Stroke Prediction Analysis

PROJECT REPORT

Venkata Krishnan R
ravichandran.ve@northeastern.edu

CONTENTS

Project Proposal Guidelines

1. Problem Setting:

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. A lot of factors contribute to the risk of a heart stroke. Leading risk factors for heart disease and stroke are high blood pressure, high low-density lipoprotein (LDL) cholesterol, diabetes, smoking and secondhand smoke exposure, obesity, unhealthy diet, and physical inactivity. The problem setting for this project is to develop a machine learning model to predict the likelihood of a patient experiencing a stroke based on demographic and health information such as gender, age, pre-existing medical conditions, and smoking status. The goal is to use this model to identify high-risk patients and potentially prevent or mitigate the impact of strokes, as they are a leading cause of death worldwide. The project involves utilizing a dataset of patient information and conducting exploratory data analysis, feature selection, and testing various model types including tree-based and neural network models.

2. Problem Definition:

Through a series of observations, it is found that a lot of common risk factors lead to heart stroke. Through our given dataset, the common risk factors are used as predictors and examined. The Machine Learning model is fed with predictors and respective outcome variables. The result is a binary classification for the outcome variable. Based on the series of observations, the model can predict and classify a patient in the future, if the person is likely to have a heart stroke.

3. Data Sources:

Kaggle is used as the Data Source. **https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.**

4. Data Description:

With over 5000 rows and 12 columns, the data describes the common risk factors of a heart stroke. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like id, gender age, hypertension, history of heart diseases, ever married before, kind of work, the type of residence the person lives in, the average glucose level, bmi, smoking status and an outcome variable if the person has had a stroke.

Below are the attributes of the Stroke dataset which explains the numerical and categorical variables in detail:
The numerical variables are:

- id: unique identifier (numeric)
- age: age of the patient (numeric)
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension (binary)
- heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease (binary)
- avg_glucose_level: average glucose level in blood (numeric)
- bmi: body mass index (numeric)
- stroke: 1 if the patient had a stroke(binary)

The categorical variables are:

- gender: "Male", "Female" or "Other" (categorical)
- ever_married: "No" or "Yes" (binary)
- work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed" (categorical)
- Residence_type: "Rural" or "Urban" (binary)
- smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown" (categorical)

The attributes can be used as predictor variables in our machine learning project. They can be used to build models to predict if a patient has had a stroke or not, based on the other information available about the patient such as age, gender, hypertension, heart disease, ever married status, work type, residence type, average glucose level, BMI, and smoking status.

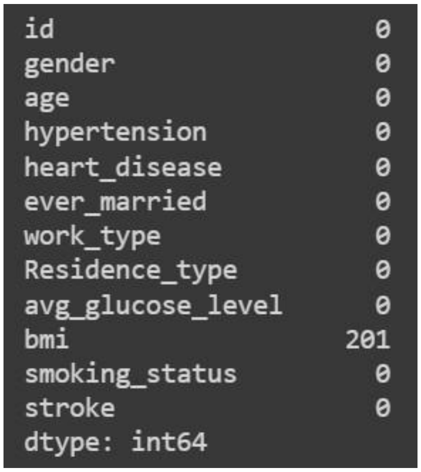
5. Project Motive:

Every year, approximately 800,000 individuals in the United States are afflicted with a stroke, with three out of four being first-time occurrences. The encouraging news is that 80% of these strokes could have been prevented through appropriate education about the signs of a stroke.

This study aims to develop a prediction model for stroke and evaluate the model's accuracy. We will explore seven different models: Decision Tree, Logistic Regression, Random Forest, Support Vector Machine, K Nearest Neighbour, Naive Bayes, and KMeans Clustering. Our goal is to find which model yields dependable and repeatable outcomes. Once we have the prediction results from these models, we will compare their performance to determine the most effective model. We will also subject the top-performing model to a cross-validation process to evaluate its repeatability. This analysis will help to identify the most accurate and reliable model.

6. Data Mining Tasks:

On the first hand, when I uploaded the dataset and started with the cleaning process, the dataset was relatively clean and easier to comprehend, however we can the number of null values in the dataset below:



```
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                201
smoking_status       0
stroke               0
dtype: int64
```

Figure 1

Below is the graphical representation of the null values in "bmi column:

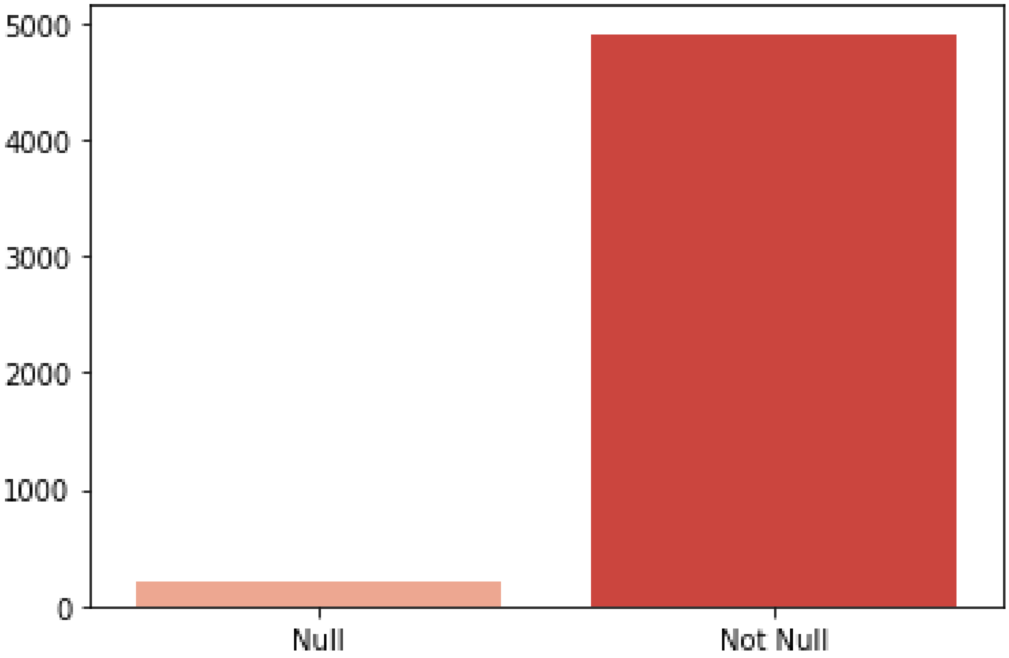


Figure 2: Graphical representation of BMI Column

Next, we decided to plot a histplot to understand the distribution of values in "bmi" column:

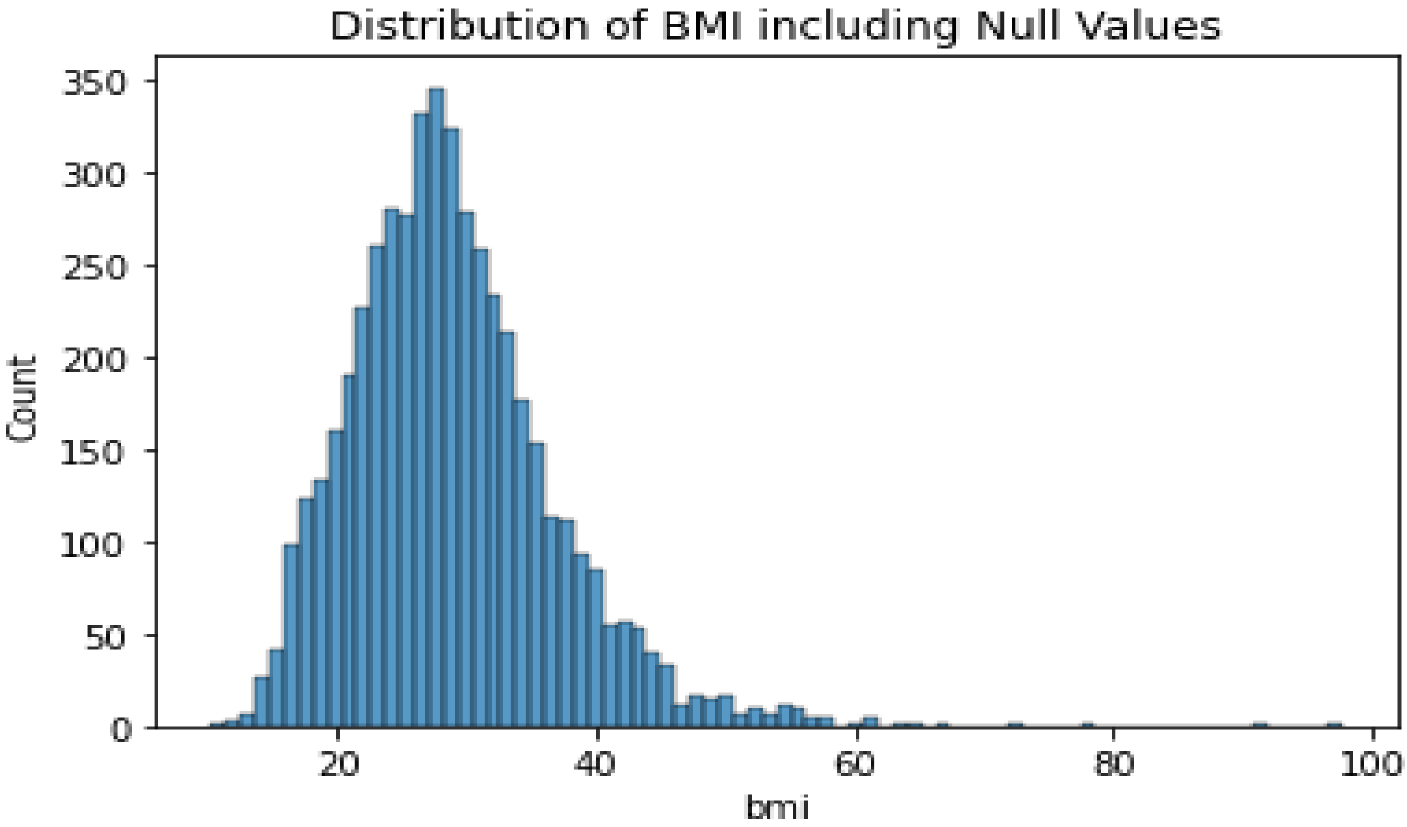


Figure 3

From this distribution we can interpret that, there are 201 missing values for the BMI variable. However, since there are 40 patients who had a stroke that have a missing BMI value, we cannot simply drop these missing values without losing valuable data. In order to continue with the analysis, we must impute values for BMI.

After analyzing the distribution of the BMI variable, we have decided to use the median value to replace the missing values. This decision was made because BMI has a long right tail, and the median is a better representation of central tendency in such cases. Using the mean could potentially skew the data, as the mean can be significantly affected by outliers in the data.

Therefore, we have used the following code to replace the missing values with the median value for BMI.

This code finds all the NaN values in the 'bmi' column of the dataframe 'df' and replaces them with the median value of the 'bmi' column. Using the 'inplace' parameter with a value of 'True' ensures that the changes are made to the original dataframe.

By imputing values for the missing BMI values, we can retain valuable information in the dataset, which is necessary for further analysis and prediction modeling. Now, that we have accounted for the null values in the dataset we decided to add an extra column for data analysis and for plotting interactive visualizations.

Since there’s a column “avg_glucose_level” that mentions the average glucose level of our dataset.

The labels for glucose levels can vary depending on the specific reference range used by healthcare providers and organizations. However, in general, the following labels are often used:

1. Low Glucose: Less than 70 mg/dL
2. Normal Glucose: 70 to 99 mg/dL
3. Medium/Borderline High Glucose: 100 to 125 mg/dL
4. High Glucose: 126 mg/dL or higher (after two separate tests)

We first defined the glucose level ranges: low sugar (< 70), medium sugar (between 70 and 140), and high sugar (>= 140). We then add a new column 'sugar_range' to the existing dataset 'df' with an initial empty string. Using the loc accessor, we set the values in the 'sugar_range' column to 'Low Sugar', 'Medium Sugar', or 'High Sugar' based on the glucose level ranges.

For further analysis and deep dive into the Strokes dataset, we decided to use Label encoder for categorical variables into numerical format so that machine learning models can better understand and utilize the data. The code replaces string values in each column with corresponding numerical values using the replace() function.

In this specific dataset, the code is used to encode five categorical variables: gender, ever_married, Residence_type, smoking_status, and work_type. The mapping of the string values to numerical values is as follows:

- Gender: Male (1), Female (0)
- Ever_married: No (0), Yes (1)
- Residence_type: Rural (0), Urban (1)
- Smoking_status: Unknown (0), never smoked (1), formerly smoked (2), smokes (3)
- Work_type: Private (0), Self-employed (1), children (2), Govt_job (3), Never_worked (4)

By converting categorical variables into numerical format, we can utilize them as features in machine learning models. This can lead to more accurate models and better insights into the data.

Lastly we decided to drop the column “id” from the strokes dataset as it was not required for data analysis and machine learning purposes.

7. Data Exploration

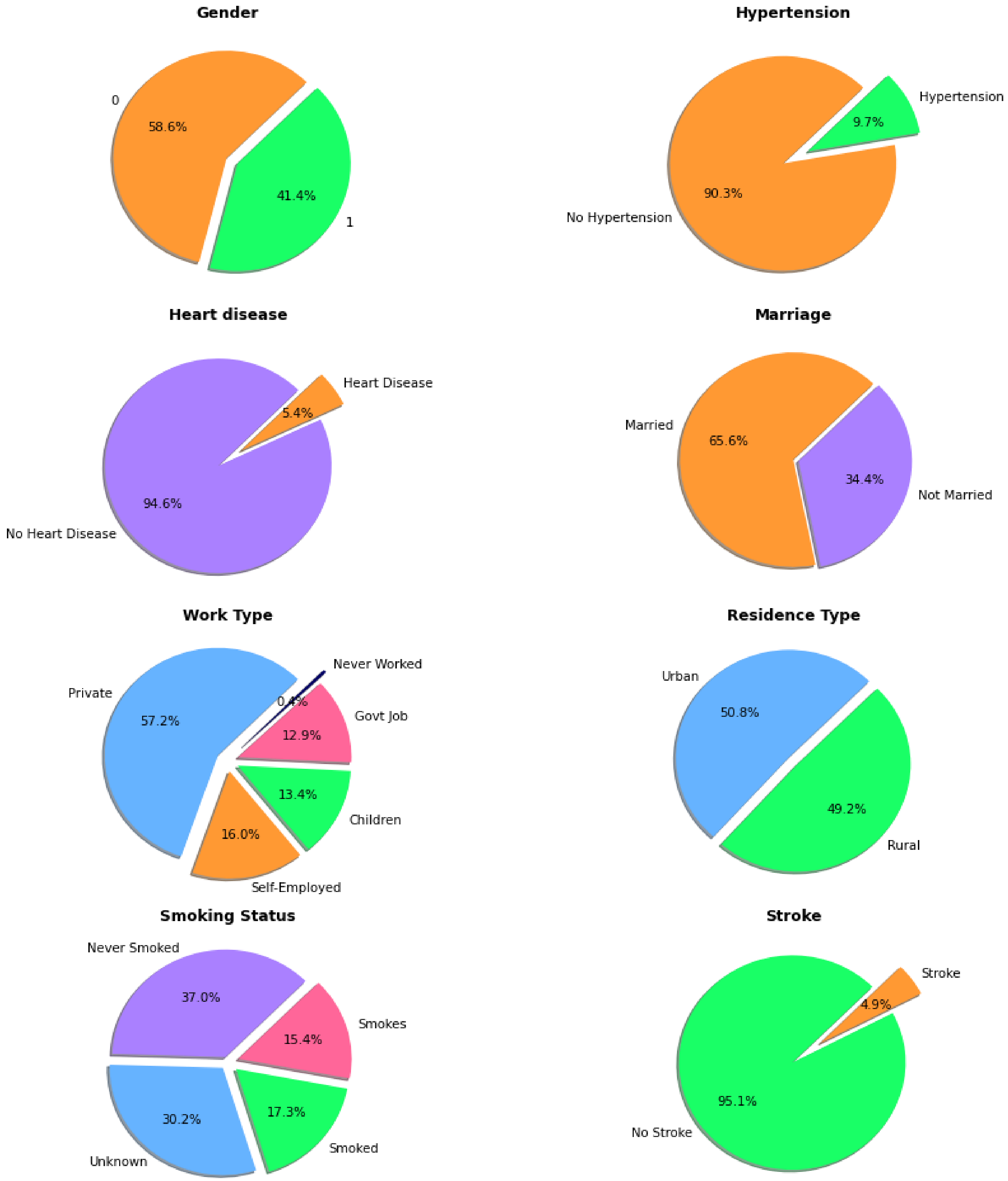


Figure 4

Pie Charts: To better understand the dataset, we created pie charts to visualize the percentage of each variable in the stroke dataset. The following are the observations that can be made from the pie charts:

- Hypertension: The majority (59.4%) of stroke patients in the dataset had previously been diagnosed with hypertension, indicating that hypertension is a significant risk factor for stroke.
- Heart disease: About 11.6% of stroke patients had a history of heart disease, which is also a significant risk factor for stroke.
- Ever married: The pie chart shows that 67.5% of stroke patients were ever married, indicating that being married may be a risk factor for stroke.
- Work type: The pie chart indicates that subjects with work experience and in government-related work (32.9%) have a higher risk of experiencing stroke, while those with no work experience (0.7%) barely experienced stroke.
- Residence type: The pie chart shows that there is no significant relationship between the residence type and the likelihood of experiencing a stroke.
- Smoking status: Being a smoker or former smoker increases the risk of having a stroke, with most stroke patients being either a smoker or former smoker.

Conclusion: From the pie charts, we can see that hypertension, heart disease, and smoking status are significant risk factors for stroke. Additionally, being married and working in government-related work may also increase the risk of experiencing a stroke. By understanding these risk factors, healthcare professionals can work to prevent and manage strokes, ultimately improving patient outcomes.

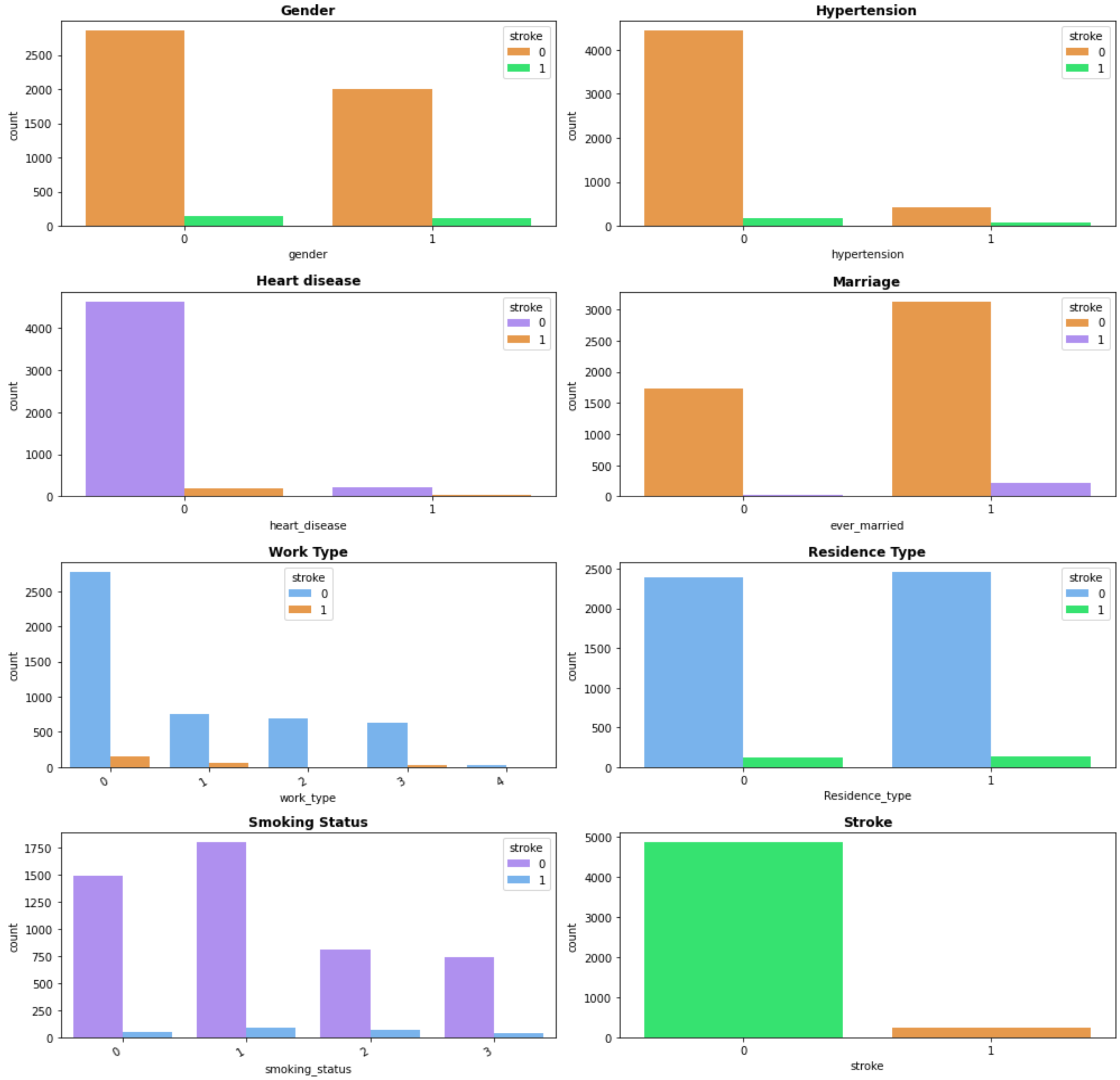


Figure 5

The count plot analysis indicates that several variables in the stroke dataset are related to the likelihood of experiencing a stroke. Individuals who have previously been diagnosed with hypertension or heart disease are at a higher risk of having a stroke. Also, those who are married and have work experience, particularly in government-related jobs, are more likely to experience a stroke. However, no apparent relationship was observed between the likelihood of stroke and the individual's residence type. Smoking, as well as former smoking, also increases the risk of having a stroke.

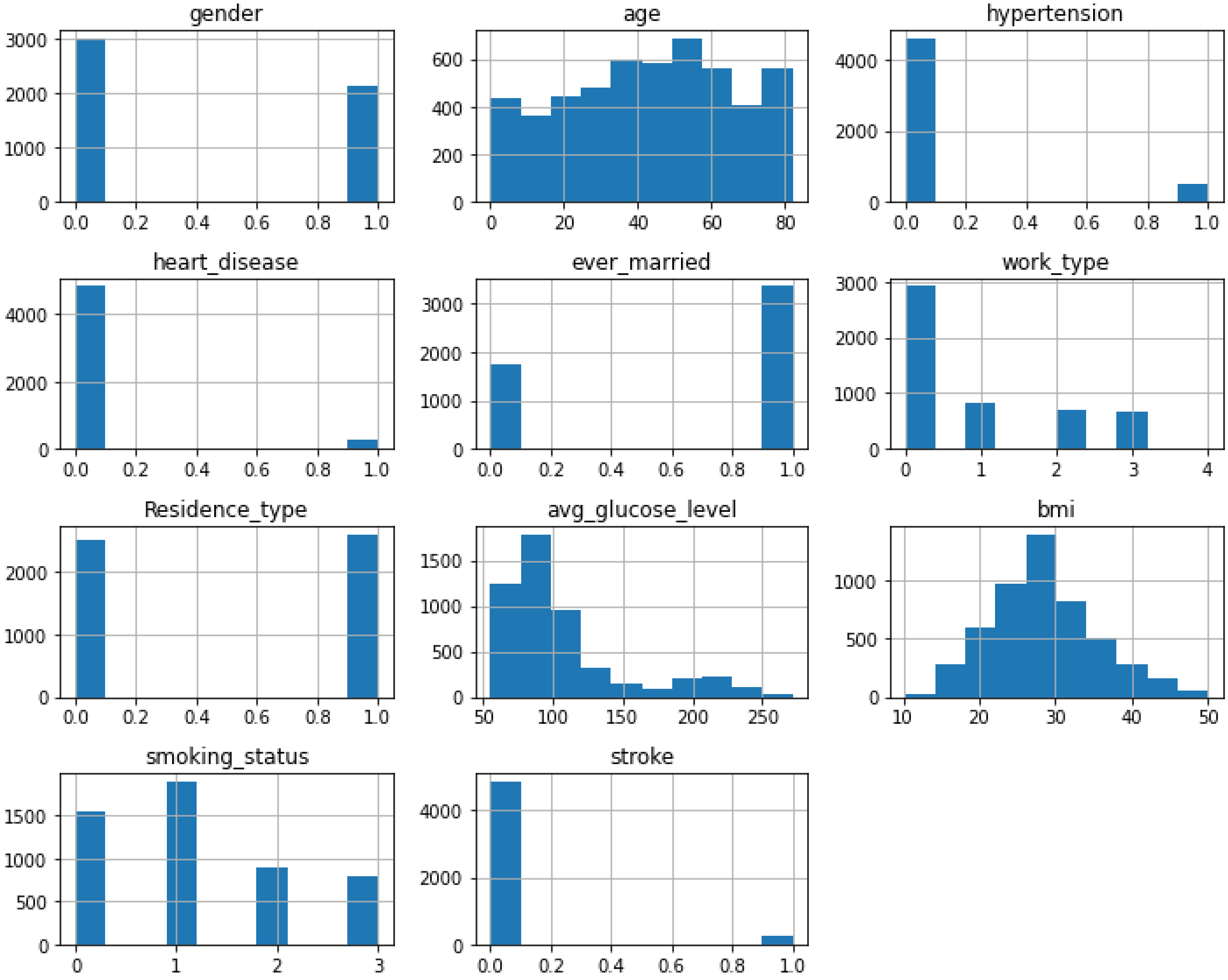


Figure 6

The above Histplot represents the distribution of the attributes in strokes dataset.

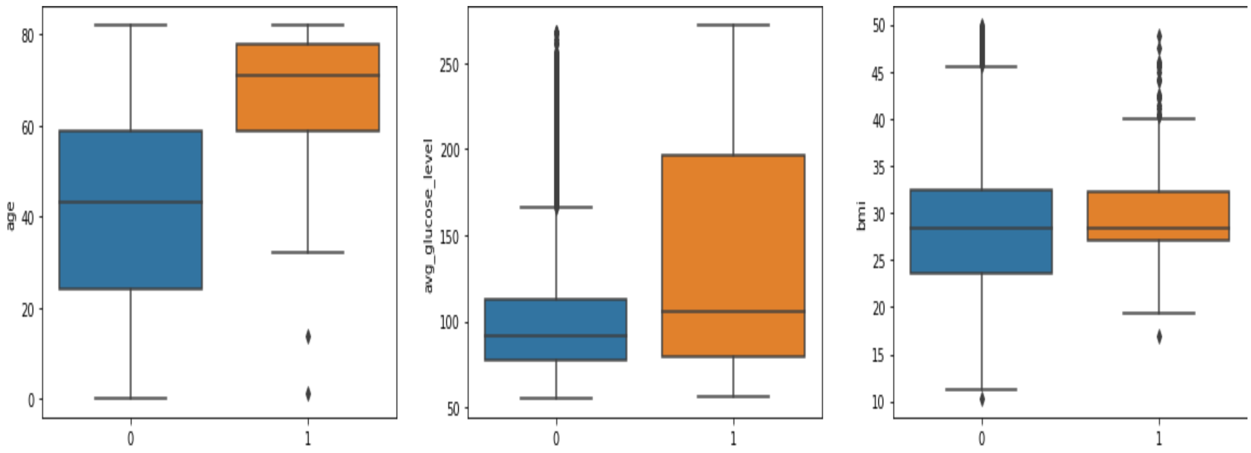


Figure 7

From the above boxplot, some observations can be drawn:

- age: Subjects with stroke tends to have higher mean age.
- ave glucose level: Subjects with stroke tends to have higher average glucose level.
- bmi: bmi index does not give much indication on the likelihood of experiencing stroke. bmi index for super obesity is 50. Outliers in this feature should be replaced to its highest limit (50).

There's a total of 79 counts of outliers detected.

Interesting Visualizations:

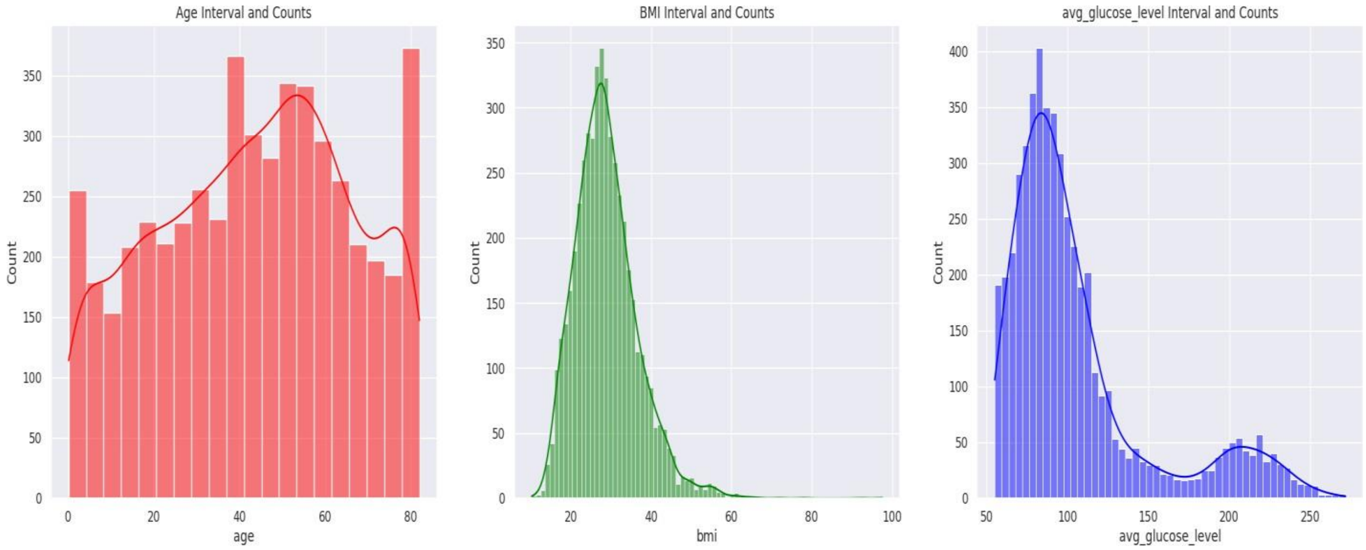


Figure 8

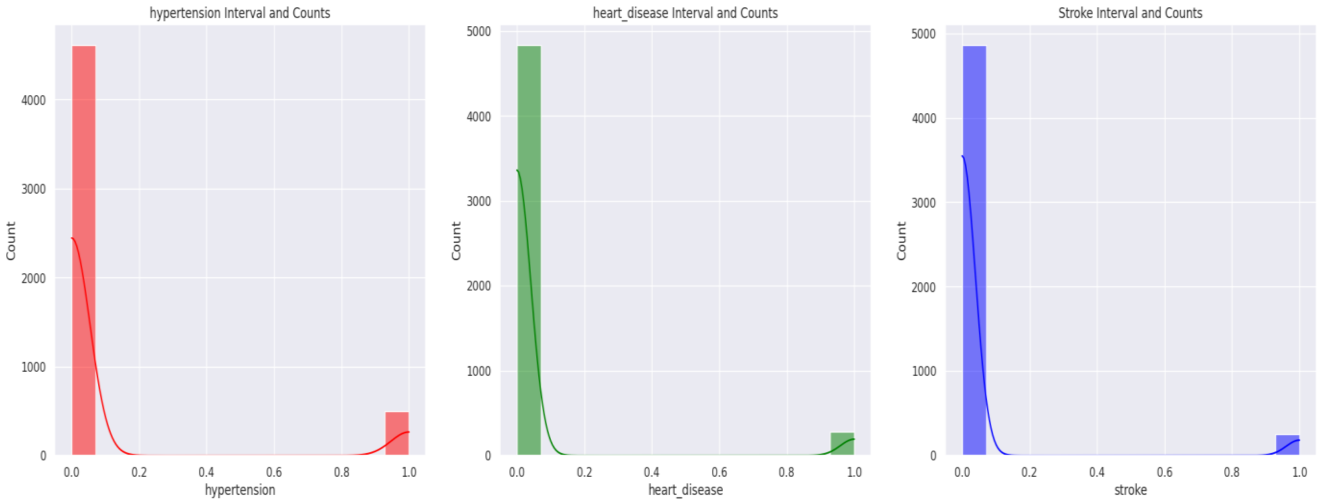


Figure 9

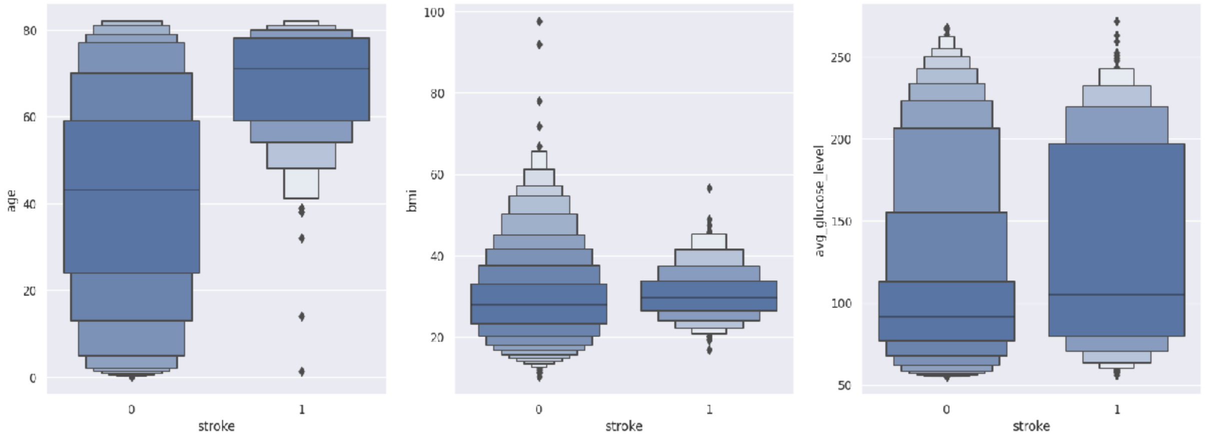


Figure 10

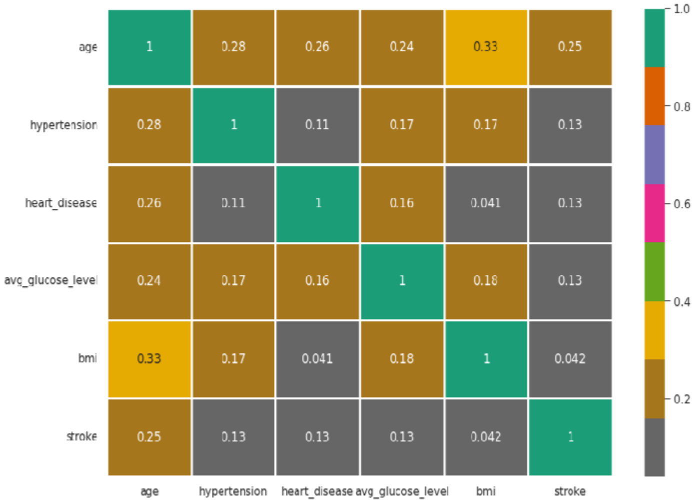


Figure 11

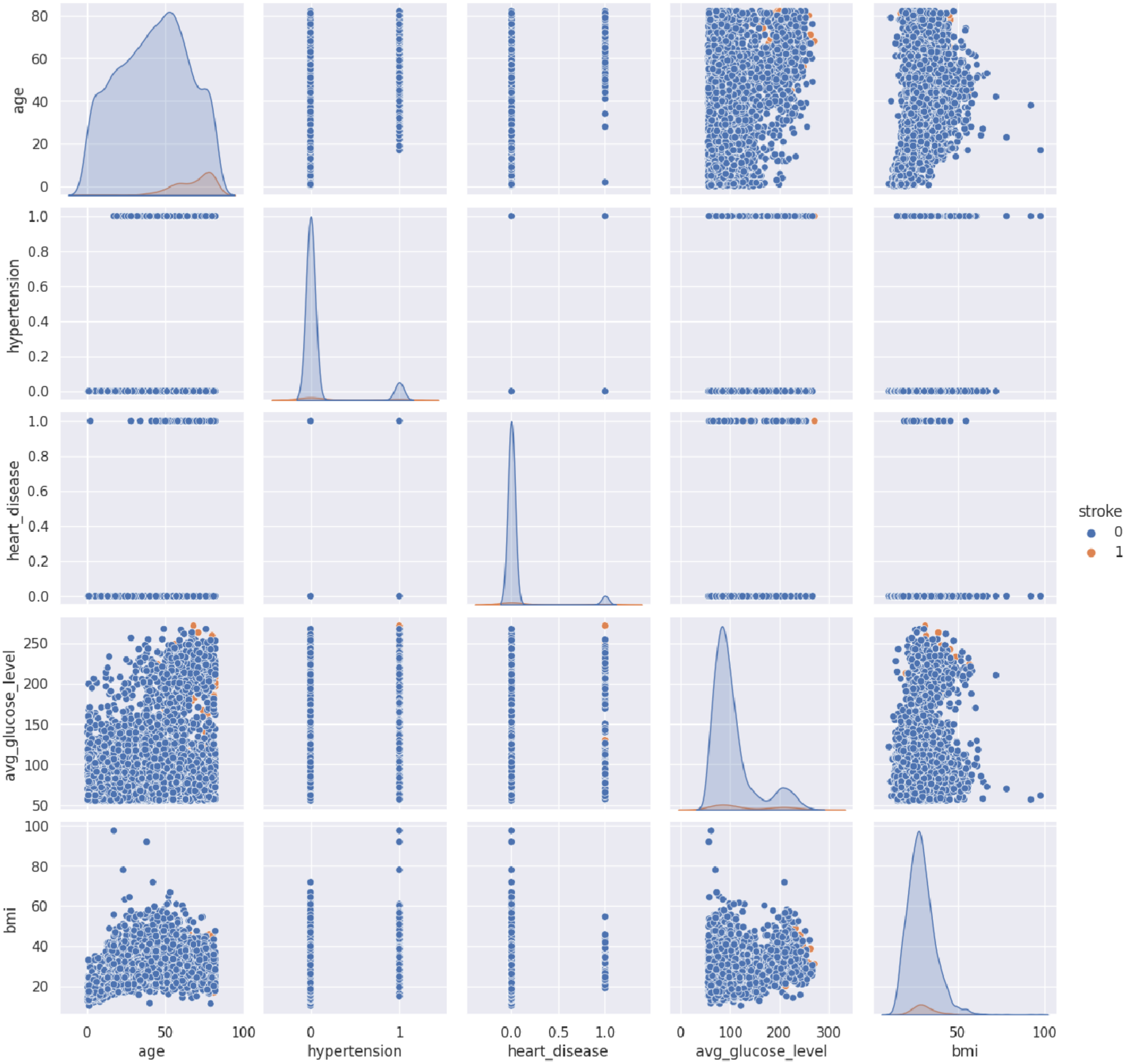

age
hypertension
heart_disease
avg_glucose_level
bmi
0
20
40
60
80
0.0
0.2
0.4
0.6
0.8
1.0
50
100
150
200
250
20
40
60
80
100
0
50
100
0
1
0.0
0.5
1.0
0
100
200
300
50
100
stroke
● 0
● 1

Figure 12

Anomaly Detection on Strokes Dataset:

Anomaly detection is a process of identifying data points that deviate significantly from other data points in the same dataset. These anomalous data points, also known as outliers, can provide valuable insights into the data and help in detecting errors, fraud, or other unusual events.

To detect anomalies, anomaly detection algorithms produce two different outputs: a categorical tag and a score or trust value. The tag categorizes the data point as either normal or abnormal based on a threshold or predefined rule. The tag simply tells us if the data point is an anomaly or not.

On the other hand, the score or trust value provides more information about the data point. It measures the degree of anomaly of a data point and indicates how far it deviates from the normal range. A high score indicates that the data point is highly anomalous, while a low score means that it is closer to the normal range.

Anomaly is one that differs / deviates significantly from other observations in the same sample. An anomaly detection pattern produces two different results. The first is a categorical tag for whether the observation is abnormal or not; the second is a score or trust value. Score carries more information than the label. Because it also tells us how abnormal the observation is. The tag just tells you if it's abnormal. While labeling is more common in supervised methods, the score is more common in unsupervised and semisupervised methods. This has also been implemented in the stroke analysis dataset to remove outliers from the data and improve the accuracy percentage.

Steps to increase the accuracy of the Stroke prediction Model:

One-Hot Encoding

- One Hot Encoding is the binary representation of categorical variables. This process requires categorical values to be mapped to integer values first. Next, each integer value is represented as a binary vector with all values zero except the integer index marked with 1.
- One Hot Encoding makes the representation of categorical data more expressive and easier. Many machine learning algorithms cannot work directly with categorical data, so categories must be converted to numbers. This operation is required for input and output variables that are categorical.

Reasons for not opting for Dimension Reduction in Stroke Detection Dataset:

Dimensionality reduction is a technique used in machine learning to reduce the number of features or variables in a dataset while preserving its important information. This technique is often used to deal with high-dimensional datasets, which can pose challenges for many machine learning algorithms, such as overfitting, computational complexity, and difficulty in visualization.

In the case of the provided dataset of 5110 rows and 11 columns, it already has a relatively low number of dimensions. Therefore, it is not necessary to perform dimensionality reduction on this dataset as it is manageable for most machine learning algorithms. Furthermore, removing any of the features in this dataset could result in the loss of important information that may be crucial for accurate prediction and analysis.

In addition, this dataset consists of numerical features that can be easily visualized and interpreted, making it easier to understand the relationships between the features and the target variable. Therefore, reducing the number of dimensions may not provide any significant benefits in terms of computational efficiency or model performance.

In summary, the provided dataset does not require dimensionality reduction as it has a low number of dimensions and the features are crucial for accurate prediction and analysis.

8. Exploration of Candidate Data Mining Models and Select the Final Model:

There have been five Candidate Machine Learning models under consideration. The models have been trained and explored using the training data.

- **Random Forest Classifier**

Advantages:

- Can handle high-dimensional datasets and large datasets with many features
- Can model complex relationships between variables and capture non-linearities
- Provides estimates of feature importance
- Robust to noise and overfitting

Disadvantages:

- Can be slow to train on very large datasets
- Can overfit if not tuned properly
- Can be difficult to interpret

- **Logistic Regression**

Advantages:

- Simple and interpretable
- Works well for linearly separable data
- Can handle high-dimensional datasets
- Provides probabilistic interpretations of the output

Disadvantages:

- Assumes a linear relationship between the input variables and the output
- Can underfit if the relationship between the input variables and the output is complex
- Can be sensitive to outliers

- **Support Vector Machines (SVMs)**

Advantages:

- Can handle high-dimensional datasets
- Can model non-linear relationships between variables
- Robust to noise and overfitting
- Can use different kernel functions to model different types of relationships

Disadvantages:

- Can be sensitive to the choice of kernel function and its parameters
- Can be computationally expensive to train on very large datasets
- Can be difficult to interpret

- **K-Nearest Neighbors (KNN)**

Advantages:

- Simple and easy to understand
- Can handle non-linear relationships between variables
- Can be effective when the decision boundary is irregular

Disadvantages:

- Can be computationally expensive to make predictions on large datasets
- Requires a metric to calculate distance between data points
- Can be sensitive to the choice of the number of neighbors (k)

5. **Gradient Boosting Classifier**

Advantages:

- Can handle high-dimensional datasets
- Can model complex non-linear relationships between variables
- Tends to have high accuracy and low bias
- Can handle missing data effectively

Disadvantages:

- Can be sensitive to overfitting if not tuned properly
- Can be computationally expensive to train on large datasets
- Can be difficult to interpret

9. Model Performance Evaluation and Interpretation

Accuracy of GaussianNB ----- > 17.477656405163852

Accuracy of BernoulliNB----- > 95.72989076464746

Accuracy of LogisticRegression ----- > 96.92154915590864

Accuracy of RandomForestClassifier ----- > 96.8222442899702

Accuracy of SupportVectorMachine ----- > 96.92154915590864

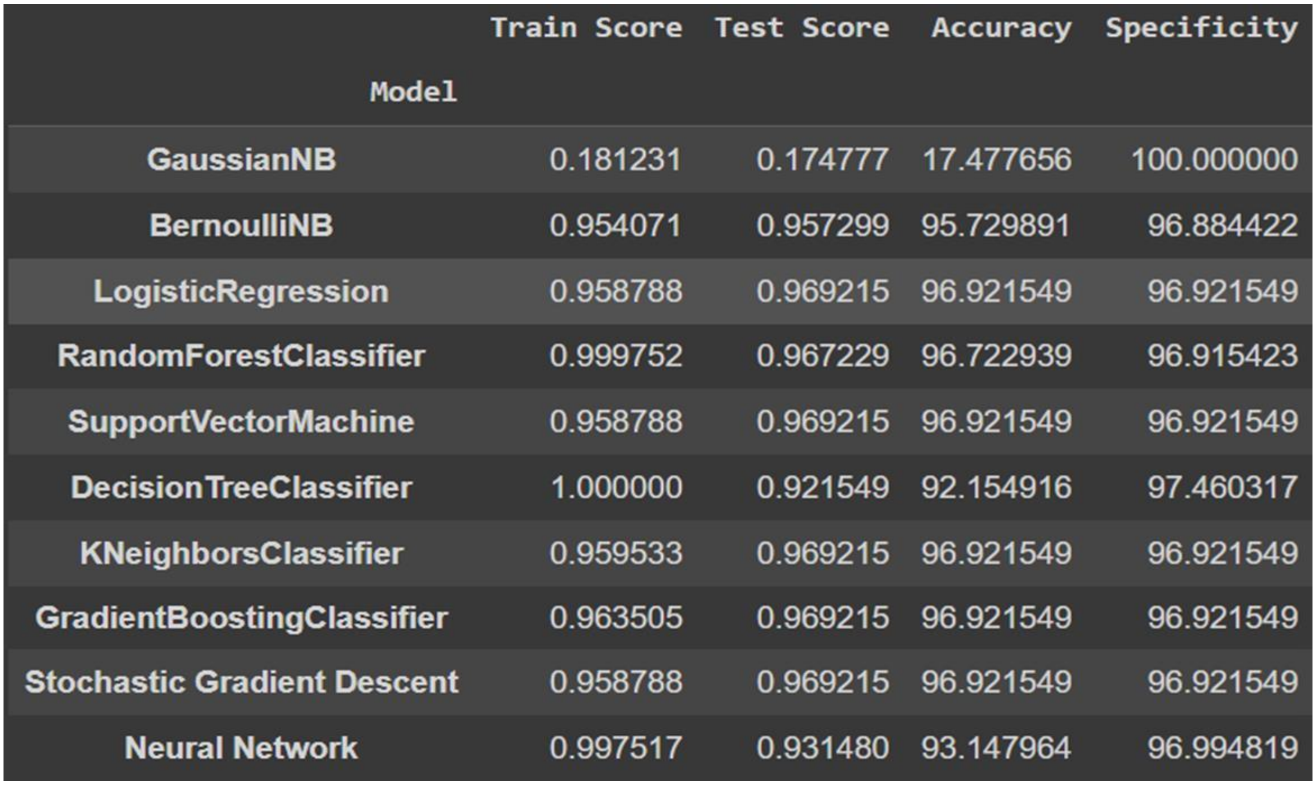
Accuracy of DecisionTreeClassifier------> 92.25422045680239

Accuracy of KNeighborsClassifier ----- > 96.92154915590864

Accuracy of GradientBoostingClassifier------ > 96.92154915590864

Accuracy of Stochastic Gradient Descent ----- > 96.92154915590864

Accuracy of Neural Nets----- > 93.14796425024826



	Train Score	Test Score	Accuracy	Specificity
Model				
GaussianNB	0.181231	0.174777	17.477656	100.000000
BernoulliNB	0.954071	0.957299	95.729891	96.884422
LogisticRegression	0.958788	0.969215	96.921549	96.921549
RandomForestClassifier	0.999752	0.967229	96.722939	96.915423
SupportVectorMachine	0.958788	0.969215	96.921549	96.921549
DecisionTreeClassifier	1.000000	0.921549	92.154916	97.460317
KNeighborsClassifier	0.959533	0.969215	96.921549	96.921549
GradientBoostingClassifier	0.963505	0.969215	96.921549	96.921549
Stochastic Gradient Descent	0.958788	0.969215	96.921549	96.921549
Neural Network	0.997517	0.931480	93.147964	96.994819

Figure 13

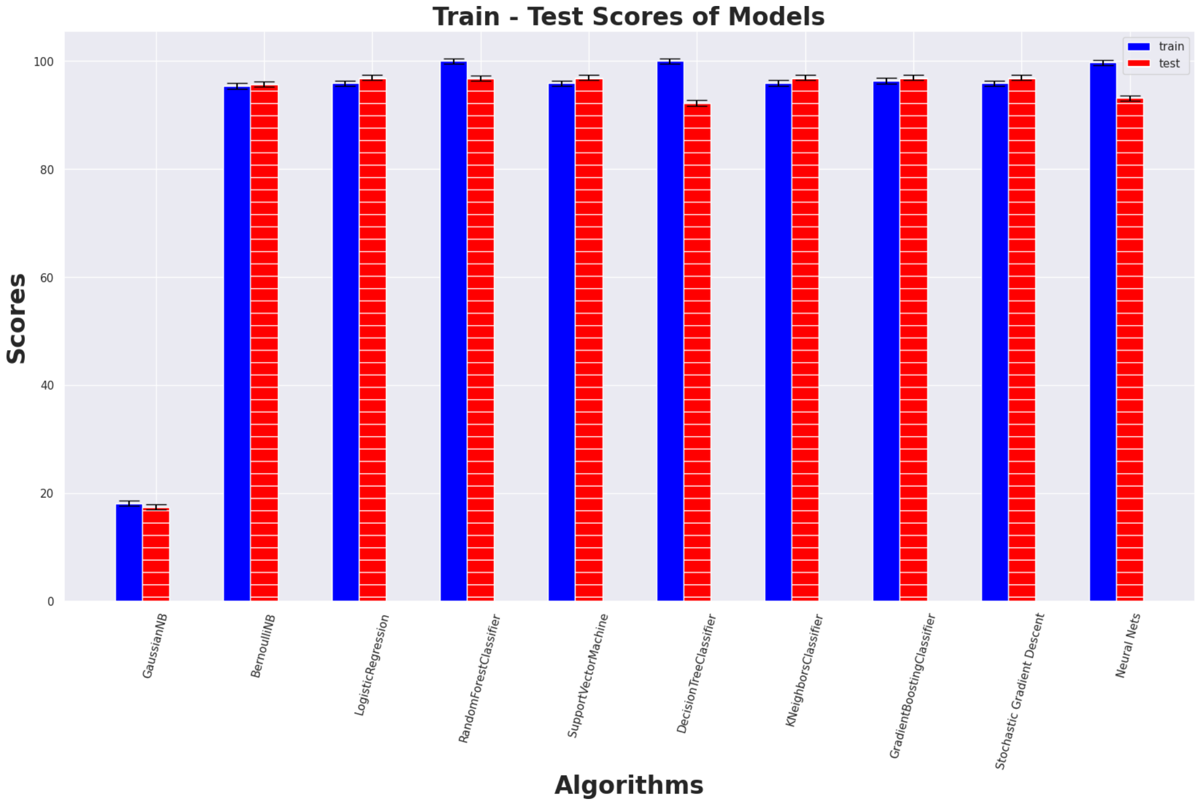


Figure 14

The Highest is the "Accuracy of KNeighborsClassifier ----- > 96.22641509433963"

10. Evaluation of Models:

We evaluated these models according to their accuracies. Best algorithm is KNN with 96.921%. So, we will make k-Fold Cross Validation and Hyper-Parameter Optimization for KNN algorithm.

Train score of trained model: 0.9595332671300894

Test score of trained model: 0.9682539682539683

Accuracy: 96.82539682539682

Confusion matrix:

```
[[485  15]
 [  1   3]]
```

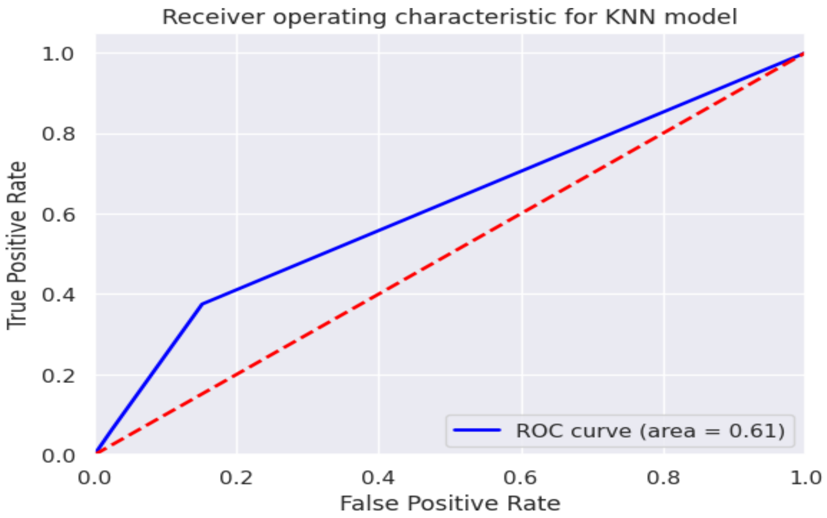


Figure 15

ROC Curve for the KNN model

Features Selection for RandomForestClassifier (Selecting the most important features from this dataset)

Old Shape: (5035, 19) New shape: (5035, 3)

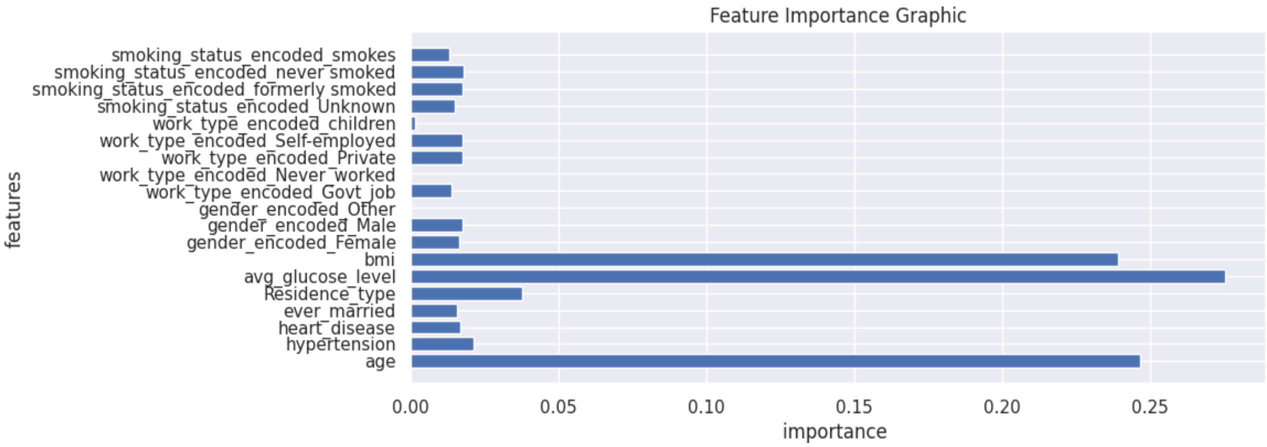


Figure 16

11. Project Impact:

- **Improved healthcare**: By accurately predicting stroke risk, healthcare providers can take proactive measures to prevent or reduce the impact of stroke on patients. This can include recommending lifestyle changes, prescribing medications, or even surgery in extreme cases. By providing early intervention, healthcare providers can improve patient outcomes and reduce healthcare costs associated with stroke treatment and rehabilitation.

- **Public health awareness**: The project can raise awareness about the risk factors for stroke and the importance of early intervention. The public can learn about the modifiable and non-modifiable risk factors, such as age, gender, hypertension, diabetes, and smoking, and how to manage them effectively.

By educating the public about stroke risk factors, individuals may take steps to reduce their risk of stroke, leading to better public health outcomes.

- **Personalized healthcare**: The prediction models can be used to develop personalized healthcare plans for individuals based on their unique risk factors. By tailoring healthcare plans to an individual's specific risk factors, healthcare providers can develop more targeted and effective interventions. This can improve patient outcomes and reduce the cost of healthcare delivery.

- **Advancements in machine learning**: The project can contribute to the advancement of machine learning techniques and their applications in healthcare. By identifying which algorithms are most effective in predicting stroke risk, researchers can develop more accurate and efficient models for use in healthcare settings. This can lead to further innovations in healthcare delivery and improved patient outcomes. Additionally, the insights gained from this project can be applied to other healthcare domains, leading to more widespread use of machine learning in healthcare.