Fundamentals of Data Science in Business and Engineering
Project Report

Merve Pakcan Tufenk

Goal: The goal of this project is to analyze and derive meaningful insights from a healthcare dataset focusing on stroke prediction and its contributing factors. By utilizing the statistical and analytical techniques covered in this course, the aim is to explore the relationships between key health indicators (such as age, BMI, glucose levels, and smoking status) and the likelihood of stroke occurrences. This project seeks to demonstrate proficiency in data analysis by uncovering patterns, relationships, and variations in stroke risk and health metrics across different demographics.

This project emphasizes not only interpreting and presenting the results of data analysis but also critically evaluating the methods used to ensure meaningful and accurate conclusions. Insights from this analysis aim to contribute to better healthcare strategies, enabling improved prevention and treatment of stroke.

Part 1 - Questions

This project will address the following key research questions:

1. **Which factors have the greatest impact on stroke risk?**
    - **Objective**: To investigate the relative importance of variables such as age, hypertension, heart disease, and BMI in determining stroke risk.
2. **What are the differences in stroke risk and contributing factors across demographic groups (age, gender, residence type, work type)?**
    - **Objective**: To analyze variations in stroke risk and its contributing factors across different demographic and lifestyle groups.
3. **How do health metrics (BMI, glucose levels) correlate with stroke risk, and are there thresholds for increased risk?**
    - **Objective**: To explore how specific health indicators influence stroke risk and identify potential cutoff points for increased likelihood of stroke.

These questions aim to provide a comprehensive understanding of the factors influencing stroke occurrences and the variations observed across different groups.

Part 2: Data

The dataset used for this project is the **Stroke Prediction Dataset** from Kaggle, which provides critical insights into predicting the likelihood of a patient experiencing a stroke. According to the World Health Organization (WHO), stroke is the second leading cause of death globally, accounting for approximately 11% of all deaths. Early identification of individuals at high risk is essential for improving prevention strategies and healthcare outcomes.

This dataset contains detailed information about 5110 patients, including demographic attributes (gender, age, and marital status), health indicators (hypertension, heart disease, BMI, and glucose levels), and lifestyle factors (smoking status). Each row in the dataset represents a unique individual, offering relevant health and lifestyle information for stroke risk analysis.

Initially, the dataset comprised 5110 rows and 12 columns. During data cleaning, 201 missing values were identified in the BMI variable. As these missing values represented a small portion of the dataset, they were eliminated to ensure the integrity and consistency of the analysis. This step ensured that the remaining data was complete and ready for meaningful analysis.After data cleaning, the dataset has 4908 rows.

This dataset is a refined and clean subset of the original dataset, which is based on the Electronic Health Record (EHR) controlled by McKinsey & Company. It includes variables such as age, BMI, glucose levels, hypertension, and smoking status, which seem to be measured through standardized health record entries. The dataset is well-structured and shows no signs of bias or imbalance, ensuring reliable and generalizable results for the analysis.

This dataset contains the following 12 key variables:

1. **id**: Unique identifier for each individual.
2. **gender**: Gender of the individual (Male/Female/Other).
3. **age**: Age of the individual in years.
4. **hypertension**: Whether the individual has hypertension (0 = No, 1 = Yes).
5. **heart_disease**: Whether the individual has heart disease (0 = No, 1 = Yes).
6. **ever_married**: Whether the individual has ever been married (Yes/No).
7. **work_type**: Type of employment (e.g., Private, Self-employed, Government Job, Never Worked).
8. **Residence_type**: Type of residence (Urban/Rural).
9. **avg_glucose_level**: Average glucose level in the individual's blood.
10. **bmi**: Body Mass Index (weight-to-height ratio).
11. **smoking_status**: Smoking habits (e.g., formerly smoked, never smoked, smokes).
12. **stroke**: Outcome variable indicating whether the individual experienced a stroke (0 = No, 1 = Yes).

With a mix of 9 categorical and 3 numerical variables, this dataset provides a robust foundation for analyzing the relationships between health indicators, demographics, and the risk of stroke.

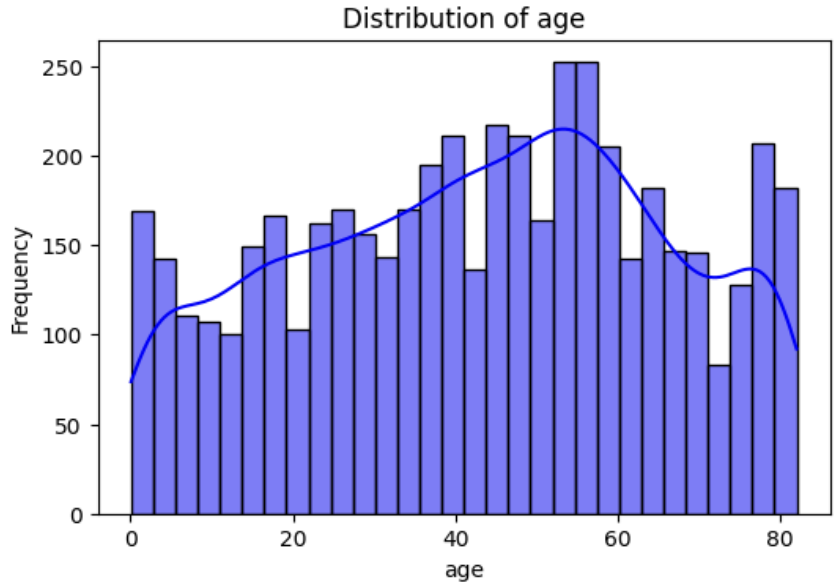
The dataset plays a significant role in advancing predictive healthcare by enabling data-driven approaches to identify key factors influencing stroke risk. Through comprehensive analysis, this project aims to uncover meaningful patterns and relationships, contributing to improved healthcare strategies and policies for stroke prevention and treatment. It is publicly available and can be accessed at the following link:

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

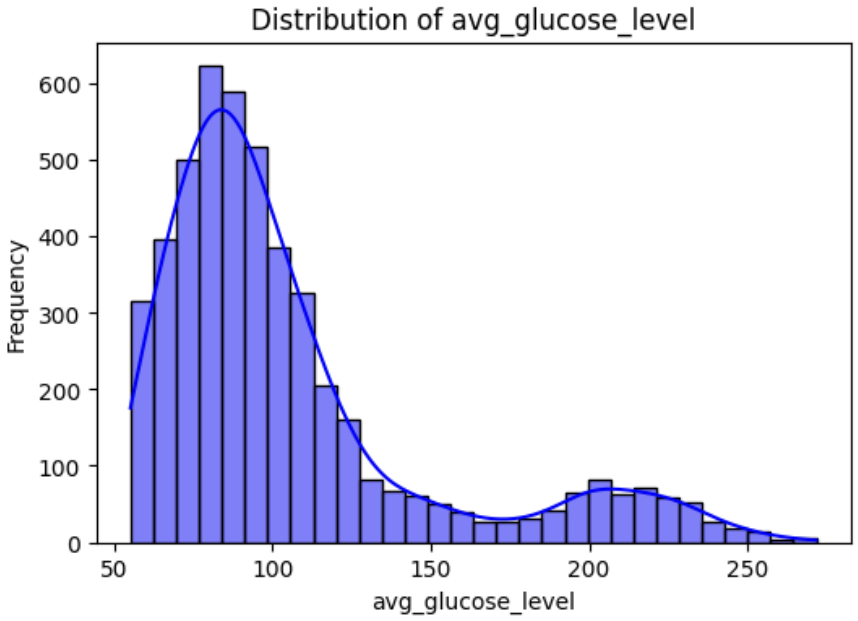
Part 3: Exploratory Data Analysis

```
Column: age
Mean: 42.87
Median: 44.00
Standard Deviation: 22.56

Column: avg_glucose_level
Mean: 105.30
Median: 91.68
Standard Deviation: 44.43

Column: bmi
Mean: 28.89
Median: 28.10
Standard Deviation: 7.85
```

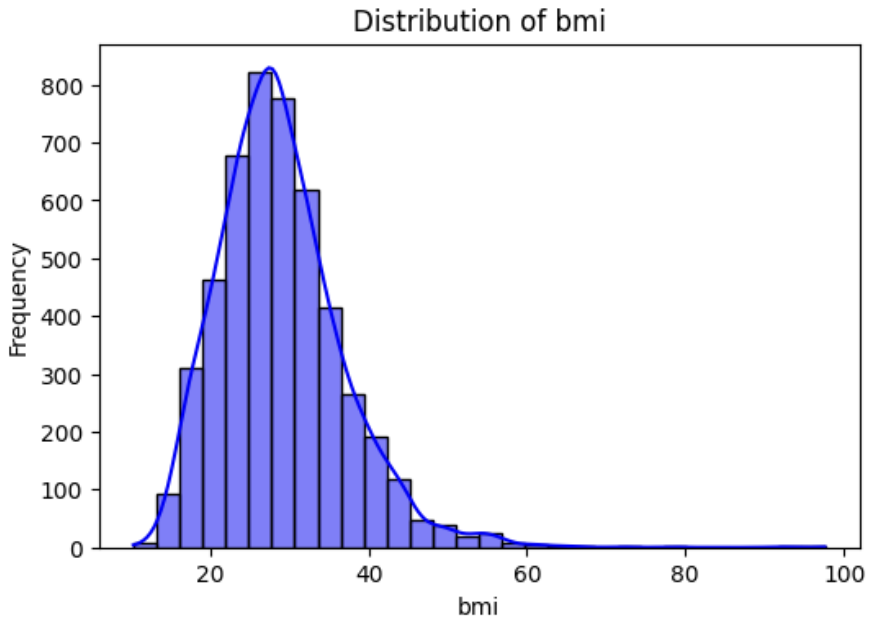
The dataset reveals that the average age is 42.87 years, with a wide age distribution (SD: 22.56). The average glucose level is 105.30 mg/dL, showing high variability (SD: 44.43) and a potential right-skewed distribution. The average BMI is 28.89, indicating a tendency toward overweight, with moderate variability (SD: 7.85). These statistics highlight a diverse population in terms of age, glucose levels, and BMI, warranting further analysis of potential health risks.



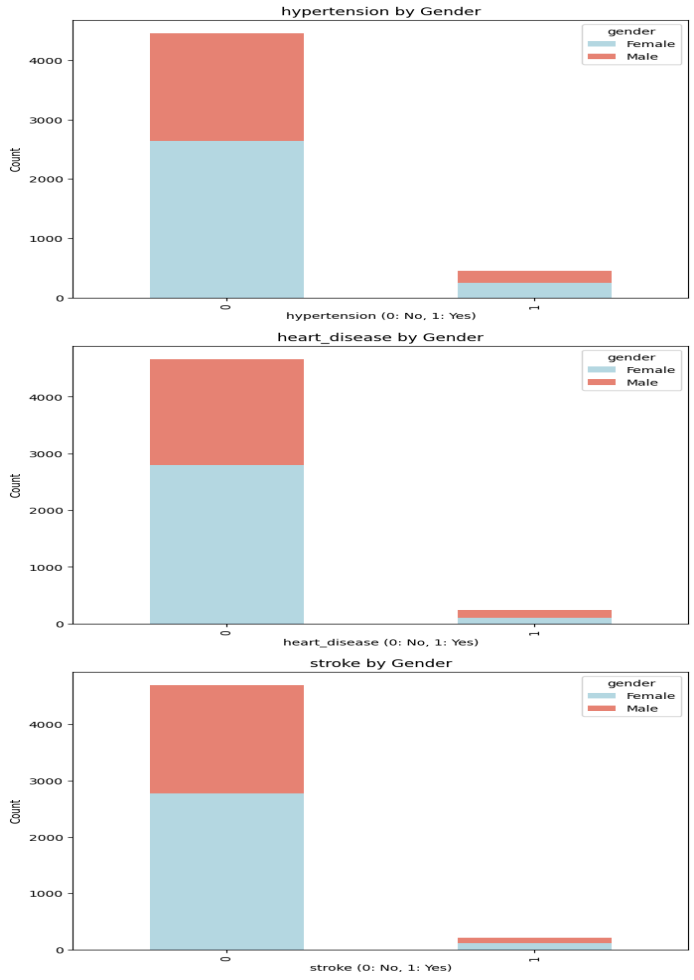
The histogram shows the distribution of age in the dataset. It indicates that the dataset includes individuals from a wide range of ages, with a notable peak around the 60-70 age group. The age distribution appears fairly balanced, though there is a slight skew toward older individuals. The presence of younger age groups is also significant, indicating that the dataset captures a diverse population across different age brackets.



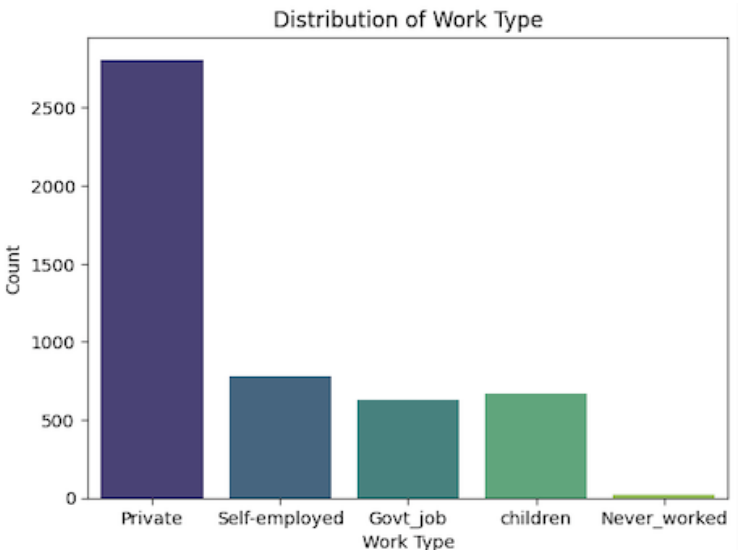
The histogram displays the distribution of average glucose levels in the dataset. It shows a right-skewed distribution, with the majority of individuals having glucose levels between 70 and 120 mg/dL. There is a sharp decline in frequency for higher glucose levels, but a smaller secondary peak is visible around 200 mg/dL, which could indicate a subgroup with elevated glucose levels, possibly individuals with diabetes or other metabolic conditions. Overall, the data highlights a significant concentration of individuals within the normal range, with fewer cases of very high glucose levels.



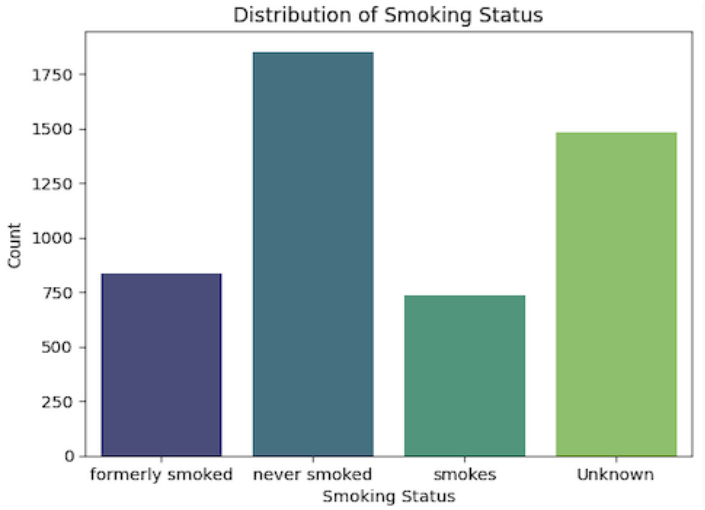
The BMI distribution is slightly right-skewed, with the majority of values falling between 20 and 35, which corresponds to the normal and overweight ranges. The peak is around 25-30, indicating that many individuals are in the overweight category. Higher BMI values above 40, representing obesity, are less frequent, and extreme cases above 60 are rare.



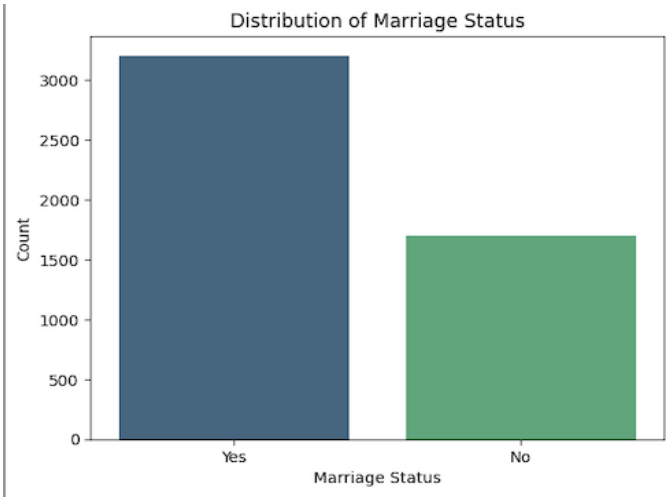
The charts show that hypertension, heart disease, and stroke are relatively uncommon in the dataset. Overall, the majority of individuals do not exhibit these health conditions.



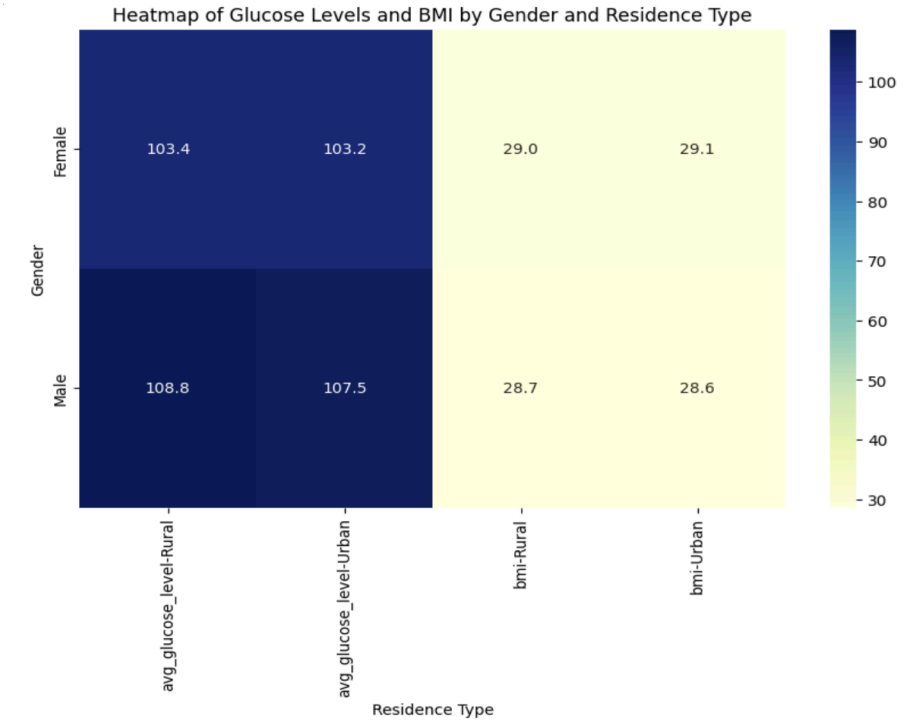
The bar chart shows the distribution of work types in the dataset. The majority of individuals work in the private sector, with a count significantly higher than other categories. Self-employed, government jobs, and children have similar but much lower counts. The "Never_worked" category is minimal, indicating that most individuals in the dataset have some form of employment. This highlights a strong representation of the private sector in the dataset.



The bar chart displays the distribution of smoking status in the dataset. The largest group consists of individuals who have never smoked, followed by those with an "Unknown" smoking status. Smaller but similar proportions are seen for individuals who formerly smoked and those who currently smoke. This suggests that most individuals either never smoked or their smoking status is not recorded, while a smaller portion has a history of smoking.



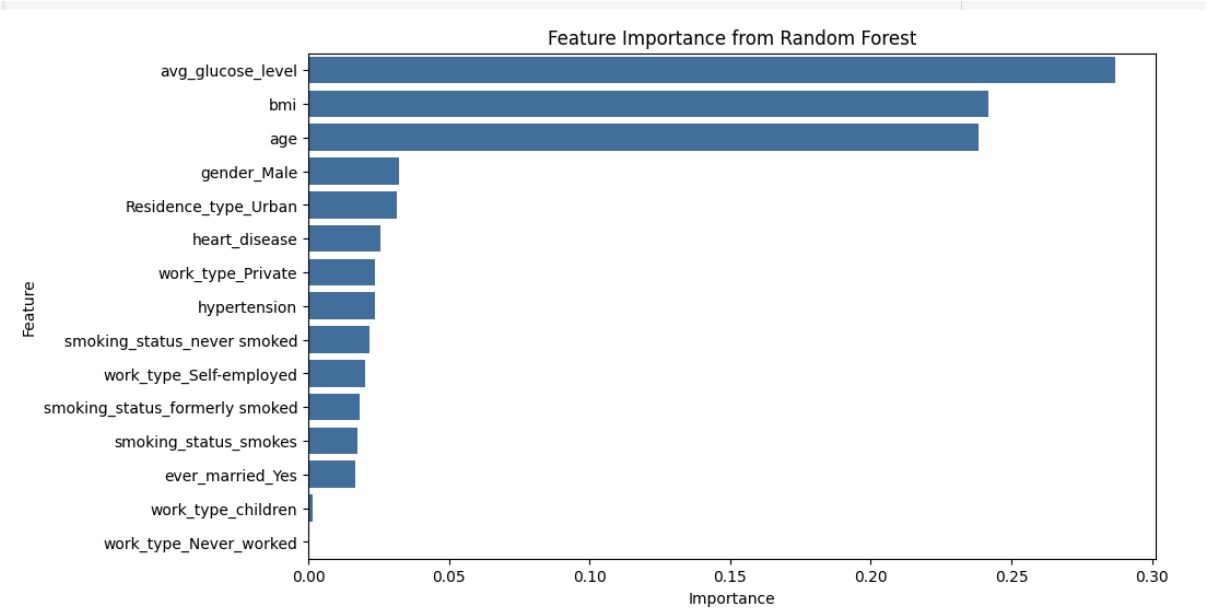
The bar chart illustrates the distribution of marriage status in the dataset. A majority of individuals are married ("Yes"), with their count significantly higher than those who are not married ("No"). This indicates that married individuals constitute a larger portion of the dataset compared to unmarried ones.



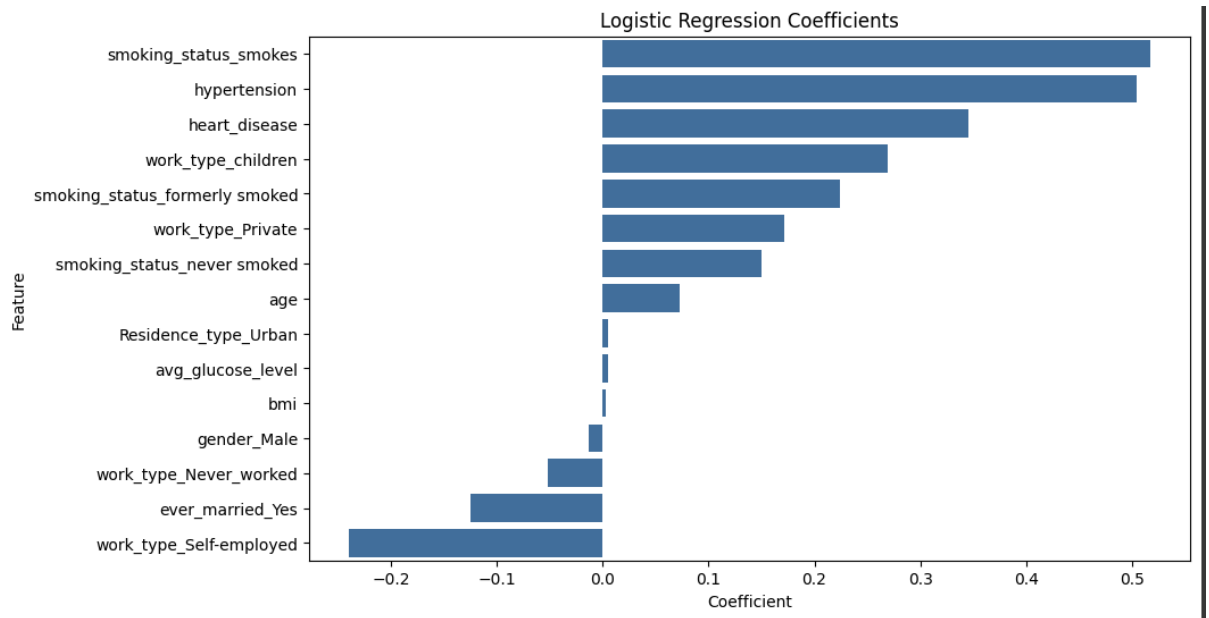
The heatmap visualizes the distribution of average glucose levels and BMI across different gender and residence types. **Key observations include**: Males have higher average glucose levels compared to females in both rural (108.8 mg/dL vs. 103.4 mg/dL) and urban (107.5 mg/dL vs. 103.2 mg/dL) settings, indicating that males may be at a greater risk for glucose-related health issues. For BMI, the differences between genders are minimal, with females slightly higher in urban areas (29.1 vs. 28.6) and males slightly higher in rural areas (28.7 vs. 29.0).These results suggest that residence type has a lesser impact on BMI, whereas gender plays a more significant role in average glucose levels.

Part 4: Data Analysis

In the data analysis section, several methods were used to address key research questions: Random Forest and logistic regression were used to determine the relative importance of factors such as age, hypertension, and BMI in stroke risk. ANOVA and Chi-Square tests analyzed differences in stroke risk across demographic groups like gender and work type. Correlation plots and ROC analysis identified the impact of health metrics, such as glucose levels, and established optimal thresholds for increased risk. These approaches provided a comprehensive understanding of stroke risk factors and their variations.



The feature importance graph from the Random Forest model highlights the key factors influencing stroke risk. The most significant variable is the average glucose level, which has the highest impact on stroke prediction. BMI and age also play a crucial role, contributing significantly to risk estimation. Hypertension and heart disease show a moderate effect. This analysis underscores the critical role of health indicators, particularly glucose levels, in predicting stroke risk, emphasizing their importance in targeted prevention and early intervention strategies.




```
One-way ANOVA result for age: F-statistic = 279.87841499632003, p-value = 3.8408903844855186e-61
Chi-Square Test for Gender and Stroke: Chi2 = 0.16955129804441268, p-value = 0.6805108914997836
```

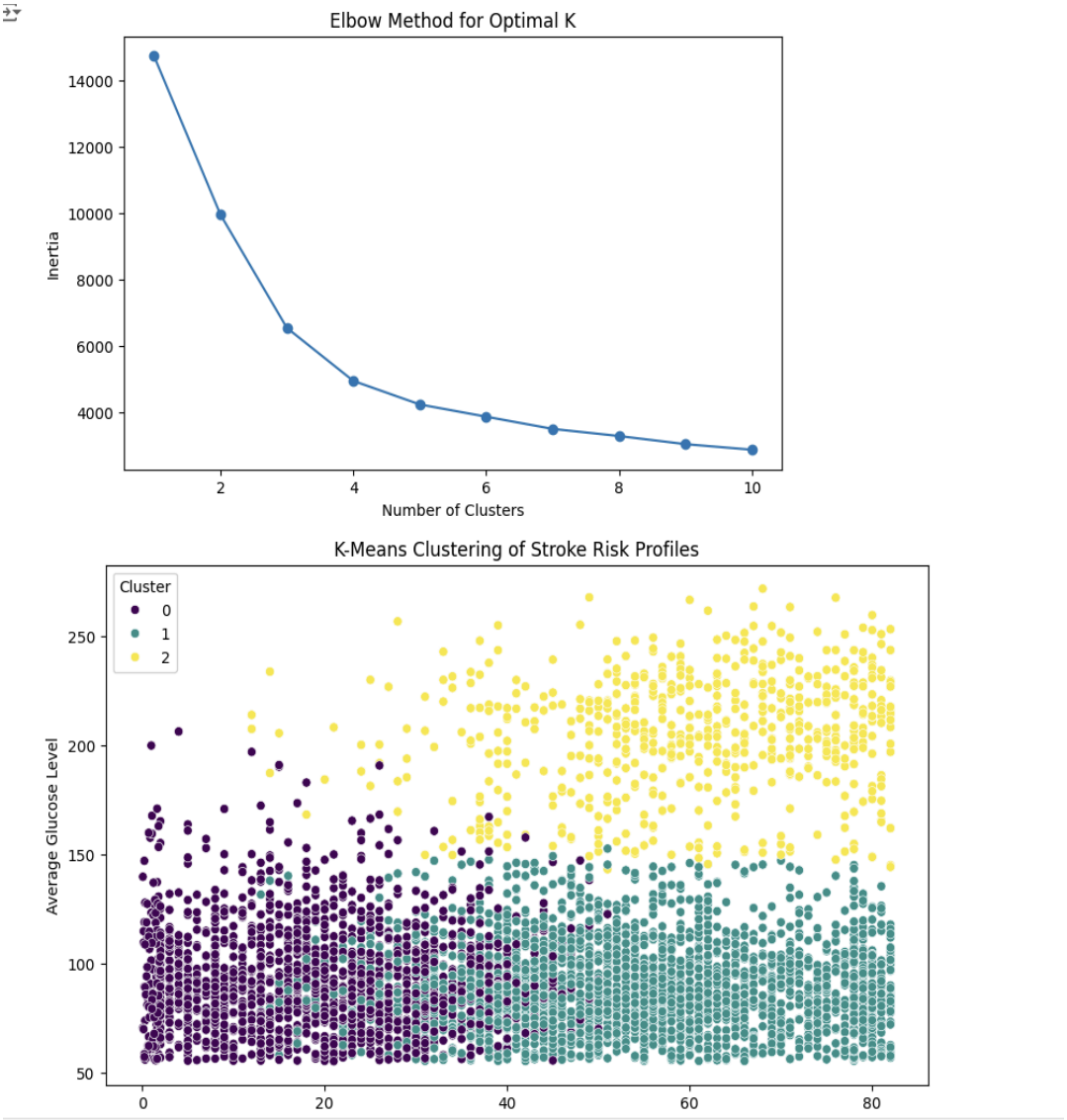
This graph displays the coefficients from the logistic regression model, which measure the relative impact of each factor on stroke risk. Positive coefficients indicate increased risk, with smoking_status_smokes showing that smokers have a higher stroke risk, followed by hypertension and heart disease. Negative coefficients suggest a reduced risk, with ever_married_Yes and work_type_Self-employed showing a negative relationship with stroke risk.

The difference in feature importance rankings between Logistic Regression and Random Forest stems from their distinct approaches. Logistic Regression assumes linear relationships and evaluates each variable's independent effect, identifying **smoking status** as the most significant factor. In contrast, Random Forest captures non-linear relationships and interactions, highlighting **average glucose level** as the top contributor. These results show that Logistic Regression is suited for linear interpretations, while Random Forest provides deeper insights into complex patterns, offering complementary perspectives on stroke risk factors.

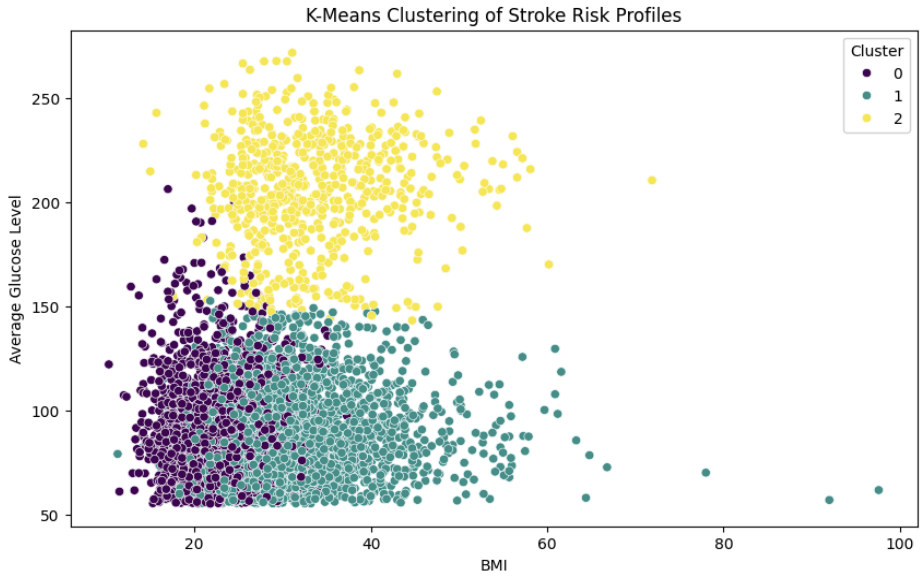
Additional Statistics

K-Means clustering was used to group individuals with similar health characteristics, focusing on **age**, **average glucose level**, and **BMI**. The goal is to identify patterns or profiles within the data that may indicate varying levels of stroke risk. This method is particularly useful for discovering hidden structures and segmenting data into meaningful clusters, enabling a deeper understanding of the relationships between these variables and their potential impact on stroke risk.

The **elbow method** was applied to determine the optimal number of clusters, which was identified as 3. This ensures that the clusters are well-defined and meaningful without overfitting or underfitting the data.



The elbow curve shows the inertia (sum of squared distances) for different numbers of clusters. The optimal number of clusters (k = 3) was selected where the curve starts to flatten, indicating a balance between cluster compactness and simplicity.



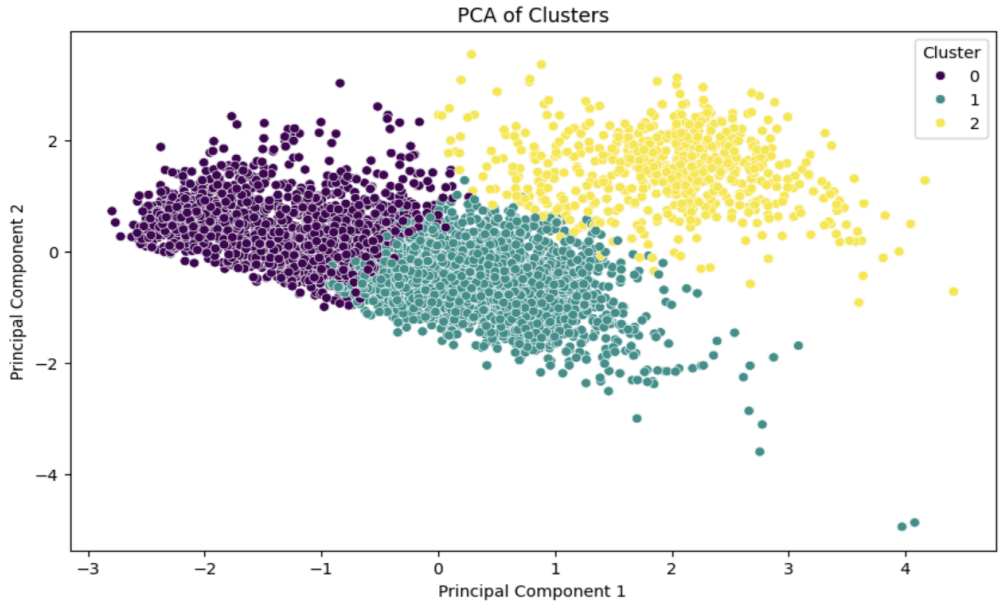
Cluster Profiles:

Cluster	age	avg_glucose_level	bmi
0	18.847053	92.719731	22.712373
1	54.021522	89.119093	31.766833
2	60.513428	205.042117	33.427172

The scatter plots reveal the distribution of individuals across three distinct clusters based on age, average glucose level, and BMI. **Cluster 0 (Purple)** includes younger individuals with the lowest average glucose levels (92.7 mg/dL), low BMI (22.7), and a mean age of 18.8 years, likely representing a low stroke risk group. **Cluster 1 (Teal)** consists of middle-aged individuals with moderate average glucose levels (89.1 mg/dL), higher BMI (31.8), and a mean age of 54.0 years, indicating a moderate risk group that may benefit from preventive measures. **Cluster 2 (Yellow)** represents older individuals with significantly elevated average glucose levels (205.0 mg/dL), higher BMI (33.4), and a mean age of 60.5 years, categorizing them as a high-risk group for stroke due to these health factors.

K-Means clustering effectively segmented the data into three distinct risk profiles. **Cluster 0** represents low-risk, younger individuals with minimal health concerns. **Cluster 1** includes moderate-risk, middle-aged individuals with elevated BMI, suggesting the need for lifestyle adjustments or preventive measures. **Cluster 2** comprises high-risk, older individuals with significantly elevated glucose levels and BMI, indicating an urgent need for targeted healthcare interventions. These findings highlight the potential for prioritizing individuals in Cluster 2 for immediate preventive strategies.

Discovering Patterns in Clustered Data



Cluster Patterns:

	age				avg_glucose_level		bmi			
	mean	std	min	max	mean	std	min	max	mean	std
Cluster										
0	18.847053	12.218730	0.08	55.0	92.719731	23.949850	55.12	206.25	22.712373	4.485052
1	54.021522	15.454623	13.00	82.0	89.119093	19.884184	55.22	152.56	31.766833	7.173703
2	60.513428	14.862350	12.00	82.0	205.042117	26.794301	143.15	271.74	33.427172	7.632394

	bmi			
	min	max	mean	std
Cluster				
0	10.3	37.3		
1	11.3	97.6		
2	14.2	71.9		

The Principal Component Analysis (PCA) and clustering were performed to simplify the high-dimensional data (age, BMI, and average glucose level) and identify distinct stroke risk profiles. By reducing the data to two principal components, visualized the clusters more clearly, allowing for a deeper understanding of how these features interact.

The One-Way ANOVA test was conducted to evaluate whether the **age** variable significantly differs between individuals with and without stroke.

- **Hypotheses**:
  - Null Hypothesis (H_0): There is no significant difference in the mean age between the stroke and non-stroke groups.
  - Alternative Hypothesis (H_1): There is a significant difference in the mean age between the stroke and non-stroke groups.

The test result ($p < 0.05$) supports rejecting the null hypothesis, indicating that age is significantly associated with stroke risk.

The **Chi-Square test** was used to examine the relationship between **gender** and stroke risk.

- **Hypotheses**:
    - Null Hypothesis (H_0): There is no relationship between gender and stroke risk (they are independent).
    - Alternative Hypothesis (H_1): There is a relationship between gender and stroke risk.

The Chi-Square test result ($p > 0.05$) failed to reject the null hypothesis, suggesting that gender does not have a significant effect on stroke risk.

T-Test:

```
T-Test for Average Glucose Level:
T-statistic: 9.830215360205345, P-value: 1.3476353968167712e-22

T-Test for BMI:
T-statistic: 2.968365485973203, P-value: 0.003008355955526417
The difference in average glucose levels between stroke and non-stroke groups is statistically significant.
The difference in BMI between stroke and non-stroke groups is statistically significant.
```

The purpose of conducting a T-Test in this project is to determine whether there is a statistically significant difference in two key health metrics—**average glucose level** and **BMI**—between individuals who experienced a stroke and those who did not. These health metrics are known to contribute to stroke risk, and understanding their differences across these groups provides valuable insights for early detection and prevention strategies.

T-Test for Average Glucose Level:

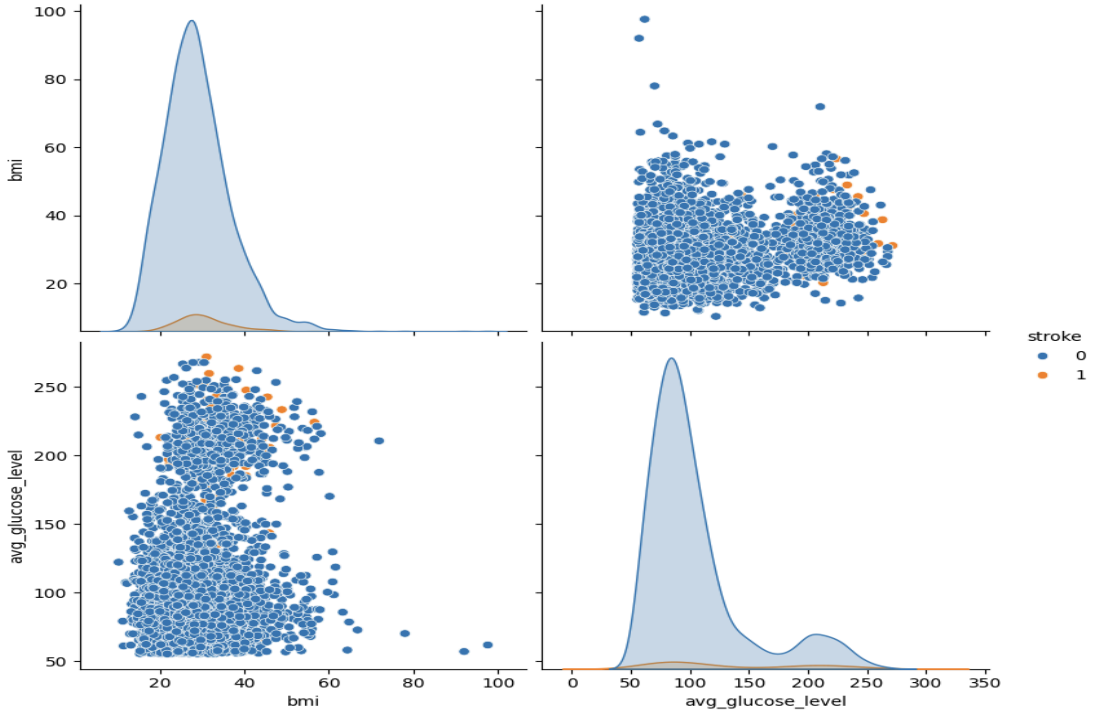
- **Hypotheses:**
- **Null Hypothesis (H_0):** There is no significant difference in average glucose levels between stroke and non-stroke groups.
- **Alternative Hypothesis (H_1):** There is a significant difference in average glucose levels between stroke and non-stroke groups.
- **T-Statistic**: 9.83, **P-Value**: 1.35e-22

Since the p-value is much smaller than 0.05, we reject the null hypothesis. This indicates that there is a statistically significant difference in the average glucose levels between stroke and non-stroke groups. Individuals who experienced a stroke tend to have higher average glucose levels, meaning that elevated glucose levels are strongly associated with stroke risk.

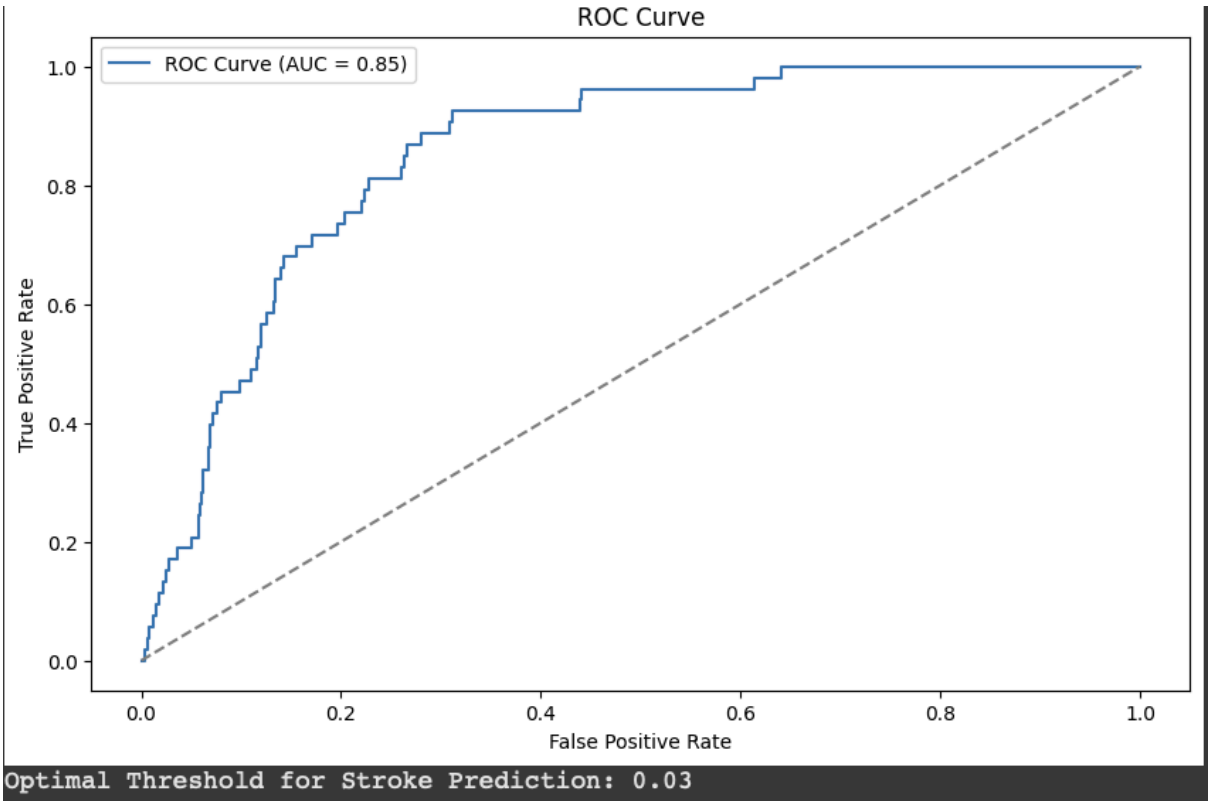
T-Test for BMI:

- **Hypothesis:**
- **Null Hypothesis (H_0):** There is no significant difference in BMI between stroke and non-stroke groups.
- **Alternative Hypothesis (H_1):** There is a significant difference in BMI between stroke and non-stroke groups.
- **T-Statistic**: 2.97, **P-Value**: 0.003

Similarly, the p-value for the BMI T-Test is less than 0.05, leading to the rejection of the null hypothesis. This shows that BMI values are significantly different between stroke and non-stroke groups. Specifically, individuals in the stroke group generally have higher BMI values, which aligns with the understanding that obesity is a contributing factor to stroke risk.



The pairplot highlights the relationship between **BMI**, **average glucose level**, and stroke risk. Individuals with stroke tend to have higher BMI and glucose levels, indicating a positive association between these health metrics and stroke risk. The histograms (diagonal plots) show that stroke cases are concentrated within a BMI range of 30-40, suggesting that obesity may contribute to stroke risk. Similarly, higher glucose levels, particularly above 150 mg/dL, are more prevalent among individuals with stroke. Although the scatterplots (off-diagonal plots) do not show a clear linear relationship between BMI and glucose level, the clustering of stroke cases in regions of BMI(30-40) and high glucose levels suggests that these factors together may significantly increase stroke risk.



The ROC curve analysis evaluates the model's ability to distinguish between individuals with and without stroke, yielding a high AUC value of **0.85**, indicating strong predictive performance. The optimal threshold for stroke prediction was determined to be 0.03, representing the point at which the model achieves the best balance between true positive and false positive rates, optimizing its classification performance.

Part 5: Results

This project provided significant insights into stroke risk factors and their impact through comprehensive data analysis and machine learning techniques.A range of statistical and machine learning methods were employed to address the key research questions.The selection of methods was driven by the need for both interpretability and robustness in our analysis.

To address the question, **"Which factors have the greatest impact on stroke risk?"** Random Forest and logistic regression were utilized to identify the most critical predictors, leveraging Random Forest's ability to capture complex, nonlinear relationships and logistic regression's strength in providing clear, interpretable insights into individual variable contributions.

To answer the question, **"What are the differences in stroke risk and contributing factors across demographic groups (age, gender, residence type, work type)?"** techniques such as ANOVA, Chi-Square tests, and T-Tests were applied, ensuring a rigorous statistical evaluation of differences across groups. Additionally, K-Means clustering and PCA were integrated to segment the data into meaningful risk profiles, uncovering patterns that inform targeted healthcare strategies. These methods were essential for providing a comprehensive understanding of the interplay between demographic factors, health metrics, and stroke risk, ultimately supporting the development of actionable insights for prevention and intervention.

The most critical predictors of stroke were identified as average glucose level, BMI, and age, with hypertension and heart disease playing moderate roles. These findings underscore the importance of controlling glucose levels and maintaining a healthy BMI as part of stroke prevention strategies.

To explore the question, **"How do health metrics (BMI, glucose levels) correlate with stroke risk, and are there thresholds for increased risk?"** demographic analyses revealed no significant relationship between gender and stroke risk, while age emerged as a crucial factor, with older individuals facing a higher risk. Work type and marital status had minimal influence on stroke likelihood. The T-Test results further highlighted that individuals with stroke tend to have significantly higher glucose levels and BMI, reinforcing these metrics as key indicators for early detection and prevention.

Machine learning models, including Random Forest and Logistic Regression, provided complementary insights. Random Forest captured complex, non-linear interactions between variables, while Logistic Regression clarified individual variable contributions, such as the heightened risk associated with smoking. The ROC analysis demonstrated strong model performance with an AUC of 0.85, identifying a threshold for stroke prediction, enhancing classification accuracy.

Additionally, K-Means clustering and PCA helped group individuals into low, moderate, and high-risk profiles, providing actionable insights for targeted healthcare interventions. The high-risk group (Cluster 2), characterized by significantly elevated glucose levels and BMI, demands immediate attention and preventive measures.

In conclusion, this project combined statistical and machine learning methods to uncover meaningful patterns and relationships in stroke risk factors, offering valuable guidance for healthcare strategies and early intervention policies. These results emphasize the need for regular monitoring of health metrics and personalized preventive strategies to mitigate stroke risks effectively.