

Stroke Prediction Using Machine Learning: Identifying Key Risk Factors

Project Overview

Stroke is one of the leading causes of mortality and long-term disability worldwide, making early identification of high-risk individuals critical. This study investigates how machine learning (ML) can enhance stroke risk prediction in real-world scenarios characterized by limited and imbalanced data. Using models such as Random Forest, K-Nearest Neighbors, Decision Tree, and Logistic Regression, the study classifies stroke cases and identifies key risk factors, providing insights that can support clinical decision-making and early preventive interventions.

Introduction

Stroke, caused by an interruption of blood supply to the brain, remains a major public health challenge globally. According to the World Health Organization (WHO), strokes account for roughly 11% of all deaths. The clinical and societal burden of stroke is significant, as survivors often face permanent disabilities that require long-term care. Early detection of individuals at high risk is therefore crucial for preventing strokes and improving patient outcomes.

Machine learning has shown promise in healthcare for its ability to analyse complex datasets and uncover patterns that may not be immediately apparent through traditional statistical methods. In the context of stroke prediction, ML can help identify individuals at high risk by leveraging demographic, clinical, and lifestyle data. This study explores the predictive power of ML models on a real-world dataset while addressing challenges such as data imbalance and feature selection.

Objectives

The objectives of this study are:

1. Develop robust machine learning models capable of predicting stroke risk using small, imbalanced datasets.
2. Identify key features and risk factors influencing stroke occurrence.
3. Enhance model interpretability to support clinical decision-making.
4. Evaluate whether complex ML models outperform a simple logistic regression baseline.
5. Reduce false negatives through appropriate handling of class imbalance.

Background

Stroke remains one of the leading causes of death and disability globally, accounting for roughly 11% of total deaths (WHO). Early identification of high-risk individuals can greatly improve prevention and treatment outcomes.

By analysing factors such as age, hypertension, heart disease, glucose level, BMI, smoking status, residence type, marital status, and work type, this project seeks to uncover the most influential predictors of stroke and provide insights that support data-driven healthcare decisions and early interventions.

Methodology

Data Source and Overview

The study used a dataset of 5,110 patients from Kaggle, containing demographic, lifestyle, and clinical variables relevant to stroke risk. Features included age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, smoking status, and stroke occurrence.

Out[2]:

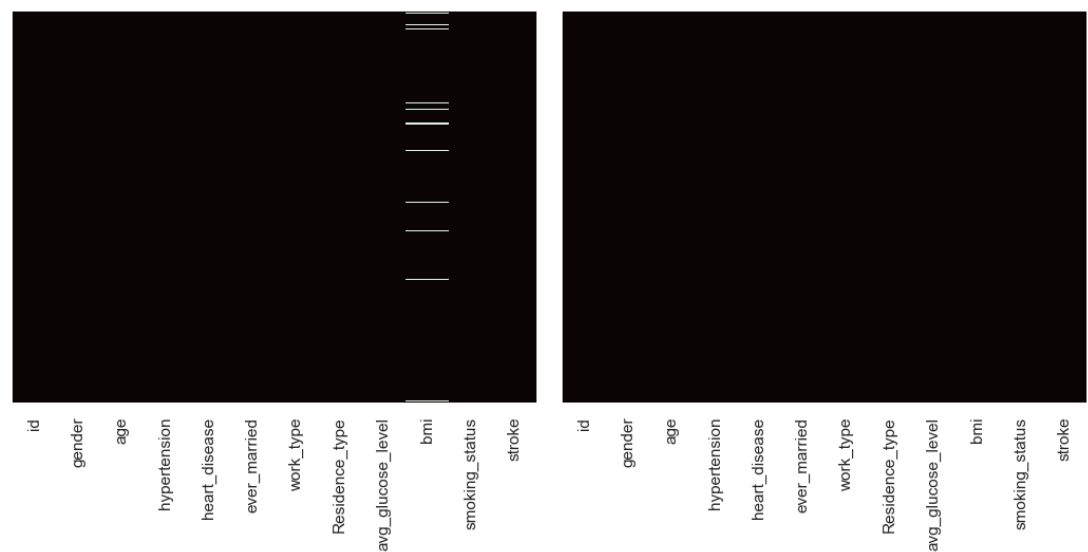
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Initial exploration revealed a severe class imbalance, with only 4.87% of patients having experienced a stroke. Handling this imbalance is crucial to prevent biased model predictions.

Data Preprocessing and Cleaning

Data preprocessing involved several steps:

- **Missing Values:** Missing BMI values (201 instances) were filled with the median.



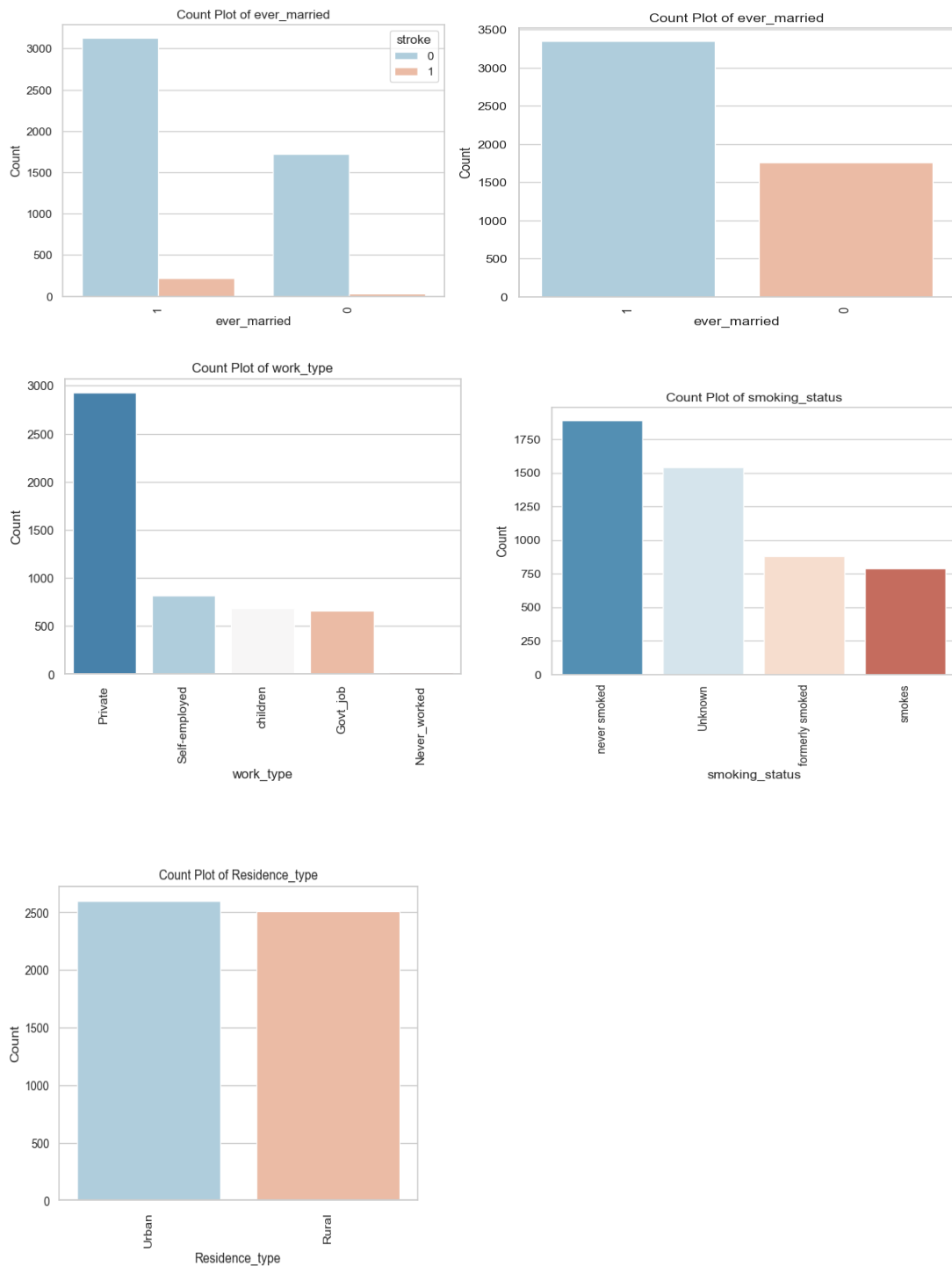
- **Feature Cleaning:** The 'id' feature was dropped as it did not contribute meaningful information.
- **Categorical Encoding:** Gender and marital status were binary encoded, while other categorical features were later one-hot encoded.
- **Age Conversion:** Age was converted from float to integer for clarity.

One erroneous record for gender labelled as “Other” was removed. After preprocessing, the dataset contained 5,109 observations.

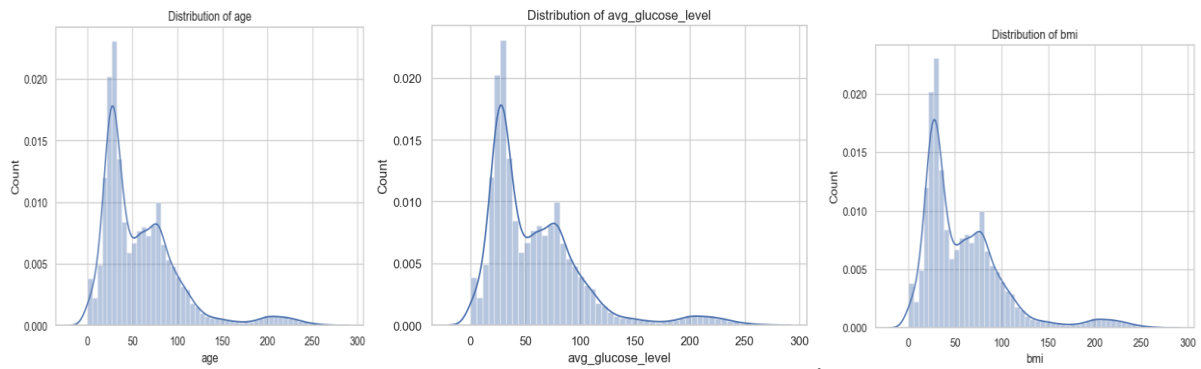
Exploratory Data Analysis

Descriptive and visual analyses were performed to understand the distribution of features and their relationship with stroke occurrence:

- **Categorical Variables:** Bar plots showed that stroke occurrence was higher in older age groups, self-employed individuals, and those with a history of smoking.



- **Numerical Variables:** Histograms and density plots for age, BMI, and glucose levels helped identify at-risk ranges.

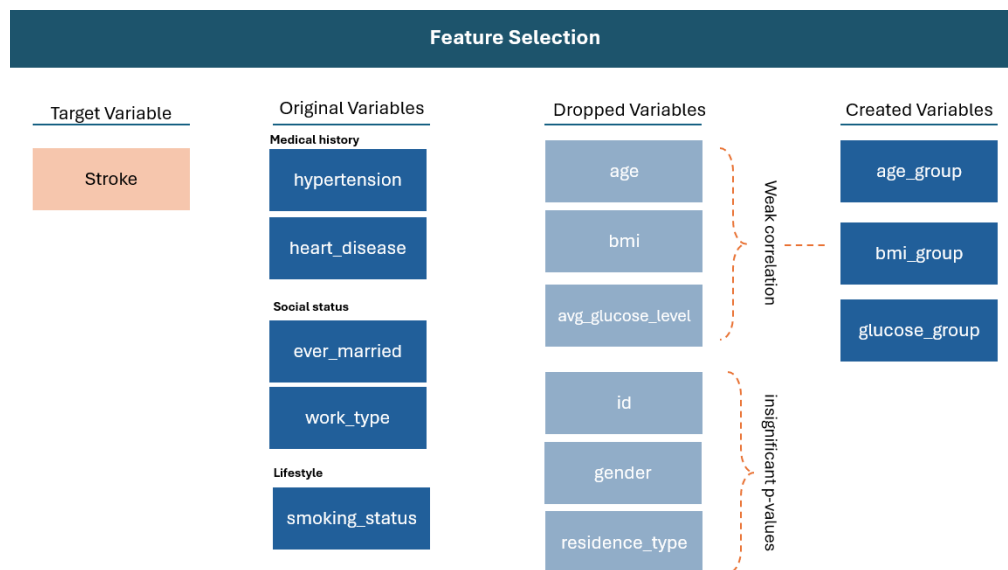


- **Probability Analysis:** Calculating the percent age of stroke occurrences across categories highlighted that age, glucose level, and BMI were among the most influential risk factors.

Feature Engineering

New features were created by binning continuous variables into clinically meaningful ranges:

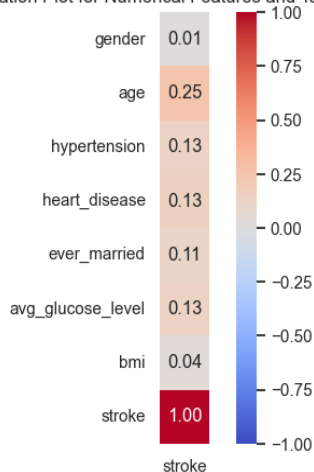
- **Age Groups:** 0–30, 31–50, 51–70, 71–80, 81–90 years
- **Glucose Groups:** Normal (<150 mg/dL), Pre-diabetes (150–200 mg/dL), Diabetes (>200 mg/dL)
- **BMI Groups:** Healthy, Overweight, Class 1–3 Obesity



Statistical Testing

A correlation analysis was conducted to examine the linear associations between the numerical predictors and stroke occurrence. The heatmap below demonstrates that all correlations with the stroke outcome were positive but generally low in magnitude, reflecting the multifactorial nature of stroke risk.

Correlation Plot for Numerical Features and Target Variable



Age exhibited the strongest correlation with stroke ($r = 0.25$), indicating that increasing age is moderately associated with higher stroke prevalence, consistent with established epidemiological evidence. Hypertension, heart disease, and average glucose level each showed small but meaningful positive correlations with stroke ($r = 0.13$ for all), highlighting their relevance as cardiovascular and metabolic risk factors. Ever-married status demonstrated a weak correlation ($r = 0.11$), which is likely attributable to age-related confounding rather than a direct causal relationship.

Conversely, BMI ($r = 0.04$) and gender ($r = 0.01$) displayed negligible correlations with the outcome, suggesting minimal linear association within this dataset.

Feature	Chi-Square	p-value	Significance
gender	0.34	0.5598	Not significant
hypertension	81.57	0.0000	Significant
heart_disease	90.23	0.0000	Significant
ever_married	58.87	0.0000	Significant
work_type	49.16	0.0000	Significant
Residence_type	1.07	0.2998	Not significant
smoking_status	29.23	0.0000	Significant
bmi_groups	40.64	0.0000	Significant
age_groups	389.92	0.0000	Significant
glucose_groups	102.86	0.0000	Significant

Statistical testing using chi-square tests for categorical features confirmed that age groups, glucose groups, BMI groups, hypertension, heart disease, marital status, work type, and smoking status were significantly associated with stroke occurrence. Features such as gender and residence type were found to be insignificant and excluded from the final analysis.

Optimizing Performance and Selecting the Best Model

Split Data into Train and Test Sets

The dataset was partitioned into training and test subsets using a 75:25 ratio to ensure robust model evaluation. Before model development, all numerical features were standardized to

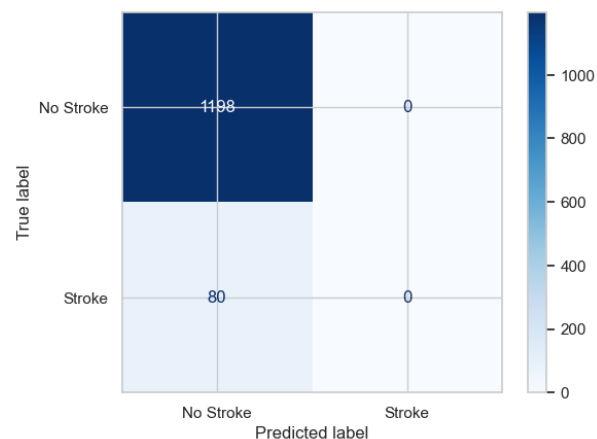
normalize their distributions, while categorical variables were transformed using one-hot encoding to enhance their compatibility with the machine-learning algorithms employed.

Performance Evaluation of Stroke Prediction Models

The performance of four machine learning classifiers, Logistic Regression, Random Forest, K-Nearest Neighbour (KNN), and Decision Tree, was evaluated to predict stroke occurrence based on clinical and imaging data. Given the dataset's severe class imbalance, with many more "No Stroke" cases than "Stroke" cases, accuracy alone was insufficient to judge effectiveness. Confusion matrices, classification reports, cross-validation scores, and ROC curves were used to provide a comprehensive evaluation.

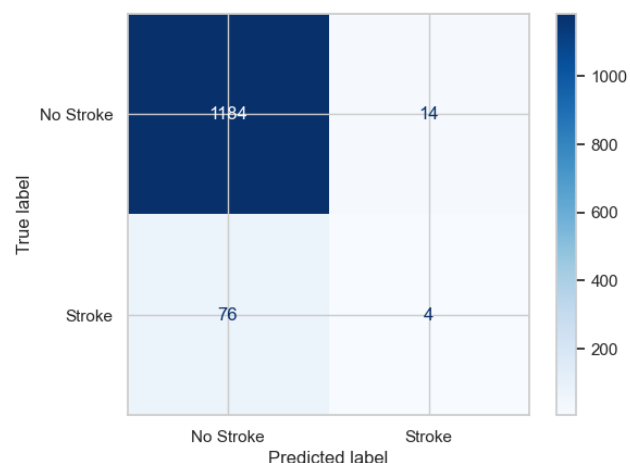
Logistic Regression

Logistic Regression achieved a test accuracy of 93.74% and a mean 10-fold cross-validation score of 95.11%. A confusion matrix revealed that while the model correctly predicted almost all "No Stroke" cases, it failed to identify stroke cases, resulting in a recall of 0 % for the Stroke class. This highlights the limitations of relying solely on accuracy in imbalanced datasets.



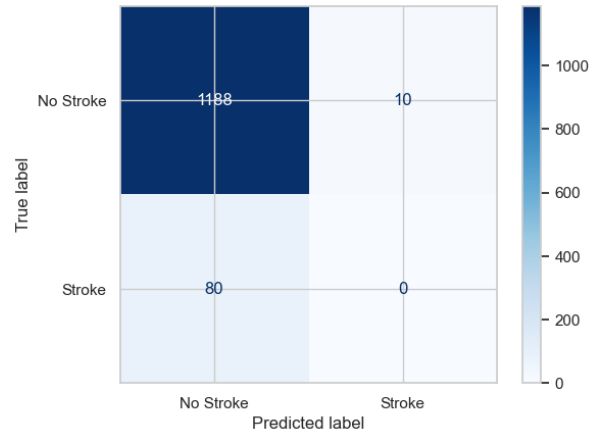
Random Forest

Random Forest achieved a test accuracy of 92.96 % with a mean cross-validation score of 94.17%. The confusion matrix and classification report showed slight improvement in identifying stroke cases, with a recall of around 5%, but performance on the minority class remained low.



K-Nearest Neighbour

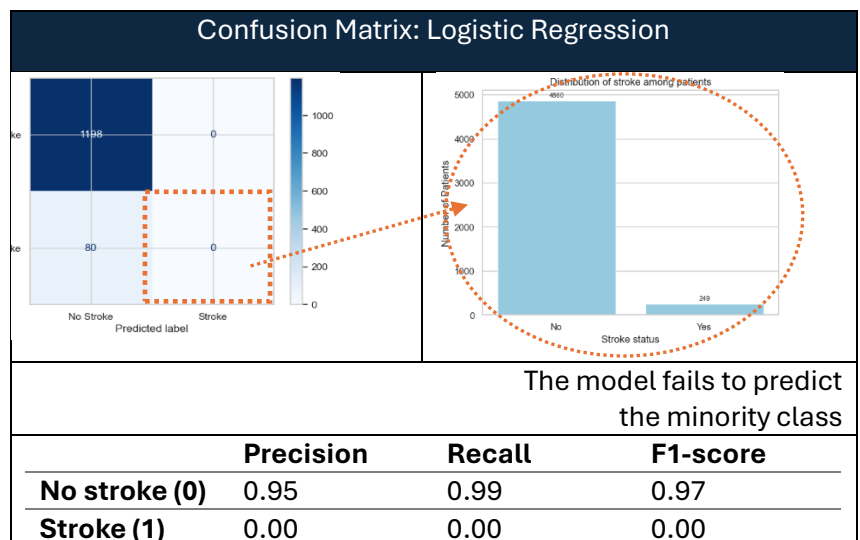
KNN, with k set to 3 and standardized features, reached a test accuracy of 93.74 %. It performed poorly in detecting stroke cases, with a recall of 0%. Cross-validation results showed a mean accuracy of 92.64 % and a slightly higher variability with a standard deviation of 1.0 %.



Decision Tree

The Decision Tree had the lowest test accuracy at 92.18%. It struggled to detect stroke cases effectively, and cross-validation confirmed consistent but weak performance.

Accuracy by Classifier Models	
Model	Score
Logistic Regression	93.74
KNN	93.74
Random Forest	92.96
Decision Tree	92.18

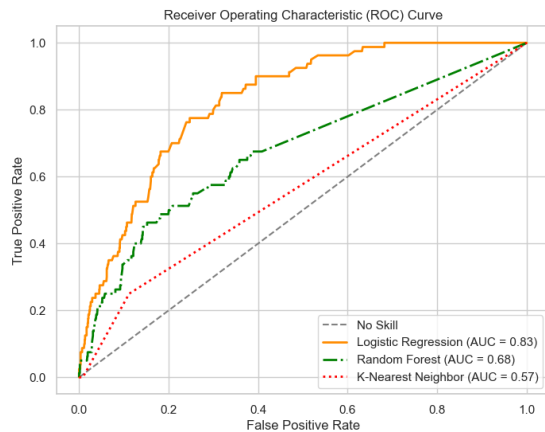


Evaluation of ROC Curve Analysis

The ROC curve provides a direct comparison of each model's ability to discriminate between stroke and non-stroke cases.

Model Comparison

In summary, Logistic Regression outperformed the other models in overall accuracy, cross-validation stability, and ROC AUC, achieving an AUC of 0.83%. Despite this strong statistical performance, it failed to detect any actual stroke cases in the test set, with a recall of 0 % for the minority class.

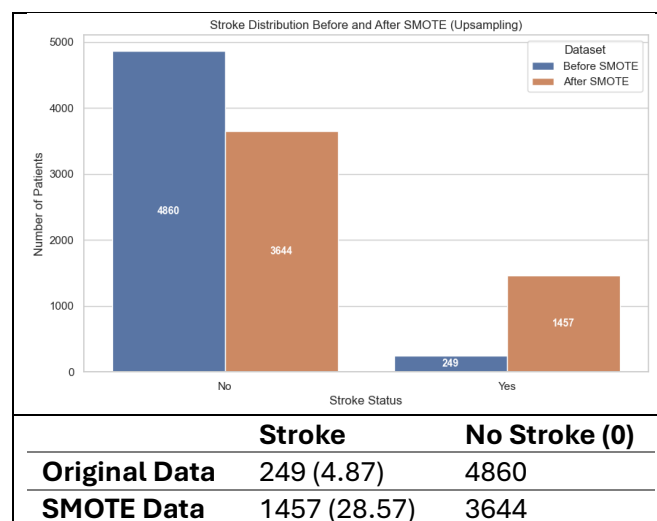


Model	AUC	Performance
Logistic Regression	0.83	Excellent
Random Forest	0.68	Poor to Fair
KNN	0.57	Fair to Poor
No Skill (baseline)	0.50	Random Guessing

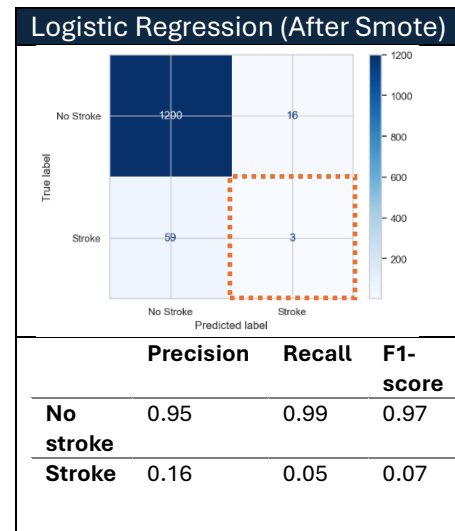
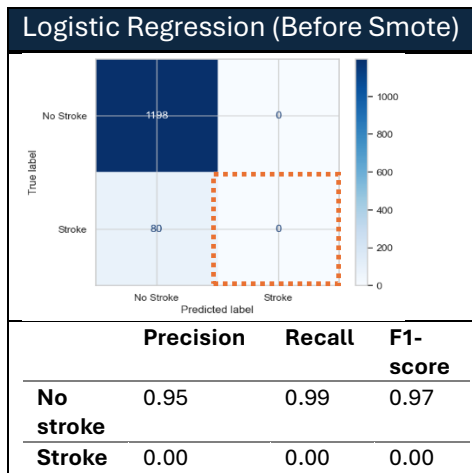
Random Forest and K-Nearest Neighbour performed slightly worse overall, with lower AUCs (0.68 and 0.57, respectively) and similarly poor detection of stroke cases. The Decision Tree had the lowest accuracy and was also unable to identify stroke cases effectively. These results highlight that while Logistic Regression shows the best discriminatory ability across all instances, all models currently struggle with the minority class, emphasizing the need for strategies such as class weighting, resampling, or feature engineering to improve clinical applicability.

Synthetic Minority Oversampling Technique (SMOTE)

To mitigate the substantial class imbalance present in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied to enhance representation of the minority class and reduce bias toward the majority class during model training.



After balancing the dataset with SMOTE, all four machine learning models showed significant improvement in their ability to classify stroke cases. The balanced training data allowed the algorithms to learn minority-class patterns more effectively, leading to notable gains in accuracy scores. Logistic Regression maintained the highest accuracy (94.91), an improvement of 1.17% from the pre-SMOTE model.



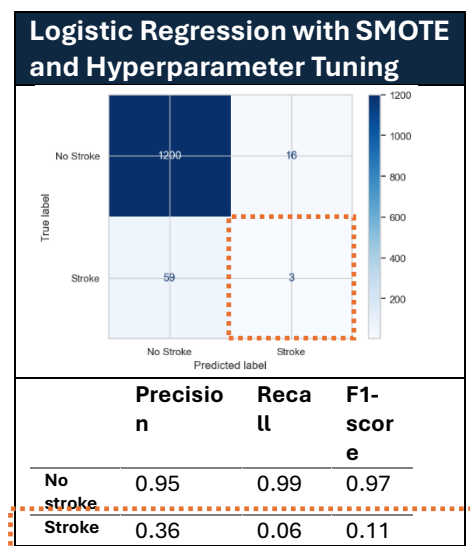
Effect of SMOTE on Stroke Classification Performance

Before applying SMOTE, the model was highly biased toward the majority class (No stroke), achieving high precision (0.95), recall (0.99), and F1-score (0.97) for that class, while completely failing to identify stroke cases (precision, recall, F1-score = 0). After applying SMOTE to balance the dataset, the model began detecting some stroke cases, reflected in low but non-zero performance metrics (precision 0.16, recall 0.05, F1-score 0.07). Performance for the majority class remained largely unchanged. These results indicate that while SMOTE helps mitigate class imbalance, the model still struggles to accurately predict cases from the minority class, highlighting the need for further feature optimization or model tuning.

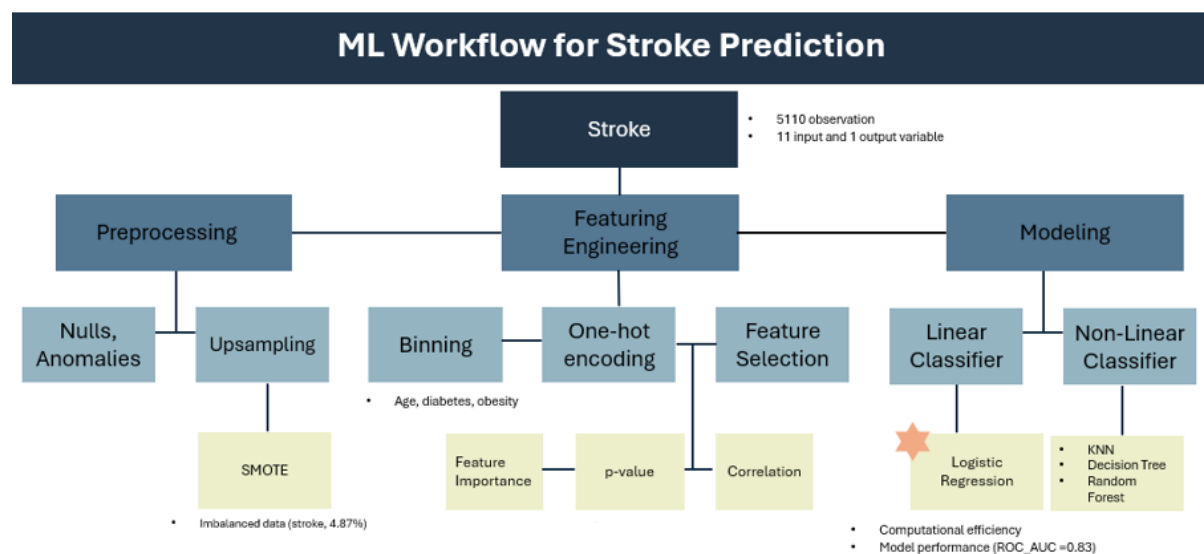
Hyperparameter Tuning and Model Evaluation

Logistic Regression was optimized using a 10-fold Grid Search over penalty and C values, with recall as the scoring metric. The best model (C=1) achieved a mean recall of 0.73 on SMOTE-balanced training data, although cross-validation scores showed high variance, indicating inconsistent detection of minority class cases. On the independent test set, the model maintained high performance for the majority class (No Stroke, recall 0.99) but poorly detected stroke cases (recall 0.06), reflecting the persistent challenge of class imbalance. These results highlight that while SMOTE improves minority class learning during training, additional strategies such as alternative algorithms or feature

optimization are needed to improve stroke detection in real-world data.

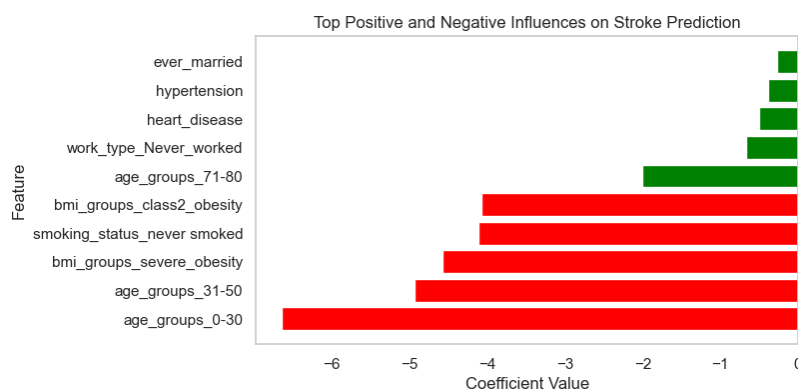


Final Model Summary



Coefficient Analysis

Coefficients in Logistic Regression represent each feature's contribution to the target variable prediction.



The model predicts stroke risk based on several features. The biggest factors that increase stroke risk are being in an older age group (especially 71–80), hypertension, heart disease, having ever been married, and never working.

The biggest factors that decrease stroke risk are being young, never smoking, and some high-BMI categories (likely due to patterns in the dataset, not because obesity is protective).

Overall, age and cardiovascular health conditions are the strongest predictors of stroke risk in your model.

Discussion

The study demonstrates that machine learning can effectively identify key risk factors for stroke and predict stroke occurrence in imbalanced datasets. Age, glucose level, and BMI emerged as the most influential features, supporting findings from prior clinical research.

Class imbalance remains a challenge, as stroke is a relatively rare event in the general population. Techniques like SMOTE are essential to ensure minority class representation and reliable model performance.

While complex models provided better predictive performance, logistic regression remains valuable for its interpretability and simplicity, which is important in clinical settings where explainability is critical.

References

1. Dataset: <https://www.kaggle.com/datasets/imaadmahmood/stroke-risk-synthetic-2025>
2. Glucose-Level Interpretation: <https://www.cdc.gov/diabetes/basics/getting-tested.html>
3. BMI Interpretation: <https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/>

In preparing my article, I consulted ChatGPT, a large language model developed by OpenAI (OpenAI. (2023). ChatGPT [Large language model]. <https://chat.openai.com>), for additional insights and information.