



# Predicting Stroke Risk: A Machine Learning Approach.

W  
E  
L  
C  
O  
M  
E

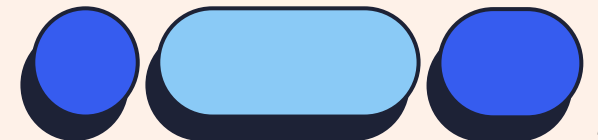
By Cindy Seth



# Stroke

## Problem Statement

*Stroke is a serious medical condition that can have devastating consequences. In the United States, stroke ranks as the fifth leading cause of death and a primary source of long-term disability. According to the American Brain Foundation, every year, over 800,000 individuals experience a new or recurrent stroke, and unfortunately, at least 140,000 succumb to this disease. Can we develop a highly accurate predictive model for stroke occurrence, utilizing a comprehensive set of patient data and advanced machine learning techniques?*





# 01

## Data Overview

For this project, I utilized the '[healthcare-dataset-stroke-data.csv](#)' dataset from Kaggle. The dataset comprises twelve columns, several of which contained missing data.

- id
- gender
- age
- hypertension
- heart\_disease
- ever\_married
- work\_type
- Residence\_type
- avg\_glucose\_level
- bmi
- smoking\_status
- stroke



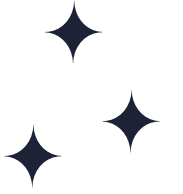


# Data Wrangling

D  
A  
T  
A

To clean up the data, I removed the **'id'** column, as it served solely as a unique identifier for each patient and was not relevant to the analysis. Additionally, I eliminated any instances with missing values to ensure data integrity and consistency.

# Problem / Solution



## Problems Encountered:

- Age column had a few random values
- BMI column had missing values
- Work type, residence type, and smoking status had multiple entries

## Problem Solutions:

- Applied `.astype(float)` to ensure all values under 'age' column matched
- Iterated values under 'gender' and 'ever\_married' columns
- Applied `get_dummies` for the columns that had multiple values to create new binary columns



# 02

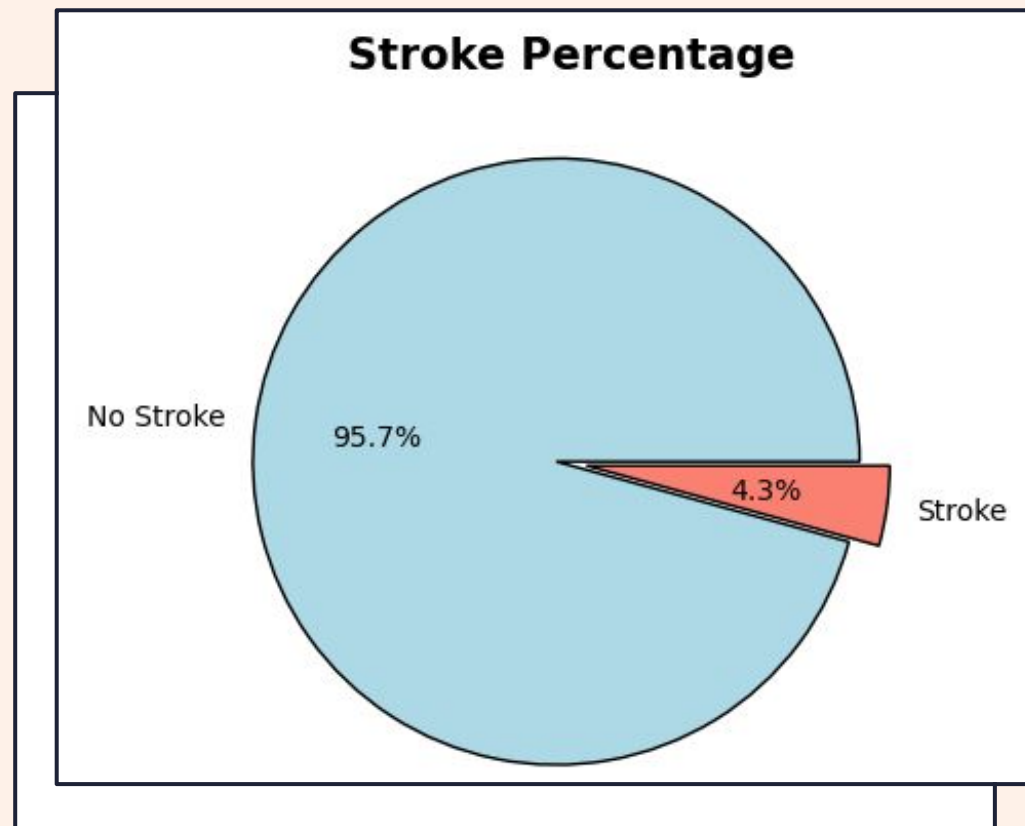
# Exploratory Data Analysis

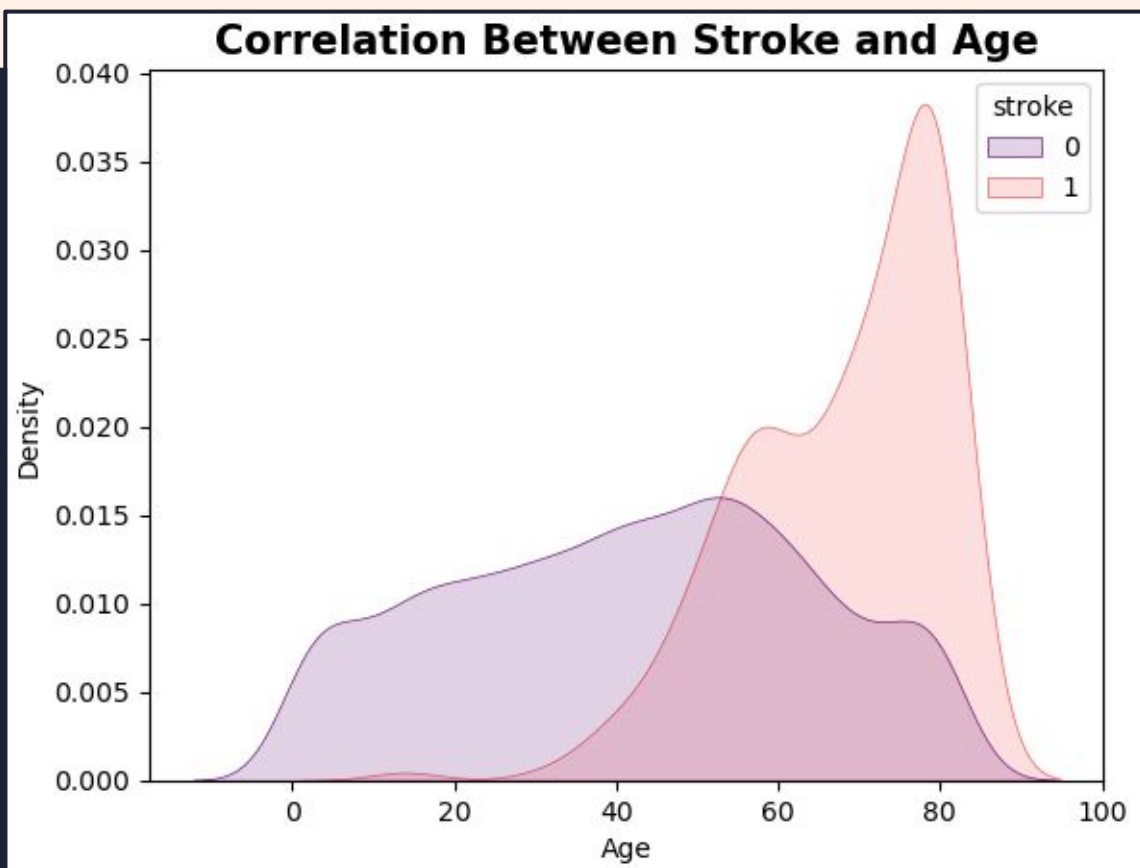
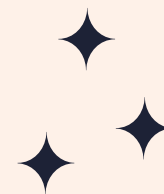
In this phase, we delved into the data to uncover hidden connections between variables and stroke.



# EDA

Analysis of the  
'**healthcare-dataset-stroke-data.csv**  
' dataset reveals a stroke  
prevalence of 4.3%, whereas a vast  
majority (95.7%) of patients did  
not suffer from stroke. After  
cleaning the data, we focused on  
209 patients who had experienced a  
stroke.





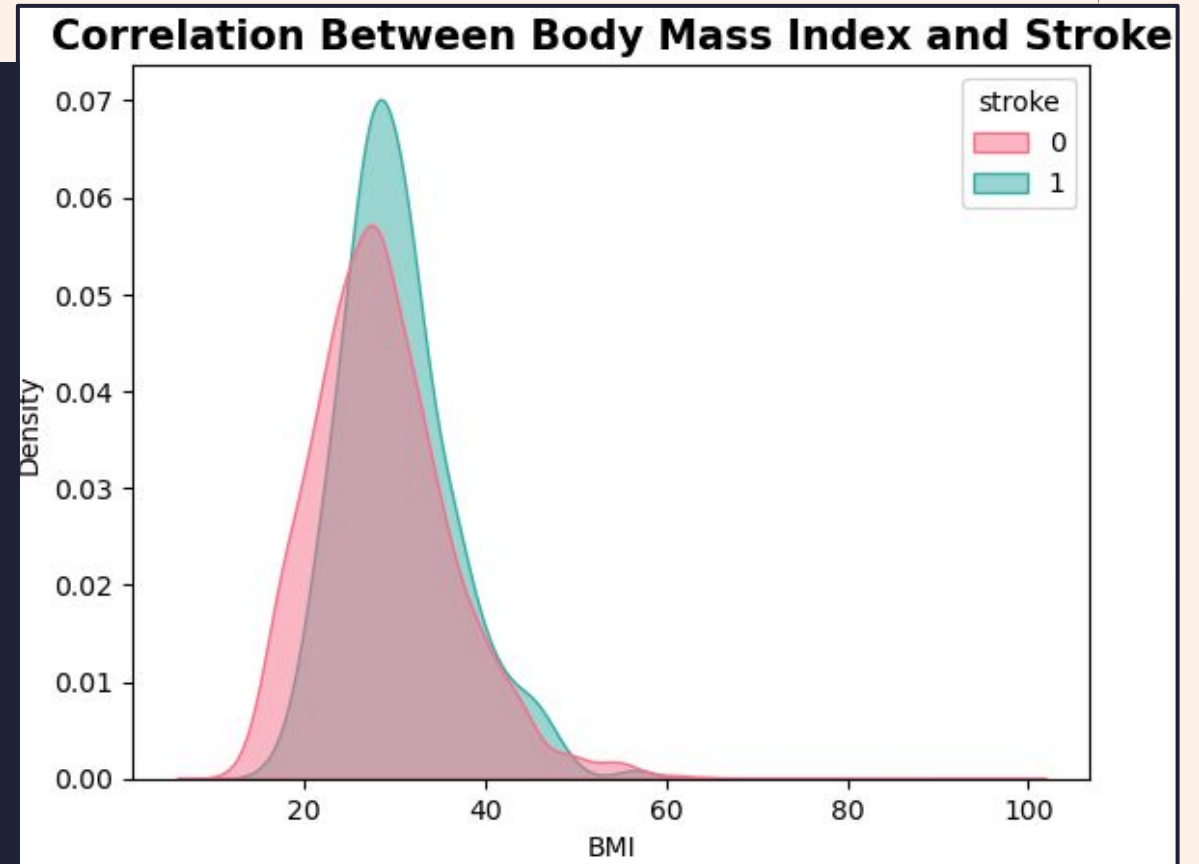
# Correlation Between **Stroke** and Age.

The risk of experiencing a stroke increases as people age. The majority of strokes occur in individuals aged 60 and older, although younger adults can also experience strokes.

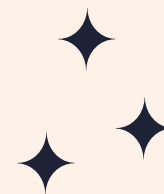




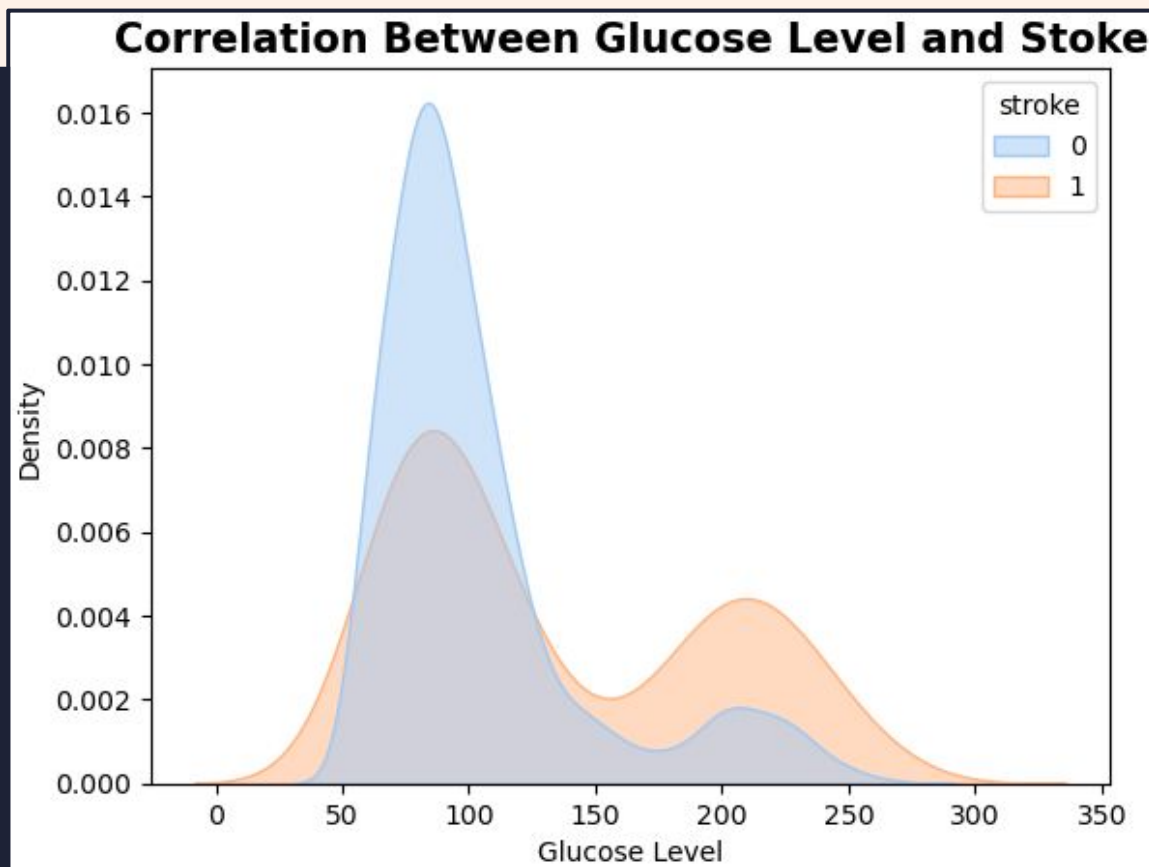
# Correlation Between BMI and Stroke.



The correlation between BMI and stroke indicates that a high BMI does not always necessarily mean the person will experience a stroke. This tells us that there may be other underlying factors that influence stroke risk.



# Correlation Between Glucose Level and Stroke.

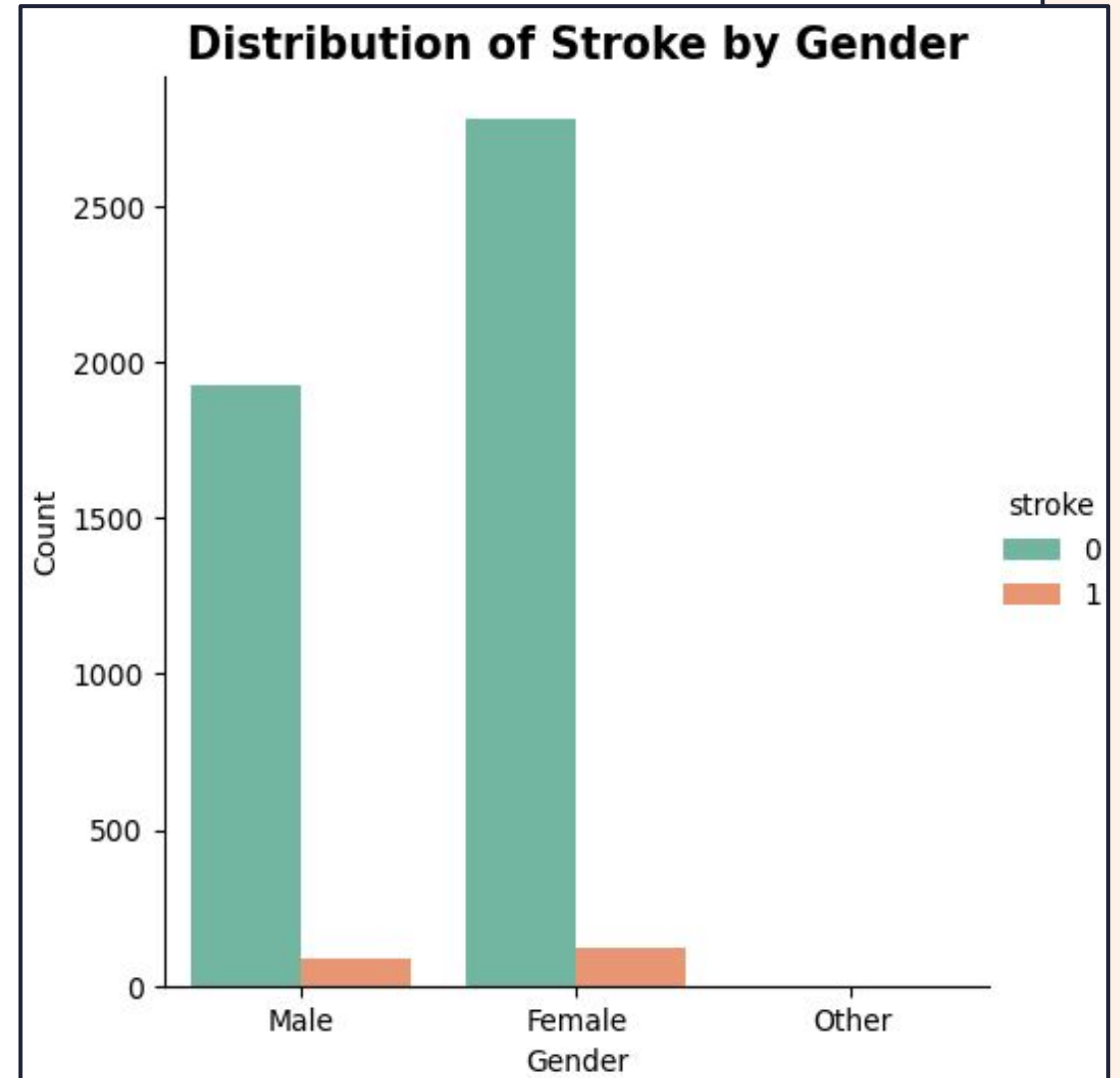


The correlation between glucose level and stroke shows that an elevated blood glucose level can significantly increase the risk of stroke.



# Distribution of **Stroke** by Gender.

Analyzing the distribution of stroke risk by gender revealed only a slight difference between the two groups.



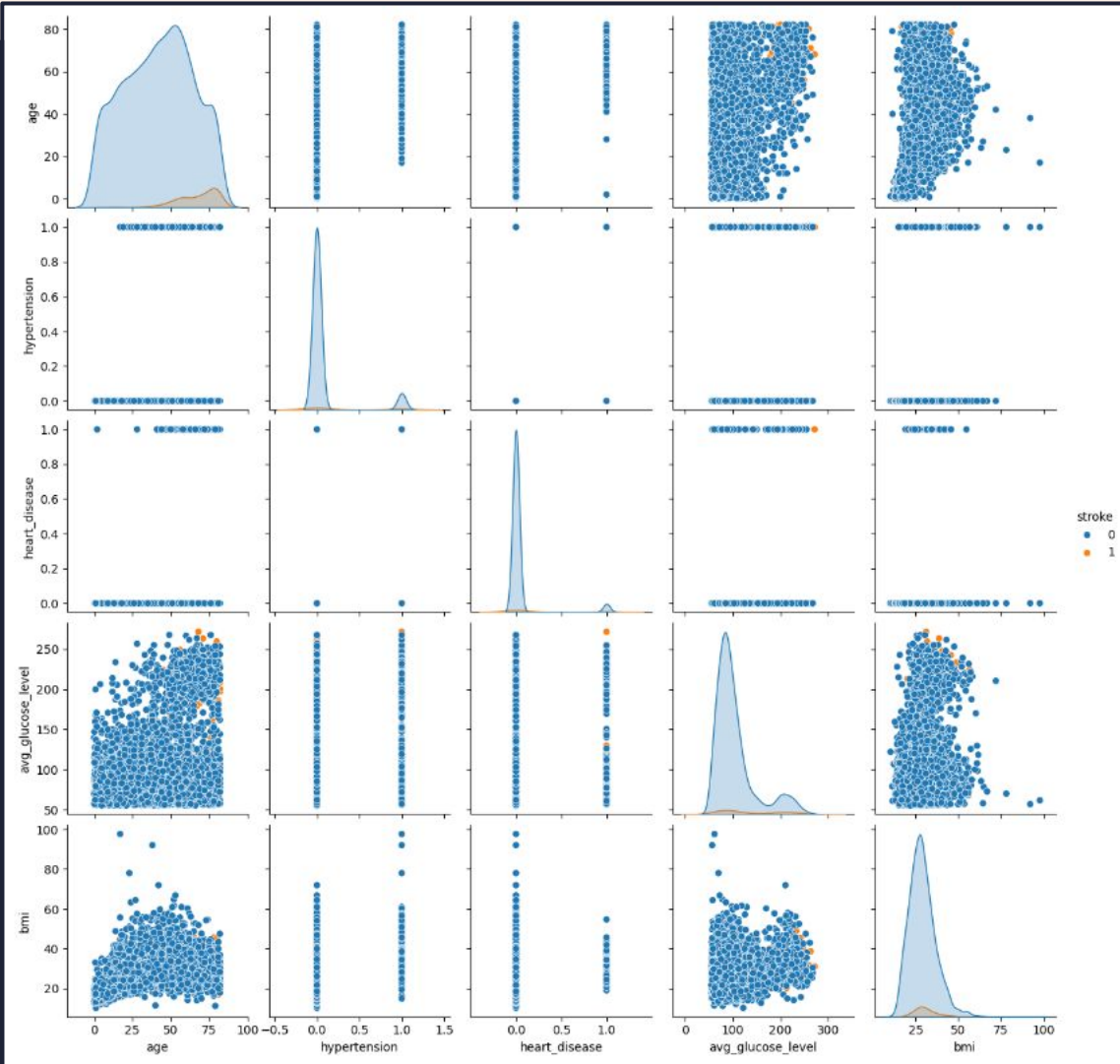


# Correlations Correlations Correlations

To find correlations among the dataset, we used a kernel density estimate (KDE) plot and a categorical plot. By doing so, it helps us visualize the given data.

This analysis revealed several key findings:

- The distribution of age is skewed to the right, indicating a higher concentration of individuals in the older age groups. There seems to be a positive correlation between age and stroke, as the scatterplot shows a general upward trend.
- The scatterplots between hypertension, heart disease, and stroke suggest potential correlations, especially between hypertension and stroke.
- There seems to be a positive correlation between average glucose level and stroke, however, it is less pronounced than the relationship between age and stroke.
- BMI and stroke shows a relationship that is less definitive with a more scattered distribution.





# Correlations

# Correlations

# Correlations

To ensure the accuracy of the data after analyzing the pair plot, a series of chi-squared tests was used to analyze the association between several variables (hypertension, heart\_disease, ever\_married, avg\_glucose\_level, bmi, and age) along with the 'stroke' variable was used.

- Based on this chi-squared test, hypertension, heart\_disease, ever\_married, and age are significantly associated with stroke by the chi-squared statistics being relatively high and the p-values being very low.
- The chi-squared statistics shows a weak correlation between BMI and stroke.

```
hypertension: Chi-squared = 97.2749949311716, p-value = 6.033751208728256e-23
heart_disease: Chi-squared = 90.2795595563918, p-value = 2.0677783295228626e-21
ever_married: Chi-squared = 53.12593819801626, p-value = 3.1283412849388787e-13
avg_glucose_level_bin: Chi-squared = 28.172755674738273, p-value = 3.341074728932425e-06
bmi_bin: Chi-squared = 28.672231211518906, p-value = 2.6242109793248913e-06
age_bin: Chi-squared = 329.1787268098242, p-value = 4.537309548319127e-68
```


```
hypertension: Chi-squared = 97.2749949311716, p-value = 6.033751208728256e-23, Adjusted p-value = 3.6202507252369534e-22, Significant: True
heart_disease: Chi-squared = 90.2795595563918, p-value = 2.0677783295228626e-21, Adjusted p-value = 1.2406669977137176e-20, Significant: True
ever_married: Chi-squared = 53.12593819801626, p-value = 3.1283412849388787e-13, Adjusted p-value = 1.8770047709633272e-12, Significant: True
avg_glucose_level_bin: Chi-squared = 28.172755674738273, p-value = 3.341074728932425e-06, Adjusted p-value = 2.004644837359455e-05, Significant: True
bmi_bin: Chi-squared = 28.672231211518906, p-value = 2.6242109793248913e-06, Adjusted p-value = 1.574526587594935e-05, Significant: True
age_bin: Chi-squared = 329.1787268098242, p-value = 4.537309548319127e-68, Adjusted p-value = 2.7223857289914758e-67, Significant: True
```



# 03

# Preprocessing and Training Data

To develop a robust machine learning model, we need to preprocess the data to ensure its quality and consistency. Then, by training the model on this prepared data, we can equip it to learn meaningful patterns and make accurate stroke predictions that generalize well to new, unseen data.





# Preprocessing and Training Data

## Columns Used:

- 'hypertension'
- 'heart\_disease'
- 'ever\_married'
- 'avg\_glucose\_level\_bin'
- 'bmi\_bin'
- 'age\_bin'

```
X_train shape: (3927, 21)
X_test shape: (982, 21)
y_train shape: (3927,)
y_test shape: (982,)
```

## Preprocessing Results:

The output shapes confirm that the dataset has been split correctly and is well-prepared for training a machine learning model.

- The training set has **3927** samples which will be beneficial for the model to learn the underlying patterns in the dataset.
- The testing set has **982** samples has a reasonable size to evaluate meaningful performance metrics





# 04

## Modeling

To determine which models to use to accurately predict the risk of stroke, I chose to test four types of models:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier





Class distribution before resampling:

```
stroke
0    4700
1     209
Name: count, dtype: int64
```

```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

      0       0.95       0.98       0.97       929
      1       0.21       0.08       0.11        53

   accuracy          0.93       982
  macro avg       0.58       0.53       0.54       982
 weighted avg       0.91       0.93       0.92       982
```

```
Decision Tree Classification Report:
              precision    recall  f1-score   support

      0       0.95       0.94       0.95       929
      1       0.13       0.15       0.14        53

   accuracy          0.90       982
  macro avg       0.54       0.55       0.54       982
 weighted avg       0.91       0.90       0.90       982
```

```
Random Forest Classification Report:
              precision    recall  f1-score   support

      0       0.95       0.99       0.97       929
      1       0.00       0.00       0.00        53

   accuracy          0.94       982
  macro avg       0.47       0.50       0.48       982
 weighted avg       0.89       0.94       0.92       982
```

```
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

      0       0.95       0.98       0.97       929
      1       0.19       0.08       0.11        53

   accuracy          0.93       982
  macro avg       0.57       0.53       0.54       982
 weighted avg       0.91       0.93       0.92       982
```

# Results for Each Model

While Logistic Regression exhibited the highest overall accuracy among the models evaluated, it encountered difficulties in accurately predicting certain classes, as evidenced by the '*UndefinedMetricWarning*' error. This suggests that the model might be struggling to capture the nuances of these specific classes, potentially due to factors such as class imbalance or insufficient data. Further investigation and potential adjustments to the model or data preprocessing was needed to address this issue and improve performance for all classes. Due to the persistent error, I applied Synthetic Minority Over-sampling Technique, or SMOTE.



# Conclusion: Best Model?

## Deep Dive Each Model

- **Random Forest** and **Gradient Boosting** have the highest accuracy, indicating that they perform well overall.
- **Gradient Boosting** has a precision of 1.000000 which indicates every positive prediction it makes will be correct. *However*, its recall is ZERO! Meaning that it does not identify any positive cases.
- **Logistic Regression** has the highest recall, meaning that it identifies a large portion of actual positive cases.
- **Logistic Regression** also has the best F1 score which indicates a better balance between precision and recall.
- **Logistic Regression** ALSO has a decent ROC AUC score which indicates a good model discrimination ability.

Therefore, to reduce the amount of false negatives and capture the best prediction as possible, **Logistic Regression** is the better choice.

