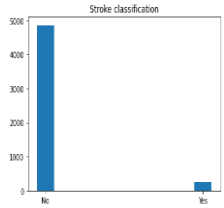
HEART STROKE PREDICTION

PETER OLUSEGUN AINA
YUSIF ISRAFILBAYOV

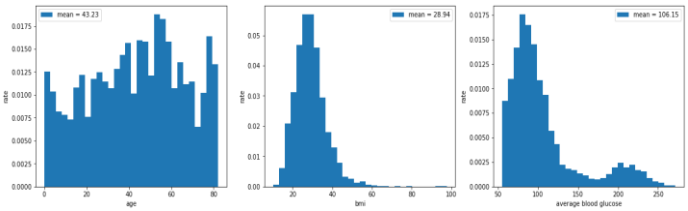
INTRODUCTION

Heart Stroke is one of the most common causes of deaths in the world. It causes physical, emotional and financial wellbeing of the people. In our project, by using machine learning (ML) algorithms, we aim to improve this issue by predicting people's chance of having stroke, so that they can take strong precautions.

THE DATA



- The data is retrieved from Kaggle. It contains 5111 entries with occurrence of stroke.
- However, it contains many **missing or "unknown"** labeled values in body-mass index (BMI) and smoking status attributes. These values consist of a large part of the data, so, we could not get rid of them.
- We used ML models for predicting the missing values; **linear regression** for BMI values and **random forest classifier** for smoking status.

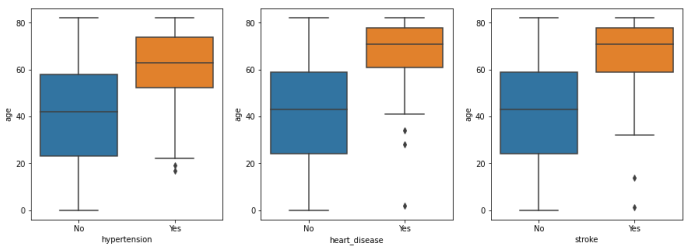


- After filling missing values, we checked the distribution of continuous variables of our dataset. We see that they resemble **normal distribution**. The mean of these variables reflect the population parameters.
- Another problem with our data is **the imbalance** in number of classes; having stroke or not. We addressed this issue in model building

STATISTICAL ANALYSIS

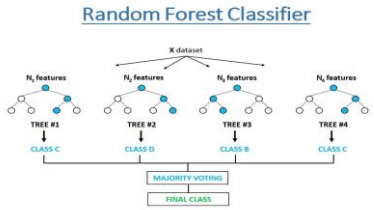
Correlation of variables

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
id											
gender	0.002										
age	0.004	-0.02									
hypertension	0.004	0.021	0.281								
heart_disease	-0.001	0.086	0.27	0.108							
ever_married	0.014	-0.031	0.664	0.164	0.115						
work_type	-0.02	0.054	-0.28	-0.034	-0.016	-0.323					
Residence_type	-0.001	-0.007	0.014	-0.008	0.003	0.006	-0.004				
avg_glucose_level	-0.053	0.049	0.145	0.114	0.106	0.098	-0.02	-0.013			
bmi	0.002	0.002	0.377	0.179	0.068	0.39	-0.324	0.002	0.125		
smoking_status	-0.001	-0.01	-0.098	-0.014	-0.015	-0.027	-0.07	0.02	-0.001	0.005	
stroke	0.006	0.009	0.25	0.128	0.135	0.108	-0.025	0.015	0.083	0.057	-0.046



MODEL BUILDING

- Our problem is a classification problem with mix of numeric and categorical data. The best model for such case is **Random Forest Classifier**.
- However, our data has **class imbalance** and Random Forest Classifier is sensitive towards this issue. Therefore, we did re-sampling.
- **Oversampling** is chosen over under-sampling because of the amount of the data available. We implemented **Synthetic Minority Oversampling Technique.**
- Best hyperparameters for the model is chosen through **hyperparameter tuning by Grid Search** algorithm.
- **5-fold Repeated Stratified Cross Validation** is performed to get the average accuracy score of the model.



MODEL EVALUATION

Accuracy	F1 score	Recall	Precision
0.94	0.95	0.96	0.93