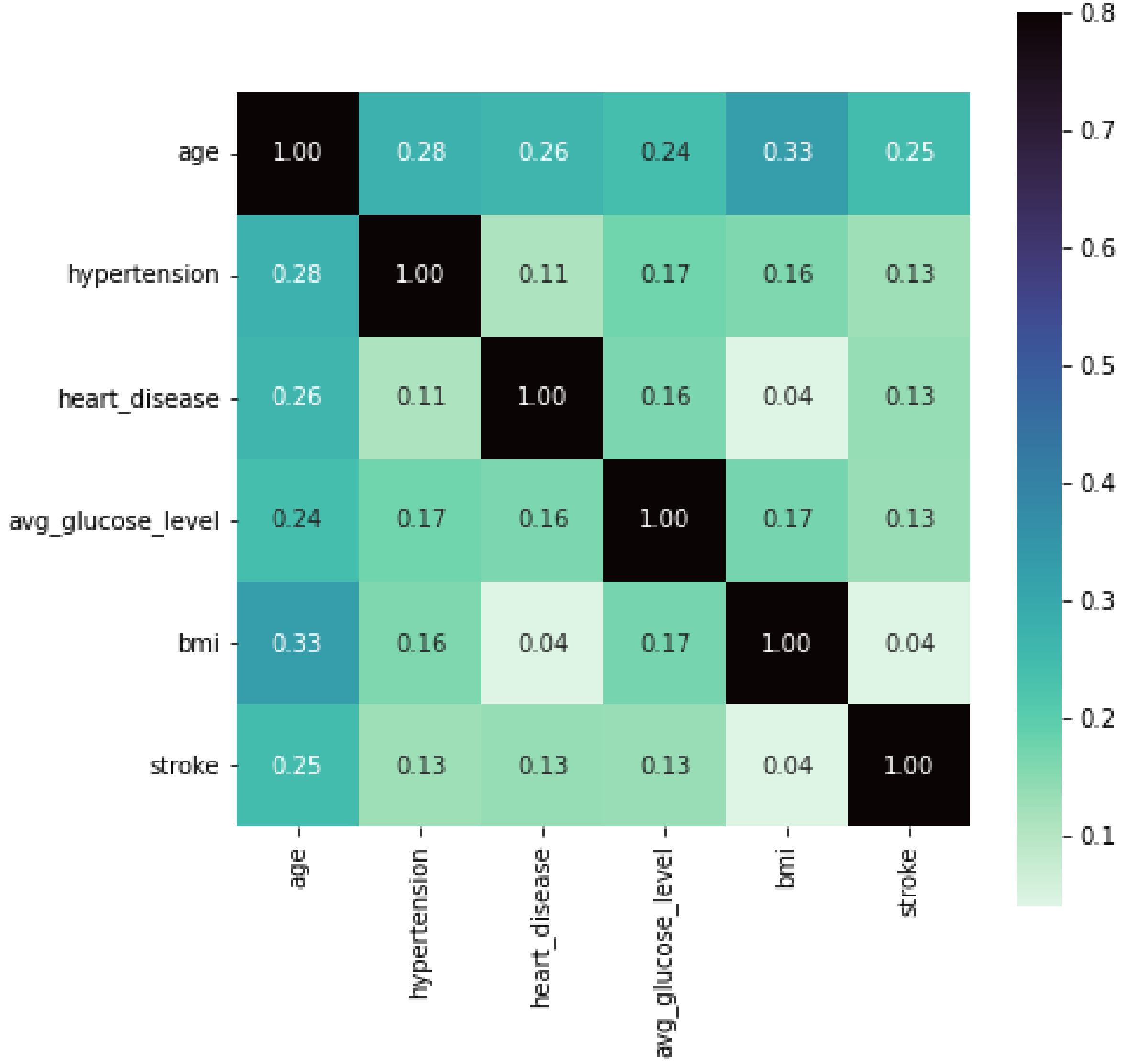
Methods and techniques used to clean the dataset and prepare it….

1. Check for the null values (missing values).
2. Check for duplications.
3. Drop unnecessary columns (id).
4. Drop 'Other' gendered individuals to simplify the mathematical computations “since there is only one patient that has ‘other’ value as a gender”
5. 'bmi' column has 201 nan values so I decided to impute them with mean.
6. Round 'age' column and convert data type to integer.

Findings(EDA)

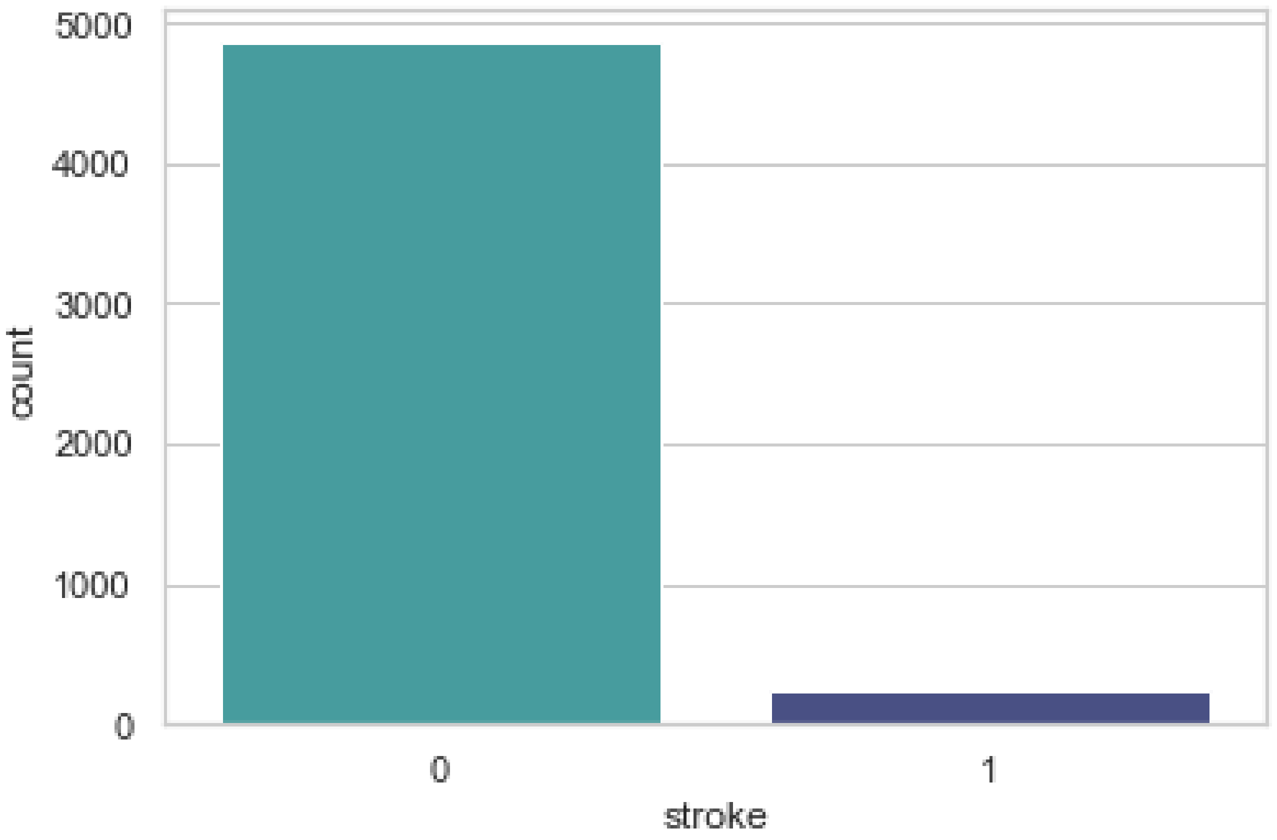
Stroke DS project

Correlation of values with each other



Stroke DS project

Visualizing stroke distribution



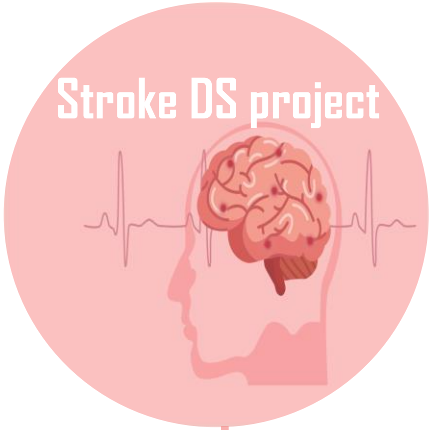
About 5% detected to have stroke. While 95% with no stroke.

"Imbalanced data"

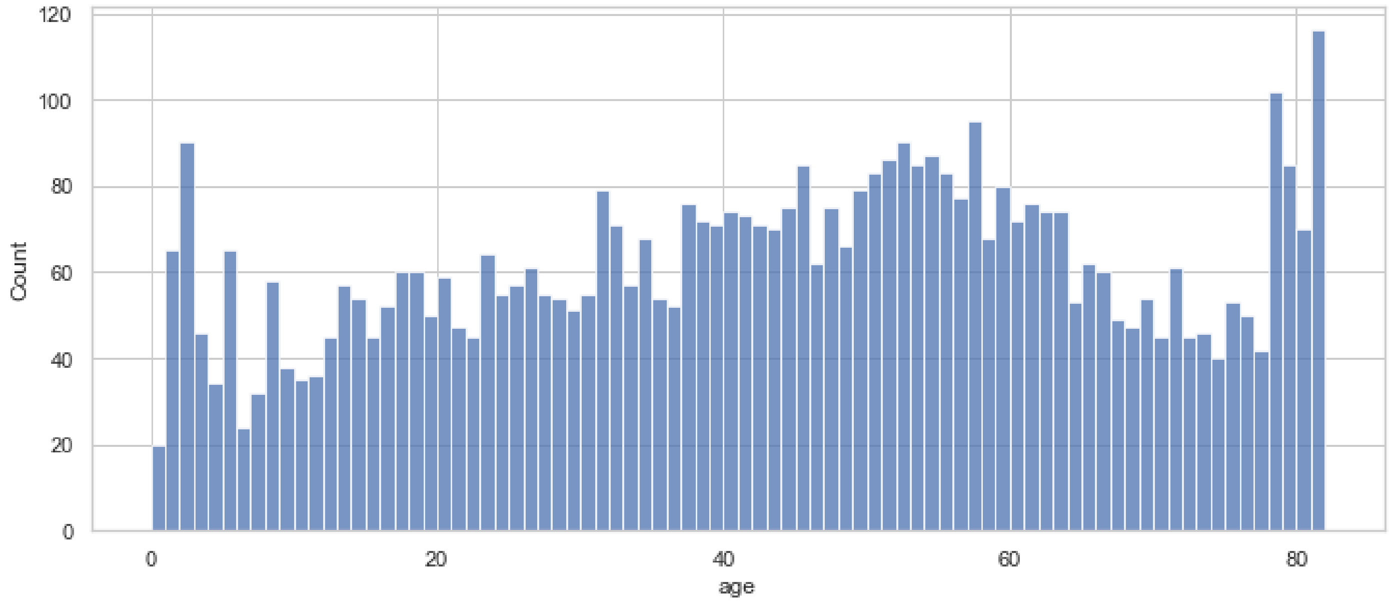
Stroke DS project

Visualizing Some Numerical Features

age	hypertension	heart_disease	avg_glucose_level	bmi



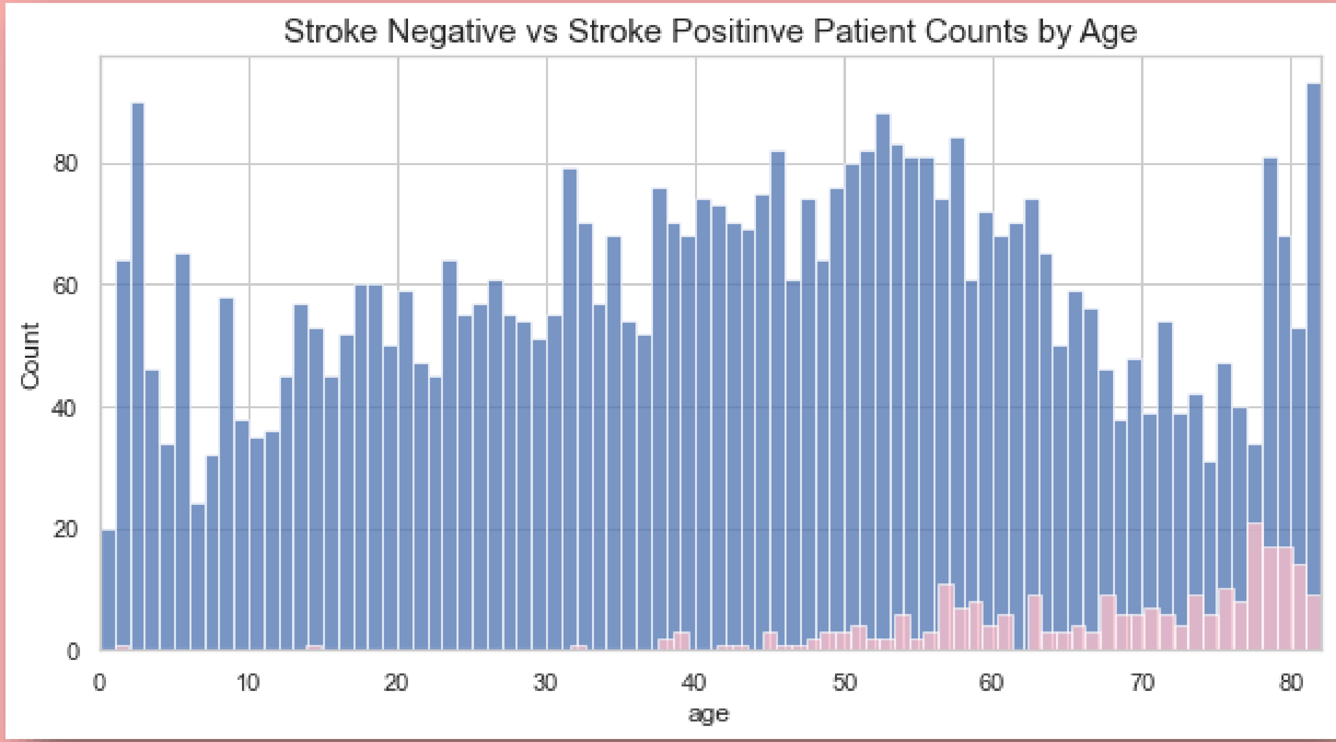
Visualizing age distribution



Age distribution of data is balanced.

People in dataset are between 0 to 82 years old.

Visualizing age and stroke correlation.



As we can see, we have a balanced age distribution in the stroke negative group.

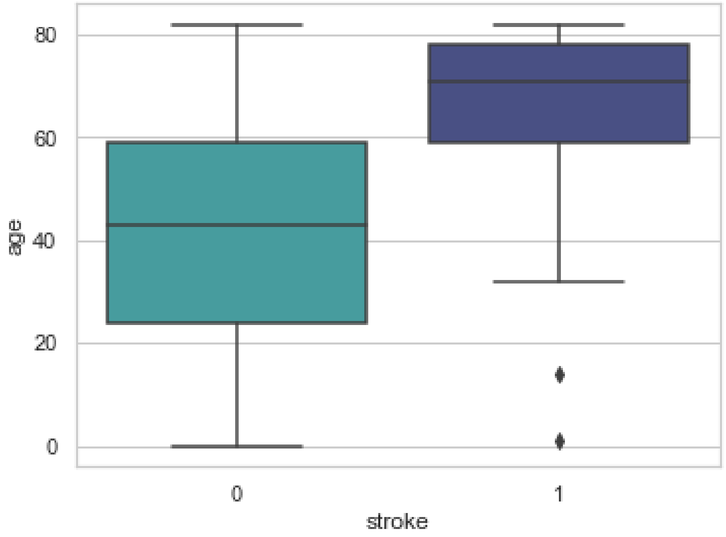
However, stroke positive patients are stacked to the right side (older people).

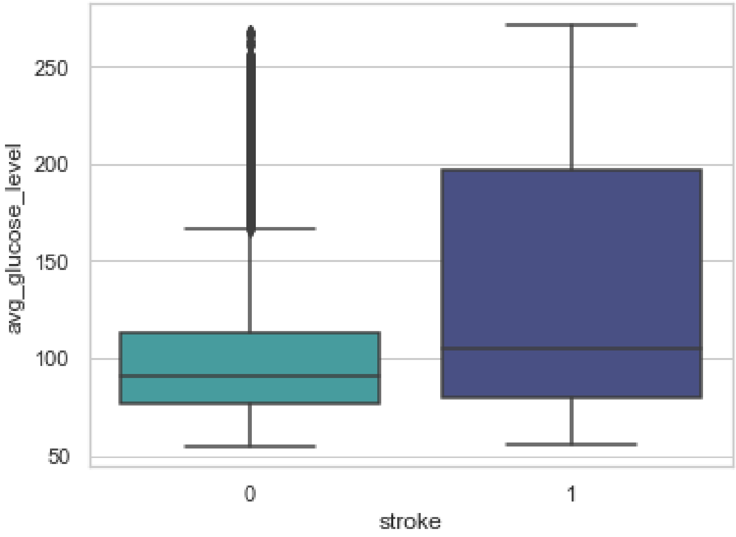
we can clearly see that age plays a huge role in predicting stroke.

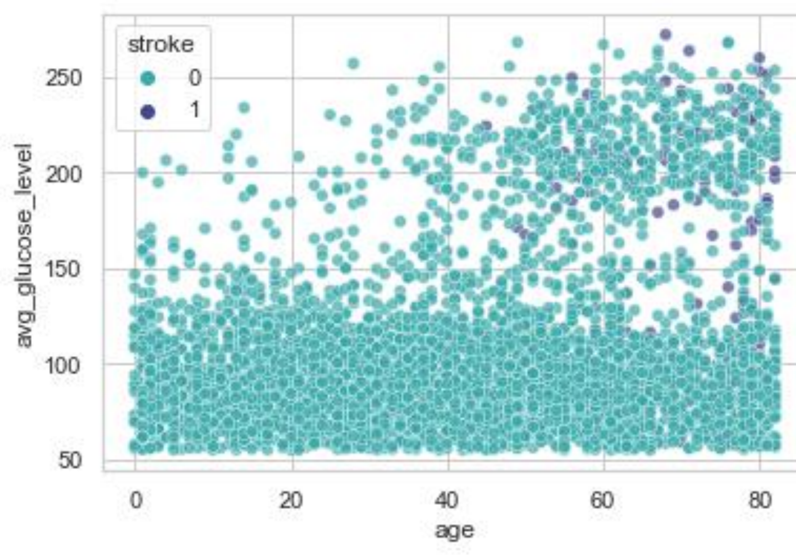
	gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi	smoking_status	stroke
162	Female	1	0	0	No	children	Urban	70.37	28.9	Unknown	1
245	Female	14	0	0	No	children	Rural	57.93	30.9	Unknown	1

Describing three variables

(Age , avg-glucose-level, stroke)





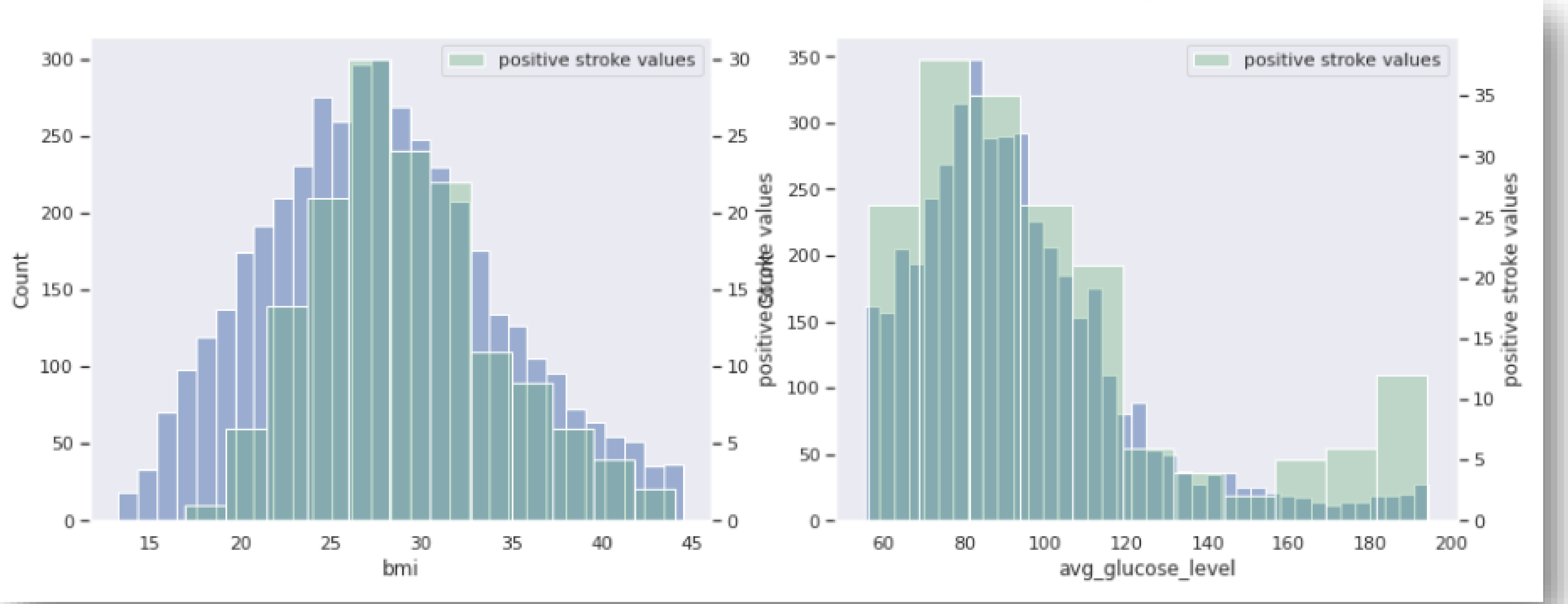


most people who have had stroke are on the right top side who are old and has a higher avg glucose level-- makes sense

Stroke with Glucose – level & Age

Describing another three variables

(bmi , avg-glucose-level, stroke)



Normal bmi and glucose levels are seen with the values of positive cases. This does not mean bmi and glucose levels don't play a role. They must have an effect but its not clear in here.

Stroke with Glucose – level & BMI

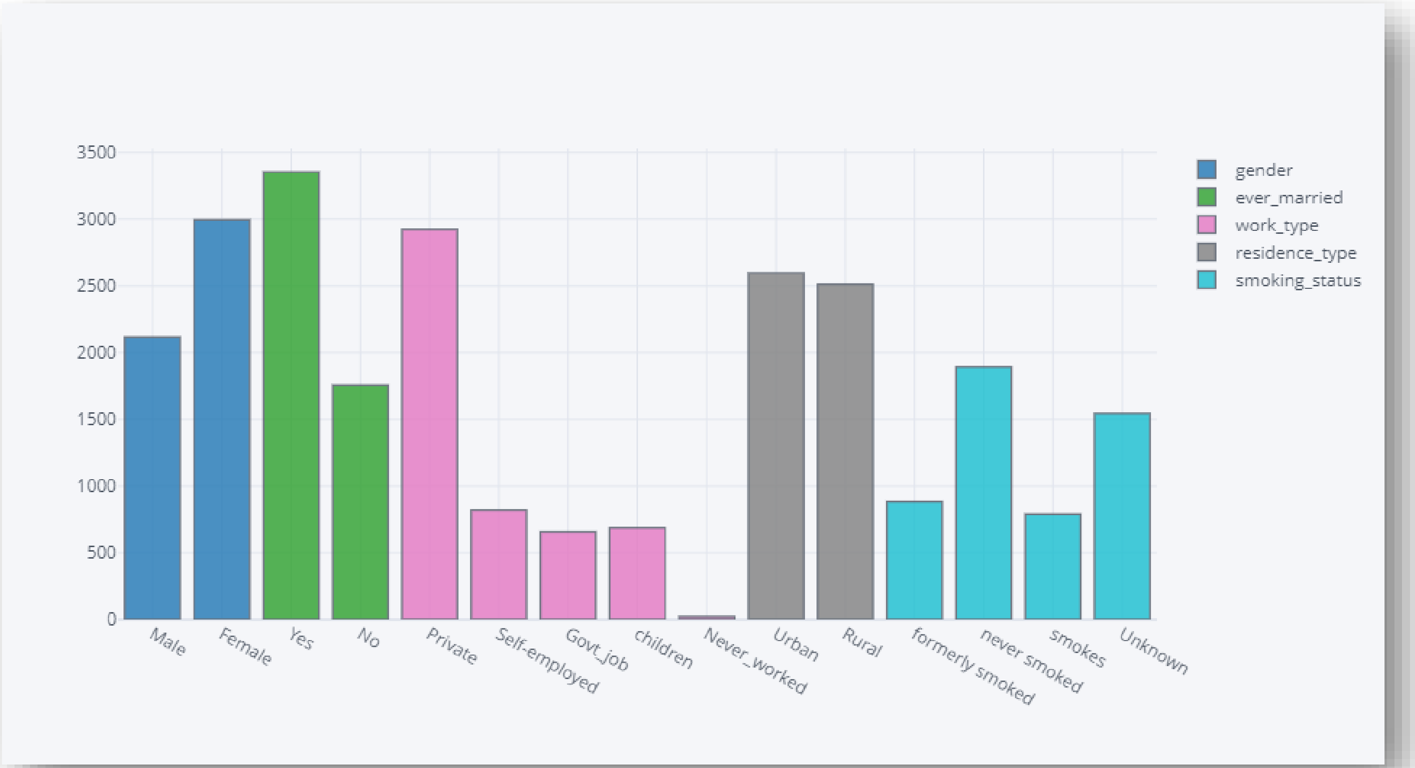
Stroke DS project

Visualizing Some Categorical Features

gender	ever_married	work_type	residence_type	smoking_status

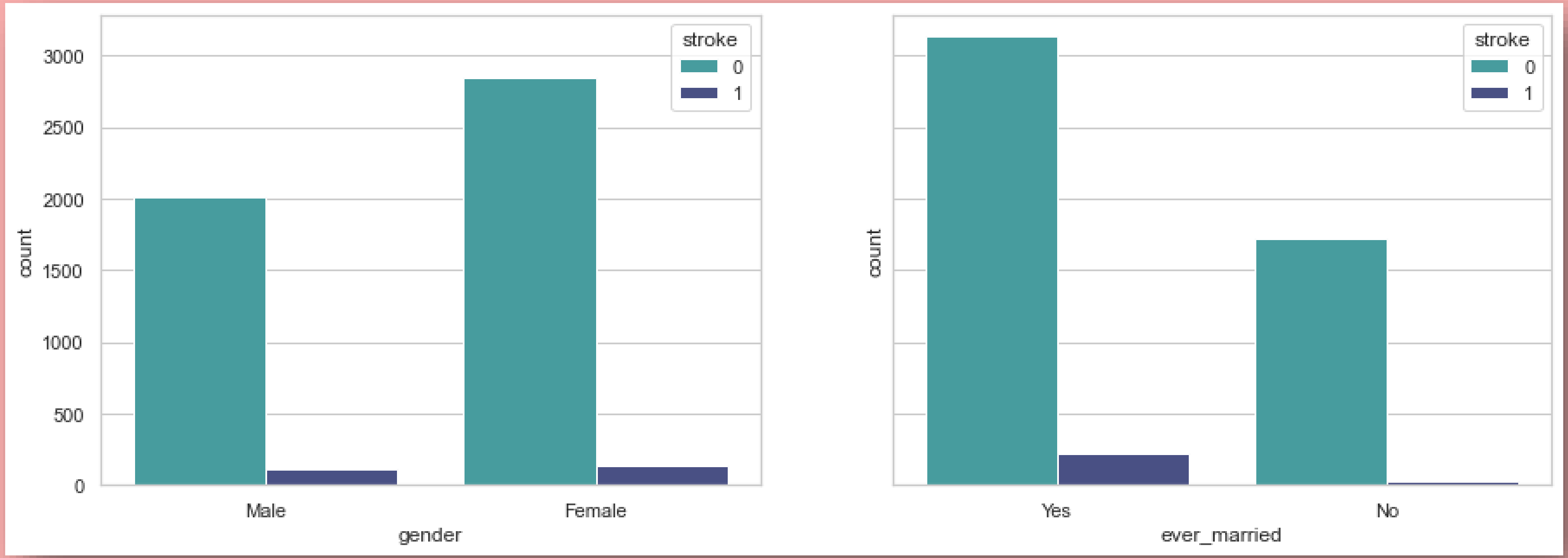
Stroke DS project

Categorical Data iplot



This plot gives a general vision of all the categorical data.

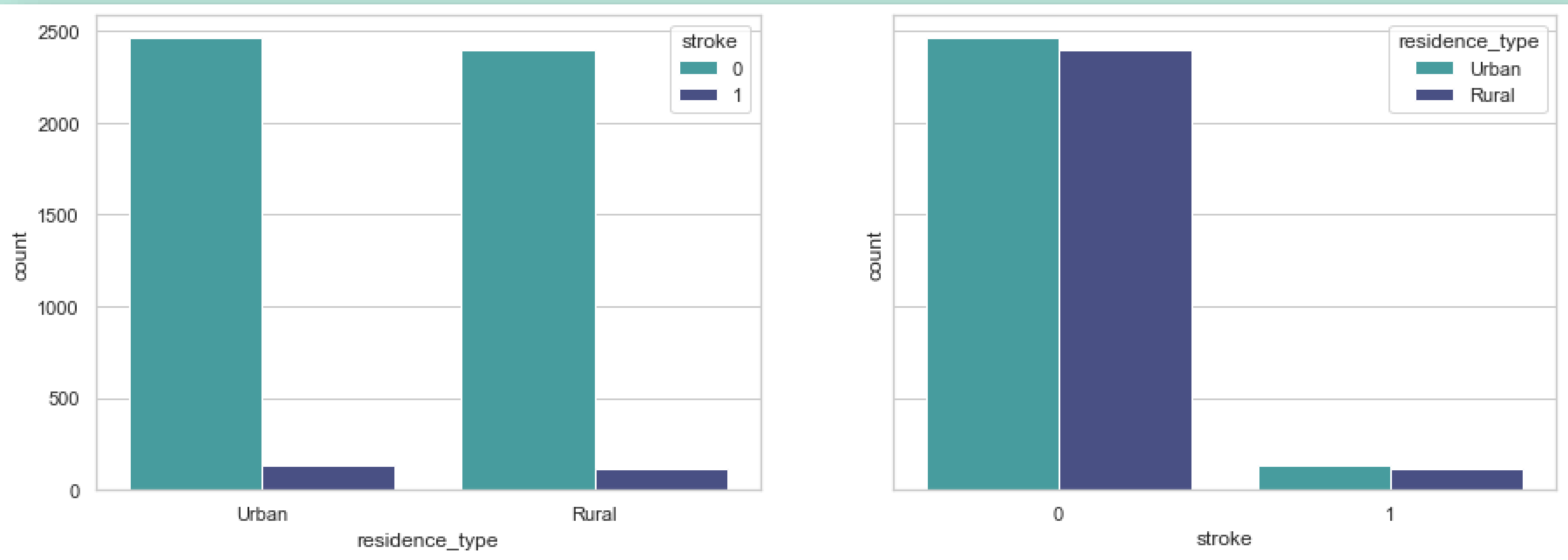
Visualizing marriage and gender correlation with Stroke



The result with married people is kinda <u>funny</u>. The <u>rate of stroke is higher in married people</u> :) but this may be a <u>result of bias</u> :/

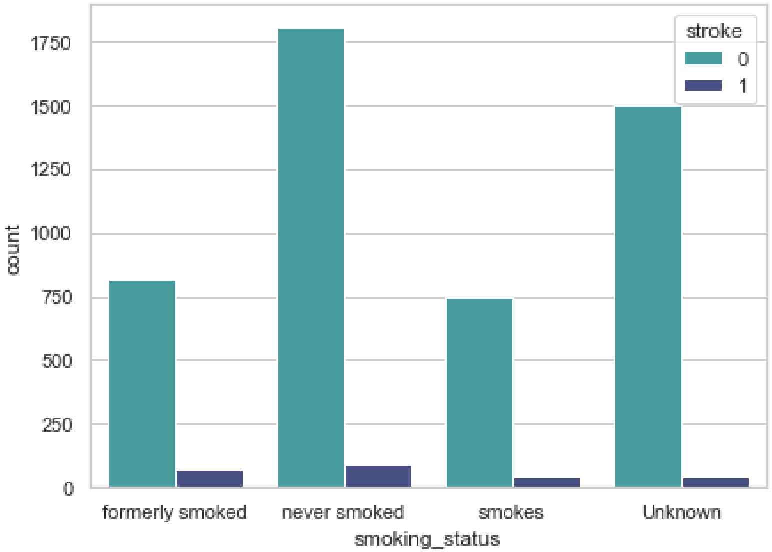
In gender there is no significant deference between males and females regarding stroke cases.

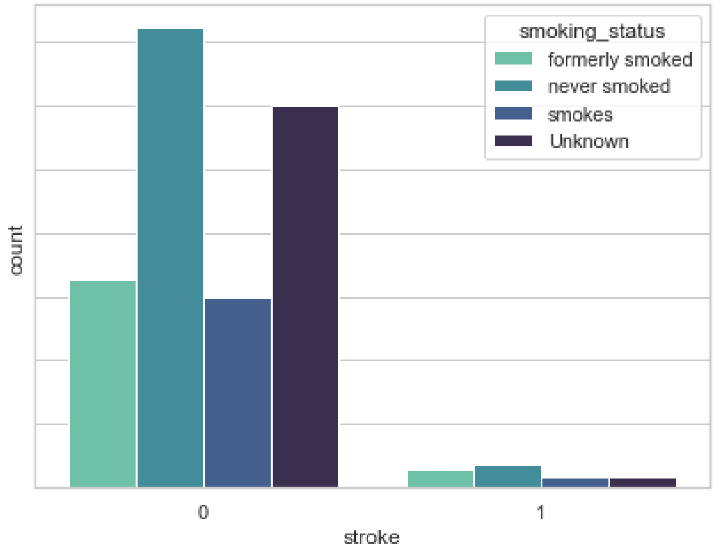
Visualizing residence type correlation with Stroke

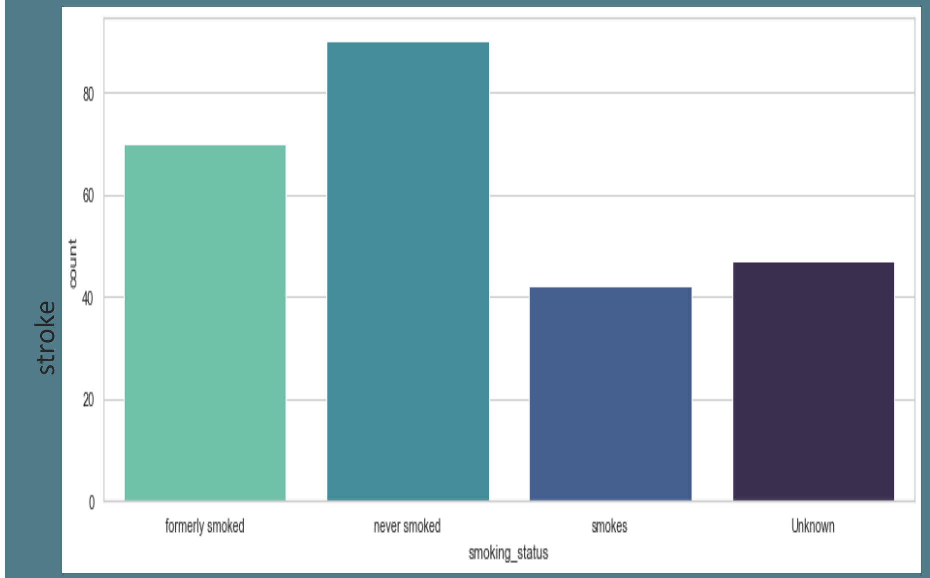


The numbers and rates of stroke and non-stroke patients are very similar in both residence types (urban, rural) .

Smoking status distribution in stroke patients





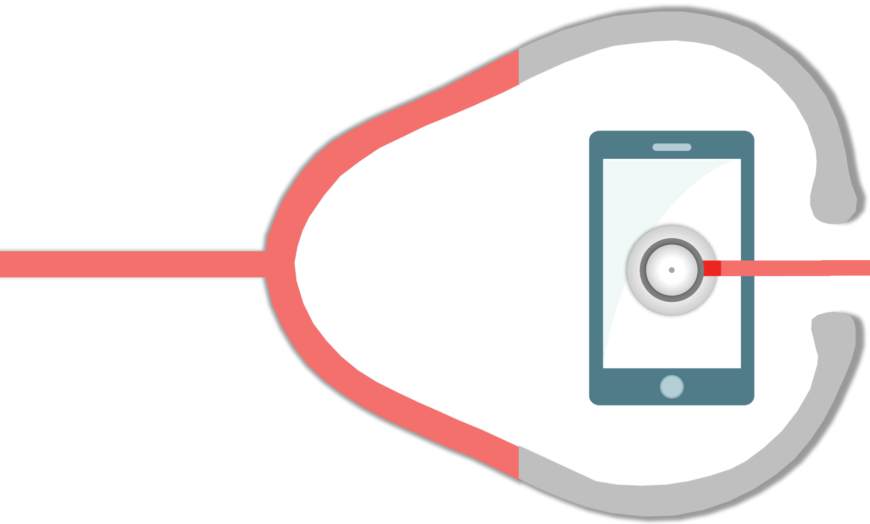


Smoking_status

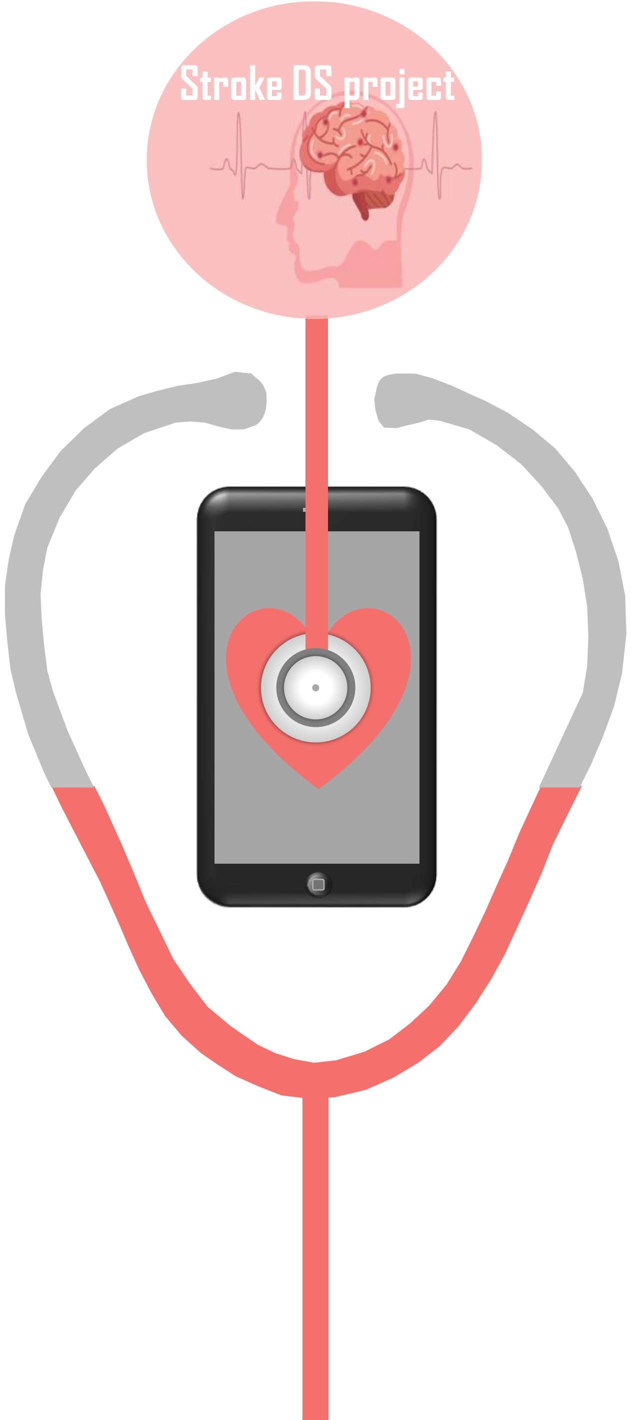
The highest stroke group is non-smokers!! Strange but true. This may be due to other factors and bias.

Stroke Relation with Smoking

Stroke DS project



Data Modelling



Data Preparation for Modelling:

1. All categorical data transformed into numerical by using dummies and converting to Boolean.
2. Encoding by "Gitting dummies“ for 'work_type', 'smoking_status' columns.
3. Convert 'gender’ column to Boolean 1 male - 0 female.
4. Convert ‘ever_married’ column to Boolean 1 married - 0 not married.
5. Convert ‘residence_type’ column to Boolean 1 urban - 0 rural.
6. scale the variance to make the data closer to normal distribution ('age’, 'avg_glucose_level’, 'bmi’).
7. Splitting data into (testing & training) and then I make resampling by using SMOTE

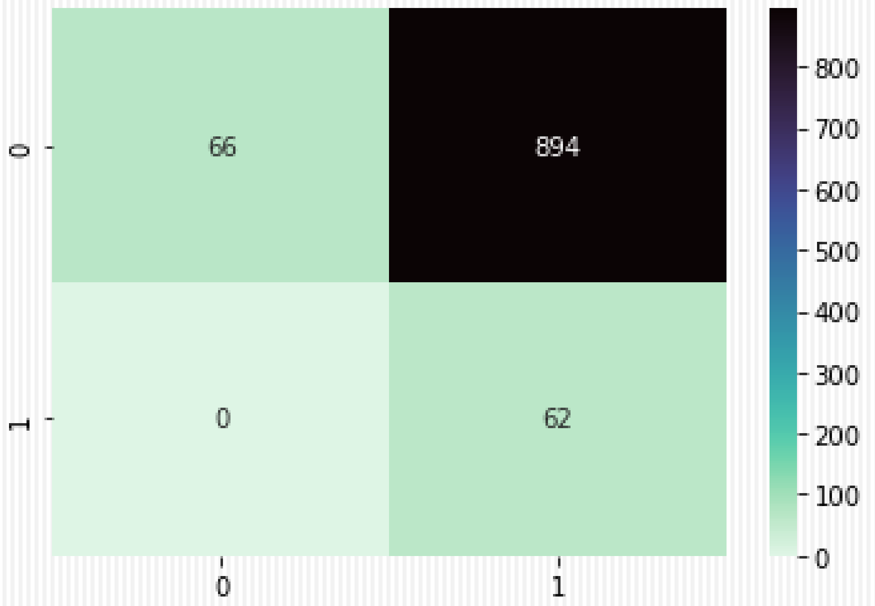
Stroke DS project

Starting to build the Models

(XGboost, Random Forest, Logistic Regression)

Logistic Regression Model

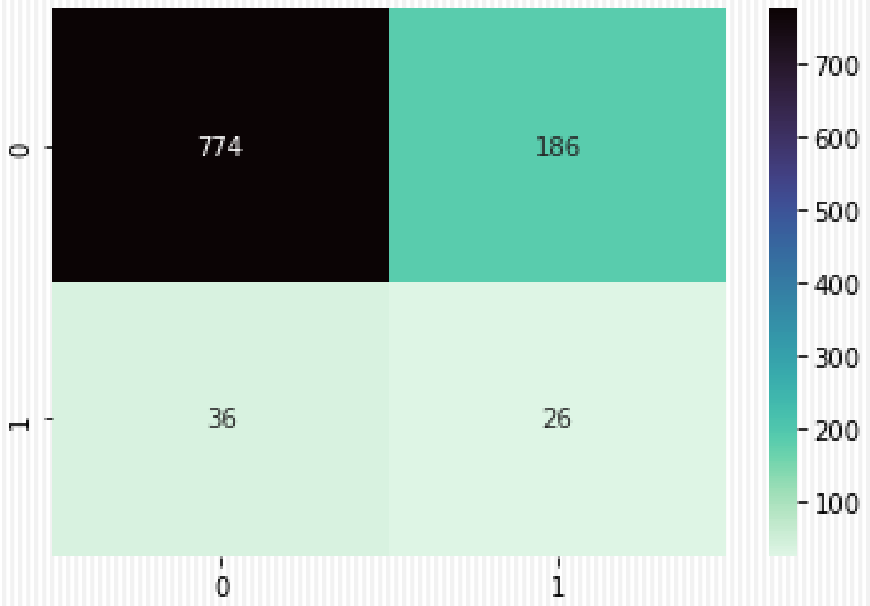
	Precision	Recall	f1-score	support
0	1.00	0.07	0.13	960
1	0.06	1.00	0.12	62
Accuracy			0.13	1022
Macro avg	0.53	0.53	0.13	1022
Weighted avg	0.94	0.13	0.13	1022



Random Forest Classifier Model

	Precision	Recall	f1-score	support
0	0.96	0.81	0.87	960
1	0.12	0.42	0.19	62

Accuracy			0.78	1022
Macro avg	0.54	0.61	0.53	1022
Weighted avg	0.91	0.78	0.83	1022

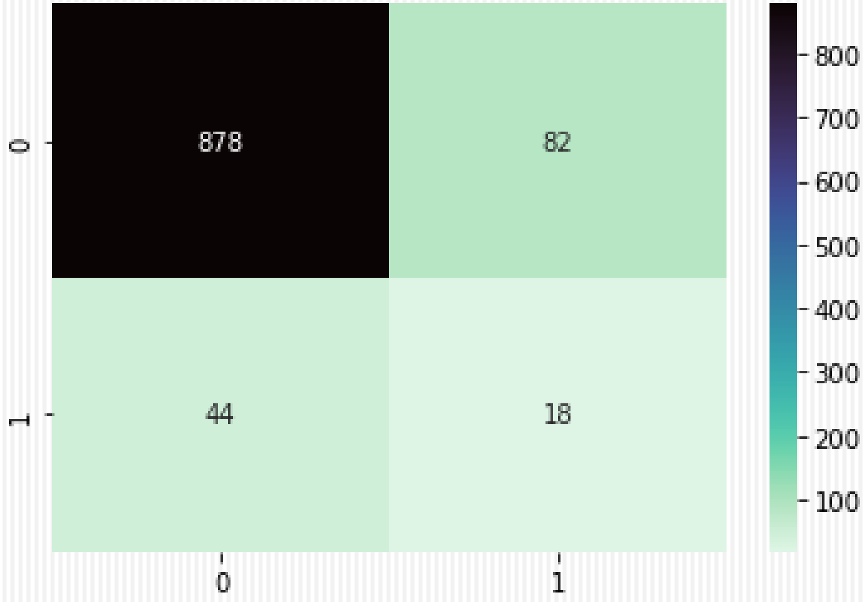


XGBOOST Model

	precision	recall	f1-score	support
0	0.95	0.91	0.93	960
1	0.18	0.29	0.22	62

Accuracy			0.88	1022
Macro avg	0.57	0.60	0.58	1022
Weighted avg	0.91	0.88	0.89	1022

XGBoost gives the best results with 0.88 accuracy and 0.91 recall



❑ Conclusion

The project has answered these predictions:

“Does age have a direct impact on stroke?”

“Are smokers more likely to have stroke?”

“Are people who live in cities at risk of having stroke more than those live in rural?”

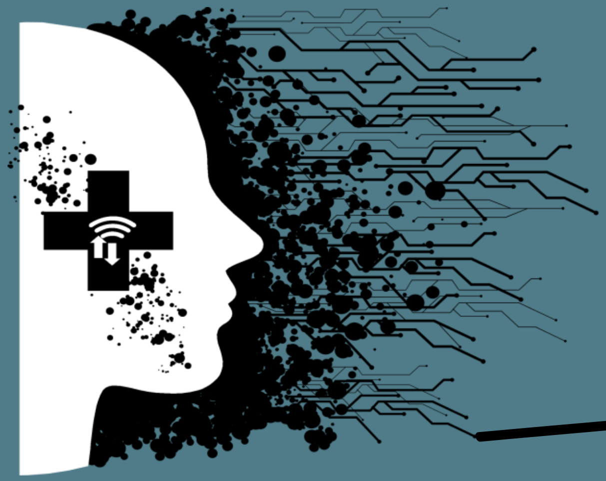
The Best model:

Logistic regression accuracy score 13%

Random forest accuracy score 78%

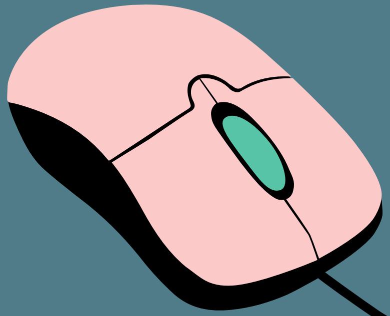
XGBoost accuracy score 88%

Which means that XGBoost is the best model for this dataset.



Thank You

I wish you enjoyed my work ☺



Eman Ahmed Alzhrani