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	BMI	Smoking Status	Stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1

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)

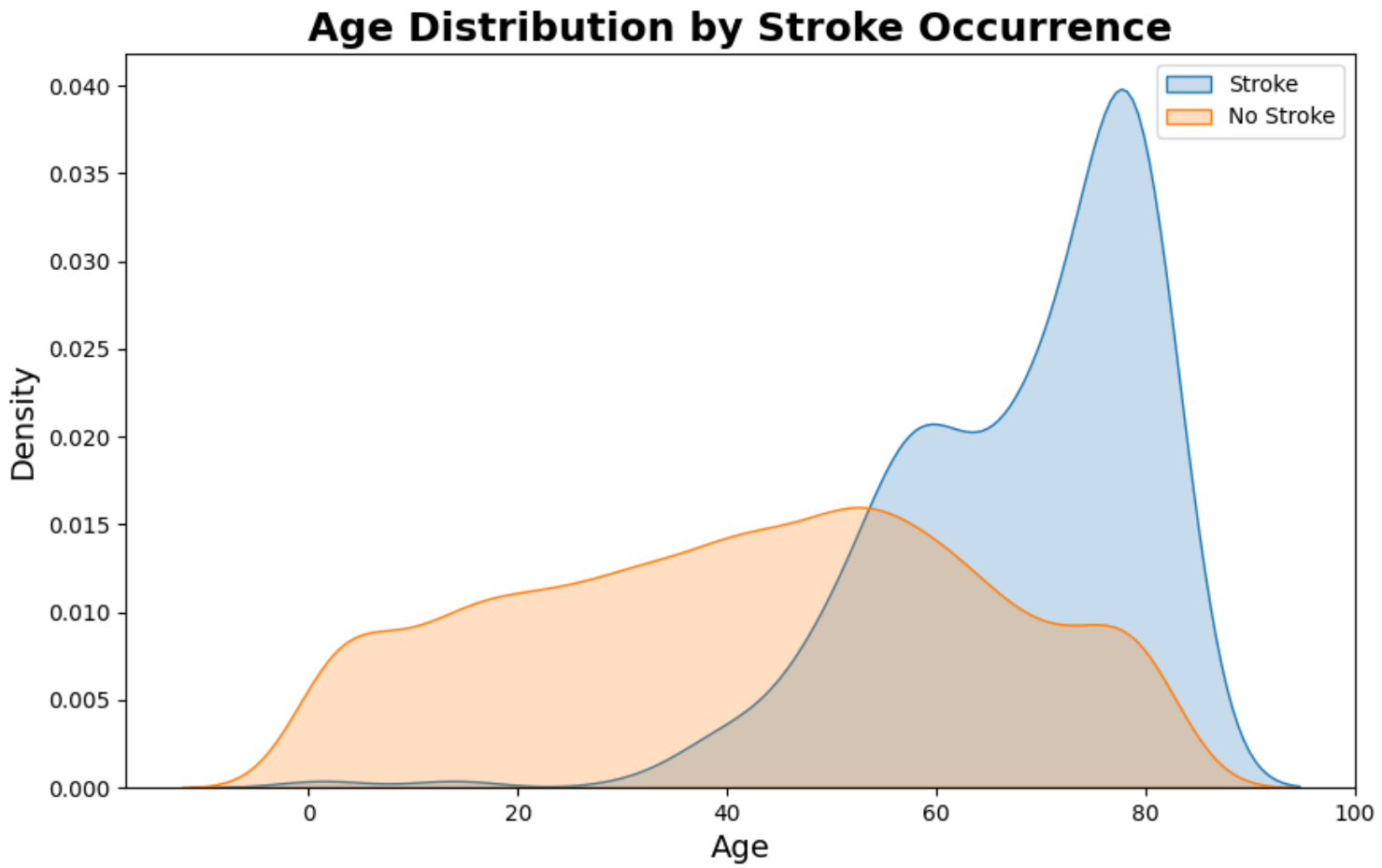
Residence Type: "Urban" is the top category (2596 cases)

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

Data Visualization and Analysis

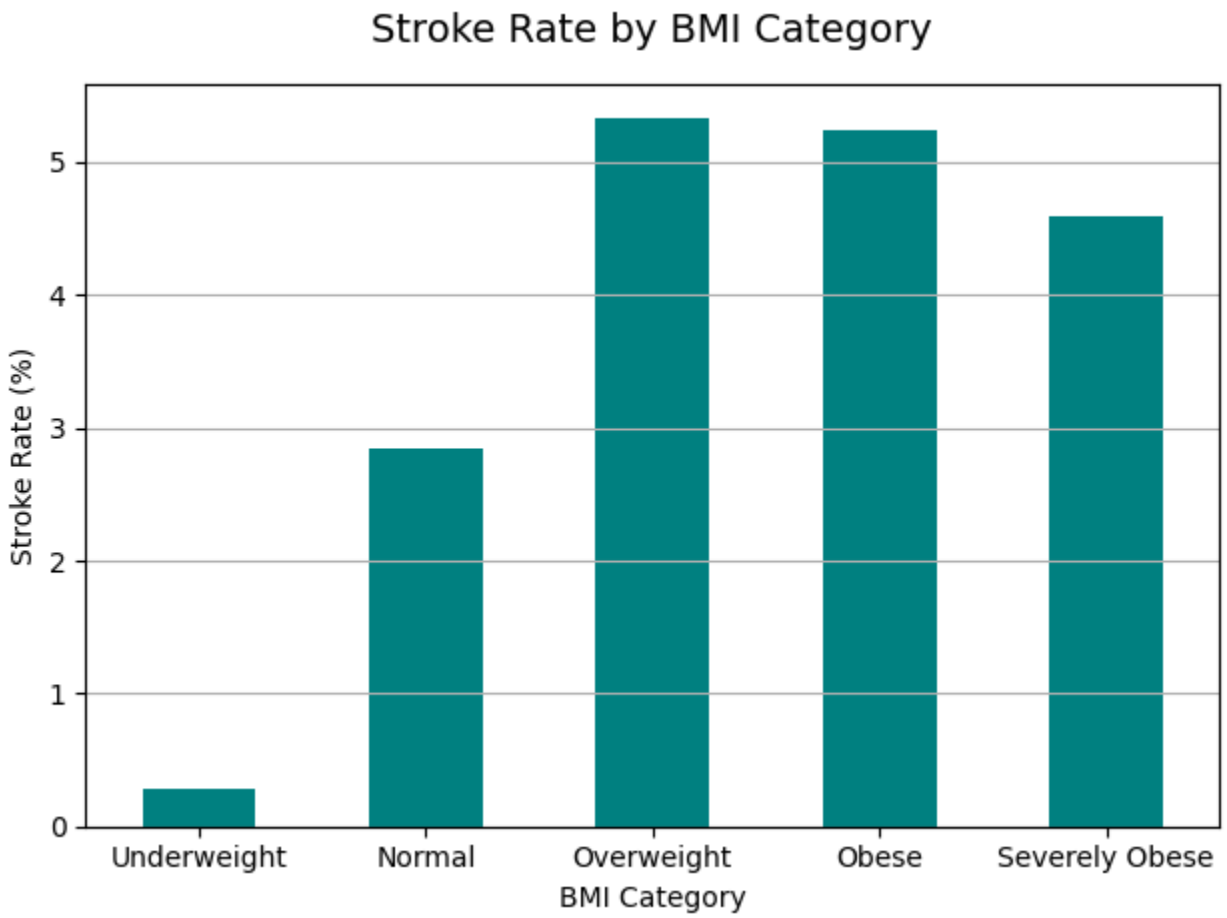
Visual analysis of the dataset reveals important patterns and relationships between various factors and stroke occurrence. The following charts highlight key insights from the data:

Age Distribution by Stroke Occurrence



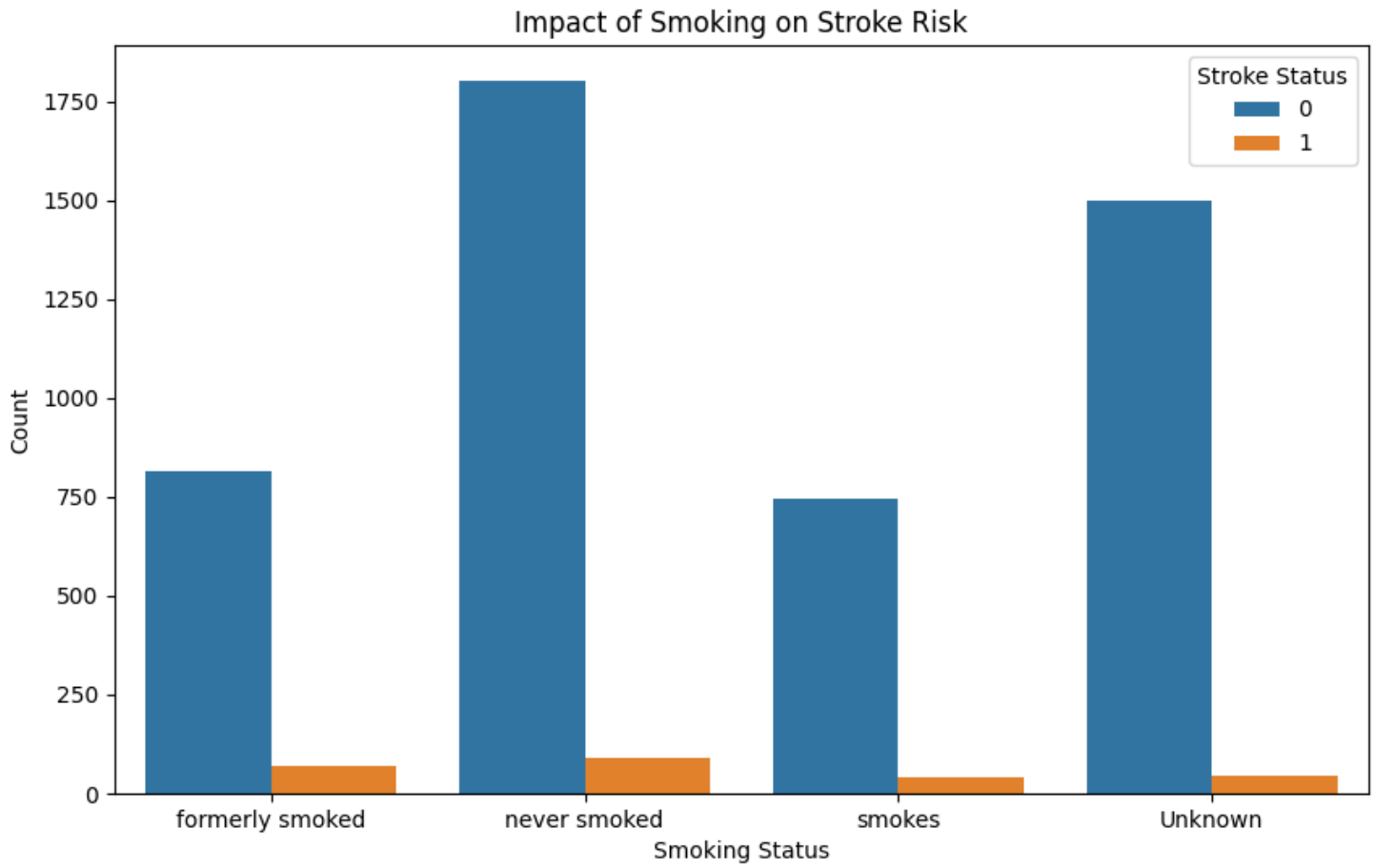
> This visualization shows the age distribution for patients with and without stroke. The density plot clearly demonstrates that stroke cases are more prevalent in older age groups, with the peak for stroke patients occurring at a much higher age compared to non-stroke patients. This confirms that age is one of the most significant risk factors for stroke.

BMI Categories and Stroke Rate



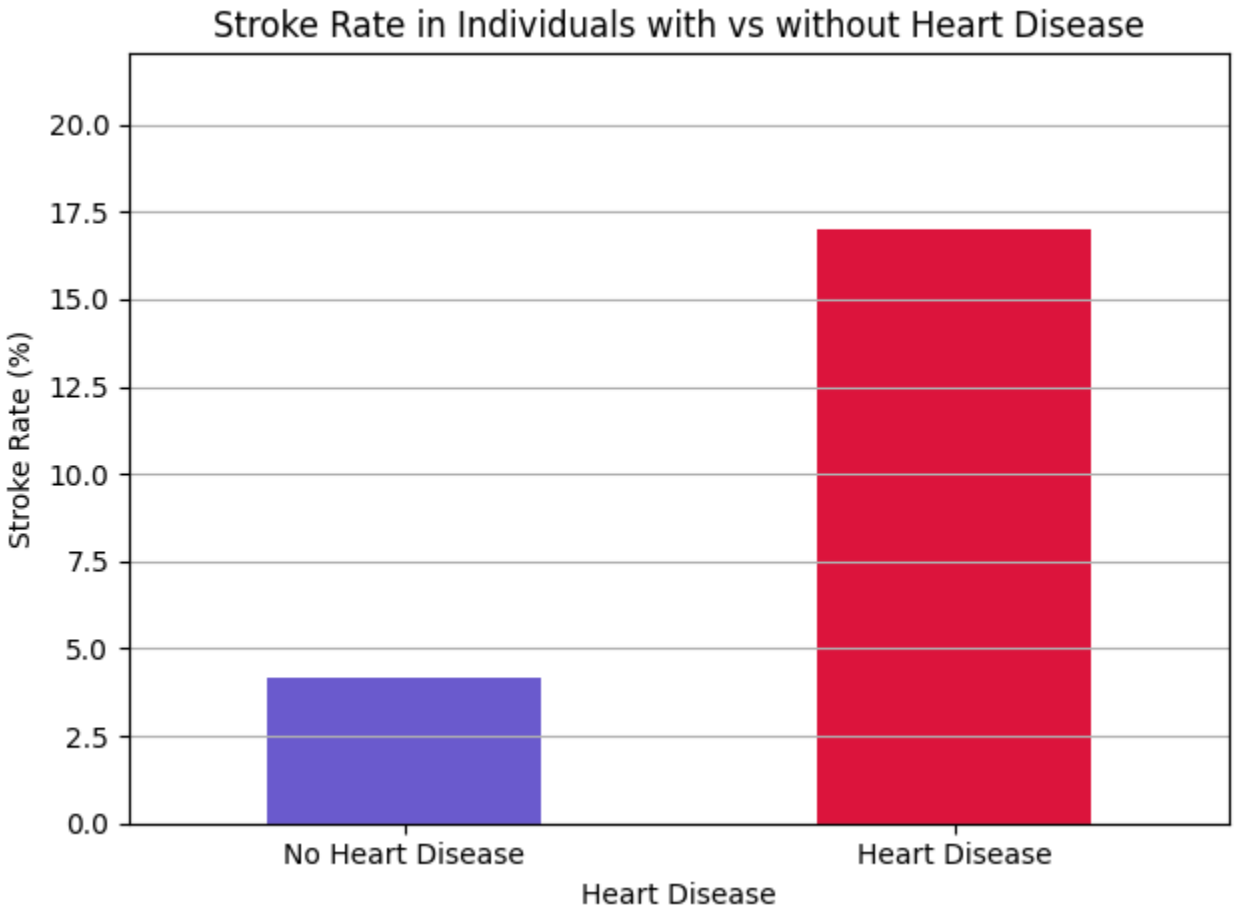
> This chart examines the relationship between BMI categories (Underweight, Normal, Overweight, Obese, Severely Obese) and stroke rate. The data shows how stroke risk varies across different body mass index ranges, providing insights into how weight management might relate to stroke prevention.

Impact of Smoking on Stroke Risk



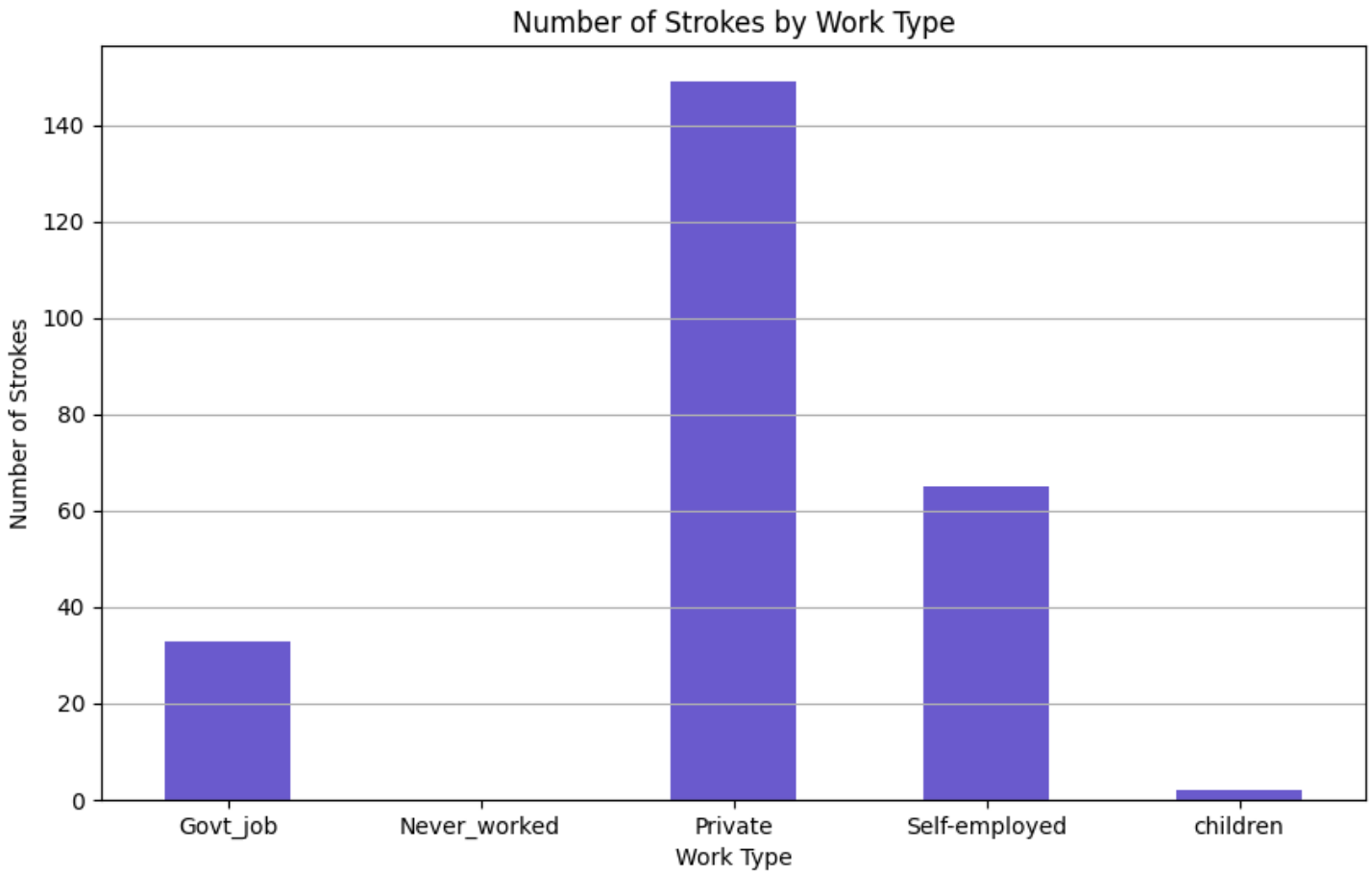
> This visualization shows the count of stroke and non-stroke cases across different smoking statuses (never smoked, formerly smoked, currently smokes, and unknown). The chart helps identify whether certain smoking behaviors are associated with higher stroke occurrence in the dataset.

Relationship Between Heart Disease and Stroke



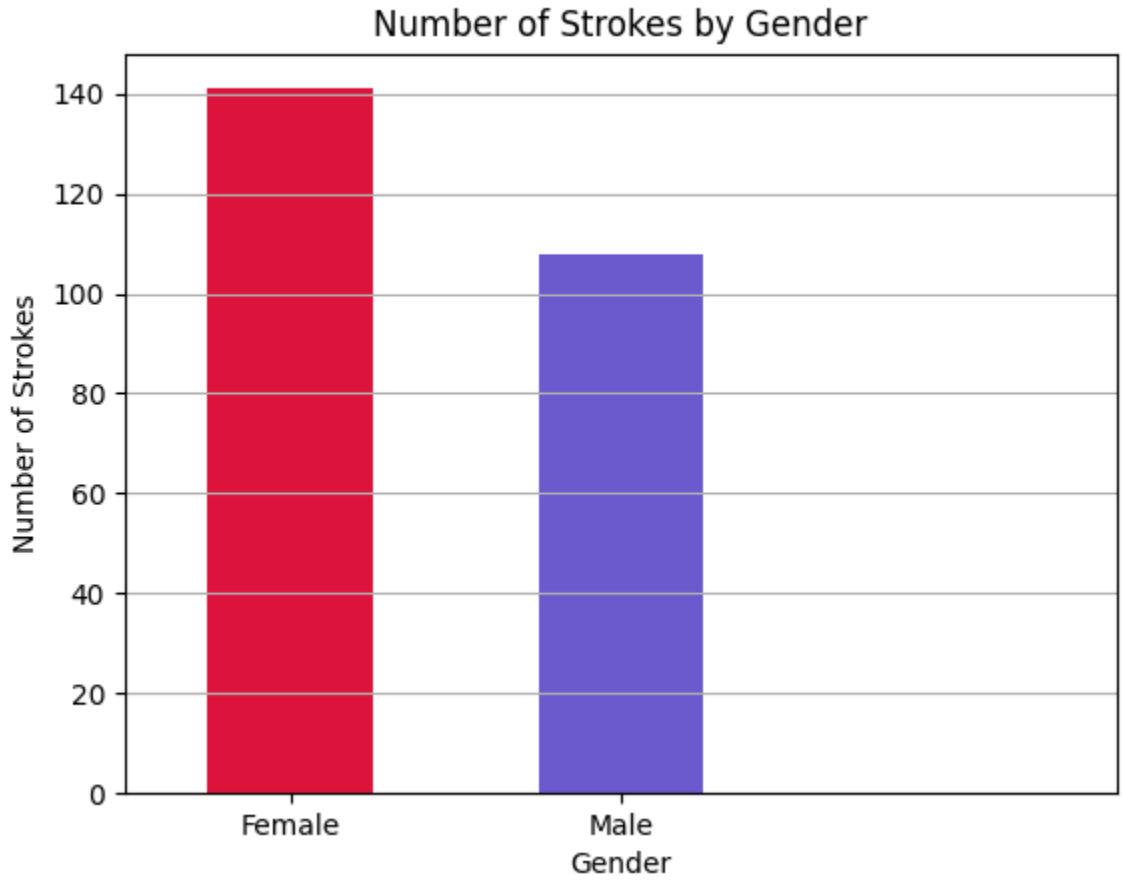
This bar chart compares the stroke rate between individuals with and without heart disease. The significant difference in stroke rates highlights the strong association between heart disease and stroke risk, emphasizing the importance of cardiovascular health in stroke prevention.

Relationship Between Work Type and Stroke



> This visualization shows the number of stroke cases across different work types (Private, Self-employed, Government job, etc.). The distribution helps identify whether certain occupational categories might be associated with higher stroke risk, potentially due to lifestyle factors, stress levels, or other work-related variables.

Stroke Occurrence by Gender



> This bar chart compares the number of stroke cases between males and females. The visualization helps identify any gender-based differences in stroke occurrence, which could be important for targeted prevention strategies and understanding gender-specific risk factors.

Logistic Regression for Stroke Prediction

For this project, Logistic Regression is chosen as the algorithm to predict the risk of stroke. Logistic Regression is usually used for two-choice (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,

Stroke Prediction Using Machine Learning

> 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.

Data Preview

ID	Gender	Age	Hypertension	Heart Disease	Ever Married	Work Type	Residence Type	Avg Glucose Level	BMI	Smoking Status	Stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1

ID	Gender	Age	Hypertension	Heart Disease	Ever Married	Work Type	Residence Type	Avg Glucose Level	BMI	Smoking Status	Stroke
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

> After applying SMOTE oversampling:
>
> - Label '1' (stroke): 3901 samples
> - Label '0' (no stroke): 3901 samples

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.
- **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.
- The model has a bias toward predicting stroke cases, likely due to the SMOTE oversampling technique used to balance the classes.

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.
- **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.

> to address class imbalance. Below is a comprehensive comparison of their performance, strengths, and limitations.

Performance Analysis

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

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.

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

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. Response to SMOTE Oversampling

The models responded differently to the SMOTE oversampling technique:

- **Logistic Regression** appears to have been more influenced by the synthetic samples, leading to its higher sensitivity for stroke cases.
- **XGBoost** seems to have been less affected by oversampling, maintaining a stronger bias toward the majority class.

3. 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-recal

- For a **screening tool** where identifying all potential stroke cases is critical, the **Logistic Regression model** would be more appropriate despite its lower overall accuracy.
- For a **confirmatory tool** where minimizing false positives is important, the **XGBoost model** might be preferred, but its low recall for stroke cases is a significant limitation.
- A **hybrid approach** could be valuable, using Logistic Regression for initial screening and then applying additional criteria to reduce false positives.

Given the life-threatening nature of strokes, the higher recall of the Logistic Regression model likely makes it more valuable in this specific healthcare context, despite its lower overall accuracy.