# PRML Course Project

## Stroke Prediction

By:--

Tanuja Tandekar(B20EE073)

Shatakshi(B20EE064)

Pratham Kumar(B20EE042)

# Introduction:--

Stroke is a medical disorder in which the blood arteries in the brain are ruptured which causes damage to the brain. In this disease, the supply of blood and other nutrients to the brain is interrupted. According to the World Health Organization , stroke is the one of greatest causes of death and disability globally. However, Early recognition of the various warning signs of a stroke can help reduce the severity of the stroke. The goal of our project is to implement various machine learning algorithms and to train a ML model to predict if the person has a stroke or not based on the given risk factors.

# Dataset:--

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

This dataset contains 5110 rows and 11 columns.

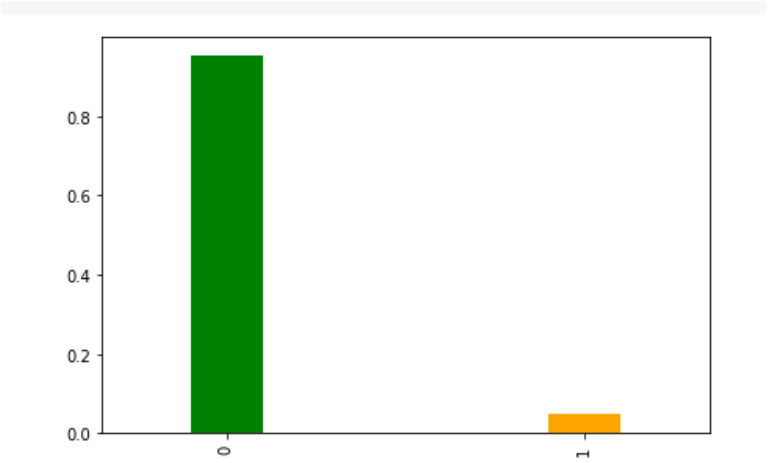
It has been split into train and test with test size of 0.3.

# Exploration and Pre-processing of the dataset:--

After importing all the necessary libraries, irrelevant columns were removed. Further, all the missing values in the dataset were filled using knnImputer and all the categorical data of string type were encoded using label encoder.

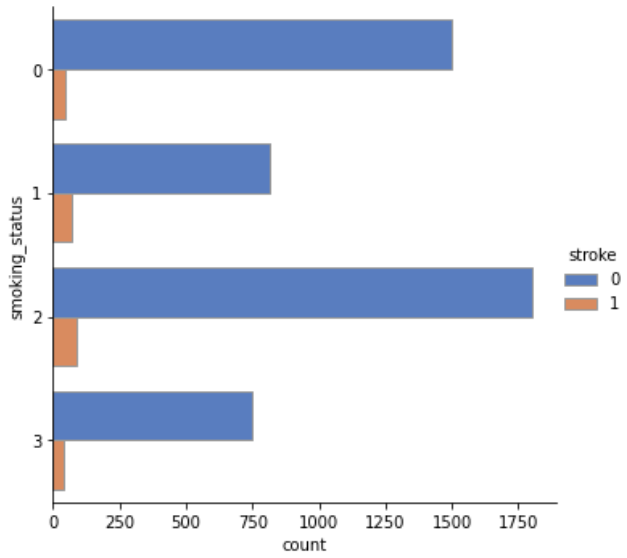
# Visualization of the Dataset:--

Plot of people having stroke(represented by yellow) and not having stroke(represented by green):

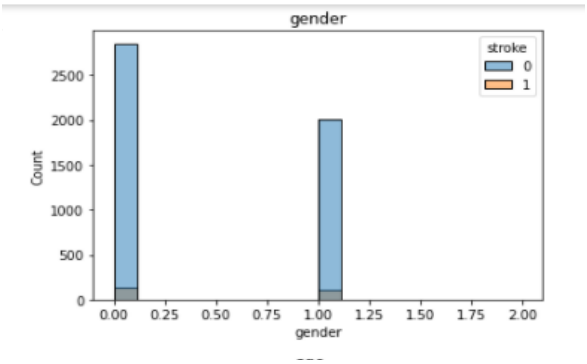



Density plot for stroke cases for various age groups:

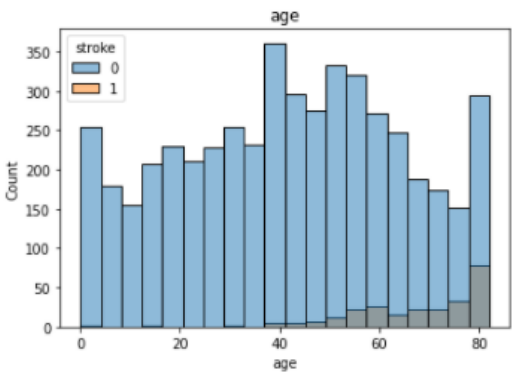
Plot of people having stroke Vs people having smoking habits:



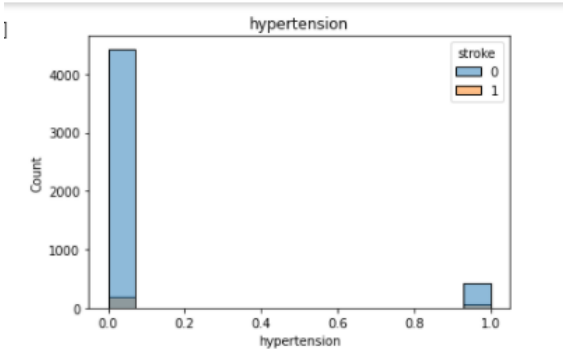

Plot of gender Vs stroke:



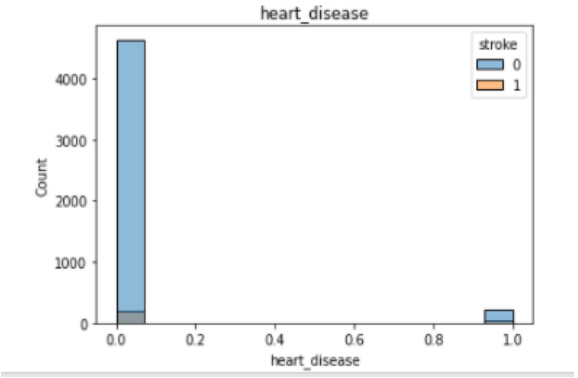

Plot of Age group Vs stroke:



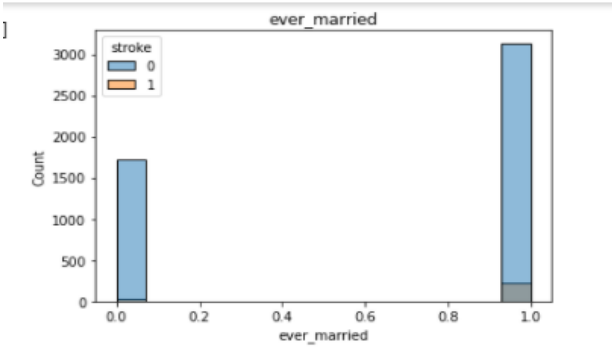

Plot of hypertension Vs stroke:



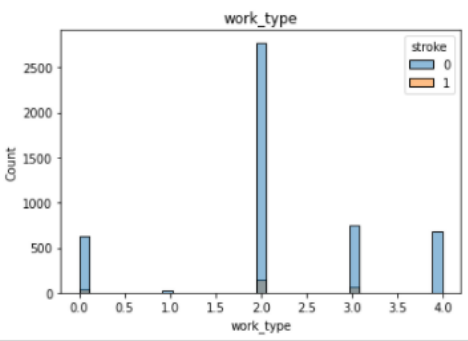

Plot of heart_disease Vs stroke:



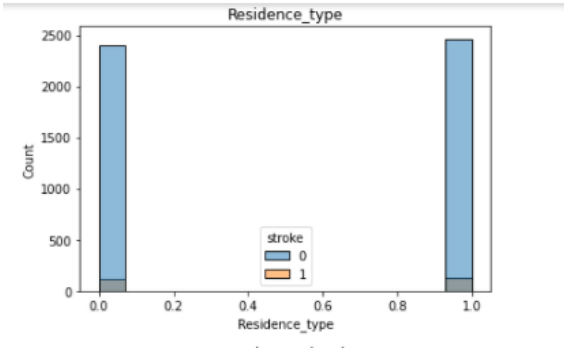
Plot of ever-married Vs stroke:



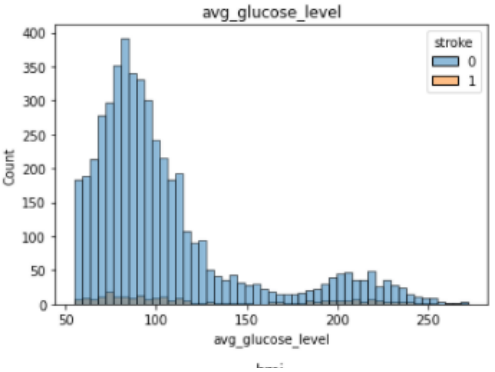

Plot of work-type Vs stroke:



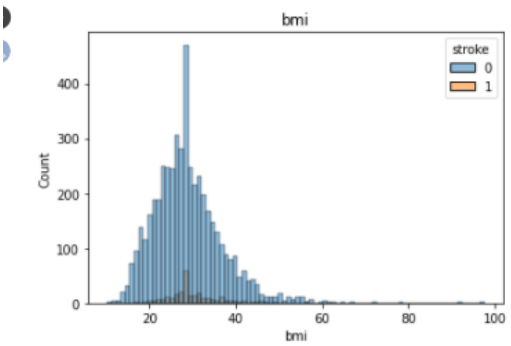

Plot of Residence-type Vs stroke:



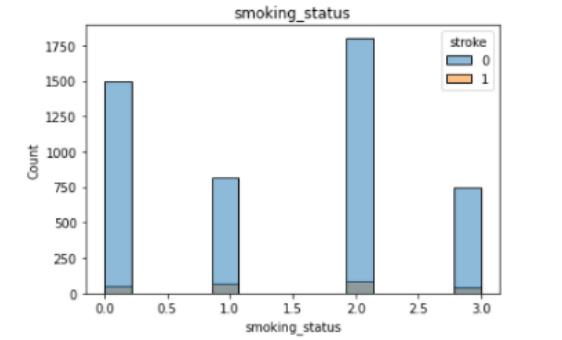

Plot of avg-glucose-level Vs stroke:




Plot of bmi Vs stroke:



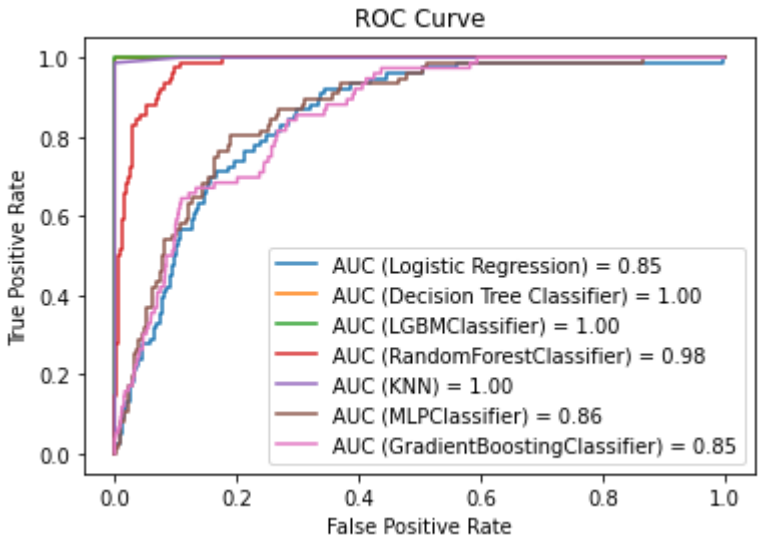

Plot of smoking-status Vs stroke:



# Implementation of classification algorithms:--

| Models | Training accuracy | Testing accuracy | roc_auc_score |
|---|---|---|---|
| Logistic regression | 76.76404 | 74.2335 | 0.85 |
| Decision trees Classifier | 99.9794 | 99.8695 | 1.00 |
| LGBM classifier | 99.8868 | 99.8695 | 1.00 |
| Random forest classifier | 97.6342 | 96.8036 | 0.98 |
| KNN | 94.1575 | 89.8891 | 1.00 |
| MLP Classifier | 77.6897 | 90.6066 | 0.86 |
| Gradient Boosting Classifier | 97.6033 | 95.1728 | 0.85 |

ROC curve:--




From the ROC curve, we can see that the Decision Tree Classifier and KNN give the best roc score.

After creating the whole ML pipeline, we get that Decision Tree Classifier gives the best accuracy.

# Contribution of Individual members:-

1) Pratham Kumar(B20EE042):-
    - Implemented some parts of code which includes visualizing the dataset.
    - Done some of the parts of the report.
    - Implemented the classifiers and compared their accuracy.
2) Shatakshi(B20EE064)
    - Implemented some parts of code, which includes, implementing the code for calculating the roc score.
    - Also, Implemented the code for plotting the roc score.
    - Created the pipeline, and trained and tested the accuracies of some models.
    - Done most of the parts of the report.
3) Tanuja Tandekar(B20EE073)
    - Implemented some parts of the code which includes, preprocessing part, code for handling imbalance classes(using smote), code for calculating the training and testing accuracy, and code for creating the pipeline and testing the various above used models.
    - Helped a little bit in creating the report.