**Project : Brain Stroke Analysis**

1. **Introduction:**

Brain strokes, also known as cerebrovascular accidents, are a critical medical condition that demands timely diagnosis and intervention. Stroke occurs when blood flow to the brain is disrupted, leading to a range of severe neurological and physical impairments. Rapid and accurate detection of stroke is crucial for ensuring appropriate medical care and minimizing potential long-term consequences.

This report focuses on the analysis and detection of brain strokes using a comprehensive dataset collected from diverse medical sources. The dataset under investigation is a compilation of medical records and patient information, spanning a range of demographic, clinical, and lifestyle attributes. It encompasses a cohort of individuals with both stroke occurrences and non-stroke instances, enabling us to build and evaluate predictive models for stroke detection.

Significance of Brain Stroke Detection:

Timely detection of brain strokes can significantly impact patient outcomes. A swift diagnosis enables medical practitioners to administer appropriate treatments, such as clot-dissolving medications or surgical interventions, which can potentially prevent further damage to brain tissue and enhance the likelihood of recovery. Additionally, accurate identification of individuals at risk of stroke empowers healthcare providers to implement preventative measures and lifestyle interventions to mitigate the risk factors associated with this condition.

Objectives:

In this report, we outline our approach to analyzing the dataset and constructing a predictive model for brain stroke detection. We delve into the exploratory data analysis (EDA) process, feature engineering, model selection, and training strategies. Through a comprehensive evaluation of our model's performance, we aim to shed light on its effectiveness in identifying patients who may be susceptible to brain strokes.

The insights derived from this analysis have the potential to contribute to advancements in medical diagnosis and patient care. Our findings can aid medical professionals in making informed decisions and improve the overall quality of life for individuals at risk of brain strokes.

1. **Dataset Overview:**

The dataset used in this analysis is sourced from Client, and it comprises a comprehensive collection of patient-related data aimed at facilitating the detection of brain strokes. The dataset contains a total of 5112 rows and 11 features, each representing an individual patient.

Features:

Features provide crucial information about the patients' health profiles. Some of the key features included in the dataset are:

1. Id : Represents the id of each patients
2. Age: The age of the patient, represented as a numeric value.
3. Gender: The gender of the patient, encoded as a categorical variable (e.g., 'Male', 'Female').
4. Hypertension: A binary indicator of whether the patient has a history of hypertension (0: No, 1: Yes).
5. Heart Disease: A binary indicator of whether the patient has a history of heart disease (0: No, 1: Yes).
6. Marital Status: The marital status of the patient, encoded categorically.
7. Work Type: The type of work the patient is engaged in, represented categorically.
8. Residence Type: The type of residence the patient lives in, categorized as 'Urban' or 'Rural'.
9. Average Glucose Level: The average glucose level in the patient's blood, represented as a numeric value.
10. Body Mass Index (BMI): The patient's BMI, represented as a numeric value.
11. Smoking Status: The smoking status of the patient, encoded categorically ('Never smoked', 'Formerly smoked', 'Smokes').
12. **Exploratory Data Analysis:**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

The Exploratory Data Analysis (EDA) phase is crucial for understanding the underlying patterns, relationships, and potential challenges present in the brain stroke detection dataset. This section aims to provide a comprehensive overview of the dataset's main features and their relevance to stroke detection.

* Bar Plot:

Bar plots are typically used to display and compare categorical data, showing the distribution or frequency of different categories. Bar plots are better suited for visualizing categorical data, like comparing the frequency of stroke occurrences across different age groups or genders, rather than for visualizing the intricate and multidimensional data often encountered in medical diagnoses.

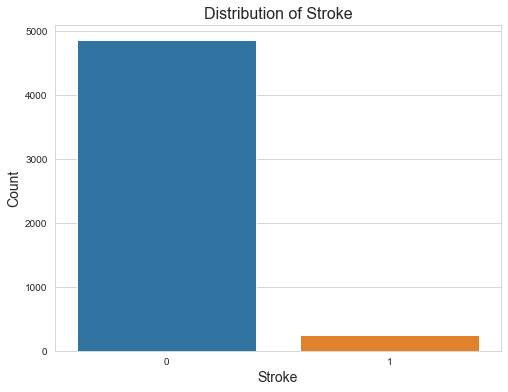


Fig1. Distribution of stroke

In the above graph we can see minimum number of people will affect for brain stroke.

* Count Plot: A count plot is a type of bar plot that shows the frequency of occurrences of different categorical values in a dataset. It's a simple and effective way to visualize the distribution of categorical data. Each category is represented on the x-axis, and the height of the bars represents the count or frequency of each category.

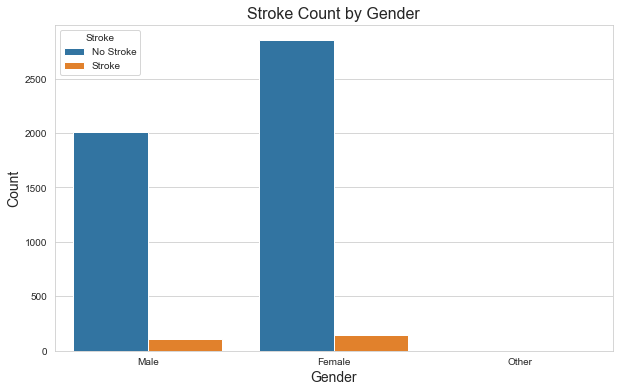


Fig.2 Distribution of stroke and gender

In the above graph we can observed that both Male and Gender are affected by brain stroke disease.

Studies have found that stroke occurrence can vary between genders. In many populations, men have a slightly higher risk of stroke than women. However, stroke can be more severe in women, leading to higher rates of disability and mortality.

Some risk factors for stroke can affect genders differently. For example, high blood pressure (hypertension) is a significant risk factor for strokes, and its impact may vary between men and women. Hormonal changes, such as those related to pregnancy and menopause, can also influence stroke risk in women.

Hormonal factors, such as the use of oral contraceptives and hormone replacement therapy, can impact stroke risk in women. Estrogen levels, which fluctuate during a woman's lifetime, can influence blood vessel health and clotting tendencies.

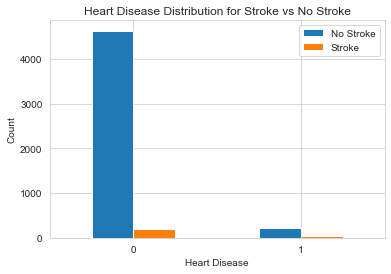


Fig3. Distribution of Heart Disease with stroke

Atherosclerosis is a condition in which arteries become narrowed and hardened due to the buildup of plaque, composed of cholesterol, fat, and other substances. This condition can occur in the arteries supplying both the heart and the brain. If a blood clot forms within a narrowed artery in the heart (coronary artery) or another part of the body and then travels to the brain, it can cause an ischemic stroke by blocking blood flow to the brain.

From the above fig.3 we can see who does not have heart disease they have more chances of suffering from brain stroke. But it may vary from dataset to dataset.

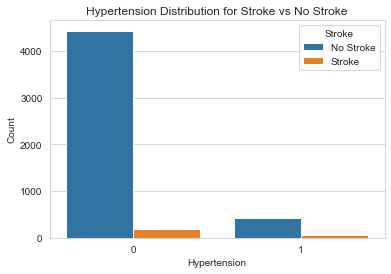


Fig4. Distribution of Hypertension and stroke

Hypertension, also known as high blood pressure, is a major risk factor for brain stroke. It can significantly increase the likelihood of both ischemic and hemorrhagic strokes. Hypertension can disrupt the normal flow of blood, leading to the formation of blood clots. These clots can block blood vessels in the brain, causing ischemic strokes, which are the result of reduced blood flow to brain tissue.

Hypertension can cause the heart's chambers to enlarge and weaken. This can lead to abnormal heart rhythms like atrial fibrillation, where blood may pool and clot within the heart. Clots formed in the heart can travel to the brain, causing strokes.

Above fig.4 is showing that who does not have hypertension they have more chances of getting brain stroke, but as per medical research people suffering from hypertension have more chances of getting brain stroke.

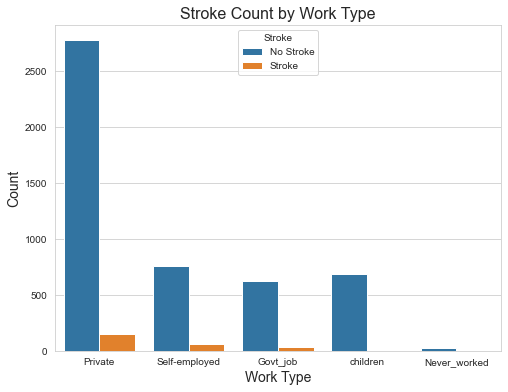


Fig5. Distribution of stroke and work type

The type of work an individual engages in can potentially affect their risk of stroke. Work-related factors can influence various aspects of health, including lifestyle habits, stress levels, physical activity, and exposure to certain risk factors.

Jobs that require prolonged sitting or minimal physical activity can contribute to a sedentary lifestyle. Lack of physical activity is a risk factor for various health conditions, including obesity, hypertension, and cardiovascular diseases, all of which can increase the risk of stroke.

High-stress occupations, such as those with demanding deadlines, high responsibilities, or constant pressure, can lead to chronic stress. Prolonged stress can contribute to the development of hypertension and other risk factors for stroke. Irregular shift work or night shifts can disrupt the body's natural sleep-wake cycle and impact overall health. Shift workers might have higher levels of stress, sleep disturbances, and metabolic issues, all of which can influence stroke risk.

From fig5 it is observed that people working in private sector have higher chances of getting brain stroke, self-employed and government job people also suffering from brain stroke.

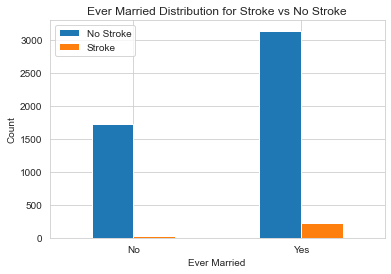


Fig.6 Count plot between Ever married and stroke

Above graph of ever married and stroke shows people are married they are suffering from brain stroke.

Marital status can have an impact on health outcomes, including the risk of brain stroke, but it's important to understand that the relationship between marital status and stroke is influenced by a variety of factors. Marriage itself is just one of many factors that can influence an individual's overall health and well-being.

Married individuals often have social support systems and someone to help monitor their health, encourage healthy habits, and provide emotional support. Strong social support can have positive effects on overall health, including stroke risk.

Married individuals might be more likely to engage in healthier behaviors, such as regular exercise, balanced diets, and avoiding risky behaviors like smoking and excessive alcohol consumption. These lifestyle choices can contribute to reduced stroke risk.

Marriage can provide emotional support and companionship, potentially reducing stress levels. Chronic stress is associated with a higher risk of stroke, so having a supportive partner can mitigate this risk.

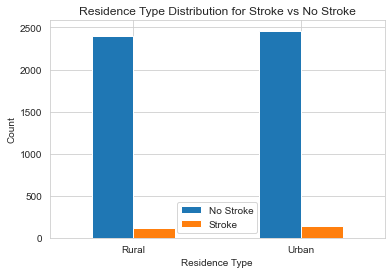


Fig.7 Count Plot of Residence type and Stroke

Urban areas often have higher levels of air pollution, which can impact cardiovascular health and increase stroke risk. Long-term exposure to pollutants is associated with various health issues, including stroke. From above fig 7 we can see urban areas have brain stroke disease.

Living in remote or rural areas might limit access to healthcare resources, including timely medical attention during emergencies. Limited access to medical services can lead to delayed treatment and higher stroke mortality rates.

* Histogram Plot:

Histograms provide a clear visual representation of how data is distributed across different ranges or bins. In the case of brain stroke detection, histograms can show the distribution of various factors such as age, time since stroke, risk factors, and clinical metrics.

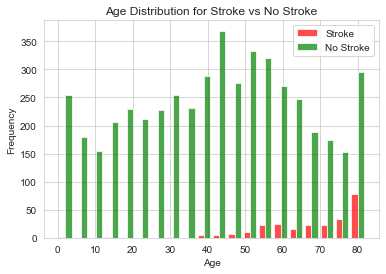


Fig.8 Histogram plot for age and stroke

The risk of stroke rises significantly with advancing age. The majority of strokes occur in individuals over the age of 65, and the risk doubles for each decade after the age of 55.

Above Histogram is giving clear visualization of between age 55 to 60 and 75 to 80 have higher chance of causing stroke.

* Heatmap :

Heatmaps are valuable visualization tools for brain stroke detection and analysis, as they allow for the representation of complex data patterns and relationships in a visually intuitive way.

Brain stroke detection involves analyzing various factors, such as patient demographics, medical history, imaging data, and clinical metrics. Heatmaps can display these multidimensional datasets, making it easier to identify correlations and trends that might not be apparent in raw data.

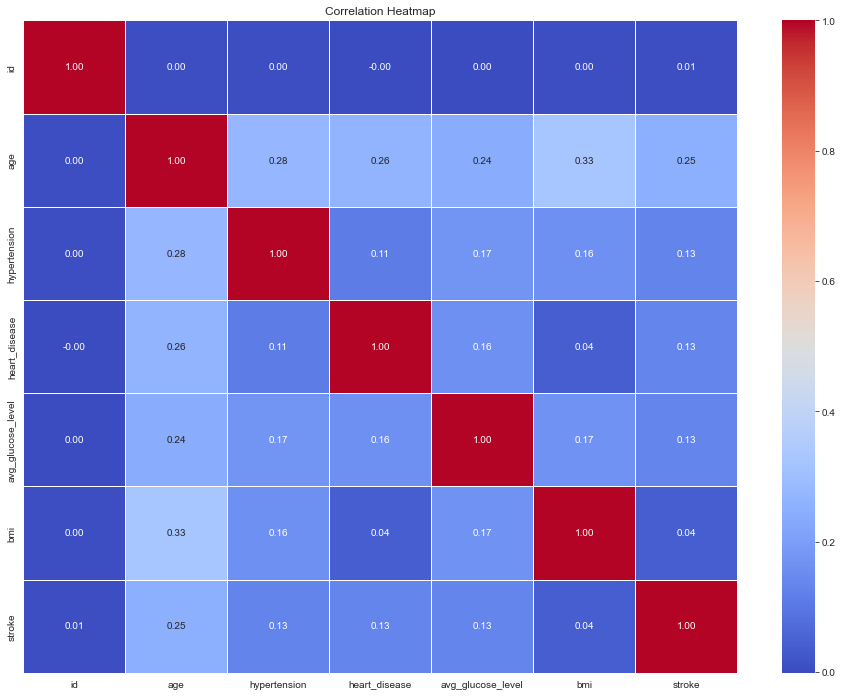


Fig.9 Heatmap for Brain Stroke

1. **Feature Engineering:**

Feature engineering involves the process of creating, transforming, or selecting features to enhance the predictive power of machine learning models. In the context of brain stroke detection, careful feature engineering can uncover hidden patterns and relationships that contribute to accurate predictions.

* Missing Data:

For various reasons, many real world datasets contain missing values, often encoded as blanks, NaNs or other placeholders. Such datasets however are incompatible with scikit-learn estimators which assume that all values in an array are numerical, and that all have and hold meaning.

In brain stroke dataset we have missing values in ‘bmi’ attribute so to remove missing value we have fill the values with mean

The syntax for removing missing value is:

dataset['bmi']=dataset['bmi'].fillna(dataset['bmi'].mean())

* Label Encoding:

Label encoding is a preprocessing technique used to convert categorical data into numerical format, which is often required for machine learning algorithms to process the data effectively.

In machine learning projects, we usually deal with datasets having different categorical columns where some columns have their elements in the ordinal variable category.

In our dataset we have column ‘gender’, ‘ever\_married’, ‘work\_type’, ‘residence\_type’ and ‘smoking\_status’ we will convert this into numerical format.

To perform this operation we will import LabelEncoder library and use .fit\_transform to convert it into numerical.

* Splitting the dataset:

We have split the data into train and test to train and test the model.

Sklearn gives these feature of splitting dataset using train\_test\_split. It will split the data into X\_train, X\_test, y\_train and y\_test.

* StandardScaler:

The StandardScaler is a common preprocessing technique used in machine learning to standardize or normalize the features of a dataset. It's applied to ensure that the features have similar scales, which can be beneficial for many machine learning algorithms.

It's important to note that not all algorithms require standardized features. Some algorithms, like decision trees or random forests, are less sensitive to feature scales. However, in general, using StandardScaler is a good practice to prepare your data for a wide range of machine learning algorithms and to ensure stable and reliable model performance.

1. **Model Selection:**

Selecting an appropriate model for brain stroke detection depends on various factors, including the nature of the data, the complexity of the problem, the availability of data, and the goals of the detection task. As our dependent variable is categorical in nature we will use classification algorithms such as,

1. Logistic Regression Algorithm
2. Support Vector Machine
3. K-Nearest Neighbors
4. Decision Tree Algorithm
5. Random Forest Algorithm
6. **Model Training:**

Train the model using above algorithms, including hyperparameter tuning, cross-validation, and any other relevant training strategies.

1. We will import the required library like SVC, DecisionTreeClassifier, LogisticRegression, etc.
2. We will do hyperparameter tuning for better model performance so it will give better accuracy.
3. We will predict the outcomes using different algorithms.
4. A confusion matrix is a fundamental tool in machine learning for evaluating the performance of classification algorithms. It provides a clear and detailed summary of the predictions made by a model and helps assess its accuracy and effectiveness.
5. Check for accuracy, bias and variance it will give clarity for model performance.
6. **Result:**

After training the model with different algorithms we got the below outcomes:

1. For Logistic Regression:

Accuracy = 0.9471624266144814

Bias = 0.9522994129158513

Variance = 0.9471624266144814

1. For Support Vector Machine:

Accuracy = 0.9461839530332681

Bias = 0.9527886497064579

Variance = 0.9461839530332681

c) For K-Nearest Neighbors:

Accuracy = 0.9461839530332681

Bias = 0.9530332681017613

Variance = 0.9461839530332681

1. For Decision Trees:

Accuracy = 0.9412915851272016

Bias = 0.9525440313111546

Variance = 0.9412915851272016

1. For Random Forest:

Accuracy = 0.9471624266144814

Bias = 0.9522994129158513

Variance = 0.9471624266144814

1. **Conclusion:**

From the results we can conclude that all the algorithms are best for detecting brain stroke. All the algorithm giving accuracy as 94% which indicates that the model is performing excellent.