# Stroke Prediction Machine Learning Project

## 1. Introduction

Stroke is one of the leading causes of death and long-term disability. Early prediction and diagnosis are crucial for effective treatment and prevention. In this project, I worked with a public dataset to build and compare several machine learning models that estimate the likelihood of a stroke based on patient information.

## 2. Data Processing and Exploratory Data Analysis (EDA)

The dataset contains information about patients, such as age, presence of hypertension or heart disease, marital status, work type, glucose levels, BMI, and whether they had a stroke.

### 2.1 Data Cleaning

The `id` column was dropped as it had no predictive value.

1 row where gender was labeled as "Other" was removed. This data point was not part of the

Missing values in the BMI column were filled using the median.

### 2.2 Encoding Categorical Data

- Categorical columns were converted into numerical format using one-hot encoding. This helps machine learning models interpret the data correctly.

### 2.3 EDA Insights

Exploratory analysis helped reveal patterns and relationships in the data:

The dataset is highly imbalanced, with only about 5% of entries representing stroke cases.

Stroke patients tended to be significantly older; the median age among stroke cases was higher than that of non-stroke cases.

Patients with strokes also showed higher average glucose levels and BMI values. Boxplots and histograms for these variables revealed a clear rightward shift in distributions for stroke cases.

Among categorical variables, married individuals and those with private or self-employed work types showed slightly higher stroke rates based on countplots.

These findings helped guide the next step—feature engineering.

## 3. Feature Engineering

Based on the patterns observed during EDA, I created new features to better capture interactions between important variables:

- `age\_glucose`: combines age and glucose level

- `age\_bmi`: combines age and BMI

- `risk\_score`: adds the binary indicators for hypertension and heart disease

- `bmi\_glucose`: combines BMI and glucose

The idea was to help the models pick up on more complex patterns that individual features might not show clearly on their own.

## 4. Feature Selection

To keep only the most relevant information, I used `SelectKBest` with an ANOVA F-test to choose the top 15 features. This step helps reduce noise and speeds up model training.

## 5. Modeling and Evaluation

### 5.1 Models Used

- Logistic Regression (with class weights to handle imbalance)

- Random Forest (with parameter tuning)

- XGBoost (adjusted for class imbalance)

- Neural Network (2 hidden layers, dropout, and Adam optimizer)

### 5.2 Dealing with Imbalanced Data

- Logistic Regression: used `class\_weight='balanced'`

- XGBoost: used `scale\_pos\_weight`

### 5.3 Evaluation Method

- Models were evaluated using the F1-score, which balances precision and recall.

- A threshold was optimized for each model based on the precision-recall curve.

### 5.4 Results

A graph with blue squares

Description automatically generated

Across all models, recall was higher than precision. This means the models were better at identifying patients who did have strokes than avoiding false alarms. In medical prediction, this is usually preferred because missing a real case can be more dangerous than a false positive.

Logistic Regression came out on top with Neural Network showing the best F1 scores. However, the Neural Network tends to fluctuate more between runs, making Logistic Regression a more consistent and interpretable choice.

## 6. Model Interpretation

To understand how the models were making their predictions, I used SHAP (SHapley Additive Explanations):

- XGBoost: SHAP values showed that age, glucose, and BMI were among the most influential features.

- Logistic Regression: SHAP was successfully applied, confirming that the same features—age, glucose, and BMI—played a key role in predictions. This supported the usefulness of the engineered features.

## 7. Testing the Impact of Feature Engineering

To check whether the new features made a difference, I trained models using the original dataset (without engineered features) and compared the results to those using the enhanced dataset.

A graph of green squares

Description automatically generated

The models performed better when the extra features were included, especially Logistic Regression.

## 8. Conclusion

This project showed that Logistic Regression, a relatively simple model, delivered the strongest and most stable performance on this dataset. While more complex models like XGBoost and Neural Networks were also tested, carefully designed feature engineering and threshold tuning had a greater impact than model complexity.

SHAP analysis confirmed that key features such as age, glucose levels, and BMI played a central role in the model's decision-making, enhancing interpretability. All models tended to have higher recall than precision, indicating their stronger ability to detect actual stroke cases — a valuable property in medical settings.

However, despite being the best performer, Logistic Regression still produced a modest F1 score, which limits its reliability as a clinical decision-making tool. This suggests that additional research and feature enrichment — possibly including genetic traits or more detailed medical histories — would be necessary to improve predictive accuracy.