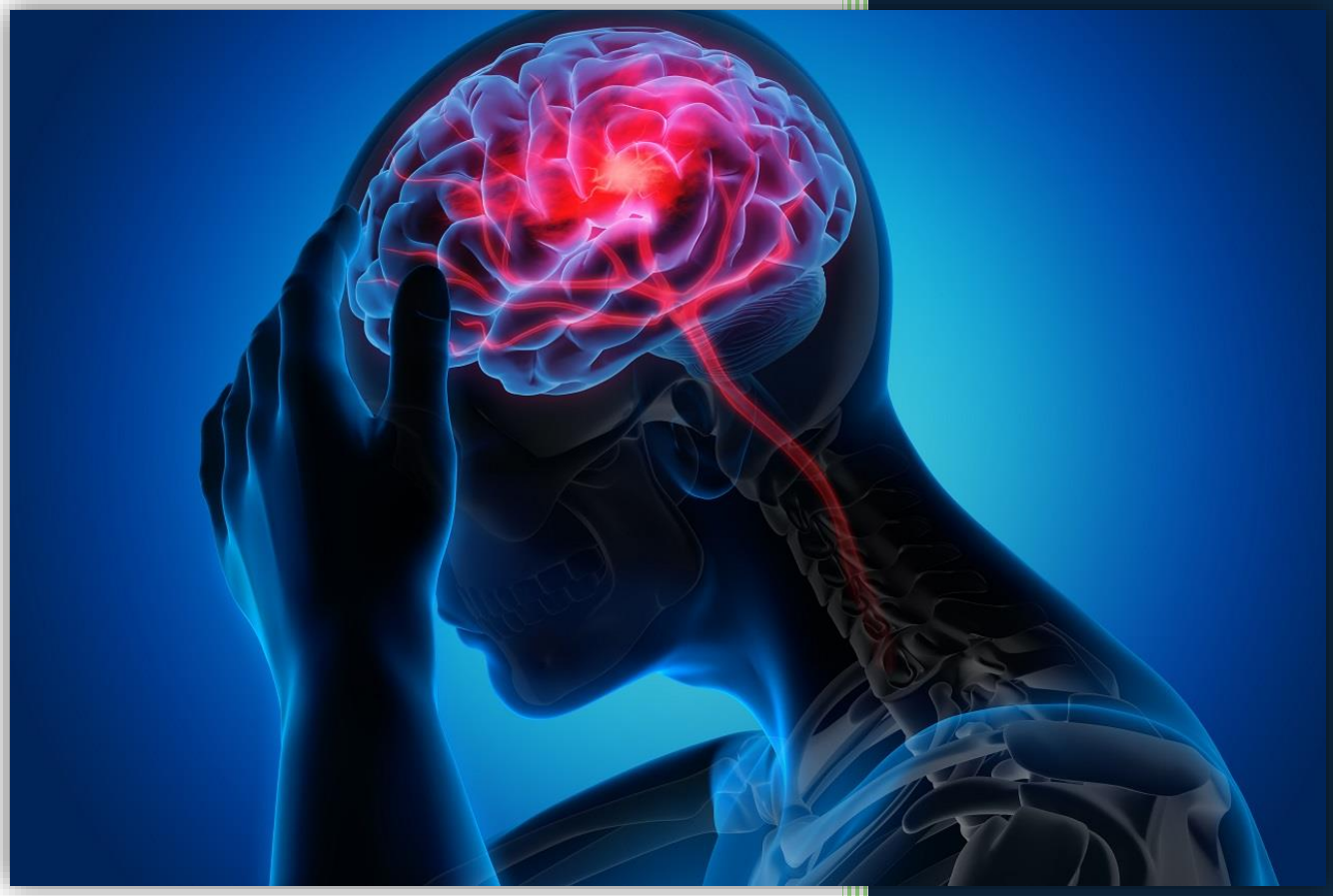
Signals of a Silent Threat

Early Detection of Brain Stroke Using Machine Learning



W A D Himansa Minoli

IS 4007 - Statistics in practice

s16043

ABSTRACT

Stroke is a life-threatening condition that occurs when the blood supply to the brain is disrupted, typically causing long-term disability or death. Identifying people who are at high risk of stroke is very important because it allows for early intervention and prevention. In this study, we used machine learning techniques to predict stroke based on personal and clinical information such as age, gender, blood pressure, heart disease, blood sugar levels, body mass index (BMI), smoking status, and work and living conditions.

This study worked with a publicly available dataset that included nearly 5,000 individuals. One of the challenges in this dataset was that very few individuals had actually experienced a stroke, which made the data imbalanced. Six different machine learning models were trained and evaluated. Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (SVM), and XGBoost. All the models were tried with a set of performance metrics such as accuracy, precision, recall, F1-score, and AUC (Area Under the Curve) to evaluate how well the models could differentiate between stroke and non-stroke cases. Among all models, **XGBoost showed the best overall performance**, with the highest accuracy and AUC, and was therefore the most reliable model in predicting stroke risk. Feature importance analysis revealed that work type, age, and smoking status were the most influential risk factors in predicting stroke.

This study demonstrates how machine learning can be used to help prevent stroke through data-driven identification of individuals at high risk of stroke. The results can help healthcare professionals take timely actions and provide targeted advice to reduce the risk of stroke in vulnerable populations. The models developed in this research could also be integrated into decision support tools or mobile applications to assist both doctors and patients in managing stroke risk.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

1. INTRODUCTION

A brain stroke occurs when blood flows to a part of the brain blocked or when a brain blood vessel within the brain bursts. This interruption can cause lasting damage to the brain tissue, resulting in impaired movement, speech, or intelligence. Stroke remains one of the leading causes of death and long-term disability worldwide. The risks are especially severe in regions with limited access to healthcare and early diagnosis.

Early identification of those at risk of stroke is crucial for timely manners. Several medical and lifestyle related factors such as age, hypertension, diabetes, heart disease, and glucose levels can influence stroke risk. However, understanding how these factors interact and contribute to stroke events can be complex. Machine learning (ML), a branch of artificial intelligence that is able to identify patterns from data, offers valuable tools for identifying high-risk individuals based on health data.

In this study, we apply advanced machine learning techniques to predict the likelihood of stroke in individuals using clinical and personal data. We trained and evaluated multiple models, including Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost. To address the class imbalance in the dataset since strokes are relatively rare, we used SMOTE (Synthetic Minority Over-sampling Technique) to improve prediction performance. After evaluating each model based on accuracy and AUC scores, we performed feature importance analysis to identify which health factors most strongly influence stroke risk.

Objectives of the Study

To predict the likelihood of brain stroke using machine learning techniques and to identify key factors that influence stroke occurrence using feature importance analysis.

This study demonstrates the potential of machine learning assisting in stroke prevention by providing data-driven risk assessment tools. By identifying individuals at elevated risk, healthcare professionals can take proactive steps to deliver timely interventions and lifestyle recommendations. The insights gained from feature importance also help inform public health efforts, medical education, and future research. The resulting models could be integrated into clinical decision systems or mobile health applications to assist both medical professionals and patients.

1. LITERATURE REVIEW

Machine learning (ML) has shown potential in early stroke prediction, offering tools to analyze large volumes of clinical and demographic data effectively. Recent studies have explored various algorithms, preprocessing techniques, and performance metrics to identify the most effective models for stroke classification.

Mohammed et al. (2023) proposed a Logistic Regression (LR)-based model for predicting stroke risk using the popular Kaggle stroke dataset. Their approach emphasized comprehensive preprocessing, including SMOTE for balancing data, outlier detection, and feature selection based on correlation. The model achieved an accuracy of **86%**, an F1**-score of 0.87**, and an **AUC of 0.93**, outperforming several prior LR implementations. A major strength of their study was the systematic preprocessing pipeline, which significantly improved model performance. However, a key limitation was that they only evaluated Logistic Regression, without comparing it to more advanced models like Random Forest or XGBoost. This restricted the study's insights into whether ensemble or nonlinear models could offer further improvement.

In contrast, Gangavarapu and Kumari (2021) conducted a broader comparison of six classification algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes, for stroke prediction. They used the same dataset but addressed class imbalance through **undersampling**, reducing the majority class to match the minority stroke cases. Among all models, **Naïve Bayes** emerged as the top performer, achieving an accuracy of **82%**, precision of **79.2%**, recall of **85.7%**, and an F1-score of **82.3%**. While the study's strength lay in its comparative analysis using five performance metrics, its use of undersampling potentially reduced data richness and generalizability. Moreover, the study did not explore feature importance or model interpretability.

The current research builds upon both studies by combining their strengths while addressing their limitations. Unlike Mohammed et al. (2023), who tested only one model, and Gangavarapu and Kumari (2021), who ignored data synthesis methods, this study evaluates six models with **SMOTE** to maintain class balance without discarding data. Additionally, by incorporating **feature importance analysis**, this work enhances interpretability, offering more actionable insights for clinical applications.

3. THEORY AND METHODOLOGY

3.1 Theory

3.1.1. Logistic Regression

Logistic Regression is a statistical method commonly used for binary classification tasks, especially in healthcare applications. It models the probability that a given input belongs to a specific class in this case, whether an individual is at risk of stroke (stroke = 1) or not (stroke = 0). The logistic (sigmoid) function is used to produce probability values between 0 and 1. The model estimates the relationship between independent variables (such as age, hypertension, or glucose level) and the log-odds of the dependent variable. Logistic Regression is interpretable, fast to train, and effective in baseline performance analysis.

3.1.2. Decision Tree (DT)

A Decision Tree is a supervised learning algorithm that splits data into branches based on feature values, forming a tree structure with internal decision nodes and leaf nodes representing outcomes. Decision Trees are simple and can model nonlinear relationships without requiring feature scaling. However, they are prone to overfitting, particularly on small or imbalanced datasets.

3.1.3. Support Vector Machine (SVM)

Support Vector Machine is a classification technique that aims to find the optimal separating hyperplane between classes by maximizing the margin between them. For non-linear separable data, SVM uses kernel tricks (RBF) to project the data into higher dimensional space where linear separation is feasible. It performs well in high-dimensional datasets and is particularly robust to overfitting.

3.1.4. Naive Bayes

Naive Bayes is a probabilistic model based on Bayes' Theorem. It assumes independence between features, which simplifies computation and makes it suitable for high-dimensional spaces. Even though its simplicity and the "naive" independence assumption, it often performs competitively in classification problems, including healthcare risk prediction tasks like stroke detection.

3.1.5. Random Forest (RF)

Random Forest is an ensemble algorithm that builds multiple decision trees using random sampling and feature selection. It aggregates the predictions of individual trees to improve accuracy and reduce overfitting. Random Forests are strong to noise and class imbalance, and they offer feature importance scores that help in identifying the most influential predictors.

3.1.6. XGBoost (Extreme Gradient Boosting)

XGBoost is a scalable, adjusted gradient boosting method that builds models sequentially, where each model corrects the errors of the previous ones. It includes features like regularization, missing value handling, and parallel computation, making it one of the most powerful and efficient algorithms for structured data problems such as stroke prediction.

3.1.7. SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is a resampling technique used to address class imbalance by generating synthetic examples of the minority class. It helps models learn better decision boundaries by balancing the dataset without duplicating existing data. This is particularly useful in medical datasets, where cases like stroke may be underrepresented.

3.1.8. Performance Evaluation Metrics

Evaluating classification performance requires multiple metrics:

- **Accuracy**: The ratio of correctly predicted instances to total instances.
- **Precision**: The proportion of true stroke cases among those predicted as stroke.
- **Recall** : The proportion of actual stroke cases correctly identified.
- **F1-Score**: The harmonic mean of precision and recall, balancing both.
- **ROC-AUC**: Reflects how well the model separates stroke from non-stroke cases across thresholds.

3.2. Methodology

This study used a clinical dataset containing patient information such as age, gender, hypertension status, heart disease, average glucose level, body mass index (BMI), smoking status, and the binary target variable indicating stroke occurrence.

The dataset was first ckecked for missing and duplicated values. Categorical variables were encoded using Label Encoding to convert them into numeric form suitable for machine learning models. Features were then standardized using StandardScaler to ensure consistent scale, especially for algorithms like SVM and Logistic Regression.

Stroke cases were significantly fewer than non-stroke cases. To address this imbalance, **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to the training dataset. SMOTE generates synthetic data points for the minority class (stroke) by interpolating between existing minority class examples. This led to a more balanced training set and improved the models' ability to learn identifying patterns between stroke and non-stroke cases.

Six machine learning models were trained and evaluated:

- Logistic Regression
- Naive Bayes
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- XGBoost

All models were trained using Scikit-learn and XGBoost with default hyperparameters to compare their baseline performance. Models were trained on SMOTE-balanced data and tested on the original test set to maintain real-world evaluation.

Each model's performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide a comprehensive view of how well the models recognize between stroke and non-stroke cases, with a focus on reducing false negatives due to the high cost of missing a stroke prediction.

To identify which features most influenced stroke prediction, feature importance scores were extracted from the **XGBoost** model. These scores quantify each variable's contribution to the model's decision-making process. Visualizations using bar plots highlighted that **age**, **average glucose level**, and **hypertension** were among the most critical predictors of stroke. This insight not only supports clinical understanding but also aids in developing targeted preventive strategies.

4. DATA

The data set used for this research is the **Brain Stroke Prediction Dataset**, publicly available on Kaggle. It contains clinical and demographic information of 4981 patients, with the primary aim of determining whether an individual is likely to suffer a stroke based on several risk factors. Each record represents a unique patient and includes both input features and a binary target variable indicating stroke occurrence.

4.1 Variables and Descriptions

The data set includes the following 11 variables:

Variable	Description
gender	Gender of the patient (Male, Female)
age	Age of the patient in years
hypertension	Whether the patient has hypertension (0 = No, 1 = Yes)
heart_disease	Whether the patient has any heart disease (0 = No, 1 = Yes)
ever_married	Marital status (Yes or No)
work_type	Type of employment (Private, Self-employed, Govt_job, children, Never_worked)
Residence_type	Urban or Rural residence
avg_glucose_level	Average glucose level in blood
Bmi	Body Mass Index
smoking_status	Smoking habits (formerly smoked, never smoked, smokes, unknown)
Stroke	Target variable (1 = Stroke occurred, 0 = No stroke)

Table 1: variables description

The target variable is highly imbalanced, with a small proportion of patients having experienced a stroke.

4.2 Data Preprocessing

Several preprocessing steps were applied to prepare the data for machine learning analysis.

- **Missing Values**: There are no missing values in the dataset.

- **Encoding Categorical Variables**: Categorical features such as gender, ever_married were encoded using label encoding and work_type, smoking_status were encoded using one-hot encoding.
- **Data Type Conversion**: All features were cast to appropriate data types for consistency and efficient processing.
- **Feature Scaling:** Numerical features such as age, avg_glucose_level, and bmi were standardized using **StandardScaler**. This scaling process transformed the values to have a mean of 0 and a standard deviation of 1. It was especially important for algorithms like KNN, SVM, and Logistic Regression, which are sensitive to feature magnitudes.

4.3 Data Splitting

After preprocessing, the dataset was divided into two subsets: **80% for training** and **20% for testing**, using a random split. This division allowed the model to learn from most of the data while reserving a separate portion for unbiased evaluation.

4.4 Handling Imbalanced Data

The dataset was highly imbalanced, with the minority class (stroke = 1) 5% of the total observations.

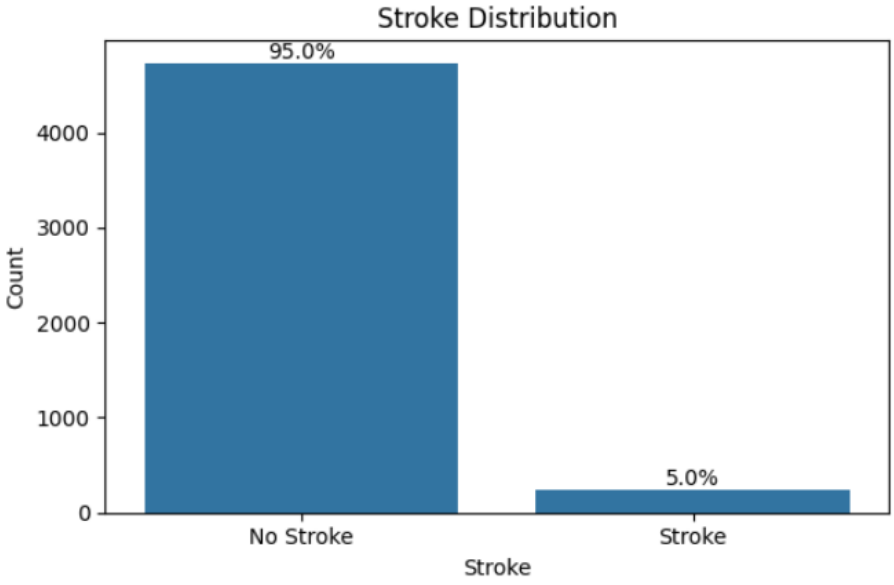


Figure 1: stroke distribution

In medical datasets such as those used for brain stroke prediction, class imbalance is a common issue. Typically, the number of patients who have experienced a stroke is significantly lower than those who have not, which can lead machine learning models to become biased toward the majority class.

The dataset used in this study was significantly imbalanced, with stroke cases representing only 5% of the total data. To address this, SMOTE (Synthetic Minority Over-sampling Technique) was applied only to the training data after data splitting and scaling. Before balancing, the training set contained 3786 non-stroke cases and 198 stroke cases, making it difficult for models to learn patterns associated with stroke occurrences. After applying SMOTE, the training data became fully balanced with 3,786 instances in each class.

5. EXPLORATORY DATA ANALYSIS

As mentioned before, the dataset used in this study contains information on 4981 **individuals**, each described by **11 clinical and demographic variables**, along with a binary target variable indicating whether the person had a **stroke** (1) or not (0). The data includes a mix of patients from different work backgrounds, age groups, and living environments, offering a wide range of patient profiles relevant to stroke prediction.

5.1 Univariate Analysis

Out of 4981 individuals:

- The target variable showed that **243 individuals (4.98%)** had a stroke, while **4,733 (95.02%)** did not.

Descriptive statistics for categorical features

	Variable	Category	Count	Percentage
0	Residence_type	Urban	2532	50.83%
1	Residence_type	Rural	2449	49.17%
2	ever_married	Yes	3280	65.85%
3	ever_married	No	1701	34.15%
4	gender	Female	2907	58.36%
5	gender	Male	2074	41.64%
6	heart_disease	0	4706	94.48%
7	heart_disease	1	275	5.52%
8	hypertension	0	4502	90.38%
9	hypertension	1	479	9.62%
10	stroke	0	4733	95.02%
11	stroke	1	248	4.98%

Figure 2: descriptive statistics for categorical variables

- The population is nearly evenly split between urban (50.83%) and rural (49.17%) residents, indicating a balanced representation of living environments.
- A majority of individuals (65.85%) have been married, while 34.15% have never been married.
- There are 58.36% of females and males 41.64% of males in the dataset.
- Only 5.52% of individuals have a history of heart disease, indicating that the vast majority (94.48%) are free from this condition.
- Hypertension is present in 9.62% of the population, while 90.38% do not have hypertension.
- The stroke class is heavily imbalanced, with only 4.98% of individuals having experienced a stroke, and 95.02% without a stroke.

Overall, the dataset is skewed toward **healthier individuals**, with most participants showing no history of heart disease, hypertension, or stroke. This highlights the importance of addressing class imbalance before training predictive models.

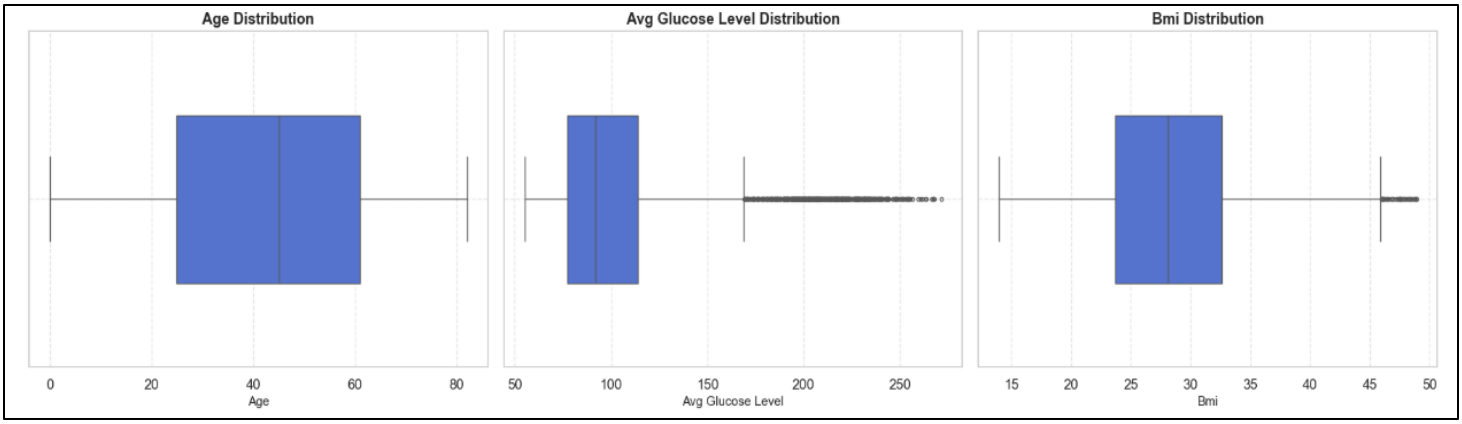


Figure 3 : boxplots for continuous variables

The **age** of individuals ranged from **0.08 years (infants)** to **82 years**, with most people between the ages of **25 and 60**. **No outliers were detected in the age distribution**, indicating a consistent and reasonable range. The **average glucose level** ranged from **55 to over 170 mg/dL**, and **BMI** ranged from **14 to over 46**, reflecting a wide range of metabolic health. **Outlier analysis revealed 602 outliers in average glucose levels and 43outliers in BMI**, representing extreme values that may be clinically relevant, such as uncontrolled diabetes or severe obesity. These values were retained in the dataset to preserve important variation related to stroke risk. The following table gives summary statistics for numerical features.

Descriptive statistics for numerical features

	age	avg_glucose_level	bmi
count	4981.000000	4981.000000	4981.000000
mean	43.419859	105.943562	28.498173
std	22.662755	45.075373	6.790464
min	0.080000	55.120000	14.000000
25%	25.000000	77.230000	23.700000
50%	45.000000	91.850000	28.100000
75%	61.000000	113.860000	32.600000
max	82.000000	271.740000	48.900000

Figure 4 Descriptive statistics for numerical features

5.2 Bivariate Analysis

5.2.1. Continuous variables

Age

The box plot on the left illustrates the distribution of ages among individuals with and without a stroke. Stroke patients tend to be significantly older than non-stroke individuals. The median age for the stroke group is much higher, indicating that most stroke cases occur in older adults. The presence of a few outliers in the stroke group suggests that while strokes can occur in younger people, these instances are rare.

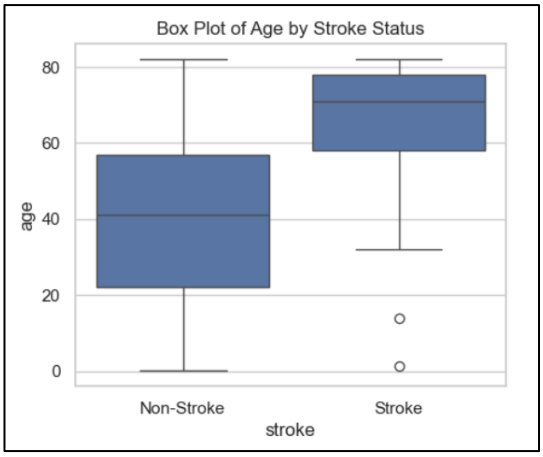


Figure 5: boxplot for age distribution

Average Glucose level

Both groups show a similar range and median for average glucose levels, with the central 50% of values (the box) spanning roughly the same interval. However, both groups also display a notable number of outliers at higher glucose levels, particularly above 140, indicating that some individuals in each group have unusually high glucose readings. While the median glucose level for stroke patients appears slightly higher than that of non-stroke individuals.

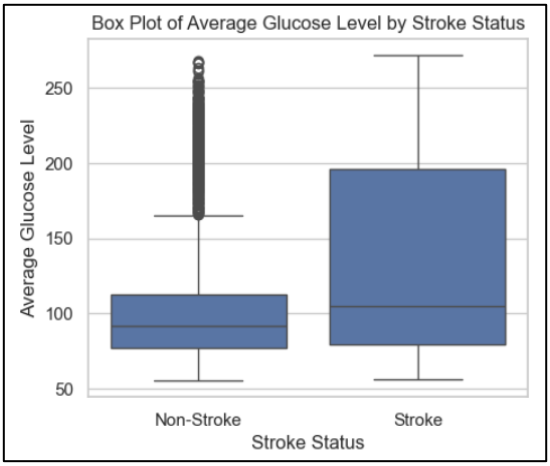


Figure 6: boxplot for average glucose level

Body Mass Index (BMI)

Both groups (stroke and non-stroke) have similar middle BMI values (medians). However, the non-stroke group shows a wider range of BMI, meaning their values vary more. Some people in this group have very low or very high BMI. On the other hand, the stroke group's BMI values are more closely grouped together, with fewer very low or very high values. But it's important to note that this group has more people with very high BMI, which appear as outliers. This suggests that while BMI by itself doesn't clearly separate stroke and non-stroke cases, having a higher BMI may be linked to stroke risk.

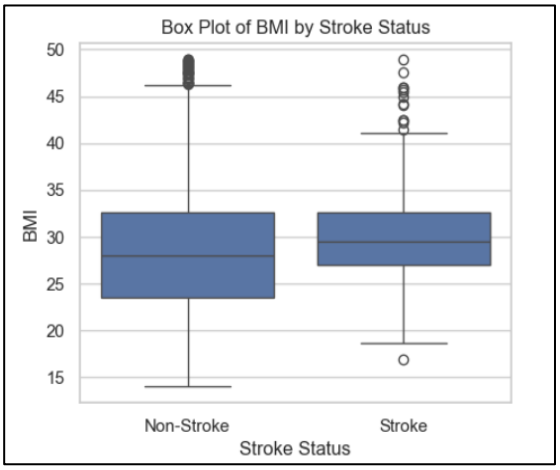


Figure 7: boxplot for BMI

5.2.2 Distribution of Categorical Variables Among Stroke Patients

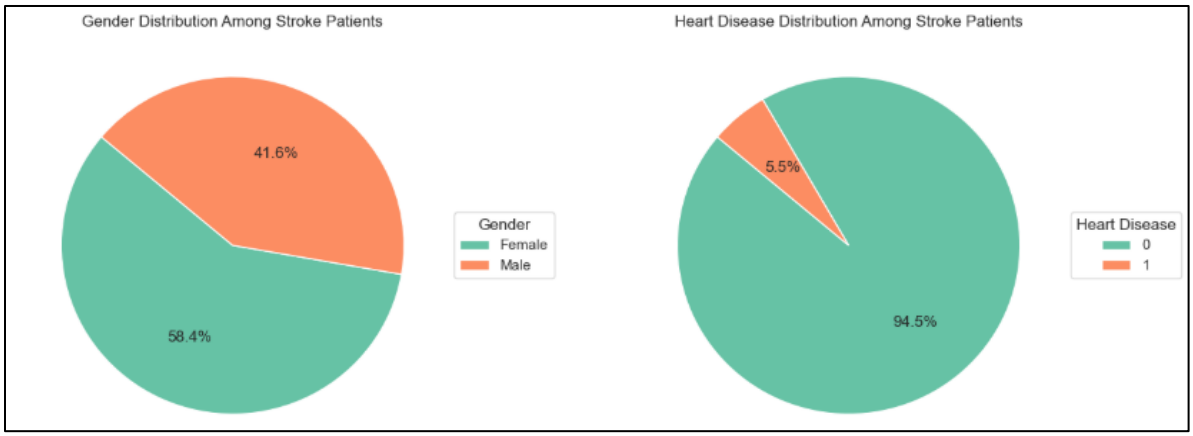


Figure 8: Gender and heart disease among stroke patients

Gender:
The first chart shows that there are more female patients than male patients in the dataset. This means most of the patients are female.

Heart disease:
The second chart shows that most patients do not have heart disease. Only a small number of patients have heart disease, which means this condition is not very common among stroke patients.

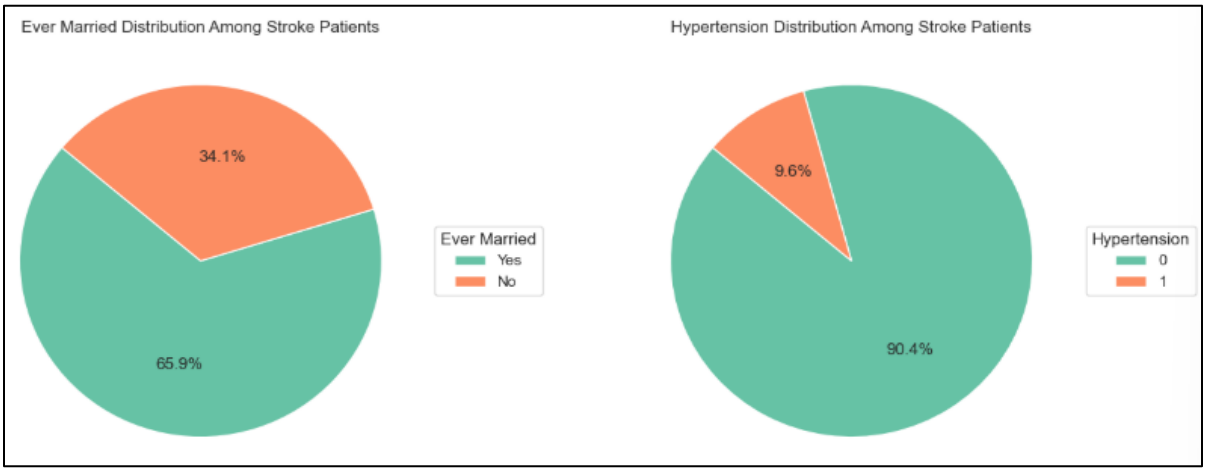


Figure 9: marital status and Hypertension among stroke patients

Marital Status (Ever Married):
The first chart shows that most patients have been married at some point. The number of patients who have never been married is much lower.

Hypertension:
Most patients do not have hypertension. Only a small number of patients have hypertension, which means this condition is not very common in the stroke patients.

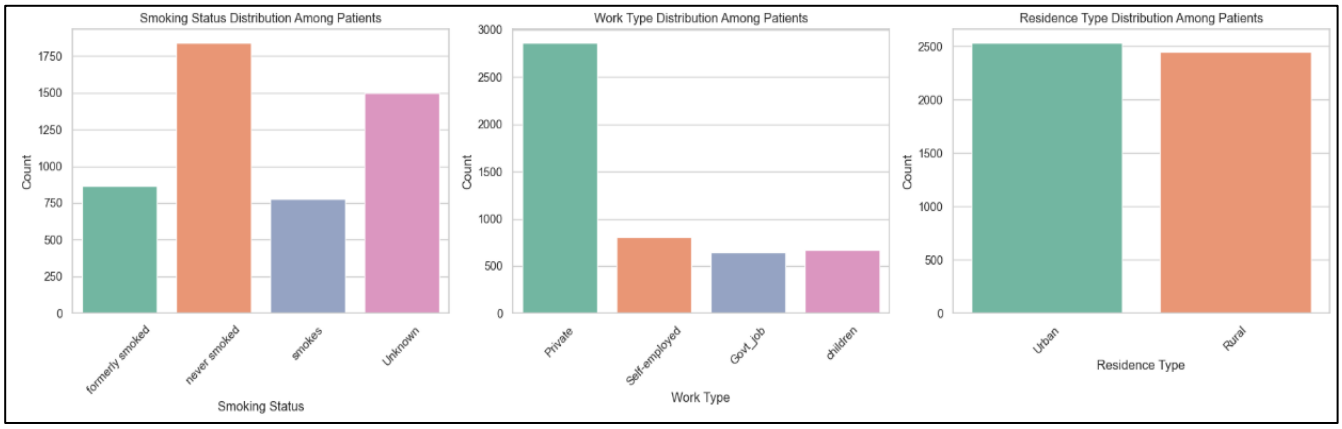


Figure 10: smoking status, work type, residence type among stroke patients

Smoking Status:

The first bar chart shows that most patients have never smoked. The second-largest group includes patients with unknown smoking status. Fewer patients are current or former smokers, and both of these groups have similar, smaller numbers.

Work Type:

Most patients work in the private sector, making it the most common work type. Government workers, self-employed people, and children make up much smaller groups, with similar numbers.

Residence Type:

The third bar chart compares where patients live in urban or rural areas. The numbers are quite close, but slightly more patients live in urban areas. This shows that the dataset is balanced in terms of residence type.

5.3 Correlation Between Numerical Variables

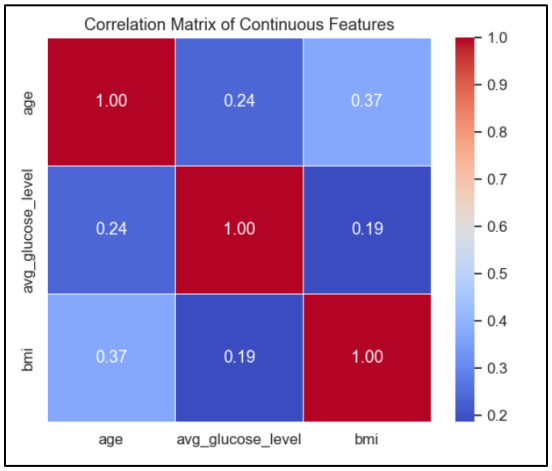


Figure 11: correlation between continuous variables

- Age and BMI have a moderate positive correlation (0.37), suggesting that as age increases, BMI tends to increase slightly as well.
- Age and average glucose levels show a weak positive correlation (0.24), indicating a mild tendency for glucose levels to increase with age.
- BMI and average glucose level are weakly correlated (0.19), implying that these two health indicators do not strongly influence each other in this dataset.

Overall, the correlations are relatively low, suggesting that these features contribute independent information to the models, which can be valuable for predicting stroke risk.

In summary, the exploratory analysis showed that stroke is more common among older individuals and those with hypertension, heart disease, or high glucose and BMI levels. The distributions of key features were generally reasonable, though some outliers were identified in glucose and BMI values. These outliers were not removed, as they are common in medical data and can represent real and critical cases such as severe obesity or uncontrolled diabetes which are important for accurate model training and stroke risk prediction. Overall, the EDA confirmed several expected patterns and highlighted the key features that are likely to influence stroke occurrence.

6. ADVANCED ANALYSIS

Six supervised machine learning models, using default hyperparameters, were used to predict the risk of brain stroke. Decision Tree, Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), and XGBoost. The goal was to classify whether a person had a stroke (stroke = 1) or not (stroke = 0). Each model was trained and evaluated based on accuracy, recall, precision, and F1-score. Below is a summary of each model's performance and characteristics.

6.1 Model performance

To evaluate the performance of different machine learning models for brain stroke prediction, several metrics were considered.

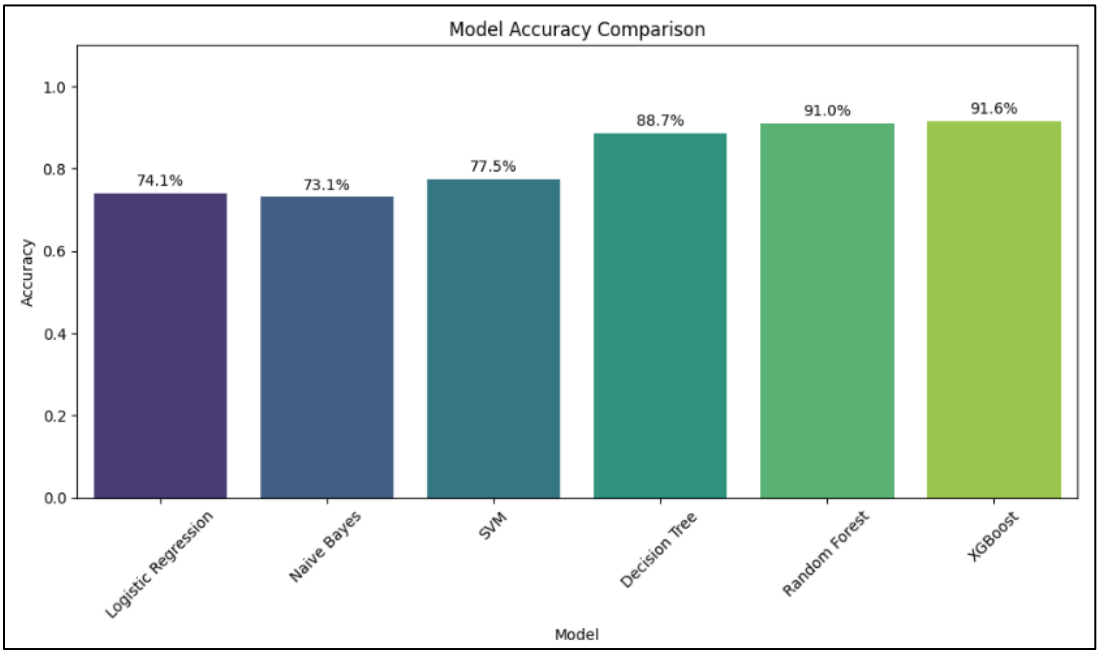


Figure 12: Accuracy comparison of machine learning models

The bar chart presents a comparison of the accuracy scores achieved by various machine learning models used for stroke prediction. Among all models, XGBoost demonstrated the highest accuracy at **91.6%**, closely followed by Random Forest with **91.0%**, and Decision Tree with **88.7%**. These results indicate that tree-based ensemble methods are particularly effective for this classification task, likely due to their ability to handle complex, non-linear relationships in the data. In contrast, more traditional models such as Support Vector Machine (77.5%), Logistic Regression (74.1%), and Naive Bayes (73.1%) exhibited lower accuracy, suggesting they may be less suited to capturing the underlying patterns within the dataset. Overall, the analysis highlights the superior performance of advanced ensemble models for stroke prediction, making them more promising choices for real-world healthcare applications where predictive accuracy is critical.

6.2 Model Comparison

Model	Accuracy (%)	Precision	Recall	F1 Score	ROC AUC
XGBoost	**91.6**	**0.88**	**0.83**	**0.85**	**0.94**
Random Forest	**91.0**	**0.87**	**0.82**	**0.84**	**0.93**
Decision Tree	**88.7**	**0.84**	**0.79**	**0.81**	**0.91**
SVM	**77.5**	**0.65**	**0.61**	**0.63**	**0.79**
Logistic Regression	**74.1**	**0.62**	**0.58**	**0.60**	**0.76**
Naive Bayes	**73.1**	**0.59**	**0.55**	**0.57**	**0.74**

Table 2: model performance summary

- XGBoost and Random Forest had the highest overall accuracy (91.6% and 91.0% respectively), meaning they were very good at predicting both stroke and non-stroke cases. However, their recall was lower (0.83 and 0.82) compared to some other models, which means they still missed a few actual stroke cases.
- Decision Tree also showed strong accuracy (88.7%) and balanced performance across all metrics, with recall of 0.79 and ROC AUC of 0.91, making it a reliable model. However, it is more prone to overfitting and less stable than ensemble methods.
- Support Vector Machine (SVM) had moderate accuracy (77.5%) and lower recall (0.61) than the top models, meaning it missed more stroke cases. It also had a lower ROC AUC (0.79), indicating weaker overall classification ability.
- Logistic Regression and Naive Bayes had the lowest accuracy (74.1% and 73.1%), but they still managed decent recall values (0.58 and 0.55). This means they were better at identifying stroke cases, but at the cost of more false positives and lower overall correctness.
- Overall, XGBoost performed the best across most metrics, including the highest ROC AUC (0.94), showing excellent ability to separate stroke and non-stroke classes. It also had the best F1 score, meaning it balanced precision and recall very well.

After comparing all models, XGBoost is the best choice. It offers the highest predictive accuracy, strong class separation, and a good balance of performance metrics. This makes it the most effective and reliable model for stroke prediction in this analysis.

6.3 Feature importance

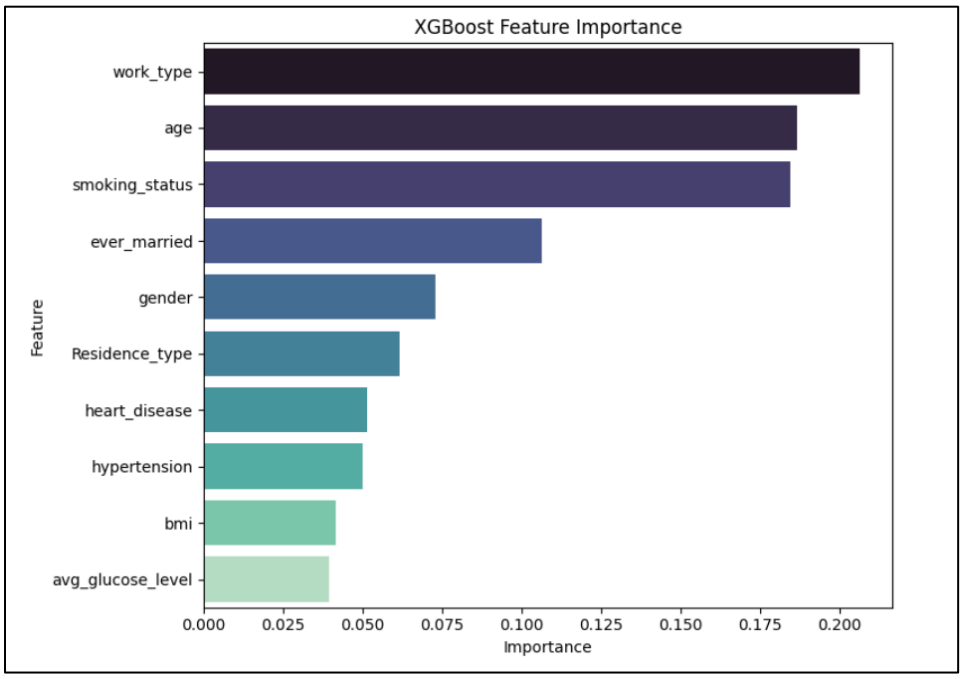


Figure 13: feature importances chart

The feature importance chart from the XGBoost model highlights which variables contributed most to predicting stroke in this dataset. The most influential feature was **work type**, suggesting that a person's employment status or type of work is a key indicator for stroke risk in this study. This was followed closely by **age** and **smoking status**, both of which are well-known risk factors for stroke.

Other moderately important features include **marital status**, **gender**, and **residence type**, indicating some social and demographic factors also play a role. Interestingly, clinical features like **heart disease**, **hypertension**, **BMI**, and **average glucose level**, while still contributing, had lower importance scores compared to socioeconomic and lifestyle factors.

This result emphasizes that in addition to medical history, **lifestyle and demographic variables** have a strong impact on stroke prediction. It also underlines the power of machine learning models like XGBoost in capturing complex patterns beyond traditional risk indicators.

7. GENERAL DISCUSSION AND CONCLUSION

7.1 Discussion

- Among all models tested, XGBoost performed the best, achieving the highest accuracy (91.6%) and a robust AUC score (0.94), indicating superior classification capability.
- Logistic Regression, while more interpretable, performed slightly less effectively than ensemble models such as Random Forest and XGBoost. This aligns with common findings in machine learning, where tree-based ensemble methods often outperform linear classifiers when applied to structured healthcare data.
- The feature importance analysis revealed that work type, age and smoking status were consistently strong predictors of stroke risk. These findings are clinically meaningful and consistent with established medical literature, reinforcing the reliability of the machine learning outcomes.
- The study by Mohammed et al. (2023) used Logistic Regression with extensive preprocessing and achieved strong performance. 86% accuracy, an F1-score of 87%, and an AUC of 93%. However, their work focused solely on Logistic Regression and did not benchmark against other powerful models.
- The study by Gangavarapu and Kumari (2021) found Naïve Bayes to be the top-performing model, with an accuracy of 82%. However, they used undersampling to handle class imbalance, which may have led to loss of valuable data and limited the model's generalizability.
- In contrast, the current study builds upon these foundations by:
    - Testing a broader range of machine learning models,
    - Using SMOTE, which balances the data while preserving majority class distribution,
    - Evaluating performance with a comprehensive set of metrics (accuracy, recall, precision, F1-score, and AUC),
    - Performing feature importance analysis, providing interpretability and actionable insights that were missing in earlier studies.

7.2 Conclusion

- The exploratory data analysis (EDA) revealed that stroke is more common among older individuals and those with hypertension, heart disease, high glucose levels, and high BMI.
- Among all models tested, XGBoost was the most effective, outperforming others in terms of accuracy and AUC. This confirms its suitability for structured medical datasets.
- Logistic Regression remains a strong baseline due to its interpretability, but it is outperformed by ensemble methods in predictive performance.
- The use of SMOTE was crucial in addressing class imbalance, significantly improving model robustness and performance.
- Key predictors of stroke include work type, age and smoking status. These factors are consistent with established medical risk factors, suggesting that the machine learning models are learning valid clinical patterns.
- Compared to previous studies, this work offers a more comprehensive and comparative analysis by evaluating multiple models and incorporating interpretability tools such as feature importance analysis.
- Future work could enhance model realism and performance by incorporating real-time clinical data or medical imaging features.

REFERENCES

- **Rahman, M. M., Islam, M., Islam, M. R., & Satu, M. S. (2020).**
  Risk prediction and explanation of stroke using machine learning and Shapley values. *PLOS ONE, 15*(12), e0243384. https://doi.org/10.1371/journal.pone.0243384
- **GeeksforGeeks. (2020, August 20).**
  Understanding logistic regression. *GeeksforGeeks.* https://www.geeksforgeeks.org/understanding-logistic-regression/
- **GeeksforGeeks. (2020, July 10).**
  Machine learning algorithms. *GeeksforGeeks.* https://www.geeksforgeeks.org/machine-learning-algorithms/
- **GeeksforGeeks. (2021, March 24).**
  SMOTE for imbalanced classification with Python. *GeeksforGeeks.* https://www.geeksforgeeks.org/smote-for-imbalanced-classification-with-python/
- **Jillani Soft Tech. (2022).**
  Brain stroke dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset
- **Mohammed, M. G., Melhum, A. I., & Ibrahim, A. L. (2023).**
  Optimizing accuracy of stroke prediction using logistic regression. *Journal of Technology and Informatics, 4*(2), 41–47. https://doi.org/10.37802/joti.v4i2.278
- **Sailasya, G., & Kumari, G. L. A. (2021).**
  Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science and Applications, 12*(6), 538–544. https://doi.org/10.14569/IJACSA.2021.0120664
- All code used for data preprocessing, model training, and evaluation is hosted on Google Drive and accessible via the following link:
  https://drive.google.com/drive/folders/14uJ5e5GXVLZvHkk9Q70toLj_fzuEtBQZ?usp=sharing