**Final Report**

Contents

[**Introduction:** 2](#_Toc163809745)

[**DATASET:** 3](#_Toc163809746)

[**Preliminary Analysis:** 3](#_Toc163809747)

[**Methods:** 5](#_Toc163809748)

[**Logistic Regression:** 5](#_Toc163809749)

[**Support Vector Machine (SVM):** 7](#_Toc163809750)

[**Results:** 8](#_Toc163809751)

[**Figures:** 9](#_Toc163809752)

[**Conclusion:** 17](#_Toc163809753)

# **Introduction:**

Strokes, also known as cerebrovascular accidents, are a major global public health problem because of their high rates of morbidity and mortality. Strokes are a major cause of long-term impairment and the second greatest cause of mortality worldwide, according to the World Health Organization (WHO). In low- and middle-income nations, where access to preventive measures may be restricted and healthcare resources may be scarce, the prevalence of stroke is especially high. Given these difficulties, applying artificial intelligence (AI) has become a viable strategy for enhancing stroke care and prediction. Artificial intelligence (AI) has the potential to improve the precision and efficacy of stroke risk assessment, early identification, and individualized treatment plans by utilizing cutting-edge computational approaches and machine learning algorithms.

The goal of this project is to determine how well two popular machine learning models—logistic regression and support vector machine (SVM) predict various stroke types using medical data. Our goal is to examine the prediction power of different models and assess how well they identify those who are at risk of stroke.

**Context and Motivation:**

The pressing need for precise and trustworthy stroke prediction tools in clinical practice is what inspired this effort. The timely identification of individuals who are susceptible to strokes is of utmost importance in order to execute preventative measures, commence suitable interventions, and reduce the related morbidity and death. While useful, traditional risk assessment techniques may not be as scalable or accurate in the long run, especially when working with huge and complicated datasets.   
Predictive models for a variety of medical illnesses, including strokes, have benefited greatly from the availability of electronic health records (EHRs), medical imaging data, and other healthcare data sources in recent years. These large datasets may yield valuable patterns and insights that may be extracted by AI algorithms, allowing for more precise and customized risk categorization for stroke prevention.

**Originality and Significance:**

Although the application of machine learning models for stroke prediction has been investigated in the past, this study adds to the body of literature by concentrating on the effectiveness of SVM and logistic regression algorithms in the context of assessing stroke risk. In order to shed light on these two models' advantages, disadvantages, and applicability for clinical settings, we will compare them using actual healthcare data.  
Moreover, the uniqueness of this work is seen in its thorough assessment of AI-based methods for predicting strokes, taking into account variables like interpretability, scalability, and model accuracy. We hope to educate policymakers, researchers, and healthcare professionals on the potential of AI to enhance stroke care and drive future research efforts in this area by addressing these important issues.

In conclusion, by utilizing AI technology, this project seeks to answer the urgent demand for precise and effective stroke prediction tools. We aim to improve our knowledge of the usefulness of logistic regression and SVM models in assessing stroke risk and help create more efficient methods for managing and preventing strokes through our research into these models.

# **DATASET:**

Our stroke prediction project is using the "Health Care Stroke Prediction Data" dataset, which we obtained from Kaggle. This dataset attempts to estimate a person's chance of having a stroke based on a number of health characteristics and offers insightful information on the causes that lead to strokes. This section will offer a thorough description of the dataset, covering its attributes, data sources, and extraction procedures.

The dataset includes a range of health-related characteristics for each participant as well as a signal indicating whether or not they have had a stroke. The elements of the dataset, which each item represents a distinct individual, include medical history, lifestyle factors, and demographic data. The binary target variable has values that represent whether or not the person has experienced a stroke (1) or not (0).

**Features:**

Numerous characteristics in the dataset may be important in the prediction of strokes. Among the salient characteristics are:

1. Gender: Denotes the person's gender (Male/Female/Other).
2. Age: Denotes the person's actual age.
3. Hypertension: A binary characteristic that indicates if a person has high blood pressure (1) or not (0).
4. Heart disease: A binary characteristic that indicates if a person has ever experienced heart illness (1) or not (0).
5. Marital Status: This indicates whether or not the person is married.
6. Work Type: Indicates the kind of work that the person is doing (private, self-employed, government job, etc.).
7. Residence Type: Denotes whether the person lives in a rural or urban setting.
8. The average glucose level in a person's blood is indicated by this number.
9. Body Mass Index, or BMI, is determined by taking a person's weight and height.
10. Smoking Status: Indicates the individual's smoking habits (e.g., never smoked, formerly smoked, smokes).

To sum up, the "Health Care Stroke Prediction Data" dataset is an invaluable tool for researching stroke risk variables and creating models that predict the likelihood of stroke. Researchers and healthcare professionals can learn more about the intricate interactions between variables that lead to strokes thanks to the dataset's many features and binary stroke indication. Our goal is to use this information to create predictive models that are useful for the early identification and treatment of stroke risk factors in people.

# **Preliminary Analysis:**

An important first phase in the data analysis process is preliminary analysis, sometimes referred to as exploratory data analysis (EDA). It entails dissecting and comprehending a dataset's structure, features, and trends before utilizing more sophisticated statistical methods or developing prediction models. Gaining an understanding of the data, spotting abnormalities or outliers, and providing guidance for further research are the goals of preliminary analysis.

This preliminary investigation sheds light on the information utilized to forecast the danger of stroke. The results are broken down as follows:

**Data Cleaning:**

* Using data.isnull().sum(), which shows the total count of null values in each column, we looked for any missing values. We see that only ‘bmi’ has more than 200 null values
* In particular, we looked into missing values in the data['bmi'].isnull().astype(int) column of the "BMI" column.

**Data Exploration:**

* Through data.info(), we were able to gather broad details about the non-null values and data types in each column.
* Using data.describe(), we generated a summary of the numerical columns, which included the mean, standard deviation, minimum, maximum, percentiles, etc.
* To illustrate the distribution of various diseases, we separated the numerical and categorical data and used histograms (nume.hist()) for the numerical columns and the top few rows (cate.head(), nume.head()) for the category columns.

**Stroke Distribution:**

* The distribution of stroke occurrences was plotted using sns.countplot(x='stroke', data=data).
* The percentage distribution of patients with and without strokes was computed and plotted (data['stroke'].\* 100 = value\_counts(normalize=True).

**Correlations and Feature Analysis:**

* To investigate the differences in age, average blood sugar, and body mass index between stroke patients and non-survivors, we made box plots (sns.boxplot()).
* We used count plots (sns.countplot()) to examine the distribution of "hypertension" and "heart\_disease" by stroke status.
* We looked into the relationship between other numerical variables and the missing values in "BMI" (num\_data.corrwith(miss\_bmi)).

**Missing Value Imputation - BMI:**

* In the "BMI" column (data['bmi'].isnull().sum()), we found missing values.
* To estimate missing BMI values based on other characteristics such as age, hypertension, heart disease, average glucose level, and stroke status, we employed a linear regression model (LinearRegression()).
* We assessed the performance of the model using the training set (optional, not displayed here).
* The projected BMI values were re-imputed into the original dataset.

**Post-Imputation Analysis:**

* We verified that missing values (data.isnull().sum()) had been removed.
* To check for any changes, we looked over the data information again (data.info()).
* Using a histogram and kernel density estimation (KDE), we were able to display the distribution of the imputed BMI data and compare it to a normal distribution (sns.histplot(), norm.fit()).
* To evaluate the normality of the BMI data distribution, we made a Q-Q plot (ggplot(data, aes(sample=data['bmi']))).

**Transformation and QQ Plot:**

* We verified that missing values (data.isnull().sum()) had been removed.
* In an attempt to maybe obtain a more normal distribution, we applied the Box-Cox transformation (yeojohnson()) to the BMI data.
* To evaluate the normality of the transformed BMI data, we made a QQ plot (ggplot(trans\_data, aes(sample='bmi'))).

This preliminary analysis lays the groundwork for additional research. It evaluates the distribution of important variables, looks at any problems with missing values, and investigates connections between characteristics and the target variable (stroke). To create a strong stroke prediction model, more data cleaning, feature engineering, and model selection procedures can be implemented in light of these findings.

# **Methods:**

## **Logistic Regression:**

A supervised machine learning approach called logistic regression predicts the likelihood of a result, an occurrence, or an observation to complete binary classification problems. The output of the model is binary, or dichotomous, with two possible outcomes: true or false, 0/1, or yes/no.

By examining the correlation between one or more independent variables, logical regression divides data into distinct groups. It is often applied in predictive modeling, in which the model calculates the mathematical likelihood of an event falling into a particular category or not. For instance, the numbers 0 and 1 stand for negative and positive classes, respectively. In binary classification issues, where the outcome variable reveals one of the two categories, (0 or 1), logistic regression is frequently utilized.

Here are a few examples of these groupings and situations where the binary response is either inferred or expected:

* **Determine the probability of heart attacks:** Using a logistic model, medical professionals can ascertain the correlation between an individual's weight, level of exercise, and other variables in order to forecast the likelihood of a heart attack or other medical issues.
* **Possibility of enrolling in a university:** By analyzing the relationship between the estimator variables, such as GRE, GMAT, or TOEFL scores, application aggregators can ascertain the likelihood that a student will be accepted to a specific institution or a degree course in a college.

**Advantages of Logistic Regression:**

* Machine learning techniques that are easier to use: Training and testing are key components in the efficient setup of a machine learning model. The process of training finds patterns in the input data (picture) and links them to an output (label). Regression algorithm training of a logistic model does not require more processing power. Compared to other ML techniques, logistic regression is therefore simpler to use, understand, and train.
* Ideal for datasets that can be separated linearly: A graph in which a straight line divides the two data classes is referred to as a linearly separable dataset. The y variable in logistic regression only accepts two values. Therefore, if linearly separable data is used, it is possible to effectively categorize the data into two distinct classes.
* Offers insightful information In addition to indicating the direction of a link or association (positive or negative), logistic regression indicates the degree of relevance or appropriateness of an independent or predictor variable (coefficient size).

**How we use this model:**

* For the initial prediction, a logistic regression model was selected.
* X\_train\_imputed and y\_train, the pre-processed training data, were used to train the model.
* Using the training data, the model's performance was assessed using:
* Accuracy Score: The percentage of accurate predictions the model makes is shown by this indicator.
* Confusion Matrix: This matrix shows how many guesses were right and wrong for every class (stroke and no stroke).
* Report on Classification: Extensive details regarding the model's performance are provided in this report, encompassing precision, recall, F1-score, and support for every class.

## **Support Vector Machine (SVM):**

The support vector machine (SVM) is a machine learning algorithm that determines boundaries between data points based on predefined classes, labels, or outputs. It uses supervised learning models to solve complex problems related to classification, regression, and outlier detection. SVMs are widely used in many sectors, including speech and image recognition, natural language processing, healthcare, and signal processing applications. The SVM algorithm's main goal, technically, is to locate a hyperplane that clearly divides the data points into separate classes. The hyperplane is positioned so that the classes being considered are separated by the greatest margin.

It is possible that SVMs were created to solve binary classification issues. But when computationally demanding multiclass problems become more common, a number of binary classifiers are built and coupled to create SVMs that can carry out these kinds of multiclass classifications using binary methods.

A better understanding of a support vector machine's operation can be gained from an example. Assume that the features represented by x and y are written on red and black labels, respectively. For these tags, we plan to have a classifier that divides data into two categories: red and black. Typically, an SVM uses the hyperplane a two-dimensional line in this case to divide these data points into red and black tags. The decision boundary line, where data points fall into the red or black category, is represented by the hyperplane.

A line that tends to enlarge the space between the two nearest tags or labels (black and red) is called a hyperplane. Data classification is facilitated by the hyperplane's largest distance to the most immediate label.

**Advantages of Support Vector Machine:**

* When there is a distinct margin of difference between classes, SVM performs reasonably well.
* In spaces with higher dimensions, SVM performs better.
* In situations where there are more dimensions than samples, SVM performs well.
* SVM uses memory reasonably efficiently.

**How we use this model:**

* Balanced class weights and an RBF kernel were used to create an SVM model.
* Prior to training the model, the training data underwent feature scaling.
* Accuracy score, confusion matrix, and classification report were used to assess the model's performance on the training set, much like in the case of logistic regression.

# **Results:**

The outcomes of training and assessing two machine learning models for stroke risk prediction are shown in this section. The models investigated were Support Vector Machine (SVM) with an RBF kernel and Logistic Regression.

**Model Performance:**

We used the accuracy score to assess each model's performance. The percentage of accurate predictions the model makes is measured by accuracy. The outcomes are broken down as follows:

|  |  |  |
| --- | --- | --- |
| **Model** | **Testing Accuracy** | **Training Accuracy** |
| **Logistic Regression** | 70% | 74% |
| **SVM** | 85% | 81% |

**Comparison and Interpretation:**The SVM model consistently outperformed Logistic Regression in terms of accuracy on both the training and testing data, as shown by the bar chart. With a training accuracy of 81% and a testing accuracy of 85%, the SVM demonstrated a strong fit to the training set and respectable generalization to new data. On the other hand, 74% of training and 70% of testing accuracy were achieved with Logistic Regression. Despite its decent performance, Logistic Regression may have overfitted based on the testing data's accuracy decline.

Given this particular case, the SVM model may be a superior option for predicting stroke risk due to its higher accuracy. But it's crucial to take restrictions into account:

* **Overfitting:** Although the SVM performs well, additional research may be necessary to make sure the training set isn't being over fit. For a more thorough assessment, methods like employing a validation set and tweaking the hyper parameters could be used.
* **Alternative Metrics:** While accuracy is a standard statistic, other metrics such as F1-score or AUC-ROC may provide additional insight into imbalanced datasets (where one class may be less frequent).

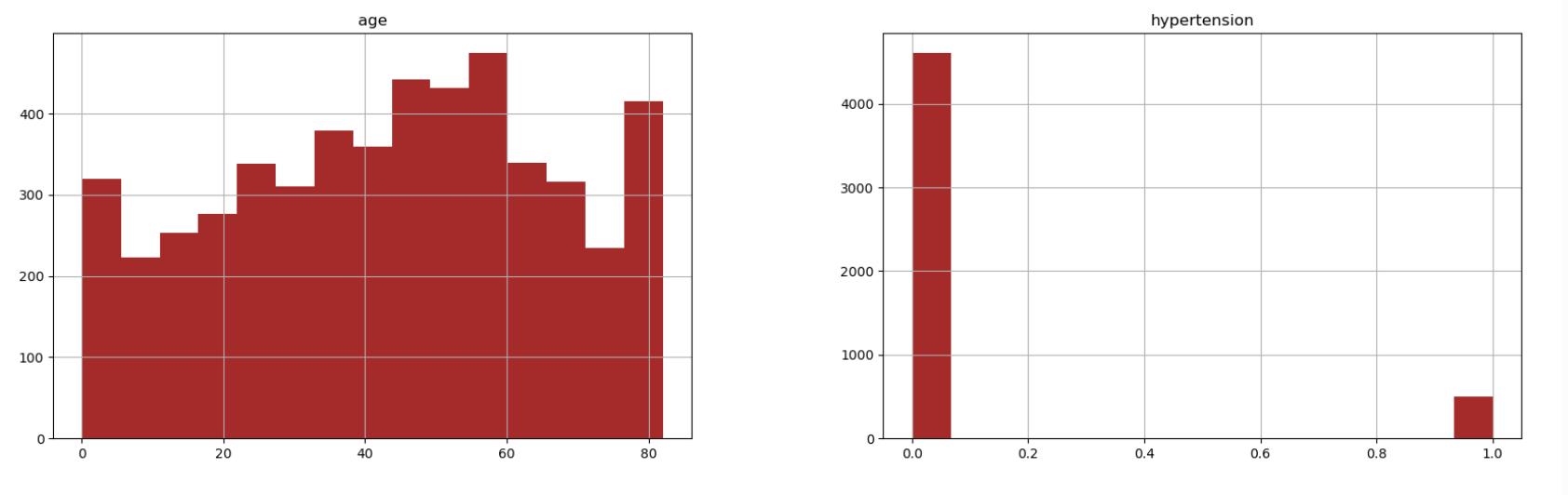
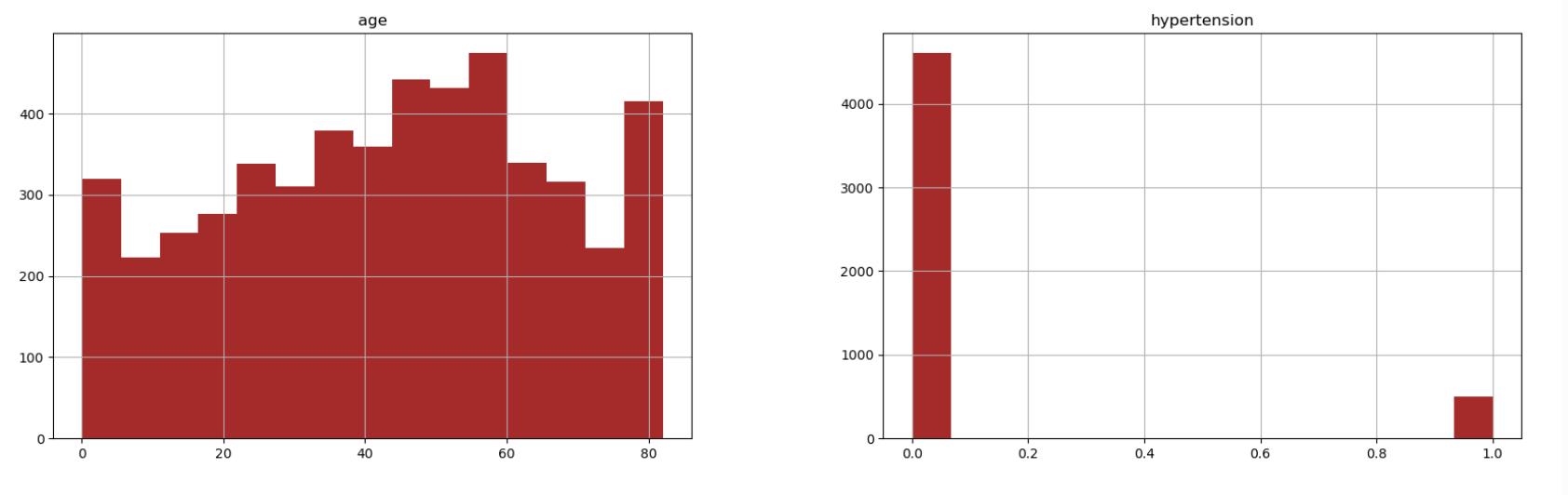
**Future Exploration:**

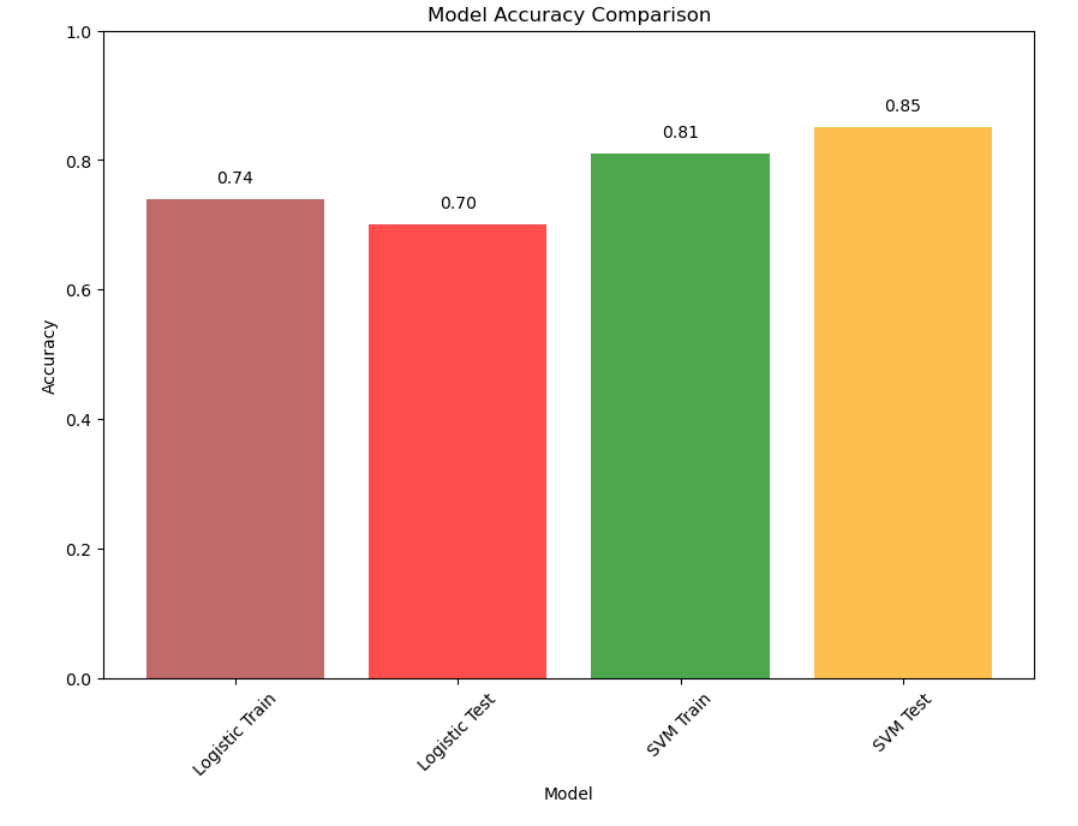
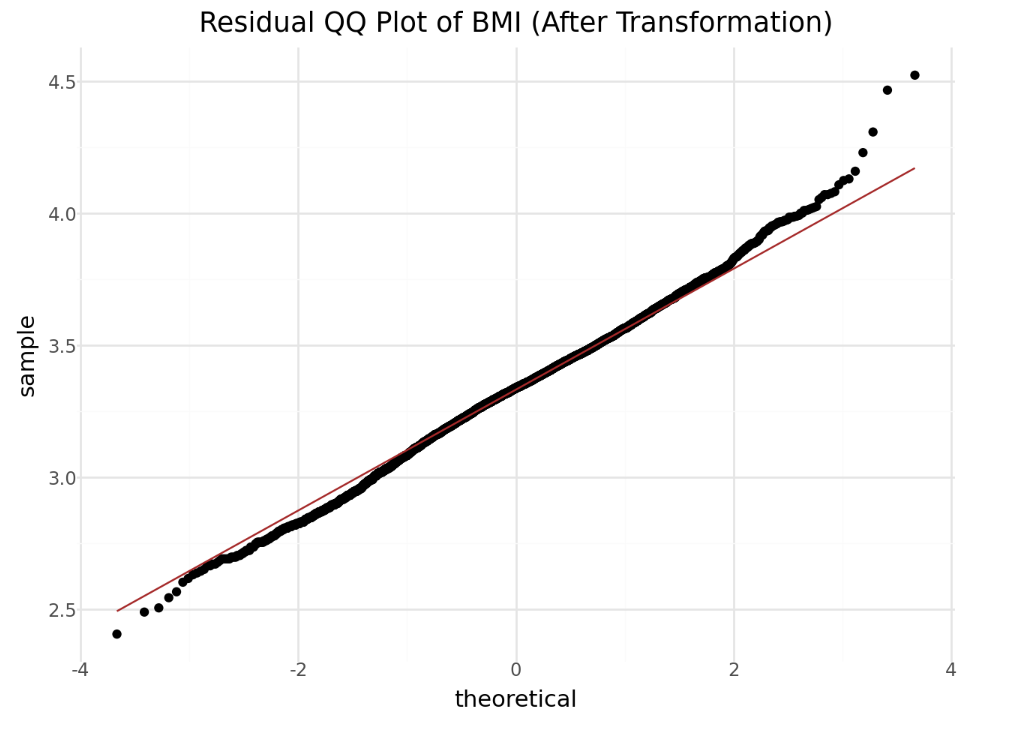
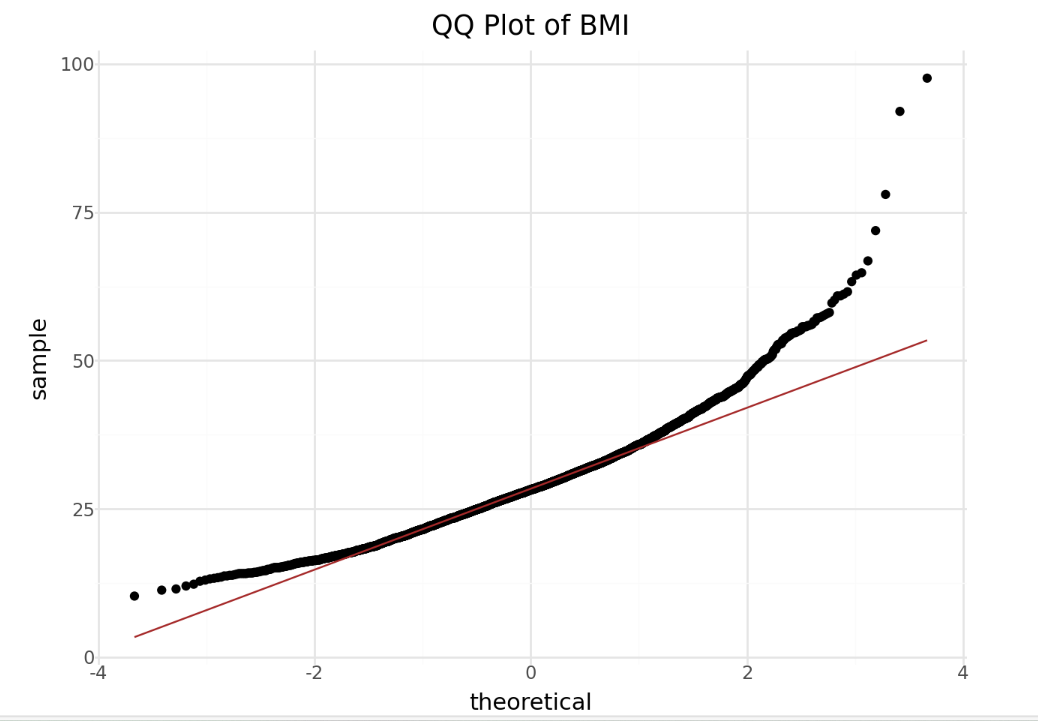
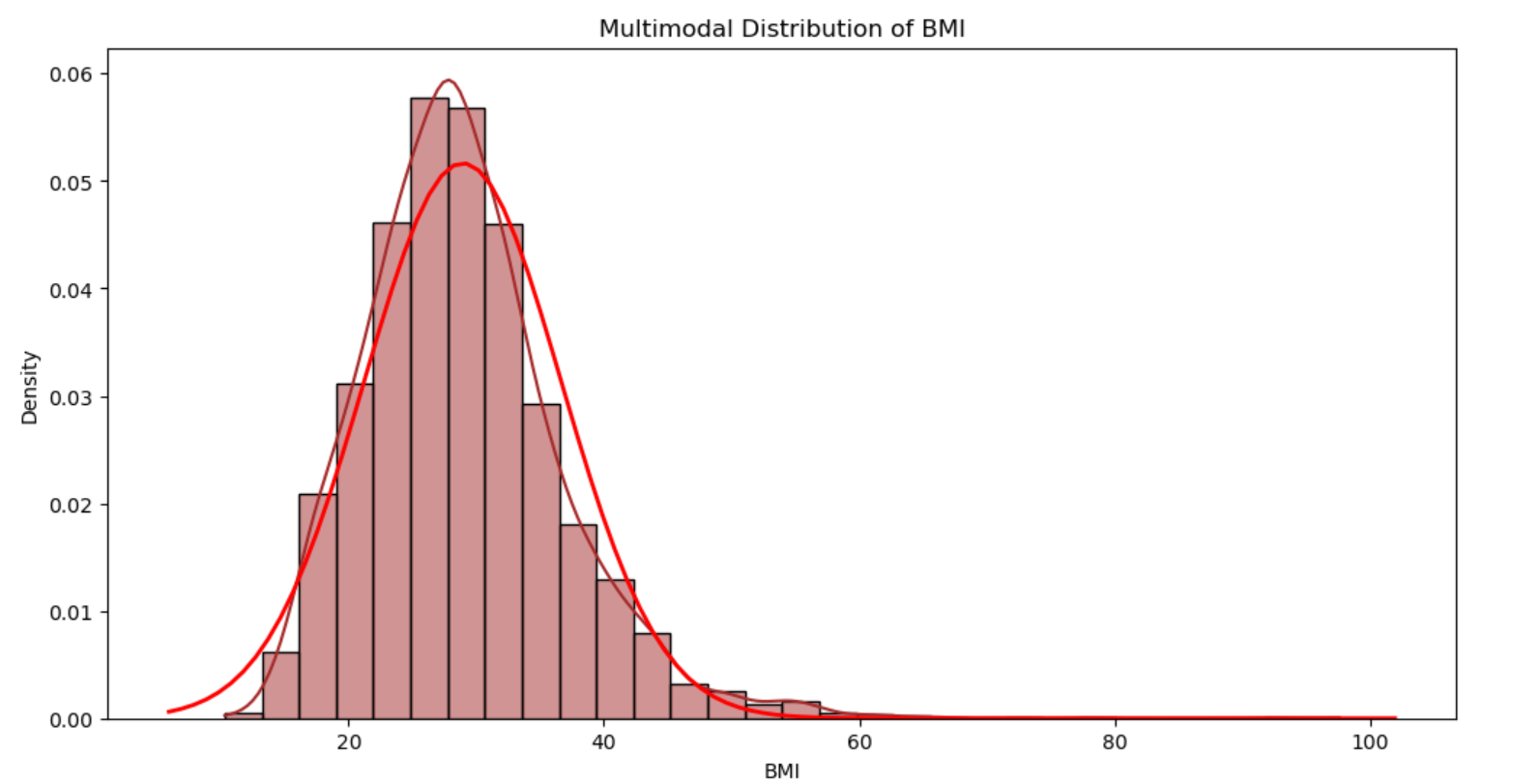
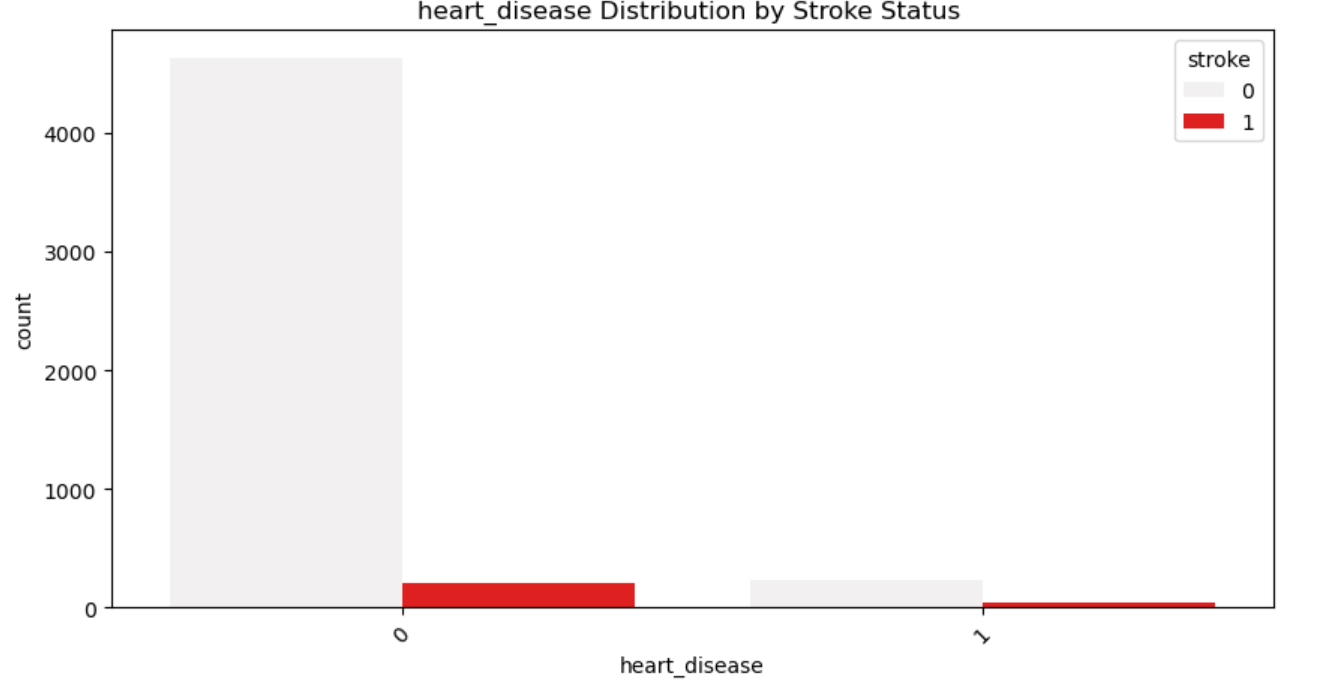
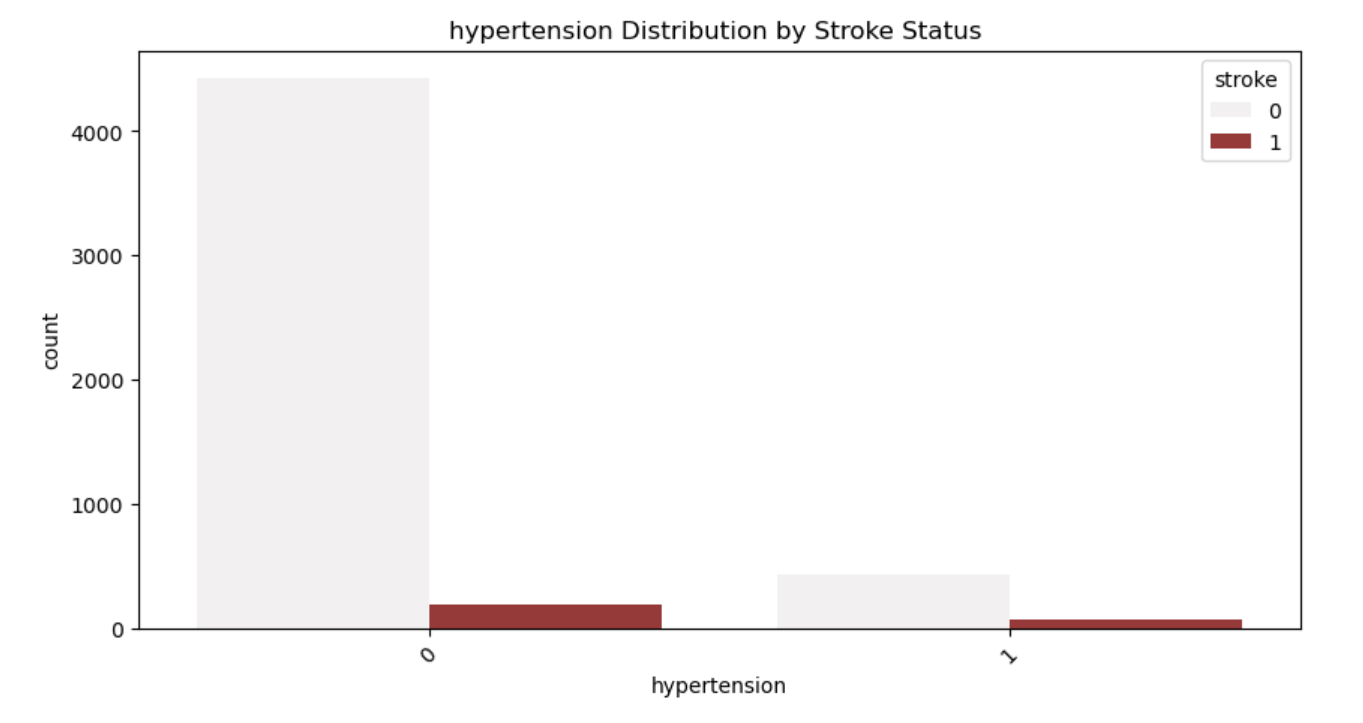
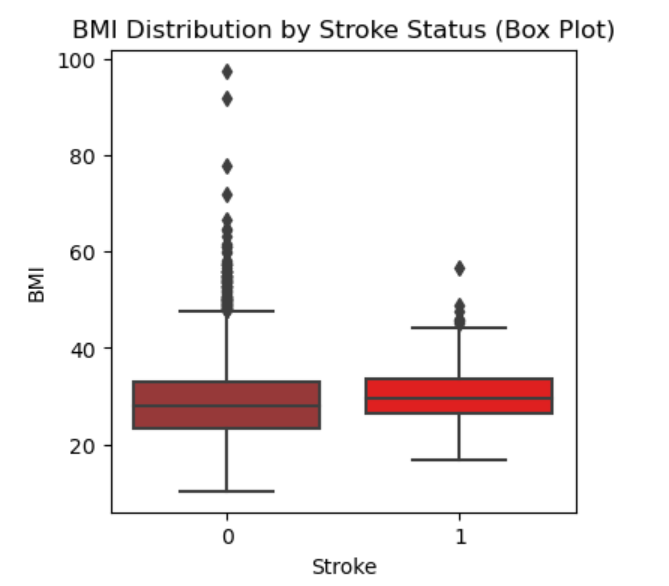
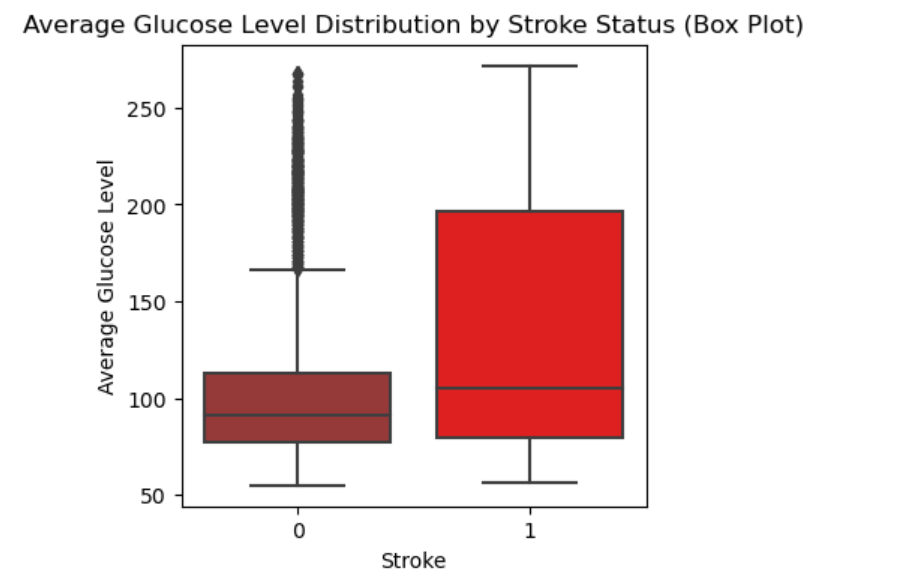
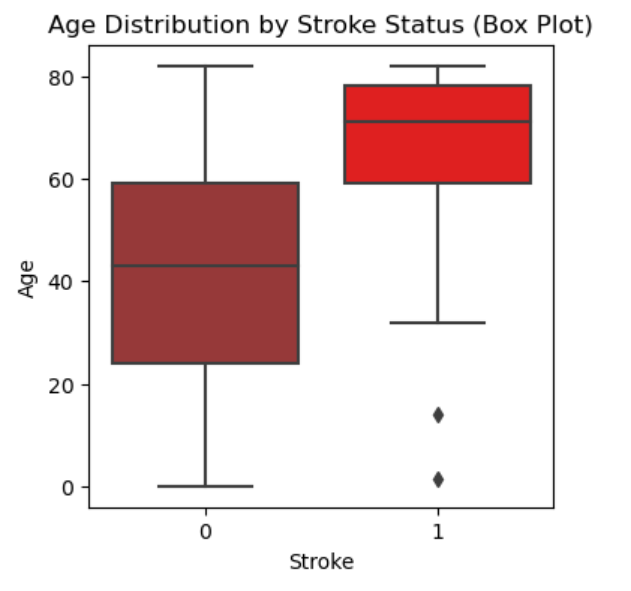
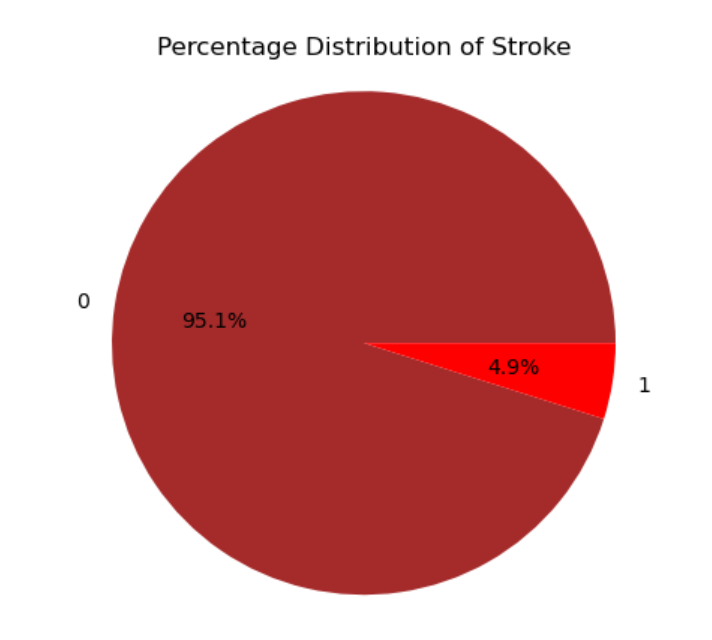
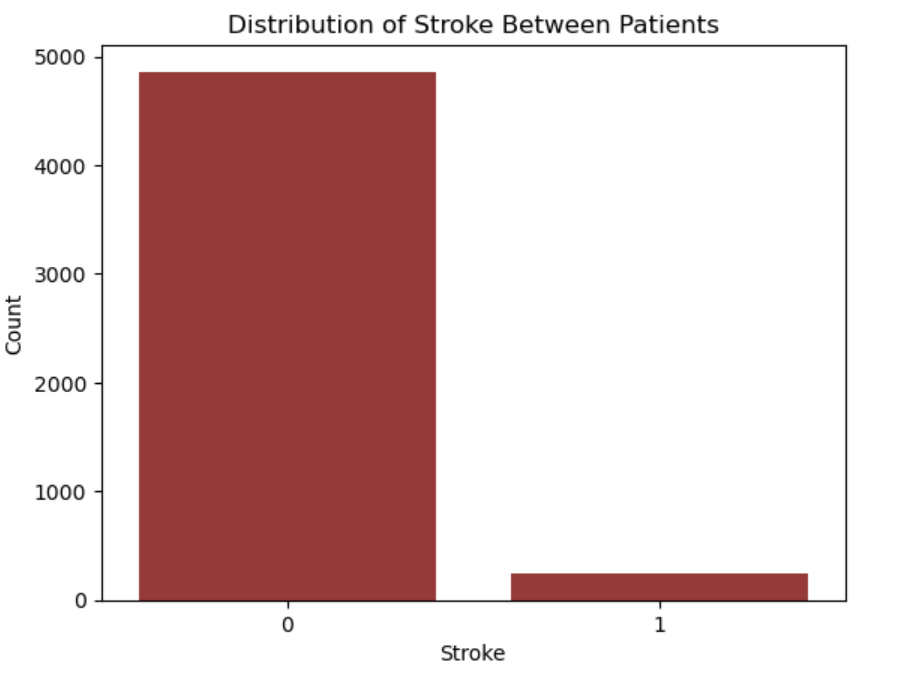
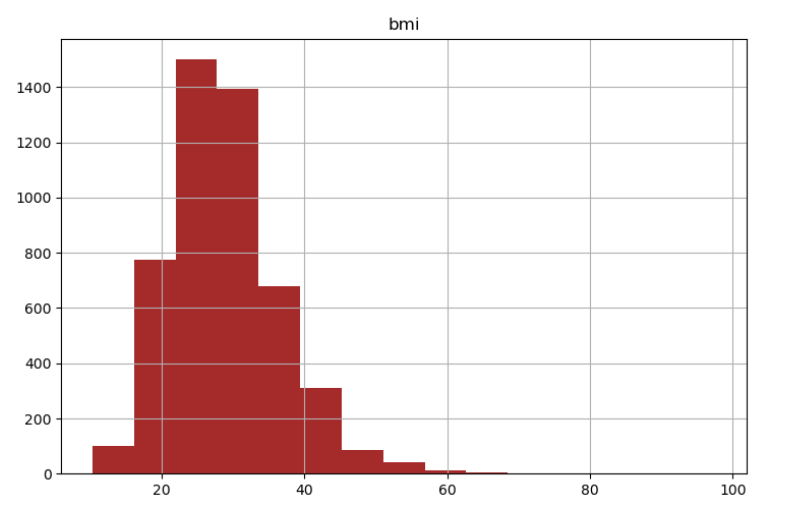
In light of these findings, additional research might focus on:

* Hyper parameter tuning: Performance may be enhanced by fine-tuning hyperparameters such as the kernel coefficient in SVM or the regularization parameter in Logistic Regression.
* Model Selection: For even better outcomes, consider looking into alternative machine learning models such as Gradient Boosting Machines or Random Forests.
* Feature engineering: By building new characteristics off of preexisting ones, more insightful data for prediction models may be obtained.

Two models for predicting strokes were successfully implemented overall by this research. Although the SVM model performed better, more research is advised for a more thorough analysis and possibly better results.

# **Figures:**





# **Conclusion:**

To sum up, this project is a major advancement in the use of artificial intelligence (AI) to meet the urgent demand for accurate and reliable stroke prediction tools. As the introduction noted, strokes pose a serious threat to global public health due to their high rates of morbidity and mortality. The incidence of strokes is startlingly high, especially in low- and middle-income nations where access to healthcare resources and preventive measures may be restricted. Given these obstacles, applying artificial intelligence (AI), which has the potential to improve accuracy and effectiveness in stroke risk assessment and prediction, seems like a good course of action.

The urgent need for precise stroke prediction tools in clinical practice is what inspired this effort. In order to reduce stroke-related morbidity and death, preventive measures must be put into place and suitable therapies must be started as soon as people who are at risk of stroke are identified. Even while they have their uses, traditional risk assessment methods can be inaccurate and unable to scale, particularly when working with vast and complicated datasets. This is why it's important to use AI-based methods, which can mine large healthcare databases for insightful patterns and insights that lead to more accurate and customized risk classification for stroke prevention.

By concentrating on the efficacy of logistic regression and support vector machine (SVM) methods in determining stroke risk, this work adds to the body of existing knowledge. The report intends to provide light on the advantages, disadvantages, and suitability of these models in clinical settings using actual healthcare data. It also highlights how crucial it is to take into account aspects like model accuracy, scalability, and interpretability when assessing AI-based techniques for stroke prediction.

The project's outcomes show that, for both the training and testing datasets, SVM predicts strokes more accurately than logistic regression. This suggests that SVM may be able to accurately detect strokes due to its better predictive capacity. To assure the models' generalizability and dependability, it is imperative to address some of their shortcomings, such as the possibility of overfitting and the requirement for alternative assessment measures. Future research directions can look into feature engineering to improve model performance, hyperparameter tuning, and alternative machine learning models. Even though the SVM model performed well in this study, more research is necessary for a thorough examination and possibly better results.

In conclusion, this report highlights how artificial intelligence (AI) can be used to tackle the difficulties involved in predicting strokes, providing important information to researchers, policymakers, and medical practitioners. Our goals are to improve patient outcomes in the field of stroke prevention and management, propel future research efforts, and develop stroke care by utilizing cutting-edge computational techniques and machine learning algorithms.