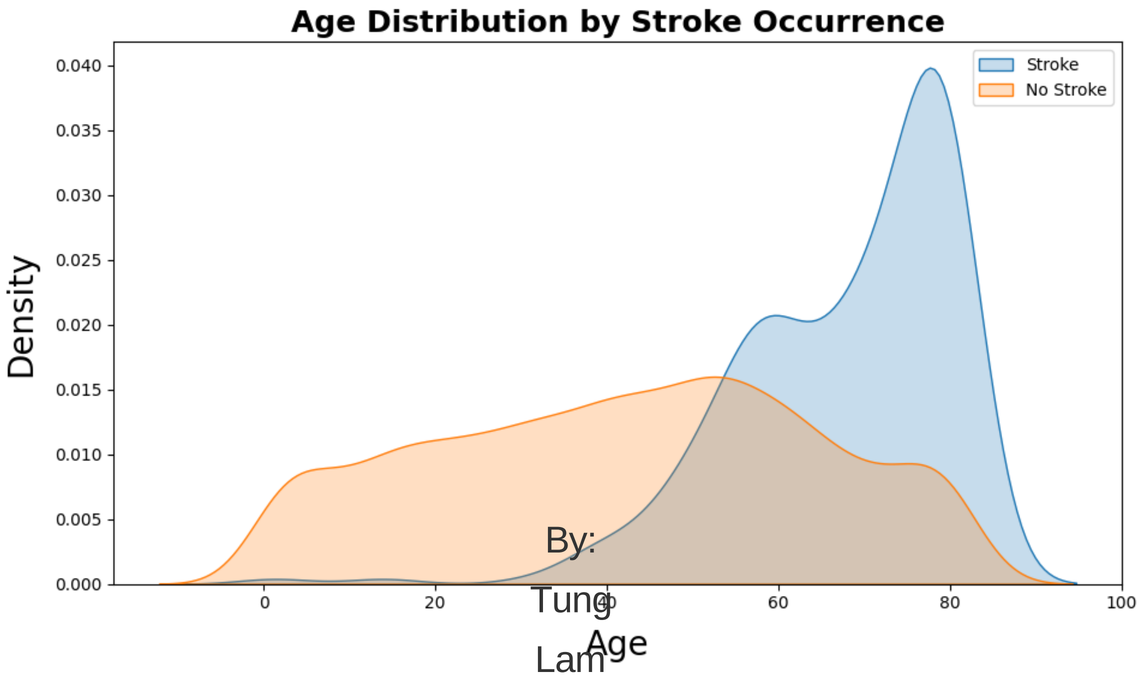
# Stroke Prediction Using Machine Learning

## A Comparative Analysis of Logistic Regression and XGBoost



By:
Tung Lam & Hoan

May 2023

# Table of Contents

# Stroke Prediction Using Logistic Regression

> According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.
>
> This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

## Dataset Overview and Preprocessing

The dataset used for this project contains detailed information for each patient, where every record represents a single patient and includes the following key features:

- This dataset contains: 5110 rows and 12 columns
- **id**: unique identifier
- **gender**: "Male", "Female" or "Other"
- **age**: age of the patient
- **hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart_disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- **ever_married**: "No" or "Yes"
- **work_type**: "children", "Govt_job", "Never_worked", "Private" or "Self-employed"
- **Residence_type**: "Rural" or "Urban"
- **avg_glucose_level**: average glucose level in blood
- **bmi**: body mass index
- **smoking_status**: "formerly smoked", "never smoked", "smokes" or "Unknown"
- **stroke**: 1 if the patient had a stroke or 0 if not

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient*

## Sample Data

| ID | Gender | Age | Hypertension | Heart Disease | Ever Married | Work Type | Residence Type | Avg Glucose Level |
|---|---|---|---|---|---|---|---|---|
| 9046 | Male | 67 | 0 | 1 | Yes | Private | Urban | 228.69 |
| 51676 | Female | 61 | 0 | 0 | Yes | Self-employed | Rural | 202.21 |
| 31112 | Male | 80 | 0 | 1 | Yes | Private | Rural | 105.92 |
| 60182 | Female | 49 | 0 | 0 | Yes | Private | Urban | 171.23 |
| 1665 | Female | 79 | 1 | 0 | Yes | Self-employed | Rural | 174.12 |

## Data Preprocessing

To create any predictive model, the process of data processing involves:

1. **Data Cleaning**: Checking for and handling missing values. For example, the BMI column may contain missing values that can be replaced with the average or middle value.
2. **Feature Encoding**: Most of the features in the dataset are categories, such as gender, work type, residence type, and smoking status. These could be converted into numbers by techniques such as One hot encoding or label encoding.
3. **Feature Scaling**: The values of age, average blood glucose level, and BMI can be scaled to be equally important during training of the model. Methods to scale these values usually standardization (z-score normalization) and min-max scaling.
4. **Data Splitting**: The dataset is divided into 2 sets: training and testing sets (usually 80/20 split). This helps observing how well the model performs on new data.

This dataset considered a supervised learning problem since it contains a labeled column stroke (following the previous lesson of distinguishing supervised learning vs unsupervised learning).

# Descriptive Statistics

### Age

- Count: 5110
- Mean (Average): ~43.23 years
- Median: 45 years
- Standard Deviation: ~22.61 years
- Range: 0.08 (min) to 82 (max)
- 25th percentile: 25 years
- 75th percentile: 61 years

### Average Glucose Level

- Count: 5110
- Mean (Average): ~106.15 mg/dL
- Median: ~91.89 mg/dL
- Standard Deviation: ~45.28 mg/dL
- Range: ~55.12 mg/dL (min) to ~271.74 mg/dL (max)
- 25th percentile: ~77.25 mg/dL
- 75th percentile: ~114.09 mg/dL

### BMI (Body Mass Index)

- Count: 4909 (missing 201 entries)
- Mean (Average): ~28.89
- Median: ~28.10
- Standard Deviation: ~7.85
- Range: 10.30 (min) to 97.60 (max)
- 25th percentile: 23.50
- 75th percentile: 33.10

### Hypertension & Heart Disease

**Hypertension:**

- 0 (No): 4658 records
- 1 (Yes): 452 records

**Heart Disease:**

- 0 (No): 4897 records
- 1 (Yes): 213 records

### Categorical Variables

**Gender:** Female (2994) is the most common, followed by Male

**Marital Status:** "Yes" is predominant (3353 cases)

**Work Type:** "Private" is most frequent (2925 records)

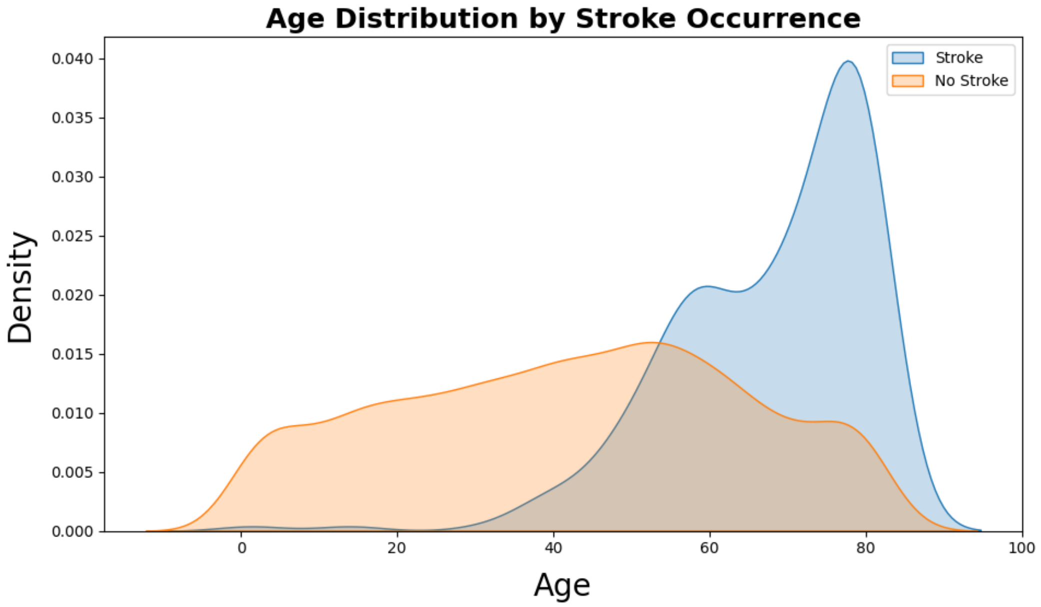
**Residence Type:** "Urban" is the top category (2596 cases)

**Smoking Status:** "never smoked" is most common (1892 cases)

## Comprehensive Analysis of Stroke Risk Factors

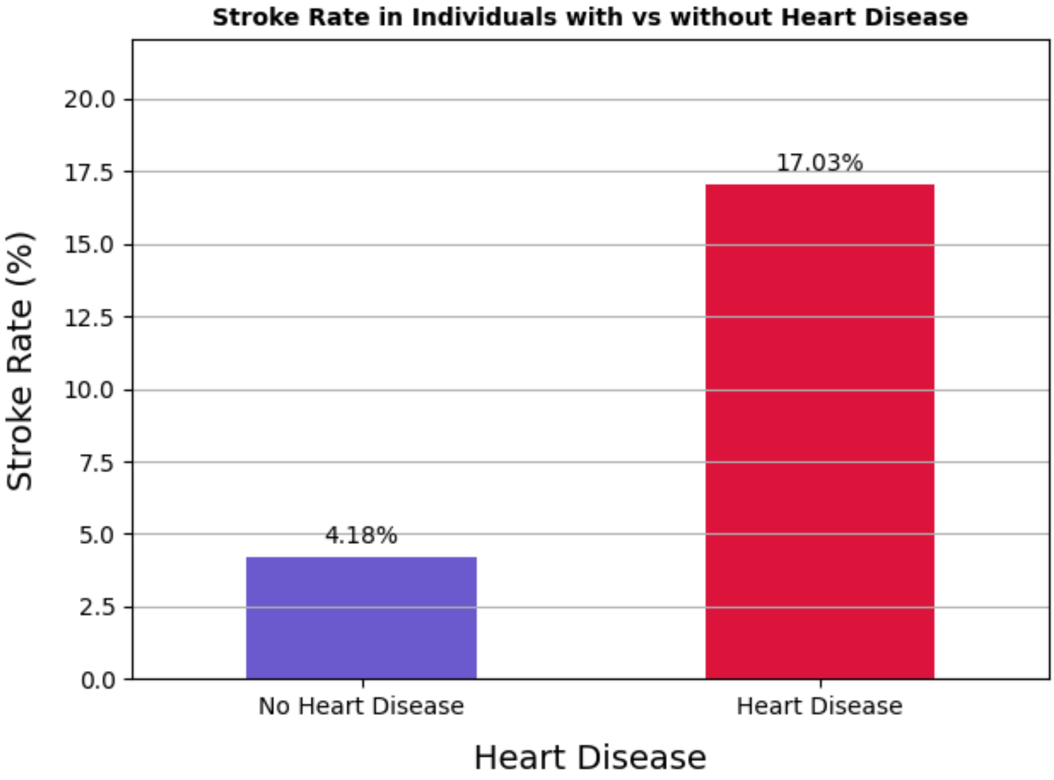
Our analysis reveals important relationships between various demographic and health factors and stroke risk. The following sections present key findings from our statistical analysis.

# Age and Stroke Risk



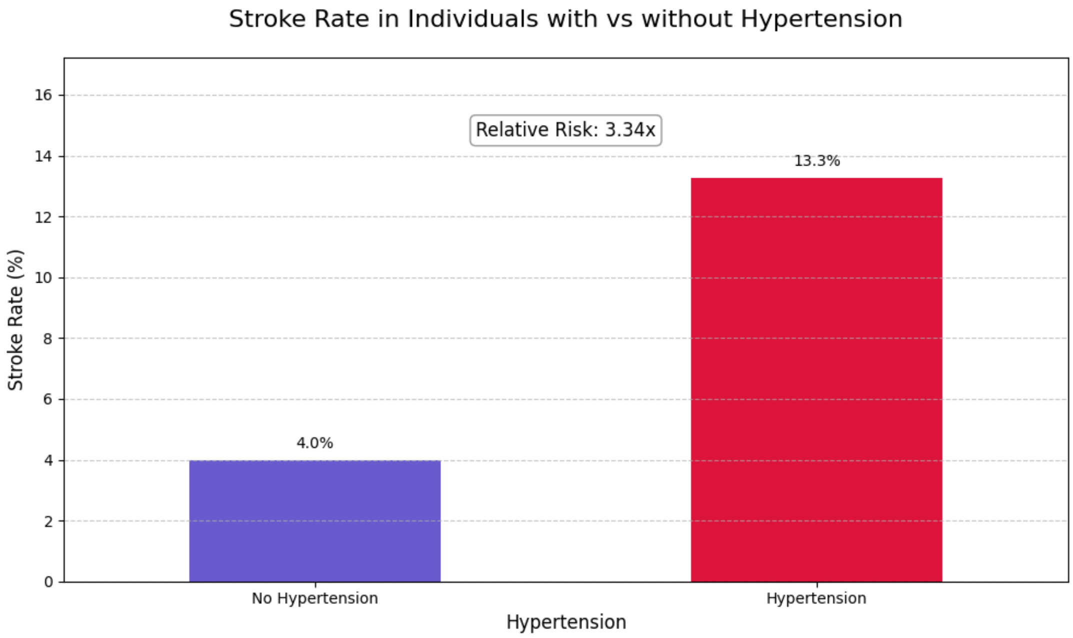
Age is the strongest predictor of stroke risk in our dataset. The average age of all patients is 43.2 years (range: 0.08-82 years), but stroke patients are significantly older on average. The age distribution shows a clear pattern: stroke risk increases dramatically with age, with the highest concentration of stroke cases occurring in patients over 60 years old.

# Heart Disease and Stroke Risk



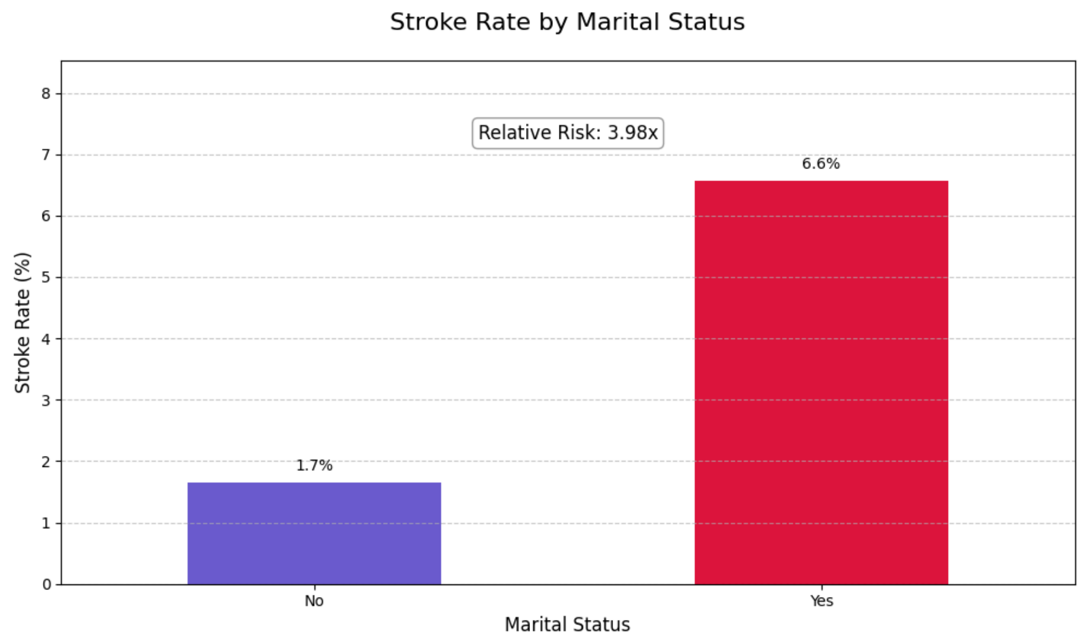
> **Heart Disease:** The stroke rate in patients with heart disease is 17.03% compared to 4.18% in those without heart disease. This makes heart disease one of the strongest risk factors in our dataset, with patients with heart disease having approximately 4 times higher risk of stroke. This highlights the critical importance of cardiovascular health in stroke prevention.

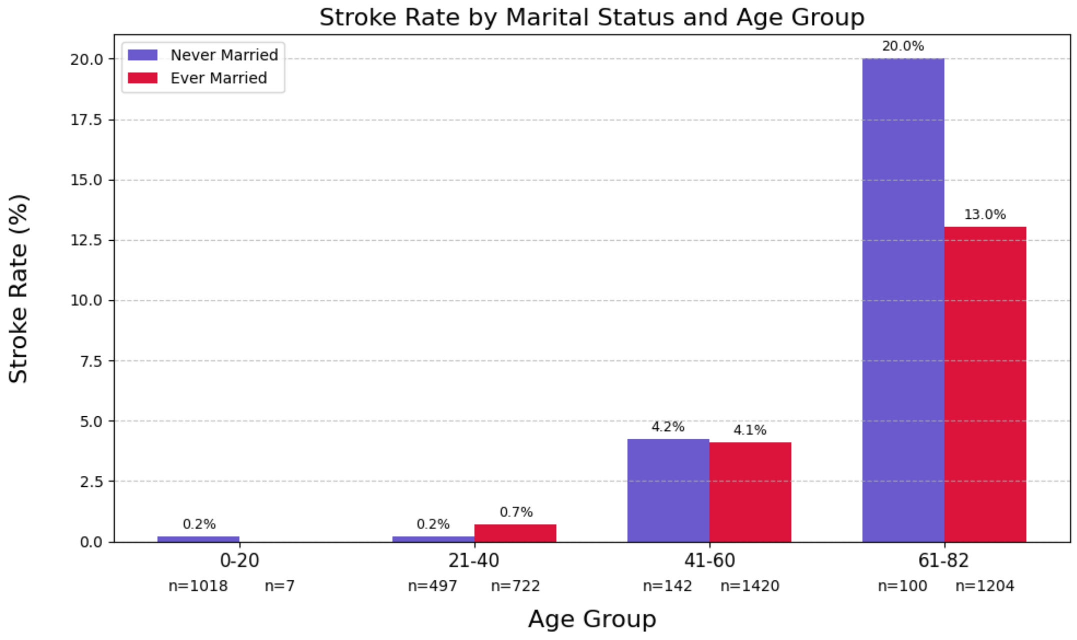
# Hypertension and Stroke Risk



**Hypertension:** Patients with hypertension have a 13.25% stroke rate compared to 3.97% in those without hypertension, representing a 3.34× higher risk. This significant difference underscores the importance of blood pressure control in stroke prevention strategies.

# Marital Status and Stroke Risk



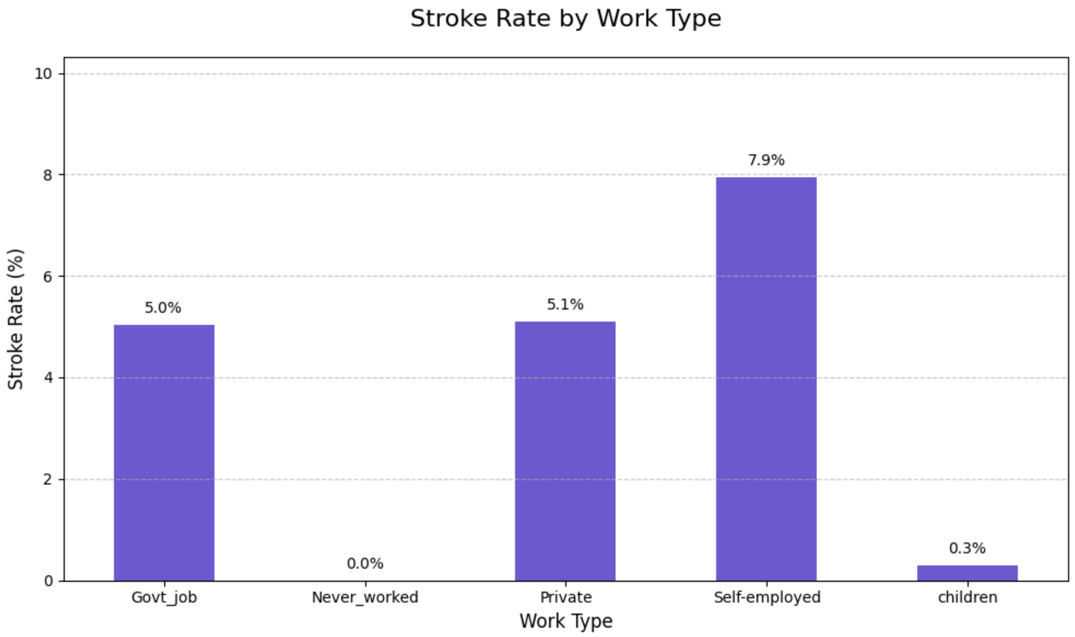


At first glance, married individuals appear to have a higher stroke rate (6.56%) compared to never married individuals (1.65%). However, this is a classic example of Simpson's Paradox, as the difference is primarily explained by age differences:

- Never married individuals are much younger (average age: 22.0 years)
- Ever married individuals are significantly older (average age: 54.3 years)

When controlling for age by examining each age group separately, the relationship changes or even reverses in older age groups (61-82), where never married individuals have a 20% stroke rate compared to 13% for married individuals. Marriage appears to be protective against stroke, especially in older age groups where the effect is strongest (7% difference in the 61-82 group).

## Work Type and Stroke Risk

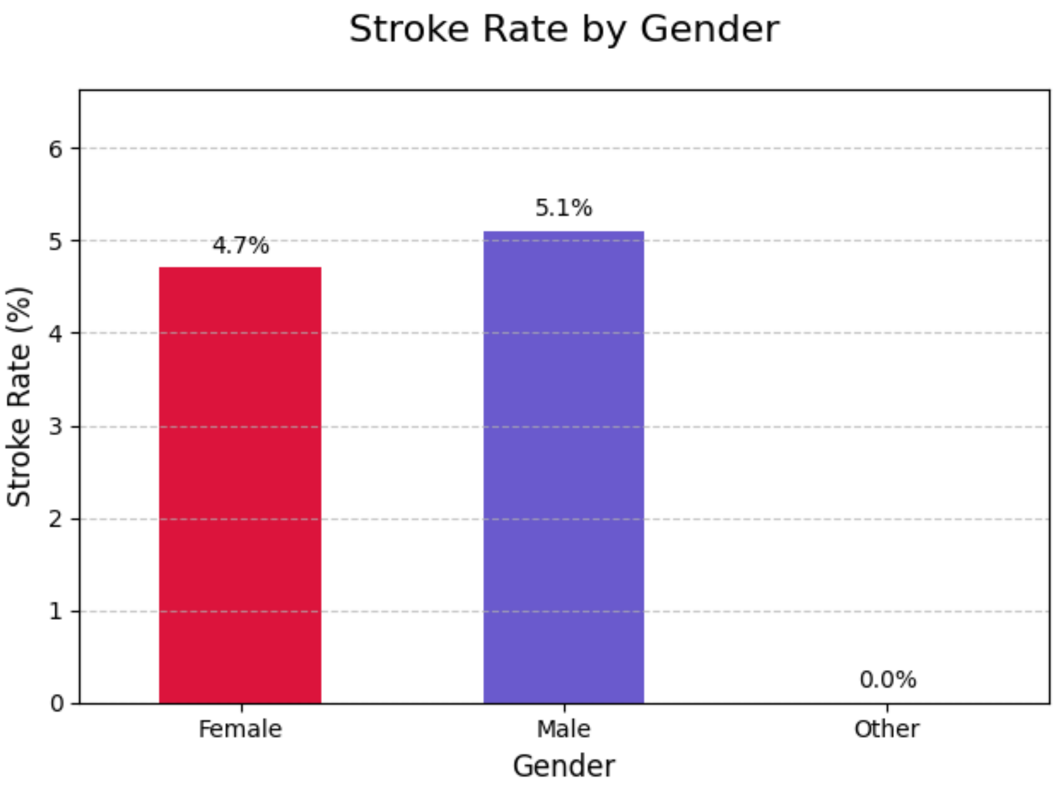


Self-employed individuals have the highest stroke rate (7.94%), followed by government employees (5.02%) and private workers (5.09%). Children and those who never worked show very low stroke rates. However, these differences are largely explained by age differences:

- Self-employed: Average age 60.2 years
- Government job: Average age 50.9 years
- Private: Average age 45.5 years
- Children: Average age 6.8 years

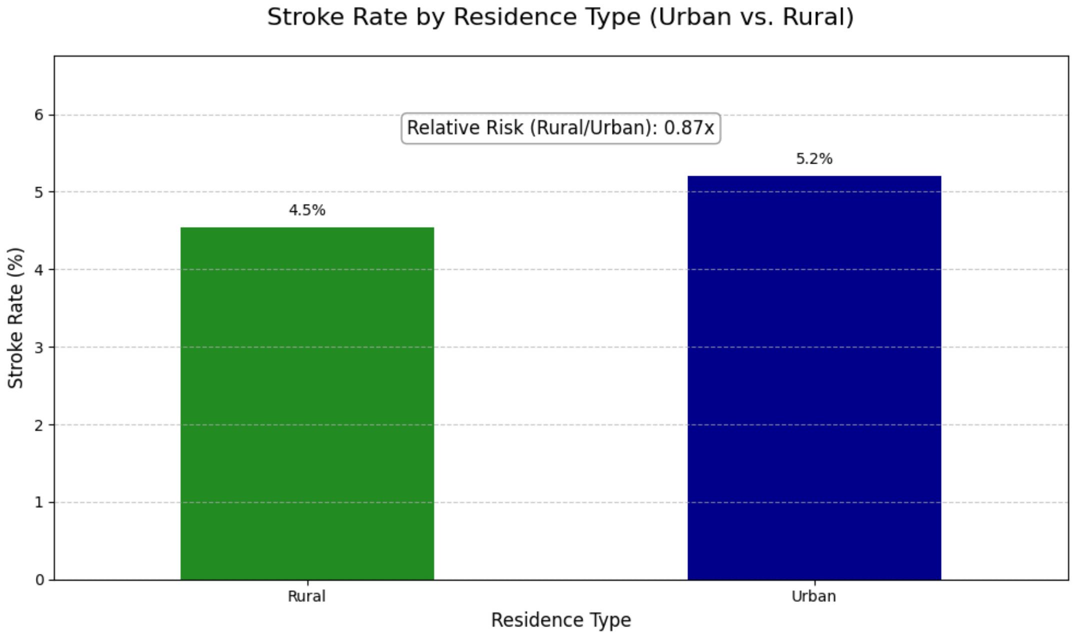
> The higher stroke rates in certain occupations primarily reflect the older age of individuals in those categories rather than occupation-specific risk factors.

## Gender and Stroke Risk



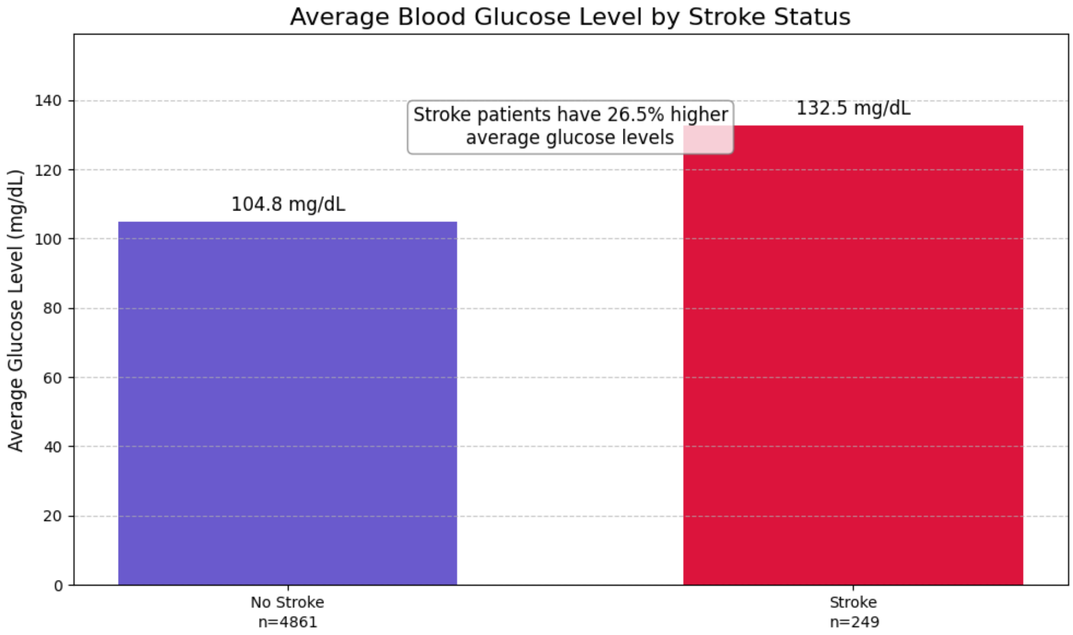
> The analysis shows relatively similar stroke rates between males (5.11%) and females (4.71%), suggesting that gender alone is not a strong predictor of stroke risk in our dataset. No strokes were recorded in the "Other" gender category, though this may be due to small sample size.

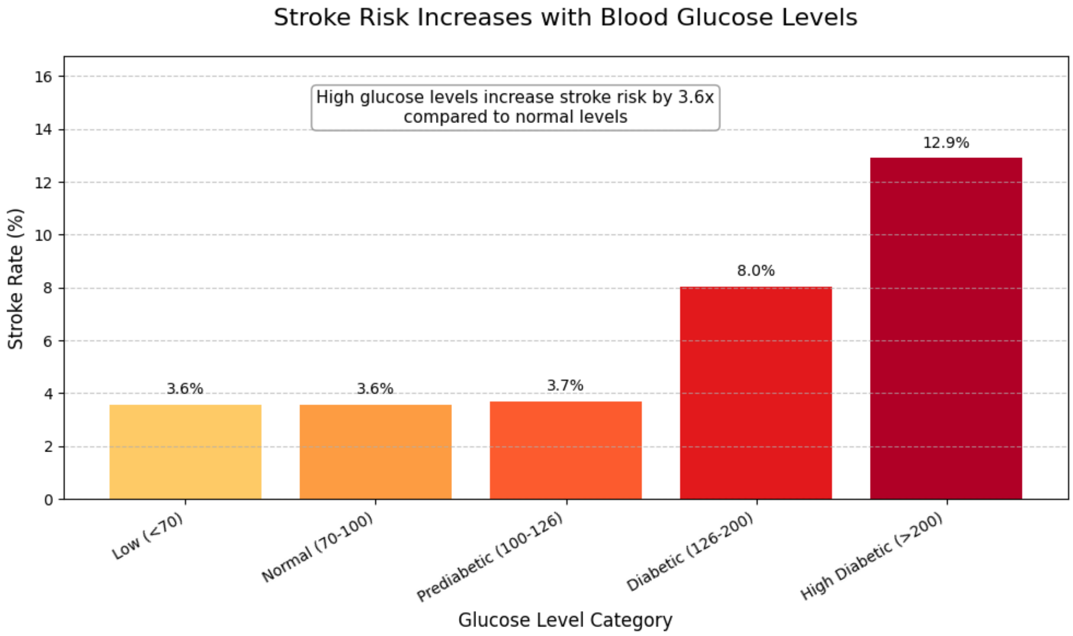
# Residence Type and Stroke Risk



Urban residents show a slightly higher stroke rate (5.20%) compared to rural residents (4.53%), with a relative risk ratio of 0.87×. This small difference suggests that residence type has a minimal impact on stroke risk compared to other factors like age, hypertension, and heart disease.

# Blood Glucose Levels and Stroke Risk





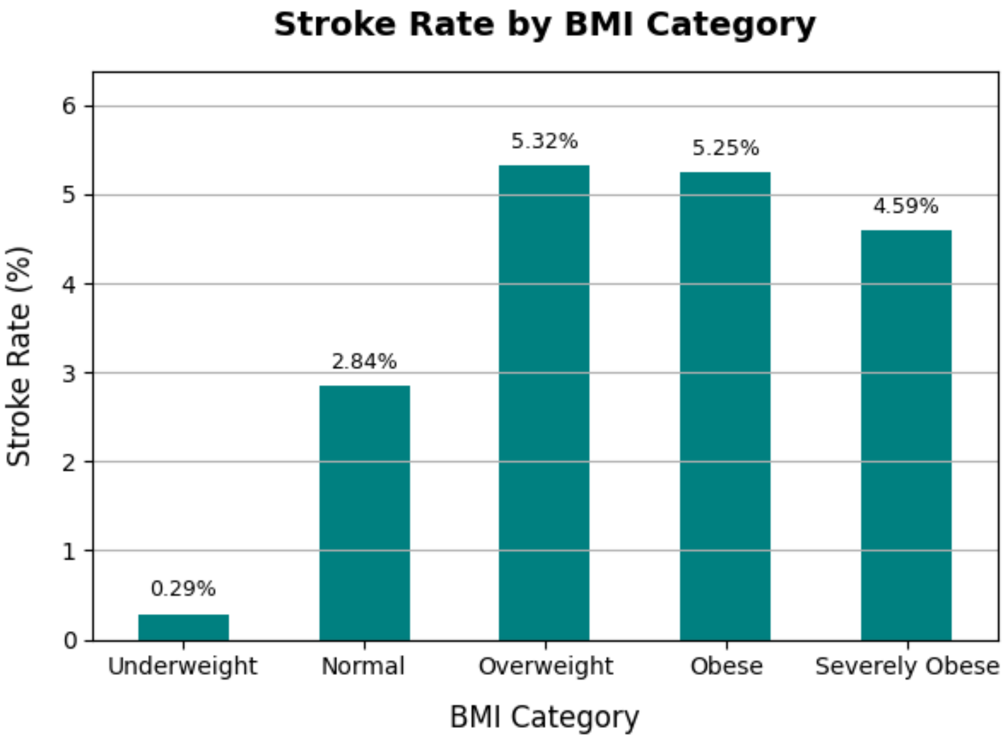
Blood glucose levels show a strong association with stroke risk:

- The average glucose level for stroke patients is 132.5 mg/dL, which is 26.5% higher than non-stroke patients (104.8 mg/dL)
- Stroke rates increase dramatically with glucose levels:
    - Low (<70 mg/dL): 3.57%
    - Normal (70-100 mg/dL): 3.58%
    - Prediabetic (100-126 mg/dL): 3.71%
    - Diabetic (126-200 mg/dL): 8.04%

- ◦ High Diabetic (>200 mg/dL): 12.90%

Patients with high diabetic glucose levels have a 3.61× higher stroke risk compared to those with normal levels, highlighting the importance of glucose management in stroke prevention.

## BMI and Stroke Risk

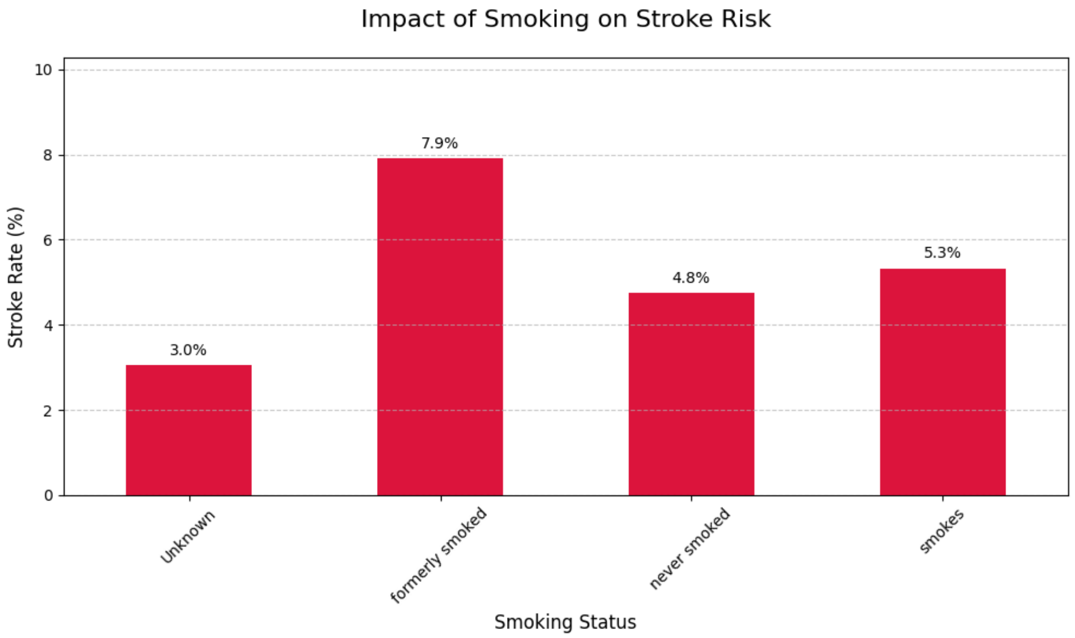


BMI shows a non-linear relationship with stroke risk:

- Underweight (BMI <18.5): 0.29%
- Normal (BMI 18.5-24.9): 2.84%
- Overweight (BMI 25-29.9): 5.32%
- Obese (BMI 30-39.9): 5.25%
- Severely Obese (BMI >40): 4.59%

> Stroke risk increases significantly from normal to overweight categories, then plateaus or slightly decreases in higher BMI categories. This pattern suggests that being overweight or obese increases stroke risk, but the relationship is not simply linear.

# Smoking Status and Stroke Risk



Former smokers show the highest stroke rate (7.91%), followed by current smokers (5.32%) and those who never smoked (4.76%). This pattern may reflect that:

- Former smokers may have quit due to existing health problems
- Former smokers tend to be older on average
- The lingering effects of previous smoking may continue to impact stroke risk

The data suggests that while quitting smoking is beneficial for overall health, the elevated stroke risk may persist for some time after cessation.

# Multivariate Risk Analysis

Our comprehensive analysis reveals that stroke risk is influenced by multiple interacting factors. Below are the key risk factors ranked by their relative impact on stroke risk:

1. **Age -** The strongest predictor of stroke risk, with incidence increasing dramatically after age 60. The average age of stroke patients is significantly higher than non-stroke patients.
2. **Heart disease -** Patients with heart disease have a 17.03% stroke rate compared to 4.18% in those without, representing a ~4.1× increased risk.
3. **Hypertension** - Patients with hypertension have a 13.25% stroke rate compared to 3.97% in those without, representing a 3.34× higher risk.
4. **High glucose levels** - Patients with high diabetic glucose levels (>200 mg/dL) have a 12.90% stroke rate, representing a 3.61× higher risk compared to those with normal levels.
5. **Former smoking status** - Former smokers have a 7.91% stroke rate compared to 4.76% for those who never smoked, representing a 1.66× increased risk.
6. **BMI in overweight/obese range** - Overweight individuals have a 5.32% stroke rate compared to 2.84% for those with normal BMI, representing a 1.87× increased risk.
7. **Current smoking** - Current smokers have a 5.32% stroke rate, representing a 1.12× increased risk compared to those who never smoked.
8. **Residence type -** Urban residents have a 5.20% stroke rate compared to 4.53% for rural residents, representing a relatively minor 1.15× difference.
9. **Gender -** Males have a slightly higher stroke rate (5.11%) compared to females (4.71%), but this difference is minimal (1.08× risk ratio).

When controlling for age, several factors remain significant independent risk factors. This ranking highlights the importance of cardiovascular health (heart disease, hypertension), metabolic factors (glucose levels, BMI), and lifestyle choices (smoking) in stroke risk assessment.

# Model Comparison: Logistic Regression vs. XGBoost

# Logistic Regression for Stroke Prediction

Logistic Regression is chosen by our group as the algorithm to predict the risk of stroke. Logistic Regression is usually used for binary problems. It estimates the probability that a given input is in a given group that whether a patient will suffer from a stroke (1) or not (0).

### Overview:

- **Binary Outcome:** Logistic Regression is designed for binary outcomes. It uses the logistic (sigmoid) function to convert predicted values to probabilities between 0 and 1.
- **Model Formulation:** It considers various factors such as age, blood pressure, smoking status, etc., and assigns a "weight" or importance to each factor. It then combines these weights to come up with a final score that shows how likely a patient will have a stroke.
- **Interpretability:** The weights it assigns can tell you which factors increase or decrease the risk of stroke. For example, if the weight for age is high, it means that the older you are, the higher your risk of a stroke is. This makes it useful, especially in healthcare, because doctors can see which factors matter the most.
- **Optimization:** The model is trained by adjusting weights for accurate predictions. It is similar to turning a radio to receive the best sound.

## Applying Logistic Regression to Stroke Prediction

### Data Preparation

- Cleaning the data by handling missing values, particularly in the BMI column.
- Encoding categorical variables into numerical formats.
- Scaling numerical features to ensure consistent input ranges.

### Model Training and Evaluation

- Once the data is cleaned, the dataset is split into training and testing sets.
- Using training set, the Logistic Regression model learns by adjusting weights for each feature. It attempts to minimize errors, which are measured by a method called log-loss.
- After training, the model is evaluated using various metrics such as:
  - Accuracy: How often the model's predictions are correct.
  - Precision and Recall: These are key for medical use. They help ensure high risk patients are identified,

## Model Performance Analysis

We implemented and compared two machine learning algorithms for stroke prediction: Logistic Regression and XGBoost. Below is an analysis of their performance on the stroke dataset.

## Logistic Regression Results

Logistic Regression with SMOTE Oversampling was used to address the class imbalance in the dataset.

### Performance Metrics

```
--- Logistic Regression with SMOTE Oversampling ---
Accuracy: 75.15 %

Classification Report:
            precision    recall  f1-score   support

        0       0.98      0.75      0.85       960
        1       0.17      0.81      0.28        62

accuracy                           0.75      1022
macro avg       0.58      0.78      0.57      1022
weighted avg       0.93      0.75      0.82      1022
```

### Confusion Matrix

```
Confusion Matrix: [[718 242] [ 12 50]]
```

|  | Predicted No | Predicted Yes |
|---|---|---|
| Actual No | TN: 718 | FP: 242 |
| Actual Yes | FN: 12 | TP: 50 |

### Analysis

- **Accuracy:** 75.15% - The model correctly predicts about three-quarters of all cases.
- **Precision for Stroke (Class 1):** 0.17 - Only 17% of predicted stroke cases are actual strokes, indicating many false positives.
- **Recall for Stroke (Class 1):** 0.81 - The model captures 81% of all actual stroke cases, which is quite good.
- **False Positive Rate:** 25.2% - About one-quarter of patients without stroke are incorrectly classified as having stroke risk.
- **F1-Score for Stroke (Class 1):** 0.28 - The harmonic mean of precision and recall is low due to the poor precision.
- **Confusion Matrix:** Shows 50 true positives, 718 true negatives, 242 false positives, and 12 false negatives.

### Strengths and Weaknesses

- **Strengths:** High recall for stroke cases (0.81) means the model is good at identifying patients who will have a stroke.
- **Weaknesses:** Low precision (0.17) means many patients will be falsely identified as at risk for stroke.

## XGBoost for Stroke Prediction

XGBoost (Extreme Gradient Boosting) is another powerful algorithm we tested for predicting stroke risk. It's designed to handle complex patterns in data that simpler models might miss.

### Overview:

- **Tree-Based Learning:** XGBoost builds many decision trees that work together as a team. Each tree helps correct mistakes made by previous trees, gradually improving predictions.
- **Pattern Recognition:** Unlike Logistic Regression which looks at factors independently, XGBoost can discover how different factors interact. For example, it might learn that age affects stroke risk differently for smokers versus non-smokers.
- **Feature Importance:** Similar to Logistic Regression, XGBoost can tell us which factors matter most for stroke prediction, but it captures more complex relationships between these factors.
- **Handling Missing Data:** XGBoost has built-in methods to deal with missing information, which is helpful for medical datasets where some patient information might be incomplete.

# Applying XGBoost to Stroke Prediction

## Data Preparation

- The same preprocessing steps used for Logistic Regression were applied: handling missing values, encoding categorical variables, and scaling features.
- SMOTE oversampling was also used to address the class imbalance issue.

## Model Training and Evaluation

- XGBoost learns by building decision trees one after another, with each new tree focusing on the errors made by previous trees.
- The model uses a technique called "gradient boosting" to minimize prediction errors.
- We evaluated XGBoost using the same metrics as Logistic Regression:
  - Accuracy: The overall percentage of correct predictions
  - Precision: How many of the patients predicted to have strokes actually had them
  - Recall: How many actual stroke cases the model successfully identified
  - F1-Score: A balance between precision and recall

### XGBoost Results

XGBoost (Extreme Gradient Boosting) was also applied with SMOTE oversampling to handle the class imbalance.

### Performance Metrics

```
--- XGBoost Classifier Results ---
Accuracy: 90.8 %

Classification Report:
                precision    recall  f1-score   support

            0       0.94      0.96      0.95       960
            1       0.14      0.10      0.11        62

    accuracy                           0.91      1022
    macro avg       0.54      0.53      0.53      1022
weighted avg       0.89      0.91      0.90      1022
```

### Confusion Matrix

```
Confusion Matrix: [[922 38] [ 56 6]]
```

| | Predicted No | Predicted Yes |
|---|---|---|
| **Actual No** | TN: 922 | FP: 38 |
| **Actual Yes** | FN: 56 | TP: 6 |

### Analysis

- **Accuracy:** 90.8% - The model has a high overall accuracy rate.

- **Precision for Stroke (Class 1):** 0.14 - Only 14% of predicted stroke cases are actual strokes.
- **Recall for Stroke (Class 1):** 0.10 - The model only captures 10% of all actual stroke cases, which is quite low.

  **False Positive Rate:** 4.0% - Only a small percentage of non-stroke patients are incorrectly classified as having stroke risk.
- **F1-Score for Stroke (Class 1):** 0.11 - The harmonic mean of precision and recall is very low.
- **Confusion Matrix:** Shows 6 true positives, 922 true negatives, 38 false positives, and 56 false negatives.

### Strengths and Weaknesses

- **Strengths:** Very high accuracy for non-stroke cases (0.96 recall for class 0) means the model rarely misclassifies healthy patients.
- **Weaknesses:** Very low recall for stroke cases (0.10) means the model misses 90% of patients who would have a stroke. This is a critical limitation in a medical context where failing to identify at-risk patients could have life-threatening consequences.

# Detailed Comparison Between Logistic Regression and XGBoost

| Aspect | Logistic Regression | XGBoost |
|---|---|---|
| **Accuracy** | 75.15% | 90.8% |
| **Precision (Class 1)** | 0.17 | 0.14 |
| **Recall (Class 1)** | 0.81 | 0.10 |
| **False Positive Rate** | 25.2% | 4.0% |
| **F1-Score (Class 1)** | 0.28 | 0.11 |
| **True Positives** | 50 | 6 |
| **False Positives** | 242 | 38 |
| **True Negatives** | 718 | 922 |
| **False Negatives** | 12 | 56 |

## Performance Analysis

### 1. Overall Accuracy

XGBoost achieves significantly higher overall accuracy (90.8%) compared to Logistic Regression (75.15%). However, this metric can be misleading in imbalanced datasets like this one, where stroke cases (class 1) are much fewer than non-stroke cases (class 0).

### 2. Class Imbalance Handling

Despite both models using SMOTE for oversampling:

- **Logistic Regression** shows a bias toward predicting stroke cases, resulting in many false positives but few false negatives.
- **XGBoost** shows a bias toward predicting non-stroke cases, resulting in few false positives but many false negatives.

**3. Stroke Detection Capability**

The models differ dramatically in their ability to detect actual stroke cases:

- **Logistic Regression** captures 81% of stroke cases (50 out of 62), making it much more effective at identifying patients at risk.
- **XGBoost** only captures 10% of stroke cases (6 out of 62), missing 90% of patients who would have a stroke.

**4. False Alarm Rate**

The models also differ in their tendency to raise false alarms:

- **Logistic Regression** has a high false positive rate, incorrectly flagging 242 out of 960 non-stroke cases as stroke risks.
- **XGBoost** has a much lower false positive rate, incorrectly flagging only 38 out of 960 non-stroke cases.

## Real-World Healthcare Applications

The choice between these models depends on the clinical priorities:

**Scenario 1: Prioritizing Detection of All Stroke Cases**

**Logistic Regression would be preferred** if the primary goal is to identify as many potential stroke patients as possible, even at the cost of false alarms. This approach is valuable when:

- Missing a stroke diagnosis could be life-threatening
- Additional tests can be performed to confirm positive predictions
- The cost of follow-up testing is relatively low compared to the cost of missing a stroke case

The false positive rate matters a lot in healthcare. With the Logistic Regression model, about 1 in 4 (25,2%) healthy patients would be wrongly flagged as at risk for stroke. This means unnecessary tests, patient worry, and higher costs. The XGBoost model has fewer false alarms (only 4%), but misses 9 out of 10 actual stroke cases - a dangerous trade-off when strokes require urgent treatment.

### Scenario 2: Minimizing False Alarms

**XGBoost would be preferred** if the primary goal is to minimize false positives and focus resources only on the most certain cases. This approach might be valuable when:

- Resources for follow-up testing are limited
- False positives could lead to unnecessary patient anxiety or costly procedures
- The model is used as an initial screening tool before more thorough evaluations

## Technical Considerations

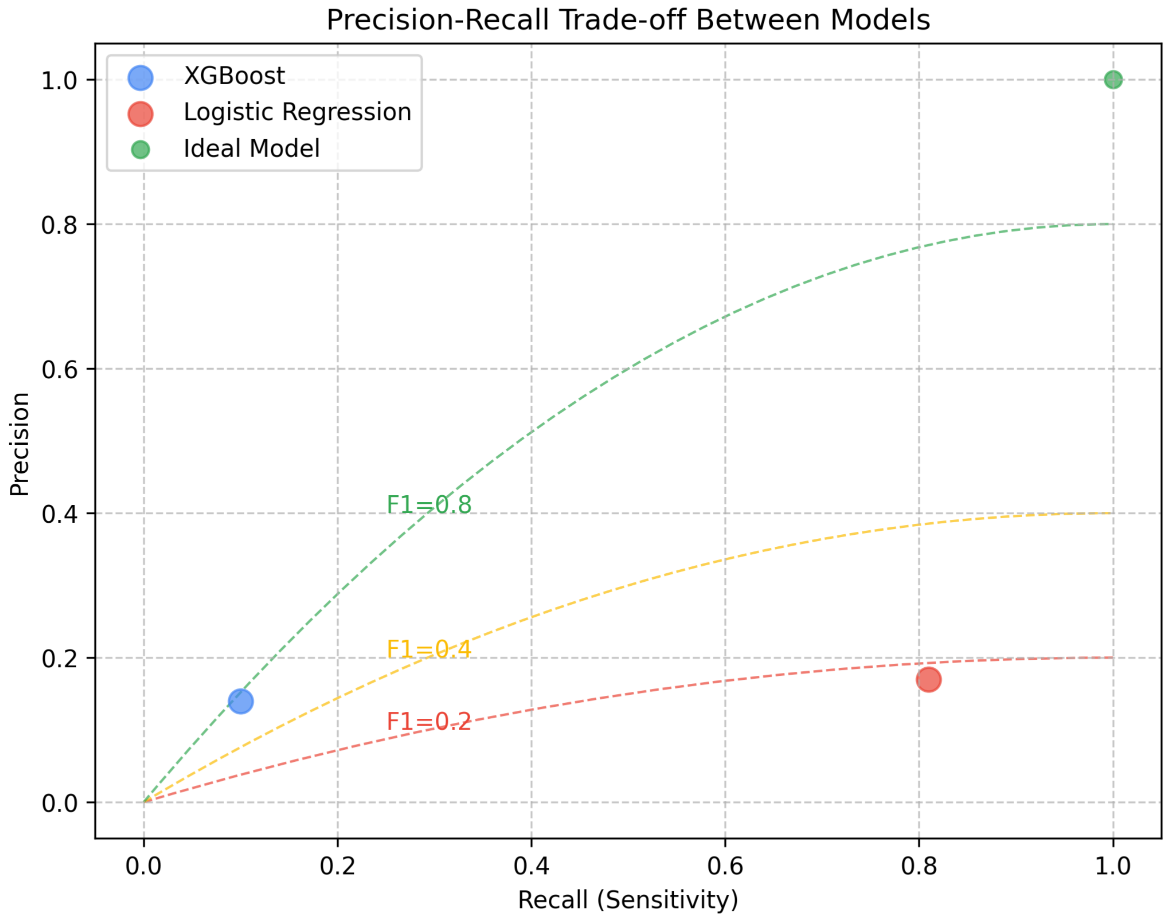
### 1. Model Complexity

- **Logistic Regression** is a simpler, more interpretable model that assigns linear weights to features.
- **XGBoost** is a more complex ensemble model that can capture non-linear relationships and feature interactions.

### 2. Potential for Improvement

- **Logistic Regression** could benefit from adjusting the classification threshold to reduce false positives while maintaining reasonable recall.
- **XGBoost** could benefit from different sampling techniques, class weighting, or cost-sensitive learning to improve its ability to detect stroke cases.

## Visualization of Model Trade-offs



*Visualization of the precision-recall trade-off between the two models. The curved lines represent constant F1 scores.*

## Recommendation

Based on the results:

- **Logistic Regression for Screening:**
  For a **screening tool**, the Logistic Regression model is recommended despite its lower accuracy (75.15%) because:
  - **High Recall (81%):** It correctly identifies 81% of actual stroke cases, missing only 19% of patients who will have a stroke. In early screening, missing a potential stroke case (false negative) can be life-threatening.

> - **Medical Context:** Strokes require rapid intervention - "time is brain." The consequences of missing a stroke are far more severe than the inconvenience of a false alarm.
>     - **Risk Management:** In medical screening, we typically prefer to cast a wider net initially, accepting some false positives to ensure we catch most true cases.
>
> - **XGBoost for Confirmation:**
>   For a **confirmatory tool**, the XGBoost model might be preferred because:
>
>     - **Low False Positive Rate (4%):** It rarely misclassifies healthy patients as having stroke risk, which means fewer unnecessary follow-up procedures.
>     - **High Specificity (96%):** It correctly identifies 96% of non-stroke cases, which is valuable when resources for advanced testing are limited.
>     - **Resource Allocation:** Confirmatory tests are often more expensive, invasive, or limited in availability, making it important to target them to patients most likely to benefit.
>
>   However, its low recall for stroke cases (10%) remains a significant limitation even for confirmation purposes.

## References

1. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
2. Brownlee, J. (2020). XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. Machine Learning Mastery.
3. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.