1 Problem Description

Stroke is a medical condition caused by a disruption in the blood flow to the brain. Today, everyone in the globe is susceptible to stroke. Stroke was identified as the second-leading cause of mortality and the third-leading cause of death and disability combined in a recent survey. The burden significantly rose in lower-middle-income nations between 1990 and 2019 (an increase of roughly 70%). Many things that can trigger a stroke. Some of the most common triggers include smoking, hypertension, high blood pressure etc.

2 Dataset Description

The dataset used in this problem solution (stroke dataset) contains 11 attributes to diagnose or predict stroke. The database's attribute information is provided below.

1. Gender :
   1. "Male"
   2. "Female"
   3. "Other"
2. Age : age of the patient
3. Hypertension :
   1. 0 if the patient doesn't have hypertension
   2. 1 if the patient has hypertension
4. Heart Disease :
   1. 0 if the patient doesn't have any heart diseases
   2. 1 if the patient has a heart disease
5. Married : "No" or "Yes"
6. Work type :
   1. "Children"
   2. "Govt\_jov"
   3. "Never\_worked"
   4. "Private"
   5. "Self-employed"
7. BMI : body mass index
8. Residency :
   1. "Rural"
   2. "Urban"
9. Average Glucose Level : average glucose level in blood
10. Smoker :
    1. "formerly smoked"
    2. "never smoked"
    3. "smokes"
    4. "Unknown"\*
11. Stroke :
    1. 1 if the patient had a stroke
    2. 0 if not

3 Choice of Algorithm

Logistic Regression Model

Logistic regression is a machine learning algorithm that builds a model using the independent factors and estimates the likelihood of the dependent variable based on prior data.

Decision Tree Model

A decision tree is a supervised learning algorithm that is used for classification and regression modeling. It is an algorithm that can create both classification and regression models.

Random Forest Model

Random Forest is a machine learning algorithm that builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

KNN Model

K-Nearest Neighbour (KNN) is a machine learning algorithm that assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

In my case I will explain the Random Forest Model.

4 Description of key STEPS

**Data Preprocessing**

**Data analyzing, pre-processing, and cleaning prior to building the model**

**Load data and shape of the data** using .shape property of pandas dataframe object

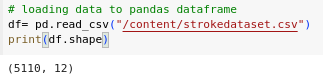


Figure-1 : Data Loading and Total No of Records

**Overview of data before** label encoding and missing value handling

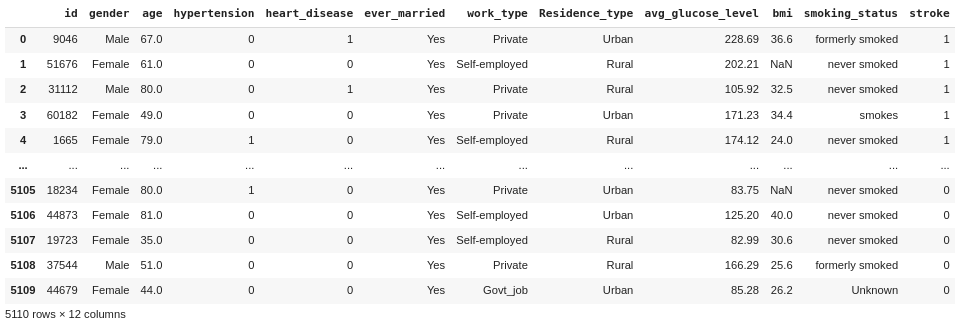


Figure-2 : Overview before label encoding and missing value handling

**Checking for the null** values and handling them using SimpleImputer

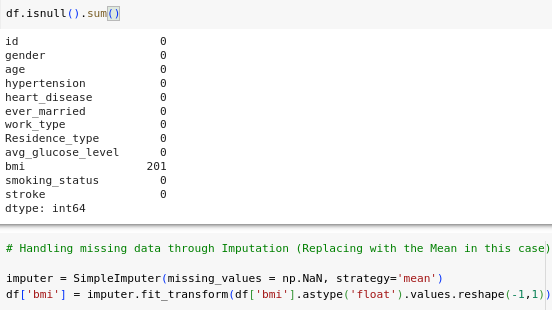


Figure-3 : Checking for Null and Missing values and Handling using simple imputer

**Label encoding** is used to label the categorical data and dropping unnecessary columns after labeling

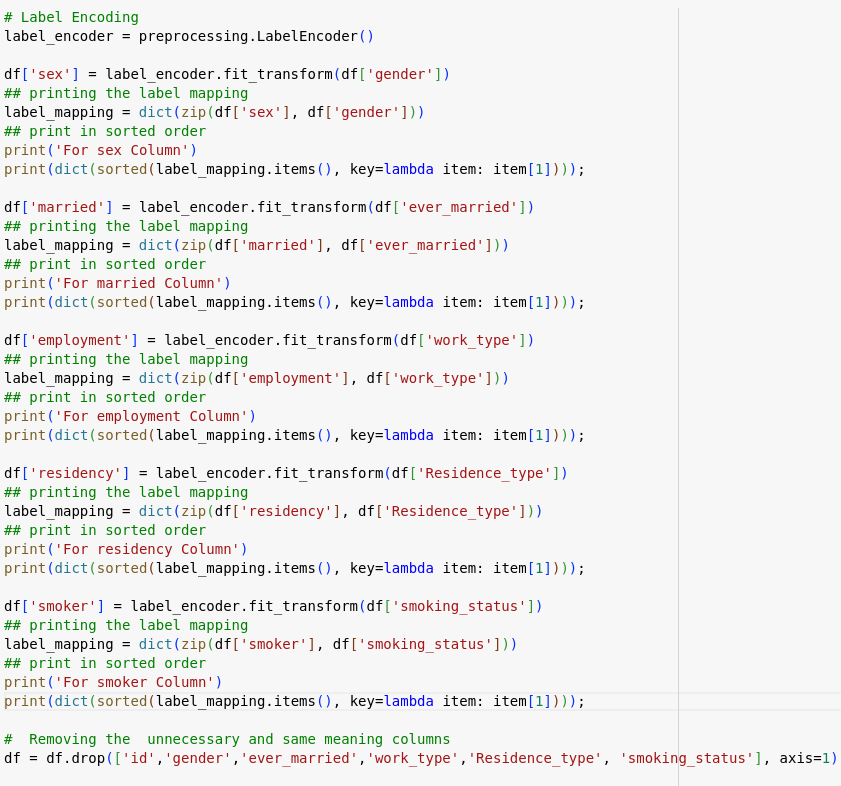


Figure-4 Label Encoding for categorical data and drop unnecessary columns

**After applying the** label encoding and removing the unnecessary columns data frame given

below

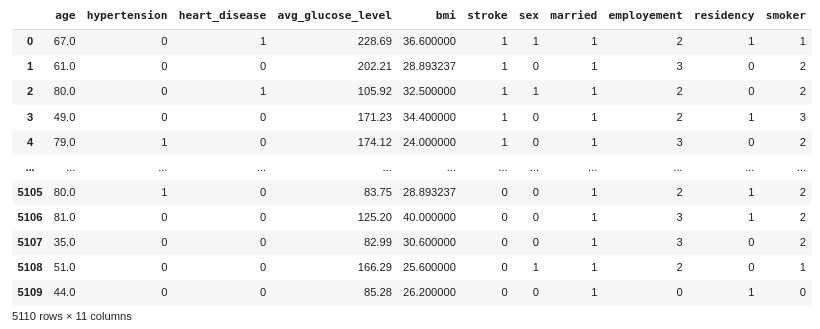


Figure-5 : Overview after label encoding and missing value handling

**Labeling description**

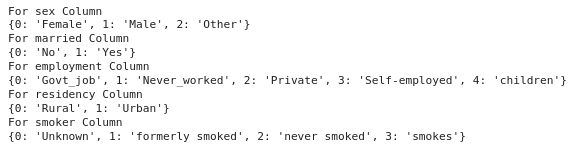


Figure-6 : Label Mapping Of Columns

**Statistical Measurements before** Scaling the numeric data

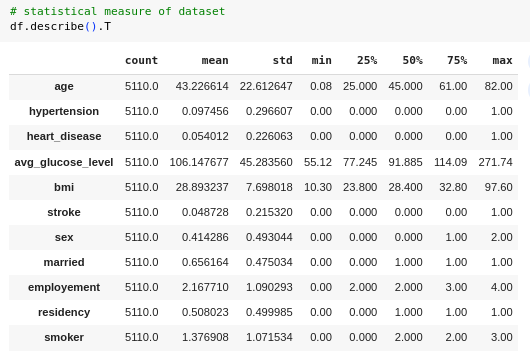


Figure-7 : Statistical Measurements before Scaling

**Statistical Measurements after** Scaling the numeric data

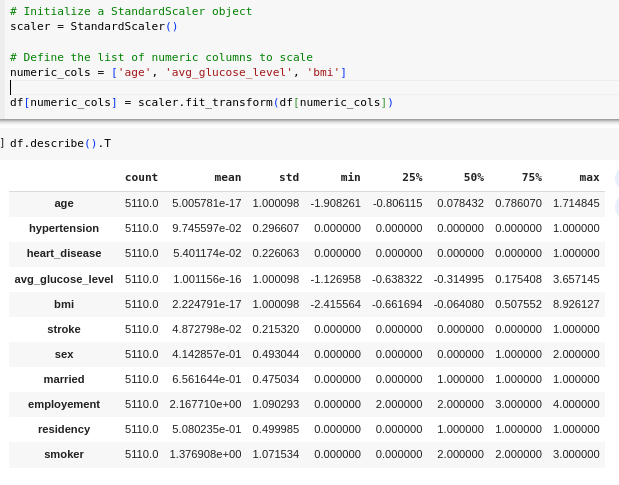


Figure-8 : Statistical Measurements after Scaling

**Correlation Matrix Generation** after preprocess the dataset

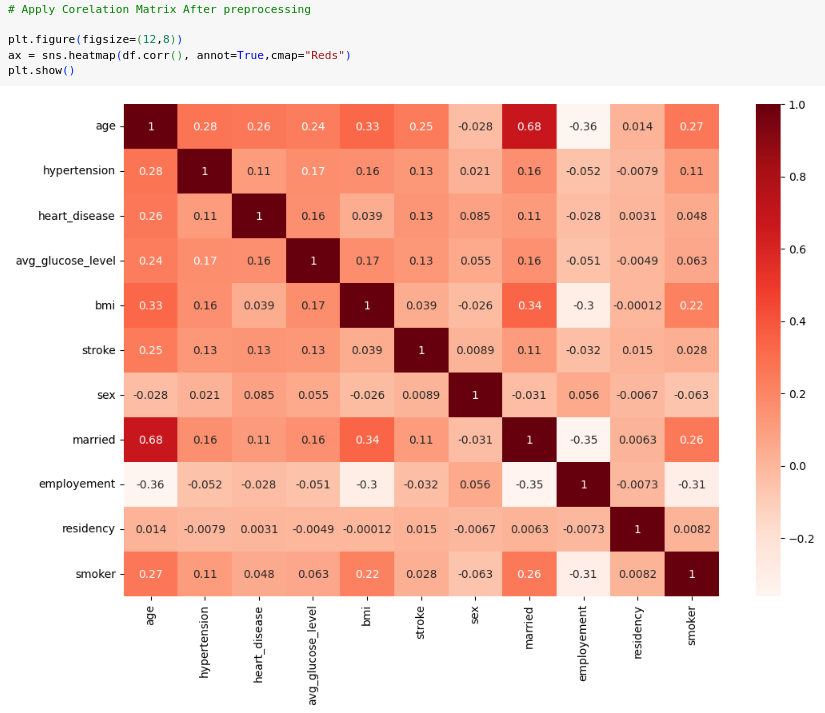


Figure-9 : Correlation Matrix from the processed dataset

**Removing Correlated** Data Columns

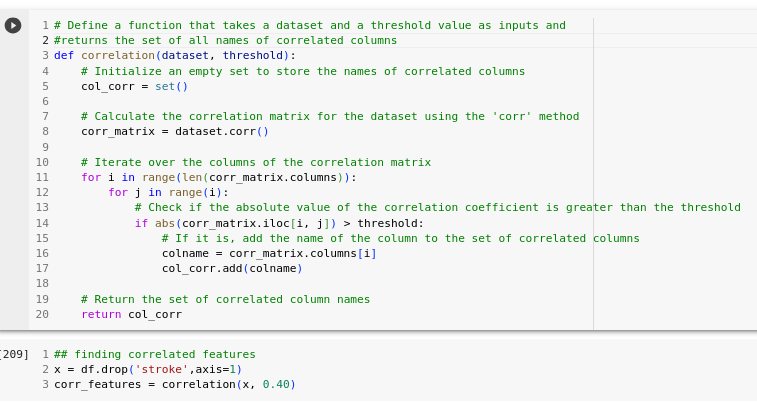


Figure-10 : Correlated Data Columns Removed

**Splitting the dataset** into train and test data.

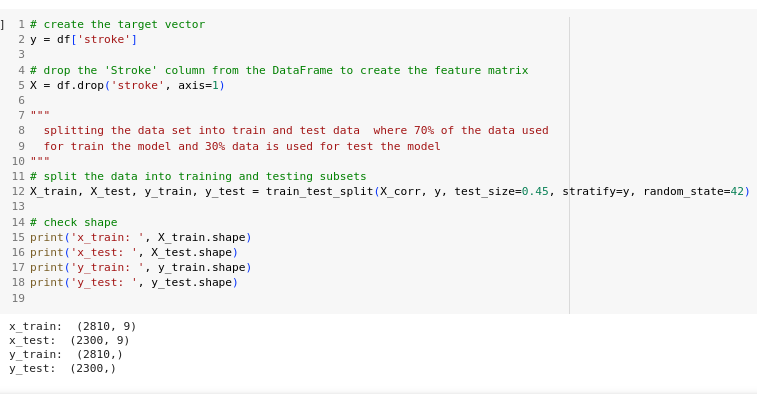


Figure-11 : Splitting the dataset

**To balance our data**

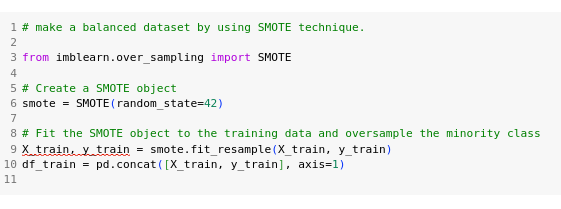


Figure-12 : Balancing The Dataset

Then we have plotted some **comparison based graphs** and charts based on the balanced data set.

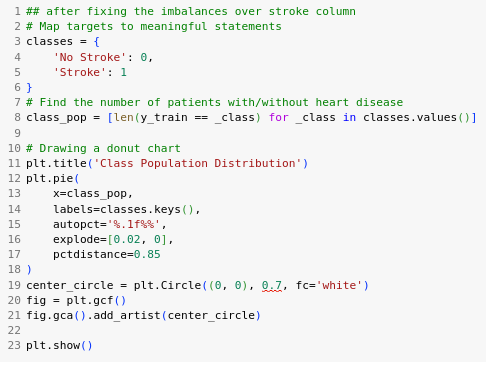


Figure-13 : Stroke Column Data Values Pie Plot

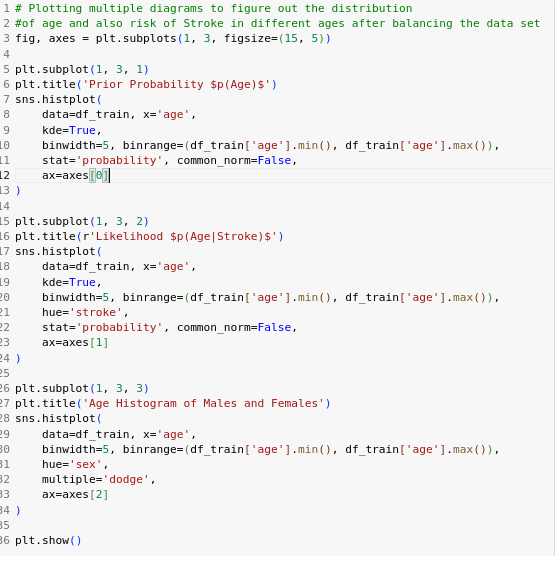


Figure-14 : Distribution of age and Risk of Stroke in different ages Plot

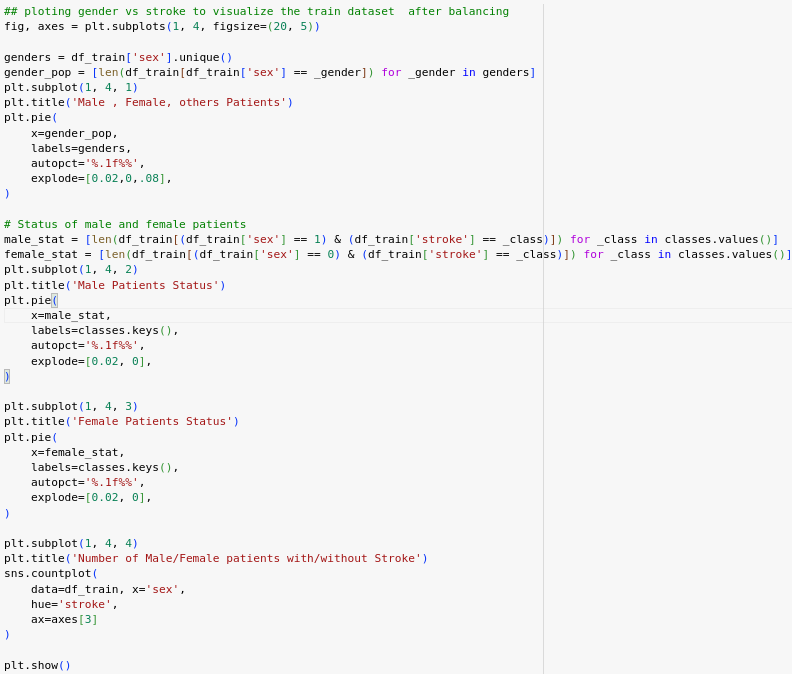


Figure-15 : Distribution of different types of Sex Respect to Stroke



Figure-16 : Distribution of different types of employes respect to Stroke

All of the models that are implemented are trained by the preprocessed dataset that I explained before in this report.

**Logistic Regression Classification Model**

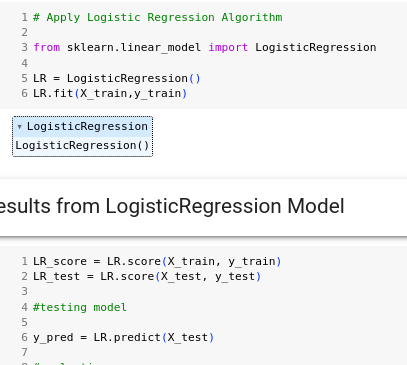


Figure-17 : Logistic Regression Model

**Decision Tree Classification Model.**

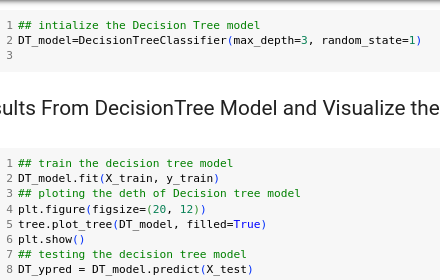


Figure-18 : Decision Tree Model

**Random Forest Classification Model**

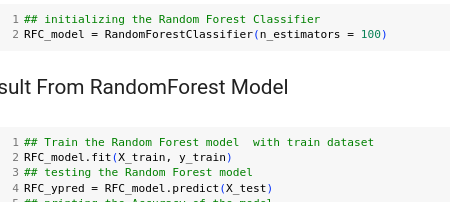


Figure-19 : Random Forest Model

**K-Nearest Neighbour (KNN) Classification Model**

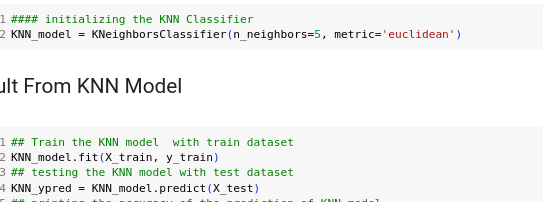


Figure-20 : KNN Model

After applying the models I have decided to tune some hyper parameters of those models to gain better results. So I have tuned the KNN classification Model and Logistic Regression Model.

**Tuned KNN Model**



Figure-21 : Tuned KNN Model

**Tuned Logistic Regression Model**

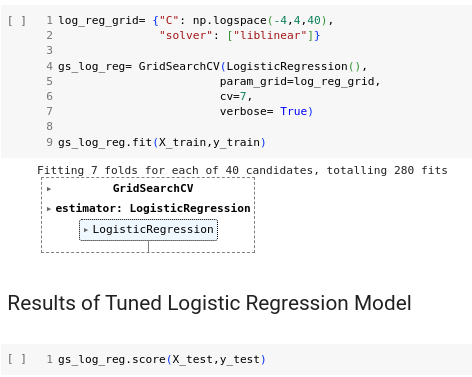
****

Figure-22 : Tuned Logistic Regression Model

5 Results obtained

Data Processing Results after balancing the dataset

**The Figures 23,24,25 illustrate the analysis of the** the data after balancing

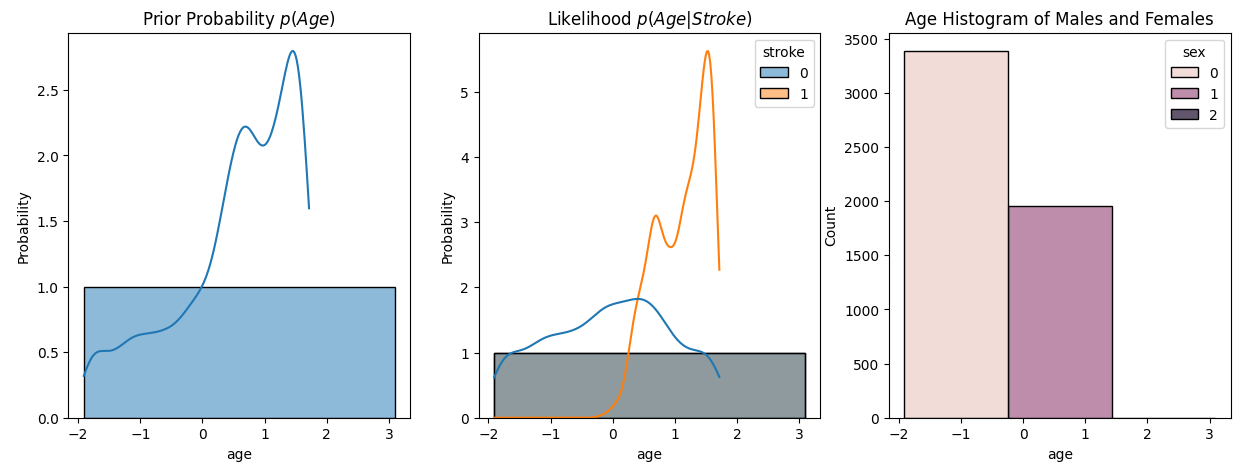


Figure-23 : Age Vs Stroke Distribution and Age Vs Sex Plot

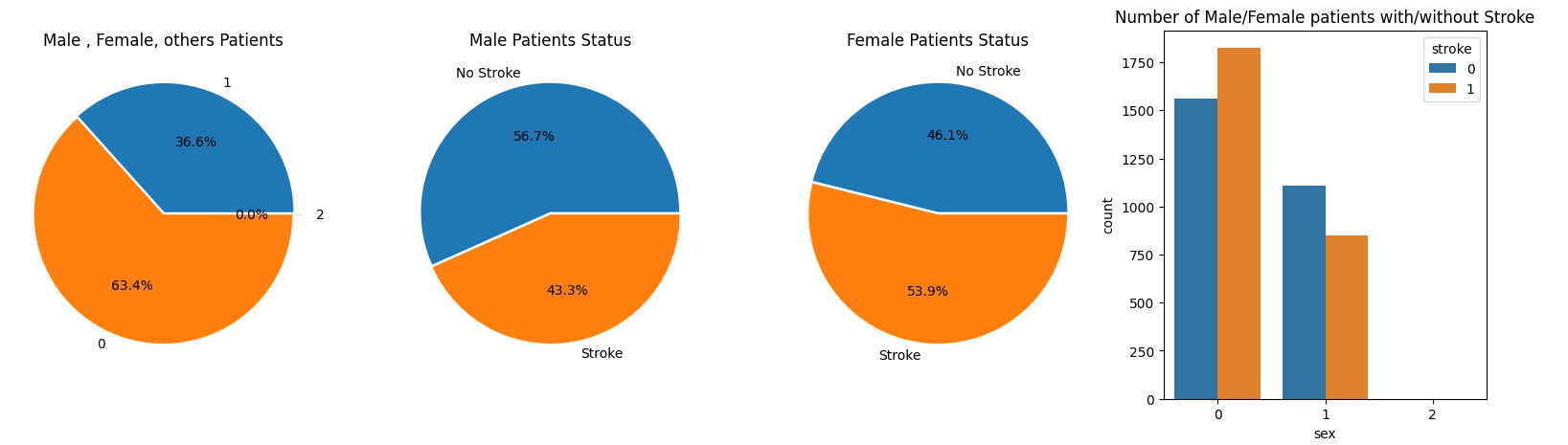


Figure-24 : Sex Vs Stroke Plot

****

Figure-25 : employment vs stroke and sex plot

According to the analysis i observed that

1. Females Has more stroke chances than Males
2. Smokers has greater chances of having stroke
3. A person who has heart disease has more chances of having stroke
4. A person who has hypertension has more chance of having stroke
5. A person who never works has less chance of having stroke
6. Stroke chances does not depends upon residency of a person

**Confusion Matrix Fo**r all of the models that I implemented

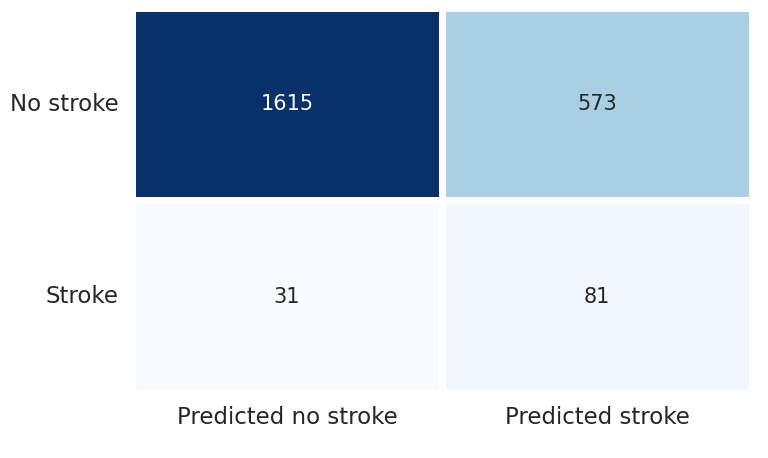


Figure-26 : Confusion Matrix of LR\_MODEL

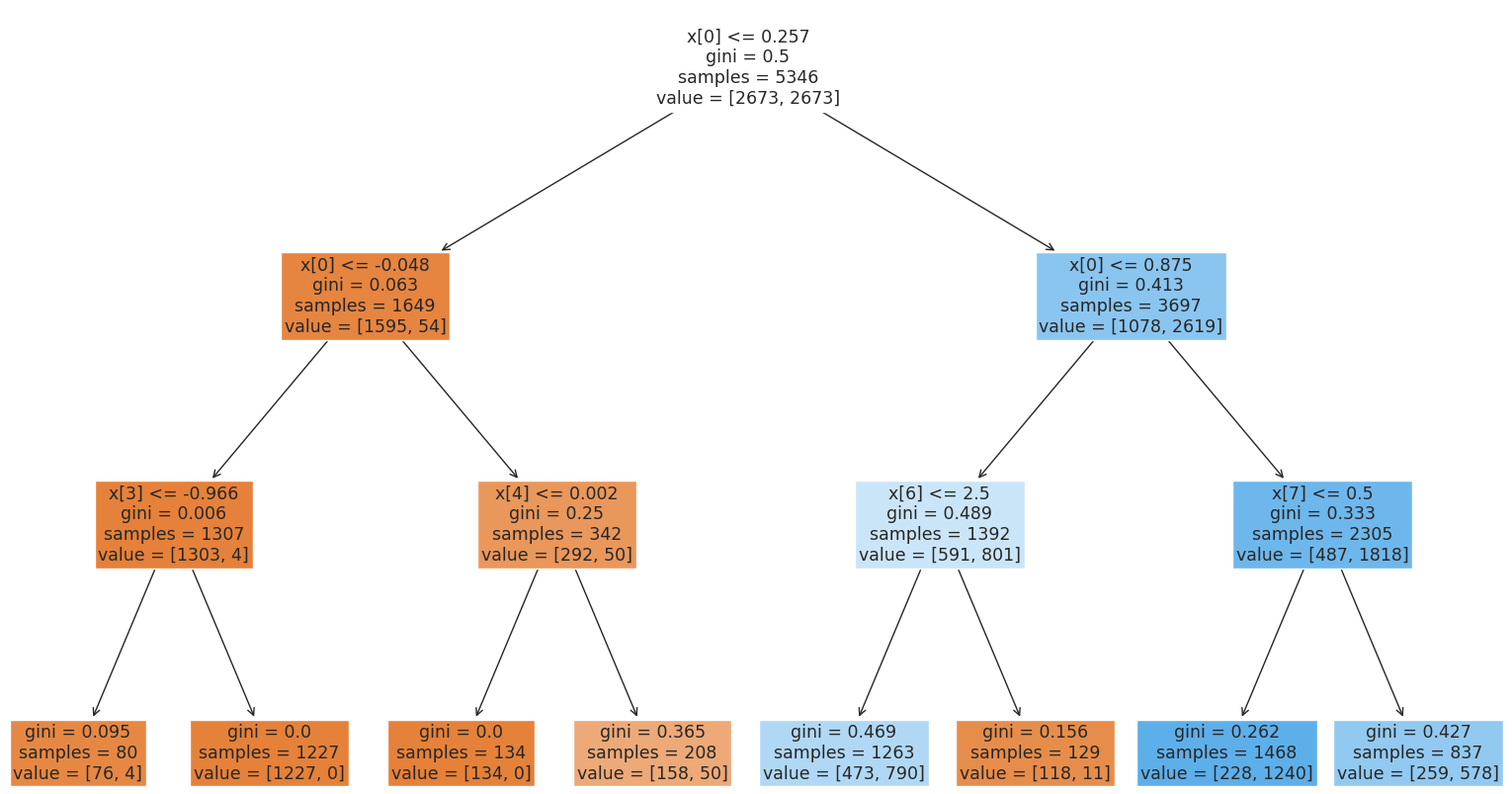


Figure-27 : Depth tree of DT\_MODEL

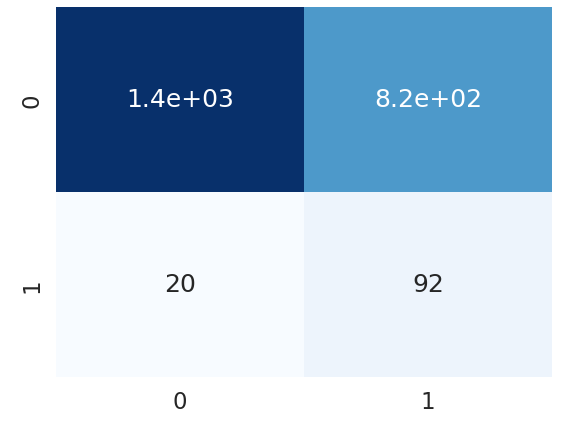


Figure-28 : Confusion Matrix of DT\_MODEL

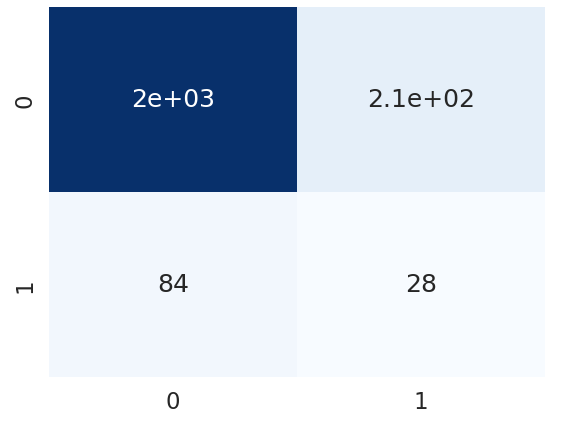
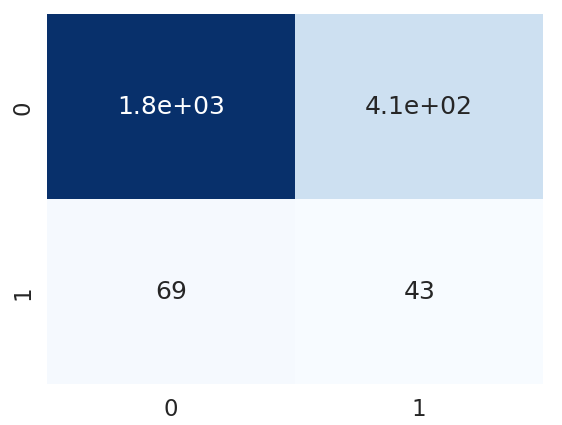


Figure-29 : Confusion Matrix of RF\_MODEL

  
Figure-30 : Confusion Matrix of KNN\_MODEL

Comparison among all models accuracy that I implemented in my code

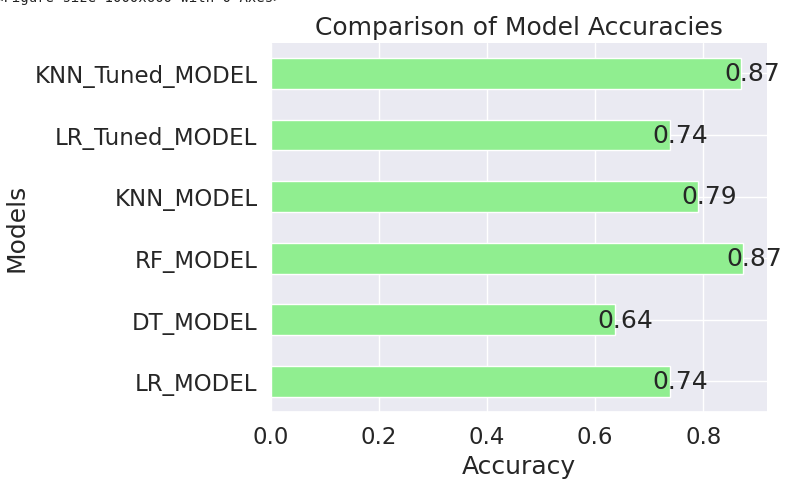


Figure-31 : Comparison among implemented Models

This figure-31 clearly shows that KNN Tuned Model and RF Model Performs better than any other models I have Implemented so far.

**References**

SAMSON TONTOYE . Stroke Dataset collected from <https://www.kaggle.com/datasets/godfatherfigure/healthcare-dataset-stroke-data>

Wso global stroke fact sheet 2022

Organization https://www.world-stroke.org/news-and-blog/news/wso-global-stroke-fact-sheet-2022