**Name : Teslim Bolaji Adeyemo**

**Student id: 13368560**

**Stroke Prediction using Machine Learning Techniques**

**Abstract**

This project aimed to develop a big data machine-learning system to predict the risk of stroke using a sample clinical and demographic dataset available on Kaggle. The dataset was evaluated and preprocessed using various techniques to improve data quality. An appropriate architecture was identified and designed for a machine learning system with a focus on performance, scalability, and reliability. Data were visualized using appropriate techniques to aid in the analysis and interpretation of the results. The machine-learning system was implemented and evaluated, achieving a high level of functionality within the project's time constraints. The performance of the system was evaluated using various performance metrics, such as accuracy, precision, recall, and F1-score. The project's results provide insights into the application of big data and machine learning in healthcare and can assist healthcare professionals in making informed decisions about stroke prevention and treatment. The project concluded with a reflection on the process, including areas for improvement and directions for future research. Overall, this project demonstrates the feasibility and potential of using big data and machine learning for stroke prediction and prevention.

Table of Contents

[Introduction 4](#_Toc131855707)

[Tools 4](#_Toc131855708)

[About The Dataset 4](#_Toc131855709)

[Preprocessing The Data 5](#_Toc131855710)

[Architecture 7](#_Toc131855711)

[Visualization 8](#_Toc131855712)

[Machine Learning System Application-Methodology 13](#_Toc131855713)

[Conclusion And Reflection 16](#_Toc131855714)

# Introduction

Stroke is a leading cause of death and disability worldwide with significant social and economic implications. The early prediction and prevention of stroke can reduce its impact and improve patient outcomes. Big data and machine-learning techniques offer powerful tools for identifying patterns and predicting outcomes in large and complex datasets. In this project, we aimed to develop a big data machine learning system to predict the risk of stroke using a large dataset of clinical and demographic data. The system will be designed to assist healthcare professionals in making informed decisions regarding stroke prevention and treatment. The project will involve various stages, including dataset evaluation, preprocessing, architecture design, data visualization, machine learning system implementation, and performance evaluation. The project outcomes will provide insights into the application of big data and machine learning in healthcare and have the potential to improve stroke prevention and treatment.

## Tools

This project utilized several tools to develop a big data machine-learning system for stroke prediction. Data cleaning and manipulation were performed using the Python Pandas library, whereas data visualization was performed using Tableau. Machine learning models were developed using Scikit-Learn, which is a powerful and widely used machine learning library in Python. Together, these tools allow for effective data processing, analysis, and visualization, ultimately enabling the development of a robust and accurate stroke-prediction system.

## About The Dataset

The dataset used in this project comprises 5110 rows and 12 columns containing clinical and demographic data of stroke patients. The data, which is publicly available on Kaggle, includes features such as:

* age
* gender
* hypertension
* heart disease
* smoking status
* ever\_married
* work\_type
* Residence\_type
* avg\_glucose\_level
* bmi
* smoking\_status
* stroke

One notable characteristic of this dataset is the relatively small number of columns, which reduces the complexity of the data and makes it more suitable for big data projects. Additionally, the dataset contains a moderate number of rows, providing sufficient data to train the machine learning models effectively. However, it is important to note that the dataset has some missing values in the "bmi" column, with 201 out of 5110 rows containing null values. Therefore, appropriate preprocessing techniques were applied to address these missing values before using the data in the machine learning system. Overall, the dataset used in this project is a suitable choice for a big data project given its moderate size and manageable complexity.

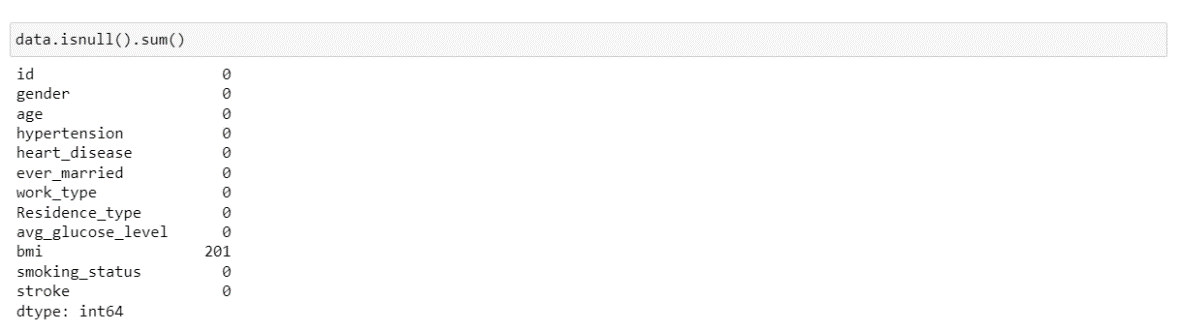
## Preprocessing The Data

The data utilized in this project was subjected to several preprocessing procedures, which are outlined below:

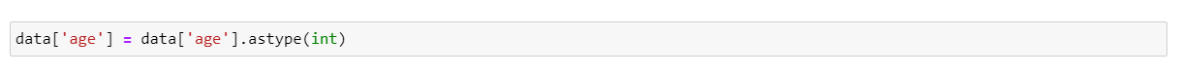
**Validating the Shape of the Loaded Dataset**

This process helps to ensure that the data has been loaded correctly without any issues such as missing values or incorrectly formatted data.

**Identifying Missing Data**

Identifying and handling missing values in the dataset is a crucial step in a machine learning project. It involves using the 'isnull()' function from the Pandas library to check for missing values and subsequently handling them by either removing the affected samples or imputing the missing values with a suitable value. By performing this step, the data quality is improved, and the accuracy and reliability of the machine learning model are enhanced.

**Transforming Column Data Types to Fit the Application**

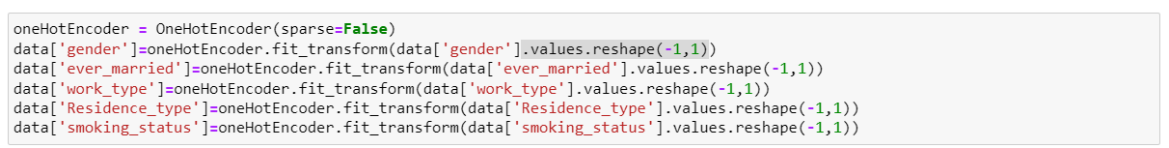
The step of converting data types of columns to the appropriate format is an essential part of data preprocessing in a machine learning project. This involves using functions such as 'astype()' to ensure accurate analysis and modeling of the data. By converting the data types of columns, the quality and reliability of the machine learning model can be improved.

**Dealing with Missing Data**

Filling the missing "bmi" feature with its mean value is a common technique that avoids introducing biases and preserves feature distribution. This approach ensured the completeness of the stroke dataset for machine learning modeling, allowing for accurate predictions and better model performance.

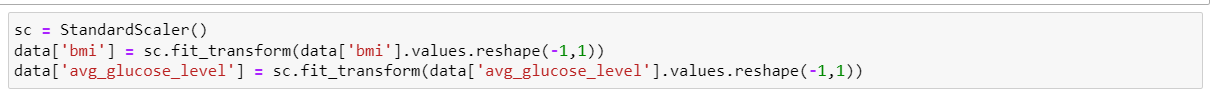
**Encoding String Data to Numeric Form**

Converted categorical data into numerical format for processing by machine learning algorithms, using techniques such as one-hot encoding or label encoding.



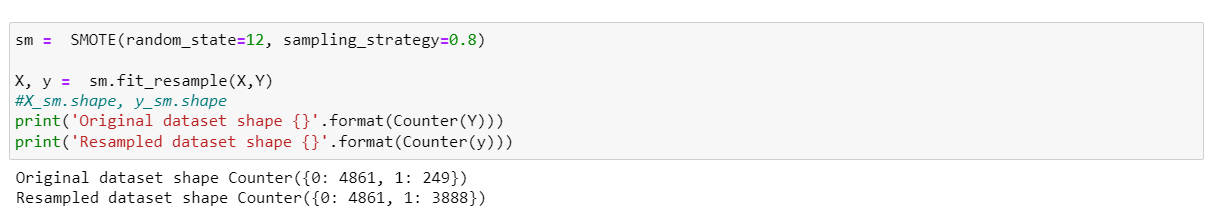
**Data Scaling**

"BMI" and "avg\_glucose\_level" features were scaled using StandardScaler to ensure they were on the same scale as other numerical features, and avoid biases arising from differing ranges. Scaling improved model accuracy and performance on the dataset.



**Balancing Dataset**

The SMOTE algorithm was applied to balance the stroke dataset, which initially had an imbalanced class distribution of 0 (4,861 instances) to 1 (249 instances). After SMOTE, the number of stroke cases increased to 3,888 due to the creation of 3,639 new synthetic instances. This approach helped mitigate the class imbalance issue and can lead to better classification accuracy in machine learning models.

****

# Architecture

The architecture for this project is:

Graphical user interface

Description automatically generated

The proposed architecture for the stroke prediction model includes several components. The first component involves pre-processing the data to ensure that it is suitable for use in the model. The pre-processed data is then split into training, validation, and testing sets using the train\_test\_split function from the scikit-learn library. The next component is the training of classification algorithms on the preprocessed data. Several classification algorithms such as logistic regression, decision trees, random forests, and support vector machines (SVMs) can be tried. After evaluating the performance of each algorithm using metrics such as accuracy, precision, recall, and F1 score, the best performing algorithm can be selected.

After the model has been trained, it is evaluated using appropriate metrics such as accuracy, precision, recall, and F1 score. This evaluation allows us to determine the effectiveness of the model and identify areas where it can be improved.

To cope with big data in the future, the architecture can be extended by incorporating big data technologies such as Hadoop and Spark. This would allow for distributed processing and storage of large datasets. To achieve fault tolerance and fast response times, techniques such as data partitioning, replication, and load balancing can be utilized.

Overall, this architecture provides a robust and scalable solution for stroke prediction, with the ability to handle large datasets and adapt to changes in technology.

# Visualization

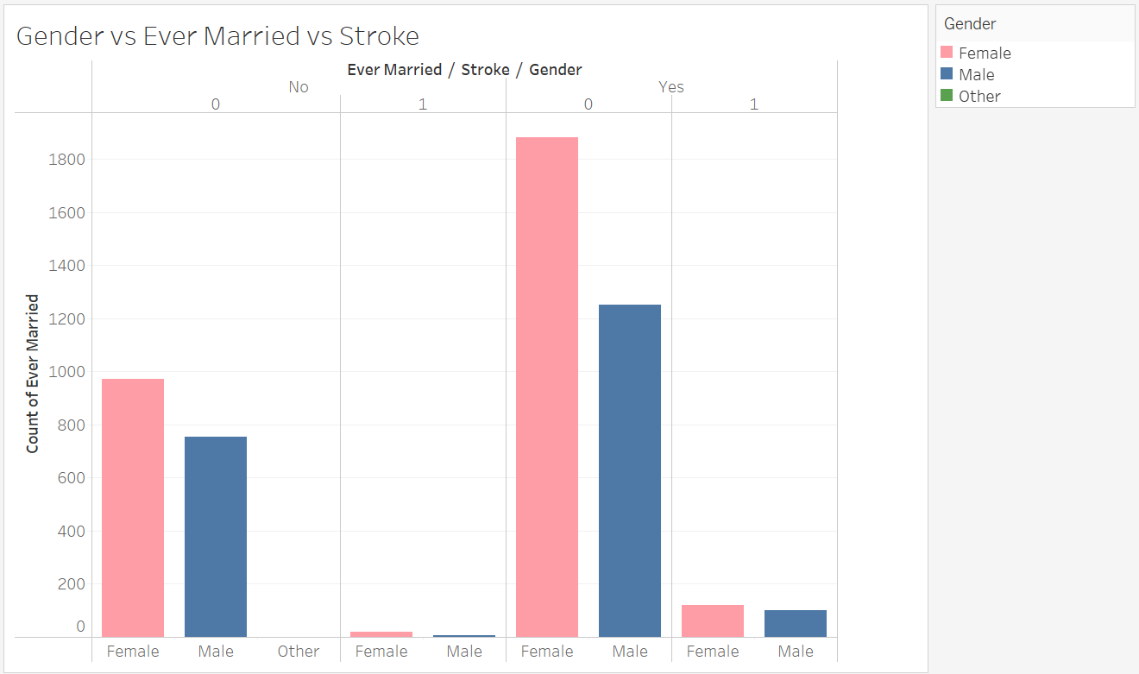
**Gender Plot**

The visualization plot summarizes the count of Gender for each Gender.

Chart, pie chart

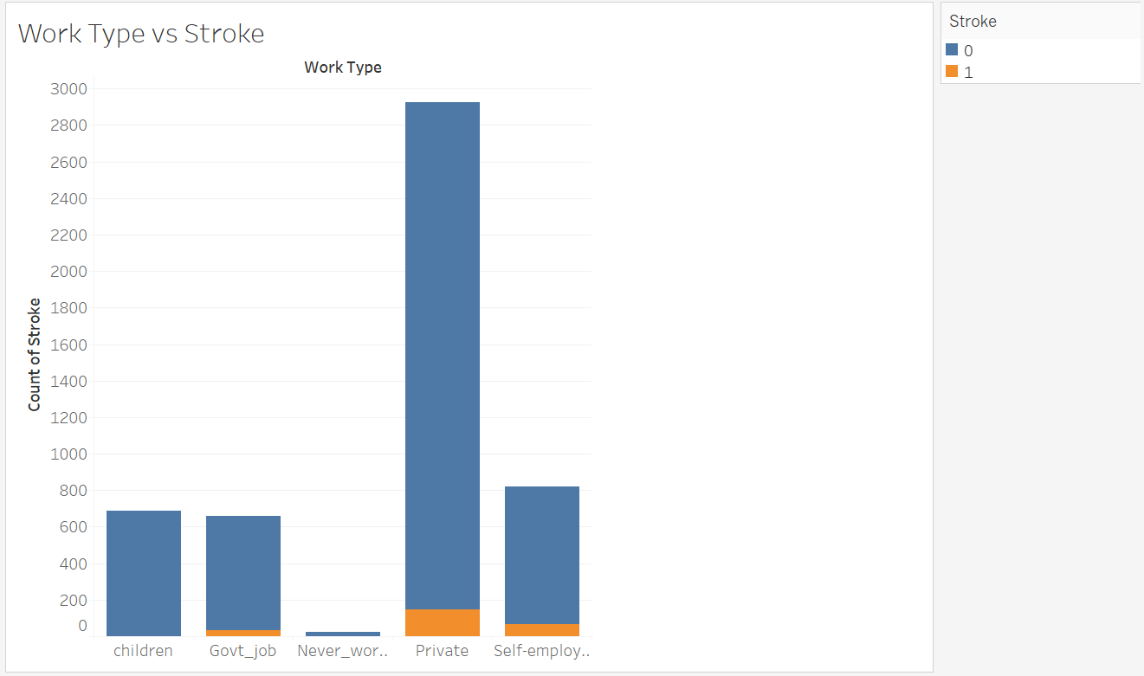
Description automatically generated

**Stroke by Gender and Ever Married**

The visualization plot summarizes the count of individuals based on their gender, ever married status, and the presence of stroke. It provides a clear representation of how many people have had a stroke or not, categorized by their gender and ever married status. This allows for easy comparison of the count of individuals with and without stroke across different gender and ever married status categories.

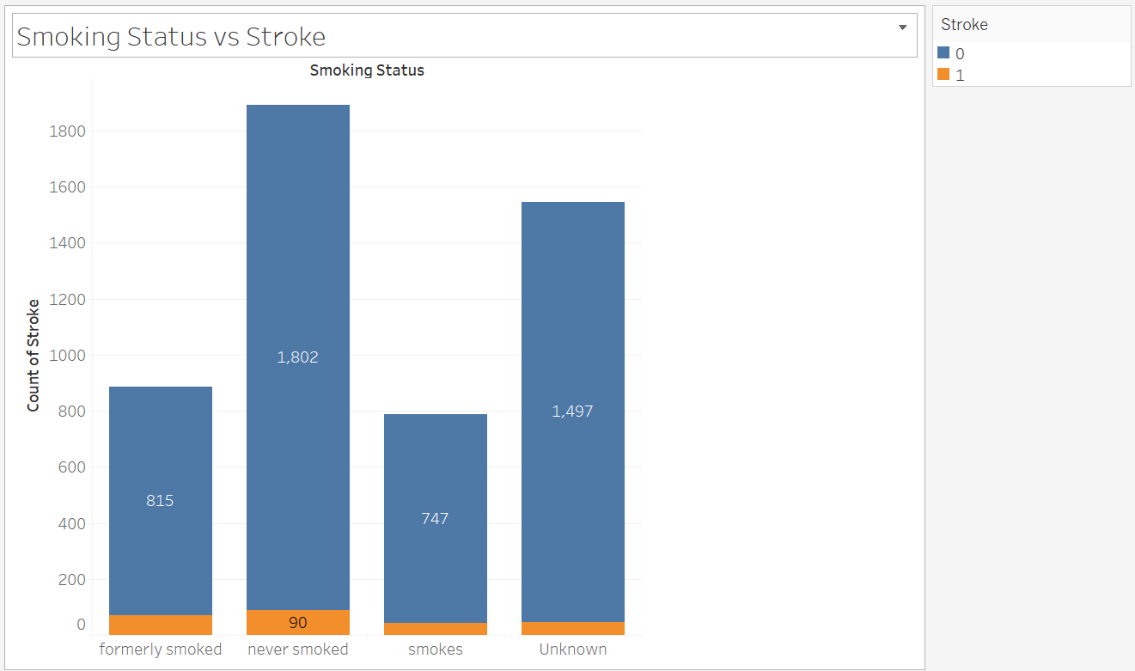
**Stroke by Work Type**

The visualization plot summarizes the count of individuals based on their work type and the presence of stroke. It provides a clear representation of how many people have had a stroke or not, categorized by their work type. This allows for easy comparison of the count of individuals with and without stroke across different work type categories.



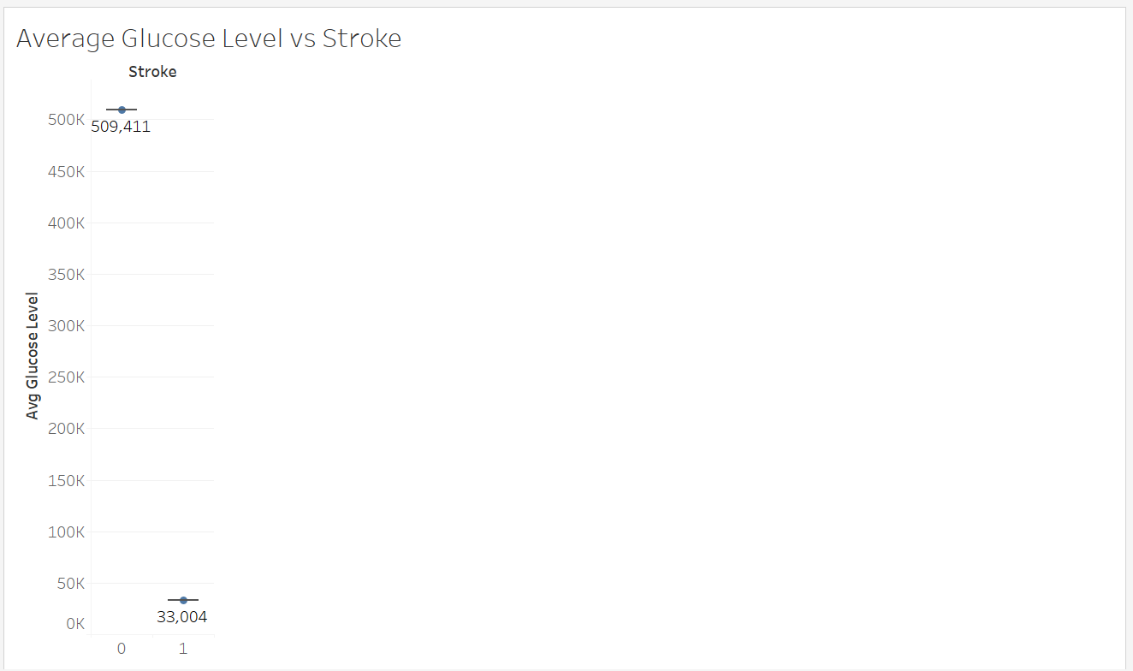
**Stroke by Smoking Status**

The visualization plot summarizes the count of individuals based on their smoking status and the presence of stroke. It provides a clear representation of how many people have had a stroke or not, categorized by their smoking status. This allows for easy comparison of the count of individuals with and without stroke across different smoking status categories.

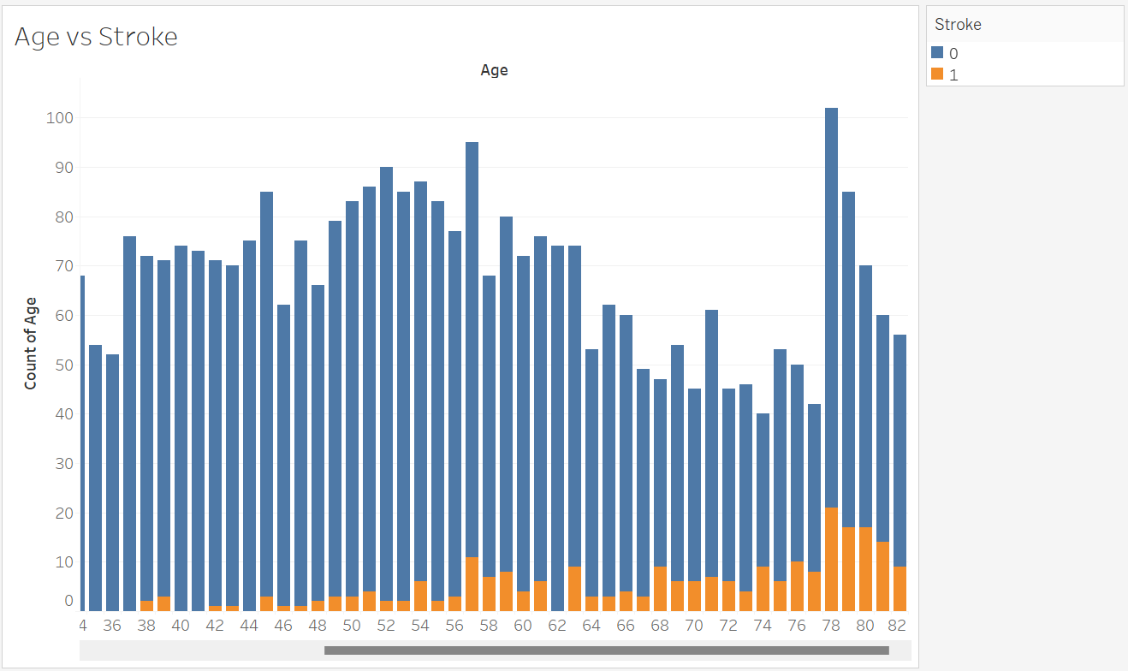


**Stroke by Average Glucose Level.**

The visualization plot summarizes the average glucose level for individuals based on the presence of stroke. It provides a clear representation of the mean glucose level for individuals who have or have not had a stroke, allowing for easy comparison of the average glucose levels between the two categories.

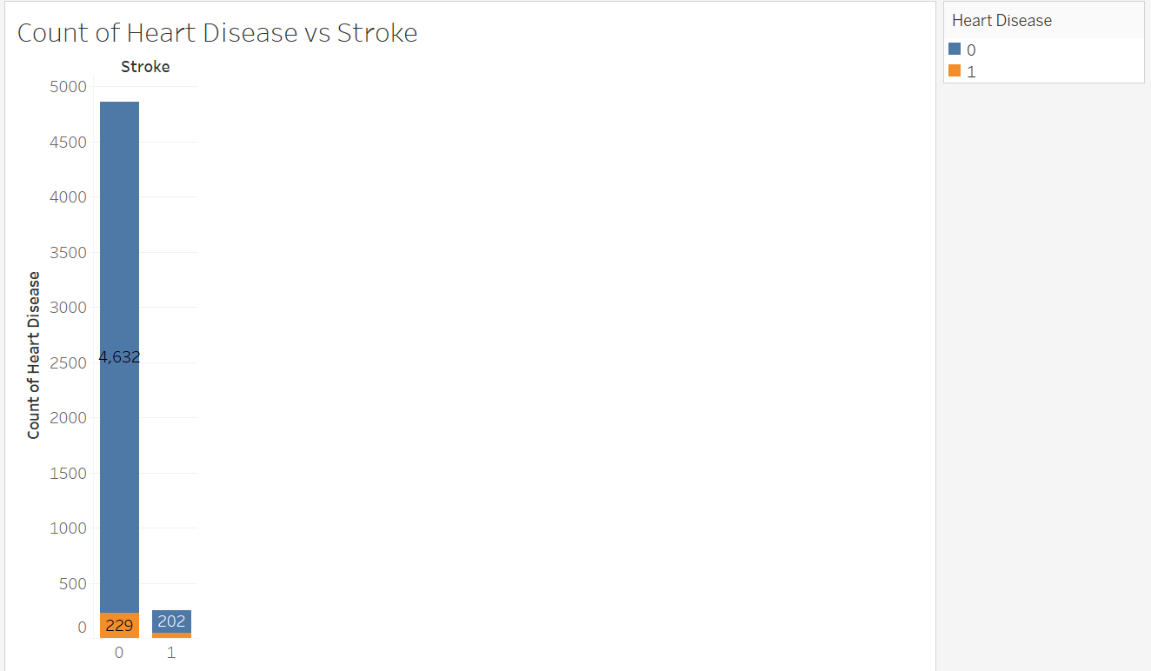


**Stroke Count by Age Group.**

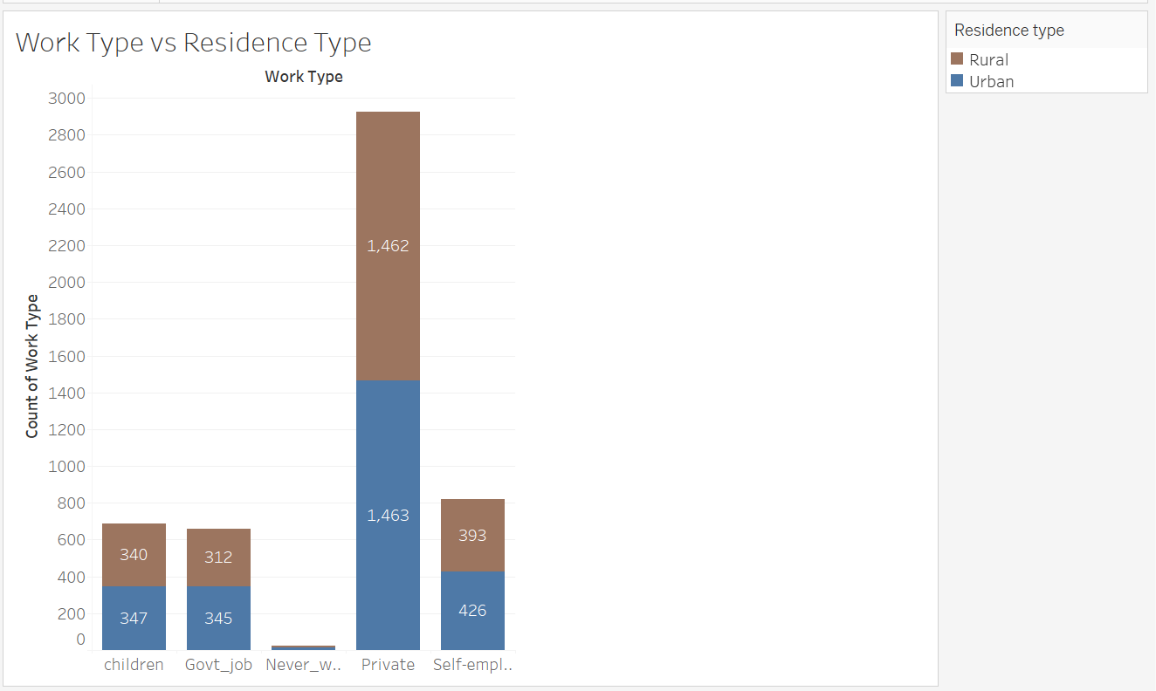
The visualization plot summarizes the count of individuals based on their age group and the presence of stroke. It provides a clear representation of the number of people in each age group who have or have not had a stroke, allowing for easy comparison of the count of individuals across different age groups.

**Stroke Count by Heart Disease.**

The visualization plot summarizes the count of individuals based on the presence of heart disease and stroke. It provides a clear representation of the number of people who have both conditions, either one of them, or neither, allowing for easy comparison of the count of individuals across different categories.

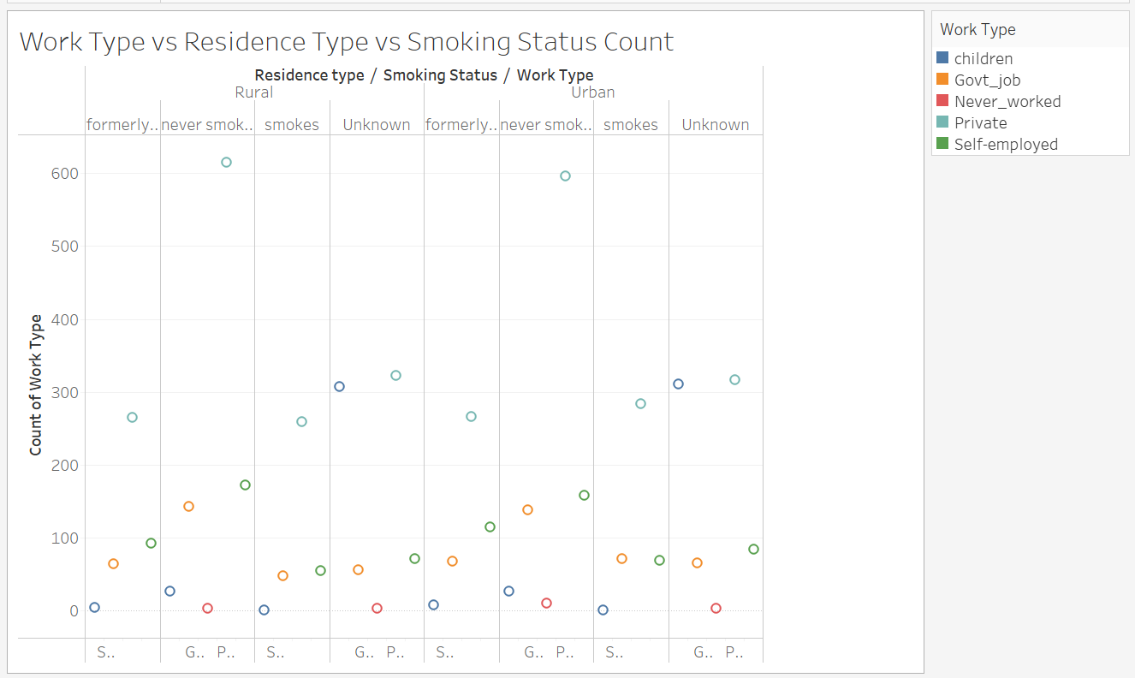


**Distribution of individuals by residence type and work type.**

The visualization plot summarizes the distribution of individuals based on their residence type and work type. It provides a clear representation of the proportion or percentage of individuals in each combination of residence type and work type, allowing for easy comparison of the distribution across different combinations.

**Count of Smoking Status by Residence Type and Work Type**

The visualization plot summarizes the count of individuals by their smoking status, residence type, and work type. It provides a clear representation of the distribution of smoking status across different combinations of residence and work types, allowing for easy comparison of the count of individuals belonging to each category.



# Machine Learning System Application-Methodology

**Data Splitting**

The data was split into training and testing sets using the train\_test\_split function from the scikit-learn library. A 80/20 split ratio was used, with 80% of the data for training and 20% for testing.



**Model Training**

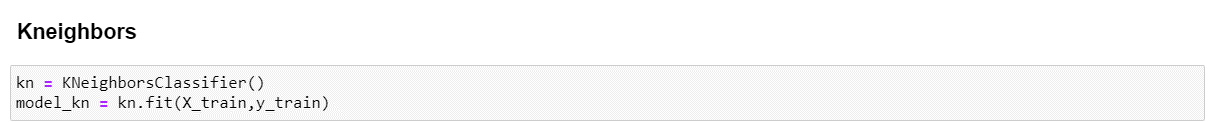
The model was trained using a classification algorithm on the 80% split of the data. Multiple algorithms, including naïve bayes, logistic regression, decision trees, random forests, and K-Nearest Neighbors (KNNs).











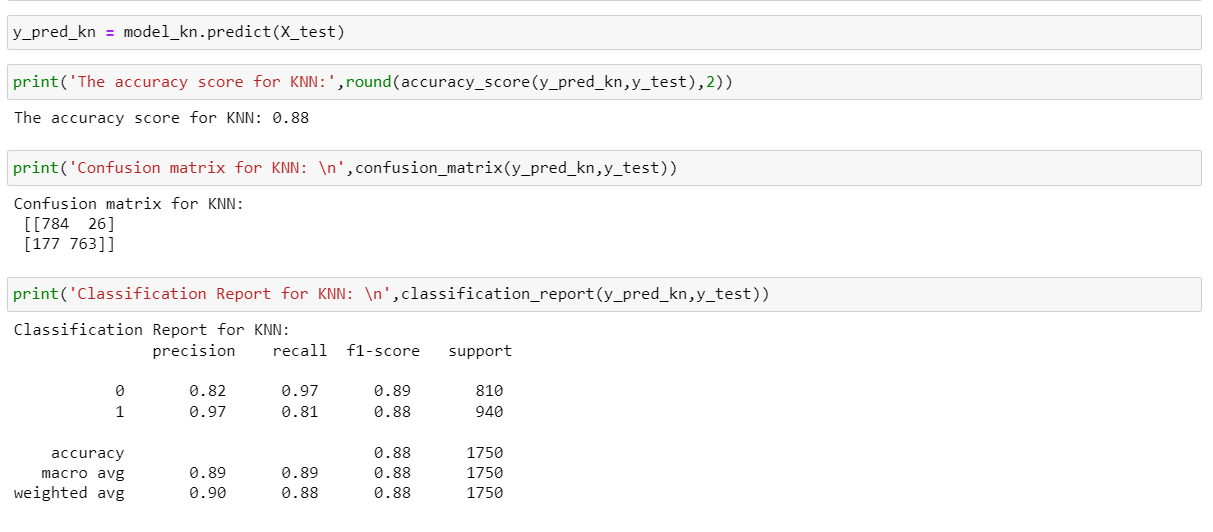
**Model Evaluation**

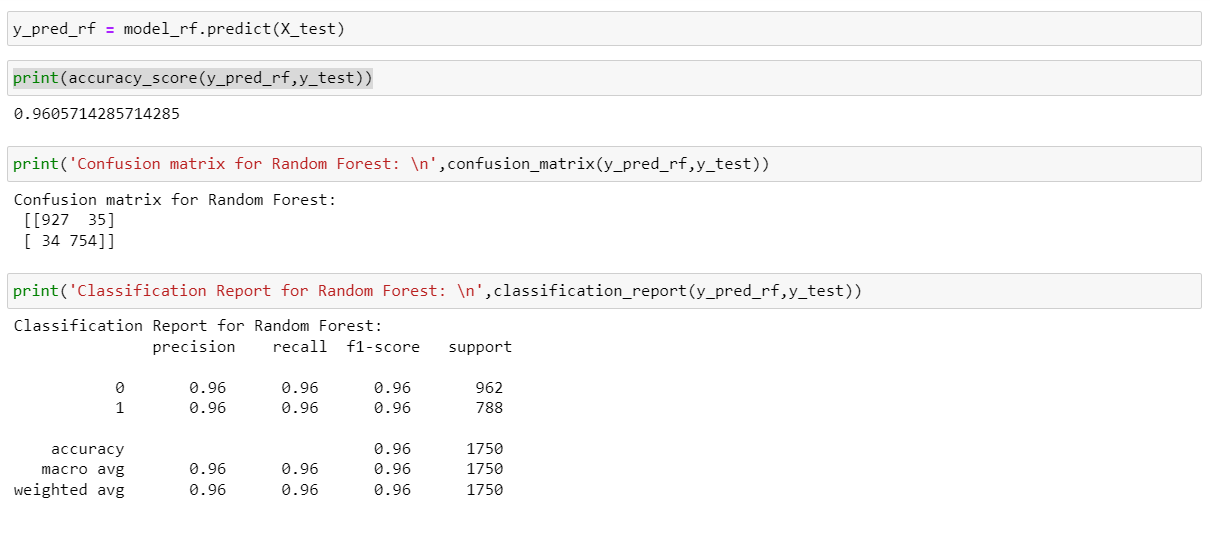
After testing the trained models on a 20% split of the data, appropriate evaluation metrics such as accuracy\_score, classification report, and confusion matrix were used to evaluate the models. Following the evaluation, the Random Forest model was determined to be the best model, with an accuracy of 96%.



Table

Description automatically generated with medium confidence





# Conclusion And Reflection

The aim of this project was to predict stroke using machine learning techniques. The dataset contained 5110 rows and 12 columns. The data preprocessing steps involved data cleaning, handling of missing values, and encoding of categorical variables. Machine learning algorithms, such as Decision Tree, Logistic Regression, Naive Bayes, and Random Forest, were trained on the preprocessed data. Random Forest had the highest accuracy score of 96%, making it the best model for stroke prediction. To handle the issue of data imbalance, the SMOTE technique was used, which improved the overall accuracy of the models.

Further work can be done to improve the accuracy of the model, such as using more advanced techniques like Deep Learning. Additionally, the dataset used in this study was relatively small, so using big data could be leveraged to increase the size of the dataset, which can lead to better prediction models. With the increasing availability of data, the use of big data can be a game-changer for healthcare analytics, particularly in stroke prediction. It can help identify new risk factors, improve accuracy, and enhance patient outcomes.

In conclusion, this project highlights the potential of machine learning in predicting stroke and the importance of leveraging big data in healthcare.