# HEART STROKE PREDICTION

MODULE LEADER
**DR ALESSANDRO DI STEFANO**
A.DISTEFANO@TEES.AC.UK
TEESSIDE UNIVERSITY

STUDENT
**SAICHANDRA MUVVA**
W9533735@LIVE.TEES.AC.UK
TEESSIDE UNIVERSITY

## Table of Contents

1.Introduction....2
2.Dataset Description....3
3.Data Wrangling....5
4.EDA (Exploratory Data Analysis)....6
5.Train and Test....10
6.Evaluation Metrics....11
7.Results....12
8.Model comparison....14
9.Conclusion....15
10.References....16

## 1.Introduction

Heart stroke is considered as one of the leading health problems in today's world and death rate is high for this disease and WHO declared nearly 10 million people die each year for this disease and causes of this stroke are lifestyle, stress, age, smoking, drinking and few other key factors and health sector must tackle this situation by early notice to patients the suggestion of this research is to avoid early heart attacks

The importance of machine learning for this project is to analyze the hidden patterns. machine learning gives the early approach of heart attack detection by application of algorithms of logistic regression, decision tree and random forest for data of the patient

## 2.Dataset Description

Data for this study is obtained from Kaggle. This dataset provides information about the patients data and it contains 43,382 thousand patients data and heart stroke column is provided with values 0 or 1 values 0 indicates he is not having the heart stroke and 1 indicates the patient having heart stroke and there are 13 columns in the dataset age column gives the age of patient and gender gives the data that patient is either male or female and column ever married has the values of either yes or no shows he is married or not and similarly column has patient details he is living in the rural or urban and column of bmi , glucose levels having the numerical data of patients and the data set is in the format of comma separate value(csv) .

In this section we discuss about the techniques and preprocessing steps that are required to apply on the dataset for the machine learning model

### Importing Libraries, Packages, and Dataset:

we have imported the dataset to google colab and loaded and then few other libraries and packages are imported because to run specific task in the sections of code(ipynb)file

### Data Preprocessing:

The dataset consists of 43,382 rows and 13 columns. The column names and column values are shown below.

| | id | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status | stroke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 30669 | Male | 3.0 | 0 | 0 | No | children | Rural | 95.12 | 18.0 | NaN | 0 |
| **1** | 30468 | Male | 58.0 | 1 | 0 | Yes | Private | Urban | 87.96 | 39.2 | never smoked | 0 |
| **2** | 16523 | Female | 8.0 | 0 | 0 | No | Private | Urban | 110.89 | 17.6 | NaN | 0 |
| **3** | 56543 | Female | 70.0 | 0 | 0 | Yes | Private | Rural | 69.04 | 35.9 | formerly smoked | 0 |
| **4** | 46136 | Male | 14.0 | 0 | 0 | No | Never_worked | Rural | 161.28 | 19.1 | NaN | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

Raw Dataset Columns (Head)

| | id | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status | stroke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 30669 | Male | 3.0 | 0 | 0 | No | children | Rural | 95.12 | 18.0 | NaN | 0 |
| **1** | 30468 | Male | 58.0 | 1 | 0 | Yes | Private | Urban | 87.96 | 39.2 | never smoked | 0 |
| **2** | 16523 | Female | 8.0 | 0 | 0 | No | Private | Urban | 110.89 | 17.6 | NaN | 0 |
| **3** | 56543 | Female | 70.0 | 0 | 0 | Yes | Private | Rural | 69.04 | 35.9 | formerly smoked | 0 |
| **4** | 46136 | Male | 14.0 | 0 | 0 | No | Never_worked | Rural | 161.28 | 19.1 | NaN | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

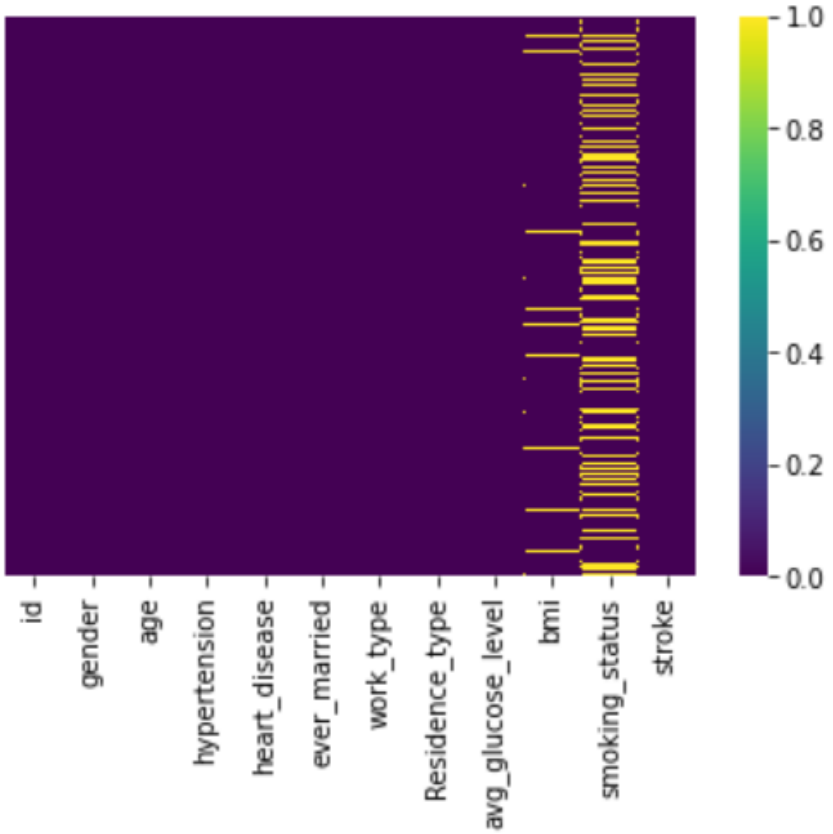
Raw Dataset Columns (Tail)

Fig 1 and 2 describes the raw dataset head and tail we can few Nan values in the columns of bmi and smoking status we can see how many non-null values and null values in the dataset we can see in the below fig3 and our theme is to remove these null values and analyze the data in the dataset

```
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 43400 non-null  int64  
 1   gender             43400 non-null  object 
 2   age                43400 non-null  float64
 3   hypertension       43400 non-null  int64  
 4   heart_disease      43400 non-null  int64  
 5   ever_married       43400 non-null  object 
 6   work_type          43400 non-null  object 
 7   Residence_type     43400 non-null  object 
 8   avg_glucose_level  43400 non-null  float64
 9   bmi                41938 non-null  float64
 10  smoking_status     30108 non-null  object 
 11  stroke             43400 non-null  int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 4.0+ MB
```

Figure 3 Raw dataset (info)

With the help of heat map, we can see the null values in dataset as shown below figure and yellow strokes on the heat map indicates the number of null values in those column



## 3.Data Wrangling

```
df.isnull()
```

| | id | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status | stroke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | False | False | False | False | False | False | False | False | False | False | True | False |
| 1 | False | False | False | False | False | False | False | False | False | False | False | False |
| 2 | False | False | False | False | False | False | False | False | False | False | True | False |
| 3 | False | False | False | False | False | False | False | False | False | False | False | False |
| 4 | False | False | False | False | False | False | False | False | False | False | True | False |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 43395 | False | False | False | False | False | False | False | False | False | False | False | False |
| 43396 | False | False | False | False | False | False | False | False | False | False | False | False |
| 43397 | False | False | False | False | False | False | False | False | False | False | False | False |
| 43398 | False | False | False | False | False | False | False | False | False | False | False | False |
| 43399 | False | False | False | False | False | False | False | False | False | False | False | False |

43400 rows × 12 columns

True values in the above figure are the null values in the dataset and mostly null vales are in bmi and smoking status column we must remove these values so that our data is accurate for algorithm

```
df.isnull().sum()
```

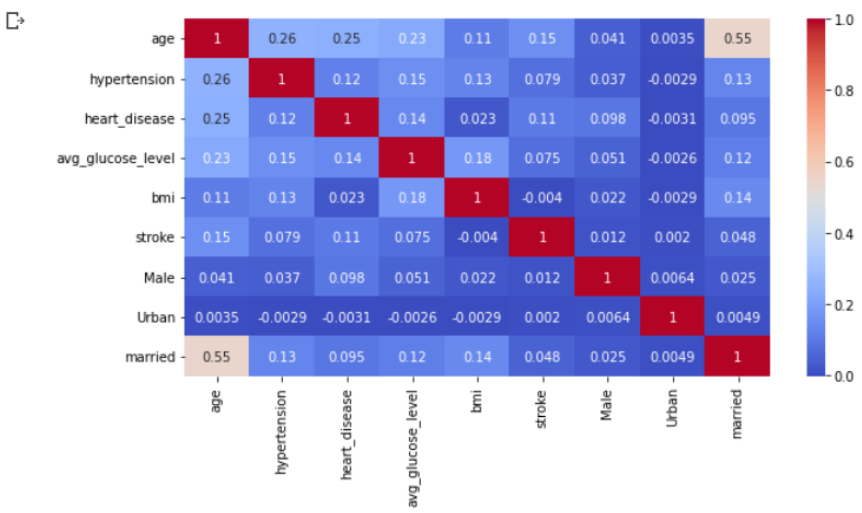
```
id                    0
gender                0
age                   0
hypertension          0
heart_disease         0
ever_married          0
work_type             0
Residence_type        0
avg_glucose_level     0
bmi                1462
smoking_status    13292
stroke                0
dtype: int64
```

After removal of null values, we plotted the bar graph for few other columns as shown in the below figures

## 4.EDA (Exploratory Data Analysis)

The null values in the columns of smoking status and bmi must be removed so that few rows are deleted, and we will attain a new dataset value without the null values after plotting the new dataset with the help of heatmap we plotted a correlation is shown in the below figure so that we can summarize the strong and corelated columns with each other

```
plt.figure(figsize=(10,5))
sns.heatmap(df2.corr(),annot=True,cmap='coolwarm');
```

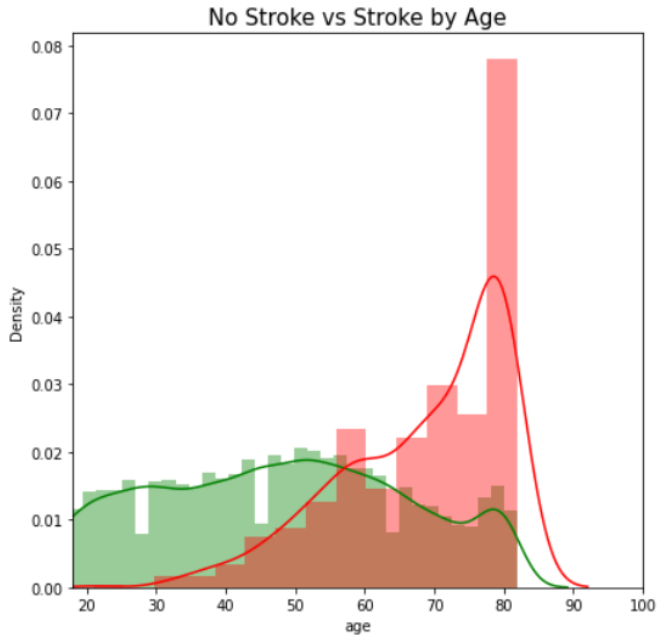



we plotted the correlation graph for attributes in the dataset and correlation helps to find hidden relations between the columns and we have sort the values of correlation for stroke below figure

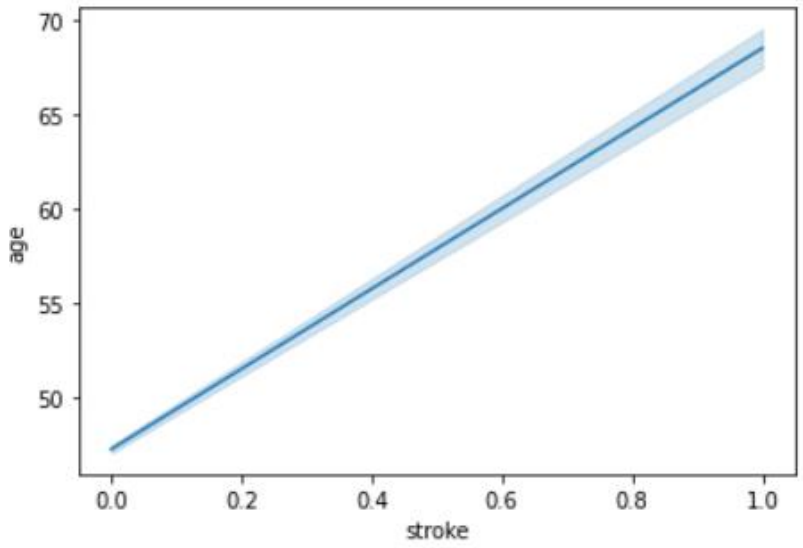
```
abs(df2.corr()['stroke']).sort_values(ascending=False)
```

```
stroke               1.000000
age                  0.154059
heart_disease        0.105142
hypertension         0.078674
avg_glucose_level    0.075447
married              0.047724
Male                 0.012257
bmi                  0.004039
Urban                0.001962
Name: stroke, dtype: float64
```

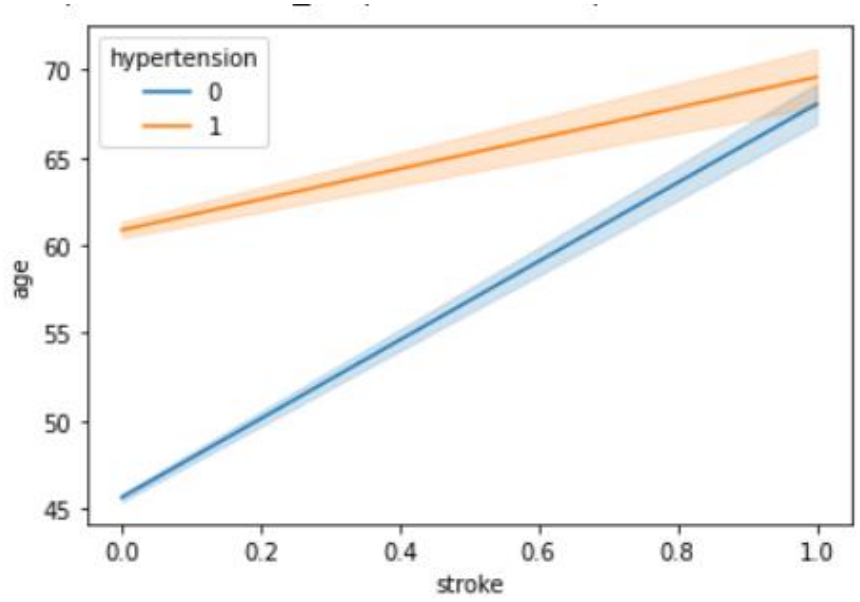
From the above attributes age is highly corelated with occurrence of stroke and then heart disease, hypertension is the primary factors for the occurrence of stroke



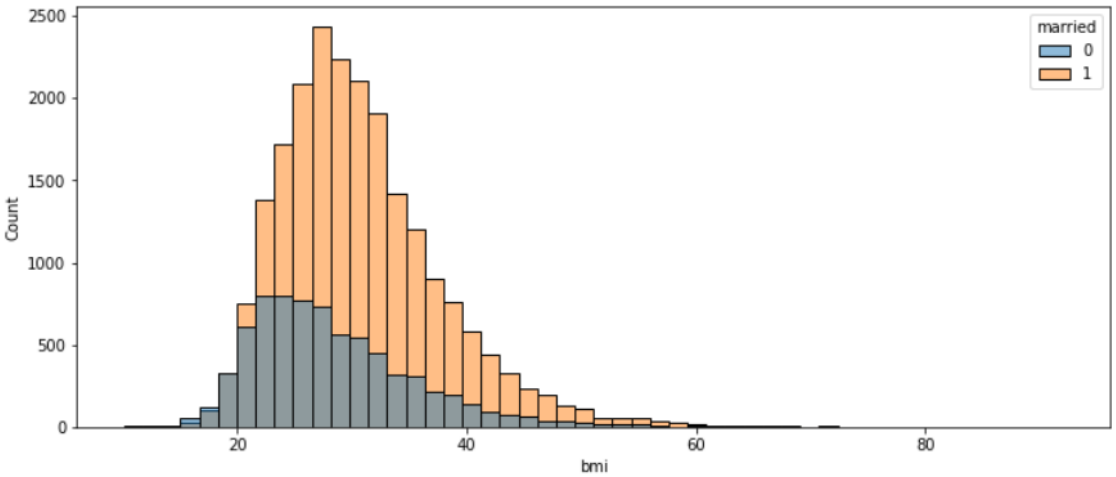

We plotted the graph for age and stroke age column is left skewed and age above 60 years having the high rate of heat stroke



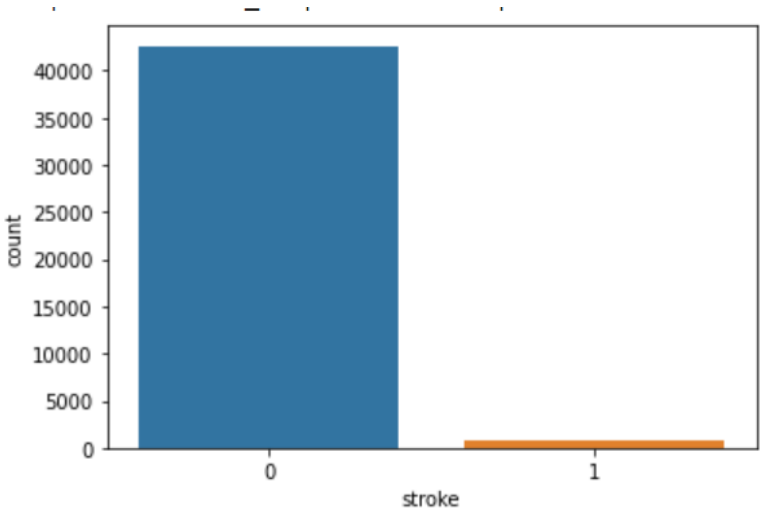

Probability of stroke increases with age it is very clear that people receive the heart stroke at older age



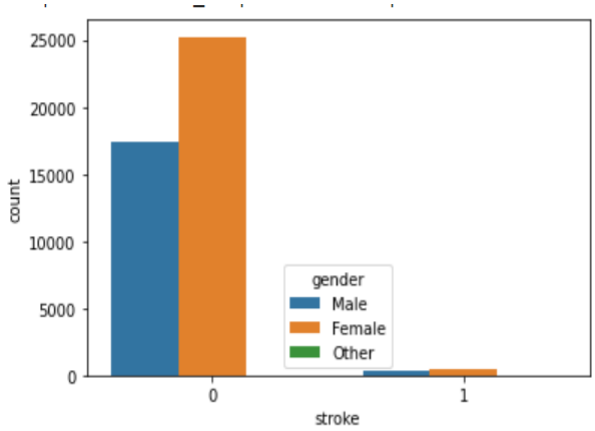

The above figure demonstrates hypertension people having the higher probability of stroke at older age



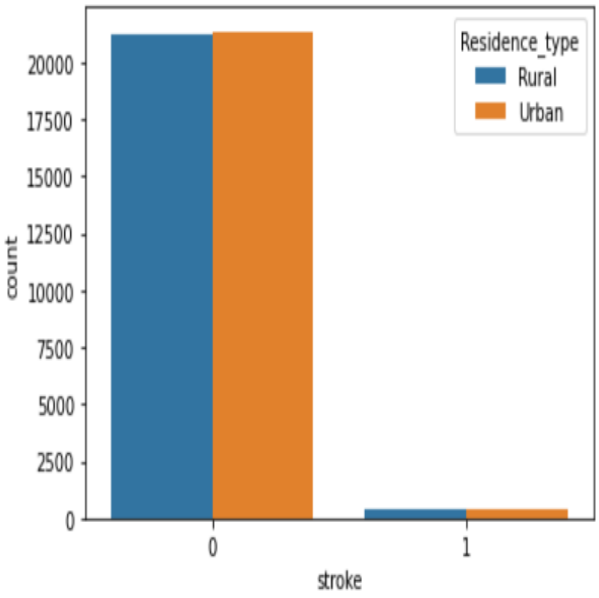

Married people are having higher bmi than unmarried people as shown in the above diagram




in the dataset we can see the greatest number of people don’t have the heart stroke a very few people having stroke in the data set shown above diagram




Above figure illustrates that most samples in the data set are female count of stroke is little higher for female people than male people




Residence type for both urban and rural people having the similar number of heart stroke count

## 5.Train and Test

In this stage the data split into train and test this method is used in the classification and regression problems. based on data first we split into percentage for training and testing and then we train the machine learning model and finally we test the model and find the accuracy for the model by use of confusion matrix, f1score, precision with this we can evaluate the model's capacity. for this dataset we use two models they are

### 1.Logistic Regression

### 2.Decision Tree classifier

For both the models we achieved the accuracy score and attained the confusion matrix so that we can conclude the prediction accuracy for test data and we compared all of these results and we will conclude in the results tab

## 6.Evaluation Metrics

This is used to define the capacity of a model these are important to criticize the model and we improve the model with help of this metrics the following metrics used in this model are as follows

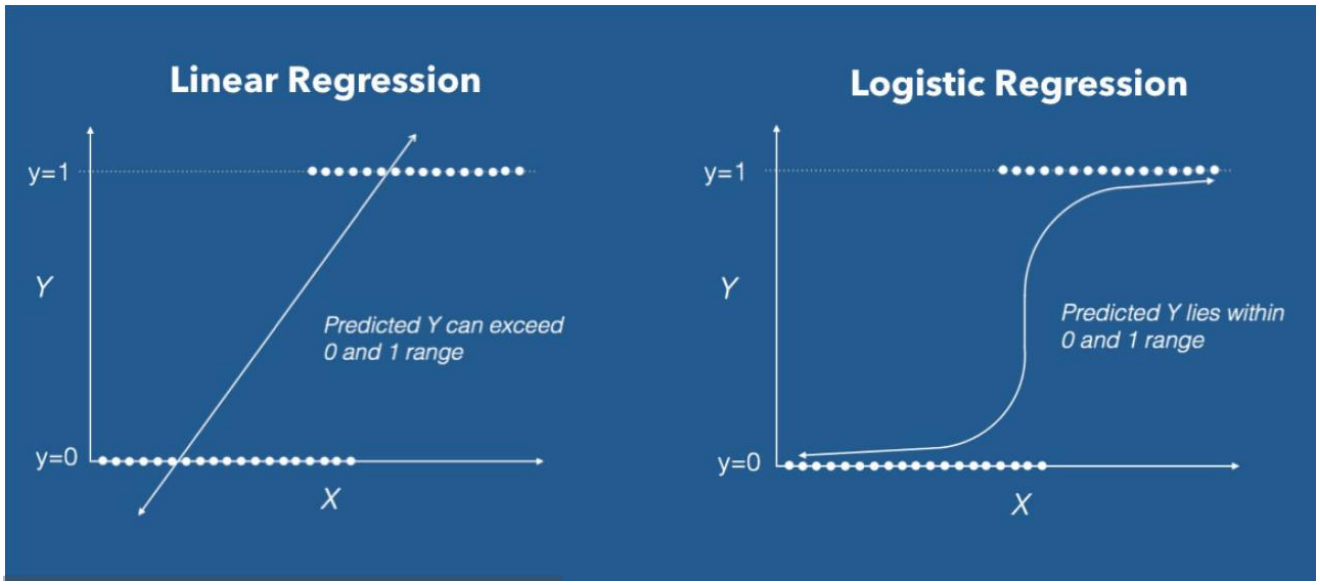
- Accuracy score
- F1score
- Precision
- Recall
- Confusion matrix
- Classification report

After the completion of testing of data with the help of algorithm we will plot the then parameters shown above and criticize the model on accurate guess and error guess, and we will conclude the model performance with the help this parameters and results of these as shown below in the results tab

## 7.Results

### 1.Logistic regression

in this model we can predict the outcome by the application of dependent and independent variables and their relationship between each other in our dataset we predict patient stroke by the relationship columns with the help of correlation matrix and we attain a confusion matrix to check whether the predictions are true or false and how much accurate this logistic regression model works .it is a supervised learning technique, and this model is used for categorical values




Linear Regression is used for continuous variables as shown above and this having the prediction having more values and while logistic regression is either o or 1 and output values lies under these two values only and accuracy for this model of heart stroke data is shown below with the help of logistic regression

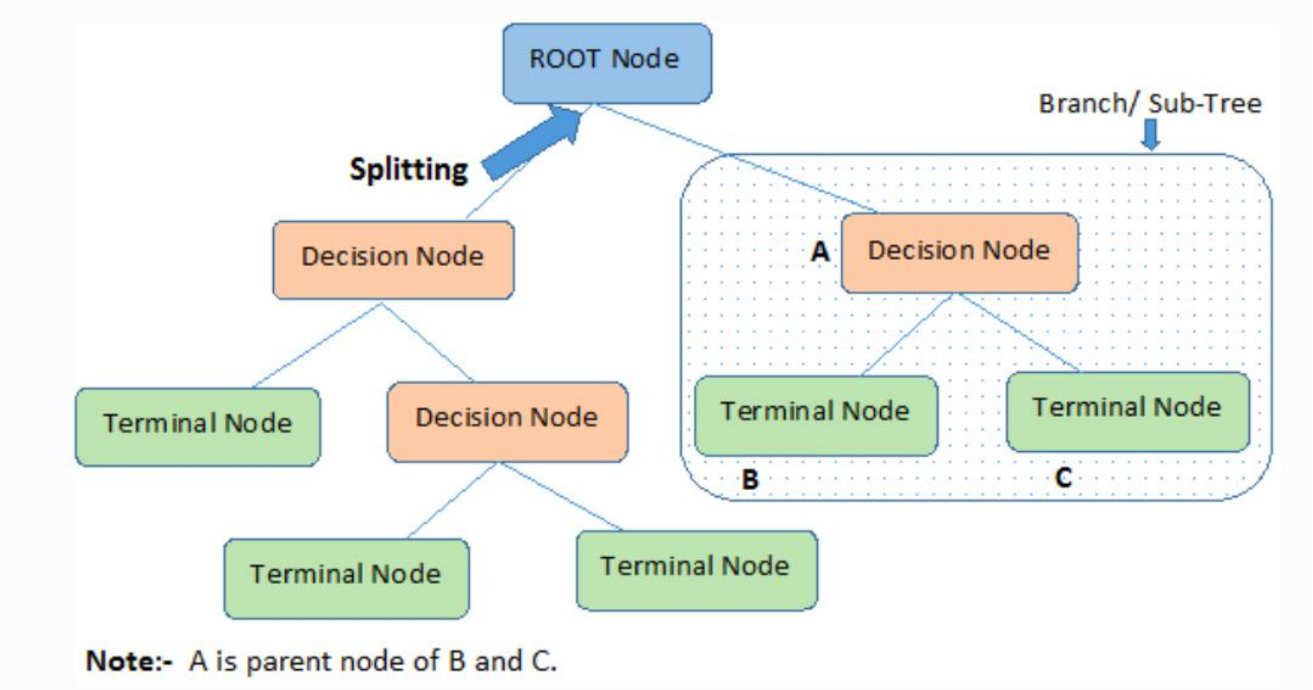
```
[ ]  confusion_matrix(y_test,predictions)

     array([[5454,  228],
            [ 708,  454]])

[ ]  from sklearn.metrics import accuracy_score

[ ]  accuracy_score(y_test,predictions)

     0.8632378725891292
```

The accuracy for this model is 86% for predicting the heart stroke

## 2.Decision tree classifier

This algorithm comes under supervised learning and this algorithm is used for regression and classification like previous model we need to train the with help of some data we feed in and predict the other data as testing and decision tree is of two types of data, we feed in they are continuous and categorized data some terms used in the decision tree are root node, split, decision node, parent node, child node

- Root node – it represents the whole population or sample of data
- Split-process of division of nodes to many nodes
- Decision node – when a node is divided into another nodes the previous node is known as Decision node
- Branch – subsection of tree is known as branch
- Parent node and child node – when node is split into further nodes split nodes are child nodes and the node that is responsible for the split is known as parent node




The above figure illustrates the hierarchy structure of decision tree, and we can see the Root node then after splitting we seen branch tree from Decision node it is represented as root node and further split into B and C these are known as child tree for A This is how Decision tree works

```
Decision tree model accuracy : 0.960
F1-score for the decision tree model is : 0.072

 Classification report for Decision tree classifier
              precision    recall  f1-score   support

  Stroke = 0       0.98      0.98      0.98      5703
  Stroke = 1       0.06      0.08      0.07       110

    accuracy                           0.96      5813
   macro avg       0.52      0.53      0.53      5813
weighted avg       0.96      0.96      0.96      5813
```

The accuracy of this model for predicting heatstroke is 96%

## 8.Model comparison

| **Model** | **Accuracy** |
|---|---|
| Logistic regression | 83 |
| Decision tree classifier | 96 |

## 9.Conclusion

Heart stroke is considered as the most fatal disease and with early symptoms we can solve this issue by applying the machine learning to this dataset values After obtain the results we have seen accuracy for both the models between these two models decision tree is the best option because the accuracy score seen is about 96 percent which is a good scenario than logistic regression

In future this study may be enhanced and prepare a application for it and large dataset is taken into consideration and deliver the better results and forecast the chance of heart stroke which would be helpful for both care department and patients it would be ideal if patients gets notified earlier than the stroke

## 10.References

[1] R. Serban, A. Kupraszewicz and G. Hu, "Predicting the characteristics of people living in the South USA using logistic regression and decision tree," 2011 9th IEEE International Conference on Industrial Informatics, 2011, pp. 688-693, doi: 10.1109/INDIN.2011.6034974.

[2] W. B. Zulfikar, Y. A. Gerhana and A. F. Rahmania, "An Approach to Classify Eligibility Blood Donors Using Decision Tree and Naive Bayes Classifier," 2018 6th International Conference on Cyber and IT Service Management (CITSM), 2018, pp. 1-5, doi: 10.1109/CITSM.2018.8674353.

[3] K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 83-87, doi: 10.1109/ICCMC48092.2020.ICCMC-00016.

[4] Y. Li, "Chinese English translation accuracy detection method based on machine learning," 2021 6th International Conference on Smart Grid and Electrical Automation (ICSGEA), 2021, pp. 185-188, doi: 10.1109/ICSGEA53208.2021.00047.

[5] H. Chen et al., "Personal-Zscore: Eliminating Individual Difference for EEG-based Cross-Subject Emotion Recognition," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2021.3137857.

[6] M. Koehler et al., "Data context informed data wrangling," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 956-963, doi: 10.1109/BigData.2017.8258015.