STROKE
PREDICTION

Data Scientist / **Mohamed Adel Hosny**

Supervisor / Doaa Mahmoud Abdel-Aty

Git hub/ https://github.com/Mohamed-adel-Hosny

Kaggle/ https://www.kaggle.com/mohamedadelhosny

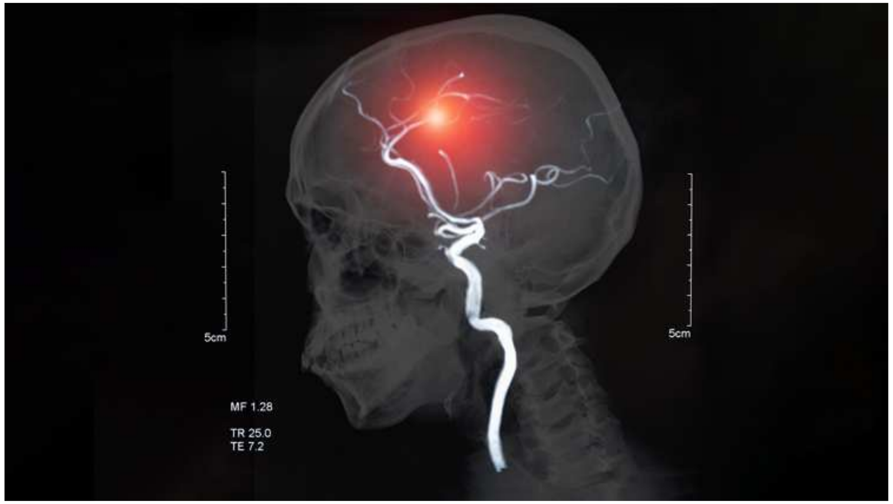
Linkdin/ https://www.linkedin.com/in/mohamed-adel-hosny-692283a3/

Data science challenges

Reveal data secrets

Agenda :

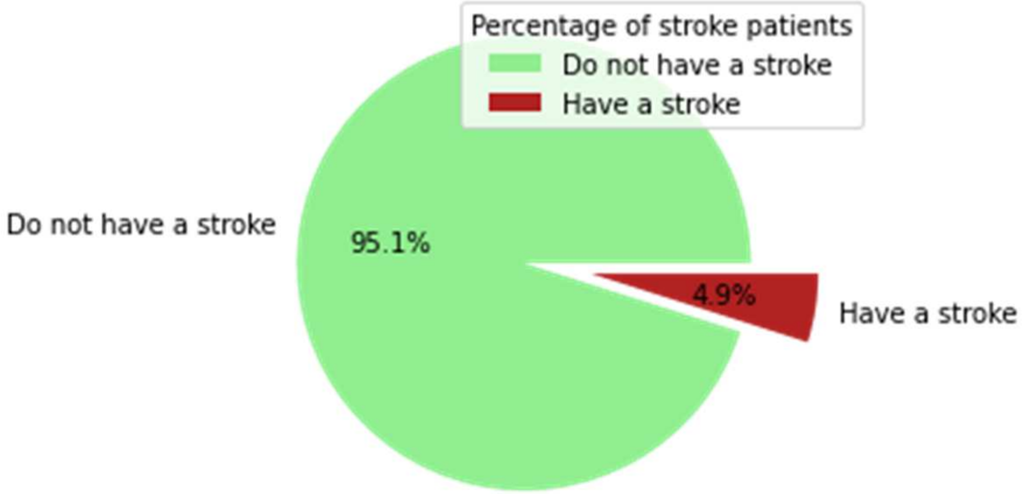
1. Case of study (Data – parameters – Goal).
2. Libraries used on Data.
3. Data cleaning and preprocessing.
4. Exploratory Data.
5. Modeling selection and accuracy.
6. Recommendations



1. Case of study :

- ❑ The stroke is the 2nd leading cause of death globally.
- ❑ It responsible for approximately 11% of total deaths.
- ❑ The data consists of 12 parameter :

  ID
  Gender
  Age
  Hypertension
  heart_disease
  ever_married
  work_type
  Residence_type
  avg_glucose_level
  Bmi
  smoking_status
  Stroke



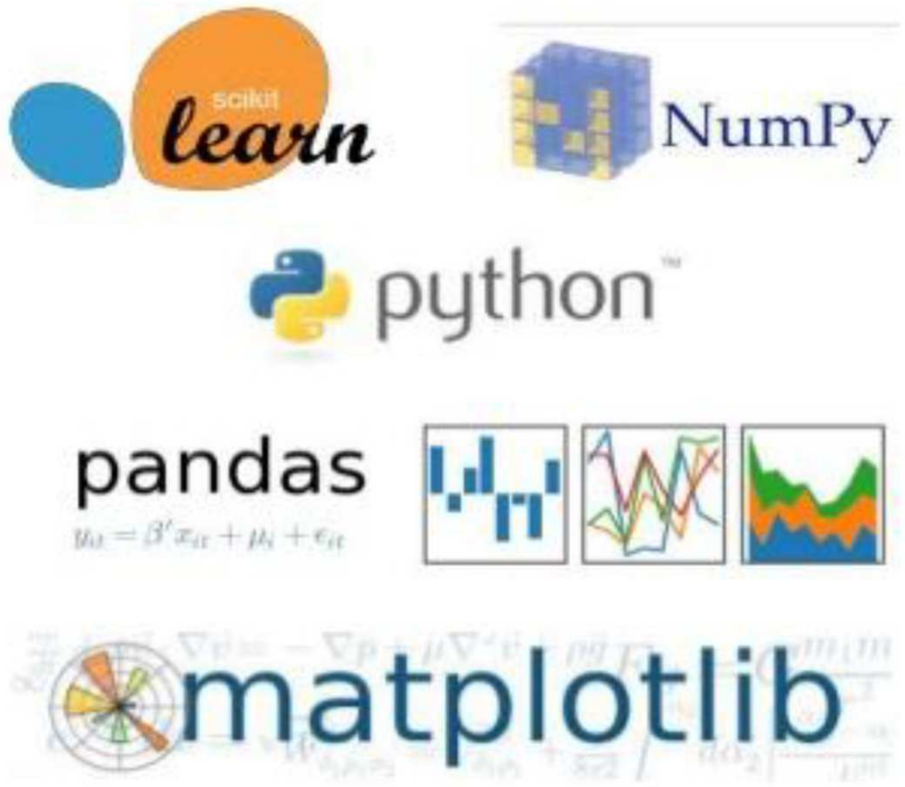
➢ Stroke patients percentage is **4.9%** of the Total Data.

Data science challenges
Reveal data secrets

2. Libraries used on Data :

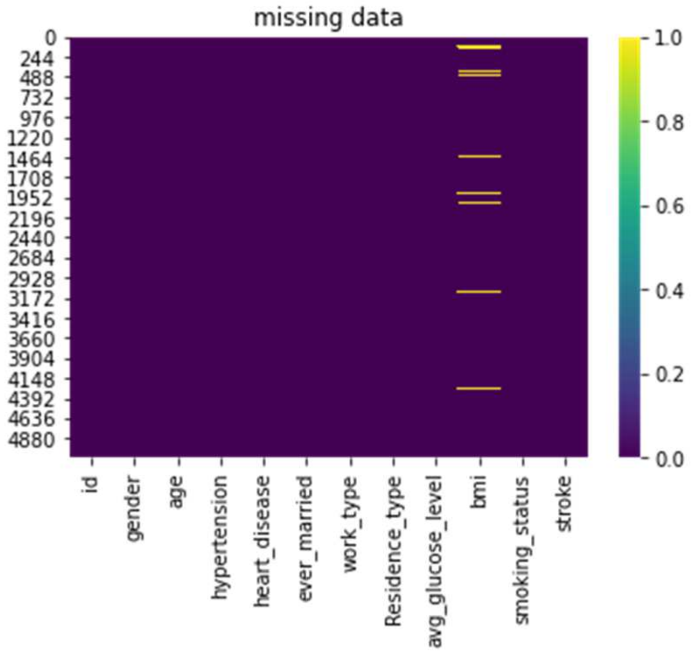
Python libraries used in analysis :

- ***Pandas***
- ***Numpy***
- ***Matplotlib***
- ***seaborn***
- ***Sklearn***
- ***Imblearn***

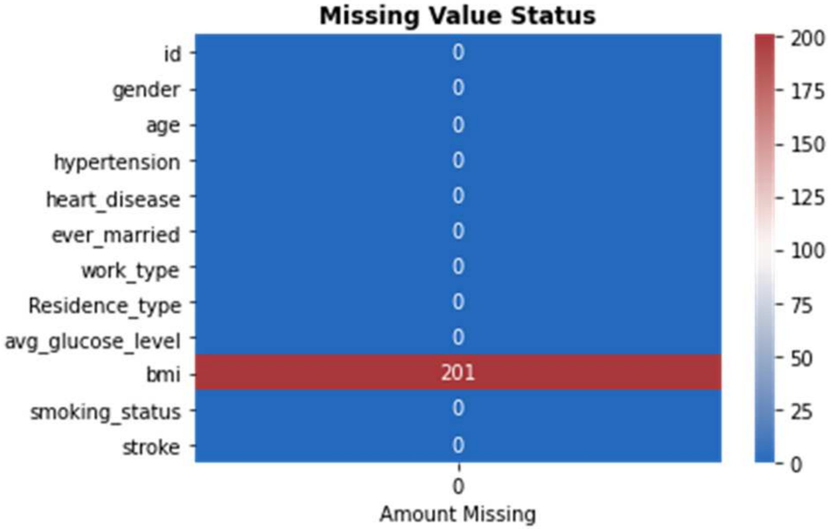


3. Data cleaning and preprocessing :

❑ Missing Value are **3.9%** of the data.



❑ All NaN values exist in **BMI** parameter (*201 values*) .

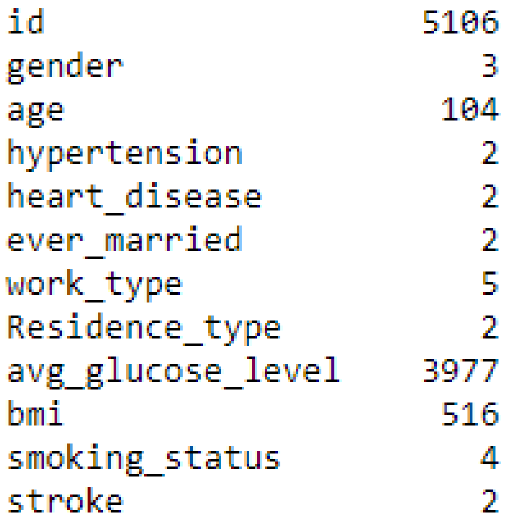


Data science challenges
Reveal data secrets

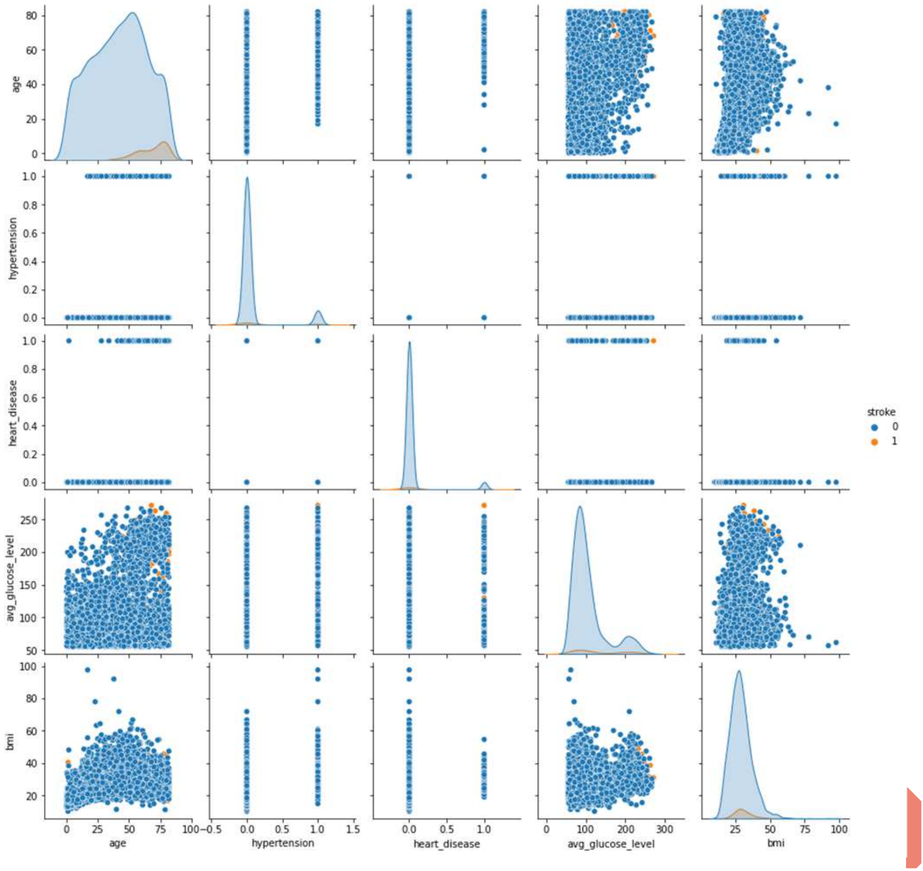
3. Data cleaning and preprocessing :

❑ No. of unique values in Data for each parameters.

❑ Most of the data is classified into categories type with two or more category.



```
id                    5106
gender                   3
age                    104
hypertension             2
heart_disease            2
ever_married             2
work_type                5
Residence_type           2
avg_glucose_level     3977
bmi                    516
smoking_status           4
stroke                   2
```

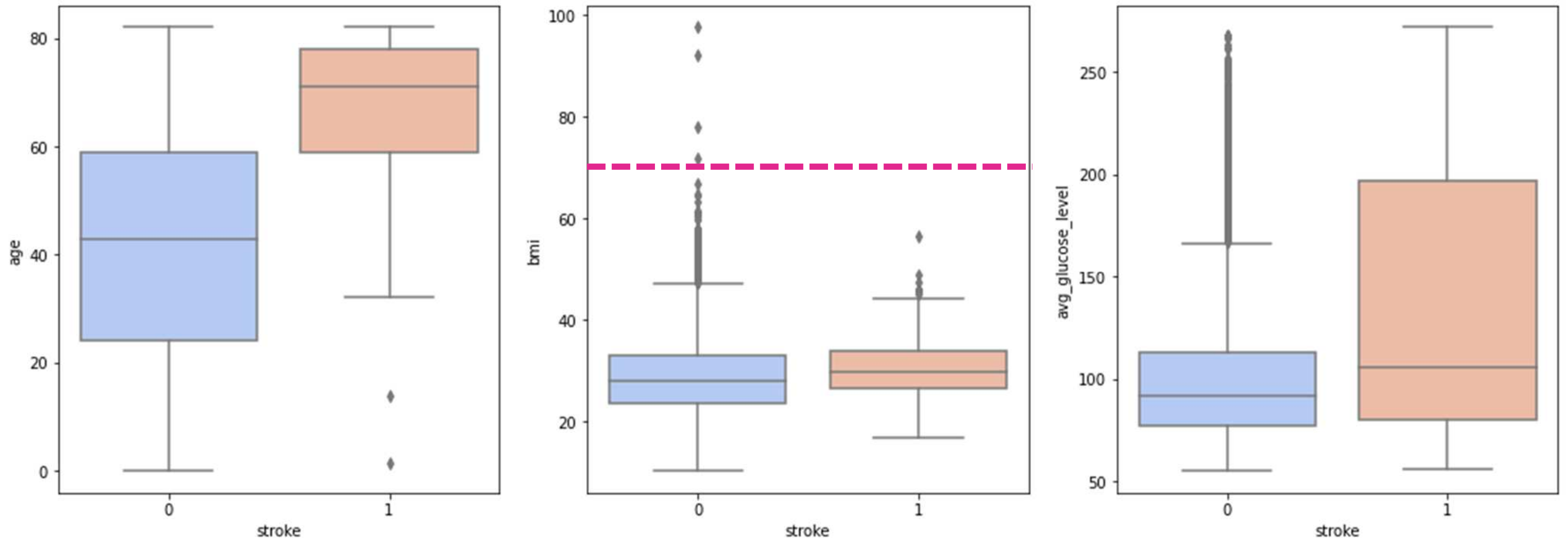


Data science challenges

Reveal data secrets

3. Data cleaning and preprocessing :

- ❑ There is 4 readings outliers in **BMI (Body Mass Index)**.
- ❑ As number of outliers is very small, we can drop it.

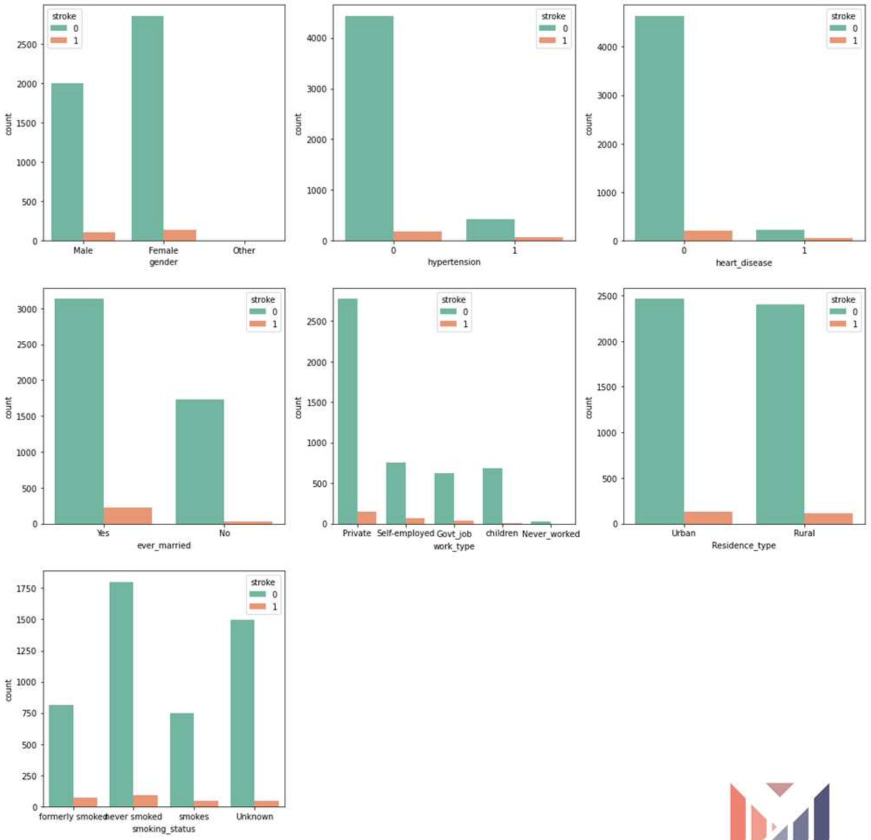


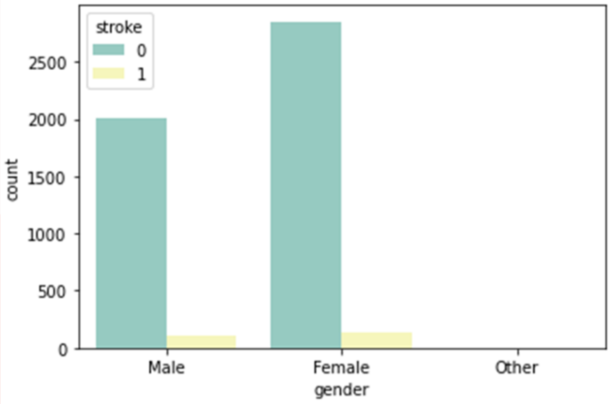
Data science challenges
Reveal data secrets

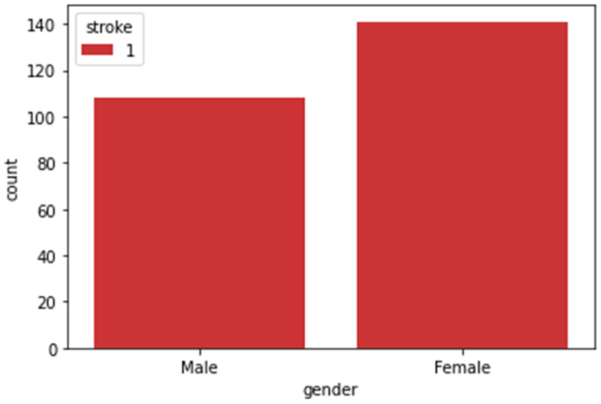
4. Exploratory Data :

- **Gender**

❑ Gender has no interference to can predict the probability for person to get a stroke or not.







Data science challenges

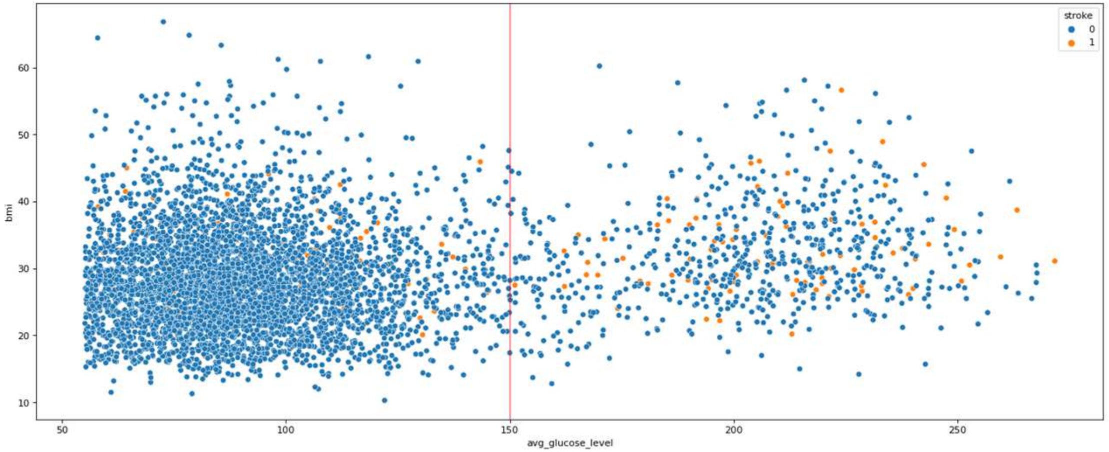
Reveal data secrets

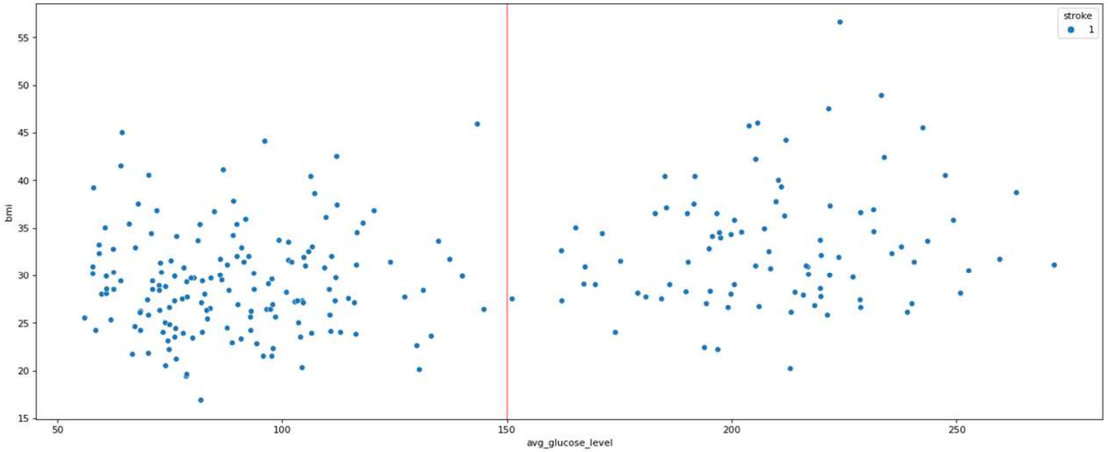
4. Exploratory Data :

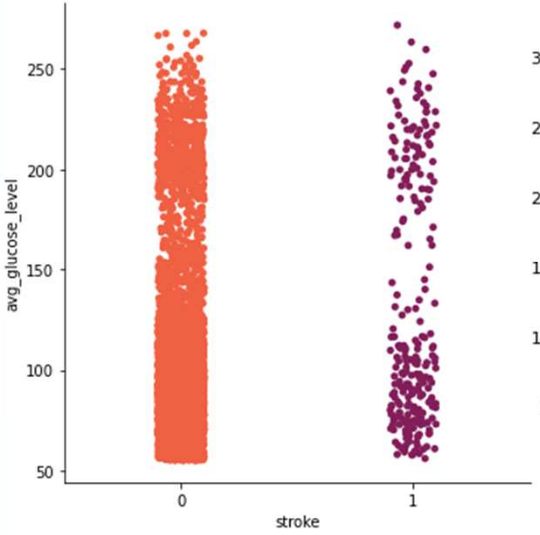
- **Avg Glucose Level**

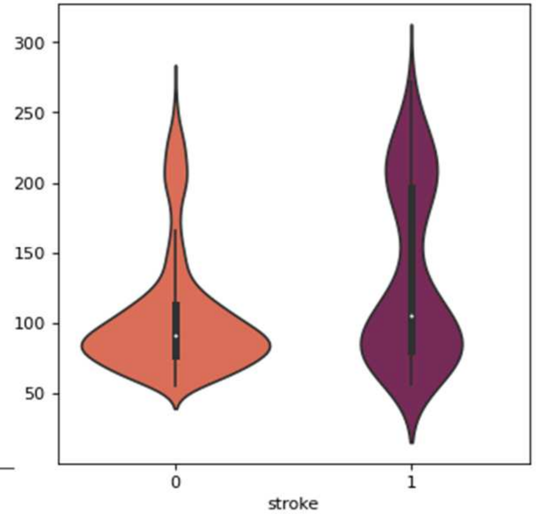
❑ The data can be split into two category :

(**Normal patient** and **Diabetes patients**)









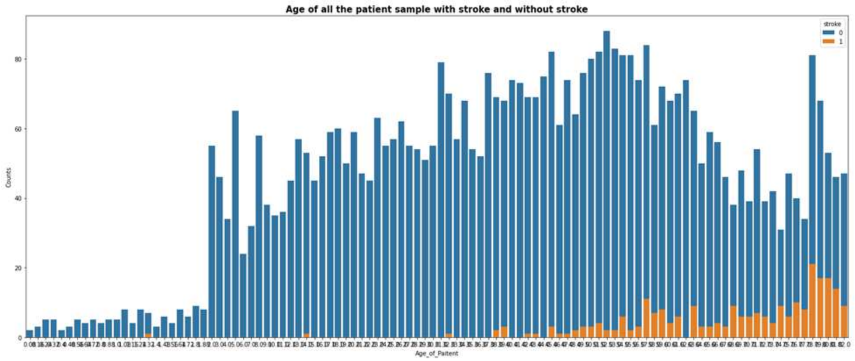
Data science challenges

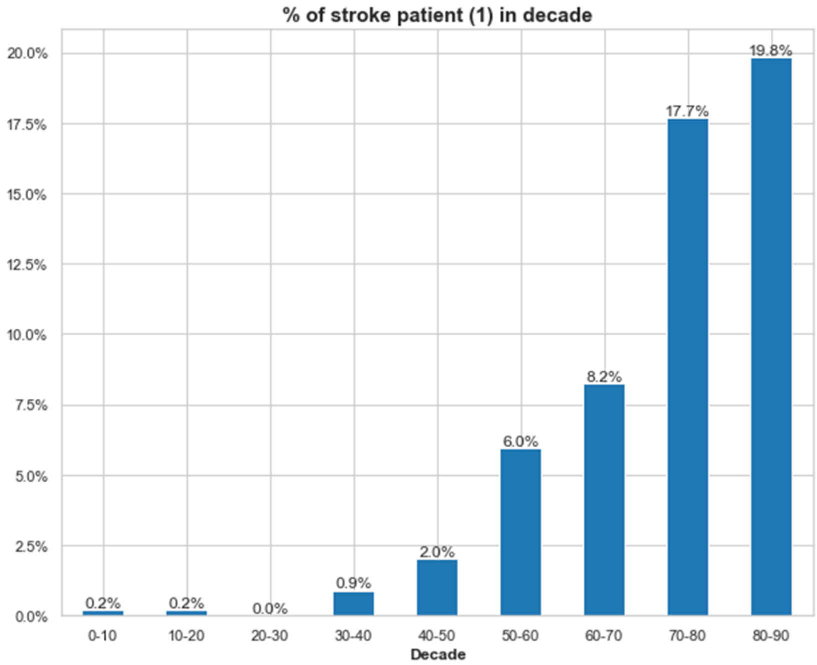
Reveal data secrets

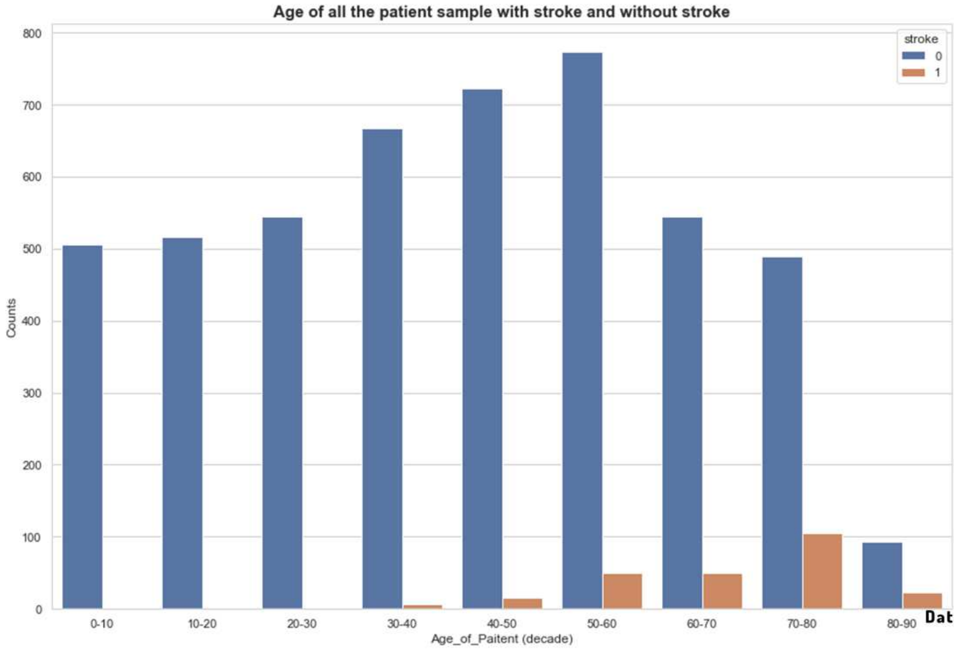
4. Exploratory Data :

- **Age**

❑ Recommendation for people with age more than 40 to check up with a doctor





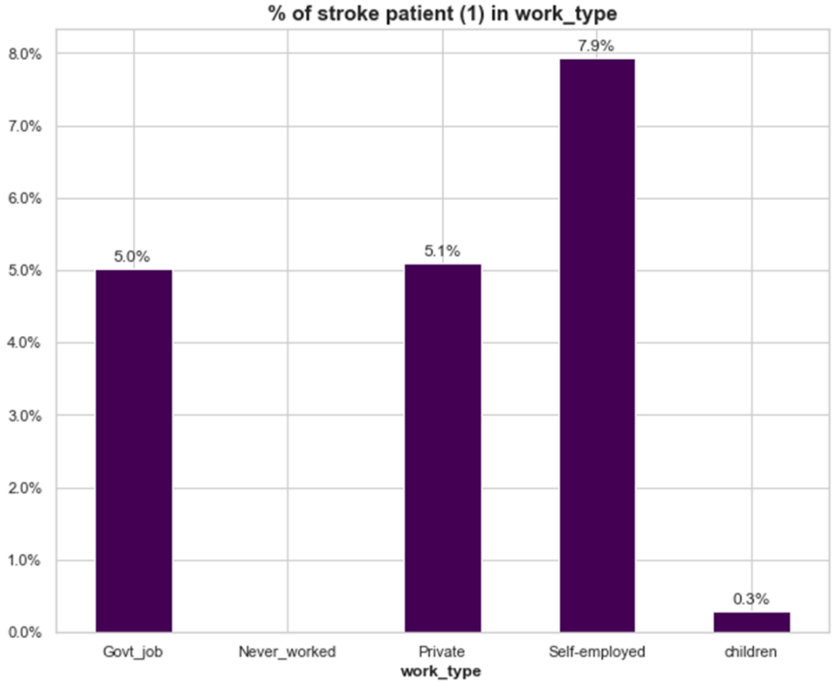


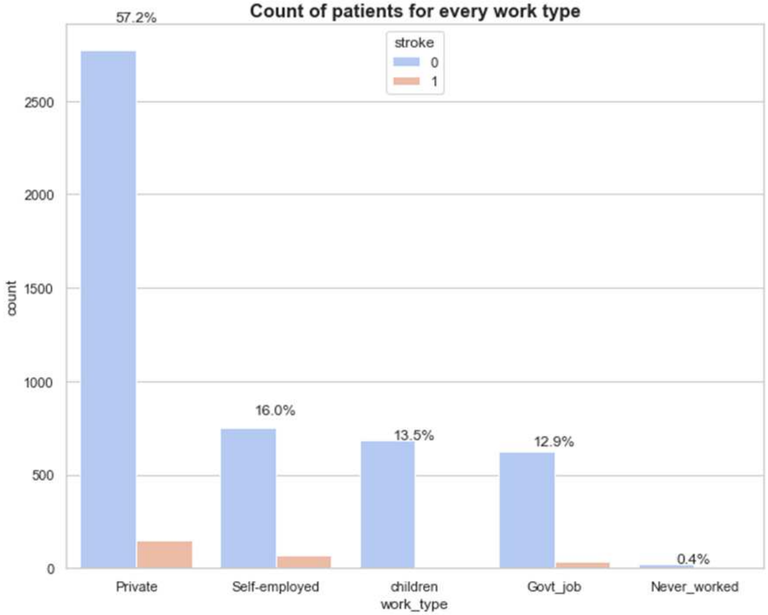
Data science challenges
Reveal data secrets

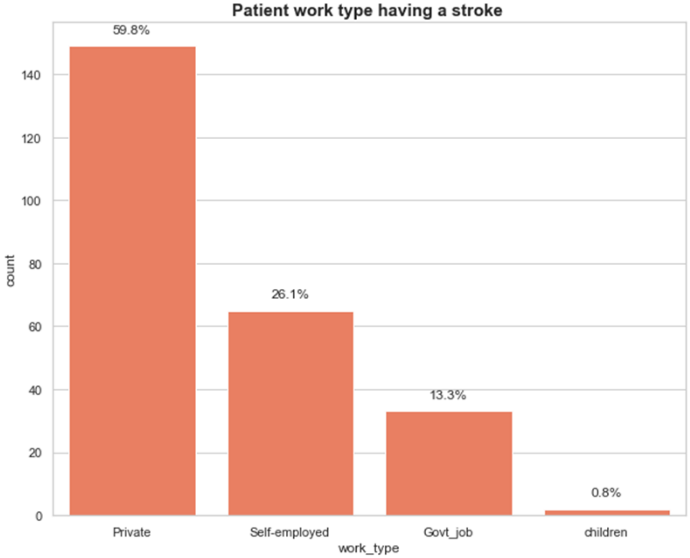
4. Exploratory Data :

- **Work type**

❑ Self employment job have the most effect on people to get a stroke.







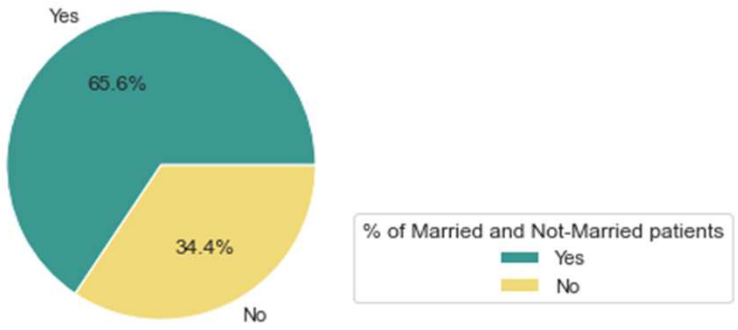
)ata science challenges

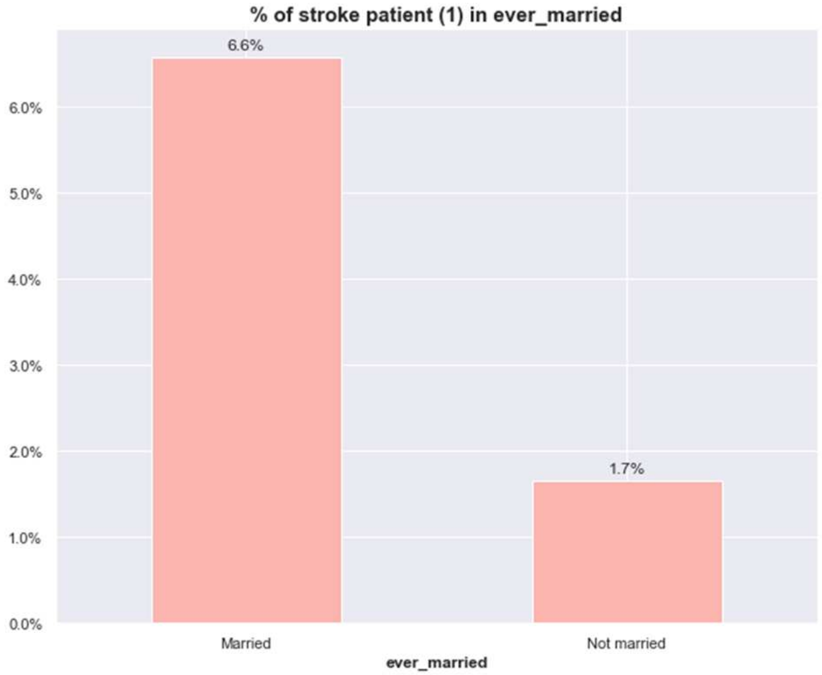
Reveal data secrets

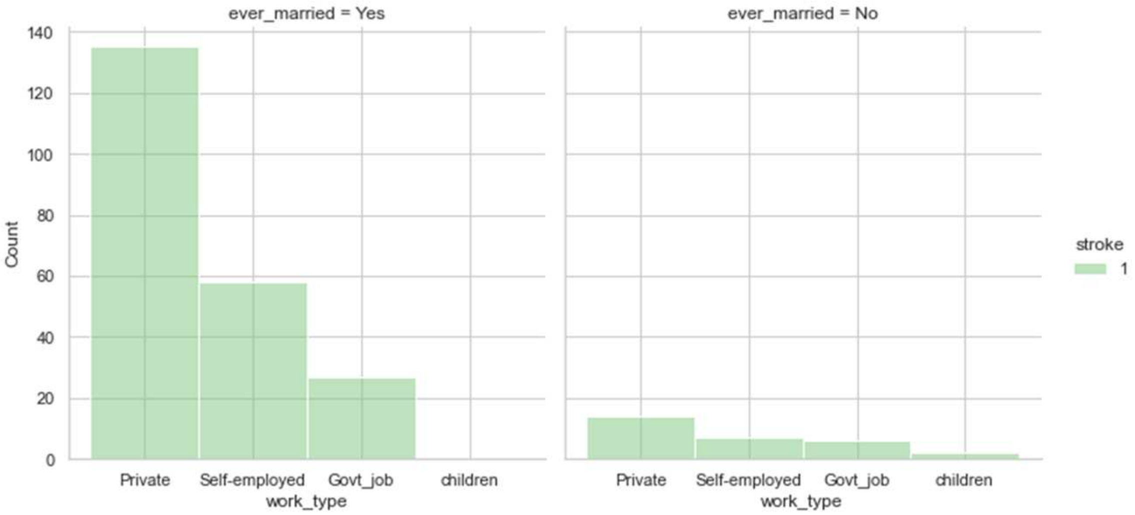
4. Exploratory Data :

- **Ever Married**

❑ The most of people get a stroke are the married one.







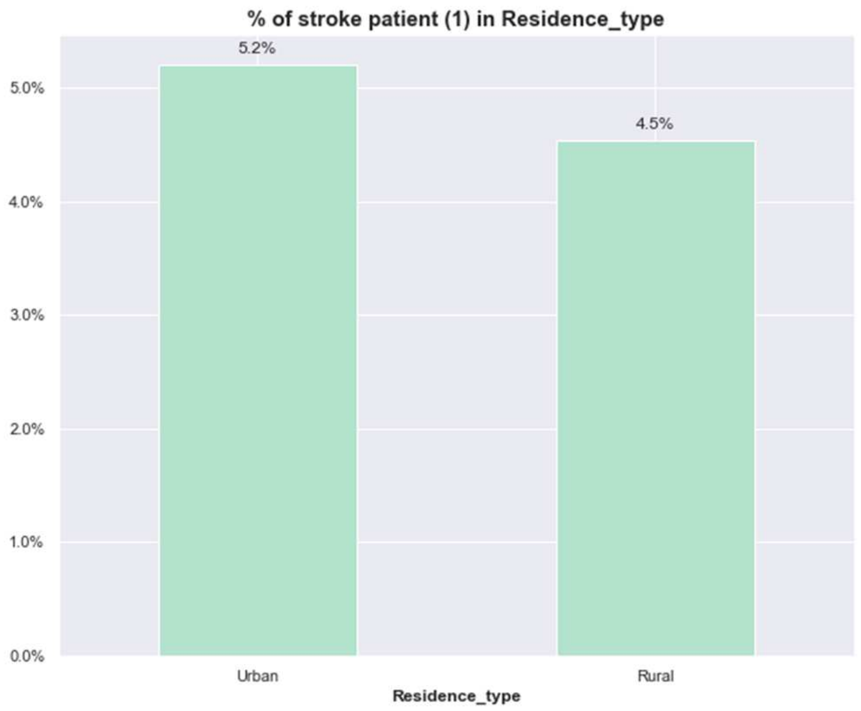
Data science challenges

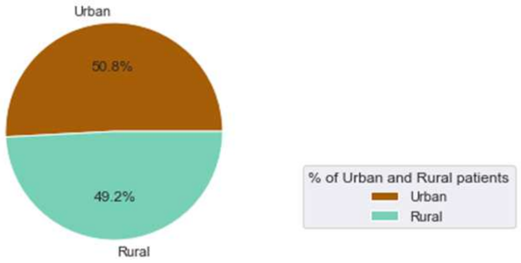
Reveal data secrets

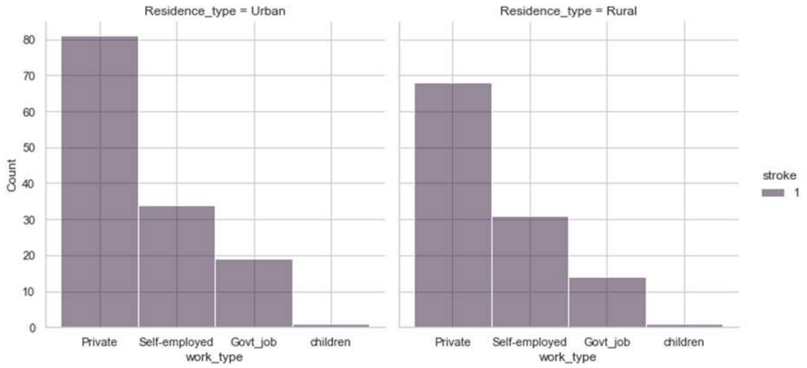
4. Exploratory Data :

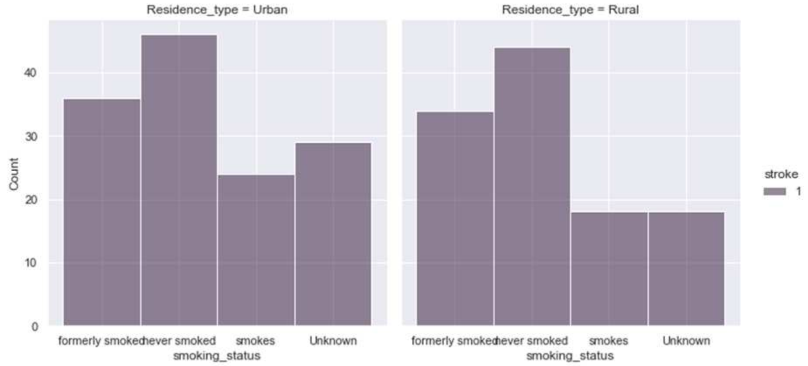
- **Residence type**

❑ The Residence type not affects the reasons to get stroke.









Data science challenges

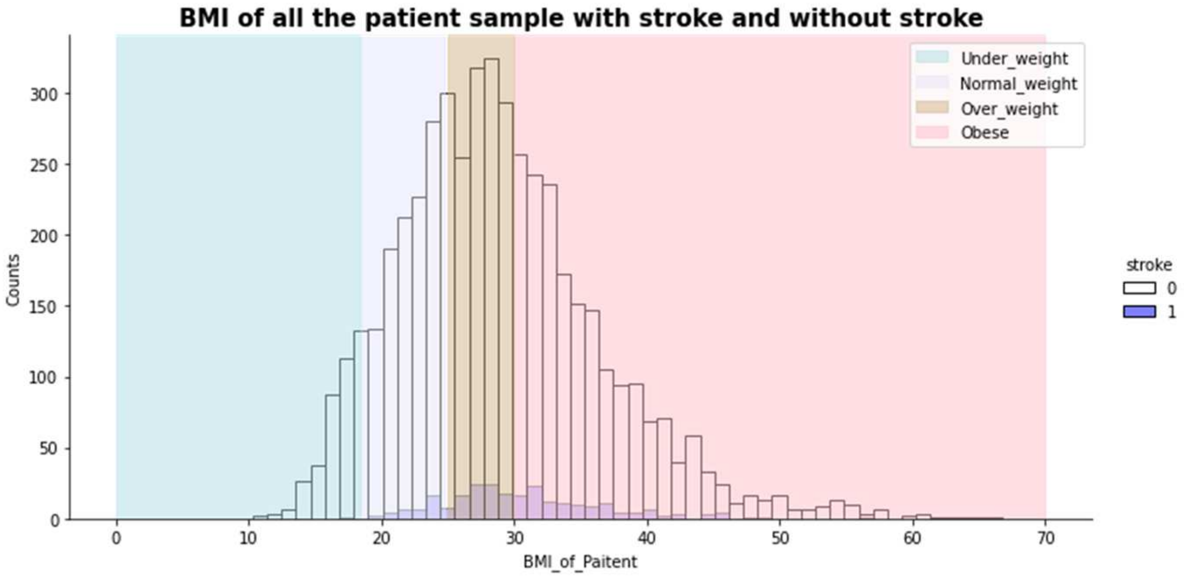
Reveal data secrets

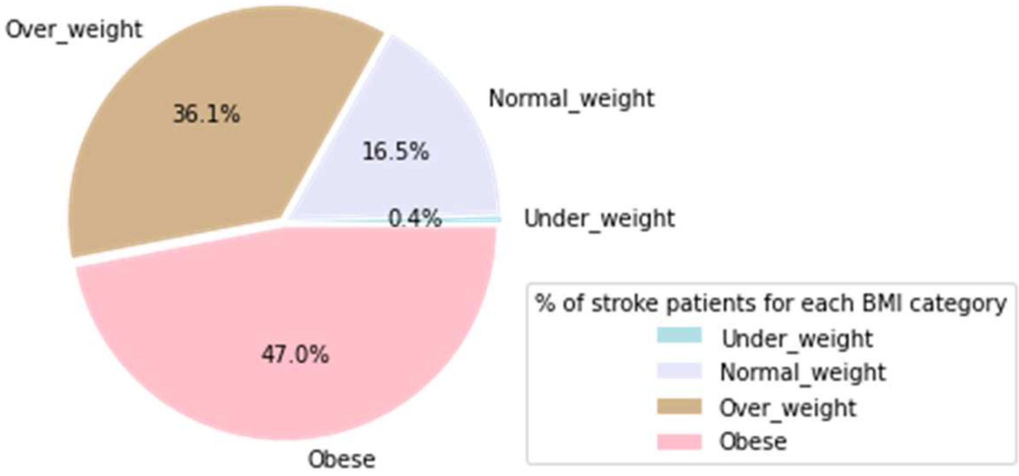
4. Exploratory Data :

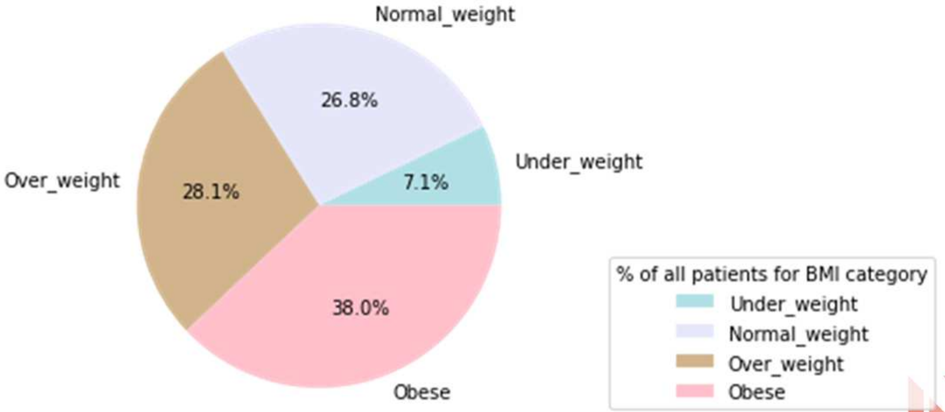
- **BMI (Body Mass Index)**

❑ The data can be split into four category :

[**Under_weight** - **Normal_weight** - **Over_weight** - **Obese**]





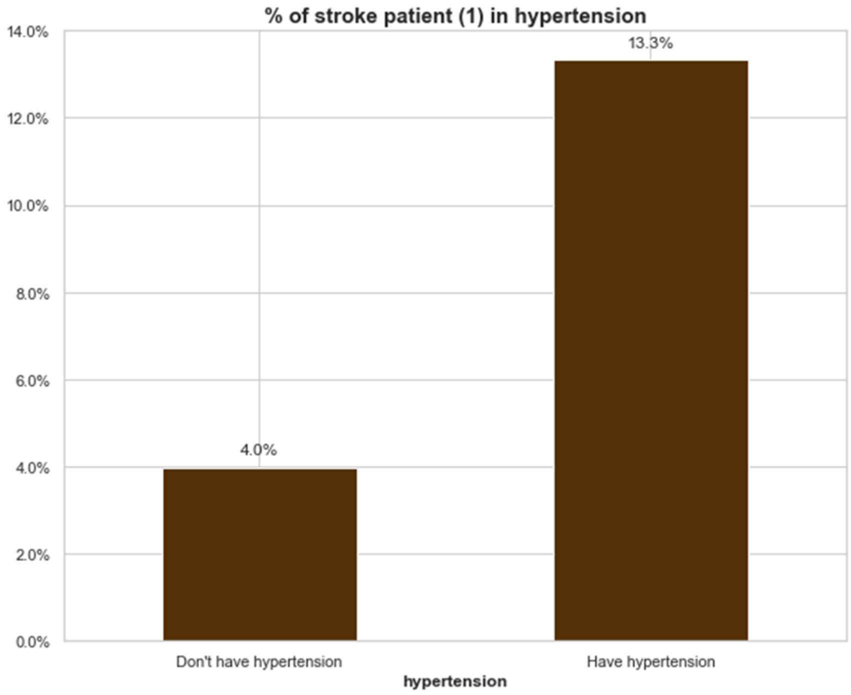


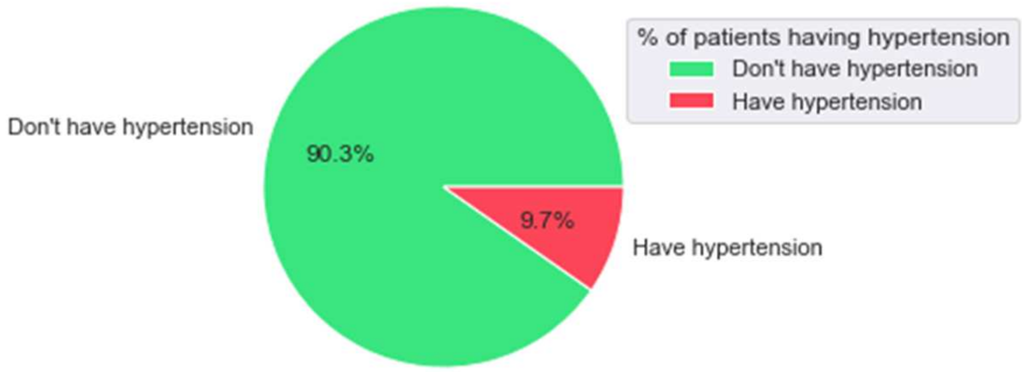
Data science challenges
Reveal data secrets

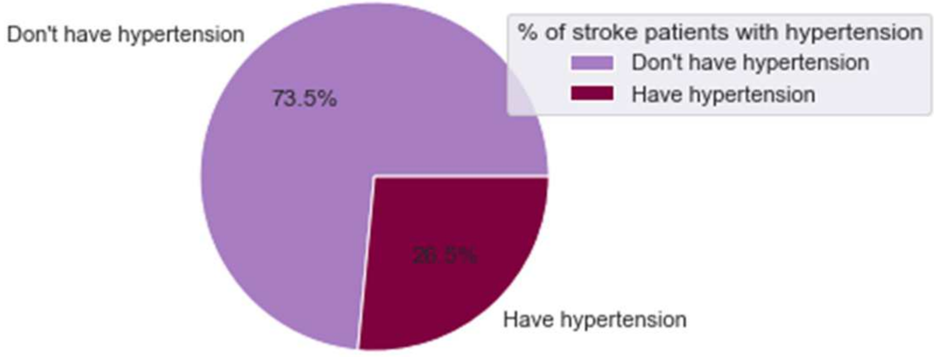
4. Exploratory Data :

- **Hypertension**

❑ The most of people get a stroke are the married one.







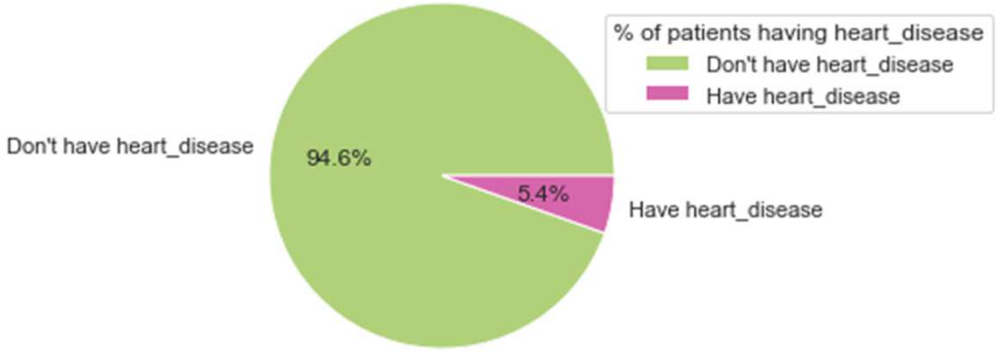
Data science challenges

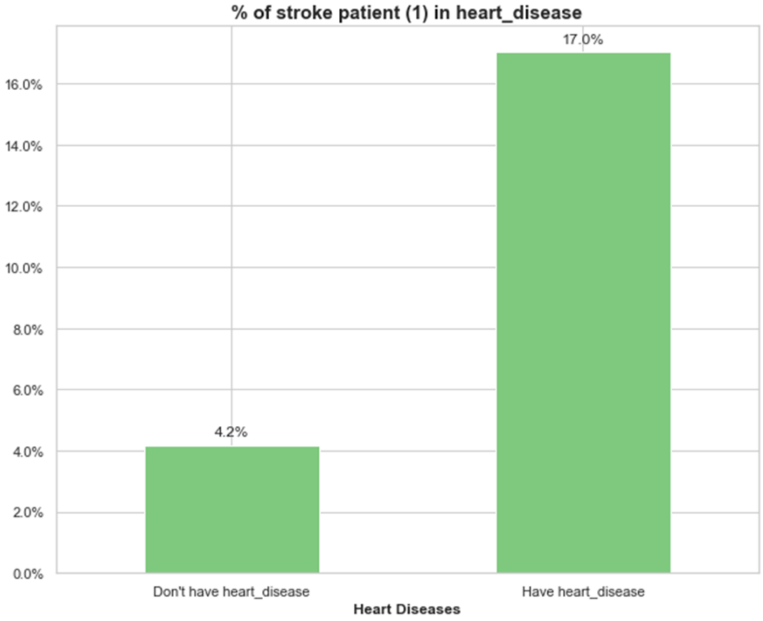
Reveal data secrets

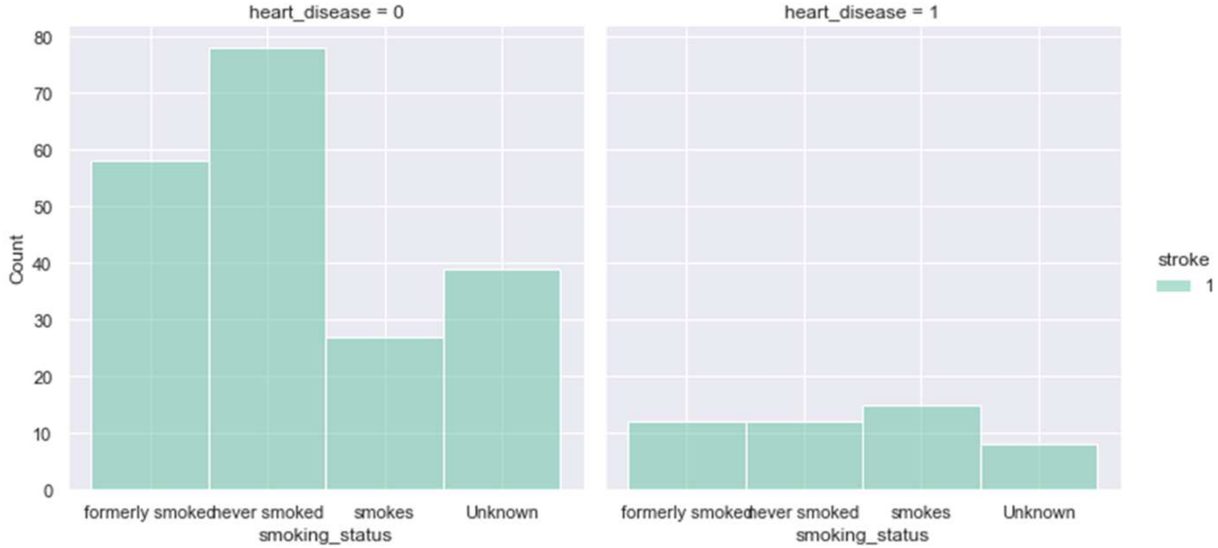
4. Exploratory Data :

- **Heart Disease**

❑ Patient with heart disease have to quit smoking as they are most likely to have a stroke.







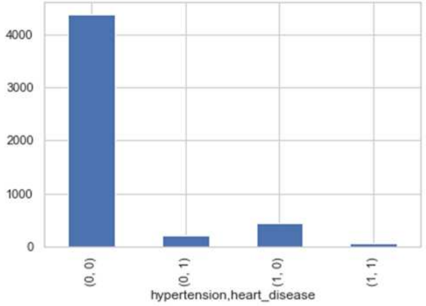
Data science challenges

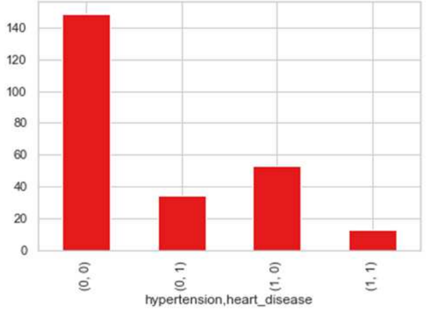
Reveal data secrets

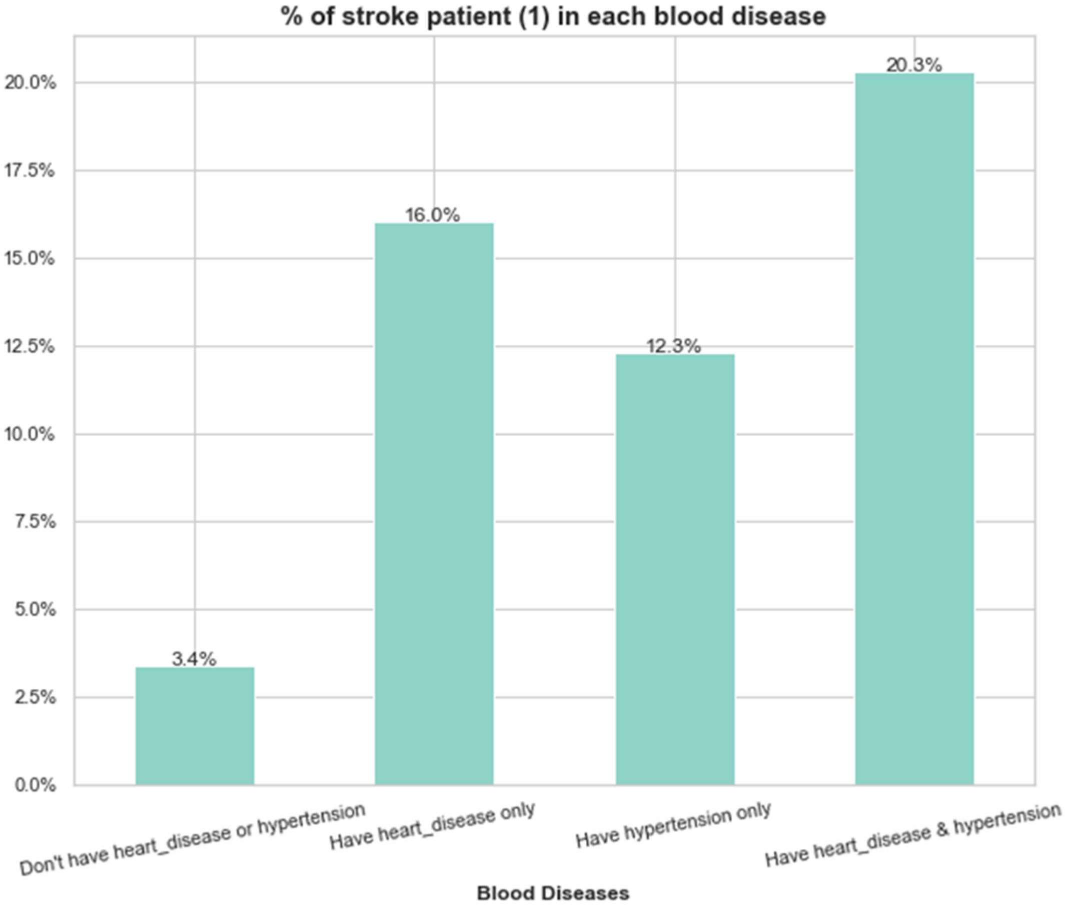
4. Exploratory Data :

- **Heart Disease & Hypertension**

❑ People who has hypertension and heart diseases are the most to get stroke.





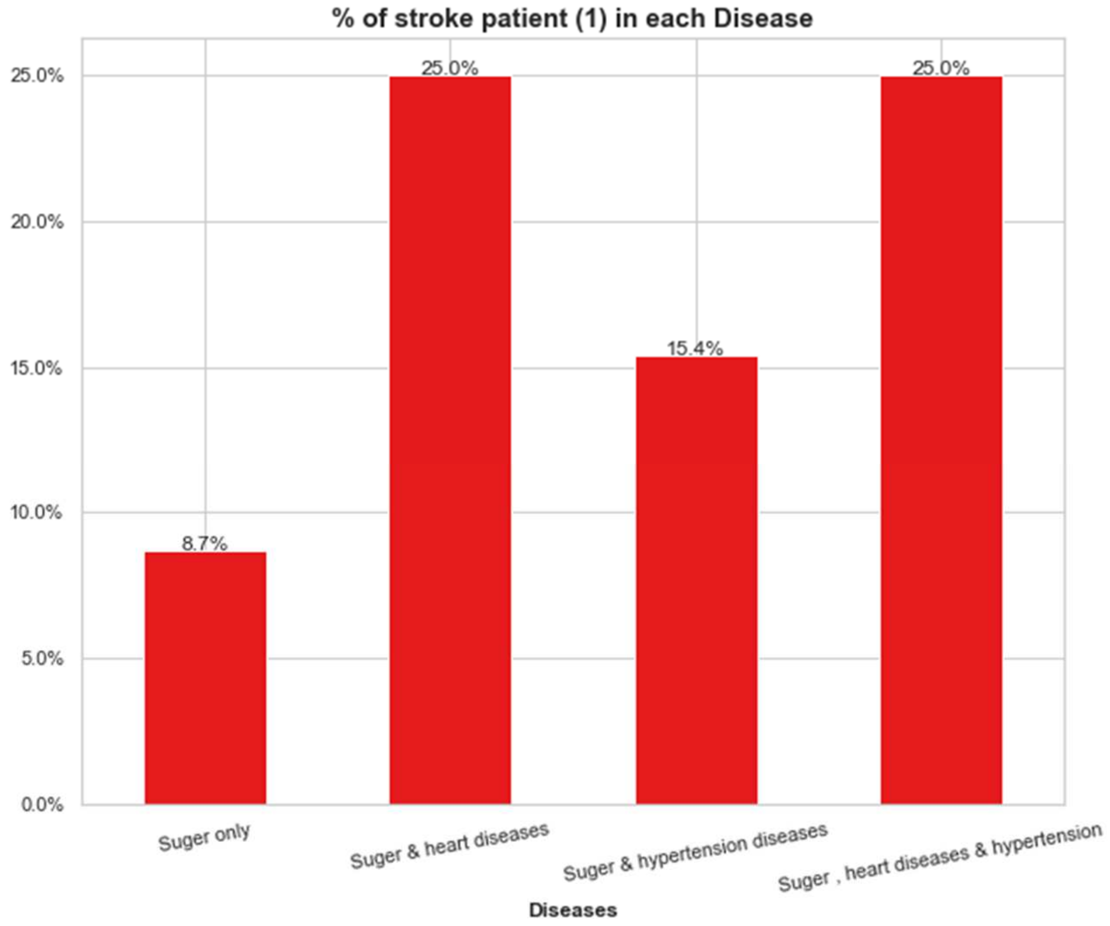


Data science challenges
Reveal data secrets

4. Exploratory Data :

- **Heart Disease & Hypertension with Diabetes patients**

❑ More than **Half** of patient with sugar disease who have stroke are also heart disease.



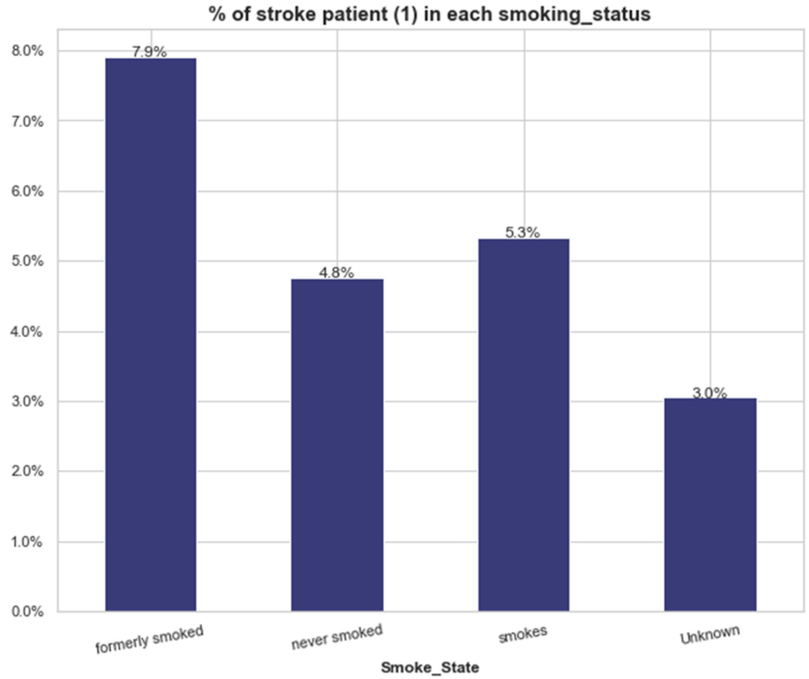
Data science challenges

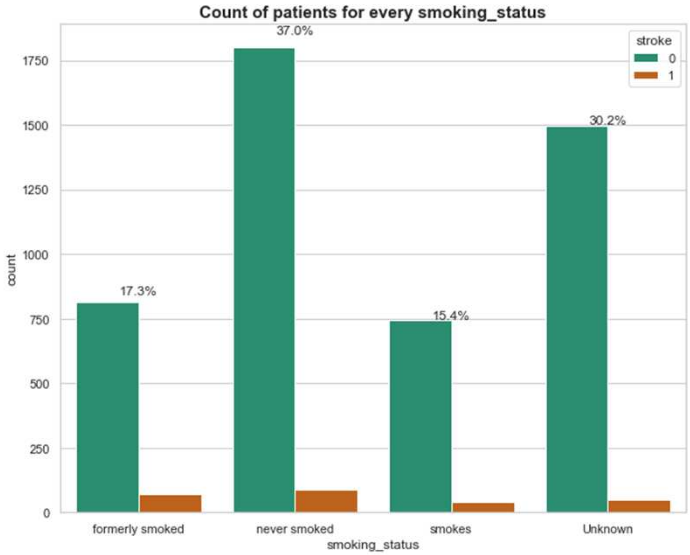
Reveal data secrets

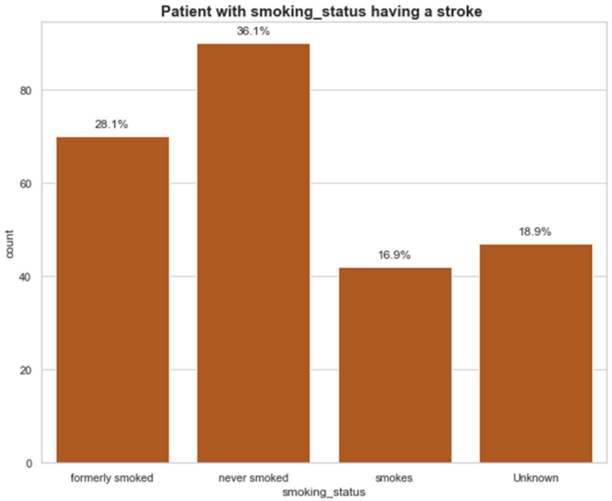
4. Exploratory Data :

- **Smoking state**

❑ The smoking people who have quite smoking or still smoking are more probably to get stroke than who don't never smoked.







Data science challenges

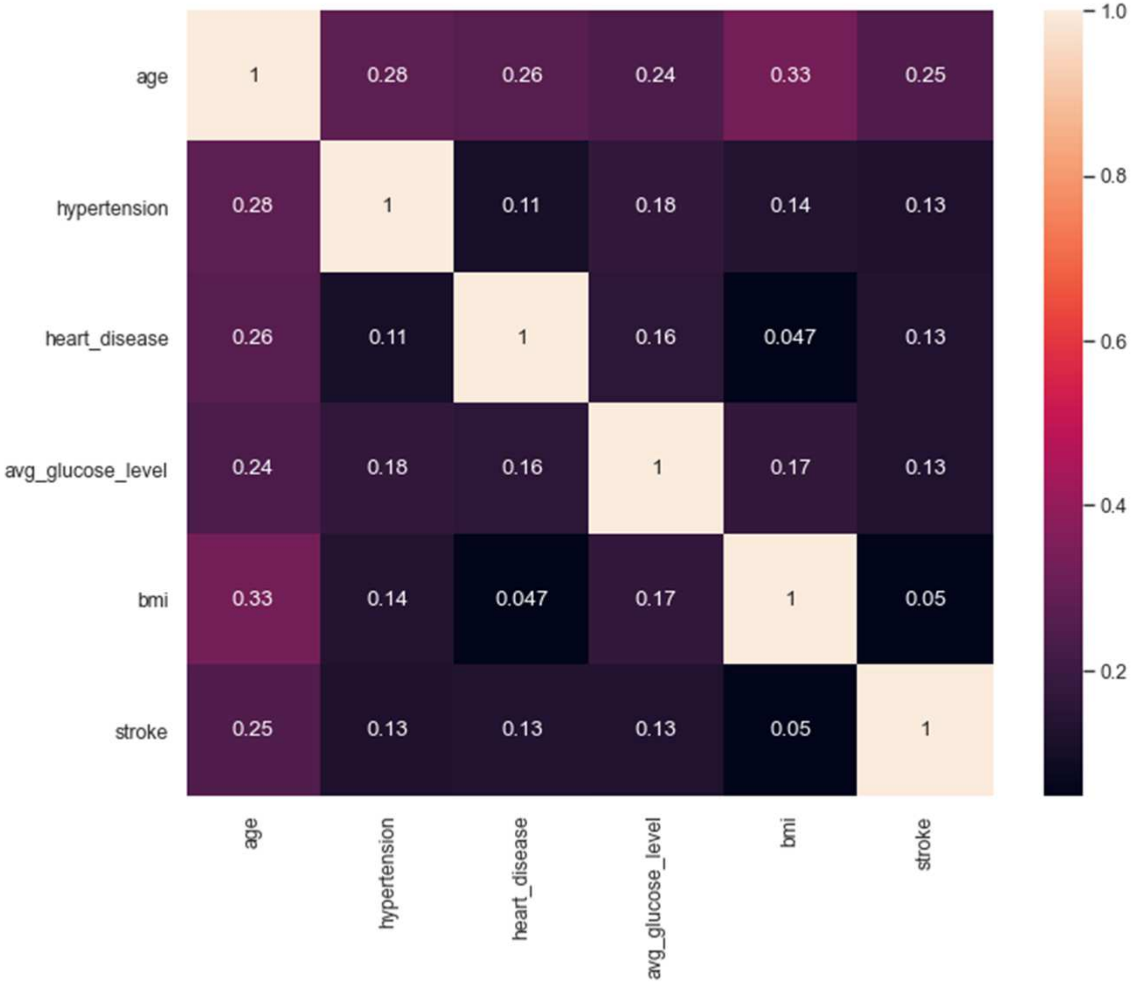
Reveal data secrets

5. Modeling selection and accuracy :

- **Feature Selection**

❑ Linear correlation between features and target prediction value is very low.

❑ ID & gender features are not effecting the probability of patients to have a stroke, So they eliminated from selection.



	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
age	1	0.28	0.26	0.24	0.33	0.25
hypertension	0.28	1	0.11	0.18	0.14	0.13
heart_disease	0.26	0.11	1	0.16	0.047	0.13
avg_glucose_level	0.24	0.18	0.16	1	0.17	0.13
bmi	0.33	0.14	0.047	0.17	1	0.05
stroke	0.25	0.13	0.13	0.13	0.05	1

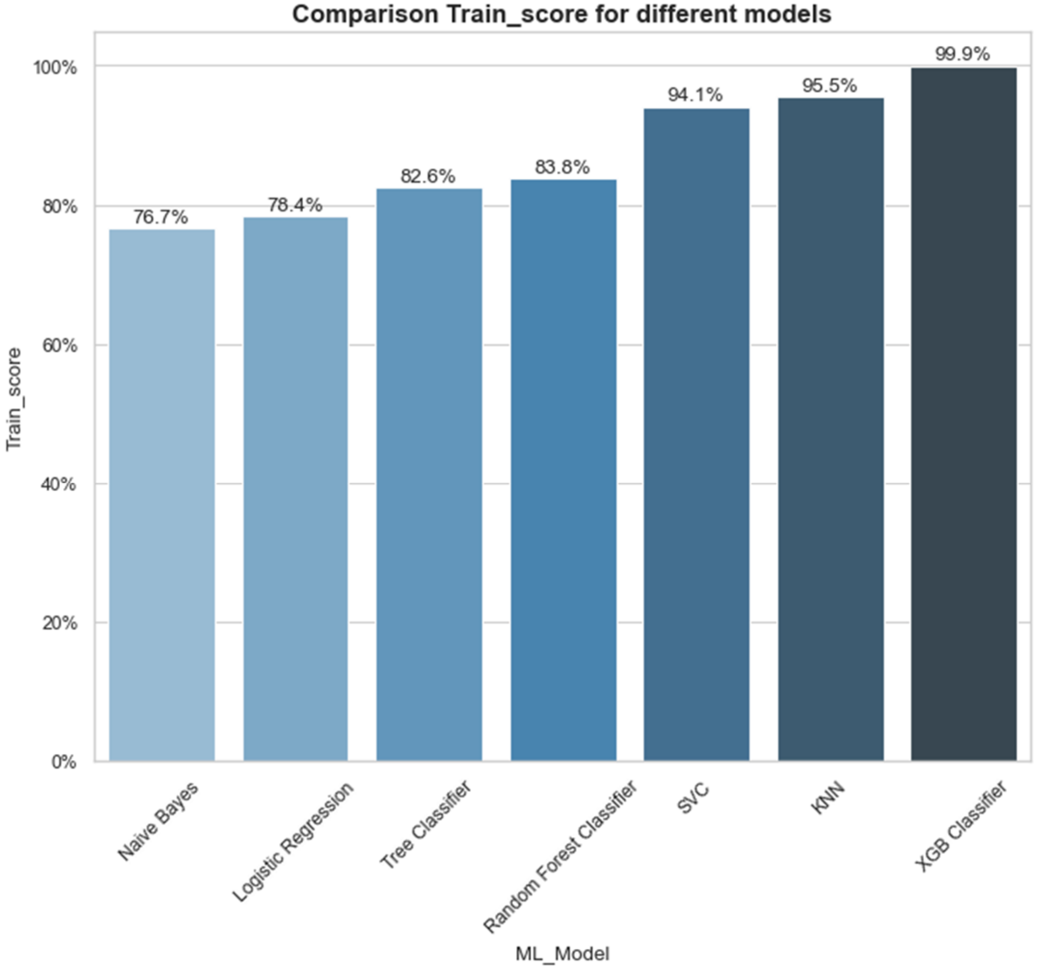
Data science challenges
Reveal data secrets

5. Modeling selection and accuracy :

- **Model Selection**

❑ Different model have been tested with Data as showing in the graph the accuracy for XGB classifier model is the best accuracy **(99.9%) .**

❑ KNN model and SVC have also good accuracy.

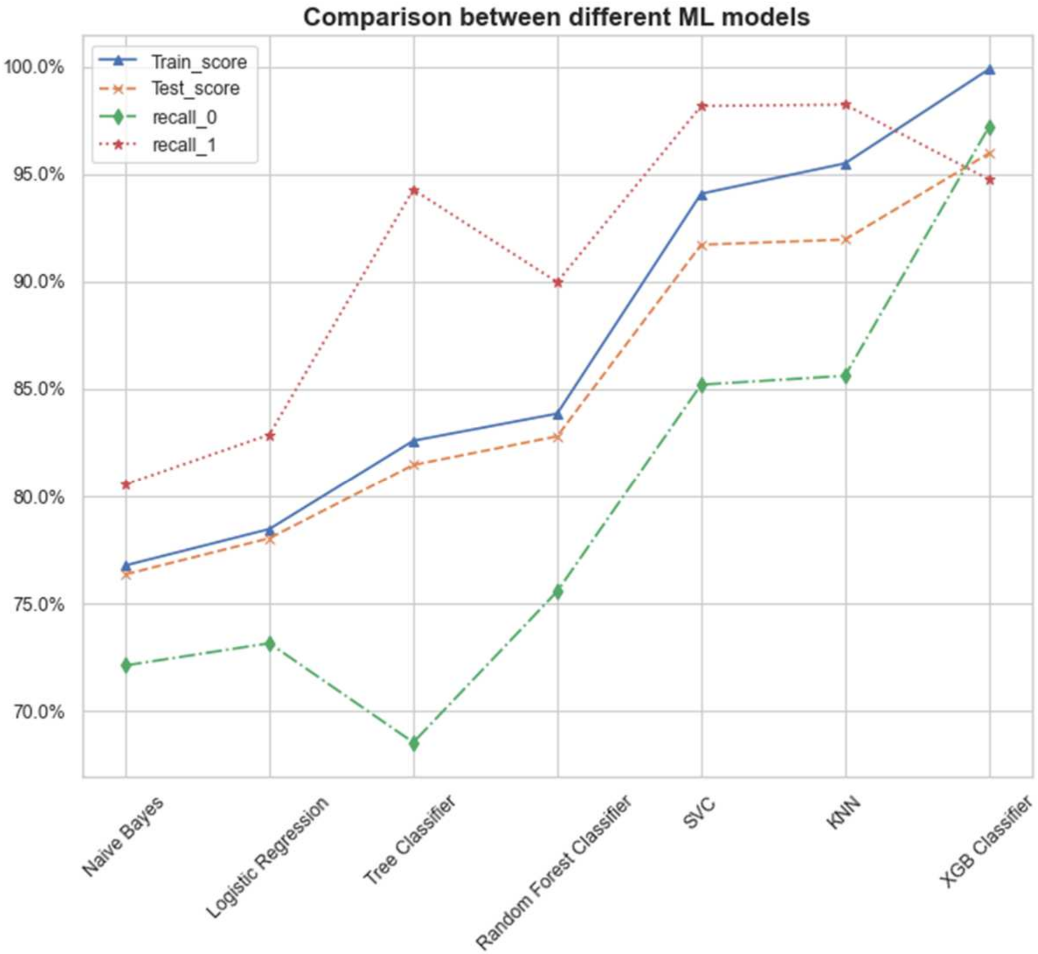


Data science challenges
Reveal data secrets

5. Modeling selection and accuracy :

- **Model Selection**

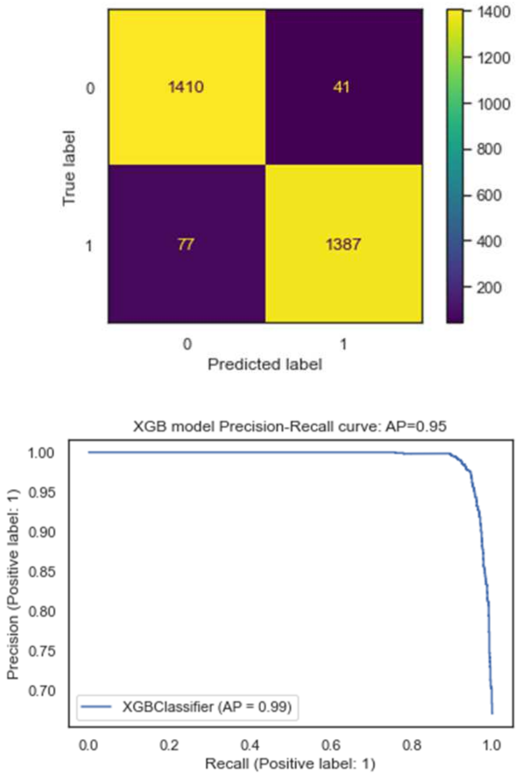
❑ Recall for XGB classifier is good beside the score of train and test data which make best model to select is **XGB Classifier**.



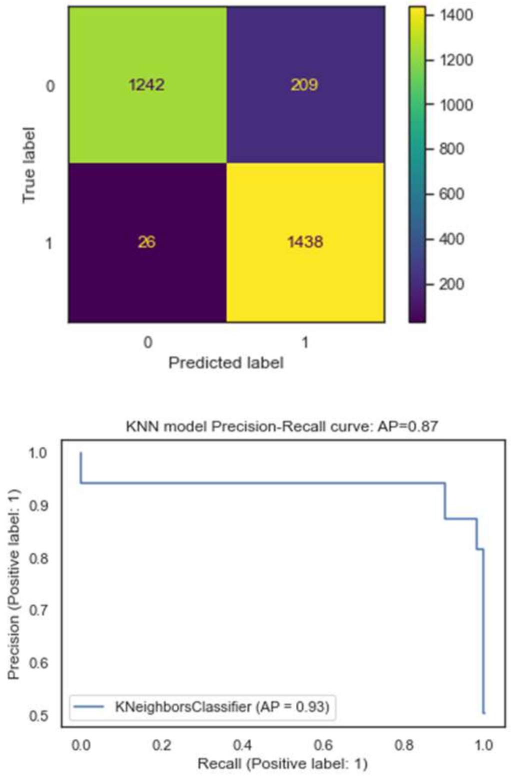
Data science challenges
Reveal data secrets

5. Modeling selection and accuracy :

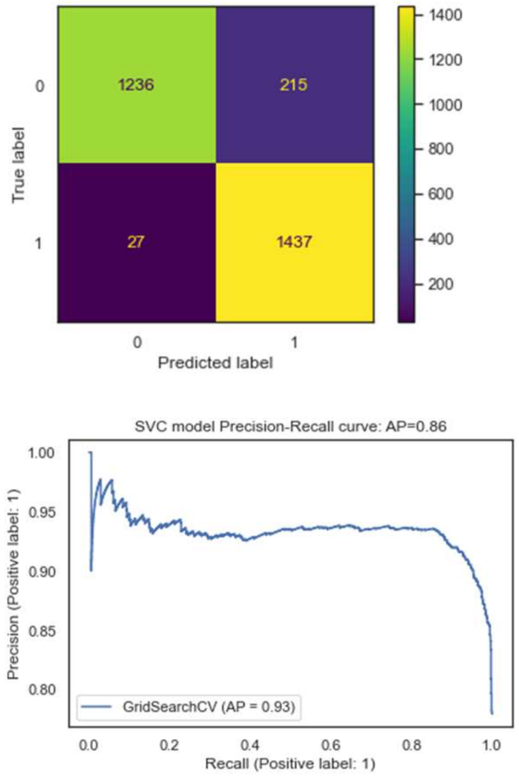
- **XGB Classifier**



- **KNN classifier**



- **SVC**



Data science challenges

Reveal data secrets

5. Modeling selection and accuracy :

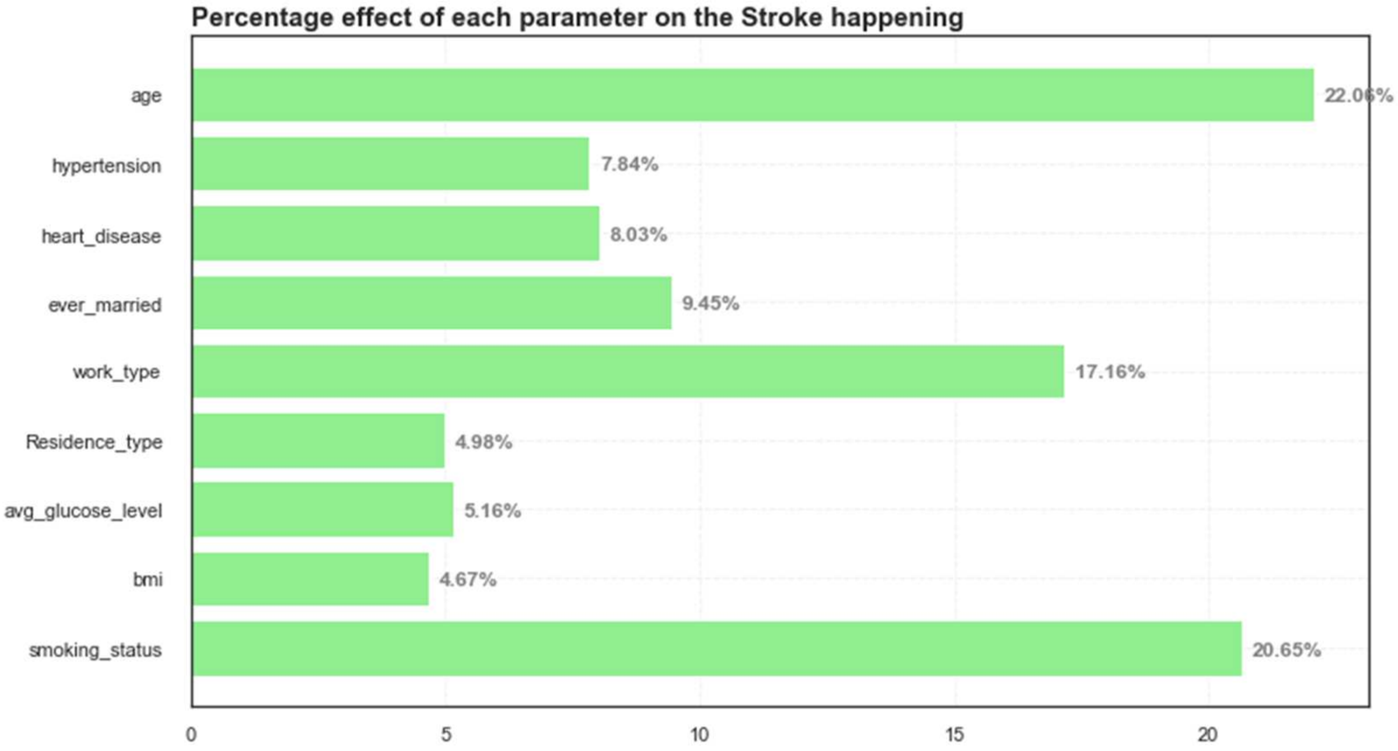
- **XGB Classifier**

❑ The Feature Importance represents Three category for importance :

I. **Strong effect**
(age – work state
– smoking state)

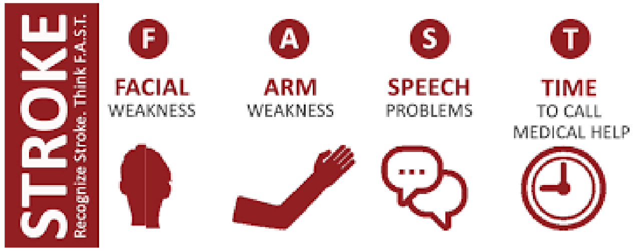
II. **Moderate effect**
(hypertension – heart disease
– ever married)

III. **Weak effect**
(Residence type – BMI
– avg. glucose level)



Data science challenges
Reveal data secrets

6. Recommendations :



- The smoking people try to quite smoking.
- If you work as self employed, try to make a frequent medical check up for any of heart disease or hypertension.
- The smoking people who have quite smoking or still smoking are more probably to get stroke than who don't never smoked.
- People with Overweight need to try health food and daily exercises.
- Patient with heart disease have to quit smoking as they are most likely to have a stroke.
- If you feel of weakness in your face or arm and have a speech problems, go immediately to nearest hospital for medical care.

Data science in challenge projects

Reveal data secrets

Thank You

Tele: 01118669802
Mail: m_adel_hosny@hotmail.com
3/8/2021