**Brain Stroke Prediction Using Machine Learning**

# Submitted By Siva prasad Rangineni

(EBEON1222708348)



Table of Contents:

1. [AIM](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#1)
2. [Tools](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#1)
3. [Importing Libraries](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#1)
4. [Importing Dataset](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#2)
5. [Data Visualization](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3)  
   3.1. [Heat Map Correlation](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.1)  
   3.2. [Count Plot](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.2)  
        a. [Gender](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.2.1)  
        b. [Hypertension](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.2.2)  
        c. [Marriage Status](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.2.3)  
        d. [Work Type](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.2.4)  
        e. [Residence Type](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.2.5)  
        f. [Smoking Status](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.2.6)  
        g. [Stroke](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.2.7)  
   3.3. [Scatter Plot](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.4)  
        a. [Age vs BMI](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.4.1)  
        b. [Age vs Avg. Glucose Level](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.4.2)  
   3.4.[Pair Plot](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#3.6)
6. [Data Preprocessing](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#4)
7. [Encoding](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#5)  
   5.1 [Categorical Encoding](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#5.1)  
   5.2 [Label Encoding](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#5.2)
8. [Model Selection](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#9)
9. [Tuning the Models](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#10)
10. [Conclusion](https://www.kaggle.com/code/siddheshera/stroke-eda-smote-9-models-90-accuracy#13)
11. References

AIM:-

The dataset aims to facilitate research and analysis to understand the factors associated with brain stroke occurrence, as well as develop prediction models to identify individuals who may be at a higher risk of stroke.

Tools:

Numpy:- It stands for "Numerical Python" and provides support for large, multi-dimensional arrays and matrices, along with a wide range of mathematical functions to operate on these arrays efficiently.

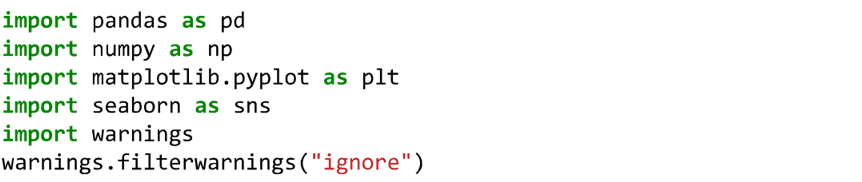
Pandas:- It provides high-performance, easy-to-use data structures, such as DataFrame and Series, which are designed to efficiently handle structured data.

Matplotlib:- It provides a wide range of functions and features to generate various types of plots, charts, histograms, and more

Seaborn:- It provides a high-level interface for creating attractive and informative statistical graphics. Seaborn is designed to work seamlessly with Pandas DataFrames and integrates well with the data manipulation capabilities of Pandas.

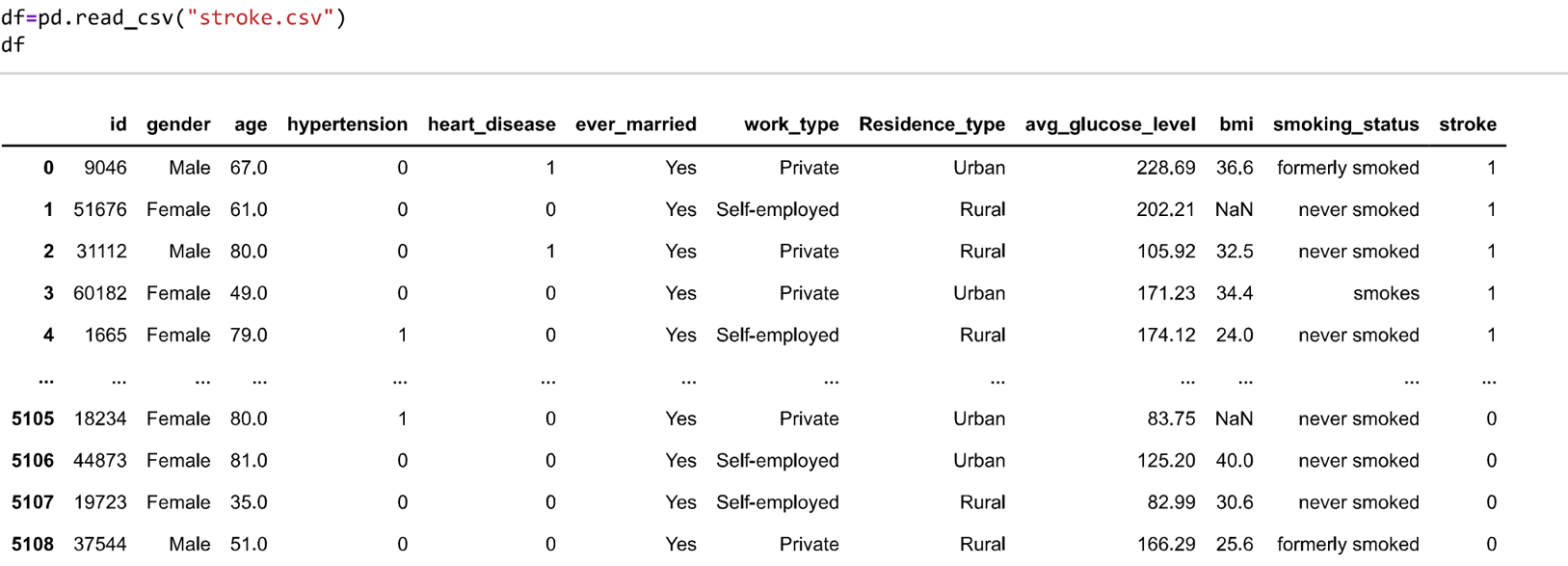
IMPORTING LIBRARIES:

Importing libraries in programming serves the purpose of gaining access to pre-defined functions, classes, and other resources that are not available in the base programming language.



IMPORTING DATSET:

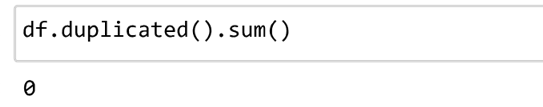
Importing a dataset involves loading the dataset into your programming environment or application so that you can access and analyze the data. The specific steps for importing a dataset may vary depending on the programming language or tools you are using

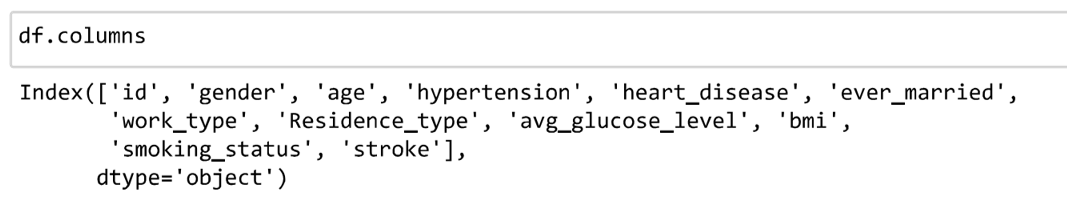


To Find the how many columns and size of the dataset:

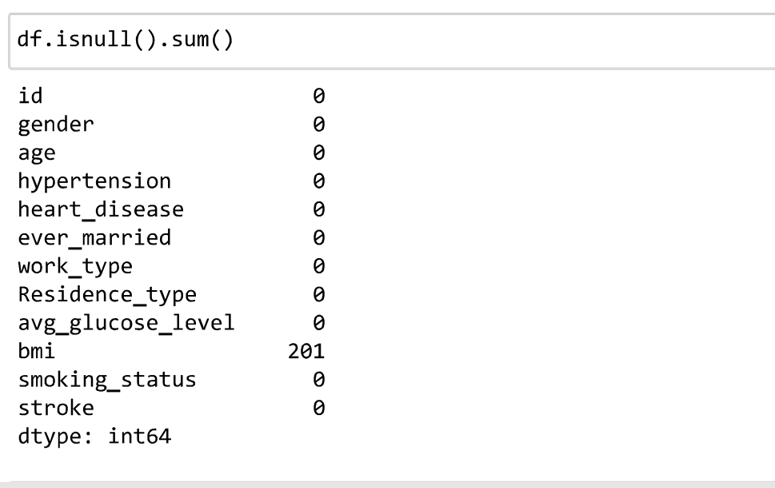


To check weather Duplicated values present in the dataset:





To check the null values in the dataset:

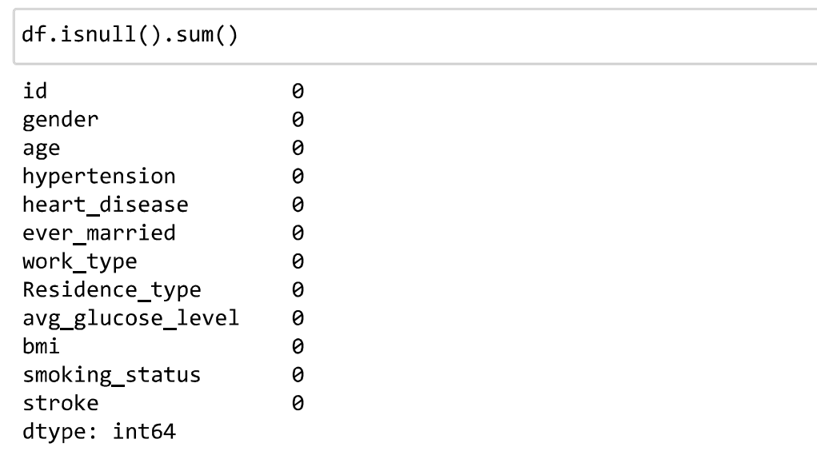


**There are null values present in 'bmi'.**

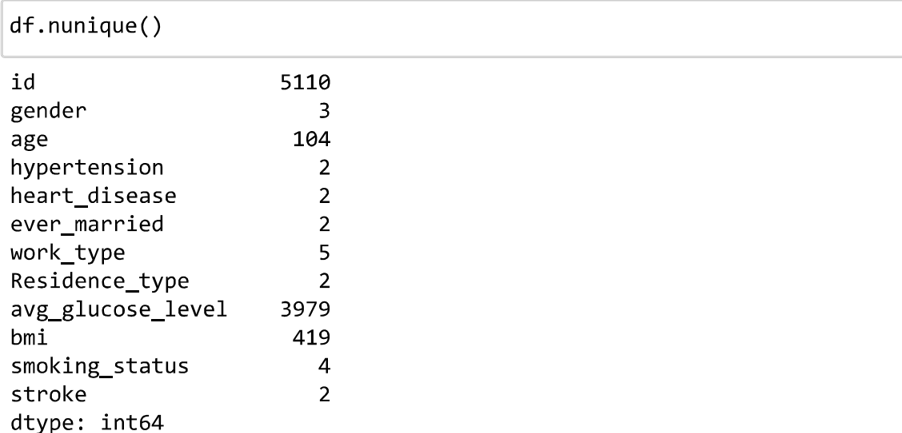


**We replaced null values of 'bmi' with mean in that column.**

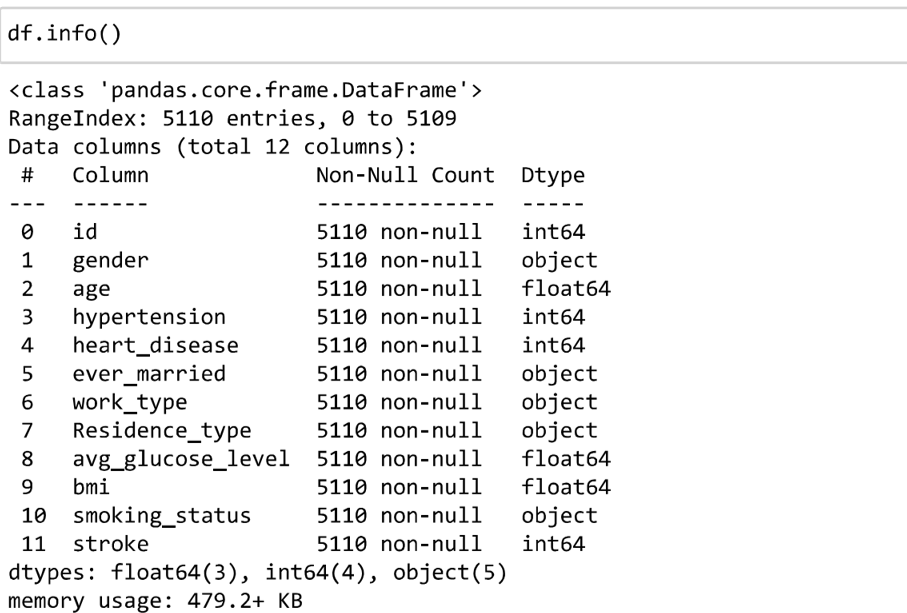
**After replacing , as you can see there are no null values present in our column**



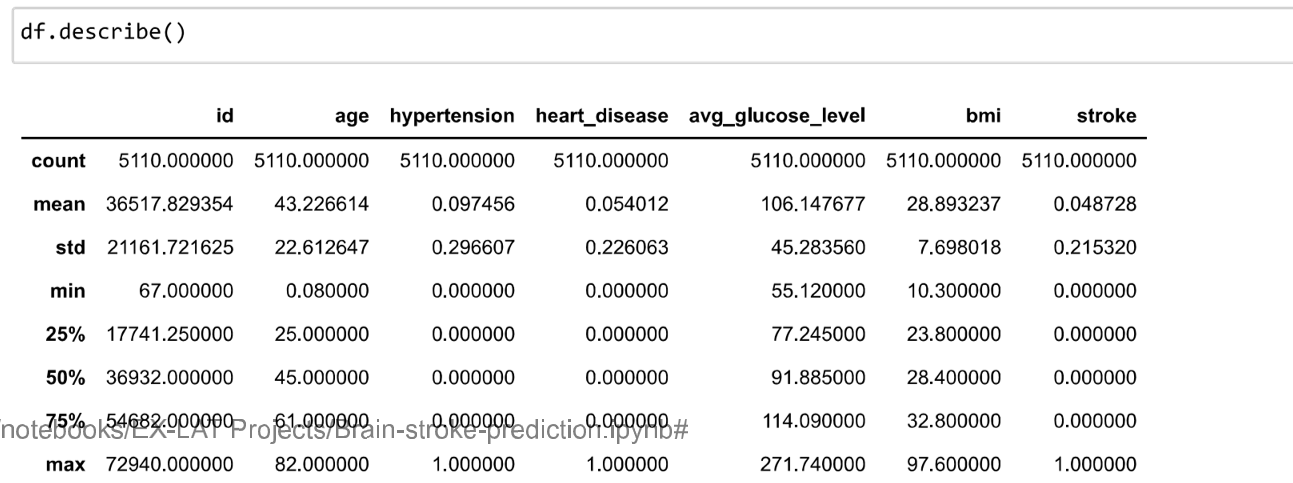
To check the Unique values in the Dataset

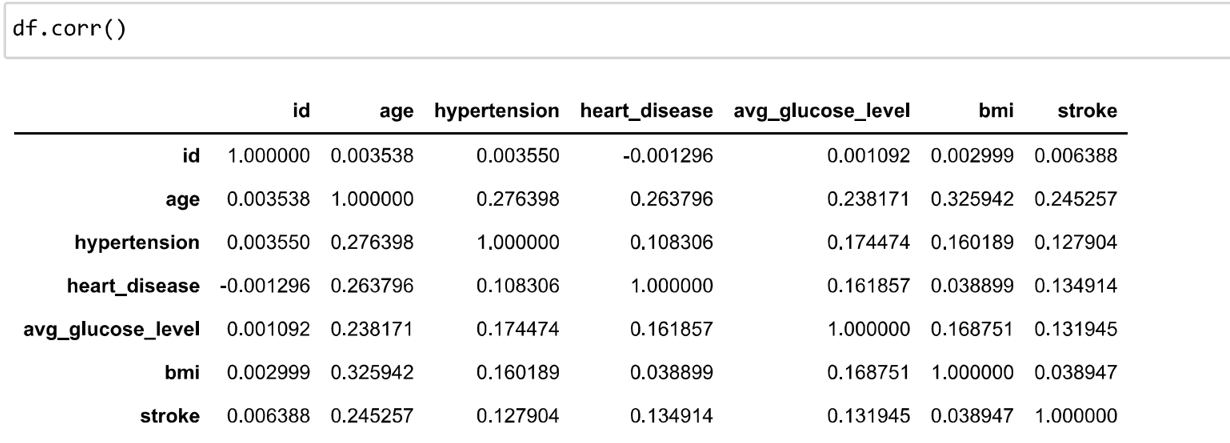


To Check the information about Datatypes present the dataset



Statical analysis:





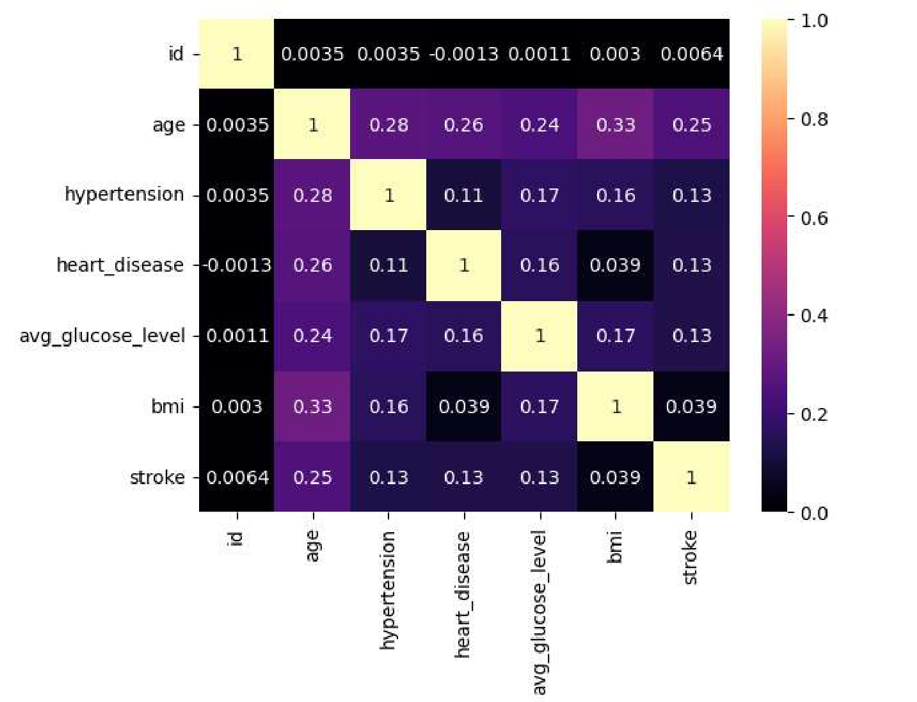
Data visualisation:

Data visualization refers to the representation of data in graphical or visual formats, such as charts, graphs, maps, or plots. It aims to provide a visual summary of information and patterns in data, making it easier to understand, interpret, and communicate insights.

Data visualization plays a crucial role in exploratory data analysis, storytelling, and presenting findings in various fields, including data science, business intelligence, research, and more. It helps to uncover patterns, relationships, trends, outliers, and correlations within the data that might be challenging to identify through raw numbers alone.

## **Heat Map Correlation :**





Count plot:

A count plot is a type of data visualization that displays the count or frequency of each category in a categorical variable. It is commonly used to analyze the distribution or occurrence of different categories in a dataset.

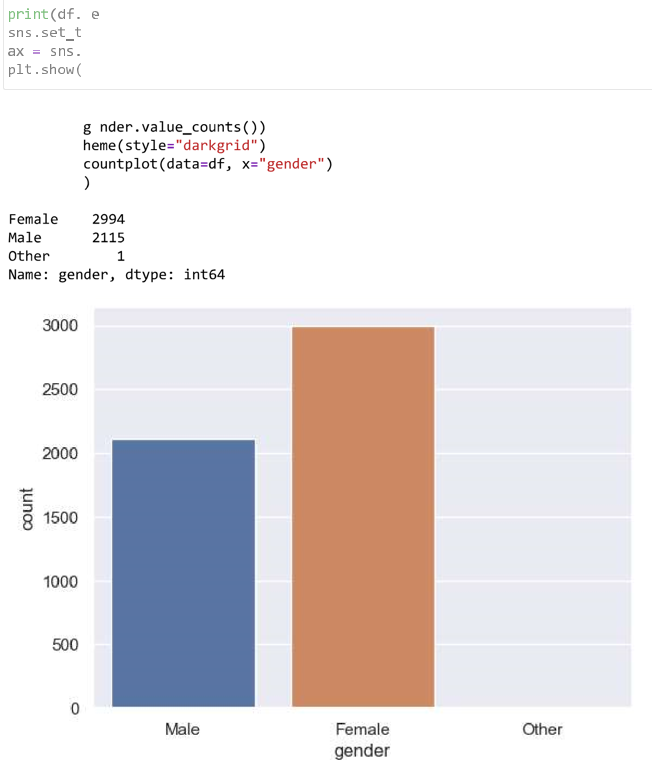
ount plots can be useful for comparing the counts of different categories. By visualizing the bars side by side or in a stacked manner, you can identify the relative frequencies of categories and observe any disparities.

Count plots can be combined with other plots or visual elements to provide more comprehensive insights. For example, you can overlay a count plot with a line plot to visualize the trend in category counts over time

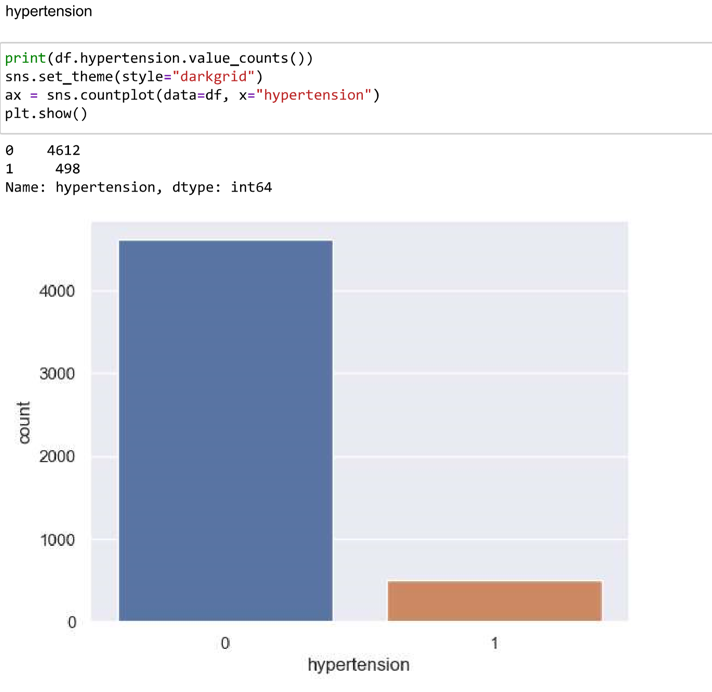
Count plots provide a basic yet effective way to visualize the distribution of categorical data. They offer a quick overview of the frequency of different categories and allow for easy comparisons.

Gender:

*Below, you can see the Females present in our dataset is higher than males.*

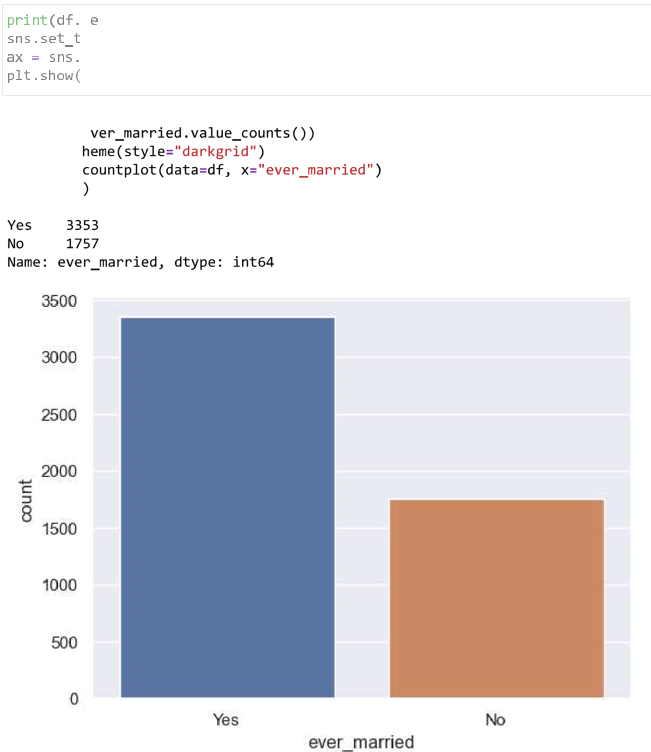


Hypertension:



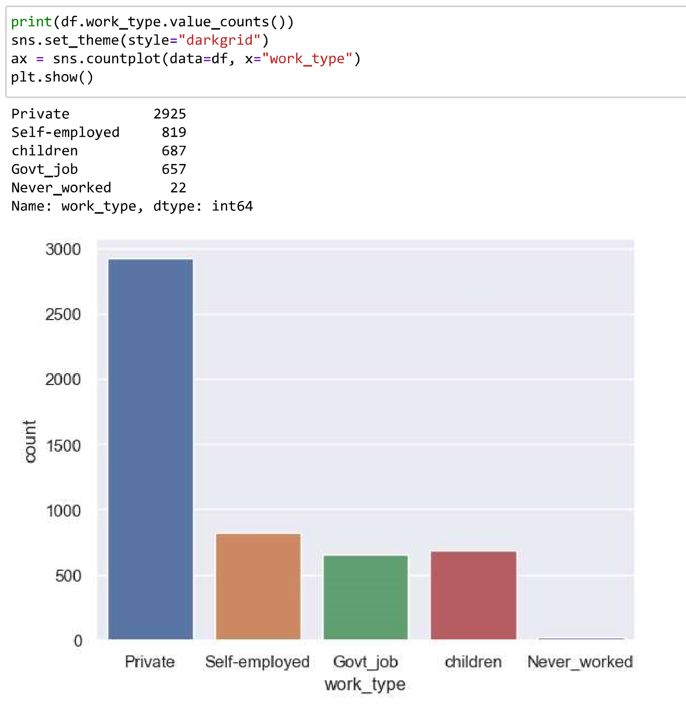
*From above, it shows that less people are suffering from hypertension.*

**Marriage Status:**



*The ratio can seen from above is around 2:1 for being ever married.*

Work Type:



*A lot of people works in Private sector.*

### **Residence Type:**

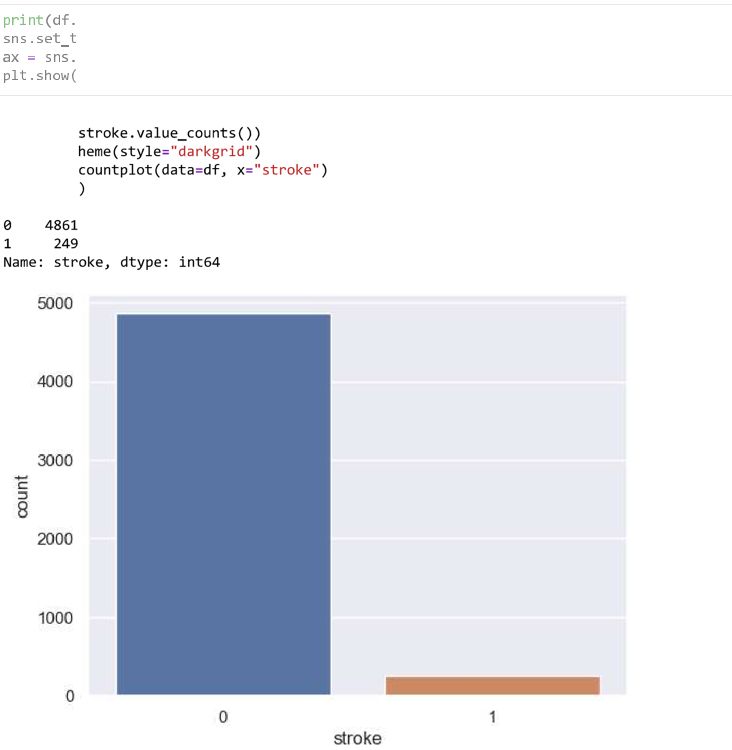


*The residence type is same for people present in our dataset.*

### **Smoking Status:**



Stroke:



*From above dependent variable, we have really less peoples who suffered stroke. But, this also means that our dataset is imbalance. We likely have to use sampling techniques to make the data balance.*

**Scatter plot:**

A scatter plot is a type of data visualization that displays the relationship between two continuous variables. It uses a Cartesian coordinate system to plot individual data points as dots on a graph, with one variable represented on the x-axis and the other variable represented on the y-axis

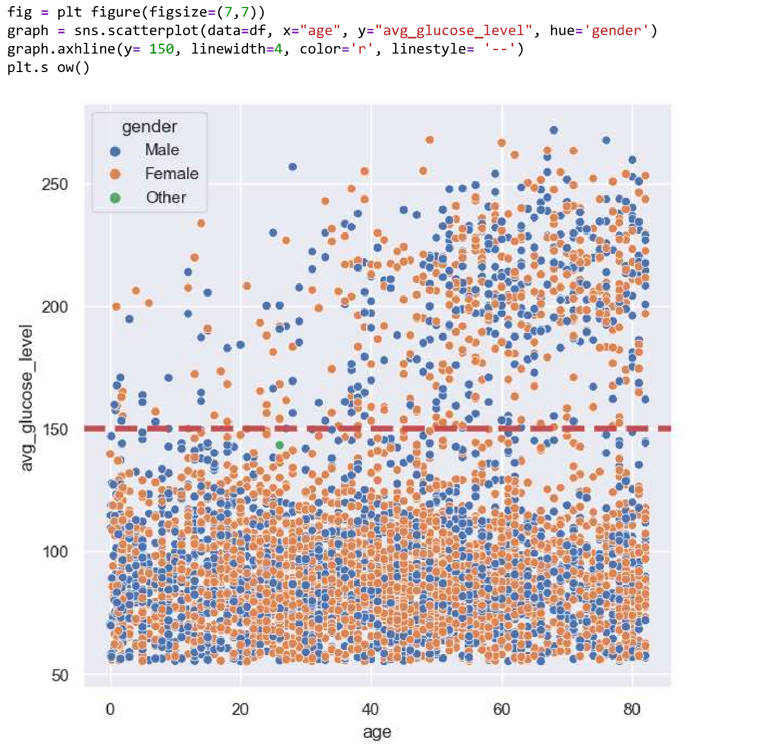
Scatter plots also provide insights into the distribution of data points along the x-axis and y-axis. They show the spread, concentration, or dispersion of the data points and can indicate whether the data is evenly distributed or skewed.

Scatter plots are used to visualize and analyze the relationship between two variables. They can reveal patterns, trends, correlations, clusters, or outliers in the data. By plotting the data points, you can examine whether there is a linear or nonlinear relationship, positive or negative correlation, or any other interesting patterns between the two variables.



*From above plot, we can see that there are lot of people having BMI above 25 are overweight and obese.*

**AVG vs AVG Glucose Level:**

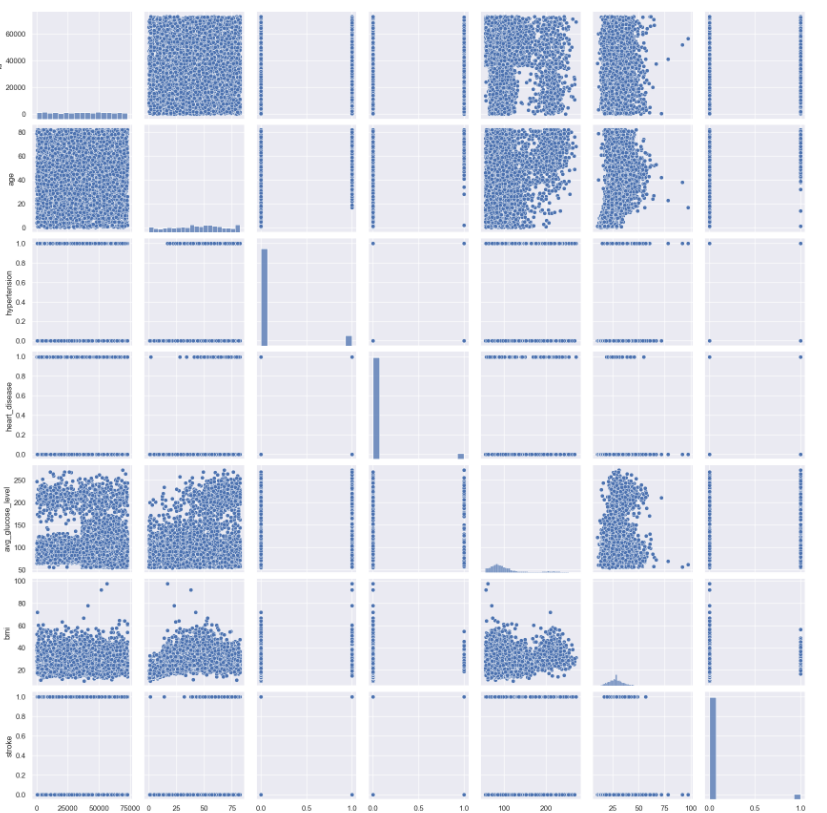


*From above plot, we can see that people having glucose level above 150 are relatively less as compare one below. So, we can say that people above 150 might be suffering from diabetes.*

**Pair Plot:**

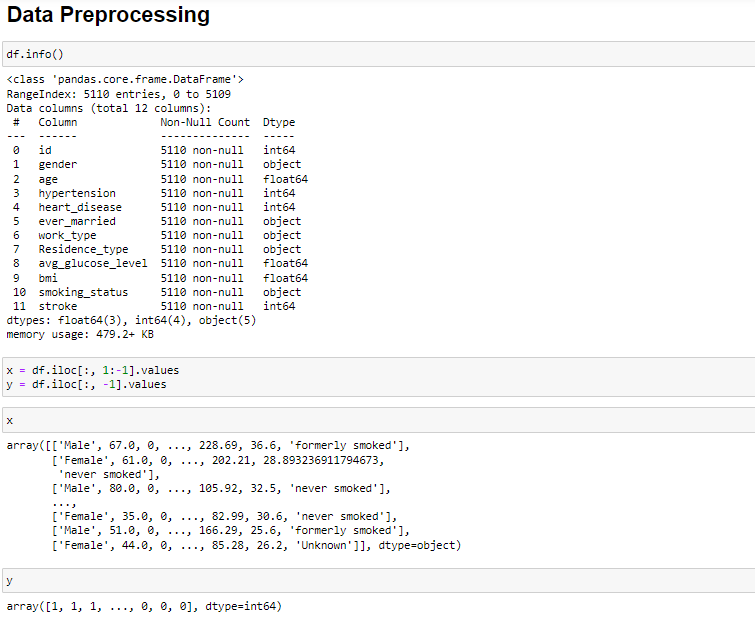
A pair plot, also known as a scatter plot matrix, is a data visualization technique that allows you to explore the relationships between multiple variables simultaneously. It creates a matrix of scatter plots, where each scatter plot represents the relationship between two variables in the dataset.





Data Preprocessing:

Data preprocessing refers to the preparation and transformation of raw data before it can be used for analysis or machine learning algorithms. It involves a series of steps to clean, format, and organize the data to ensure its quality, consistency, and compatibility with the chosen analytical or modeling techniques

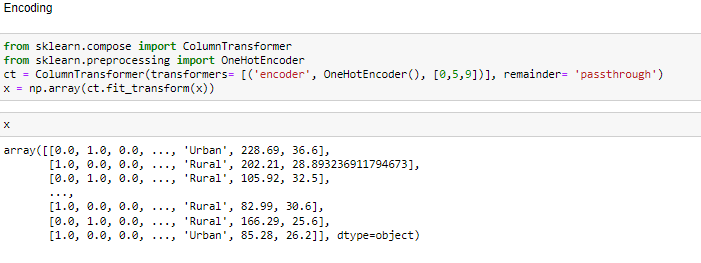


Encoding:

Categorical Encoding

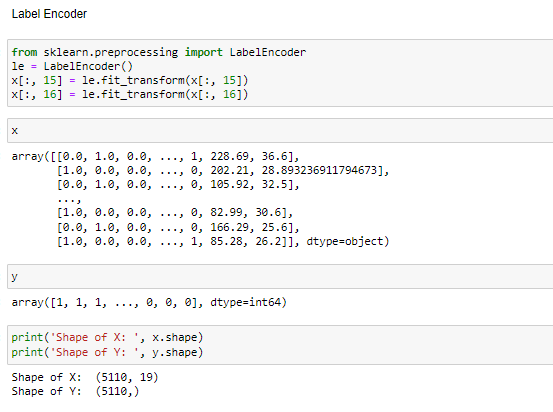
Categorical encoding is the process of converting categorical variables into a numerical representation that can be used as input for machine learning algorithms or other statistical models. Categorical variables are those that represent discrete categories or groups, such as color, gender, or product type. Since most machine learning algorithms require numerical inputs, categorical encoding is necessary to transform these variables into a suitable format.

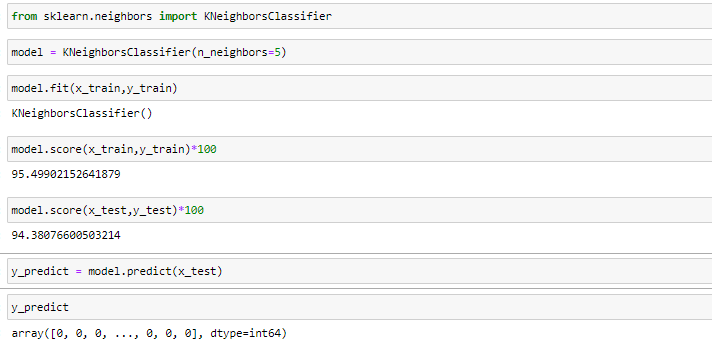
We are using **OneHotEncoder()** to encode the categorical columns: '**gender**', '**work\_type**' and '**smoking\_status**



Label encoding:

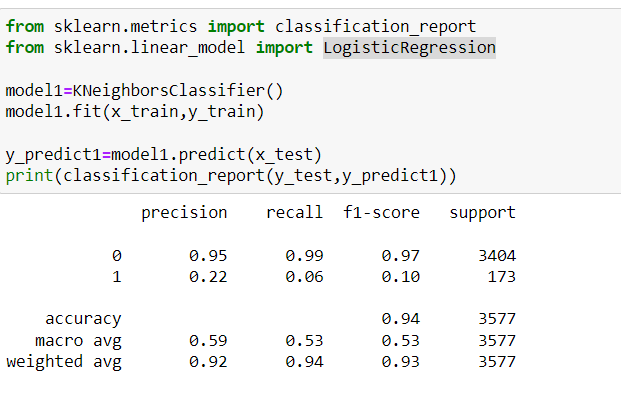
We are using **LabelEncoder()** to encode binary columns: '**ever\_married**' and '**residence\_type**'





**Logistic Regression :**

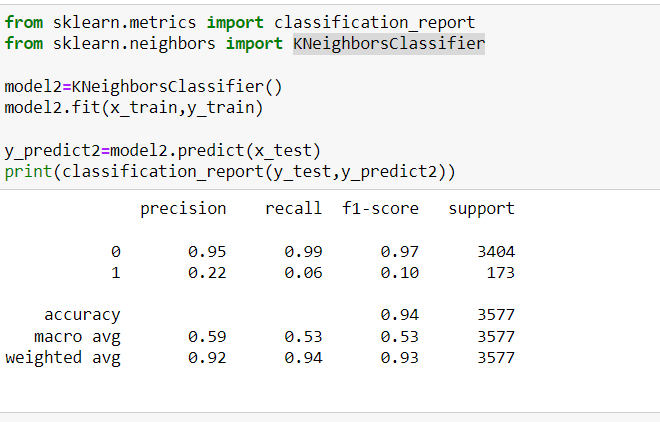
Logistic regression is a statistical modeling technique used to predict binary outcomes or probabilities of an event occurring. It is a type of regression analysis that is widely used in various fields, including machine learning, statistics, and social sciences.

Accuaracy:94%

KNeigborsClassifier:

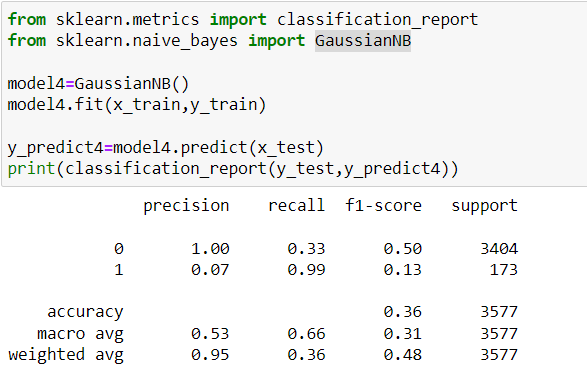
The KNeighborsClassifier is a machine learning algorithm used for classification tasks. It belongs to the family of supervised learning algorithms and is a type of instance-based or lazy learning algorithm. It classifies new instances by comparing them to the labeled instances in its training

dataset. The "K" in KNeighborsClassifier refers to the number of nearest neighbors considered for classification.



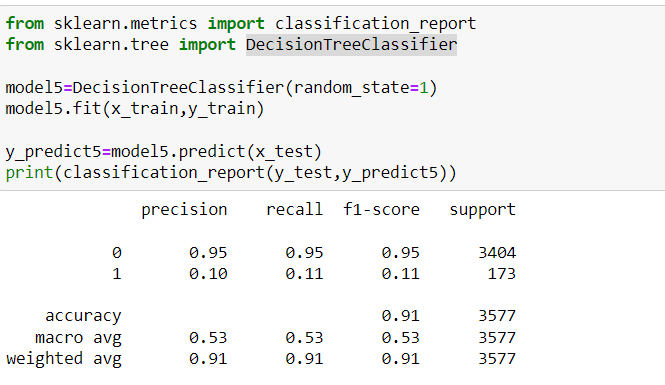
Accuaracy:94%

GaussianNB:GaussianNB, short for Gaussian Naive Bayes, is a popular algorithm for classification tasks, particularly in situations where the features are continuous and assumed to follow a Gaussian (normal) distribution. It is based on the application of Bayes' theorem with the assumption of independence between the features.



Accuaracy:36%

DecisionTreeClassifier: DecisionTreeClassifier is a supervised machine learning algorithm used for classification tasks. It builds a decision tree model that predicts the class label of an instance by following a series of binary decisions based on the feature values.



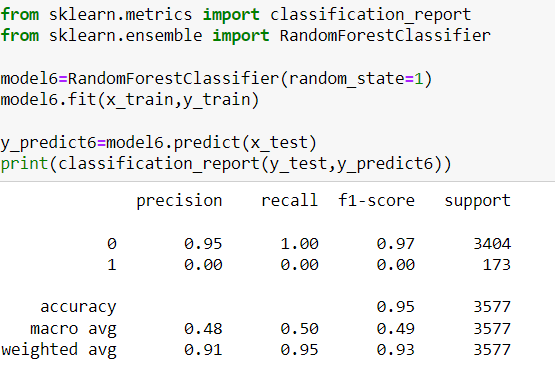
Accuarcy:91%

RandomForestClassifier: RandomForestClassifier is a supervised machine learning algorithm that combines multiple decision trees to create a more robust and accurate model. It belongs to the ensemble learning methods and is particularly effective in handling complex classification tasks.

RandomForestClassifier builds an ensemble of decision trees, where each tree is trained on a random subset of the training data. In addition to using random subsets of the training data, RandomForestClassifier also considers random subsets of features at each split in the decision trees

When making predictions, RandomForestClassifier aggregates the predictions of all the decision trees in the ensemble The final predicted class label is determined by a majority vote or averaging (in the case of regression) among the individual tree predictions

RandomForestClassifier is widely used for various classification tasks, such as image recognition, fraud detection, and medical diagnosis, due to its accuracy, versatility, and resistance to overfitting. It is an effective algorithm for both binary and multi-class classification problems.



Accuaracy:95%

Confusionmatrix:-

A confusion matrix is a performance measurement tool that provides a detailed breakdown of the predictions made by a classification model. It is a square matrix that summarizes the performance of a classification algorithm by comparing the predicted class labels to the actual class labels.

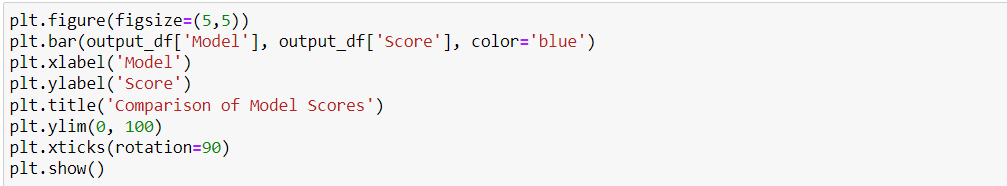
Note: A good rule of thumb is that any accuracy above 70% is considered good, but be careful because if your accuracy is extremly high, it may be too good to be true (an example of Overfitting). Thus, 80% is the ideal accuracy!

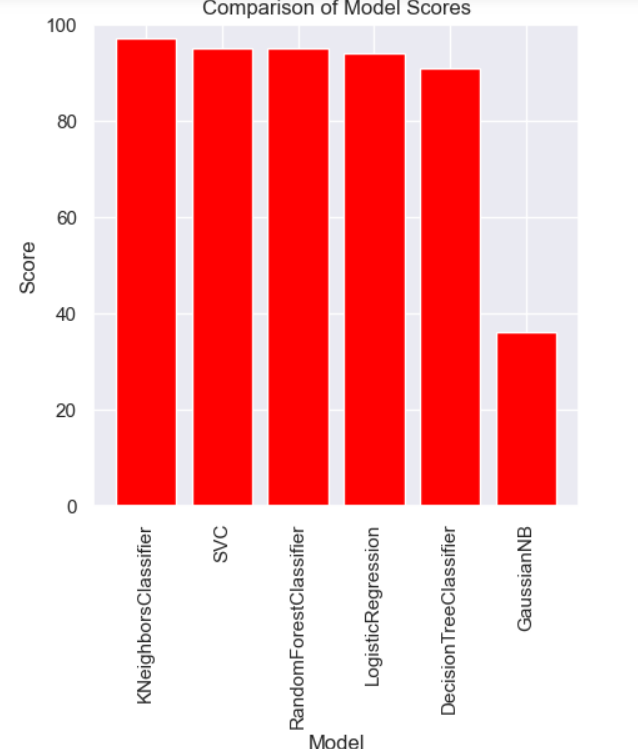


Final Observation:

* Based on the accuracy scores,the RandomForestClassifier is the best choice predicting Brain-stroke.The RandomForestClassifier has a train accuracy of 95%.This model can predict accuaracy when it is trained on data and when it is tested on new data.
* Comparsion of all models RandomForestClassifier is the best model for predicting the stroke because it has the highest accuracy.







Conclusion:

Therefore, after the multiple visualizations of our and going through all the performance of the models.. After that, I came to conclusion that RandomForestClassifier is best model for this dataset.

Reference:

<https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset>.