Stroke Analysis and Prediction

-Harshitha Chollangi

Overview

- Introduction
- Problem Description
- Algorithms considered
- Dataset Description and Analysis
- Performance Analysis
- Conclusion and Future Scope

Introduction

01

According to the World Health Organization (WHO), stroke is the greatest cause of death and disability globally.

02

Stroke, the 2nd leading global cause of death, prompts the need for predictive analytics.

03

The goal is to build a robust stroke prediction model to enhance clinical decision-making.

04

The dataset encompasses 5110 observations with 12 attributes capturing key patient information.

05

Input parameters include gender, age, diseases, and smoking status to predict stroke probability.

Problem Description

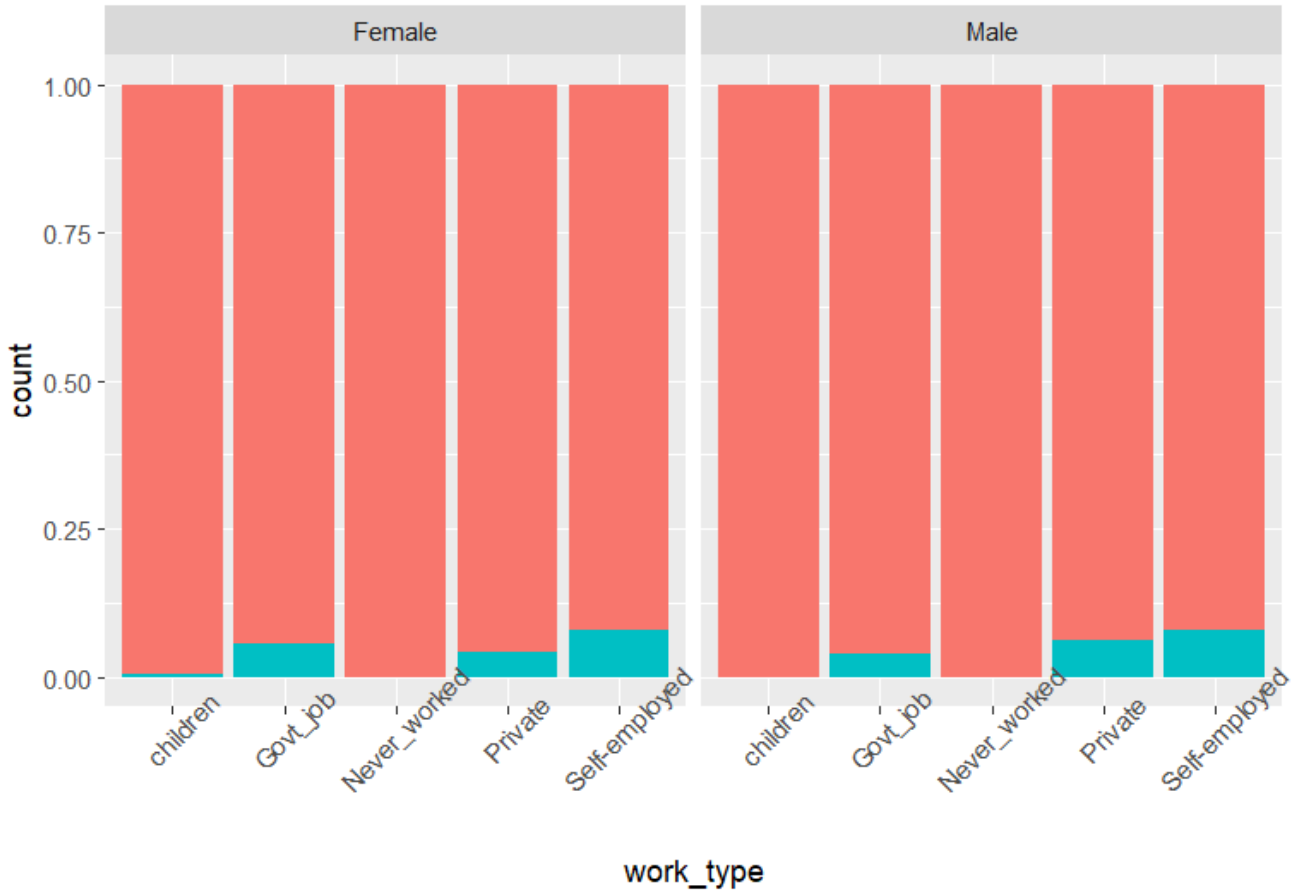
- Early recognition of the various warning signs of a stroke can help reduce the severity of the stroke.
- Stroke can be induced by factors such as
    1. Smoking
    2. Drinking
    3. Body Mass Index (BMI)
    4. Average Glucose Level and
    5. Heart Problems.

Algorithms

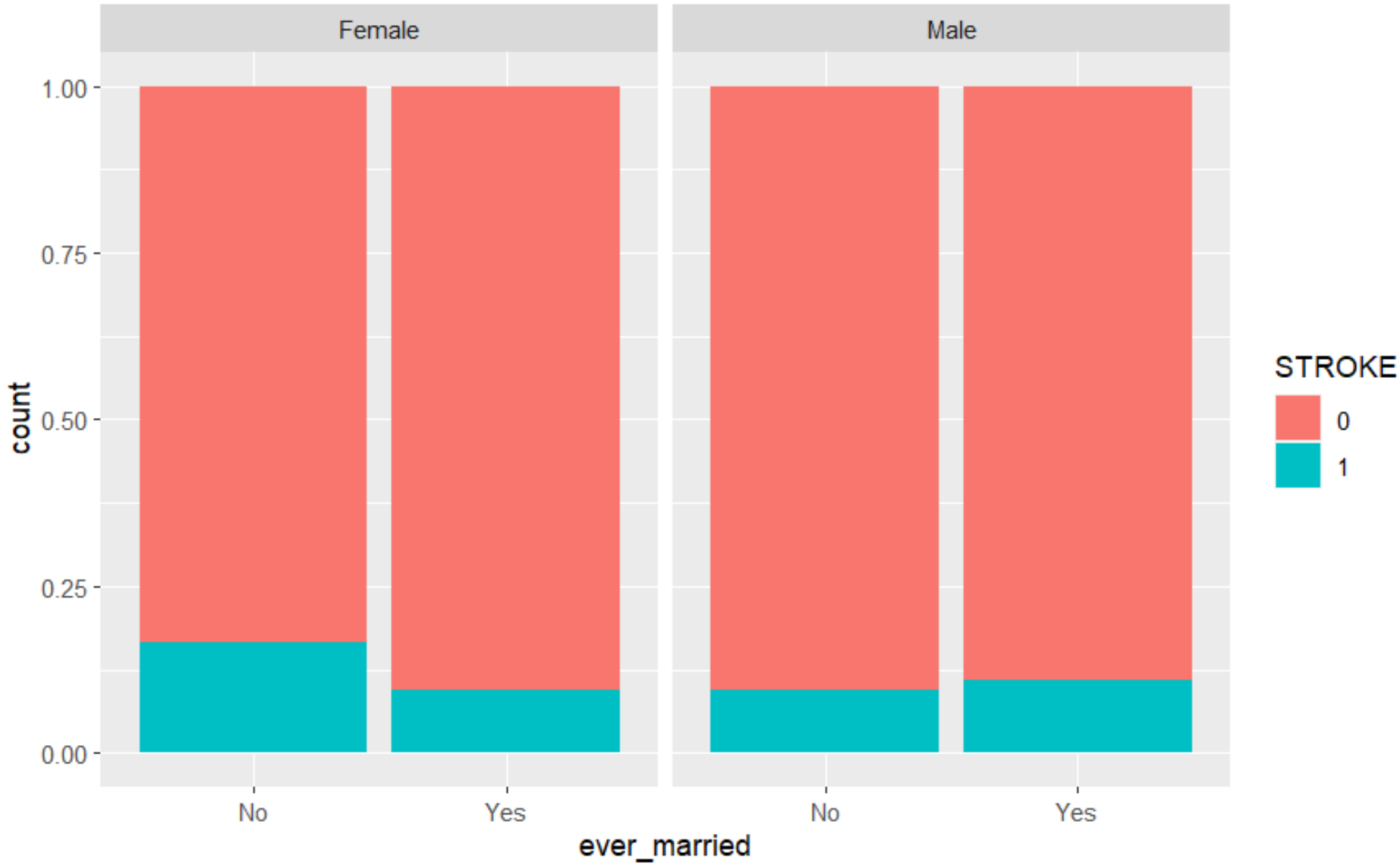
- Different machine learning models have been adapted to predict stroke
    1. Linear Model (including multilinear model, poly model, improved LM)
    2. Random forest model
    3. logistic regression model
    4. Naive Bayes Classification
    5. K-Nearest Neighbor classification
    6. Linear SVM*
    7. Simple Decision Tree Modelling*
    8. XGBoost

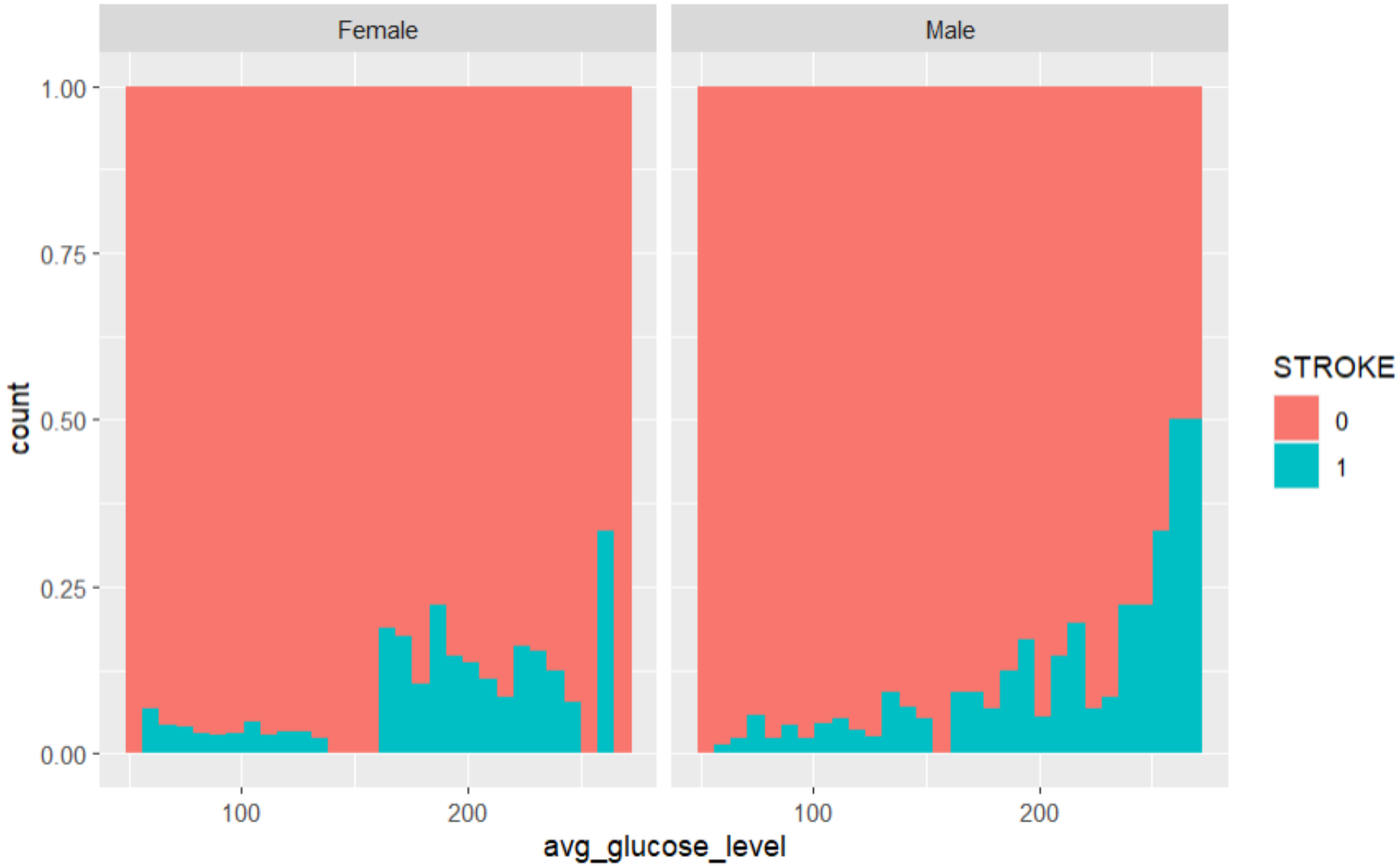
Dataset Description

S.NO	Feature	Type
1	Gender	Categorical
2	Age	Numeric
3	Has Hypertension	Binary
4	Has Heart Disease	Binary
5	Is married	Binary
6	Work Type	Categorical
7	Residence Type	Categorical
8	AVG Glucose	Numeric
9	BMI	Numeric
10	Smoking habits	Categorical
11	Stroke	Binary

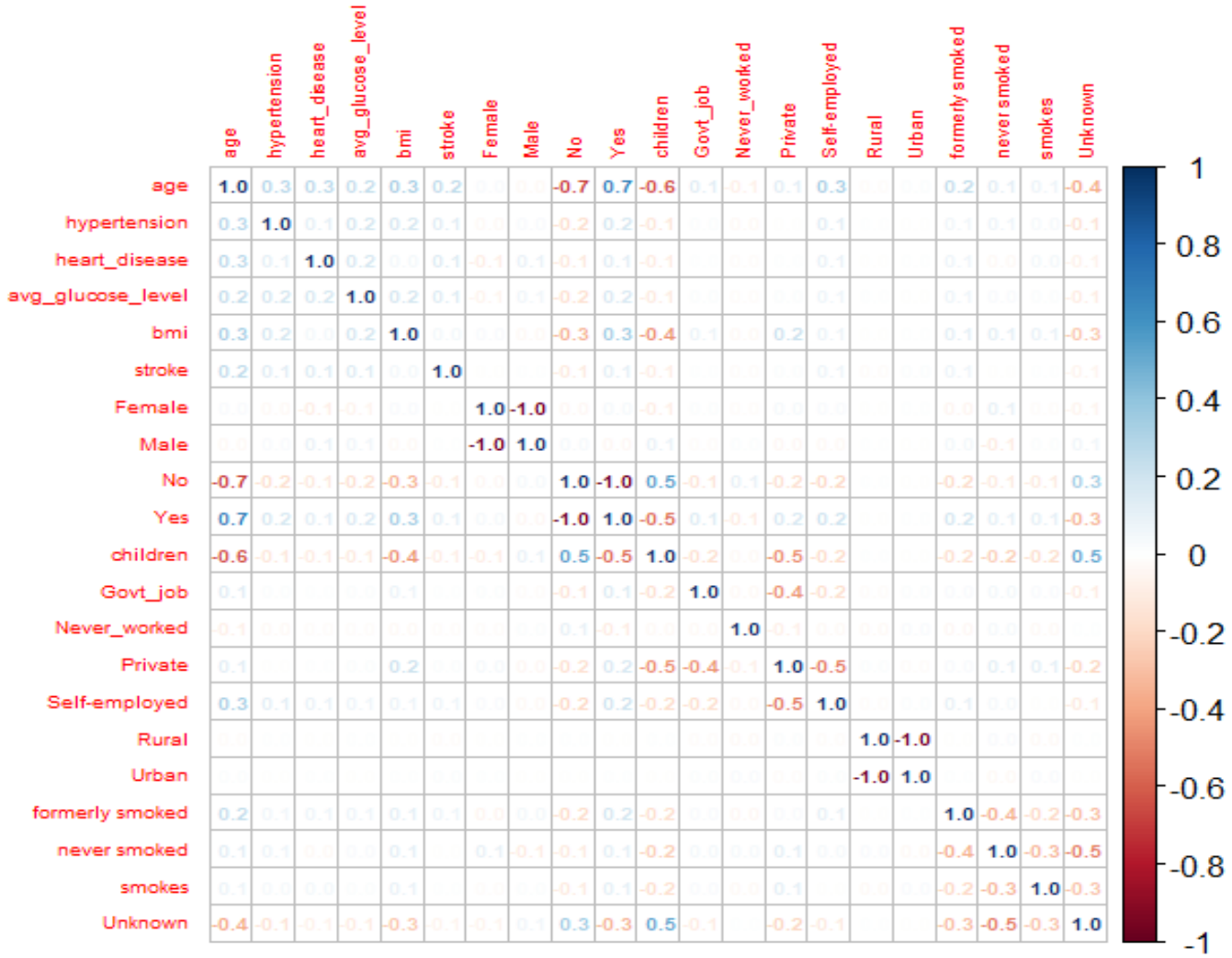


Self-employed people have higher stroke percentage than people working on private sector and governmental sector respectively, this means low security jobs have higher impact on having stroke.


Female
Male
1.00
0.75
0.50
0.25
0.00
count
No
Yes
No
Yes
ever_married
STROKE
0
1


Female
Male
1.00
0.75
0.50
0.25
0.00
count
100
200
100
200
avg_glucose_level
STROKE
0
1

Correlation matrix (Post Pre-Processing)



Model Metrics

Model	Accuracy	Recall	Precision	F1_Score	AUC	RMSE
Linear	0.7903107	0.1571730	0.7198068	1	0.5693479	0.4579184
Poly	0.8235870	0.1787500	0.6908213	1	0.5796397	0.4200155
GLMNet	0.8245657	0.1772152	0.6763285	1	0.5784468	0.4188488

Performance Results

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.9559902	0.37500000	0.05769231	0.10000000
KNN	0.9250204	0.02380952	0.01923077	0.02127660
SVM	0.9576202	0.50000000	0.01923077	0.03703704
Naive Bayes	0.5607172	0.07652174	0.84615385	0.14035088
Decision Tree	0.9576202	0.95762021	1.00000000	0.97835137
Random Forest	0.9437653	0.16000000	0.07692308	0.10389610
XGBoost	0.9568052	0.95758564	0.99914894	0.97792586

Measures the overall correctness of the model's predictions. The models with the highest accuracy are Logistic Regression (0.956), SVM (0.958), and XGBoost (0.957).

https://harshithachollangi.shinyapps.io/AppliedDataMiningProjectR/

Conclusion

- **Age and Stroke Risk:**

  Stroke percentage increases with age and peaks near the end of the 70s. Notably, there are no observations of males having a stroke below the age of 40, while several females experience strokes before reaching their 40s, including cases around 30, 15, and even below 1 year old.

- **Gender Disparities:**

  The dataset reveals a gender-related pattern where hypertension has a slightly higher effect on females than males. Additionally, married individuals, particularly females, seem to have a lower risk of stroke compared to unmarried individuals. The impact of heart disease is similar for both genders.

- **Occupation and Stroke Risk:**

  Self-employed individuals have a higher stroke percentage compared to those working in the private and governmental sectors. This suggests that occupations with lower job security may have a higher impact on stroke risk.

- **Useful Classification:**

  The classification based on gender appears to be particularly useful in understanding stroke risk factors. The observations regarding age, hypertension, marital status, and occupation also provide valuable insights into potential predictors of stroke.

Future Scope for Improvement:

Refinement of Age Analysis:

Further investigation into the specific age ranges associated with increased stroke risk could enhance the model's precision. This might involve creating age brackets or exploring nonlinear relationships between age and stroke risk.

Exploration of Lifestyle Factors:

Consider incorporating additional lifestyle factors, such as diet, physical activity, and stress levels, to provide a more comprehensive understanding of stroke risk. This could improve the model's predictive accuracy and contribute to more targeted preventive measures.