Georgia State University | J. MACK ROBINSON COLLEGE OF BUSINESS

MACHINE LEARNING FOR STROKE PREDICTION: CAN WE DETECT STROKE BEFORE IT STRIKES ?

Ngoc Nguyen, Phan Anh Nguyen

LINKEDIN — Ngoc Nguyen, Phan Anh Nguyen

GITHUB

References

1. Soriano, F. (2021). *Stroke Prediction Dataset.* Kaggle. Retrieved from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
2. World Health Organization (WHO). (2024). *Global Health Estimates: Stroke Mortality and Disability Statistics.* Retrieved from https://www.who.int
3. Centers for Disease Control and Prevention (CDC). (2024). *Stroke Facts and Statistics.* Retrieved from https://www.cdc.gov/stroke
4. American Heart Association (AHA). (2024). *Heart Disease and Stroke Statistics—2024 Update.* Retrieved from https://www.heart.org

A. INTRODUCTION

Stroke remains a leading cause of death and disability worldwide, with over **7.6 million stroke-related deaths annually** (WHO, 2024). In the **United States**, an estimated **795,000 people experience a stroke each year**, with direct and indirect costs exceeding **$56 billion** (CDC, 2024). Despite advances in treatment, **nearly 87% of strokes are ischemic**, making early risk prediction crucial for timely intervention and improved patient outcomes (AHA, 2024).

Machine learning (ML) has shown promise in enhancing stroke risk assessment by identifying high-risk individuals based on health indicators. This study utilizes a dataset of **4909** patients, incorporating key features such as **gender, age, hypertension, heart disease, glucose levels, BMI, smoking status, marital status, work type**, and **residence type**. We selected **Logistic Regression, Random Forest, LightGBM**, and **XGBoost** for model training due to their **suitability for small datasets**, their ability to provide **interpretable feature importance**, and their **computational efficiency**. These models strike a balance between predictive power and explainability, which is crucial in a healthcare context where model transparency and trust are essential for decision-making.

This study highlights the potential of machine learning models in improving stroke prevention strategies while serves as an exploration of how to effectively handle highly imbalanced medical datasets and develop strategies to improve predictive performance in real-world applications.

B. METHODS

Exploratory Data Analysis → Data Preprocessing → Model Fitting → Evaluation →

- **Exploratory Data Analysis**: Null, outlier detection; Univariate / Bivariate/ Interaction Analysis; Correlation; Odds Ratio
- **Data Preprocessing**: One-hot Encoding; Down-sampling majority class; Robust Scaler; Feature Selection
- **Model Fitting**: Logistic Regression; Random Forest; XGBoost; LightGBM
- **Evaluation**: Confusion Matrix; Accuracy, Precision, Recall, F1-score; AUC – ROC Curve

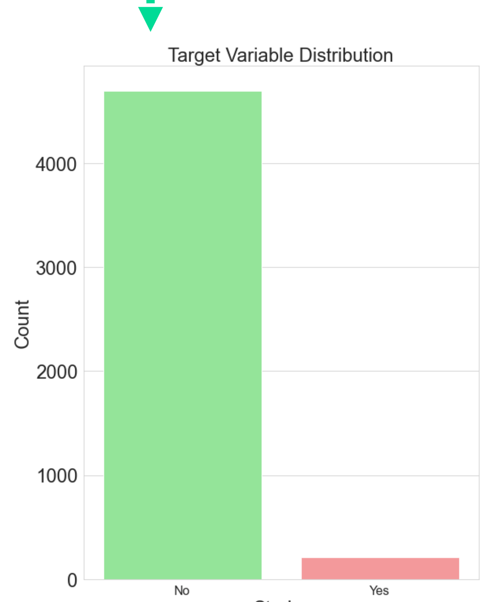

Figure 1. Target Variable Distribution

- Stroke patients make up approximately 4% of the dataset, meaning that for **every 100 patients in this population, 4 have experienced a stroke**. The dataset is **highly imbalanced**, in order to address this, the majority class was down-sampled.
- Approximately 11.6% of the data points in average glucose level and 2.2% in BMI are outliers, necessitating the use of **RobustScaler** for all numeric variables.
- **Feature selection**: two datasets choices were decided: full dataset, and reduced dataset without gender and residence type variables.

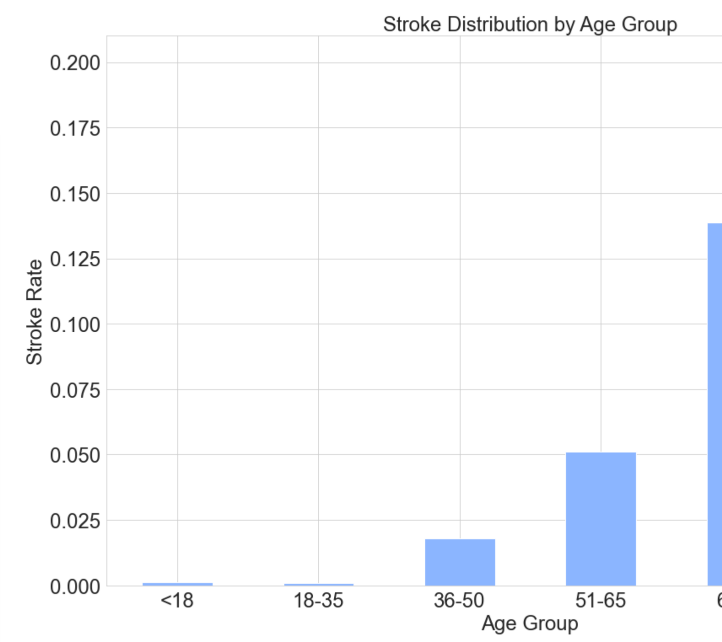

Figure 2. Stroke Distribution by Age Group

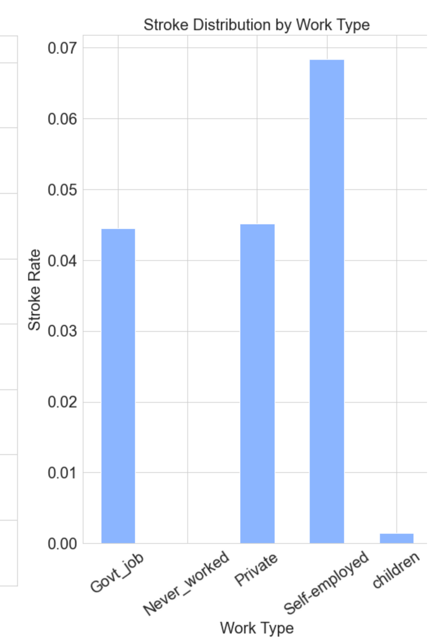

Figure 3. Stroke Distribution by Work Type

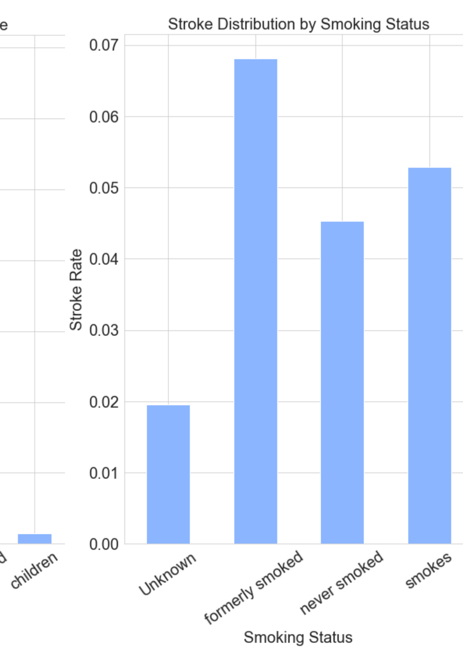

Figure 4. Stroke Distribution by Smoking Status

Stroke risk increases with age, peaking at **20% for individuals aged 80+**, demonstrating a clear upward trend.

Different **work type** also shows different risks of stroke, **self-employed** is **the highest stroke rate group.**

Interestingly, regarding smoking status, the **formerly smoked group** is at **higher stroke risk** compared to both current smokers and those who never smoked.

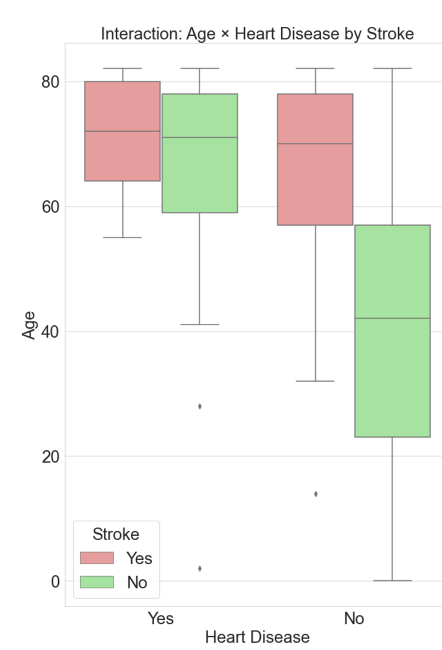

Figure 5. Interaction: Age x Heart Disease by Stroke

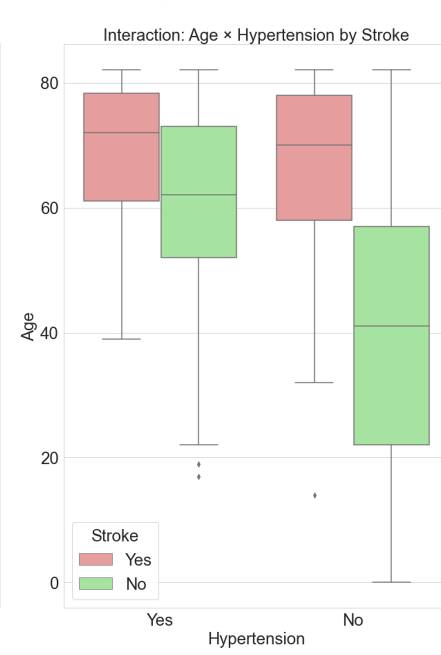

Figure 6. Interaction: Age x Hypertension by Stroke

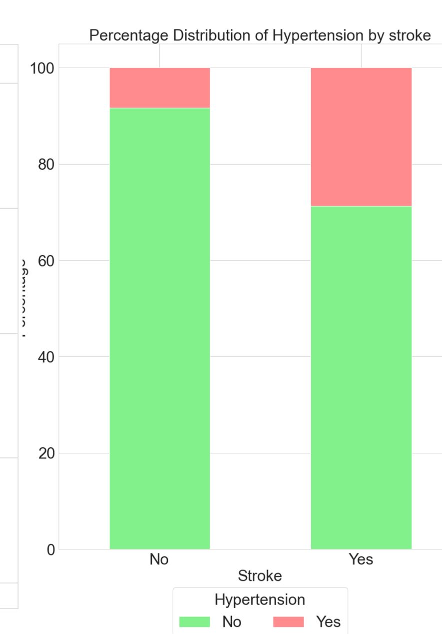

Figure 7. Percentage Distribution of Hypertension by Stroke

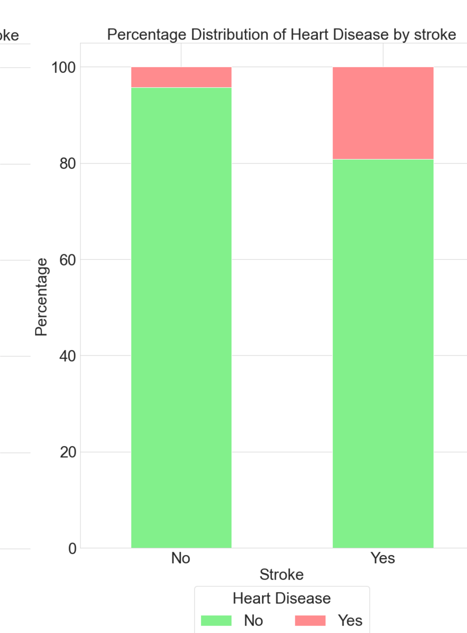

Figure 8. Percentage Distribution of Heart Disease by Stroke

Age plays a significant role in stroke risk, especially for those with heart disease. People with **heart disease** are **5.24 times more likely to have a stroke**, and those with **hypertension** are **4.44 times more likely** compared to individuals without these conditions.

VARIABLE	ODDS RATIO
Hypertension	4.437769
Heart disease	5.243245

Table 1. Odds Ratio of Binary Variables

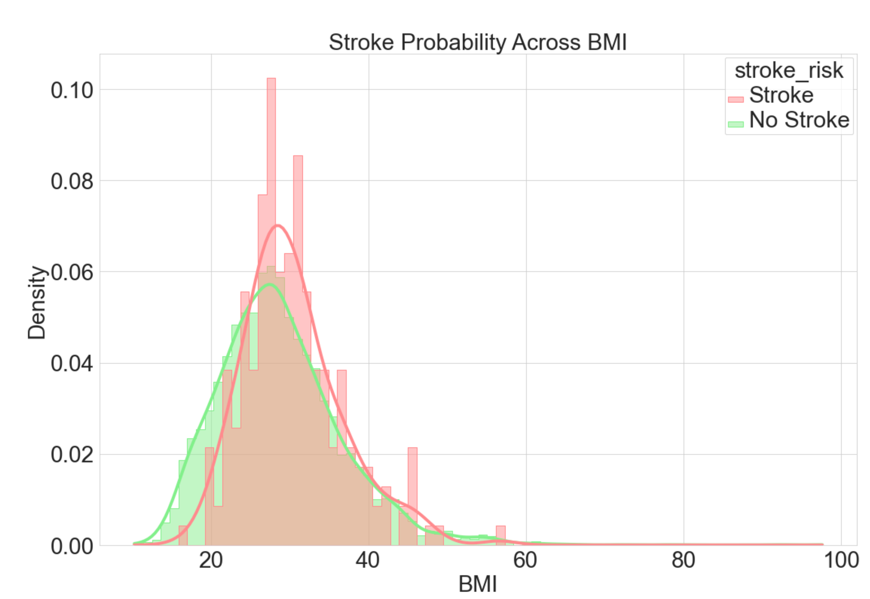

Figure 9. Stroke Probability across BMI

There's a significant overlap between the two distributions of the risk groups across BMI, indicating that **BMI index alone is not a perfect predictor of stroke risk**.

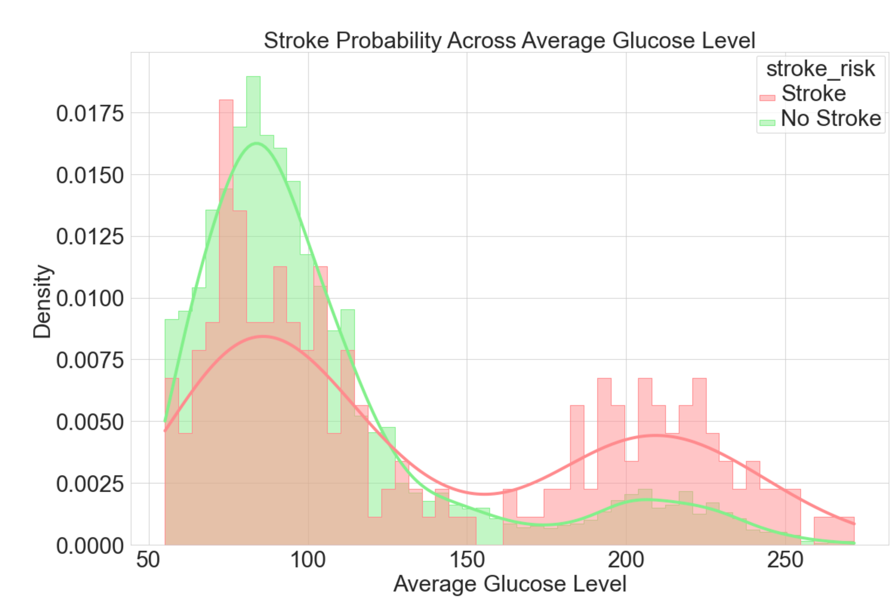

Figure 10. Stroke Probability across Average Glucose Level

There is a noticeable shift towards higher glucose levels in the stroke group, which suggests a **potential association** between **higher average glucose levels and increased stroke risk.**

C. RESULTS & DISCUSSION

CHOSEN MODEL	Random Forest
DATASET	Full
CLASSIFICATION THRESHOLD	0.4

Table 2. Chosen model

CLASS	PRECISION	RECALL	F1-SCORE	SUPPORT
0 (No Stroke)	0.99	0.69	0.82	940
1 (Stroke)	0.11	0.83	0.19	42
Accuracy			0.70	982
Macro Avg	0.55	0.76	0.50	982
Weighted Avg	0.95	0.70	0.79	982

Table 3. Classification report of the chosen model

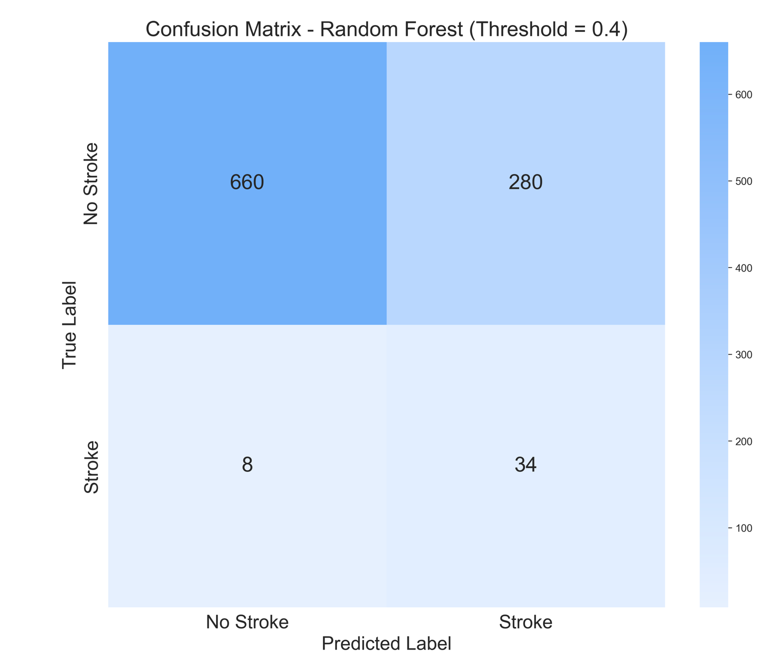

Figure 11. Confusion Matrix

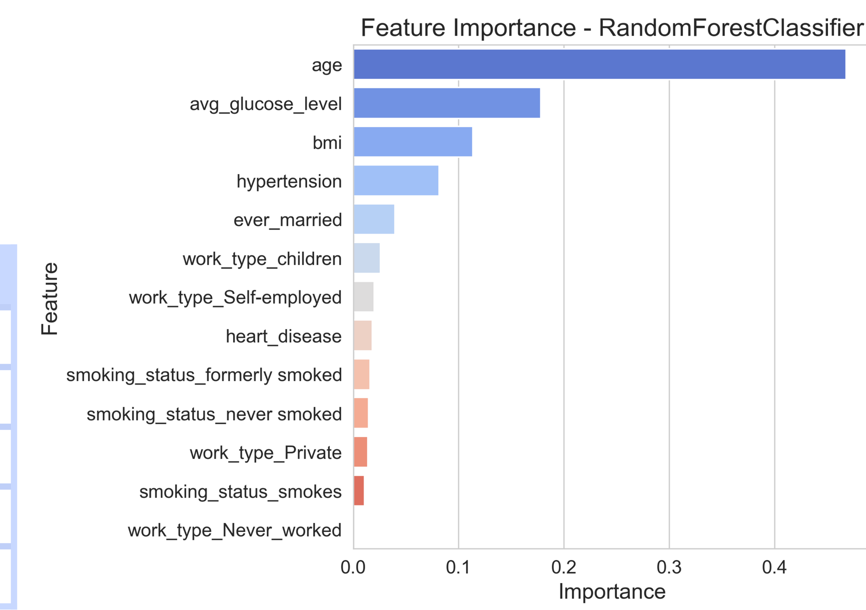

Figure 12. Feature Importance – Random Forest

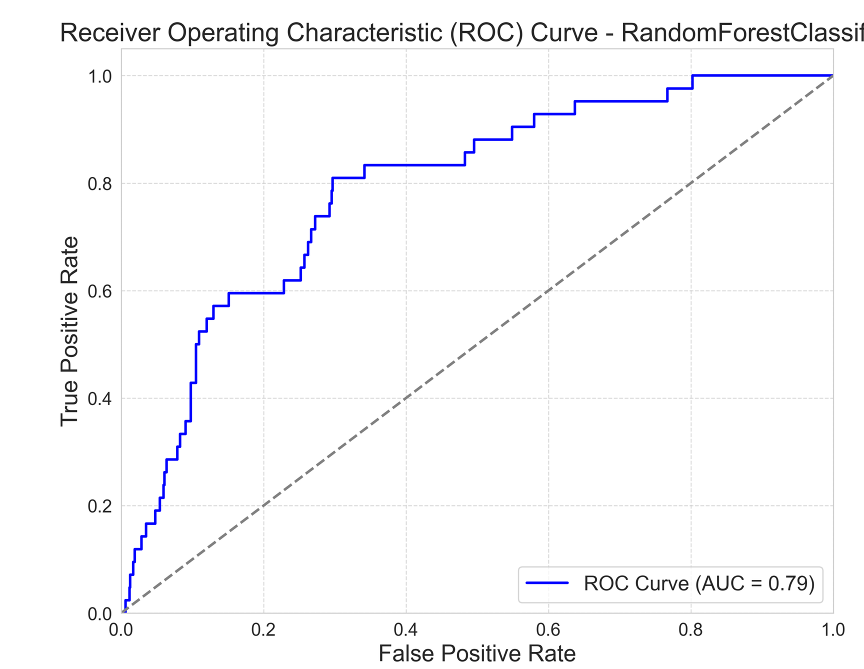

Figure 13. AUC – ROC Curve – Random Forest

MODEL PERFORMANCE METRICS

- **High recall for stroke cases (0.83):** The model effectively detects stroke cases, reducing the risk of false negatives, which is critical in healthcare.
- **Low precision for stroke cases (0.11)**: Many non-stroke patients are incorrectly classified as having a stroke, leading to excessive false positives.

CLASSIFICATION THRESHOLD

Multiple classification thresholds from 0.1 to 0.5 were tested, the optimal value was set to **0.4** so the model could reduce the number of false negatives.

CONFUSION MATRIX

False Negatives (8 cases): This is relatively low, meaning the model is effective at capturing strokes.

False Positives (280 cases): A significant number of non-stroke cases are classified as strokes, leading to unnecessary alarms and potential misallocation of medical resources.

FEATURE IMPORTANCE

The feature importance analysis highlights that **age** is the most influential predictor, reinforcing the well-established correlation between aging and stroke risk. **Average glucose level** and **BMI** also play significant roles, indicating that metabolic factors are crucial in stroke prediction. **Hypertension** emerges as another critical factor, aligning with medical knowledge of its impact on cardiovascular conditions. While **smoking status, work type, marital status**, and **heart disease** contribute to the model, their influence is comparatively smaller.

AUC-ROC CURVE

The AUC score is **0.79**, which suggests a **moderately strong** ability to distinguish between stroke and non-stroke cases.

DISCUSSION

Due to the highly imbalanced nature of our dataset and the critical implications in healthcare, where **false negatives are significantly more dangerous and costly than false positives,** minimizing missed stroke cases is our top priority. Failing to identify a stroke can lead to severe health consequences, making the trade-off between false negatives and false positives a key consideration in our study.

Our approach prioritizes **reducing the number of patients falsely classified as non-stroke**, rather than optimizing for correctly classifying non-stroke cases. The chosen model was selected based on a careful balance between **precision and recall.** In practice, this challenge can be mitigated by using more balanced datasets or ensuring sufficient representation of stroke cases.

In terms of real-world application, we hope that this model could help detect strokes early, enabling timely intervention. It can also be integrated into health screenings or wearable devices for proactive risk management.