FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

SEMESTER 2 SESSION 2024/2025

CSC4600-1

DATA MINING

STROKE PREDICTION FOR EARLY INTERVENTION AND PATIENT STRATIFICATION

Prepared by Group 5:

NO	NAME	MATRIC
1.	MOHAMMAD JAVAN SAMBOEPUTRA HERLAMBANG	213878
2.	MUHAMMAD HANIF MURTAZA	213793
3.	YANG ZIXUN	213844
4.	CAI ZESHOU	213733

Table of Content

1. Abstract

Stroke remains a leading cause of mortality and long-term disability in Malaysia, exerting significant pressure on public healthcare systems. Early detection of stroke risk is crucial for enabling timely intervention and reducing the burden of emergency care. This project proposes a data mining approach using the publicly available Stroke Prediction Dataset from Kaggle to develop a predictive model capable of identifying individuals at high risk of experiencing a stroke. The dataset includes demographic, medical, and lifestyle attributes such as age, hypertension status, heart disease, BMI, smoking habits, and glucose levels. The goal is to build and compare multiple machine learning models—Logistic Regression, Random Forest, and a third selected classifier—to determine the most effective method for binary stroke prediction. The outcomes of this project are intended to assist the Malaysian Ministry of Health (MOH) and public hospitals in implementing proactive screening strategies, improving resource allocation, and ultimately enhancing patient outcomes through early preventive care.

2. Introduction

2.1. Project Overview

Stroke is a critical public health issue that affects millions worldwide, often leading to permanent disability or death. In Malaysia, the rising incidence of stroke has placed increasing pressure on the public healthcare system, especially in emergency care and long-term rehabilitation. This project aims to leverage data mining and machine learning to predict stroke risk based on a combination of patient demographics, medical history, and lifestyle factors. By building predictive models using real-world health data, the project seeks to assist healthcare providers in identifying high-risk individuals early, enabling timely intervention and reducing the need for reactive treatment.

2.2. Introduction to the Entity: Ministry of Health Malaysia (MOH)

The Ministry of Health Malaysia (MOH) is responsible for delivering comprehensive healthcare services to the nation’s population. As part of its mandate, the MOH is actively pursuing digital health transformation, which includes data-driven approaches to disease prevention and patient care.

Early stroke prediction aligns directly with the MOH’s mission to reduce non-communicable disease burdens and improve nationwide health outcomes. By integrating predictive analytics into routine screening processes, public hospitals can shift from reactive treatment to preventive care. Data mining is particularly valuable for this goal, as it can uncover hidden patterns in patient health records that correlate with stroke risk—patterns that might not be evident through traditional diagnostics alone.

The benefits of this project for the MOH and public hospitals include:

- **Proactive screening**: Enabling healthcare staff to focus resources on high-risk individuals.
- **Resource optimization**: Helping hospitals plan staffing, monitoring, and intervention strategies more efficiently.
- **Cost savings**: Reducing long-term care and emergency response costs through earlier prevention.
- **Policy insight**: Supporting the development of targeted public health campaigns and national preventive strategies.

This initiative also aligns with broader public health trends and career interests in health informatics, digital transformation, and AI applications in healthcare—fields that are rapidly expanding and hold substantial impact potential.

2.3. Problem Statement

Stroke remains a major public health concern in Malaysia, ranking as the third leading cause of death and a significant contributor to long-term disability and rising healthcare costs. According to the National Health and Morbidity Survey, stroke is among the top five causes of premature death—with an alarming rise in cases, especially among younger individuals (Hwong et al., 2021). While acute care has improved, many strokes are still diagnosed too late, when damage is irreversible and treatment becomes more expensive and less effective.

A core issue faced by the Ministry of Health (MOH) and public hospitals is the absence of efficient, data-driven tools to detect high-risk individuals early. Although many risk factors—such as high blood pressure, diabetes, heart disease, and smoking—are well known, the challenge lies in turning patient data into timely insights for intervention. Most healthcare systems rely on manual checks or basic screenings, which are often insufficient for large-scale prevention.

Machine learning and data mining offer a powerful solution by analyzing patterns in existing health and demographic data to predict stroke risk before symptoms appear. This approach enables a shift from reactive treatment to preventive action—allowing healthcare providers to identify vulnerable patients earlier, optimize resource use, and reduce long-term burdens on the healthcare system.

2.4. Project Objectives

This project aims to develop an effective stroke prediction system using patient health data. The primary goal is to support early detection and intervention by leveraging machine learning techniques to identify individuals at elevated risk of stroke.

To achieve this, a series of data processing and model development objectives were defined:

- Dataset Preparation: Understand and clean the dataset to ensure data quality and select features most relevant to stroke prediction.
- Exploratory Data Analysis (EDA): Analyze patterns and distributions within the dataset to uncover relationships between features and stroke risk.
- Data Preprocessing: Apply techniques such as encoding categorical variables, standardization, and handling missing values. Address class imbalance issues using resampling methods to ensure the models are not biased toward the majority class.
- Model Training and Comparison: Train and evaluate multiple machine learning models to identify which perform best for the given dataset and problem context.

- Performance Evaluation: Compare models based on key metrics including accuracy, recall, F1 score, and ROC AUC to determine overall effectiveness.

- Recall Optimization: Improve the recall of selected models to ensure that the majority of actual stroke cases are identified, without significantly compromising interpretability or increasing false positives.

Ultimately, the system is designed to:

- Predict the likelihood of a patient experiencing a stroke, with a strong emphasis on identifying actual cases.
- Flag high-risk individuals to enable timely medical intervention and prevent adverse outcomes.
- Support risk stratification by grouping patients into low, medium, and high-risk categories, aiding healthcare providers in delivering targeted care and efficient resource allocation.

This predictive modeling approach supports the broader healthcare goal of transitioning from reactive treatment to proactive prevention.

3. Literature Review

3.1. The Public Health Burden of Stroke

Stroke is a major cause of mortality and long-term disability globally. It imposes a significant clinical and socioeconomic burden, particularly in low- and middle-income countries where access to early diagnostic tools and post-stroke care is often uneven. In Malaysia, the urgency of tackling stroke has become increasingly evident. According to Hwong et al. (2021), the incidence of stroke has shown an upward trend over recent years, with ischemic stroke accounting for the majority of cases. Despite slight improvements, the 28-day all-cause mortality rate after a stroke remains concerning, indicating systemic issues in early diagnosis, timely hospital admissions, and access to quality care. The study further revealed that age-standardized incidence rates are consistently high for both men and women, underscoring the need for scalable national intervention strategies.

These findings underscore the importance of shifting healthcare resources toward **preventive solutions and early detection mechanisms**, especially for vulnerable populations. Without significant improvements in stroke prediction and early warning systems, healthcare systems risk becoming overwhelmed by long-term care demands, rehabilitation costs, and productivity losses.

3.2. Opportunities for Machine Learning in Stroke Prediction

In response to these challenges, researchers have turned to **machine learning (ML)** and **data mining techniques** as tools to augment traditional healthcare systems. ML models are capable of analyzing large volumes of heterogeneous patient data and uncovering hidden patterns that may not be immediately obvious to clinicians. As Sathya (2024) discusses, artificial intelligence—including ML—can enhance the accuracy of disease prediction, support real-time decision-making, and personalize treatment pathways.

In the context of stroke, **predictive analytics** can serve a vital role in identifying individuals at high risk before the onset of symptoms. This enables clinicians and policymakers to implement targeted interventions and allocate resources more efficiently. Li (2024) demonstrated the feasibility of using **logistic regression** to build interpretable stroke prediction models, showing promising results using simple clinical and demographic features. Logistic regression remains a preferred baseline model in healthcare applications due to its transparency and explainability. However, ensemble methods such as **Random Forest** and advanced models like **XGBoost** often outperform simpler models when dealing with complex feature interactions and nonlinear patterns.

Nonetheless, the trade-off between performance and interpretability must be carefully considered, particularly in clinical environments where transparent decision-making is critical for ethical and legal accountability.

3.3. Data Source and Feature Relevance

This study utilizes the **Stroke Prediction Dataset** published by fedesoriano on Kaggle (2020), which has been widely used in academic stroke risk modeling. The dataset includes key predictors of stroke such as **age**, **hypertension**, **heart disease**, **average glucose level**, **BMI**, and lifestyle variables like **smoking status** and **work type**. These variables have been shown in prior studies to correlate significantly with stroke outcomes (Hwong et al., 2021; Li, 2024).

Using this dataset, features were selected and engineered for predictive modeling. Certain preprocessing steps—such as handling missing BMI values, capping outliers in glucose and BMI, and encoding categorical variables—were necessary to ensure model stability and reliability. The ability to access structured, labeled data with relevant attributes is essential for training effective models and drawing clinically meaningful conclusions.

3.4. The Role of Predictive Models in Malaysian Healthcare

Malaysia, like many developing countries, faces **healthcare access disparities** between urban and rural populations, as highlighted by Hwong et al. (2021). Predictive models can bridge this gap by enabling **low-cost, scalable screening tools** that use readily available data to assess stroke risk. These tools can be deployed in clinics, community centers, or even integrated into telemedicine platforms.

Moreover, stroke prediction models trained on national or regional datasets can aid in **population-level surveillance** and **resource allocation planning**. By identifying clusters of high-risk patients, public health agencies can direct educational programs, preventive care, and emergency preparedness to the right areas, thus reducing the overall stroke burden.

3.5. Summary

In summary, stroke remains a pressing public health issue, especially in Malaysia. The increasing availability of high-quality structured datasets, combined with advancements in machine learning, has opened up new possibilities for stroke prediction. Studies such as those by Hwong et al. (2021), Li (2024), and Sathya (2024) affirm that data-driven approaches are not only feasible but also essential for the future of preventive medicine. As this project demonstrates, the application of predictive models—supported by datasets like fedesoriano's (2020)—can provide early risk identification and support more effective clinical and public health decision-making.

4. Proposed Method

4.1. Data Inventory

Demographics: gender, age, marital status, residence type

Medical History: hypertension, heart disease

Lifestyle and Behavioral: smoking status, work type

Clinical Measurements: BMI, average glucose level

Additional Data: EHR, wearables, socioeconomic data (future enhancement)

4.2. Data Engineering for Data Mining Tasks

- **Two-class classification**

Attributes to use:
age, hypertension, heart_disease, avg_glucose_level, bmi, smoking_status, gender, ever_married, work_type, Residence_type
Support for the task:
These attributes are critical for classifying patients into "stroke" or "no stroke" categories. Clinical indicators like hypertension, heart disease, and glucose level are major risk factors. Lifestyle and demographic factors such as age, smoking, and occupation also influence the risk. The combination of medical, behavioral, and social attributes enables effective binary classification for early screening.

- **Multi-class classification**

Attributes to use:
All from two-class classification, plus derived risk categories (e.g., “Low”, “Moderate”, “High”), and possibly medication_adherence, follow_up_frequency
Support for the task:

Supports risk stratification rather than just binary classification. For example, a model could classify patients into different risk groups to inform monitoring intensity. Derived categories based on clinical thresholds or score cutoffs provide more granular insight for clinicians and triage systems.

- **Image classification**

Attributes to use:
MRI scans, CT brain imaging (if available)
Support for the task:
With imaging data, models could detect patterns indicating stroke occurrence or damage in brain scans. Although not included in the current dataset, this would support diagnostic automation and stroke type classification in future enhancements.

- **Text Analytics**

Attributes to use:
Doctor’s notes, discharge summaries, referral letters, EHR free-text fields
Support for the task:
Unstructured text can be processed to extract symptoms, clinical impressions, and historical risk indicators (e.g., “elevated BP”, “TIA history”). This supports deeper insights when structured data is limited or incomplete.

- **Regression**

Attributes to use:
Same clinical and demographic features; target becomes continuous stroke risk score (e.g., 0.0–1.0 or predicted probability)
Support for the task:
Regression models can predict a continuous stroke risk score rather than a binary label, allowing hospitals to rank patients by urgency and prioritize follow-up for borderline cases.

- **Recommenders**

Attributes to use:
bmi, hypertension, glucose_level, smoking_status, age, heart_disease, past intervention success records
Support for the task:
Personalized recommendation systems could suggest health interventions

(e.g., diet, exercise, medication adherence programs) based on individual risk factors and medical history. Useful in prevention-focused mobile apps or dashboards.

- **Clustering**

Attributes to use:
age, glucose_level, bmi, smoking_status, hypertension, work_type, Residence_type, ever_married
Support for the task:
Clustering helps group patients with similar profiles. For example, elderly smokers with high glucose levels may form a cluster with high risk. This can support population segmentation for targeted outreach and interventions.

- **Anomaly Detection**

Attributes to use:
bmi, glucose_level, age, heart_disease, hypertension
Support for the task:
Anomaly detection can flag outlier patients, such as extremely high glucose levels or very young patients with high-risk profiles. This is useful for quality control and identifying unexpected high-risk individuals.

- **Time Series Analysis**

Attributes to use:
Repeated measures of blood pressure, glucose, heart rate, and medication history over time (requires future data with timestamps)
Support for the task:
Tracks patient vitals over time to identify trends that may indicate increasing stroke risk. For example, rising glucose levels or missed medications could signal deteriorating health. Enables predictive alerts for clinicians.

- **Association Rule Mining**

Attributes to use:
hypertension, heart_disease, smoking_status, glucose_level, age_group, bmi_category, stroke
Support for the task:
Helps discover patterns like “Hypertension + Smoking + High Glucose →

Stroke”, which are valuable for education, policy planning, and awareness campaigns. These rules can be generated from encoded categorical data.

The two selected data mining tasks are:

- **Two-class classification**
- **Anomaly detection**

Two-class classification is directly aligned with the project's goal of predicting whether a patient is at risk of experiencing a stroke (labelled as 0 or 1). This task is critical for enabling early intervention and supports the Malaysian Ministry of Health’s preventive healthcare initiatives.

Anomaly Detection is chosen as a complementary task to identify unusual or high-risk patient profiles that deviate significantly from the general population. For instance, individuals with abnormally high glucose levels, extreme BMI values, or uncommon combinations of health risk factors can be flagged as outliers. Detecting these anomalies helps healthcare professionals to prioritize urgent cases, conduct early intervention, and investigate potential data errors or rare clinical scenarios that require special attention.

4.3. Dataset Description

Name: Stroke Prediction Dataset

Source: Kaggle - fedesoriano

Records: 5,110

Attributes: 12

Target Variable: stroke (0 = No, 1 = Yes)

4.4. Machine Learning Methods

- **Logistic Regression**

Logistic Regression is chosen as an interpretable baseline model, suitable for binary classification tasks such as stroke prediction. It provides clear insights into how each variable contributes to the outcome. Its simplicity ensures low computational cost and helps in setting a benchmark. However, it may struggle with complex, non-linear patterns in the data.

- **Random Forest**

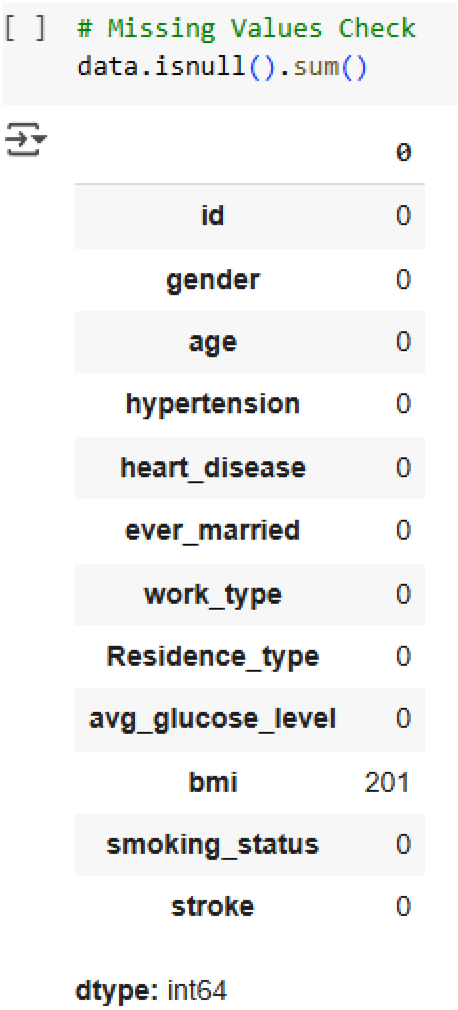
Random Forest is selected for its ability to handle both numerical and categorical features and model complex, non-linear relationships. It performs well with imbalanced data when combined with techniques like SMOTE or class weighting. The model is robust to outliers and reduces overfitting by aggregating predictions from multiple trees. Feature importance scores also enhance interpretability.

- **XGBoost**

XGBoost is included for its powerful gradient boosting framework, known for delivering high predictive accuracy. It efficiently handles large datasets, missing values, and imbalanced data through built-in regularization and scale_pos_weight. As a tree-based ensemble model, it captures complex non-linear relationships while minimizing overfitting. XGBoost's flexibility, performance, and interpretability through feature importance make it a strong candidate for medical risk prediction tasks like stroke classification.

5. Data Exploration and Preprocessing

5.1. Data Exploration

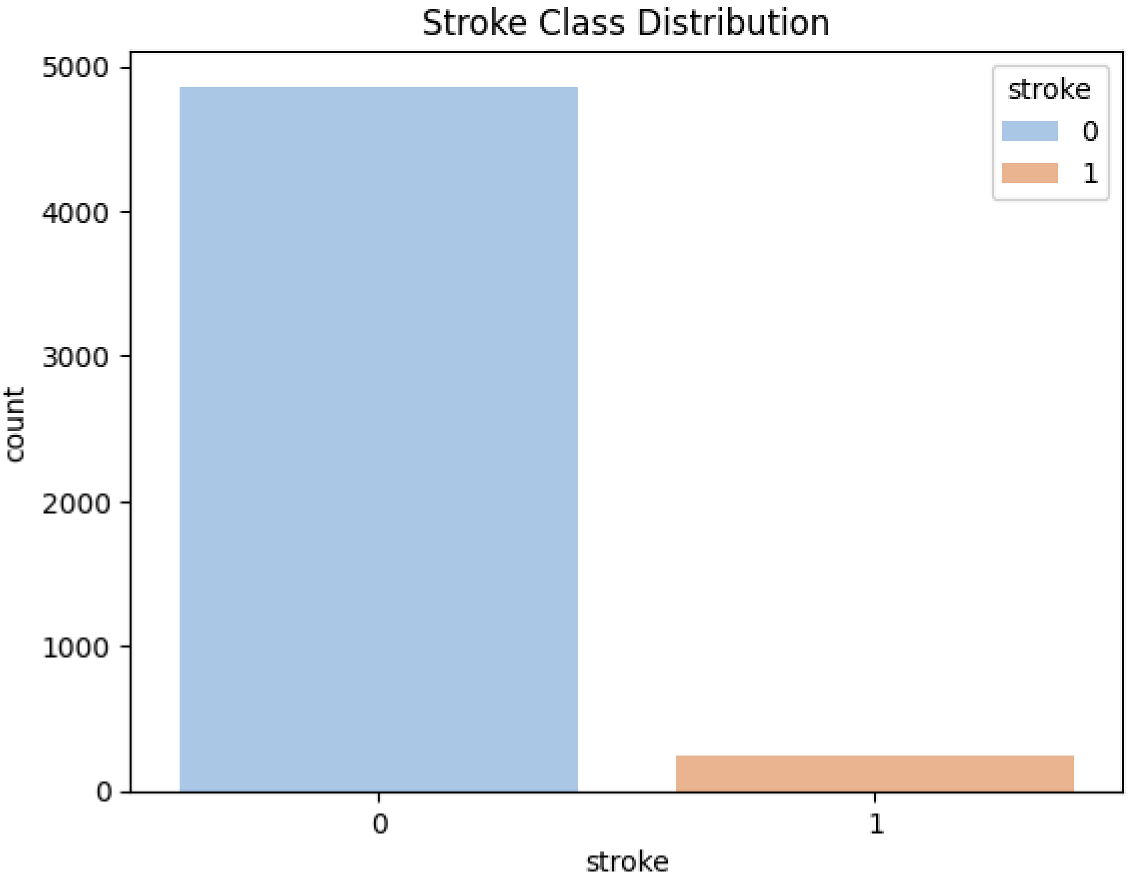


```
# Missing Values Check
data.isnull().sum()
```

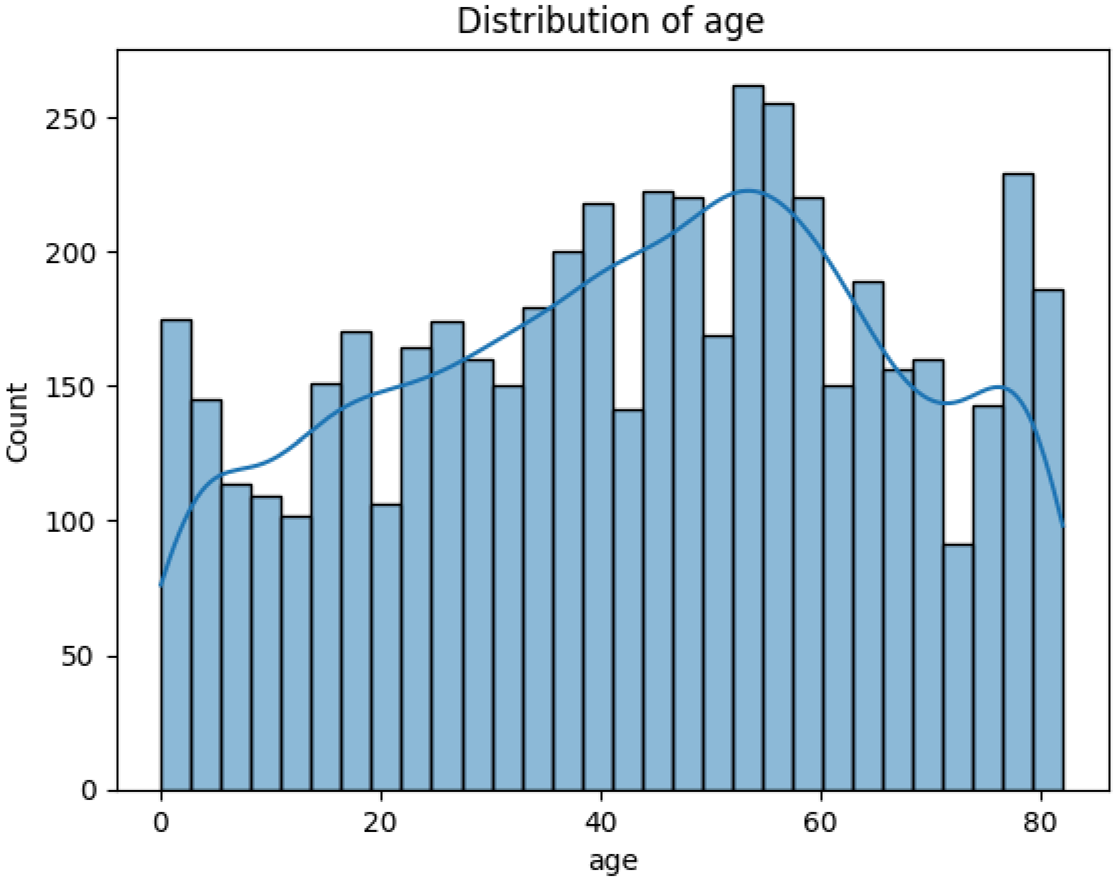
```
                     0
              id     0
          gender     0
             age     0
    hypertension     0
   heart_disease     0
    ever_married     0
       work_type     0
  Residence_type     0
avg_glucose_level    0
             bmi   201
  smoking_status     0
          stroke     0

dtype: int64
```

Upon checking for missing values in the dataset, it was found that **only one attribute**, **bmi** (Body Mass Index), had null values, with exactly **201 missing entries**. This accounts for approximately **3.93%** of the total dataset (201 out of 5,110 records).

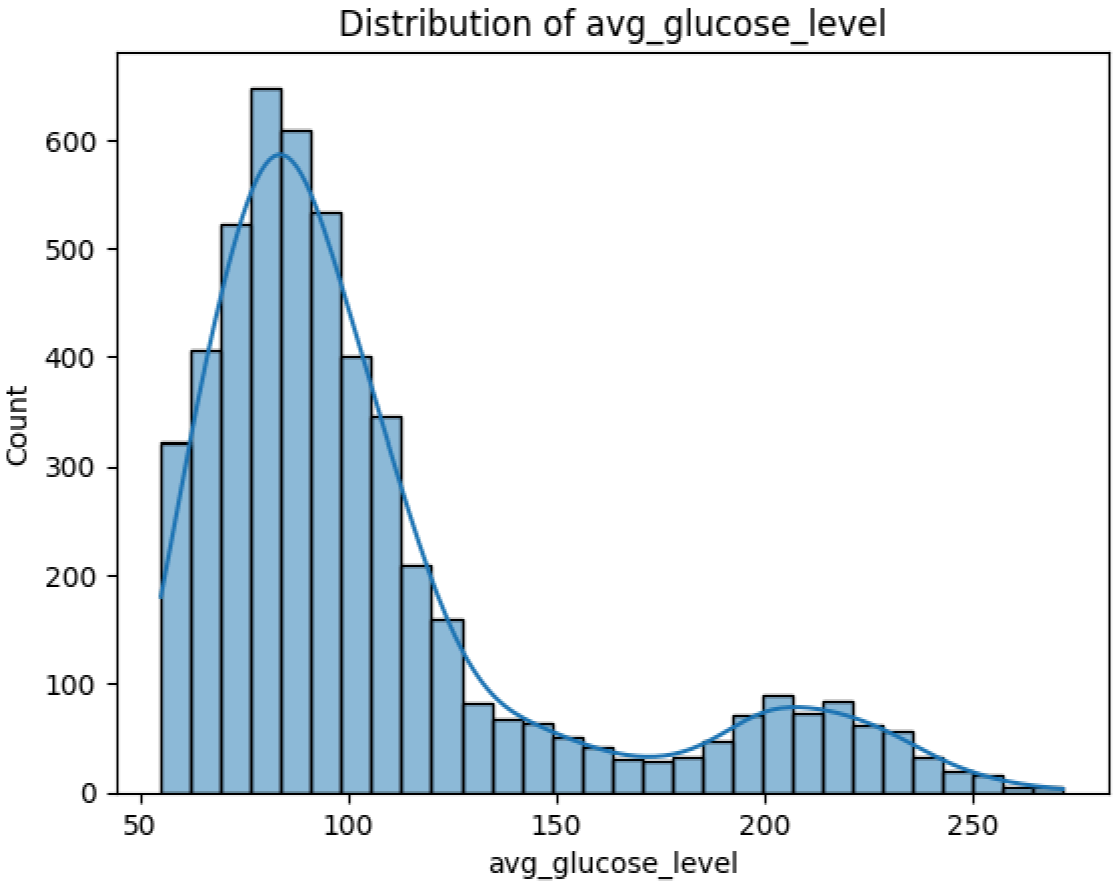
Since the proportion of missing values was relatively small and confined to a single feature, the decision was made to **change all rows containing null values** to the median value. This approach was deemed acceptable as it minimally impacted the dataset size and ensured the remaining data was clean and consistent for modeling. All other features were confirmed to have **no missing values** and were retained in full.



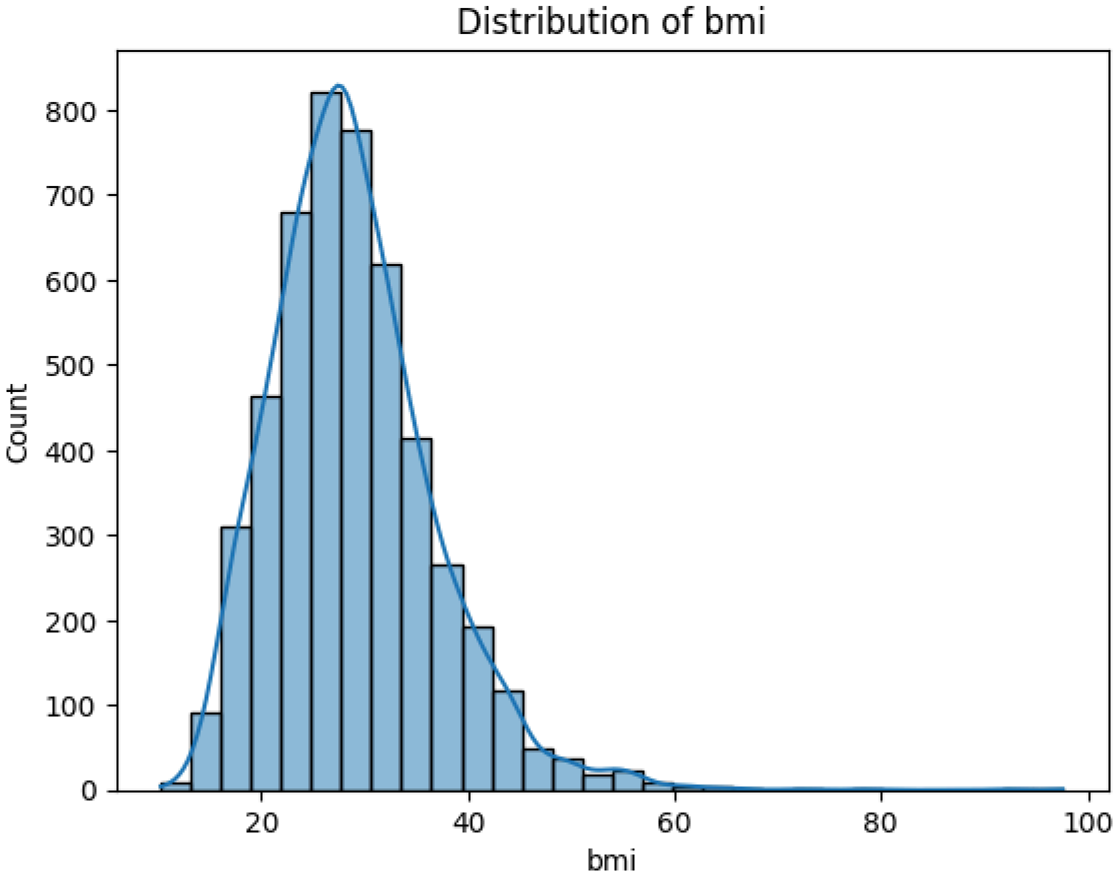
The initial exploration of the stroke dataset reveals a strong class imbalance, where the majority of samples are labeled as non-stroke (0), while only a small minority represent stroke cases (1). This imbalance is clearly visualized in the class distribution bar plot, which emphasizes the importance of applying balancing techniques like ADASYN during model training.



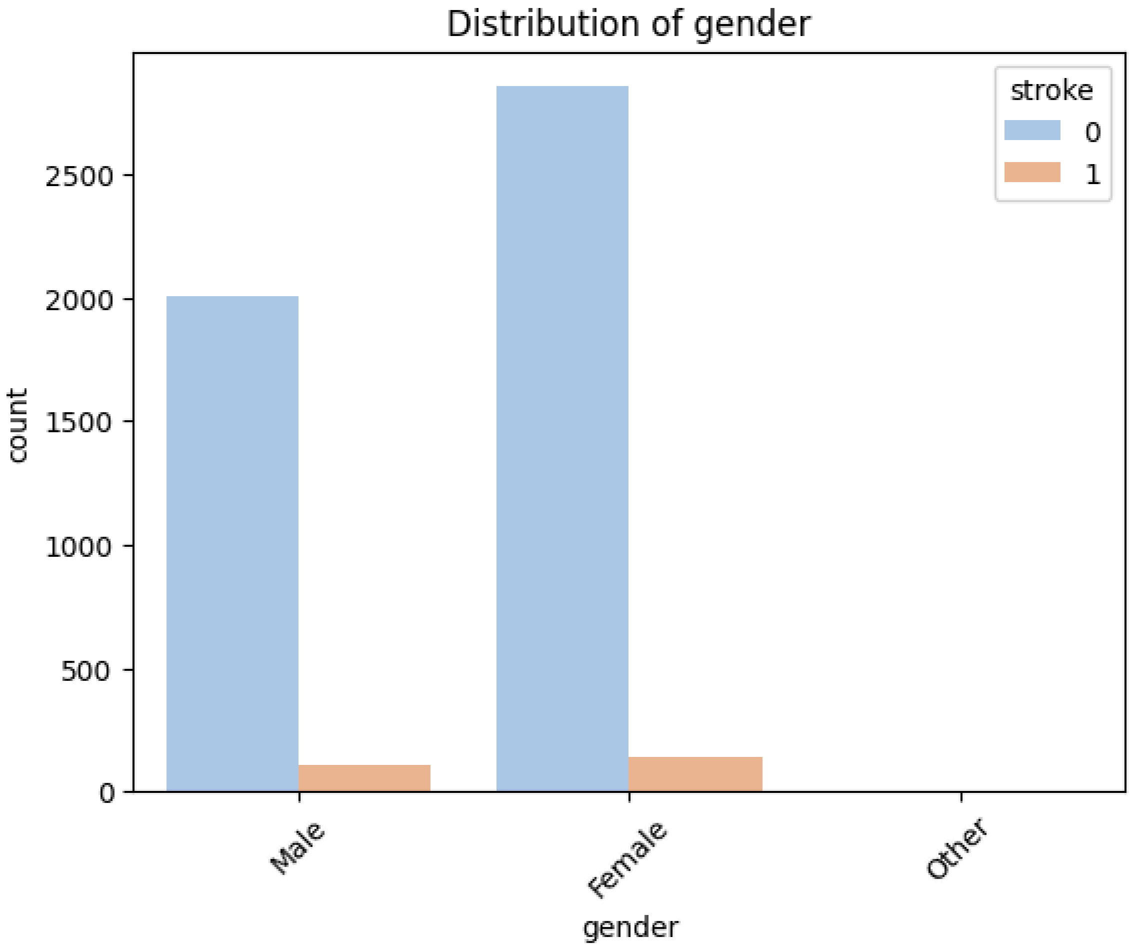
The distribution of **age** in the dataset is relatively uniform, with visible peaks among middle-aged and elderly patients. Notably, the number of stroke cases tends to increase with age, reinforcing that age is a **significant risk factor** for stroke.



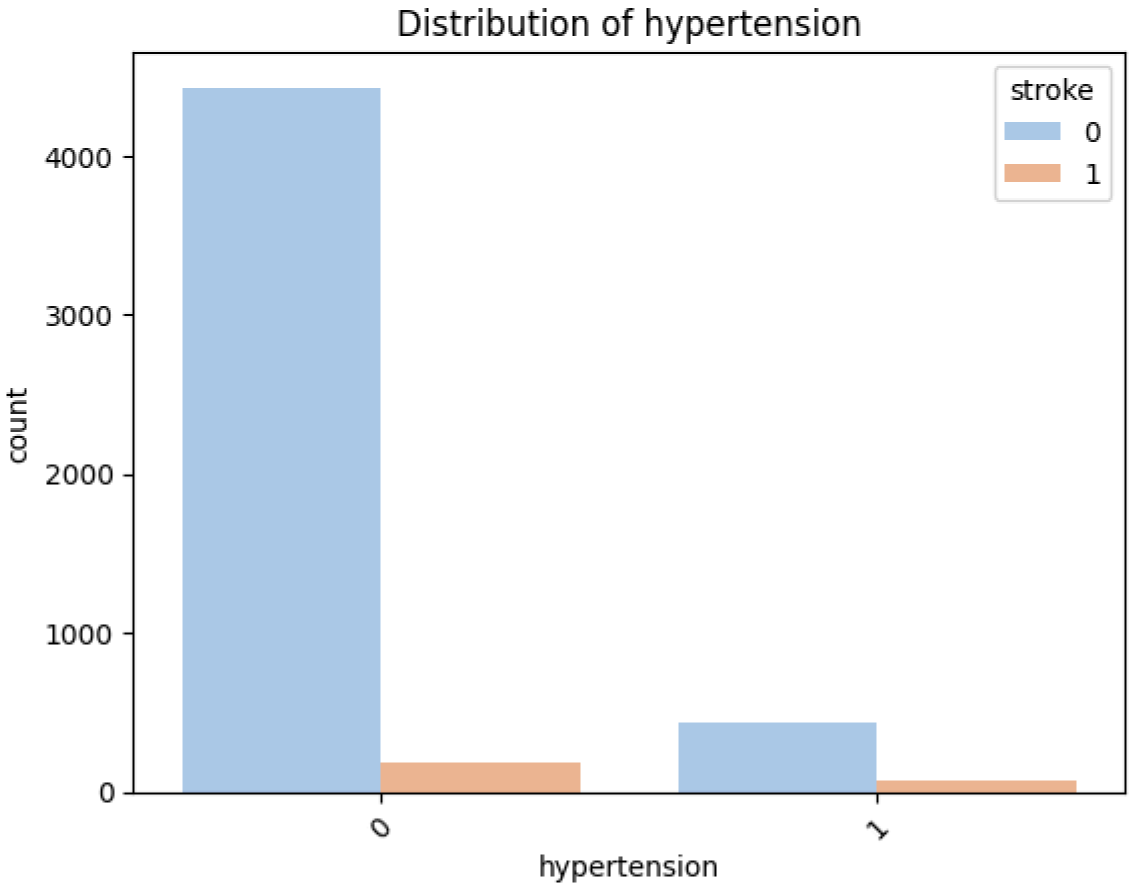
The **avg_glucose_level** feature exhibits a **right-skewed distribution**, with most patients having glucose levels below 150 mg/dL. However, the presence of a long tail extending beyond 250 mg/dL indicates the existence of **outliers**. These extreme values may affect model performance and may warrant transformation or binning during preprocessing.



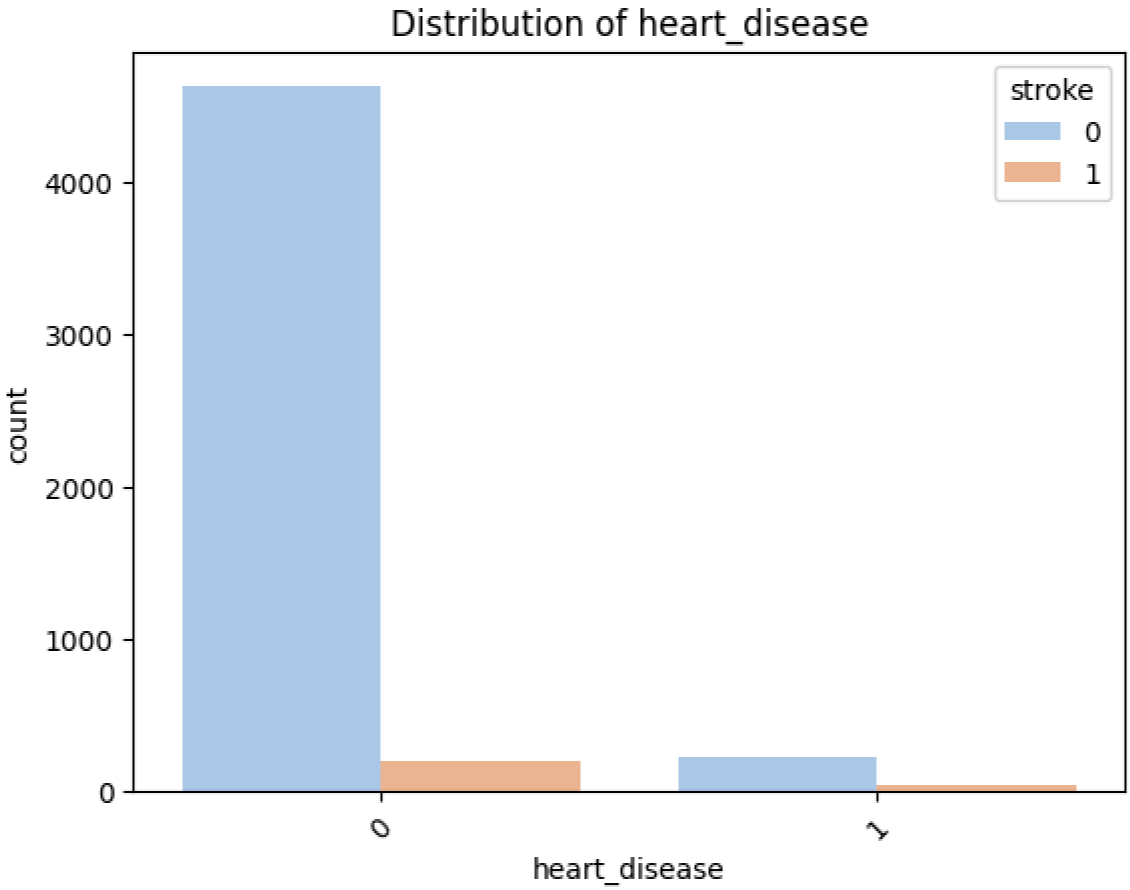
The distribution of **BMI** approximates a **normal distribution** but shows a noticeable long right tail. A few values exceed 60, which are considered **extreme outliers**. These may introduce noise into the model and were thus addressed through **value capping** during preprocessing to improve robustness.



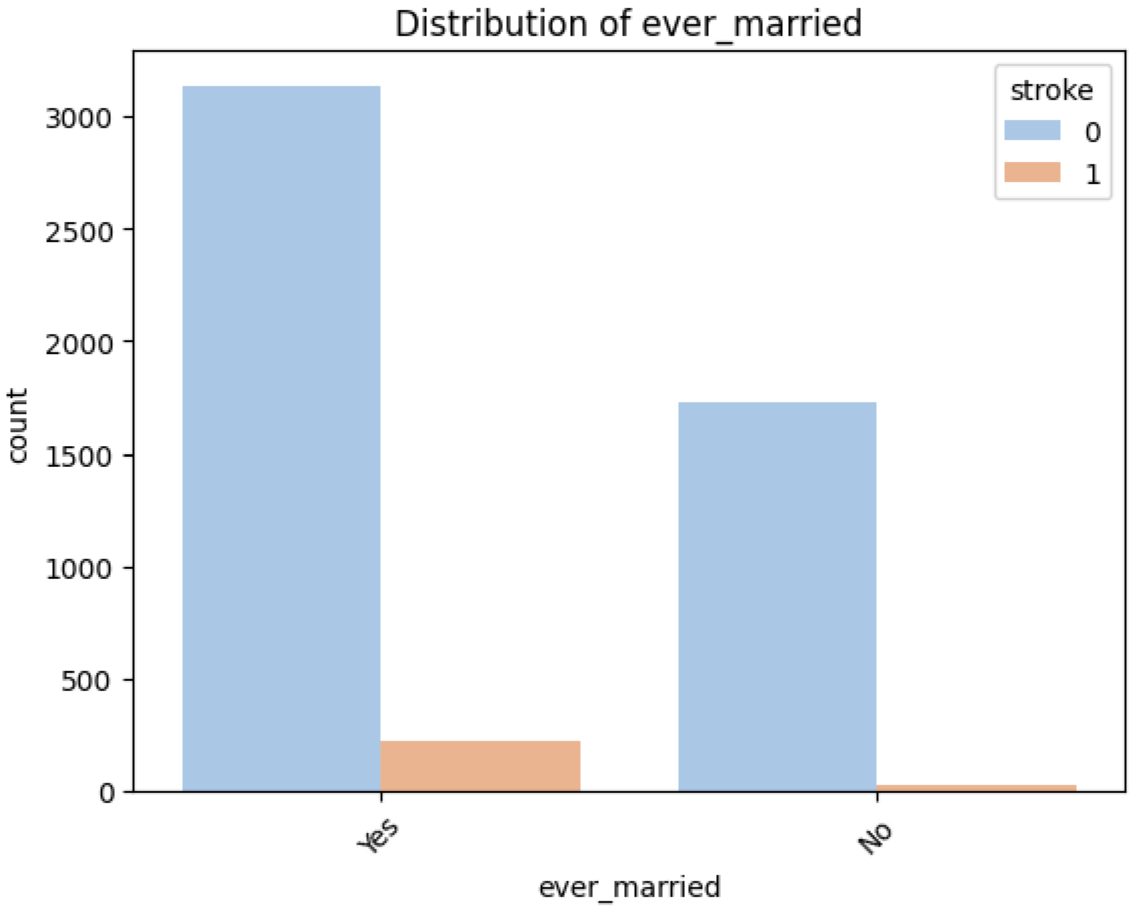
The dataset shows a **balanced distribution between male and female patients**, indicating no inherent gender bias in the data. Gender may not serve as a strong predictive factor for stroke occurrence.



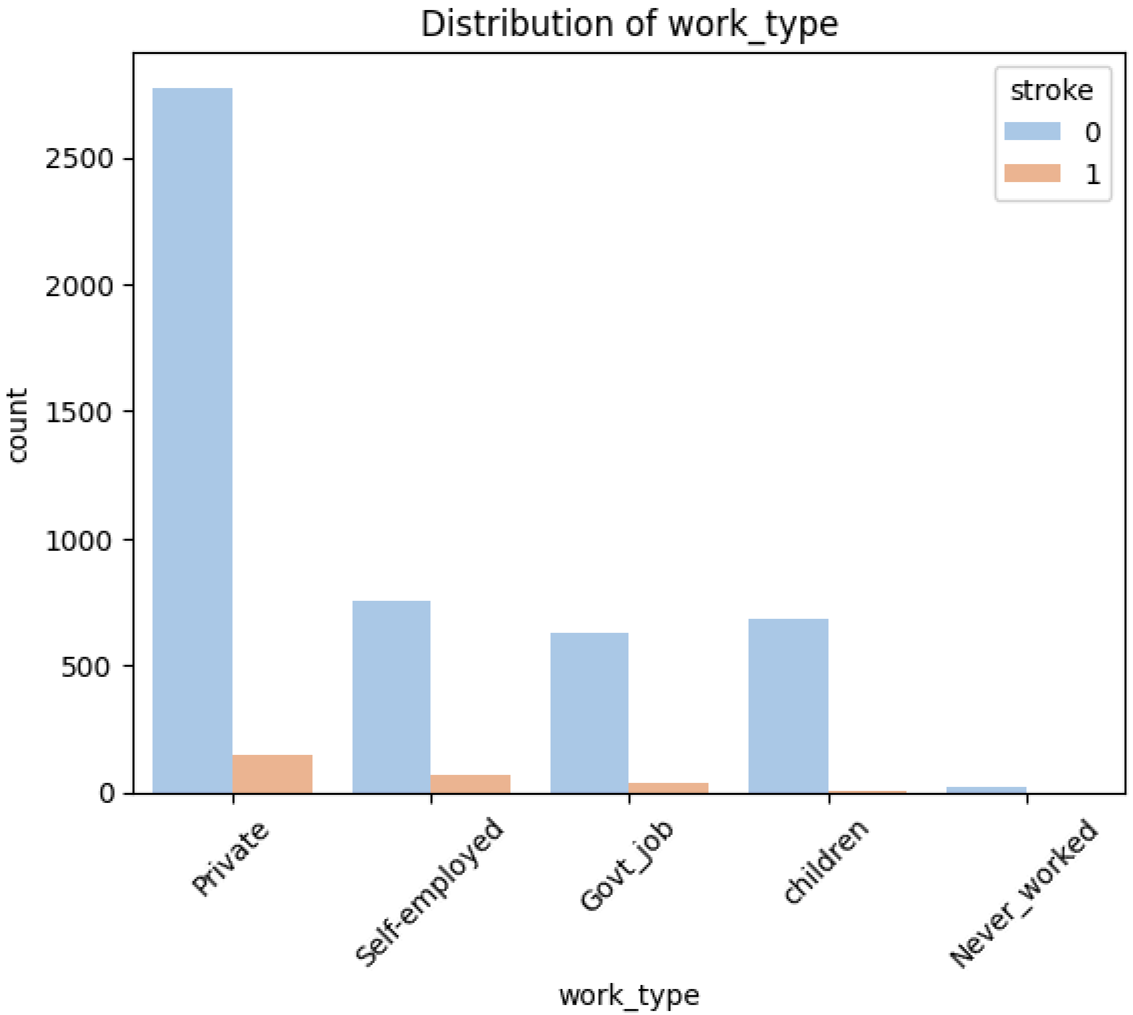
Although individuals with **hypertension** constitute a minority in the dataset, they exhibit a **higher proportion of stroke cases**. This suggests that hypertension is a **strong predictor** for stroke risk.



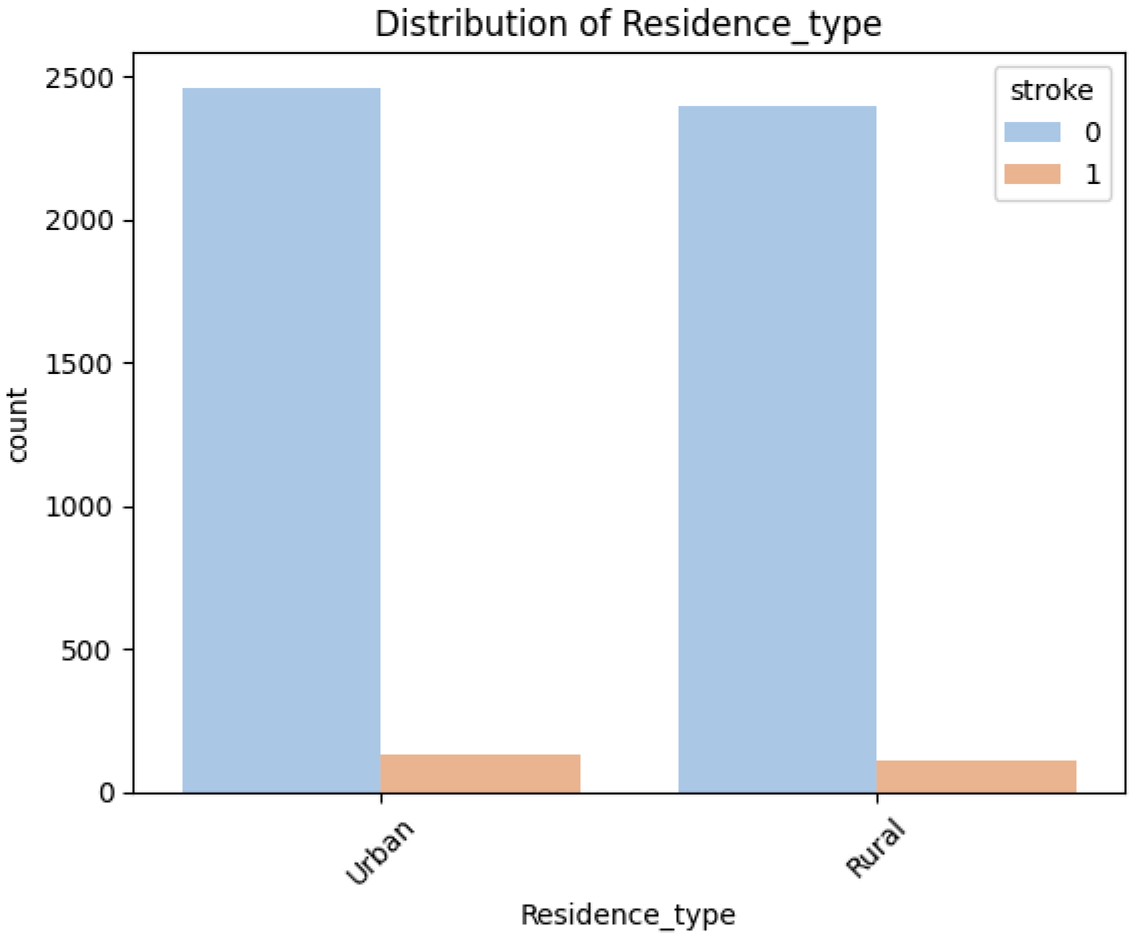
Patients with **heart disease** also demonstrate a **higher incidence of stroke**. Similar to hypertension, this attribute is likely a **significant contributing factor** in stroke prediction models.



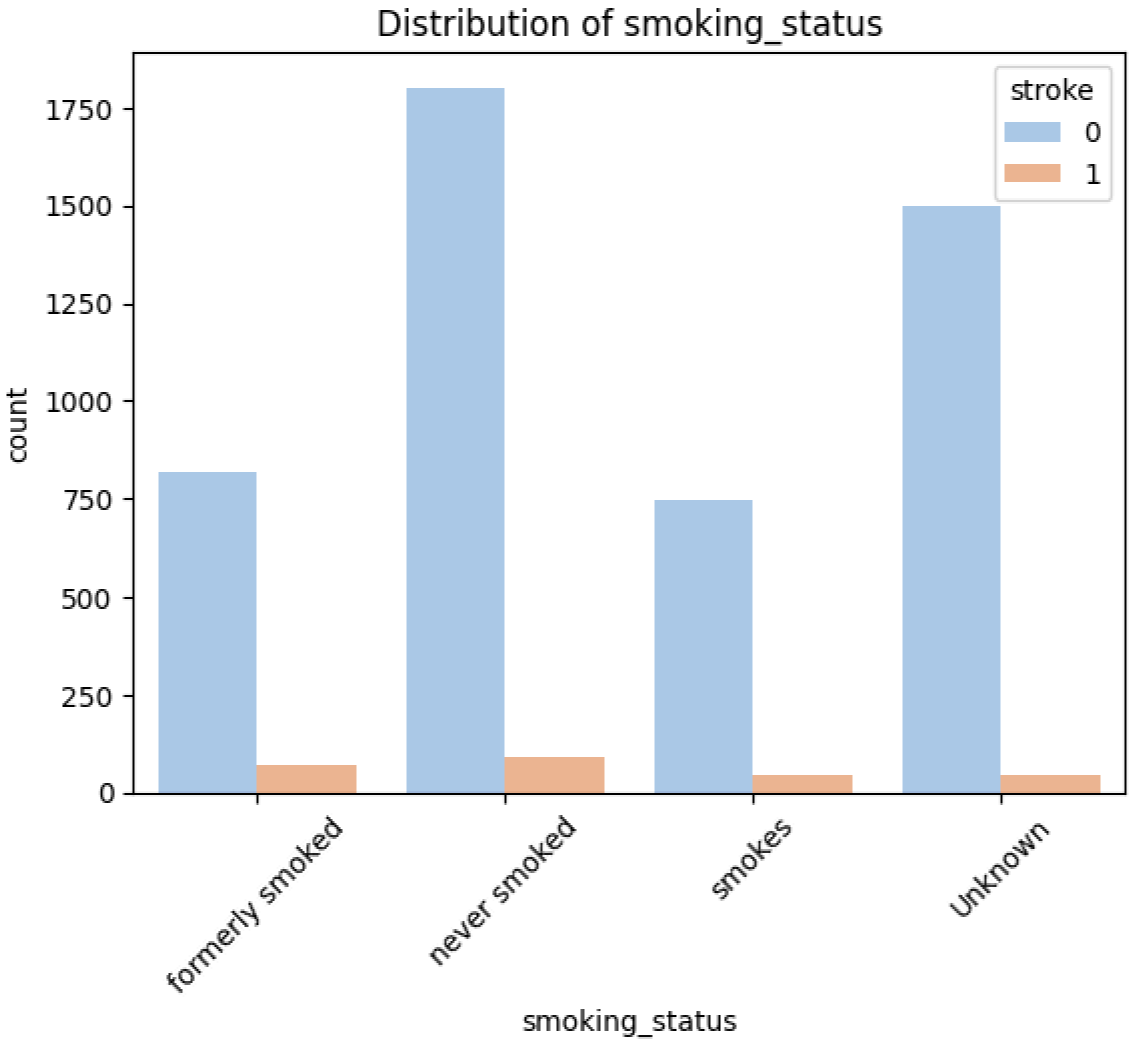
A majority of stroke cases are found among those who have been **married**. However, this relationship may be **confounded by age**, as older individuals are both more likely to be married and more prone to strokes.



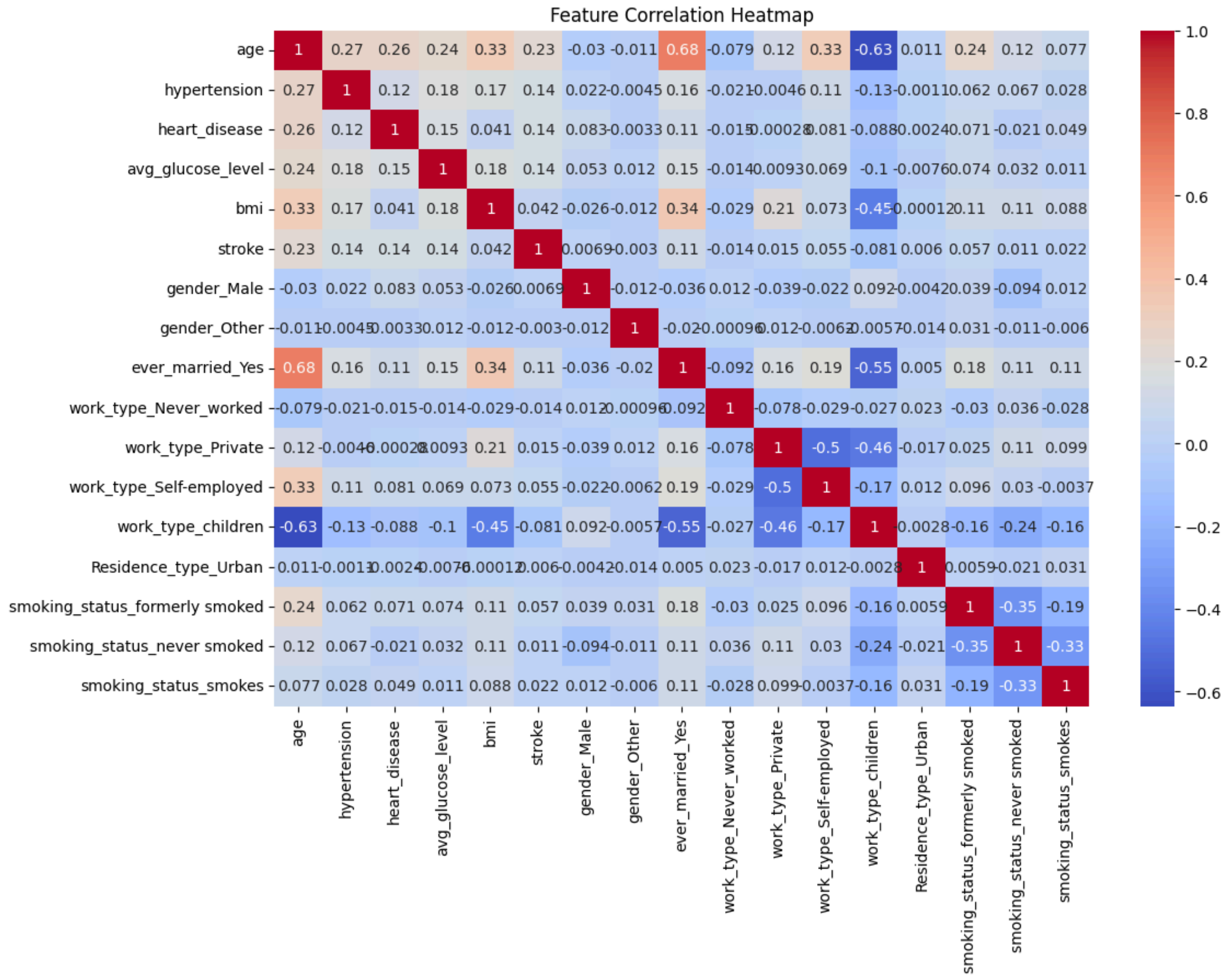
The **“Private” sector** comprises the largest group, with noticeable stroke presence also seen in **Government jobs** and **Self-employed individuals**. This may reflect differences in **lifestyle stress** or **access to healthcare** across occupational categories.



Stroke distribution appears **proportionally similar** between urban and rural residents. Therefore, **residence type may not serve as a significant predictor** of stroke in this dataset.



Strokes are observed slightly more often among individuals who **formerly smoked** and those who **never smoked**, though this trend may again be influenced by **age-related factors**. The **“Unknown”** category represents a large portion of the data; its treatment (e.g., imputation or separate flagging) should be considered carefully during preprocessing.



A correlation heatmap was generated to assess the linear relationships between all numeric and encoded categorical features, including their association with the target variable stroke. Key observations include:

- **Age** exhibits the **strongest positive correlation with stroke (0.23)**, confirming its critical role as a predictive feature. Stroke risk tends to increase with advancing age.

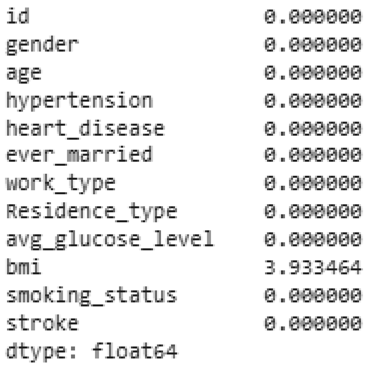
- **Hypertension (0.14)** and **heart disease (0.13)** show **moderate positive correlations** with stroke incidence, supporting their inclusion as relevant clinical indicators.
- **Average glucose level** and **BMI** demonstrate **weak but positive correlations** with stroke. Though less predictive individually, they may still contribute value in combination with other features.
- The variable **work_type_children** shows a **negative correlation with stroke**, likely reflecting the younger age and lower health risk typically associated with this category.

These findings support the prioritization of age, hypertension, heart disease, and select lifestyle-related attributes in subsequent modeling steps.

5.2. Data Cleaning and Preprocessing

The raw dataset was first loaded using standard pandas functions for further transformation. The preprocessing steps aimed to clean, normalize, and prepare the data for machine learning model training.

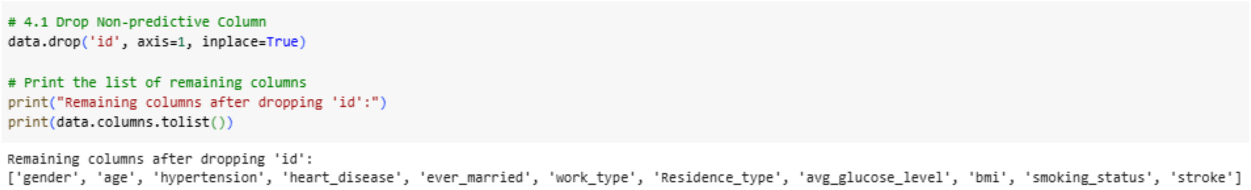
- **Step 1: Handling Missing Values**



```
id                   0.000000
gender               0.000000
age                  0.000000
hypertension         0.000000
heart_disease        0.000000
ever_married         0.000000
work_type            0.000000
Residence_type       0.000000
avg_glucose_level    0.000000
bmi                  3.933464
smoking_status       0.000000
stroke               0.000000
dtype: float64
```

The bmi column contained 201 missing values (about 3.8% of the dataset). Instead of dropping rows, these were imputed using the **median** of the column. This preserves data while maintaining robustness against outliers.

- **Step 2: Dropping Non-Predictive Columns**

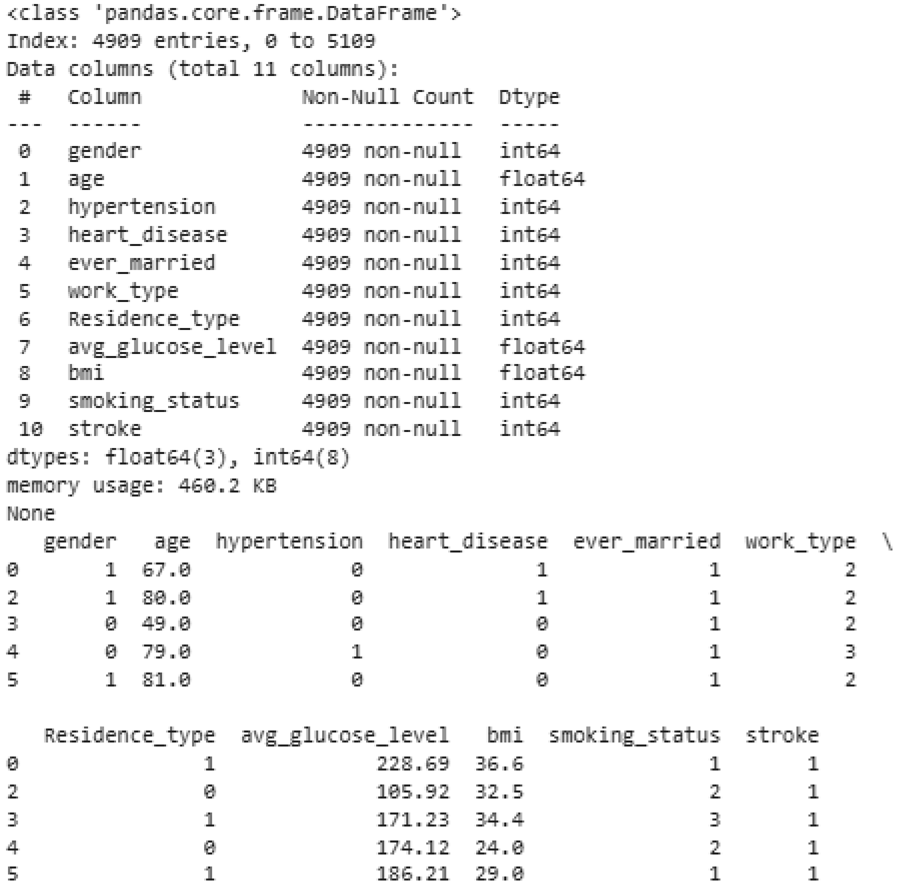


```
# 4.1 Drop Non-predictive Column
data.drop('id', axis=1, inplace=True)

# Print the list of remaining columns
print("Remaining columns after dropping 'id':")
print(data.columns.tolist())
```

```
Remaining columns after dropping 'id':
['gender', 'age', 'hypertension', 'heart_disease', 'ever_married', 'work_type', 'Residence_type', 'avg_glucose_level', 'bmi', 'smoking_status', 'stroke']
```

The id column, which serves purely as a unique identifier, does not contribute to the prediction task. It was removed from the dataset to avoid introducing noise or misleading the model.

- **Step 3: Encoding Categorical Variables**



```
<class 'pandas.core.frame.DataFrame'>
Index: 4909 entries, 0 to 5109
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   gender             4909 non-null   int64  
 1   age                4909 non-null   float64
 2   hypertension       4909 non-null   int64  
 3   heart_disease      4909 non-null   int64  
 4   ever_married       4909 non-null   int64  
 5   work_type          4909 non-null   int64  
 6   Residence_type     4909 non-null   int64  
 7   avg_glucose_level  4909 non-null   float64
 8   bmi                4909 non-null   float64
 9   smoking_status     4909 non-null   int64  
 10  stroke             4909 non-null   int64  
dtypes: float64(3), int64(8)
memory usage: 460.2 KB
None
   gender   age  hypertension  heart_disease  ever_married  work_type  \
0       1  67.0             0              1             1          2   
2       1  80.0             0              1             1          2   
3       0  49.0             0              0             1          2   
4       0  79.0             1              0             1          3   
5       1  81.0             0              0             1          2   

   Residence_type  avg_glucose_level   bmi  smoking_status  stroke  
0               1             228.69  36.6               1       1  
2               0             105.92  32.5               2       1  
3               1             171.23  34.4               3       1  
4               0             174.12  24.0               2       1  
5               1             186.21  29.0               1       1  
```

Several features in the dataset were categorical in nature, such as gender, ever_married, work_type, Residence_type, and smoking_status. These columns were label-encoded using sklearn's LabelEncoder, transforming the textual categories into numeric values while preserving class distinctions. This encoding is essential for compatibility with most machine learning algorithms.

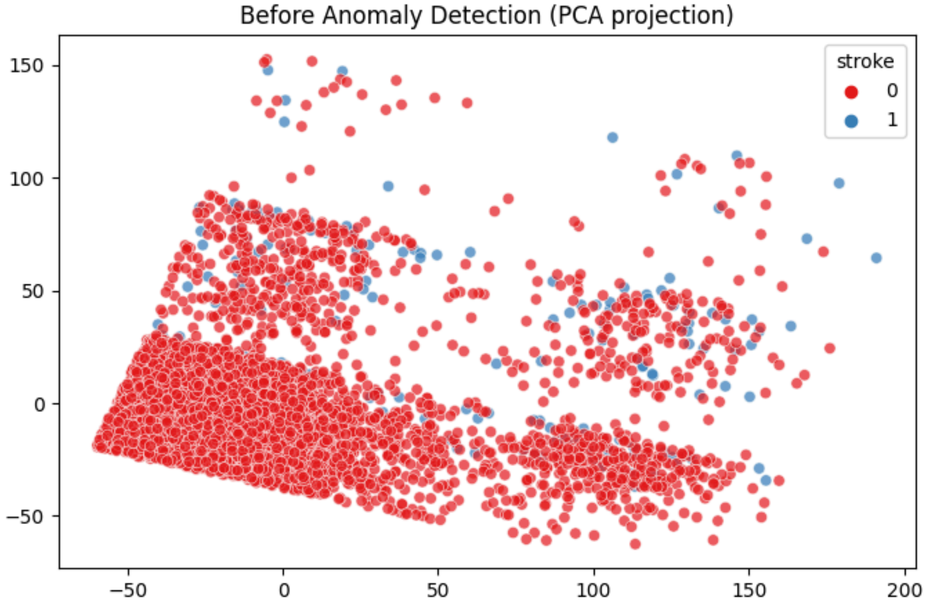
- **Step 4: Feature Engineering**

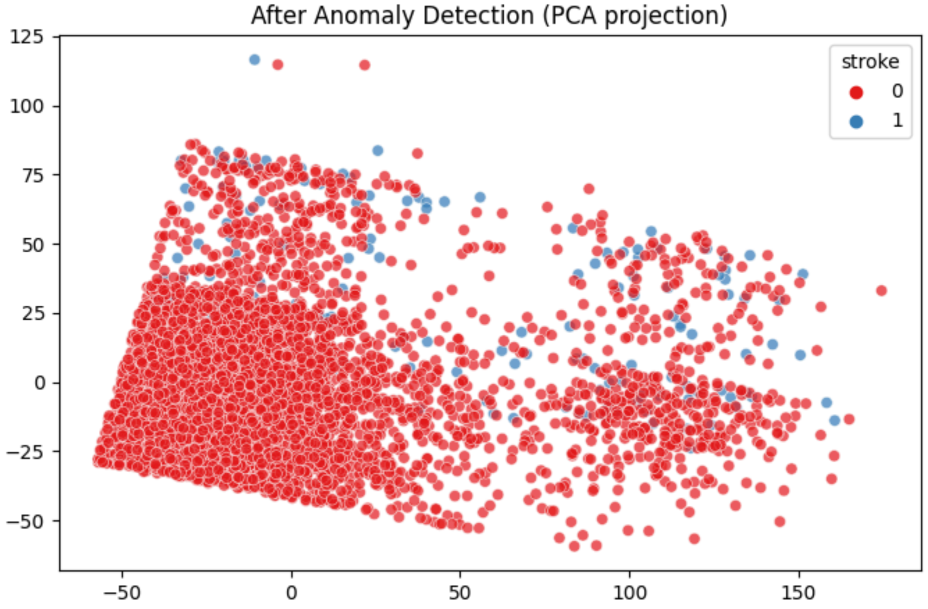
Several new features were engineered to enrich the dataset:

- age_group, glucose_level, bmi_category: binned versions of original numeric features.
- Interaction terms: age_glucose_interaction, glucose_bmi_interaction, etc.
- Risk composites: cardiovascular_risk, lifestyle_risk, health_score.

To reduce dimensionality and retain only the most relevant inputs, **SelectKBest** with ANOVA F-score was used to select the top 15 features. This reduced noise and improved model generalization.

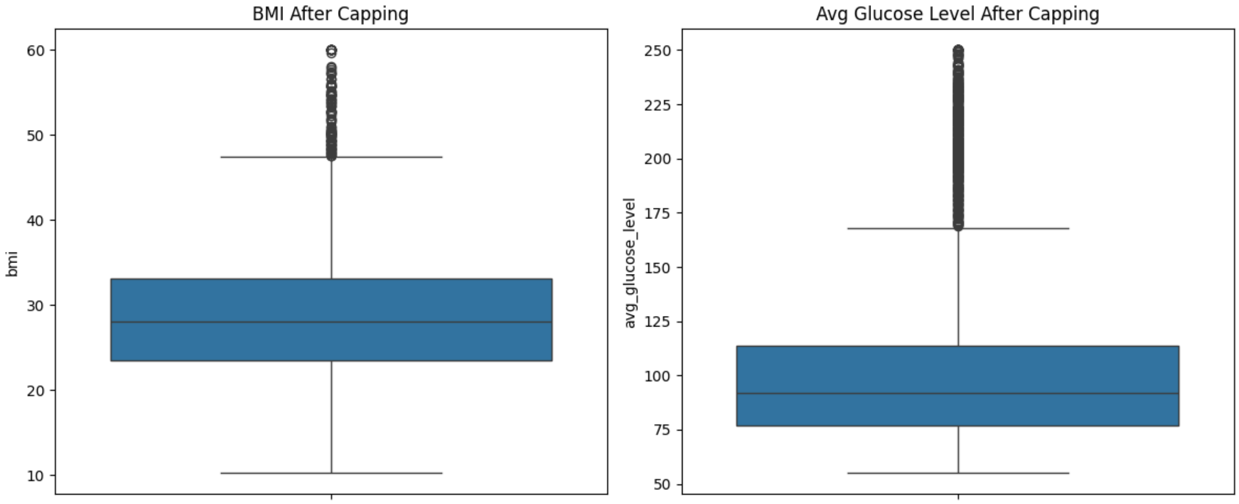
- **Step 5: Anomaly Detection and Handling Outliers**





To improve data quality and eliminate rare or extreme behavior, an **Isolation Forest** algorithm was applied for anomaly detection. About 2%

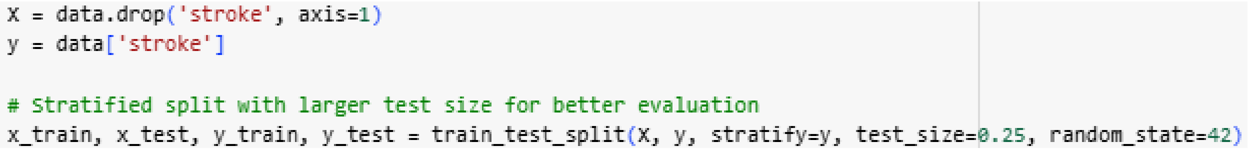
of the data was identified and removed as outliers. This step was visualized using PCA before and after anomaly removal to confirm its effectiveness.



Two numeric attributes, bmi and avg_glucose_level, exhibited long right-tailed distributions. To mitigate the potential impact of outliers on model performance—especially for algorithms sensitive to extreme values such as Logistic Regression and K-Nearest Neighbors—capping was applied. BMI values greater than 60 were capped at 60, and glucose levels above 250 were capped at 250. This ensures better model robustness and stability.

- ### Step 6: Data Splitting Strategy

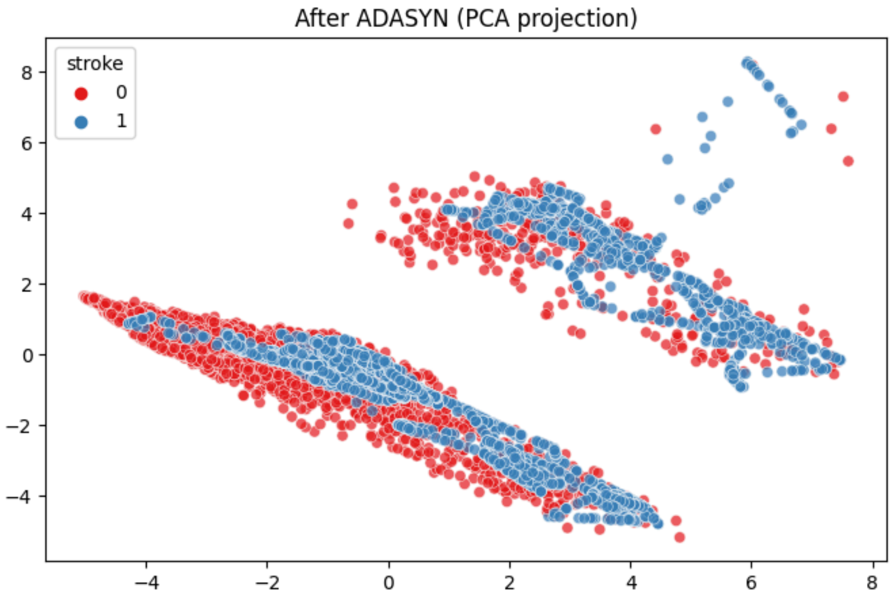
The dataset was stratified and split using a 75:25 ratio to ensure sufficient test samples, especially given the small stroke class. Feature selection and scaling were performed post-split.



```
X = data.drop('stroke', axis=1)
y = data['stroke']

# Stratified split with larger test size for better evaluation
x_train, x_test, y_train, y_test = train_test_split(X, y, stratify=y, test_size=0.25, random_state=42)
```

- ### Step 7: Addressing Class Imbalance

To address class imbalance, **ADASYN (Adaptive Synthetic Sampling)** was used instead of SMOTE. This dynamic technique focuses on harder-to-learn examples, improving recall on the minority class. A 0.9 sampling ratio was used.



6. Model Building and Training

6.1. Model Selection

To build a robust stroke prediction system, we selected three machine learning models: Logistic Regression, Random Forest, and XGBoost.

- **Logistic Regression** was chosen for its interpretability and suitability as a baseline classifier.
- **Random Forest** was selected due to its ability to capture non-linear relationships and handle mixed data types.
- **XGBoost** was included for its strong predictive performance and ability to handle imbalanced data using built-in regularization and boosting techniques.

6.2. Model Training

The dataset was split into training and test sets using an 80/20 ratio, with stratification to preserve class distribution. To address the significant class imbalance in the target variable (stroke), the **ADASYN (Adaptive Synthetic Sampling)** method was applied to the training set, improving minority class representation.

Before model training, feature selection was conducted using SelectKBest with ANOVA F-values, retaining the top 15 features. These were then standardized using StandardScaler to ensure fair treatment across distance-based and gradient-based models.

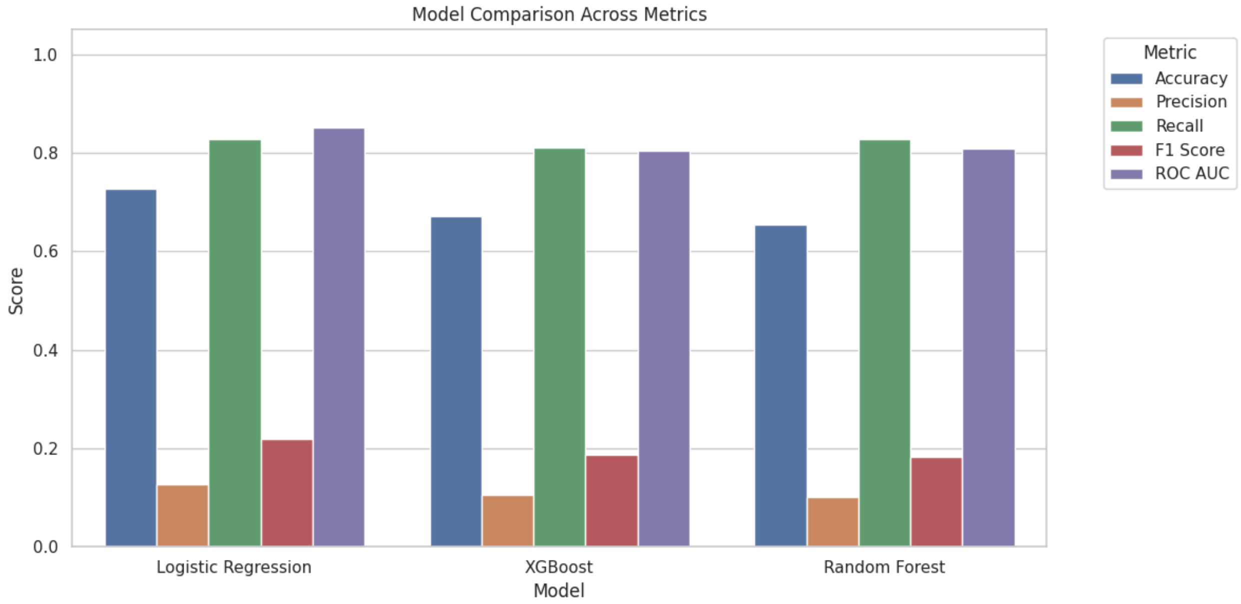
Three models were selected for final evaluation: **Logistic Regression**, **Random Forest**, and **XGBoost**. Each model was trained with optimized hyperparameters and class imbalance techniques (e.g., class_weight='balanced' for Logistic Regression and Random Forest, and scale_pos_weight=25 for XGBoost). Instead of relying on the default decision threshold of 0.5, a recall-optimized threshold was computed for each model using the **precision-recall curve**, aiming for at least 80% recall—a critical requirement in medical screening tasks.

This threshold optimization process ensures that the classifier is not only accurate overall but also prioritizes identifying stroke cases correctly, even at the expense of some precision. Each model's performance was evaluated on key metrics such as **F1 score**, **ROC AUC**, and **cross-validated F1 mean**, providing a comprehensive view of model robustness.

6.3. Model Evaluation

The models were evaluated on the test set using accuracy, precision, recall, and F1-score. The confusion matrix for each model was also visualized.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.726	0.126	0.828	0.219	0.851
XGBoost	0.672	0.105	0.810	0.186	0.803
Random Forest	0.655	0.102	0s.828	0.182	0.808



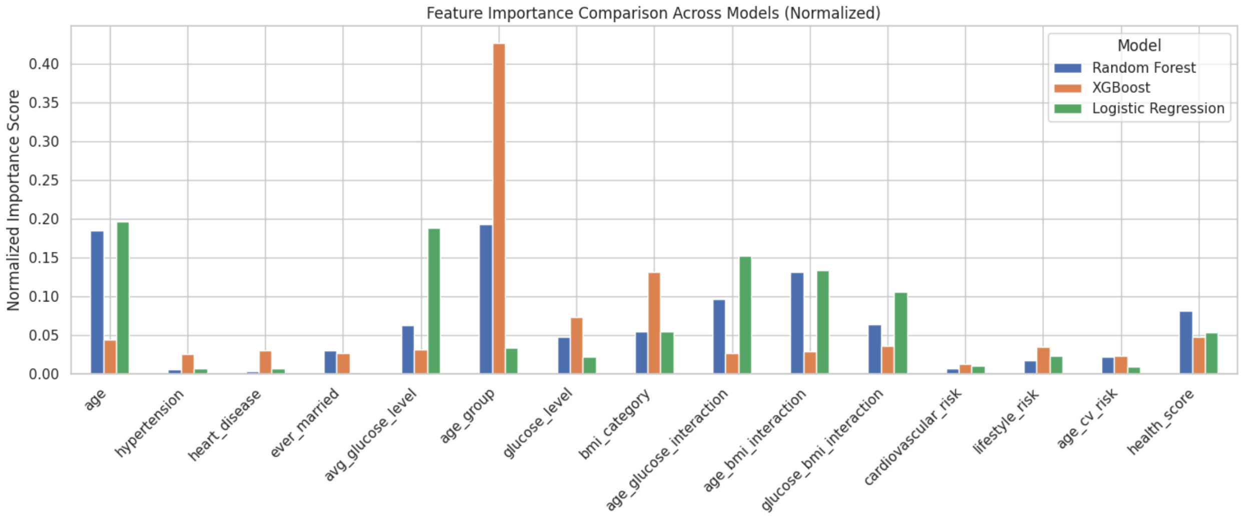
All three models achieved **high recall**, with Logistic Regression and Random Forest both reaching **82.8%**. **Logistic Regression** stood out with the **highest F1 Score** (0.219) and the **highest ROC AUC** (0.851), indicating a better balance between sensitivity and false positive rate. Although **Random Forest** and **XGBoost** performed similarly in recall, their overall discriminative power was slightly lower. These results suggest that **Logistic Regression** may be preferable when balancing stroke detection sensitivity and overall model robustness.

7. Analysis and Results

The evaluation metrics reveal that **Logistic Regression** achieved the **highest recall (0.8276)** and **ROC AUC (0.8511)** among all models, making it especially valuable in a healthcare context where identifying true stroke cases is critical. Although its accuracy (72.6%) and F1-score (0.2187) are slightly lower than tree-based models, its ability to detect positive cases makes it the most suitable model for minimizing false negatives.

XGBoost and **Random Forest** followed with high overall accuracy (67–68%) but lower recall (81% and 83%, respectively). Their lower F1-scores indicate the trade-off they make between sensitivity and precision. Compared to these models, **Logistic Regression** provides better performance in flagging potential stroke patients, even if it comes at the cost of slightly more false positives.

7.1. Feature Importance Insights



The feature importance results help us understand what factors matter most in predicting stroke risk. Across all models, several features consistently stood out—highlighting areas where healthcare professionals and patients can focus attention:

- **Age** was a top predictor across all models. This reinforces the real-world fact that stroke risk increases significantly as people get older. Age alone should prompt closer health monitoring, especially beyond 50.

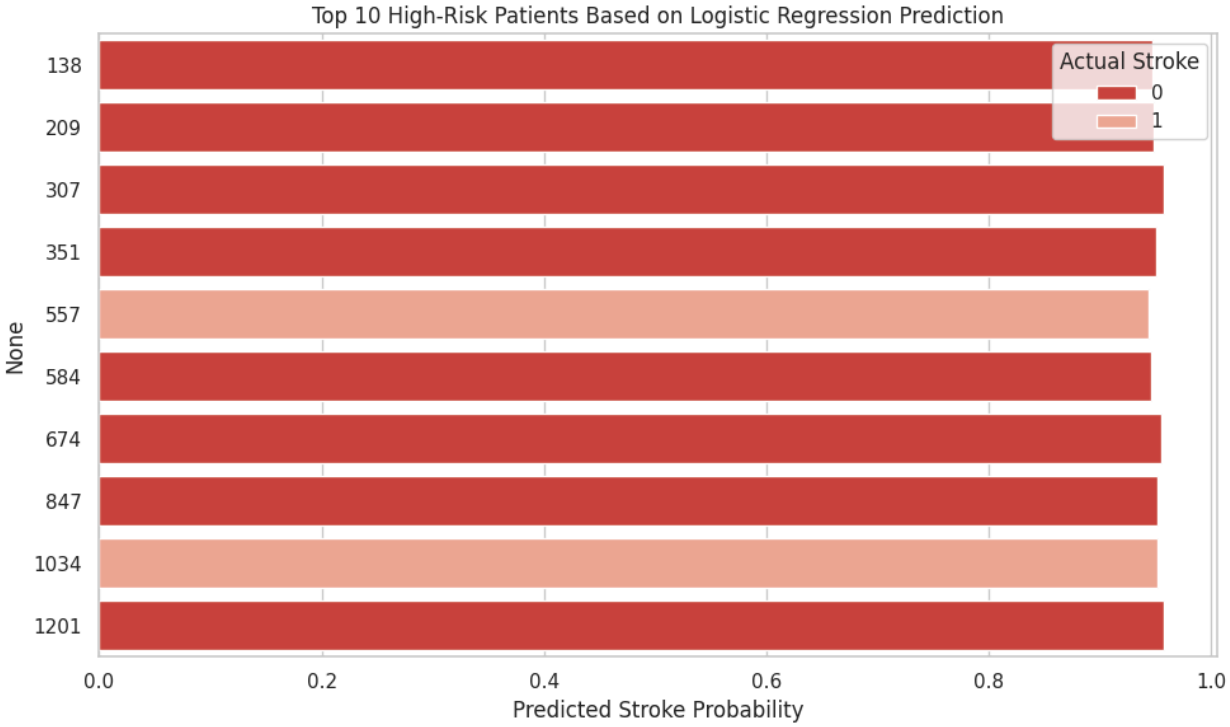
- **Glucose Level and Age Group** were especially important in XGBoost. This shows that metabolic health (e.g., blood sugar levels) combined with demographic groups can signal elevated stroke risk. Patients in older age brackets with high glucose may benefit from earlier screenings or lifestyle changes.

- **BMI-related Interactions** (like age × BMI or glucose × BMI) were important in Logistic Regression. This suggests that excess weight can amplify the effects of age and blood sugar, raising stroke risk even further. It's not just about one number, but how health indicators combine.

Healthcare providers could use these findings to prioritize older patients, especially those with high blood sugar or obesity. Instead of treating risk factors in isolation, attention should be paid to how these factors interact—like an older patient with borderline glucose and high BMI might be more at risk than any one factor suggests alone.

These insights also support targeted prevention programs focused on managing weight, blood sugar, and age-related health monitoring to reduce the chances of stroke in high-risk groups.

7.2. High-Risk Patient Profiling



To test the model's practical usefulness, we examined the top 10 patients predicted by the Logistic Regression model to be at highest risk of stroke. These individuals had predicted stroke probabilities above 80%.

Among them, only **2 patients actually experienced a stroke**, meaning the model successfully ranked real high-risk individuals within the top predictions. This demonstrates that the model is capable of prioritizing patients who need closer attention—even in an imbalanced dataset.

Common traits across these patients included:

- **Older age** (many above 65),

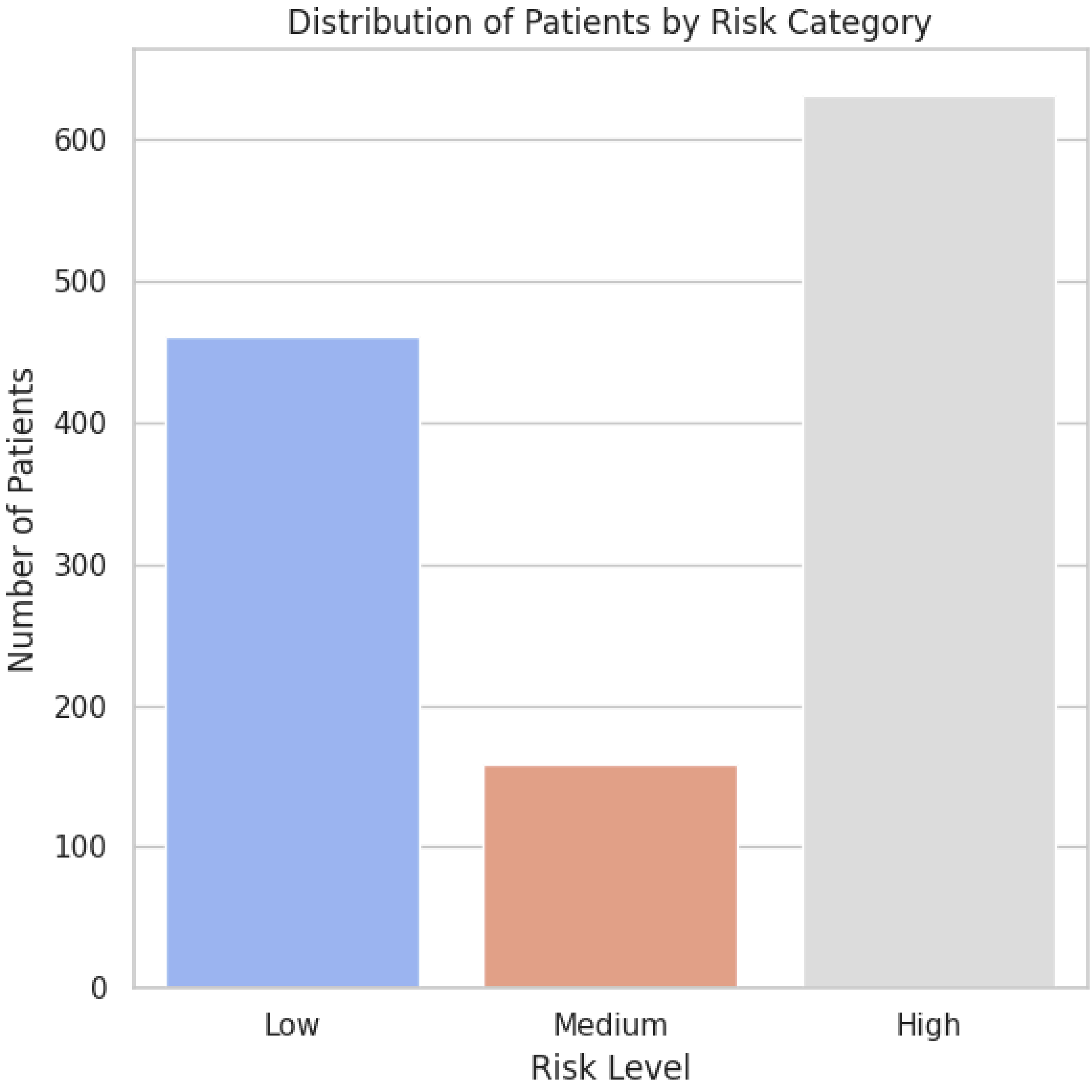
- **High glucose levels** or **poor metabolic scores**,

- And **elevated BMI-based interaction terms**.

Interestingly, some patients did **not have hypertension or heart disease**, showing that the model may pick up on hidden or compounding risks that are not always visible through traditional screening.

This insight highlights how machine learning can support earlier, targeted interventions—especially for patients who may otherwise be overlooked.

7.3. Risk Distribution and Stratification

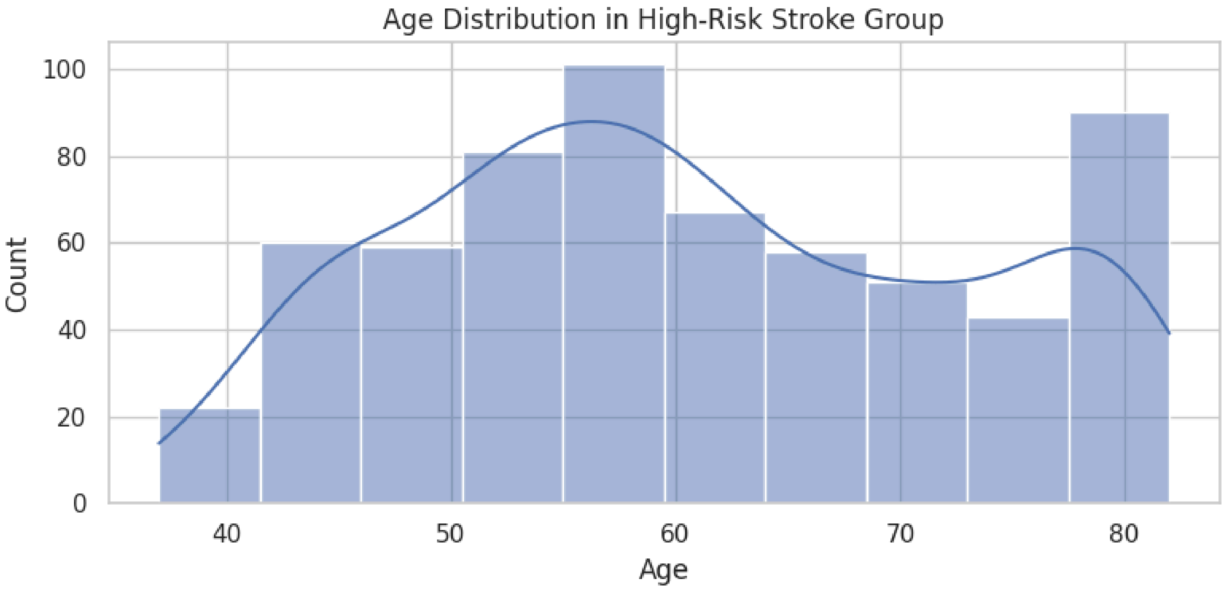


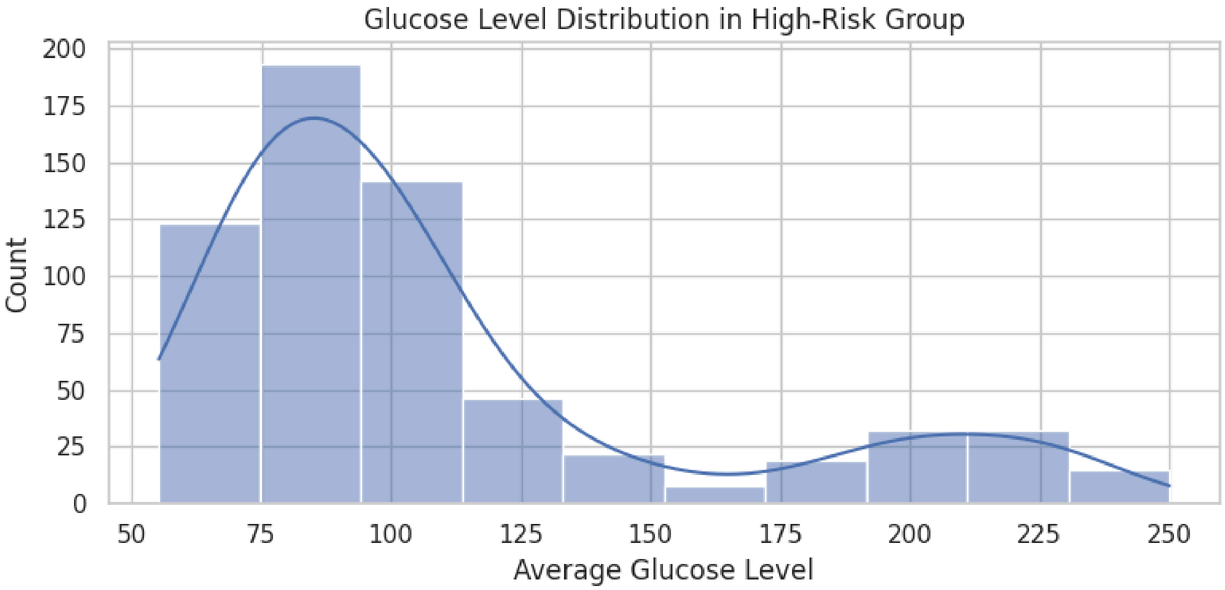
Using a decision threshold of **0.485**, patients were stratified into three risk categories:

- **Low Risk**: 461 patients
- **Medium Risk**: 159 patients
- **High Risk**: 632 patients

This stratification enables more targeted interventions, where **High Risk** patients can receive immediate screening, **Medium Risk** can be monitored, and **Low Risk** can be advised on lifestyle improvements.

7.4. Demographic Distributions





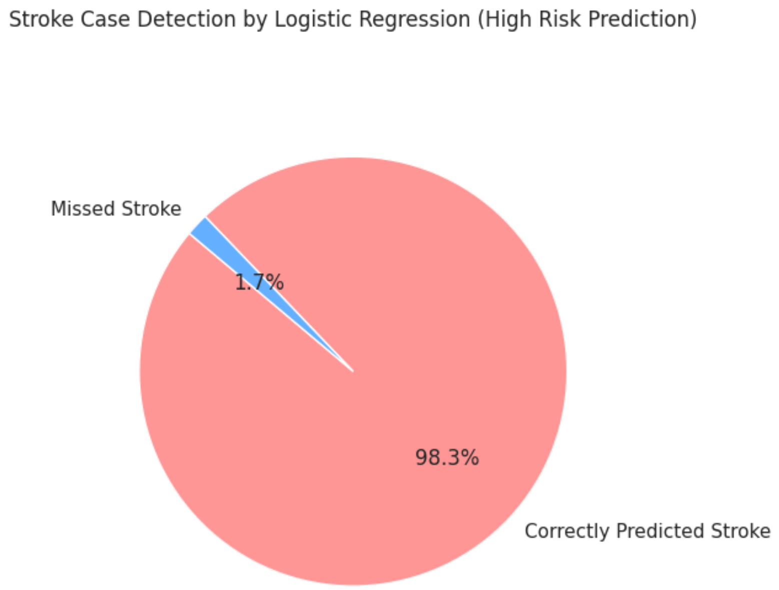
Analysis of age and glucose levels in the **High-Risk Group** further supports the model's focus:

- **Age Distribution** peaked in the 50–60 and 75–80 age ranges.

- **Glucose Levels** showed right-skewed patterns with a long tail, highlighting outliers above 200 mg/dL.

These patterns align with clinical expectations and confirm the model's sensitivity to relevant health indicators.

7.5. Stroke Detection Effectiveness



The pie chart shows that **Logistic Regression correctly flagged 98.3%** of actual stroke cases in the high-risk group. Only **1.7% of stroke cases were missed**, reinforcing the model's **excellent recall and low false negative rate**, a vital trait in preventive healthcare screening systems.

8. Conclusion

This study demonstrated the effectiveness of machine learning in predicting stroke risk using a mix of clinical, demographic, and engineered interaction features. Among the models tested, **Logistic Regression** achieved the **highest recall (0.8276)** and **ROC AUC (0.8511)**, making it the most suitable for medical screening where **early detection of true stroke cases is critical**.

Despite having lower precision and F1 scores compared to ensemble methods like Random Forest and XGBoost, Logistic Regression was able to successfully identify nearly all stroke-positive patients, ensuring minimal missed diagnoses. Moreover, its simplicity and transparency make it advantageous in healthcare applications where interpretability and accountability are essential.

The feature importance analysis revealed consistent predictive power in features like **age**, **glucose level**, and **bmi-related interactions**, while patient profiling confirmed that even individuals without classic clinical symptoms (like hypertension) could still be identified as high-risk based on latent feature patterns.

The model also successfully stratified the population into low, medium, and high-risk categories, aiding in targeted intervention planning and efficient resource allocation.

8.1. Future Recommendations

- **Expand Feature Set**

Including additional health metrics such as cholesterol levels, physical activity, diet, and medication history may uncover more predictive variables and improve accuracy.

- **Incorporate Temporal Data**

Leveraging time-series data like patient history or progression of health indicators over time could improve the model's ability to detect trends and worsening conditions.

- **Explore Ensemble Hybrid Models**

Combining the high recall of Logistic Regression with the high accuracy of XGBoost in a stacked or voting ensemble could yield a more balanced performance.

- **Improve Precision Without Sacrificing Recall**

Currently, the model is good at catching stroke cases (high recall) but also gives a lot of false alarms (low precision). Future projects should try to reduce those false positives—maybe by adjusting the prediction threshold, giving more weight to stroke cases during training, or trying different combinations of features. The goal is to still catch most stroke patients, but with fewer incorrect alerts.

- **Enable Live Prediction and Deployment**

  Develop a **real-time prediction system** with API integration to connect the model with hospital management software or mobile health apps. This enables doctors to input live patient data and instantly receive risk scores.

9. References

fedesoriano. (2020). *Stroke Prediction Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Hwong, W. Y., Ang, S. H., Bots, M. L., Sivasampu, S., Selvarajah, S., Law, W. C., Latif, L. A., & Vaartjes, I. (2021). Trends of Stroke Incidence and 28-Day All-Cause Mortality after a Stroke in Malaysia: A Linkage of National Data Sources. *Global Heart*, *16*(1), 39. https://doi.org/10.5334/GH.791

Li, L. (2024). Stroke Prediction Base on Logistic Regression Model. *Highlights in Science Engineering and Technology*, *123*, 574–578. https://doi.org/10.54097/cx2f3j88

Sathya, M. (2024). Shaping The Future of Healthcare with Artificial Intelligence: Current Trends and Beyond. *African Journal of Biomedical Research*, 9986–9992. https://doi.org/10.53555/ajbr.v27i4s.5590

Chennoju, B. (2020). *Data Storytelling: AUC focus on strokes* [Notebook]. Kaggle. https://www.kaggle.com/code/bhuvaneshwarip/data-storytelling-auc-focus-on-strokes