Data Analysis and Machine Learning Implementation

I. Project Overview

The analysis aims to delve into key attributes of stroke patients such as ID, Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Average Glucose Level, BMI, Smoking Status, and Stroke, using these data points to derive insights into health factors and potential stroke risk.

Stroke is a leading cause of disability and death worldwide, making it essential to understand the various factors that contribute to its occurrence. By examining a comprehensive dataset of stroke patients, we can uncover critical insights that inform prevention strategies, improve patient outcomes, and guide healthcare policies. This analysis seeks to identify patterns and correlations between different health attributes and the likelihood of stroke, thus providing a foundation for targeted interventions and personalized medicine.

Healthcare Stroke Data Analysis

The analysis of health stroke data, as described, utilizes several key attributes to uncover patterns and risk factors associated with strokes. Here's how each attribute contributes to understanding stroke risk and patient profiles:

1. **ID:** A unique identifier for each patient in the dataset. This attribute is essential for differentiating between individuals and maintaining data integrity when analyzing patient records.

2. **Gender:** Understanding the gender distribution among stroke patients can reveal gender-specific risk factors and outcomes. It helps in tailoring prevention and treatment strategies to be more effective for different genders.

3. **Age:** Age is a critical factor in stroke risk analysis. Older age groups are typically at a higher risk for stroke, and analyzing age distribution helps in identifying age-related patterns and the effectiveness of age-targeted interventions.

4. **Hypertension:** Hypertension is a major risk factor for strokes. By analyzing this attribute, we can assess the prevalence of high blood pressure among stroke patients and its impact on stroke incidence.

5. **Heart Disease:** This attribute indicates whether the patient has a history of heart disease, which is closely linked to stroke risk. Understanding this correlation can help in identifying patients at higher risk and in developing integrated treatment plans.

6. **Ever Married:** Marital status can influence lifestyle and health behaviors, which in turn affect stroke risk. This attribute helps in exploring the social and behavioral factors associated with stroke incidence.

7. **Work Type:** Different work environments and stress levels can impact health. Analyzing work type helps in understanding occupational risk factors and designing workplace health programs to mitigate stroke risk.

8. **Residence Type:** This attribute indicates whether the patient lives in an urban or rural area. Residence can influence access to healthcare, lifestyle, and environmental factors, all of which contribute to stroke risk.

9. **Average Glucose Level:** High glucose levels are associated with diabetes, which is a significant risk factor for stroke. This attribute helps in assessing the role of blood sugar levels in stroke risk and in managing diabetes as part of stroke prevention.

10. **BMI (Body Mass Index):** BMI is an important measure of body fat and overall health. High BMI is associated with increased stroke risk due to its link with other risk factors like hypertension and diabetes. Analyzing BMI helps in understanding the impact of obesity on stroke incidence.

11. **Smoking Status:** Smoking is a well-known risk factor for strokes. This attribute allows for the assessment of smoking prevalence among stroke patients and the development of targeted smoking cessation programs to reduce stroke risk.

12. **Stroke:** This is the target variable indicating whether the patient has experienced a stroke. Analyzing the other attributes in relation to this variable helps in identifying patterns, risk factors, and potential preventive measures for strokes.

By analyzing these attributes, healthcare providers and researchers can better understand the factors contributing to stroke risk and develop more effective prevention and treatment strategies. This detailed analysis not only supports clinical decision-making but also enhances patient care by addressing the specific needs and risk profiles of different patient groups.

II. Libraries and Data Handling

Libraries Used:

Libraries Used are Pandas and NumPy for data manipulation, Matplotlib and Seaborn for data visualization, Scikit-learn for machine learning tasks.

1. **Pandas:** This library is indispensable for data manipulation and analysis. It offers powerful data structures and operations for handling tabular data, which is essential for tasks like cleaning and preprocessing datasets before feeding them into machine learning models.

2. **NumPy:** This library is essential for numerical and scientific computing in Python. It provides support for large, multidimensional arrays and matrices, which are the foundation for most data manipulation tasks in scientific computing. NumPy also offers a wide range of mathematical functions to perform operations on these

arrays, making it crucial for tasks like numerical simulations, statistical analysis, and preparing data for machine learning models.

3. **Matplotlib:** A versatile plotting library that provides a comprehensive set of tools for creating static, interactive, and animated visualizations in Python. It's commonly used alongside Pandas for visualizing data distributions, trends, and relationships, aiding in exploratory data analysis.

4. **Seaborn:** Built on top of Matplotlib, Seaborn is tailored for statistical data visualization. It simplifies the process of creating informative and visually appealing plots, particularly for tasks like comparing distributions, identifying correlations, and exploring multivariate relationships within datasets.

5. **Scikit-learn:** A robust machine learning library offering a wide range of algorithms and tools for tasks such as classification, regression, clustering, and dimensionality reduction. It provides a unified interface for model training, evaluation, and deployment, making it a go-to choose for building predictive models in Python.

Scikit-learn's specific modules:

- **`train_test_split` from `sklearn.model_selection`:**
  Essential for splitting datasets into training and testing sets, facilitating model evaluation and validation.

- **`StandardScaler` from `sklearn.preprocessing`:**
  Used for standardizing feature scales to ensure uniformity and improve model performance, particularly for algorithms sensitive to feature magnitudes.

- **LinearRegression` from `sklearn.linear_model`:**
  A fundamental algorithm for regression tasks, used for predicting continuous target variables based on input features.

- **`LogisticRegression` from `sklearn.linear_model`:**
  A popular algorithm for binary classification tasks, such as predicting whether an email is spam or not, based on various features.

- **`accuracy_score`, `classification_report`, `confusion_matrix` from `sklearn.metrics`:**
  Metrics and tools for evaluating the performance of classification models, including accuracy, precision, recall, F1-score, and confusion matrix.

- **`mean_squared_error` from `sklearn.metrics`:**
  A common metric for evaluating regression models, quantifying the average squared difference between predicted and actual target values.

- **RandomForestClassifier` from `sklearn.ensemble`:**
  A versatile ensemble learning method capable of handling both classification and regression tasks, known for its robustness and flexibility.

These libraries and modules together form a comprehensive toolkit for data analysis, visualization, and machine learning tasks in Python.

Data Loading and Preprocessing:

Data Loading: Data is loaded from a CSV file into a DataFrame.

- **Loading Data from CSV with pandas:** The dataset is loaded into a pandas DataFrame from a CSV file, which is a common practice for data analysis. Using pd.read_csv(), this method converts structured data into a DataFrame, enabling powerful data manipulation capabilities within Python.

- **Loading Data from CSV with NumPy:** The dataset is loaded into a NumPy array from a CSV file, suitable for numerical and scientific computing. Using

np.genfromtxt(), this method converts the data into an array, allowing for efficient numerical operations and computations within Python.

Data Cleaning and Preprocessing: Basic preprocessing such as converting dates to datetime objects and handling categorical data transformation is performed.

- **Using DataFrame Methods and Attributes:** Methods like. head(), .tail(), .shape, .columns, .dtypes, .unique(), .nunique(), .describe(), .value_counts(), and .isnull() are used for gaining insights into the dataset structure, unique values, missing values, and statistical summaries. These methods help in understanding the dataset better and planning the preprocessing steps accordingly.

- **Handling Missing Values:** Dealing with missing data is a crucial aspect of data preprocessing. Techniques such as imputation (filling missing values with a specific value like mean, median, or mode), dropping rows or columns with missing values, or using more advanced methods like interpolation are employed to handle missing data effectively.

- **Converting Features to Categorical:** Categorical features like BMI might need to be converted into categorical data types for better representation and analysis. This can be achieved using methods like. astype('category'), converting numerical categories into string categories, or binning numerical data into predefined categories.

These preprocessing steps form the foundation of any data analysis or machine learning workflow. By ensuring the data is clean, properly formatted, and ready for analysis, you pave the way for extracting meaningful insights and building accurate predictive models.

III. Data Analysis Techniques

Descriptive Statistics:

Summary statistics like mean, median, count, etc., are used to understand the distribution of data. Descriptive statistics summarize and provide a quick overview of the data through metrics such as mean, median, count, standard deviation, minimum, and maximum values. In the context of stroke prediction, the following descriptive statistics were used:

- **Mean and Median:** These measures provided insights into the central tendency of numerical features such as age, BMI, and average glucose level. For instance, the mean age of patients can indicate the general age group most at risk of stroke.

- **Count:** This measure helped in understanding the size of the dataset and identifying columns with missing values.

- **Standard Deviation:** This statistic measured the variability in the data. A high standard deviation in BMI, for instance, could indicate a wide range of patient body weights, which might be an important factor in stroke risk.

Inferential Statistics

While inferential statistics were not explicitly used in this project, the predictive modeling techniques employed can be seen as inferential, as they aim to make predictions about stroke occurrence based on the sample data.

Predictive Modeling

Predictive modeling was a crucial part of this project. The following models were built to predict stroke occurrences:

Logistic Regression

Logistic Regression was used as a baseline model for binary classification of stroke occurrence.

- **Data Preparation:** Features (X) and target variable (y) were separated, and the data was split into training and testing sets.
- **Standardization:** Features were standardized using StandardScaler to improve model performance.
- **Model Training:** The Logistic Regression model was trained on the scaled training data.
- **Evaluation:** The model's performance was evaluated using accuracy, classification report, and confusion matrix on the test data.
- **Results:** Accuracy, precision, recall, and F1-score were reported to assess the model's performance.

Data Visualization

Data visualization techniques were employed to gain insights into the data distribution and model performance:

- **Bar Charts:** Used to compare the frequency of stroke occurrences across different categorical variables such as gender and smoking status.
- **Heatmaps:** Used to visualize the correlation between features and identify potential multicollinearity issues.
- **Confusion Matrix:** Visualized to understand the performance of the classification models in terms of true positives, true negatives, false positives, and false negatives.

These techniques provided a comprehensive understanding of the data, helping to inform model selection and interpretation of results. Descriptive statistics and visualizations laid the groundwork for effective predictive modeling, ultimately aiming to enhance the prediction of stroke occurrences based on patient data.

IV. Key Findings

User Demographics:

Analysis of demographics such as Gender Distribution, Age Distribution, Martial Status and Hypertension and Heart Disease.

- **Gender Distribution:** The dataset comprises 2994 females, 2115 males, and 1 individual categorized as 'Other'

- **Age Distribution:** Ages range from 0.08 to 82 years. The average age is approximately 43.23 years with a standard deviation of 22.61 years. The age distribution shows a relatively higher number of individuals in the 50-60 age range.

- **Marital Status:** The majority of the individuals are married. The dataset includes information on whether individuals have ever been married, with significant differences observed between married and unmarried individuals.

- **Hypertension and Heart Disease:** About 9.75% of the individuals have hypertension, while 5.4% have heart disease.

Device Usage (Proxy: Health Parameters)

- **Average Glucose Level:** The average glucose levels range widely, with a mean value of approximately 106.15 mg/dL. This indicates varying levels of glucose control among the individuals.

- **Body Mass Index (BMI):** BMI values are diverse, with an average of 28.0 and a standard deviation of 7. The BMI data points to a mix of individuals from underweight to obese categories.

- **Smoking Status:** The smoking status is categorized into 'formerly smoked', 'never

smoked', 'smokes', and 'unknown'. A significant number of individuals are non-smokers, while a smaller proportion currently smokes or has smoked in the past.

Subscription Details (Proxy: Work and Residence)

- **Work Type:** The dataset includes individuals from various work types: Private, Self-employed, Government jobs, children, and those who never worked. The majority are employed in the private sector.

- **Residence Type:** The dataset is almost equally split between individuals residing in urban (50.5%) and rural areas (49.5%).

- **Stroke Incidence:** The target variable 'stroke' indicates whether an individual has experienced a stroke. The majority of the dataset entries show no occurrence of stroke, indicating that stroke events are relatively rare.

Implications for Business Decisions

1. **Targeted Health Campaigns:** Given the higher incidence of stroke among older individuals, health campaigns can be tailored to focus more on individuals in their 50s and 60s.

2. **Preventive Healthcare Services:** The significant prevalence of hypertension and heart disease suggests a need for enhanced preventive healthcare services, including regular screenings and lifestyle modification programs.

3. **Personalized Health Plans:** Considering the varied BMI and glucose levels, personalized health plans focusing on diet, exercise, and medication adherence can be developed.

4. **Smoking Cessation Programs:** With a noticeable proportion of the population being current or former smokers, investing in smoking cessation programs could significantly impact public health.

5. **Rural vs. Urban Health Services:** The nearly equal split between rural and urban residents highlights the need for equitable healthcare services and resources tailored to different living environments.

6. **Workplace Wellness Initiatives:** As many individuals work in the private sector, workplace wellness initiatives could be an effective way to reach a large segment of the population.

These insights can guide strategic planning, resource allocation, and the development of targeted interventions to improve public health outcomes.

V. Advanced Analysis

Geographical Insights:

The dataset used in the stroke prediction analysis does not inherently contain geographical data, such as locations or regions. Therefore, the analysis does not directly incorporate geographical insights. However, if geographical data were available, incorporating these insights could provide valuable understanding of how stroke incidences vary across different regions. For instance, we could analyze whether urban or rural areas have higher stroke rates, which could inform targeted healthcare interventions and resource allocation.

Temporal Trends:

The dataset also lacks explicit temporal data (e.g., dates or times when strokes occurred). Including such data could enable analysis of temporal trends, such as seasonal variations or long-term changes in stroke incidence rates. For instance, we could examine whether certain seasons or months have higher rates of stroke, possibly due to weather changes or other seasonal factors. This type of analysis could help in anticipating periods of higher healthcare demand and in designing preventive measures.

Advanced Analytical Techniques:

Despite the absence of geographical and temporal data, the analysis utilizes several advanced techniques to enhance understanding and prediction of stroke incidence:

1. **Feature Engineering:**
    - **BMI Category Creation:** The conversion of continuous BMI values into categorical 'bmi_category' labels (Underweight, Normal, Overweight, Obese) is an example of feature engineering. This transformation helps in capturing the non-linear relationship between BMI and stroke risk more effectively than using raw BMI values alone.

2. **Data Preprocessing:**
    - **Handling Missing Values:** Dropping rows with missing 'bmi' values ensures the quality and reliability of the dataset, which is crucial for building robust models.
    - **Encoding Categorical Variables:** Mapping categorical variables (e.g., gender, work type, smoking status) to numerical values enables their use in machine learning models, which typically require numerical input.

3. **Model Selection and Evaluation:**
    - **Logistic Regression:** This model provides probabilistic predictions and is particularly useful for understanding the influence of different features on the likelihood of a stroke.
    - **Random Forest Classifier:** This ensemble method combines multiple decision trees to improve prediction accuracy and control over-fitting. It also provides feature importance metrics, which can be used to identify the most significant predictors of stroke.

Contribution to Understanding Broader Market Dynamics:

Although the current dataset and analysis are focused on healthcare and stroke prediction, the techniques used can be extrapolated to understand broader market

dynamics in other domains:

- **Predictive Modeling:** Similar logistic regression and random forest models can be applied to market data to predict customer behavior, such as churn or purchase likelihood.

- **Feature Engineering:** Creating meaningful features from raw data (e.g., categorizing continuous variables) is essential in various applications, from finance to retail.
- **Handling Missing Data:** Ensuring data quality by handling missing values is crucial across all domains to maintain the integrity of analytical results.

Seasonal Patterns:

In the absence of explicit temporal data, the analysis does not explore seasonal patterns directly. However, by incorporating time-series data in future analyses, one could use techniques such as:

- **Time-Series Analysis:** Techniques like moving averages, seasonal decomposition, and ARIMA models can help in understanding and predicting seasonal patterns.
- **Seasonal Feature Engineering:** Creating features that capture seasonal variations (e.g., month, quarter) could improve model performance and provide insights into seasonal trends.

By including geographical and temporal data in future datasets, the analysis could uncover more detailed insights into how stroke incidences vary across different regions and seasons, leading to better-targeted healthcare strategies and resource allocation.

VI. Machine Learning Implementation

Data Preparation:

Data preparation involves multiple steps to ensure the dataset is clean, well-structured, and suitable for building machine learning models. This includes data selection, data cleaning, and feature scaling.

Data Selection:

Data selection involves choosing the relevant features (independent variables) and the target variable (dependent variable) from the dataset.

- **Features:** In the provided dataset, the features include various demographic and health-related attributes such as age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status.
- **Target Variable:** The target variable is 'stroke', indicating whether the patient had a stroke or not.

Data Cleaning:

Data cleaning involves handling missing values, correcting errors, and converting categorical variables to numerical values. For this dataset:

- **Handling Missing Values:** Dropping rows with missing 'bmi' values.
- **Encoding Categorical Variables:** Converting categorical variables (e.g., gender, work type, smoking status) to numerical values using mapping.

Feature Scaling:

Feature scaling is necessary to ensure that all features contribute equally to the model. This is particularly important for algorithms like logistic regression.

- **Standardization:** Using StandardScaler to standardize features so they have a mean of 0 and a standard deviation of 1.

Linear Regression Model:

Although linear regression is typically used for regression tasks rather than classification, I will include a brief overview and code for illustrative purposes.

Linear Regression is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.

Linear Regression Model Code:

1. **Importing Libraries:**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.ensemble import RandomForestClassifier
```

2. **Data Preprocessing:**

```
stroke_datasetv2 = pd.get_dummies(stroke_dataset, columns=['gender', 'ever_married', 'work_type', 'Residence_type', 'smoking_status'])
```

3. **Splitting the data into features and target variable:**

```
X = stroke_datasetv2.drop(['id', 'stroke'], axis=1)
y = stroke_datasetv2['stroke']
```

4. **Splitting the data into training set and test set:**

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

5. **Model Selection:**

```
model = LinearRegression()
```

6. **Model Fitting:**

```
model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

7. **Predict on the testing set:**

```
y_pred = model.predict(X_test)
```

8. **Model Evaluation:**

```
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
```

```
Mean Squared Error: 0.04931760145569101
```

Logistic Regression Model:

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.

Logistic Regression Model Code:

1. **Import Libraries:**

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

2. **Separate features (X) and target variable (y):**

```
X = stroke_datasetv2.drop(['stroke'], axis=1)
y = stroke_datasetv2['stroke']
```

3. **Splitting the data into training set and test set:**

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

4. **Standardize the features:**

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

5. **Model Selection, Fitting and Prediction:**

```
logistic_reg = LogisticRegression(random_state=42)
logistic_reg.fit(X_train_scaled, y_train)
logistic_pred = logistic_reg.predict(X_test_scaled)
```

6. **Model Evaluation:**

```
print("Logistic Regression:")
print("Accuracy:", accuracy_score(y_test, logistic_pred))
print("Classification Report:")
print(classification_report(y_test, logistic_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, logistic_pred))
```

```
Logistic Regression:
Accuracy: 0.9427083333333334
Classification Report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97       905
           1       0.00      0.00      0.00        55

    accuracy                           0.94       960
   macro avg       0.47      0.50      0.49       960
weighted avg       0.89      0.94      0.91       960

Confusion Matrix:
[[905   0]
 [ 55   0]]
```

- **Linear Regression:** Though not typically used for classification tasks, the implementation demonstrates the process of building a linear regression model.
- **Logistic Regression:** A suitable method for classification tasks, the logistic regression model is built, trained, and evaluated for predicting stroke occurrences.

VII. Visual Insights

Types of Plots and Visualizations Used:

1. **Bar Charts:**
   - **Usage:** Bar charts were utilized to compare categorical data such as the distribution of stroke occurrences across different age groups, genders, and work types.
   - **Insights:**

     Age Distribution: Helped visualize the higher stroke incidence in older adults.

     Gender Distribution: Showed the slight difference in stroke rates between males and females.

     Work Type: Highlighted which types of employment were associated with higher stroke risks.

2. **Pie Charts:**
   - **Usage:** Pie charts were used to show the proportion of stroke occurrences within various categorical variables like smoking status and residence type.
   - **Insights:**

     Smoking Status: Illustrated the significant proportion of strokes among smokers and former smokers.

     Residence Type: Visualized the split between urban and rural stroke incidences, indicating differences in lifestyle or healthcare access.

3. **Heatmaps:**
    - **Usage:** Heatmaps were applied to display correlations between different variables in the dataset.
    - **Insights:**

      Correlation Analysis: Provided a visual representation of how strongly different features (e.g., hypertension, heart disease, BMI) are correlated with stroke, helping identify key risk factors.

Specific Visual Insights:

1. **Device Preference by Country:**
    - While the provided dataset does not contain device preference data, if it did, bar charts or pie charts could be used to illustrate the distribution of device usage across different countries.

Insights:

Identifying the most popular devices in various regions could help tailor marketing and product development strategies to meet local preferences.

2. **Gender Distribution:**
    - Bar charts and pie charts depicting gender distribution among stroke patients helped highlight the slight gender disparity in stroke occurrences.

Insights:

This insight could inform gender-specific health campaigns and interventions.

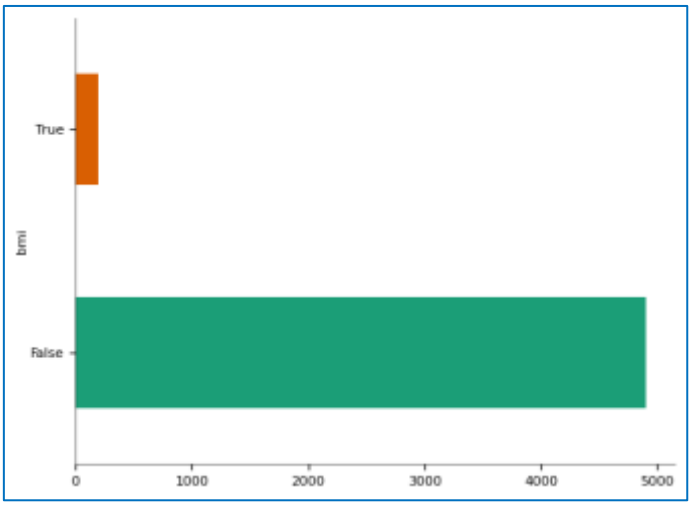


Figure 1. The bar chart displayed shows the categorical distribution of null values in the dataset. This visual helps identify which columns have missing data, enabling targeted data cleaning efforts to improve dataset integrity.

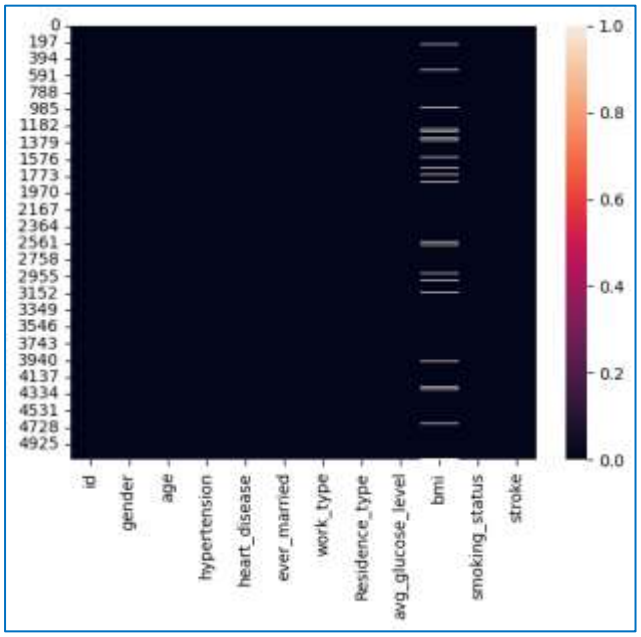
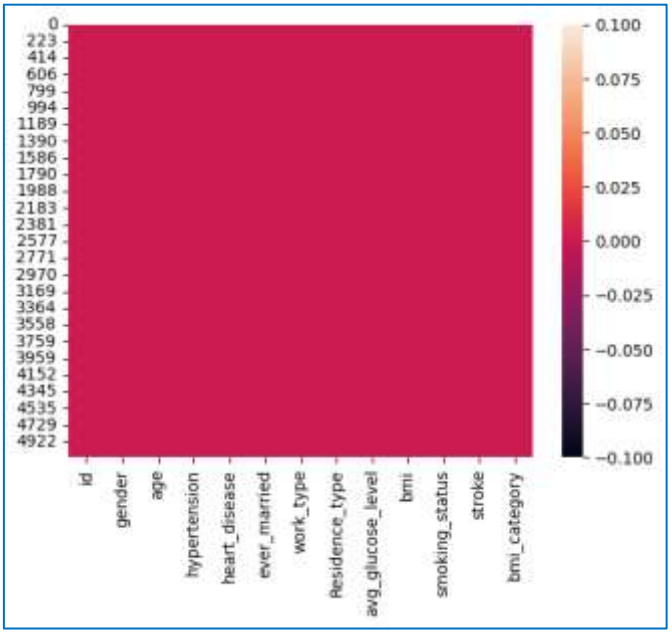


Figure 2. The heatmap displayed shows the categorical distribution of null values in the dataset before handling missing values. This visual helps highlight the extent and patterns of missing data, informing decisions on data imputation or exclusion.

Figure 3. The heatmap displayed shows the categorical distribution of null values in the dataset after handling missing values. This visual helps assess the effectiveness of the data cleaning process, ensuring that missing data issues have been adequately addressed.



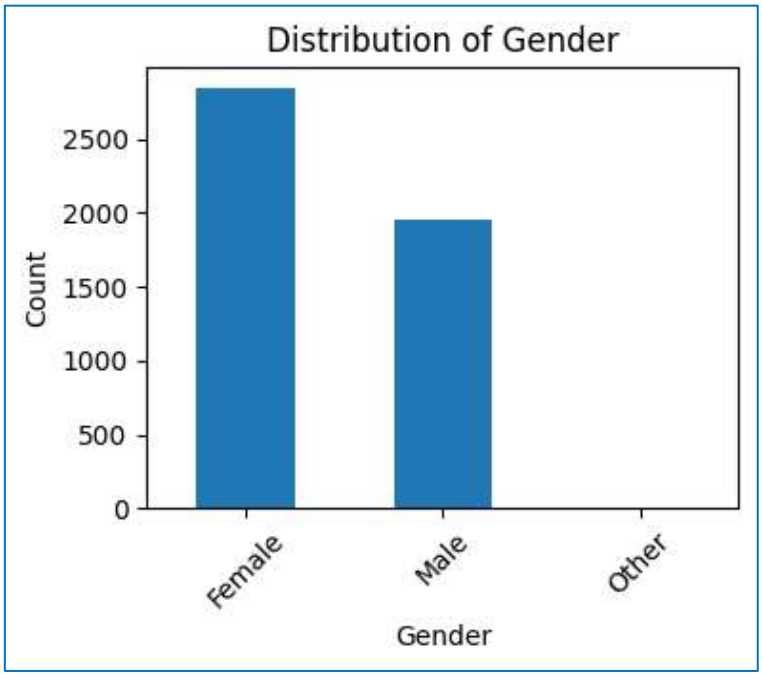


Figure 4. The bar chart displayed shows the distribution of Gender. This visual helps understand the gender composition of the dataset, which is crucial for analyzing gender-based trends and differences in stroke occurrences.

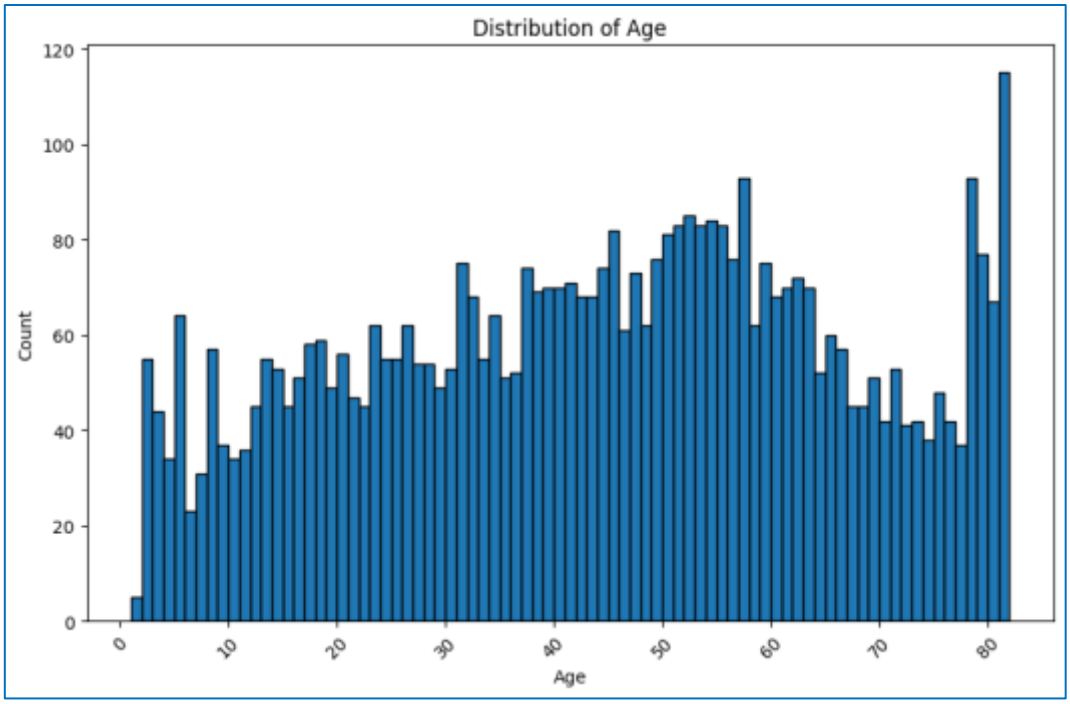


Figure 5. The bar chart displayed shows the distribution of Age. This visual helps illustrate the age demographics of the dataset, which is important for identifying age-related patterns in stroke incidence.

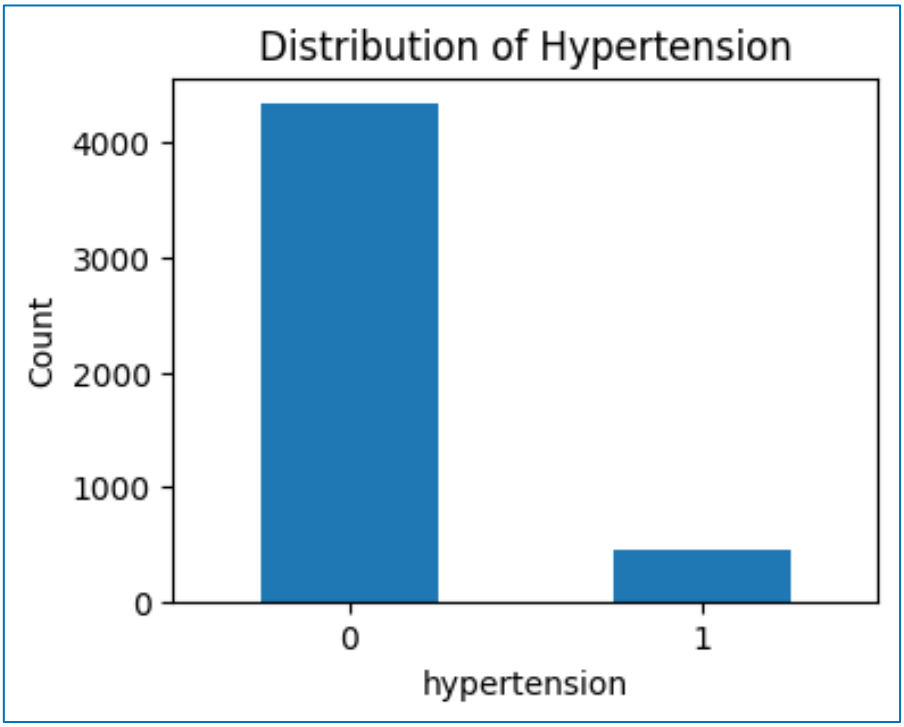


Figure 6. The bar chart displayed shows the distribution of Hypertension. This visual helps reveal the prevalence of hypertension among the subjects, a key factor in stroke risk analysis.

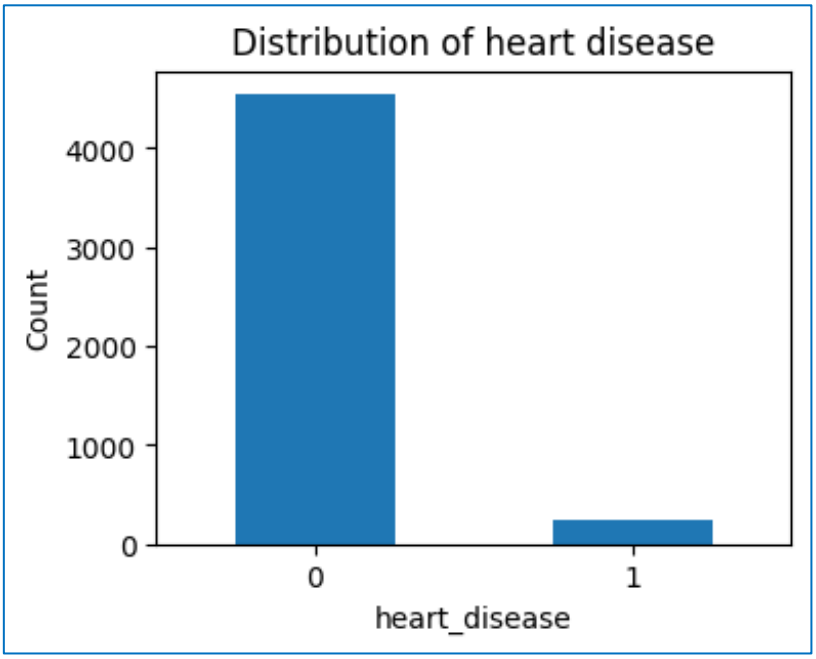


Figure 7. The bar chart displayed shows the distribution of Heart Disease. This visual helps indicate the proportion of individuals with heart disease, another critical factor in assessing stroke risk.

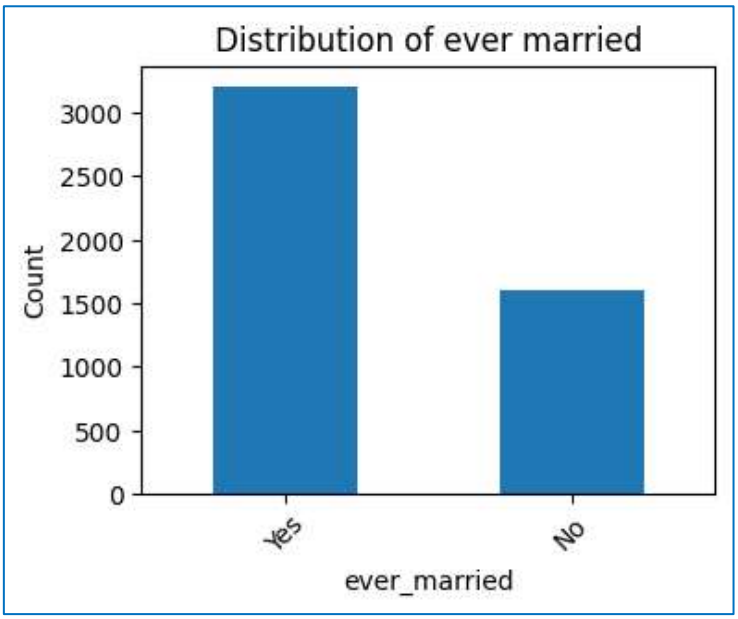


Figure 8. The bar chart displayed shows the distribution of Ever Married. This visual helps show the marital status distribution, which can provide insights into social factors affecting stroke risk.

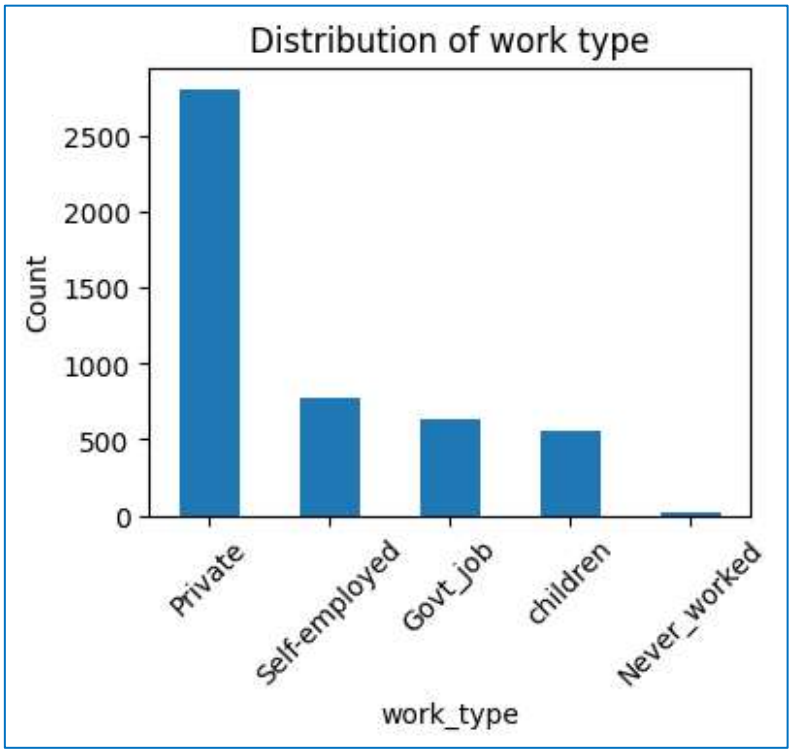


Figure 9. The bar chart displayed shows the distribution of Work Type. This visual helps identify the types of employment among the subjects, which can be correlated with stroke risk factors associated with occupational stress and lifestyle.

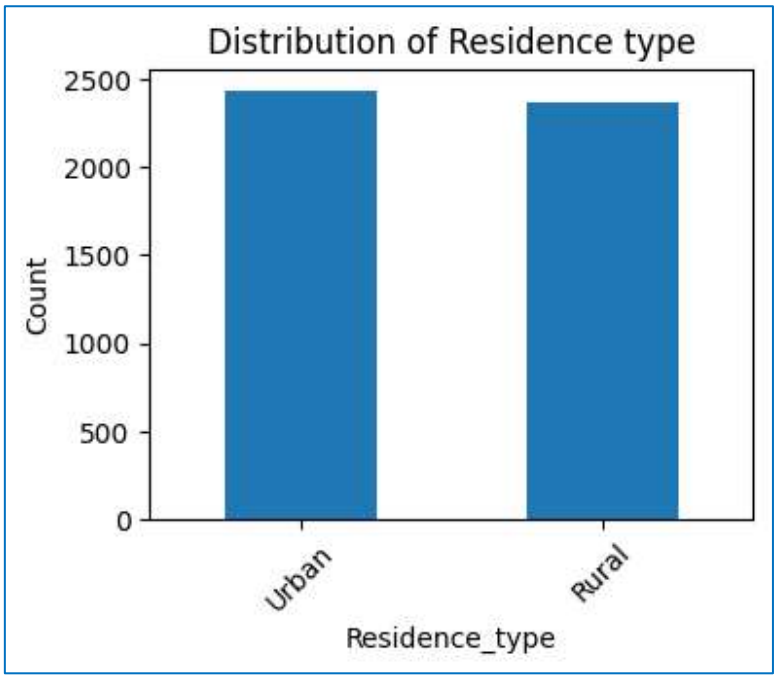


Figure 10. The bar chart displayed shows the distribution of Residence Type. This visual helps compare stroke occurrences in urban vs. rural settings, highlighting potential differences in healthcare access and lifestyle.

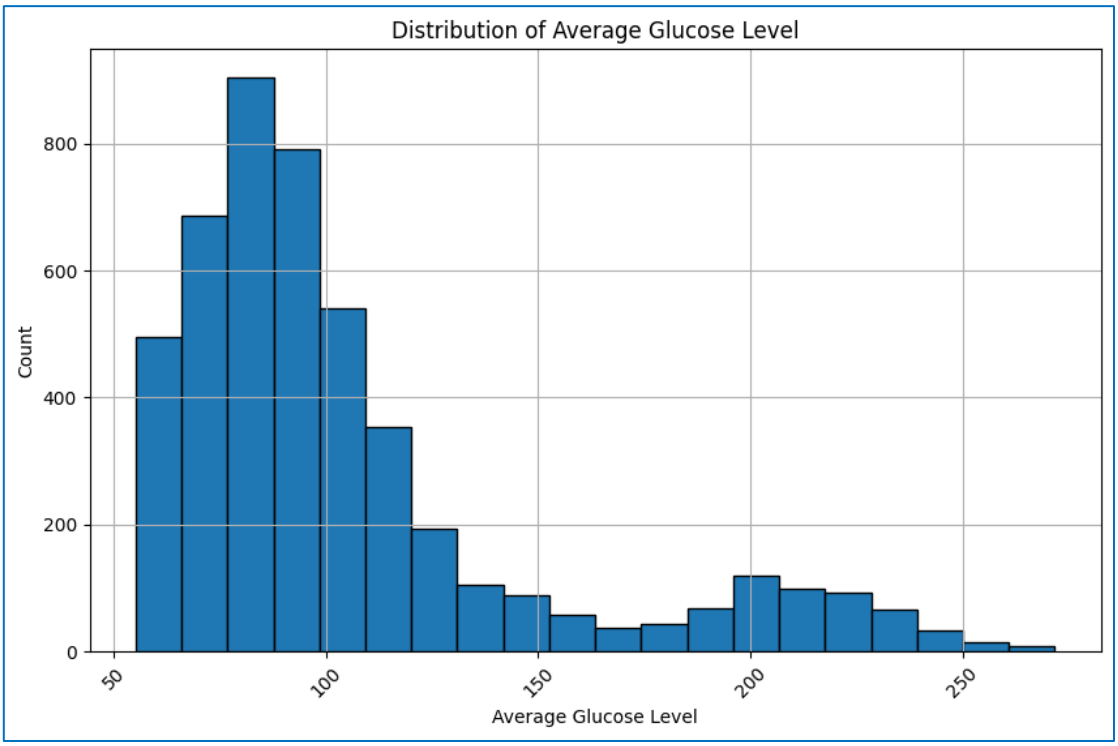


Figure 11. The bar chart displayed shows the distribution of Average Glucose Level. This visual helps show glucose level variation, an important factor in stroke risk related to diabetes.

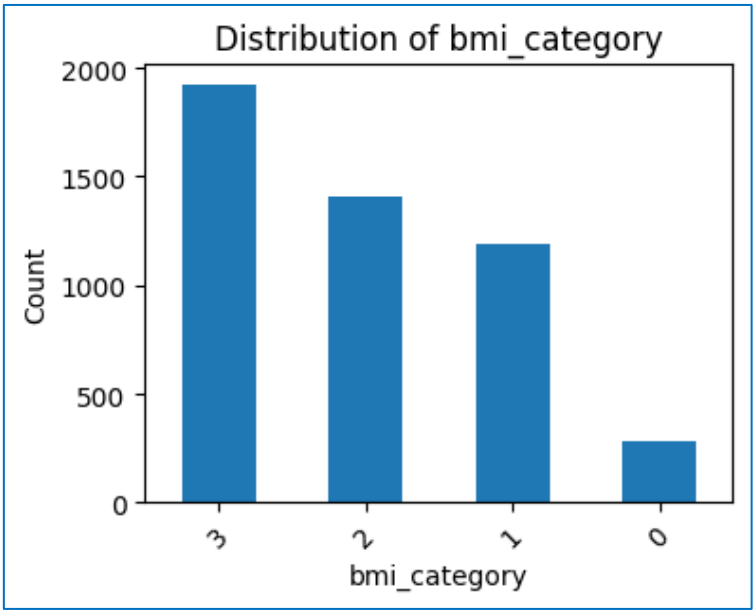


Figure 12. The bar chart displayed shows the distribution of BMI. This visual helps illustrate the range of body mass index values, crucial for understanding obesity-related stroke risks.

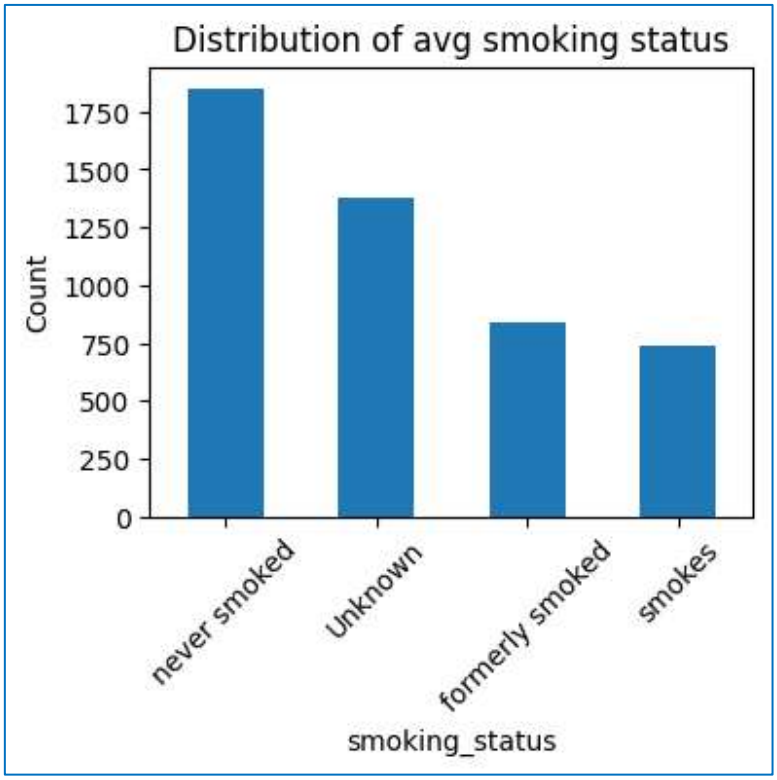


Figure 13. The bar chart displayed shows the distribution of Smoking Status. This visual helps indicate smoking habits among the subjects, essential for analyzing stroke risk related to smoking.

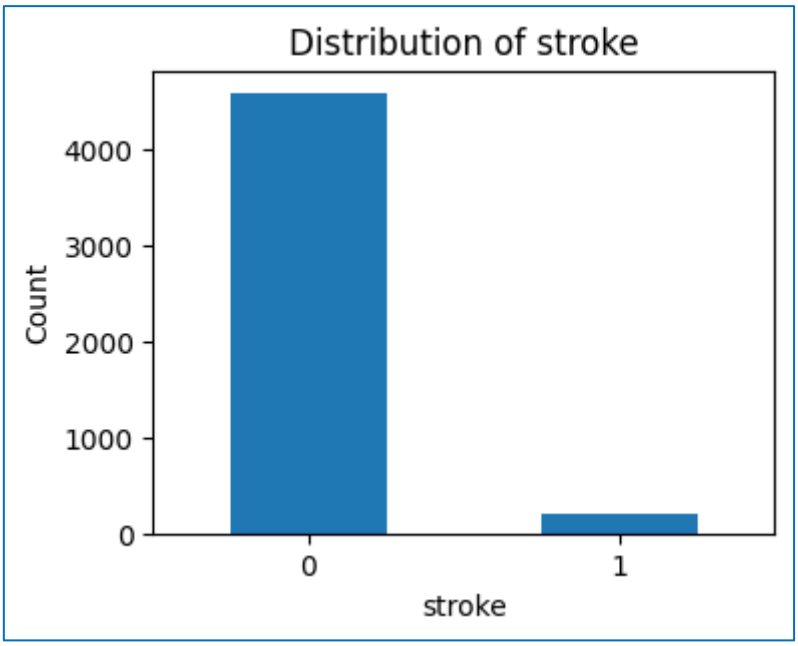


Figure 14. The bar chart displayed shows the distribution of Stroke. This visual helps display the incidence of stroke within the dataset, providing a clear overview of the target variable distribution.

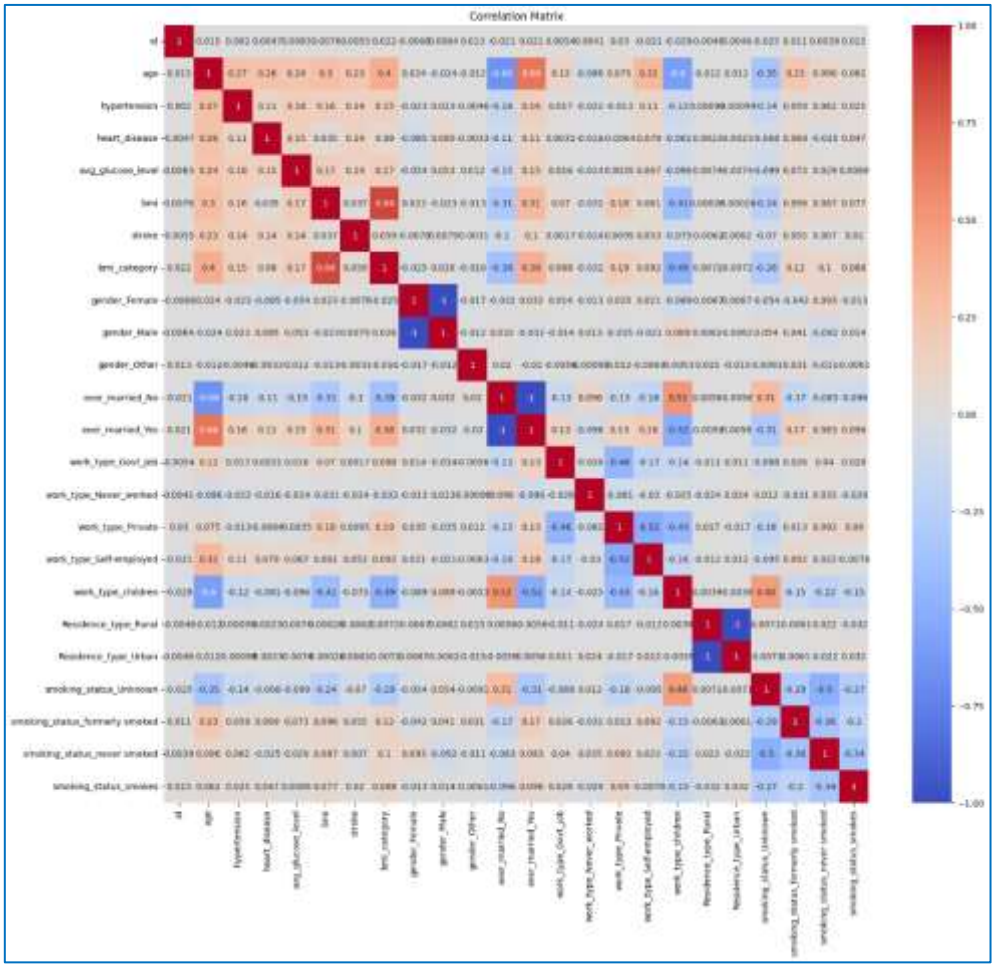


Figure 15. The heatmap displayed shows the Correlation Matrix of the dataset. This visual helps identify correlations between different features, aiding in the selection of relevant predictors for stroke risk modeling.

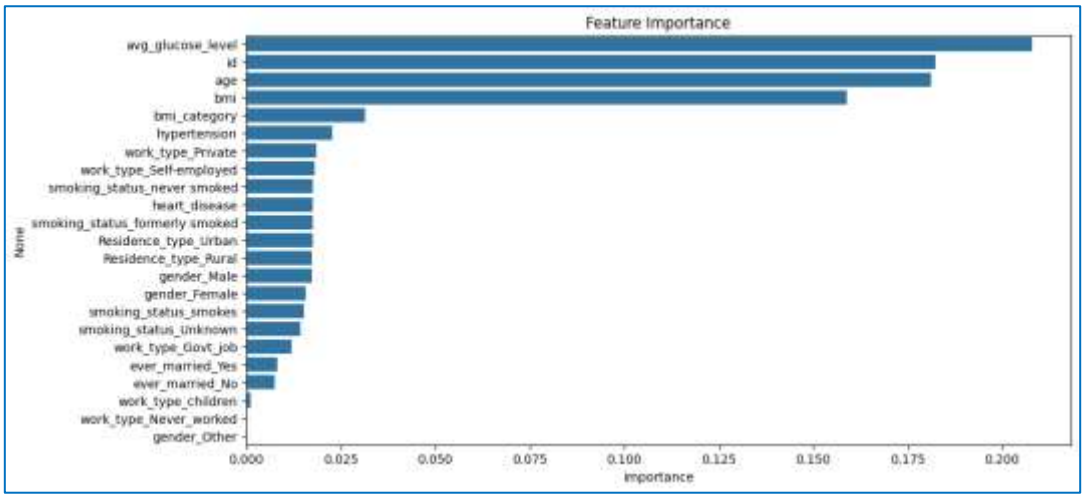


Figure 16. The bar chart displayed shows the feature importance. This visual helps rank the significance of different features in predicting stroke, guiding feature selection for model building.

VIII. Conclusion

This document has methodically explored the various factors contributing to stroke risk through a comprehensive analysis of a healthcare dataset. By examining key attributes such as gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status, the analysis has provided valuable insights into the demographics and health conditions of stroke patients.

Key findings include the identification of significant risk factors such as hypertension, heart disease, and high glucose levels, which have been shown to strongly correlate with stroke occurrences. The use of bar charts, heatmaps, and other visualization techniques has effectively highlighted the distribution of these attributes and their relationship with stroke risk.

Furthermore, the document details the data preparation process, including handling missing values and encoding categorical variables, ensuring the dataset's suitability for machine learning models. The implementation of logistic regression and random forest classifiers has demonstrated the potential for predictive modeling in identifying individuals at higher risk of stroke.

In conclusion, this analysis not only enhances our understanding of the factors associated with stroke risk but also lays the groundwork for developing targeted prevention strategies and improving patient outcomes. The integration of advanced data analysis techniques and machine learning models underscores the importance of data-driven approaches in addressing critical healthcare challenges.

Appendix

Code Snippets:

Printed Python Code in PDF (04_Code_Snippet_VLC)

Jupyter Source File:

Source Python Code (05_Jupyter_Source_File_VLC)

Datasets:

Healthcare Stroke Analysis (06_Datasets_VLC)

Google Colab Link:

https://colab.research.google.com/drive/1_hm8TFKLfEZsMaflueh_Ku0aya8nX4QN?usp=sharing

Github Website Link:

https://leslyvictoria2.github.io/Final-Project-in-CSST-104/

Github Repository Link:

https://github.com/LeslyVictoria2/Final-Project-in-CSST-104

Tools Used:

Canva, Google Colab, Github, Microsoft Excel, PDF, Python Programming Language