Stroke Analysis & Prediction

By:   
Anas Hany 19100081   
Abdelrahman Ragheb 202002421  
Adham Mohamed 19105744  
  
Under Dr. Mostafa ElAttar’s Supervision

Contents

[1. Introduction 3](#_Toc105452436)

[2. Methodology 4](#_Toc105452437)

[1. Data Preprocessing: 4](#_Toc105452438)

[2. Data Analysis 6](#_Toc105452439)

[3. Stroke Prediction 13](#_Toc105452440)

[3. Conclusion 14](#_Toc105452441)

# Introduction

This project is about analyzing and predicting whether a person can have a stroke depending on multiple variables such as; gender, age, marital status, smoking habits, and so on and so forth. Our demographic was people of both genders with ages varying from 10 to 60 years old. Using this information we used logistic regression, SVM and random forest methods to predict the possibility of a person having a stroke.

# Methodology

## Data Preprocessing:

1. Listing the details of our demographic and their data type:

gender

Female 2994

Male 2115

Other 1

Name: gender, dtype: int64

hypertension

0 4612

1 498

Name: hypertension, dtype: int64

heart\_disease

0 4834

1 276

Name: heart\_disease, dtype: int64

ever\_married

Yes 3353

No 1757

Name: ever\_married, dtype: int64

work\_type

Private 2925

Self-employed 819

children 687

Govt\_job 657

Never\_worked 22

Name: work\_type, dtype: int64

Residence\_type

Urban 2596

Rural 2514

Name: Residence\_type, dtype: int64

smoking\_status

never smoked 1892

Unknown 1544

formerly smoked 885

smokes 789

Name: smoking\_status, dtype: int64

stroke

0 4861

1 249

Name: stroke, dtype: int64

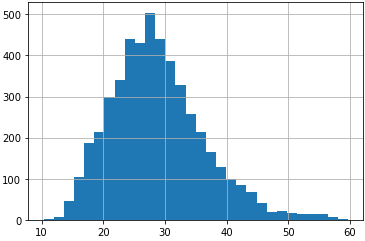
1. Removing the unwanted data:

We removed the “Others” from the genders.

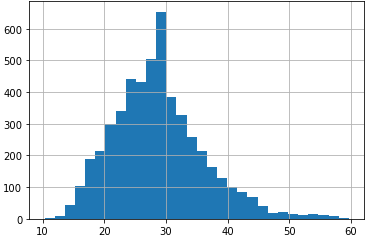


1. Checking for missing values:

Only 201 BMI values were missing, and people with BMI more than 60 were considered as missing values which gives us the below graph:



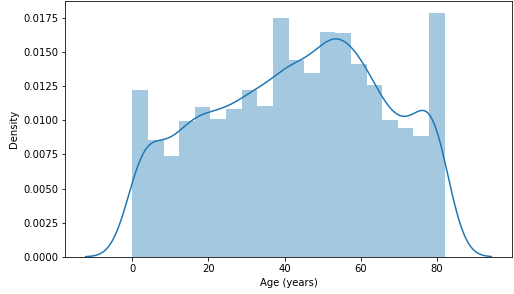
After that we imputed the total missing values by average BMI which gave us the below graph:



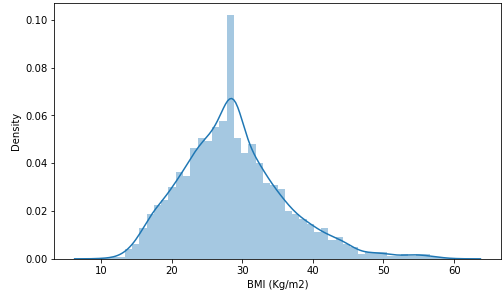
1. Last step was to impute the “unknown” category to “occasional smoking”.

## Data Analysis

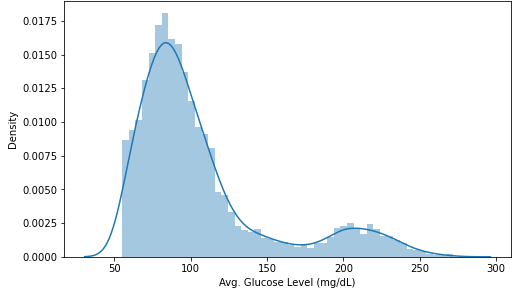
1. Creating and analyzing histograms of numerical variables:



The distribution of age resembles the uniform distribution with a small peak between 50 & 60.

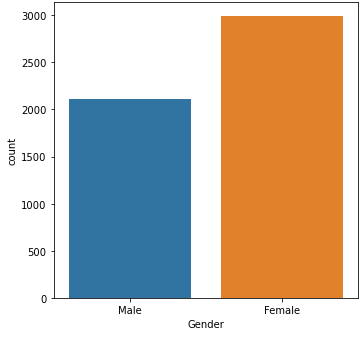


The distribution of BMI is almost normally distributed. The large number of patients around 30 is likely due to mean imputation

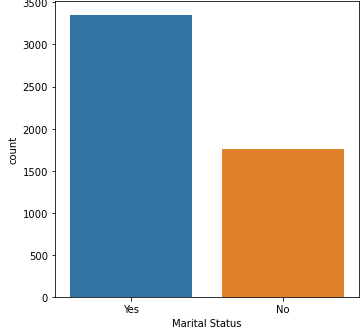


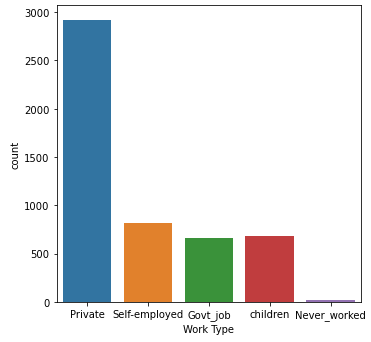
The distribution of average glucose level is bimodal.

1. Plotting the affecting variables of our patient demographic

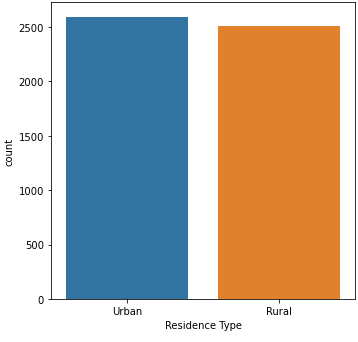


The number of the 2 genders is similar with a slightly more females than males.



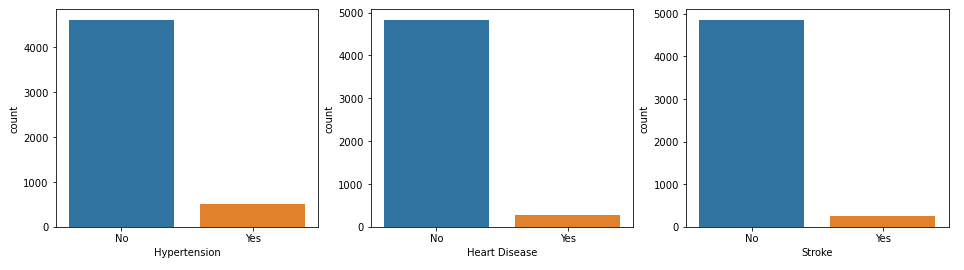


Most of the participants are married and work privately.



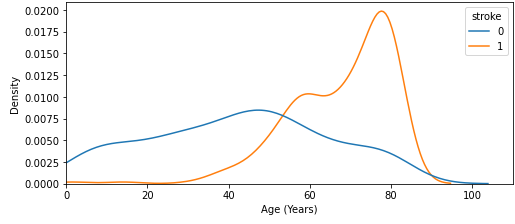
Almost half of the subjects reside in urban areas.

1. Plotting the diseases of our demographic

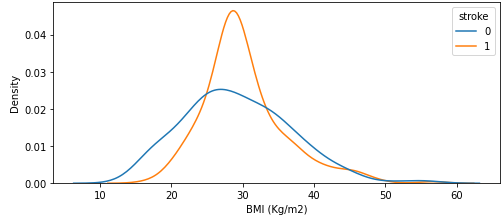


Most of the subjects do not have hypertension, heart disease or stroke.

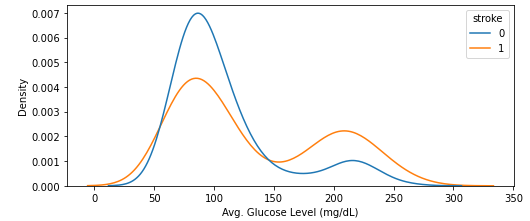
1. Comparison between stroke patients and healthy subjects



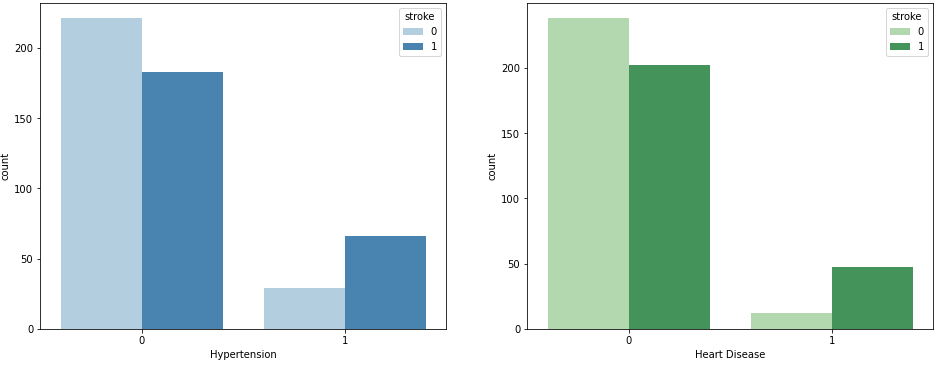
The age of stroke patients is often centered around (60-80), while healthy patients can be of any age. Therefore, old age may increase the risk of stroke.



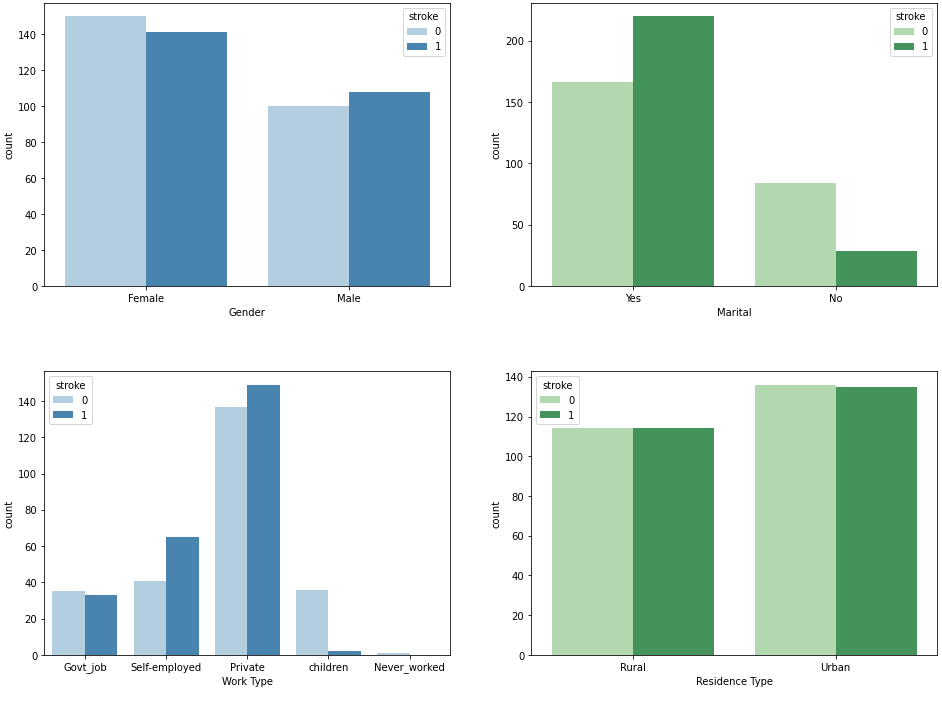
BMI is similar for both stroke patients and healthy subjects. However, the curve of healthy subjects is more spread.



The average glucose level of stroke patients is more shifted to the left compared to the healthy subjects, which might suggests that it can increase the risk of stroke.

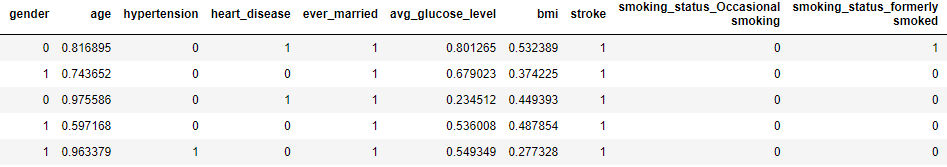


People who have hypertension and heart disease seem to be more likely to have stroke.



## Stroke Prediction

1. Preprocessing  
   We start by dropping the unnecessary columns (ID, Work Type, Residence Type) and turned gender and marital status to binary numbers and then the next step was to transform categorical variables (Smoking Status) with multiple values into one-hot-encoded columns and finally Scale the numerical variables into the same range (0 - 1) using min-max scaling technique.

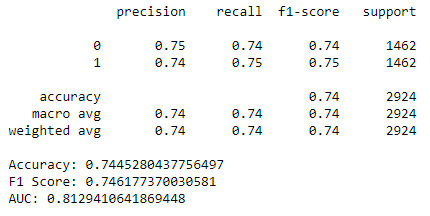


Then we partitioned the data into training set and test set with sizes 70% and 30% respectively and oversampled the data to avoid imbalance in the target column, which gave us the following result:

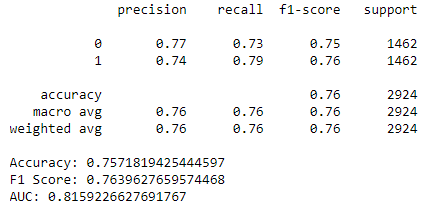
Percent of Stroke Patients before Oversampling: 4.977628635346756

Percent of Stroke Patients after Oversampling: 50.0

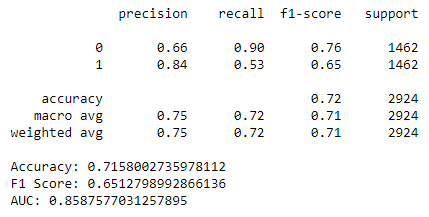
1. Training & Prediction
2. Logistic Regression:



1. SVM



1. Random Forest



# Conclusion

It appears from the 3 different algorithms above that Random Forest is the best classifier in terms of ROC-AUC. However, SVM seems to be the highest performing in terms of accuracy and F1 score. As for the stroke prediction, it seems that, from the graphs, married people are more subject to have a stroke more than people who are not married, same thing for people who work in private work places, as they seem to have more possibility of having a stroke than others who are self-employed or work a government job. People who have heart diseases and hypertension are more likely to have a stroke than others.