Stroke Prediction

11 clinical features for predicting stroke events

Context

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 relevant information about the patient.

Attribute Information

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) 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

Approaches

- Preprocessing Data before Exploratory Data Analysis

- Exploratory Data Analysis on Stroke Prediction Data

- Preparing the Data for Prediction

- Creating a Model for Stroke Prediction

Preprocessing Data before Exploratory Data Analysis

1. Using round() to round off Age.

2. Setting values to NaN where BMI is less than 12 and greater than 60 as these can be considered as outliers.

3. Sorting the Data Frame first based on Gender then on Age and using Forward Filling to fill those missing BMI values.

Exploratory Data Analysis

1. Checking if the Data is Balanced.

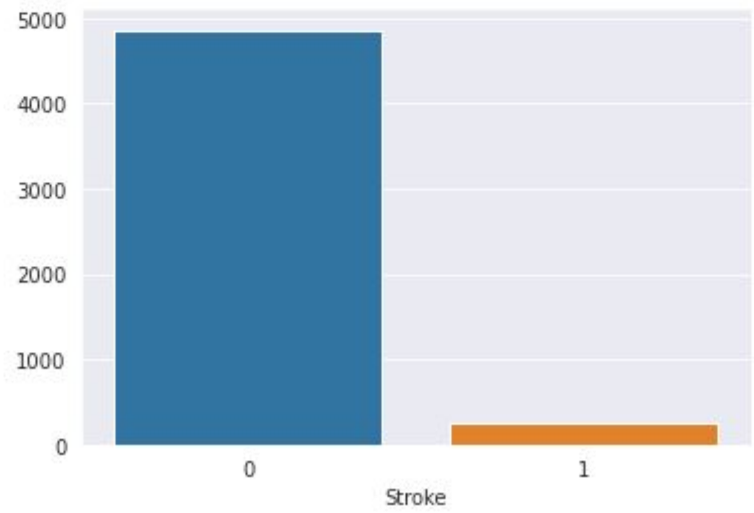
2. Plotting various graphs to check relation between the each column with respect to stroke.

    Age vs BMI with Stroke

    BMI vs Avg Glucose Level

    Percentage of people who got stroked in each category

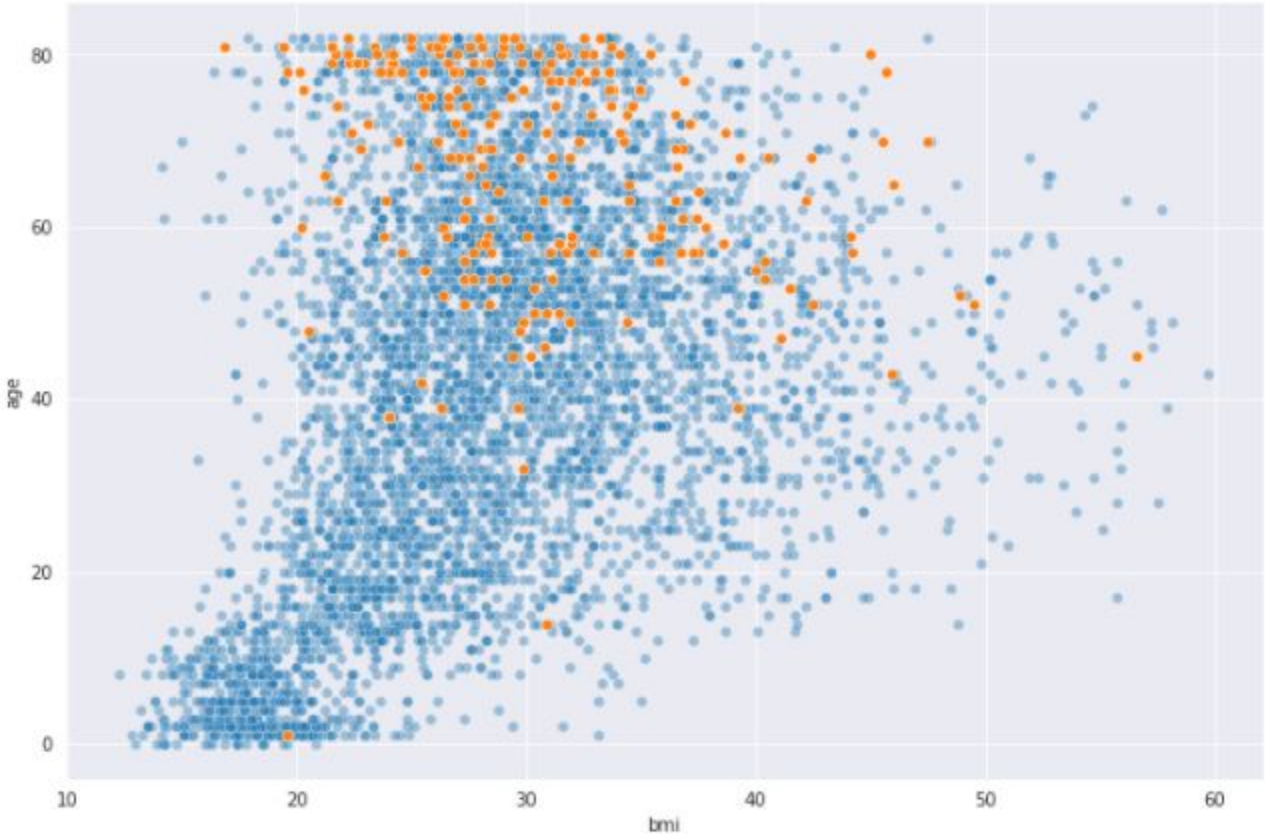
Beside plot shows that the Data is not balanced which will result in a bad Model.



So a balancing technique called ***SMOTE*** has been used to balance the Data.

This was done before fitting our data to the model.

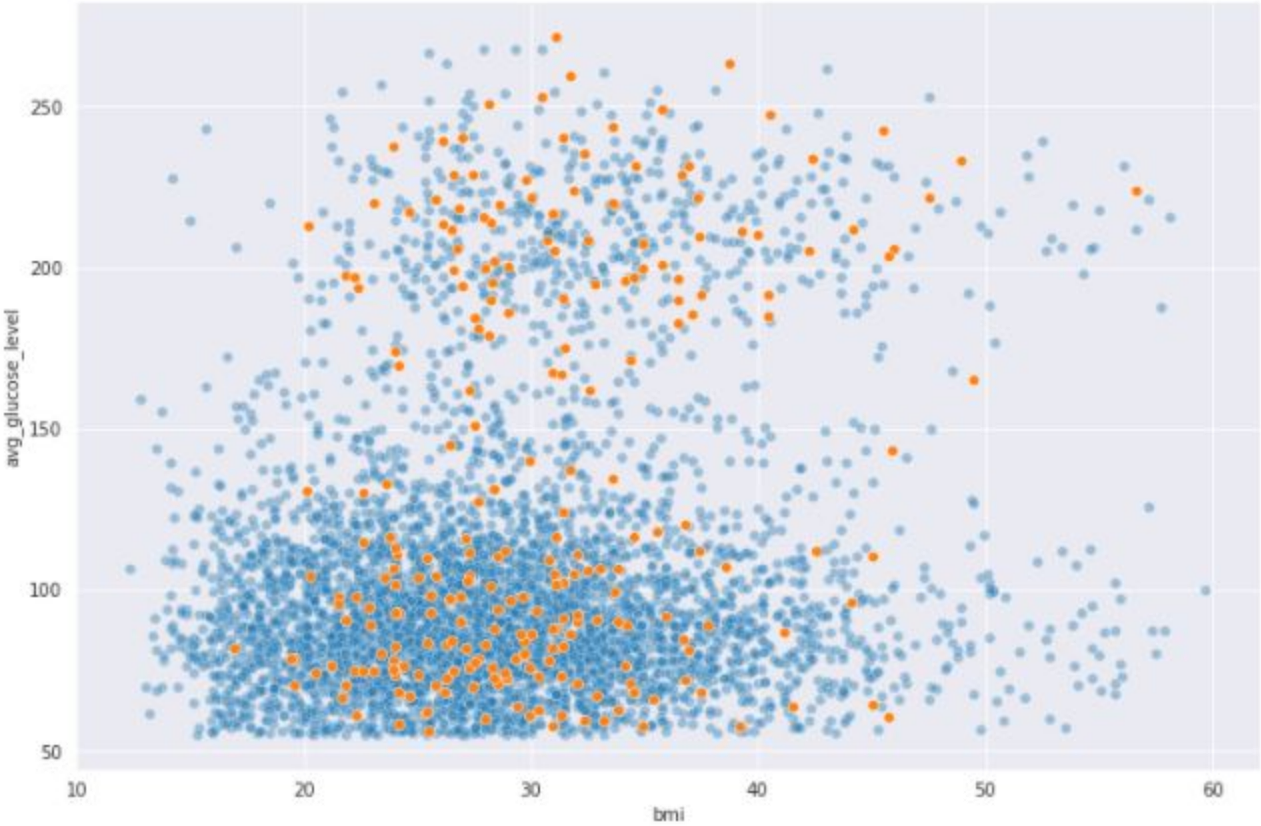
Age vs BMI plot



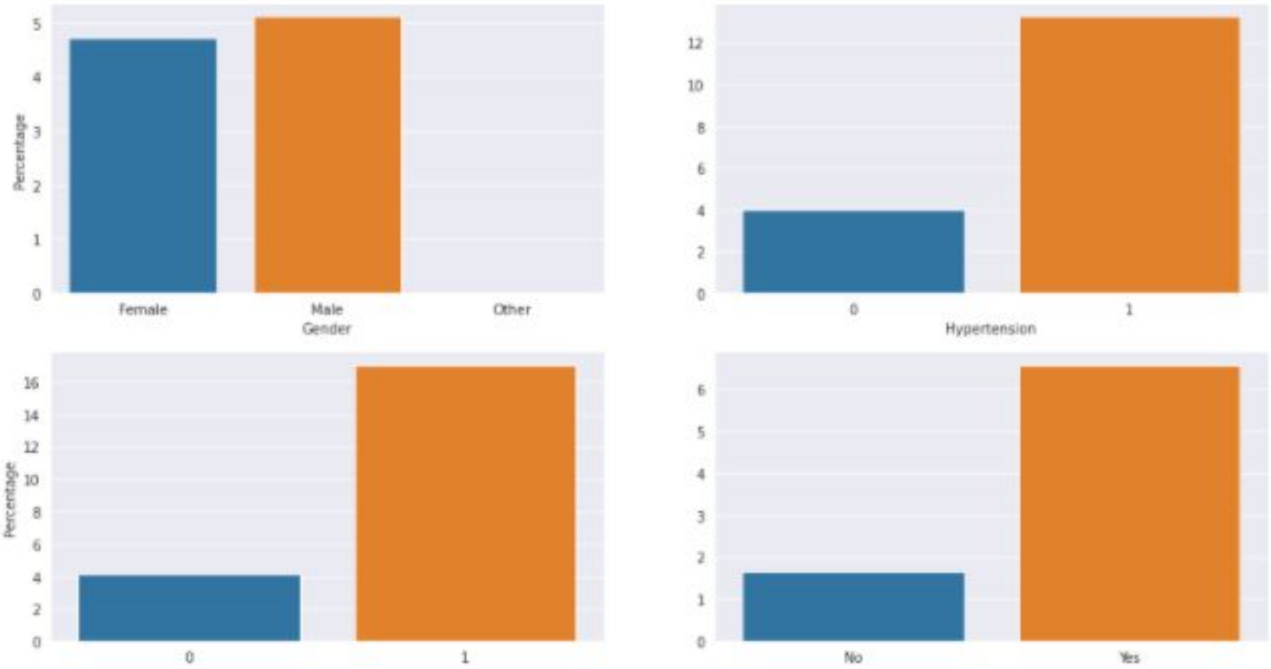
From the above Age vs BMI plot we saw that when people attain an age of 40 or greater the chances of getting a stroke increases and after 60+ it tends to increase even more. Also, people with a BMI of 25+ have shown a higher chances of encountering a stroke.

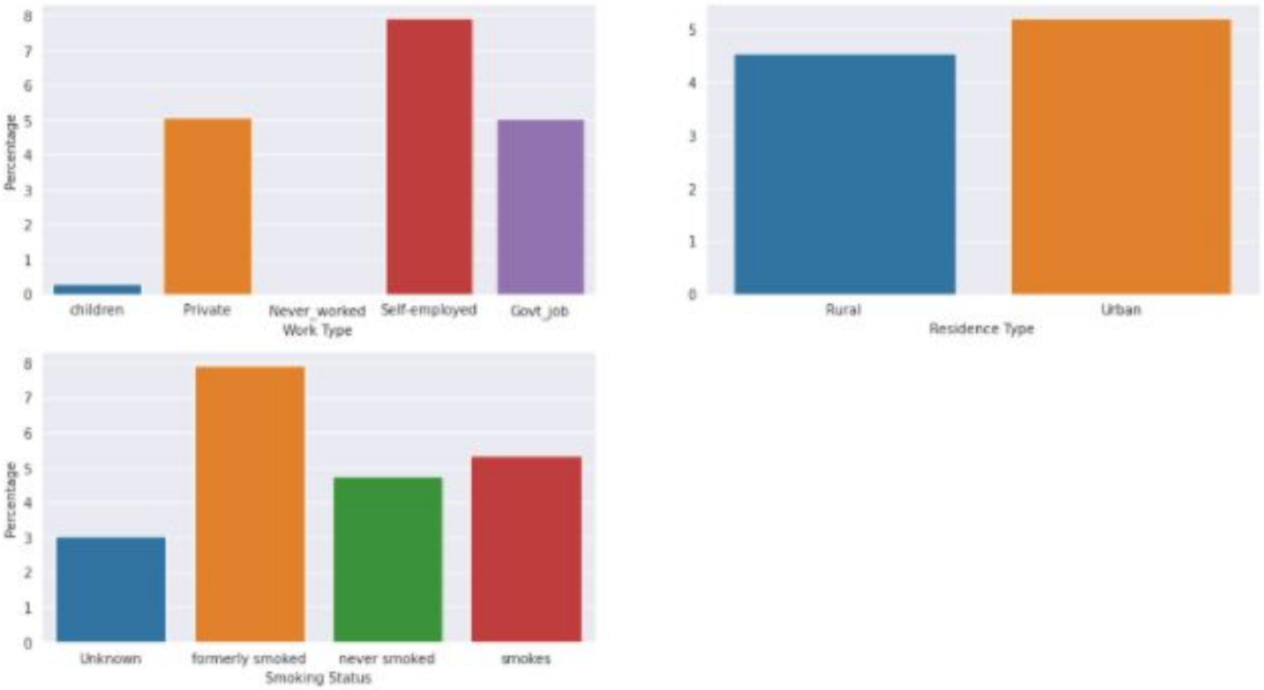
So, people with 40+ years and BMI of 25+ have a greater probability of encountering a stroke.

Avg Glucose Level vs BMI plot



Plot of % of Stroke in each category




Percentage
0
1
2
3
4
5
6
7
8
children
Private
Never_worked
Self-employed
Govt_job
Work Type
0
1
2
3
4
5
Rural
Urban
Residence Type
Percentage
0
1
2
3
4
5
6
7
8
Unknown
formerly smoked
never smoked
smokes
Smoking Status

Conclusions drawn from the above plot with respect to the Stroke Data

- *Both the Genders have around 5% chance.*

- *People with history of Hypertension and Heart Disease have shown an increase in percentage of Stroke with around 12.5% and 16.5% respectively.*

- *Married/Divorced people have a 6.5% chance of stroke. No wonder why people these days choose to stay single.*

- *Self Employed people have a higher chance compared to Private and Govt Jobs.*

- *Rural and Urban doesn't show much difference.*

- *For some reason people who once used to smoke have higher chance compared to people who are still smoking. If you have already started smoking, don't stop. JK, do as you wish.*

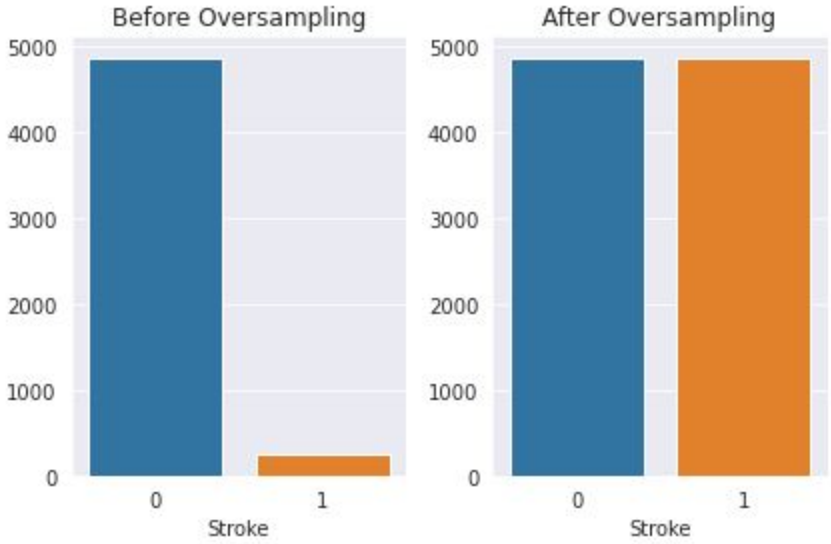
Preparing the Data for the Prediction

1. Converting the Categorical Columns into Numerical by Mapping each category to an integer value using map() on pandas series object.

2. Using a balancing technique called SMOTE (*Synthetic Minority Oversampling Technique*) to balance the data if unbalanced.

3. Splitting the Data in Training and Testing Samples

Bar plots before and after balancing the unbalanced data



Finally we created a Model for the Prediction

Use of Random Forest Classifier