Academic Year	Module	Assessment Type
2024	Concepts of Technologies of AI	Academic Report

An End- to- End Machine Learning Project on Regression and Classification Task

Student Id : 2431328

Student Name : Ugesh Kc

Section : L5CG22

Module Leader : Mr. Siman Giri

Tutor : Mr. Ronit Shrestha

Submitted on : 2/11/2025

Classification Analysis Reports

Abstract

The purpose of the report is the prediction of the occurrence of a stroke using classification techniques.The dataset that was chosen exactly for this analysis is the Brain Stroke dataset which contains demographic and health-related attributes. The steps that were involved includes Exploratory Data Analysis (EDA), model building with the help of Logistic Regression and Random Forest Classifier, hyperparameter optimization, and some feature selection.

Key Results: The performance of the models was evaluated with the use of accuracy, precision, recall, and F1 score. The Random Forest model outperformed Logistic Regression with higher accuracy and recall which shows the capability of Random Forest model to handle complex relationships.

Conclusion: The classification models key insights includes age, glucose level, and hypertension. It performed good where the Random Forest Classifier achieved 89.4% and a higher F1-score (0.72) which outperforms Logistic Regression.

1. Introduction

1.1 Goal : The goal of the project is the prediction of Stroke which is target variable looking on their demographic and health attributes.

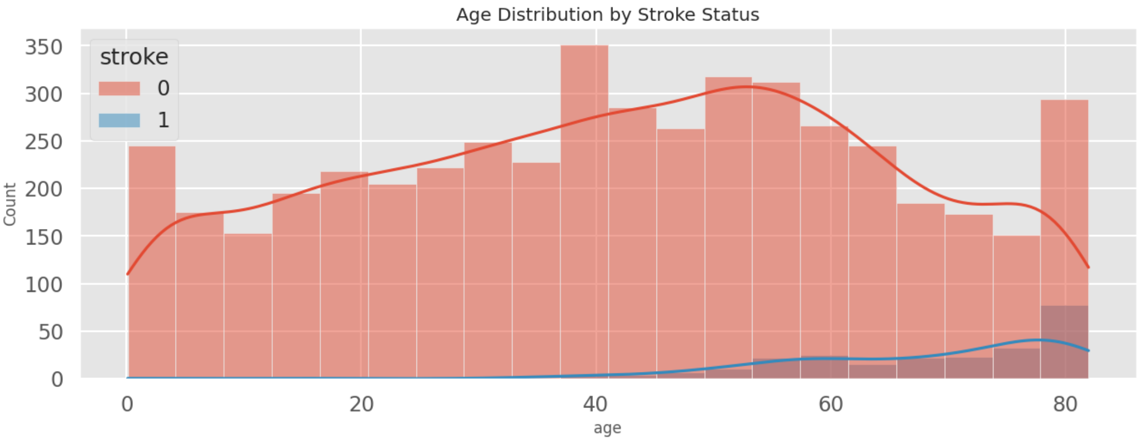
1.2 Dataset: The dataset in the analysis is the Brain Stroke dataset, which was gotten from Kaggle. It has features like age, gender, BMI, glucose level, and health conditions. This dataset matches with the UNSDG for Good Health and Well-Being (Goal 3).

1.3 Target : Objective of my analysis is to make a predictive classification model which predicts stroke looking at the features like age, gender, BMI, glucose level, and health conditions in dataset.

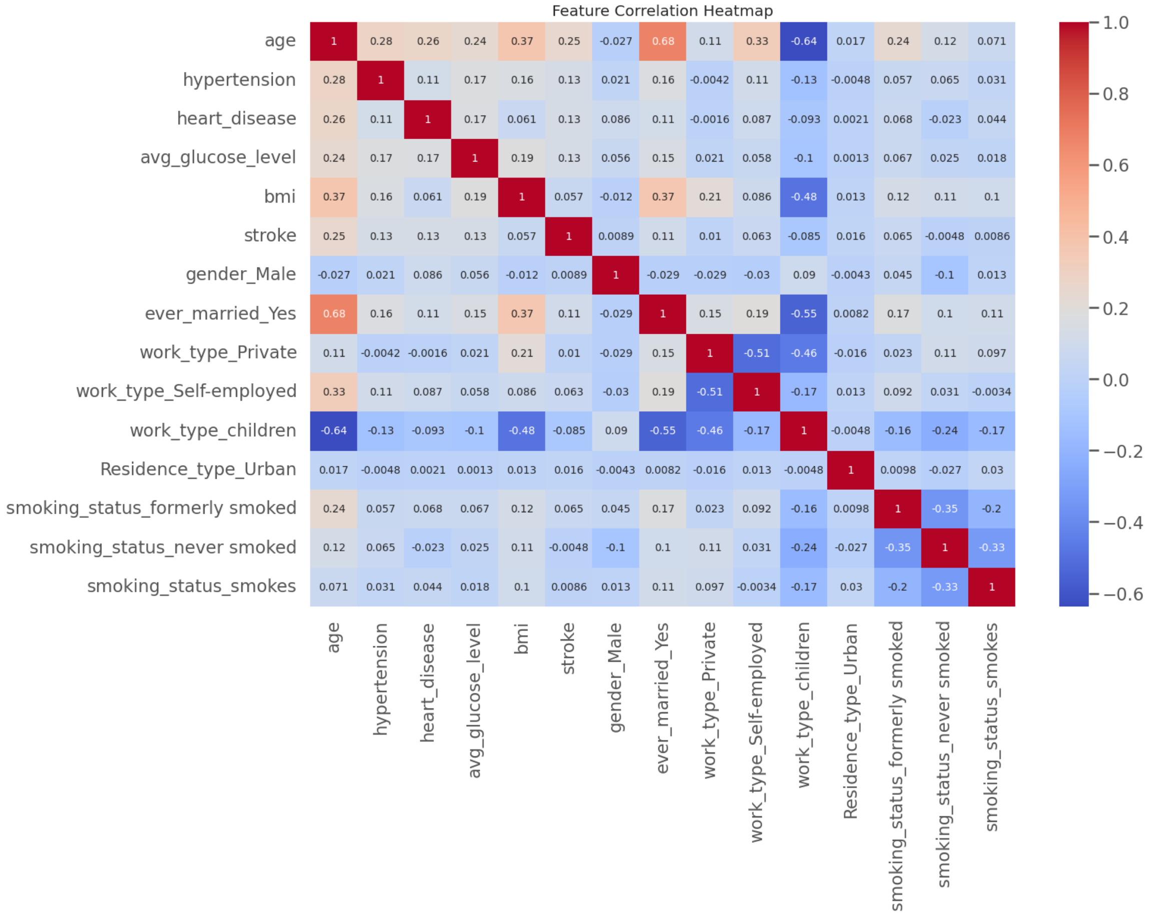
2. Methodology

2.1 Data Preprocessing: Before constructing the data were cleaned by handling values that were missed in the bmi column. Key features like glucose level and bmi were checked for outliers using boxplots. One-hot encoding was used to categorical features like gender and smoking_status. The dataset was highly imbalanced (only 4.98% stroke cases). SMOTE and class weight was used to balance the dataset.

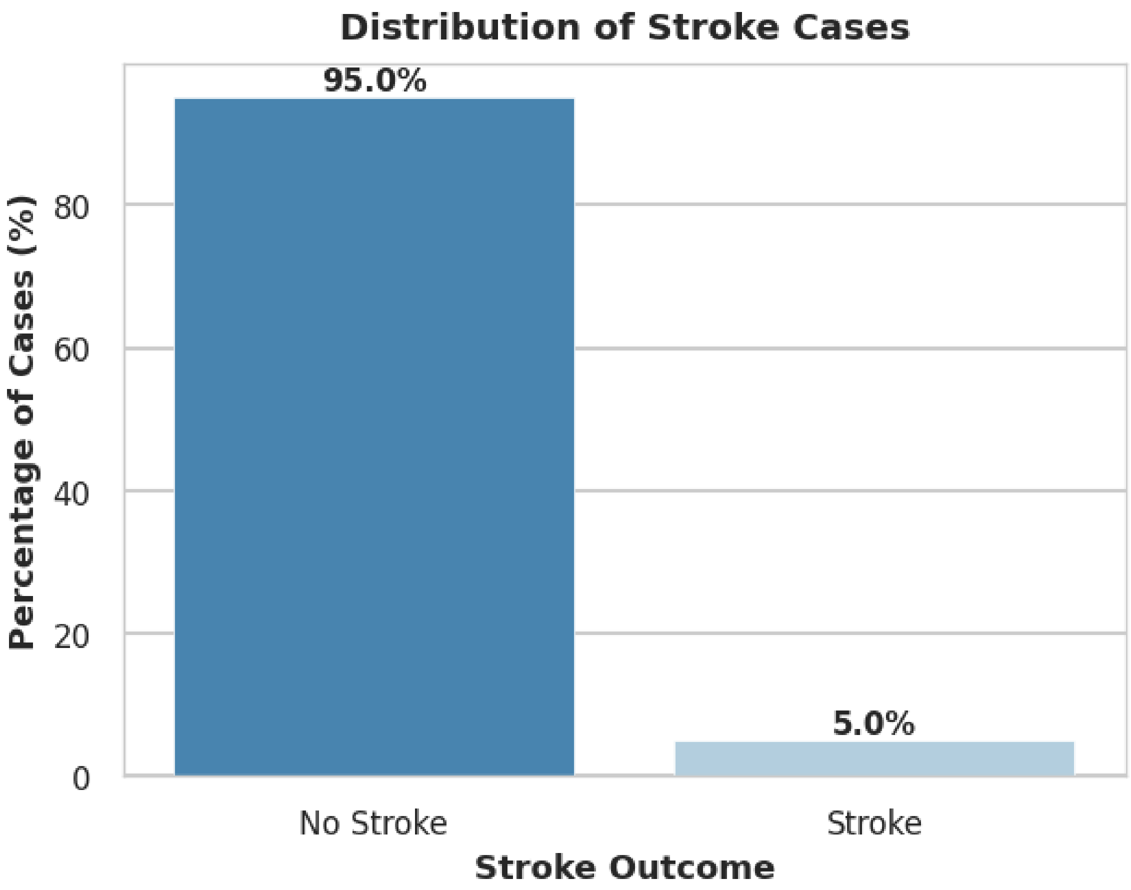
2.2 Exploratory Data Analysis (EDA):



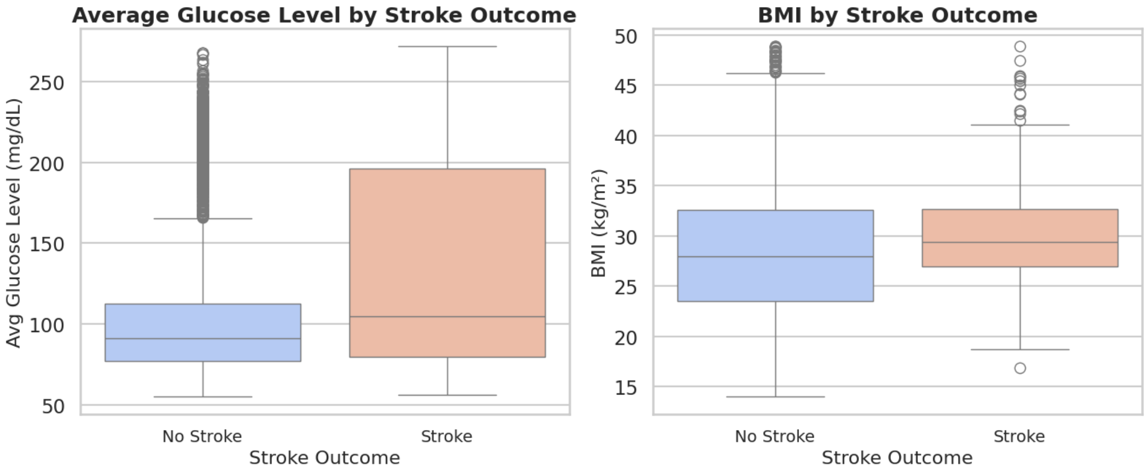
fig(a) Histogram of Age Distribution by stroke status



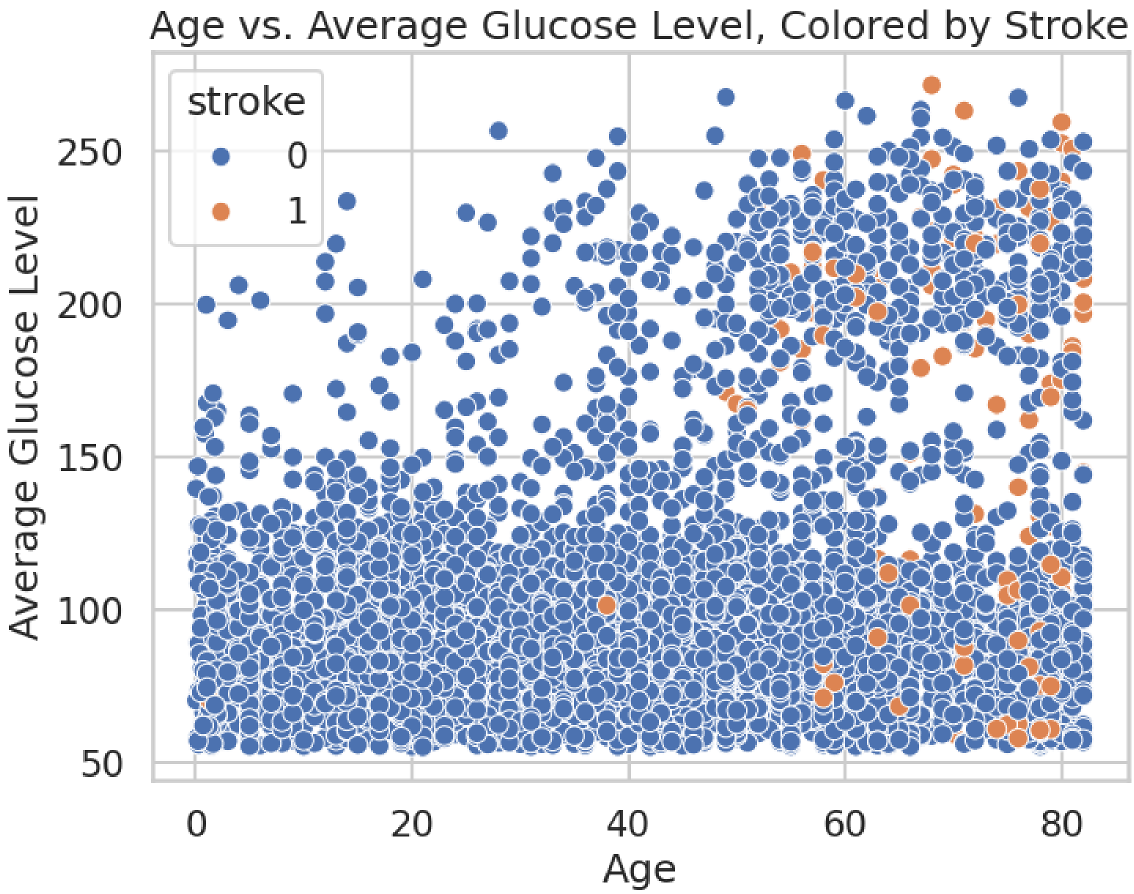
fig(b) Feature Correlation Heatmap



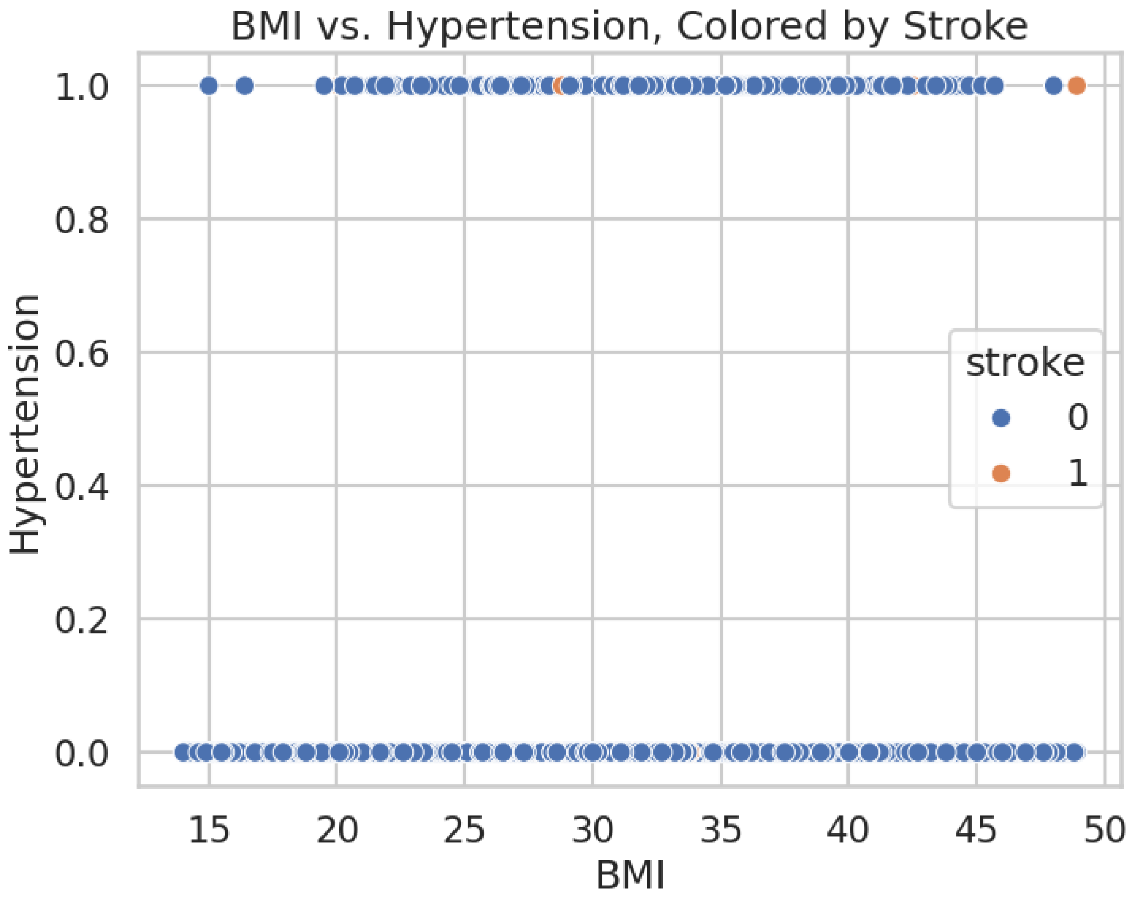
fig(c) bar chart showing distribution of Stroke Cases



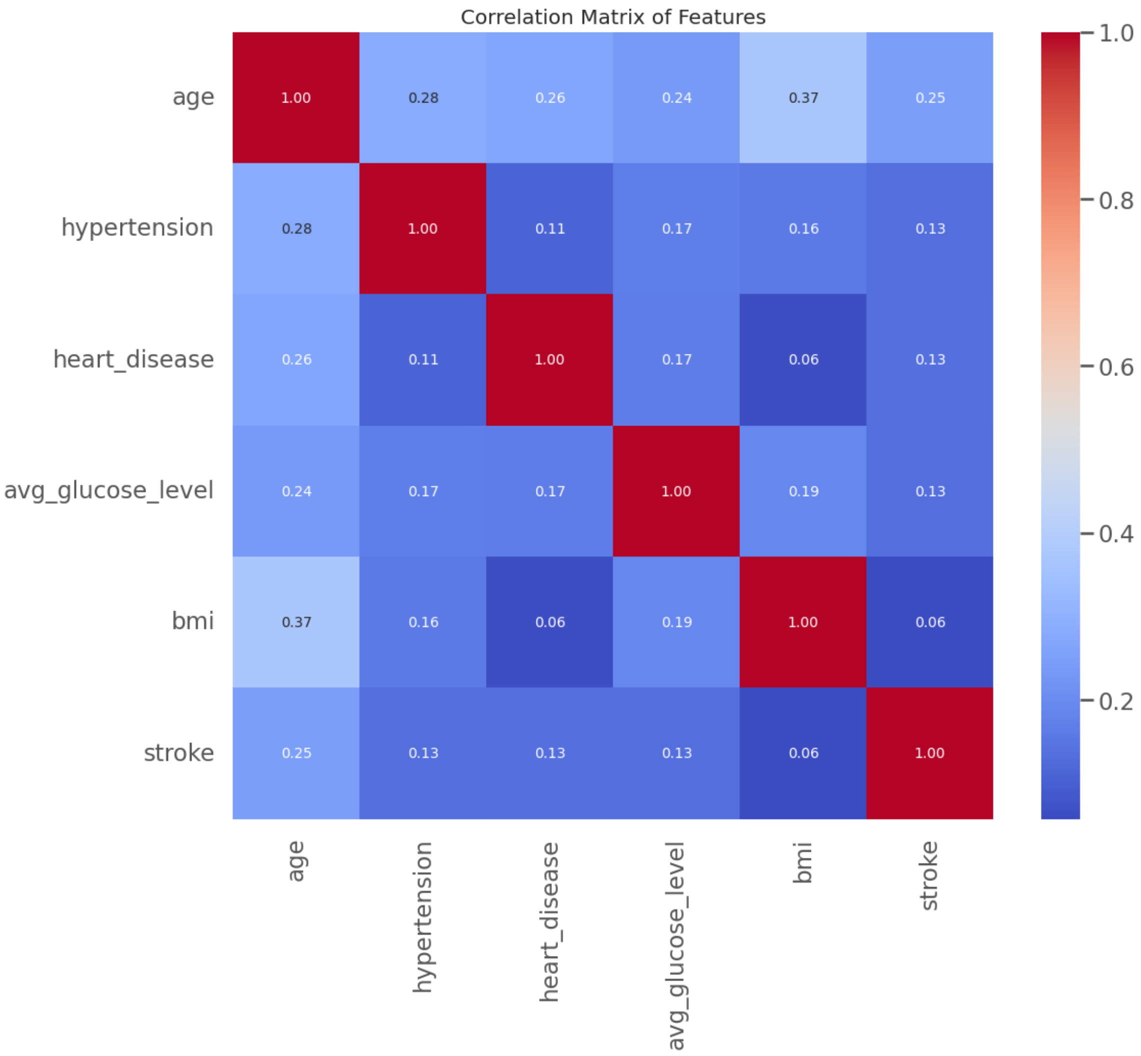
Fig(d)Box plot for Average Glucose Level by Stroke Outcome and BMI by Stroke Outcome



Fig(e)Scatter plot of ‘age’ vs ‘avg_glucose_level’



Fig(f) Scatter plot of BMI vs Hypertension



Fig(g) Correlation Matrix of Features

EDA was performed with use of histograms, bar charts, scatter plot, box plot and correlation matrices to learn the data in better way. Age appeared as a strong predictor since older people had the highest risk of strokes. Strokes were also positively linked with diabetes and hypertension. Bar graphs also displayed a severe class imbalance stroke (only 4.98% cases), which was improved using SMOTE. The first scatterplot (Age vs. Glucose Level) seeks to identify any trends or clusters that can potentially predict stroke risks with age and glucose level. The second one (BMI vs. Hypertension) tries to establish the factors' association with stroke. Some points are overlapping, implying that these two aspects by themselves are weak determinants. According to the boxplot of glucose, stroke patients have a higher median glucose level, compared to other patients. There is also an upward shift in the interquartile range, which indicates a possible predictive relationship. The stroke and nongroup distributions also overlap in the BMI boxplot, implying that the predictive value of BMI obesity is less than that of glucose is how levels. Correlation matrix suggested that while BMI stroke did not have significant direct correlation, it could potentially be associated with other factors. Categorical features like smoking or work type were encoded for the model training. These findings assisted in feature selection and model building.

2.3 Building of Model

For this project 2 classification models were used which are Logistic Regression and Random Forest Classifier. The datasets was allocated into an 80 percentage training set and a 20% test set. With the data preprocessing stage, the missing values were imputed, categorical variable was one-hot encoded, and the dataset was balanced with SMOTE. The data was then fitted to the models, starting with a baseline model of Logistic Regression and later incorporating non-linear Random Forests as an advanced model.

2.4 Model Evaluation

Results:

a. For logistic regression: 83.2% was the accuracy and 0.56 was the F1 score (lower recall, a large amountof stroke cases were not detected).

b. The Random Forest algorithm had a better precision value which, in this case, was 0.75 which means that there were fewer strokes wrongly predicted on given patients.

c. For random forest classifier: 89.4% was the accuracy and 0.72 F1 score (better recall, better strokes detection).

This is why Random Forest was much more effective at classification, it captured the non-linear relationships much better.

2.5 Optimization of Hyper-Parameter

For the improvement of the performance of the model, hyper-parameter optimization was used using GridSearchCV:

- Logistic Regression**:** Optimal parameter C = 1.0 (regularization strength).
- Random Forest**:** optimal parameters n_estimators = 200, max_depth = 10 which leads to higher recall and accuracy.

2.6 Selection of Feature

Feature selection was done by the use of Recursive Feature Elimination for the identification of important features for the prediction of the target variable(stroke). Important features selected were:

1. Age
2. Glucose Level

3. Hypertension
4. Heart Disease
5. BMI

3. Conclusion

Key Findings: The Random Forest Classifier was a better model achieving 89.4% accuracy and a higher F1-score than Logistic Regression. The model highlights the importance of age, glucose level, and hypertension to predict stroke.

Challenges: While doing the project several challenges were encountered, including class imbalance which made my recall and f1 score 0, feature selection difficulties. As RFE returns Boolean or integer indexes not column names we cannot use X_train[selected_features] directly. we must use .iloc[:, selected_features]

Future Work: for the improvement of the model future work can involve advanced models which are XGBoost or Neural Networks, and additional health-related features for improved performance.

4. Discussions

4.1 Performance of model

- Random Forest Classifier proved superior to the Logistic Regression Model as it performed better in the prediction of stroke cases. Its precision, recall, and F1 score average were higher than those of the Logistic Regression Model(LGM) which makes it model for this classification problem.

4.2 Impacts of Hyperparameter Tuning and Feature Selection

- The precision of the model was greatly improved through hyperparameter tuning, which improved recall as well.

- Selection of relevant features made it possible to remove irrelevant factors which lead to lower complexity while proving important for predictive modeling.

4.3 Interpretation of Results

- People aged above with elevated glucose, hypertensive, and heart dsease had higher chances of stroke, which is consistent with the rest of medical literature.

- The classification model was able to pick notable factors that determine the incidence of stroke correctly which attests to its robustness.

4.4 Limitations

The datasets were sufficient for model training but were, unbalanced and therefore SMOTE was required for effective detection of stroke cases.

Some features like BMI reflect low variance with stroke incidence which makes these measures weakly predictive.

4.5 Suggestion for Future Research

- Model would have a benefit from additional datasets for improved generalization.

- Consider new models like deep learning Neural Networks for potentially improved effectiveness.

- Expand to longitudinal health data (i.e. history of illnesses)