**Stroke Risk Assessment by Age and Gender and Early detection High-Risk Patients of patients**

**Introduction:**

Stroke is a major global health concern, frequently resulting in severe impairments and even death. Early detection and risk assessment are critical for stroke prevention and management. This research focuses on the key components of stroke risk assessment by age and gender, as well as the early identification of high-risk patients. We hope to establish a rigorous and data-driven approach for identifying individuals at high risk of stroke by incorporating machine learning techniques into stroke risk assessment.

Age and gender have long been recognized as risk factors for stroke. This study aims to understand how these demographic variables influence the chance of stroke by studying a large dataset of patients. Understanding risk profiles based on age and gender is critical for customizing preventative treatments and healthcare initiatives.

Machine learning is a strong technique for early detection because of its ability to process large information and discover complicated patterns. Our research looks into how machine learning algorithms can be used to classify and forecast high-risk patients, allowing for prompt interventions and individualized care. This method has the potential to greatly improve the effectiveness of stroke prevention and patient outcomes.

In this context, our findings have the potential to transform stroke risk assessment and early identification, therefore lowering the burden of stroke-related morbidity and mortality. We believe that our findings will be useful to healthcare practitioners, policymakers, and researchers, resulting in a better educated and targeted approach to stroke prevention and management.

**Objectives:**

* **Determine the Age at Which Stroke Occurs Most Frequently According to the Data:**

We will examine the age distribution of stroke occurrences using a large dataset of stroke cases. This data-driven research will identify the age groups with the highest stroke incidence, enabling for more exact targeting of age-specific stroke preventive and intervention programs.

* **Assess Gender-Specific Stroke Risk Based on the Available Data:**

Using the available patient data, we will conduct a data-driven analysis of gender-specific stroke risk. The purpose of this analysis is to determine whether the evidence supports the presence of gender discrepancies in stroke risk. The findings will help to shape evidence-based strategies for tackling gender differences in stroke prevention and management.

* **Predict Early Detection of Stroke Based on Hypertension, Heart Disease, Average Glucose Level, and BMI, Using the Available Data:**

Leveraging the available patient data, we will employ machine learning techniques to develop a predictive model for early stroke detection. By utilizing hypertension, heart disease, average glucose levels, and BMI as input features, the model will offer data-driven predictions, facilitating timely interventions and personalized care.

**Machine learning problem**

**Regression Problem (Predicting Stroke Risk):**

Regression is a key machine learning problem in the context of stroke risk assessment. It entails forecasting a continuous variable, such as the probability or risk score linked with a person's risk of having a stroke. Regression models estimate a numeric value that quantifies the risk level based on numerous input parameters such as age, gender, medical history, lifestyle factors, and more. A regression model, for example, can estimate a 10% chance of stroke for a certain individual based on their features. In order for healthcare practitioners and policymakers to identify and prioritize high-risk patients for preventative measures and individualized care, accurate regression models are required.

**Classification Problem (Categorizing Risk Groups):**

Another fundamental machine learning challenge in this context is classification. It entails putting people into distinct risk groups based on their characteristics. It can, for example, be used to categorize people as low-risk, moderate-risk, or high-risk for stroke. This is critical for identifying patients who require immediate care or specialized interventions in a timely and efficient manner. Machine learning algorithms examine input data, taking into account characteristics such as age and gender, and classifying individuals into established risk groups. Classification models are useful for streamlining healthcare processes and allocating limited resources to people who require them the most, hence lowering the burden of stroke-related morbidity and mortality.

**Application used:**

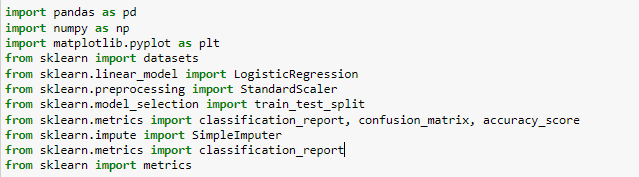
-Jupyter Notebook

**Data Set:**

“Stroke Prediction Dataset” by Fedesoriano

[**https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset**](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)

**Libraries and modules used in the model**

****

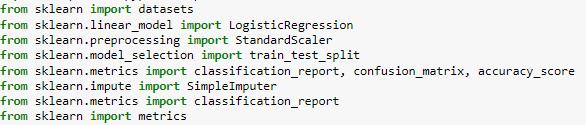
**Data sets Processing**

****

**Data Visualization**

****

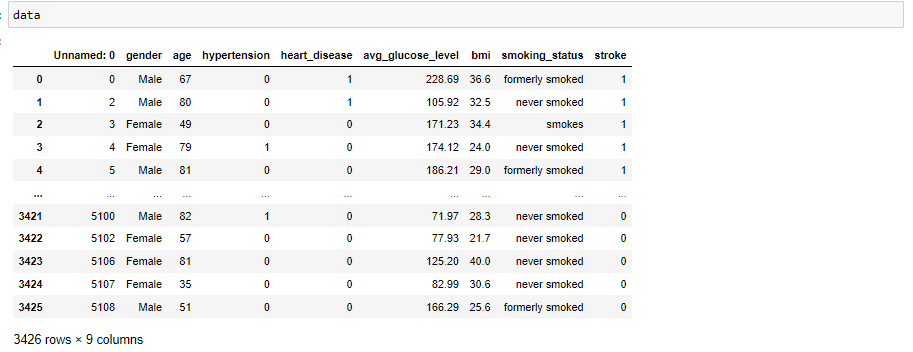
**Machine Learning and Model Building**

****

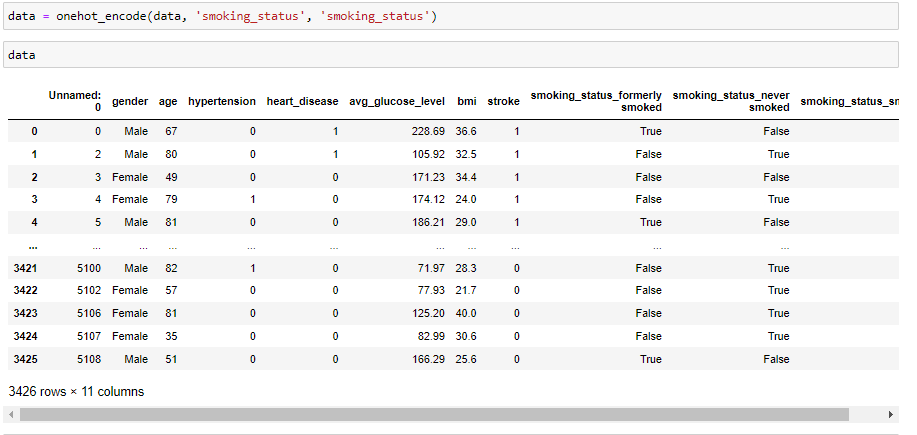
**This code is to read the Data set**

****

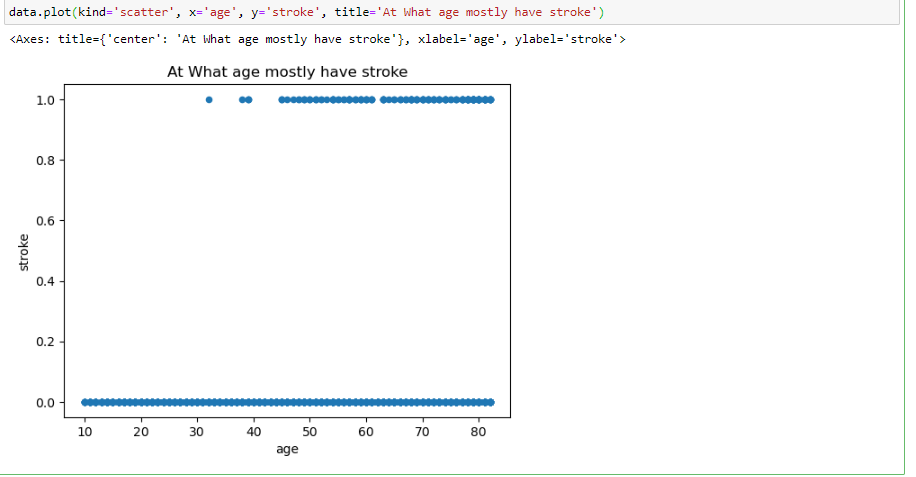
**The Data set is already read**



**We one hot encode the data so the smoking status will be True or False**

****

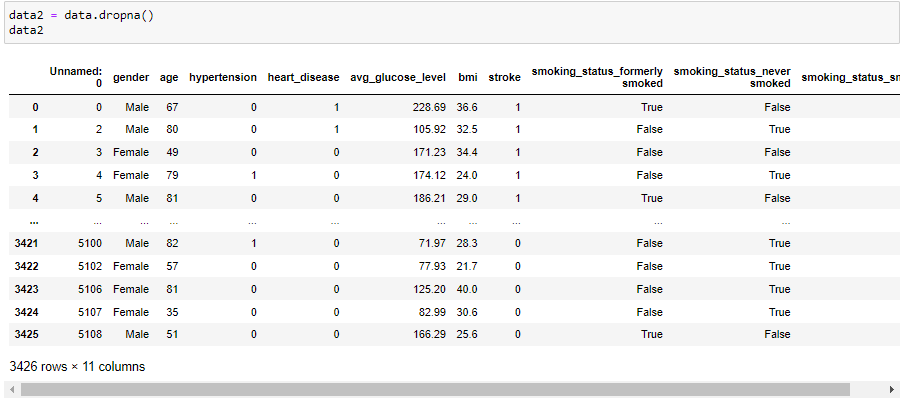
**Problem 1 (What age mostly have stroke)**

****

To acquire a better understanding of the association between age and stroke incidence, we used data visualization approaches. We sought to build a visual representation of the data that would help us to discover patterns and trends by using scatter plots as our visualization tool. We were able to visually evaluate the distribution of stroke cases across different age groups using the scatter plot. We were able to determine whether there were specific age ranges where stroke incidences were more prevalent by observing the concentration and dispersion of data points. Furthermore, we were able to discover potential age-specific risk patterns that could influence targeted preventative efforts by overlaying demographic information such as gender and risk variables.

According to the data analysis, the age group ranging from 45 to 80 has a significantly greater frequency of stroke occurrences. This age group appears to be a pivotal phase when strokes are most commonly reported. The scatter plot data points are clustered within this age range, indicating a significant risk of stroke during these years. These findings highlight the need of focusing preventative efforts and healthcare interventions on people aged 45 to 80. Understanding this age-specific tendency can help to build more effective stroke prevention and management programs for this susceptible population.

**Problem 2 (Predict early detection of stroke base on hypertension, heart disease, average glucose level, BMI)**



As part of our data preprocessing, we implemented a stringent data cleaning step. We decided to delete all data entries with missing values to ensure the robustness and correctness of our study. This comprehensive effort was taken to eliminate potential sources of bias and to ensure that our analytical method would be carried out on a dataset devoid of incomplete or untrustworthy data.

By removing records with missing values, we hoped to generate a dataset free of gaps and inconsistencies, allowing our machine learning algorithms to work with complete and reliable data. This data cleansing stage is critical for avoiding potential distortions in our results, as missing data can skew the analysis and jeopardize the validity of our conclusions. As a result, we have a more solid and trustworthy dataset to base our stroke risk assessment and early detection models on.

**Logistic Regression**



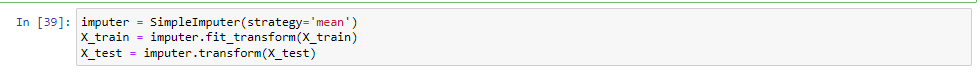
The purpose of this code is to separate the input features (hypertension, heart disease, avg glucose level, and BMI) into a DataFrame X and the target variable (stroke) into a Series y for use in a machine learning model.



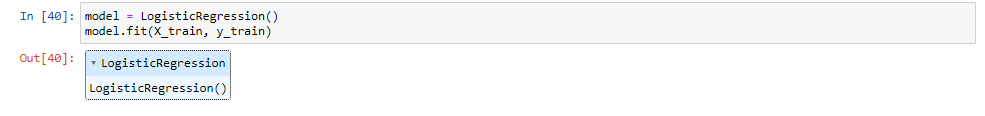
We've just split our data into two sets: X\_train and y\_train for training the machine learning model, and X\_test and y\_test for evaluating its performance. We can now use an appropriate machine learning approach to train the model on the X\_train and y\_train datasets. Once the model has been trained, we will proceed to the testing step, where we will utilize X\_test to make predictions and compare them to the actual results in y\_test. This will allow us to evaluate how well the model performs on untested data. The test\_size parameter is set to 0.2, which means that 20% of the data is set aside for testing, and the random\_state parameter is set to 42 to ensure that the data split remains consistent for reproducibility.



We went through this process to standardize the feature data and ensure that all features have the same scale. This phase is critical for the proper operation of several machine learning algorithms. We maintained data integrity and consistency by uniformly scaling both the training and test data, making the data suitable for subsequent processing and model training.



Using the SimpleImputer, we built this code to address missing values in our dataset. This method fills in any missing data points using the mean values calculated from the available data for each feature. Using the same imputer methodology on both training and test data ensures homogeneity, avoiding any biases caused by different imputation methods. As a result, our dataset is now ready for further data processing and machine learning model training, which ensures data consistency and reliability.



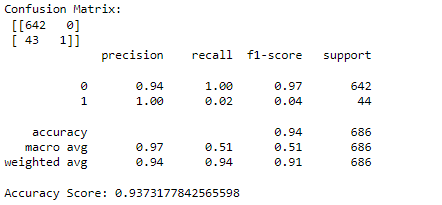
A logistic regression model is developed and trained in this code to predict the occurrence of strokes. It learns the association between characteristics and stroke occurrence using training data (X\_train containing feature inputs and y\_train containing known stroke outcomes). This trained model can then be used to predict stroke outcomes on new data.



This line of code predicts the test data (X\_test) using the trained logistic regression model (model) and assigns the expected outcomes to the variable y\_pred.



These lines, when combined, allow an evaluation of the model's ability to classify stroke outcomes, providing a comprehensive performance summary that includes accuracy as well as precise precision and recall measures.

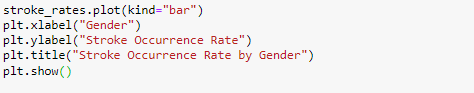


The summary provided provides a snapshot of the model's performance in stroke classification. According to the confusion matrix, it correctly predicted 642 cases of 'no stroke' and misclassified 1 incidence of'stroke' out of 44. The categorization report highlights excellent precision for 'no stroke' (94%) and low recall for'stroke' (2%), resulting in an asymmetric F1-score. The overall model accuracy is around 93.7%, suggesting success in properly detecting 'no stroke' situations but less effective performance in accurately identifying'stroke' instances. This disparity could be ascribed to a class imbalance, with 'no stroke' cases much outnumbering'stroke' instances.

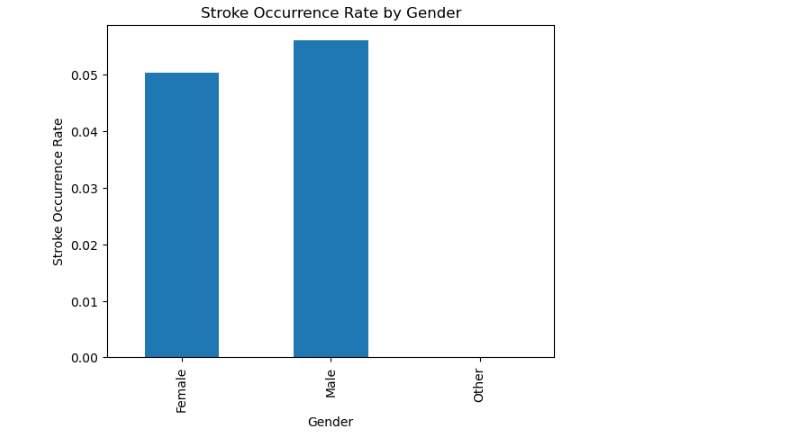
**Problem 3 (What gender has a higher risk of having stroke?)**

****

This code calculates and shows the average stroke rates for each gender in the dataset, allowing us to compare how often strokes occur on average between different genders.



This visualization aids us in immediately identifying any gender-related trends in stroke incidence, making it easier to draw insights from the data and inform decision-making.



According to the data, males are at a higher risk of stroke, with a significantly higher incidence than females. This study implies that there is a gender difference in stroke occurrence within the dataset. Understanding and acknowledging this increased risk in men is essential for personalizing preventive healthcare programs and strategies. Further research into the factors that contribute to this gender disparity can give significant insights for healthcare professionals and policymakers, enabling for more effective and targeted actions to reduce the burden of stroke-related morbidity and mortality among men.