# Stroke Analysis: A Data-Driven Exploration

Group 9 - UIT

Data Mining

December 22, 2024

# Introduction

- Stroke is one of the leading causes of death and disability worldwide, emphasizing the need for early prediction and prevention strategies.
- This project aims to analyze stroke risk factors and evaluate predictive modeling techniques to understand the key drivers of stroke occurrences.
- Objectives:
  - Investigate significant predictors of stroke through detailed data analysis.
  - Build and evaluate machine learning models for robust prediction, validating findings from data analysis.
  - Provide **actionable insights** for healthcare interventions to reduce stroke risk.

# Dataset Overview

- The analysis is based on a comprehensive dataset covering various factors associated with stroke risk.
- The dataset includes three main categories of features:
  - **Demographic Information:** Age, gender, marital status, residence type, and work type.
  - **Medical History:** Prevalence of hypertension and heart disease.
  - **Lifestyle Factors:** Smoking status, average glucose level, and BMI.
- **Target Variable:** Stroke occurrence (Yes/No), enabling binary classification.
- This dataset provides a solid foundation for understanding relationships between predictors and stroke risks.

# Research Process

1. **Data Storytelling:**
   - Formulated research questions and hypotheses to guide exploration.
   - Framed the analysis and decision-making process.
2. **Exploratory Data Analysis (EDA):**
   - Uncovered patterns and relationships using statistical methods (e.g., Chi-squared tests).
   - Visualized correlations to derive *qualitative insights* and validate with *quantitative methods*.
3. **Model Training and Evaluation:**
   - Trained ML models to validate EDA findings.
   - **Factor Evaluation:** Assessed the importance of variables by sequentially removing features.
   - **Probability Validation:** Verified trends using predicted probabilities.

*This structured approach ensured robust analysis, insightful conclusions, and actionable recommendations.*

# Data Storytelling: Assumptions

- **Key Questions to Explore:**
  - Does age significantly impact stroke occurrence?
  - Are BMI and glucose levels strong predictors of stroke risk?
  - Does smoking directly contribute to strokes?
  - Do heart disease and high blood pressure elevate stroke risk?
  - Are males more affected by stroke due to stress-related workloads?
  - Does being single reduce stroke rates?
  - Are people living in rural or urban areas more likely to have strokes?
- **Next Steps:**
  - Validate assumptions using EDA and machine learning models.
  - Identify the main causes and habits that increase stroke risk.
  - Recommend actionable solutions to reduce and prevent the risk of stroke.

# Data Summary

Table: Summary of Key Features

| Feature | Details |
|---|---|
| **Numerical** | |
| Age | Mean: 41.42, Range: 0.08–82 years |
| Avg. Glucose Level | Mean: 89.04, Range: 55.22–267.6 mg/dL |
| BMI | Mean: 28.11, Range: 10.3–80.1 |

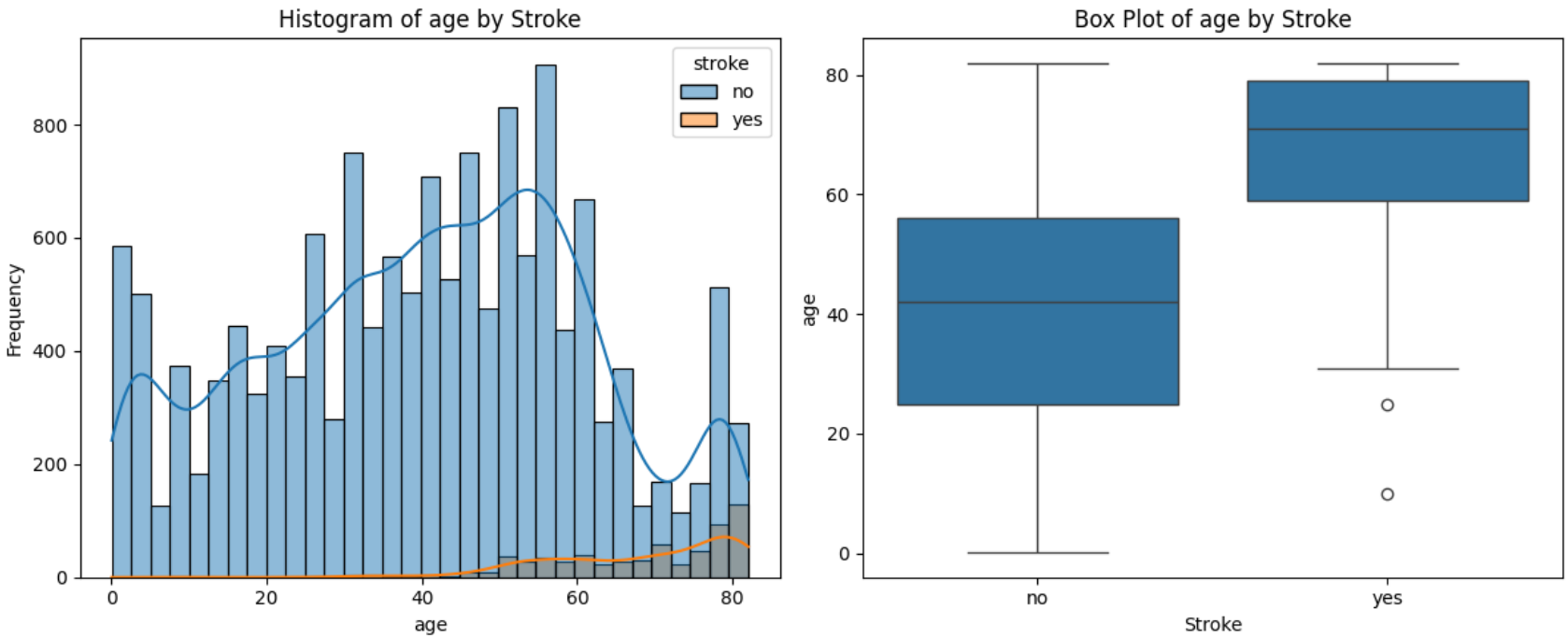
# Data Summary

Table: Summary of Key Features

| Feature | Details |
|---|---|
| **Categorical** | |
| Gender | Female, Male |
| Hypertension | No, Yes |
| Heart Disease | No, Yes |
| Ever Married | No, Yes |
| Work Type | Private, Self-employed, Govt Job Children, Never-worked |
| Residence Type | Rural, Urban |
| Smoking Status | Never, Formerly, Smokes, Unknown |
| **Target Variable** | |
| Stroke | No: 14671 (96%), Yes: 632 (4%) |

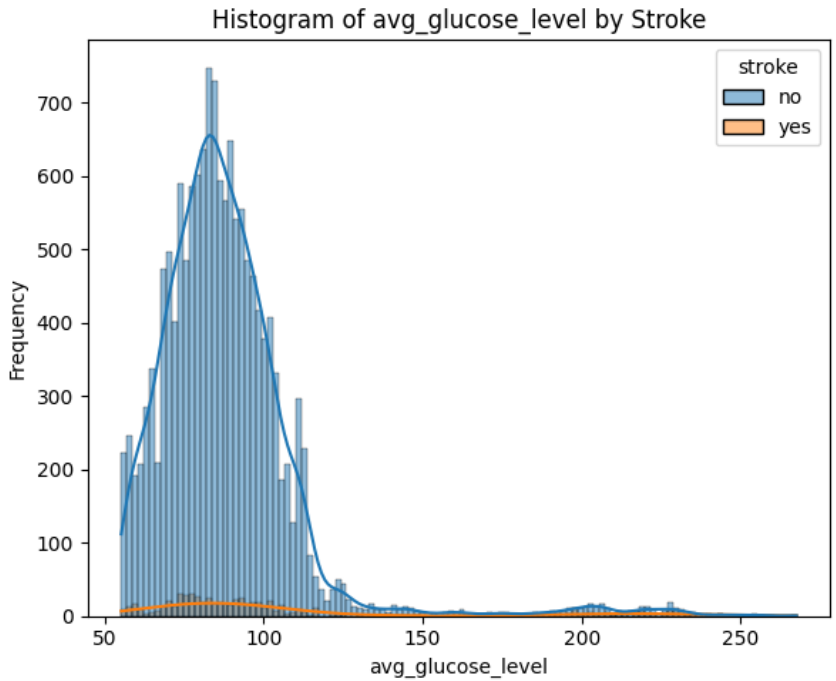
# Numerical Feature Analysis

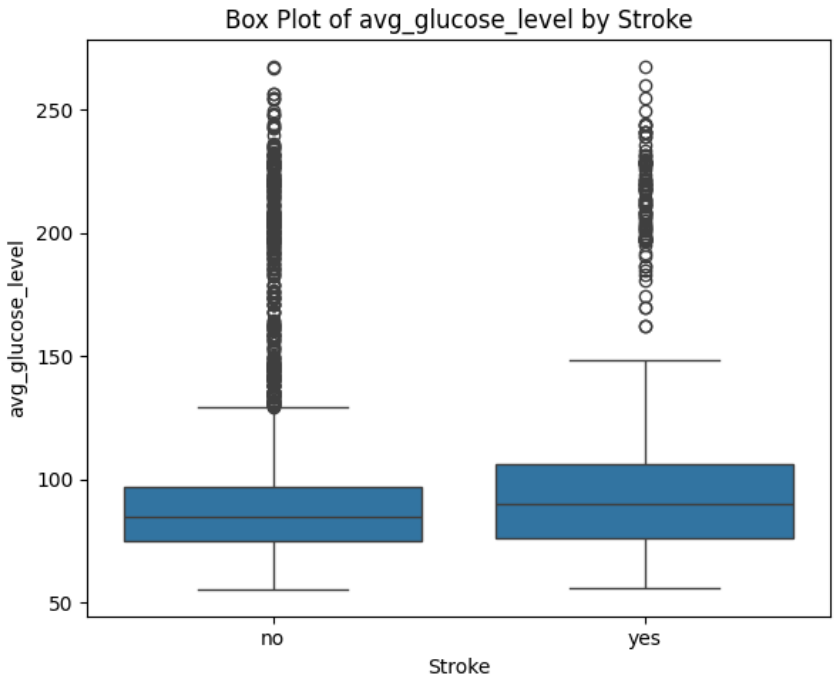
# Age Distribution and Stroke Impact




- **Key Insights:**
  - Stroke risk increases for individuals **above 50 years**.
  - Age for stroke-positive cases skews **toward higher groups**.

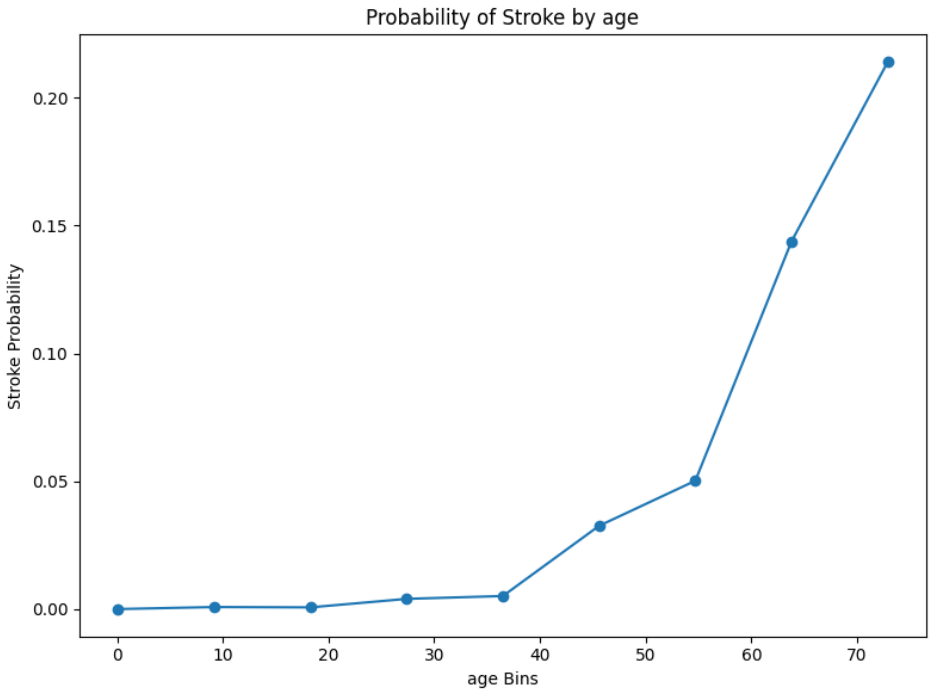
# Avg Glucose Level and Stroke Impact

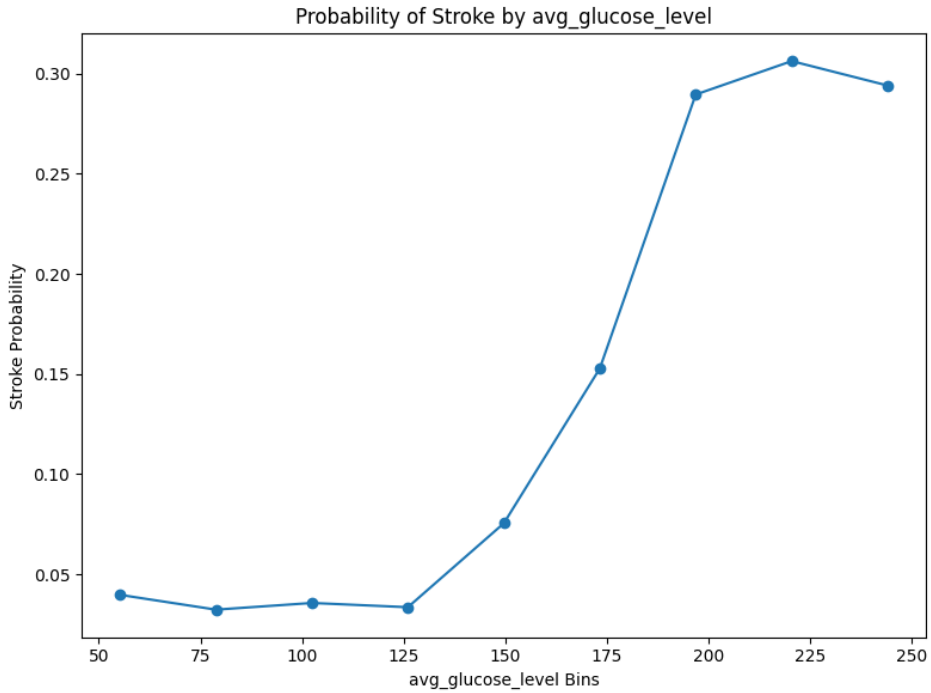



- **Key Insights:**
  - Glucose levels among stroke patients exhibit slightly higher variance and more outliers.
  - Stroke-positive cases tend to show elevated glucose levels overall.

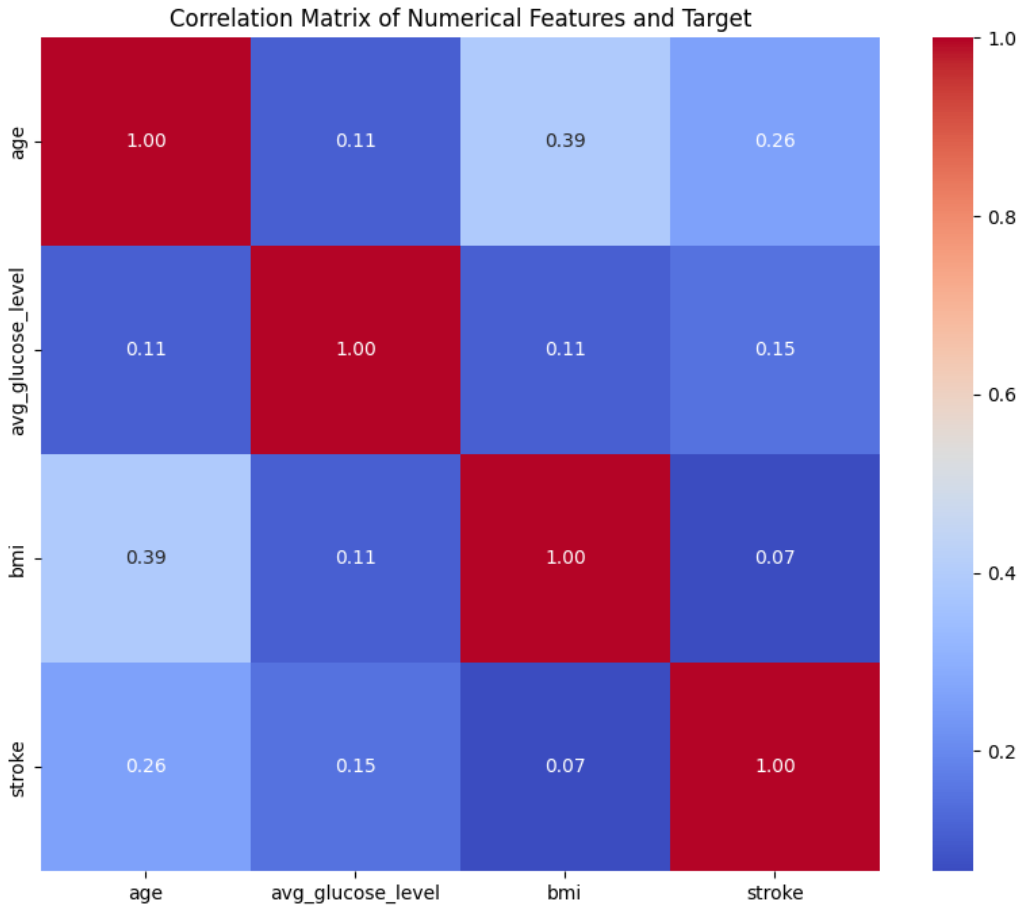
# Stroke Probability by Age and Glucose



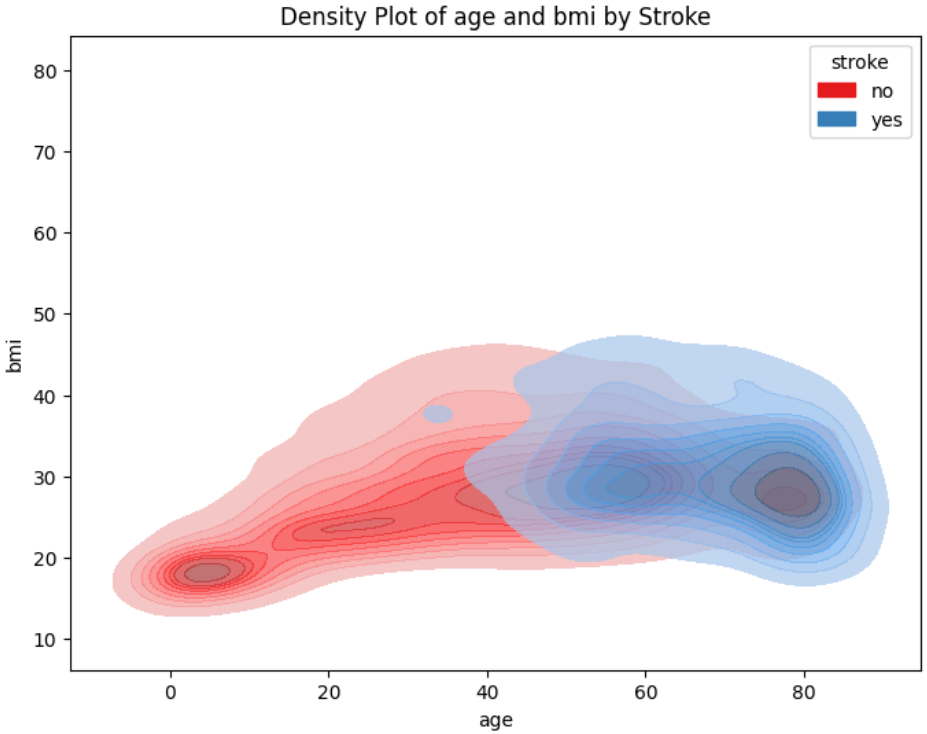

- **Conclusion:** The higher the age and glucose level, the more risk of stroke trends.

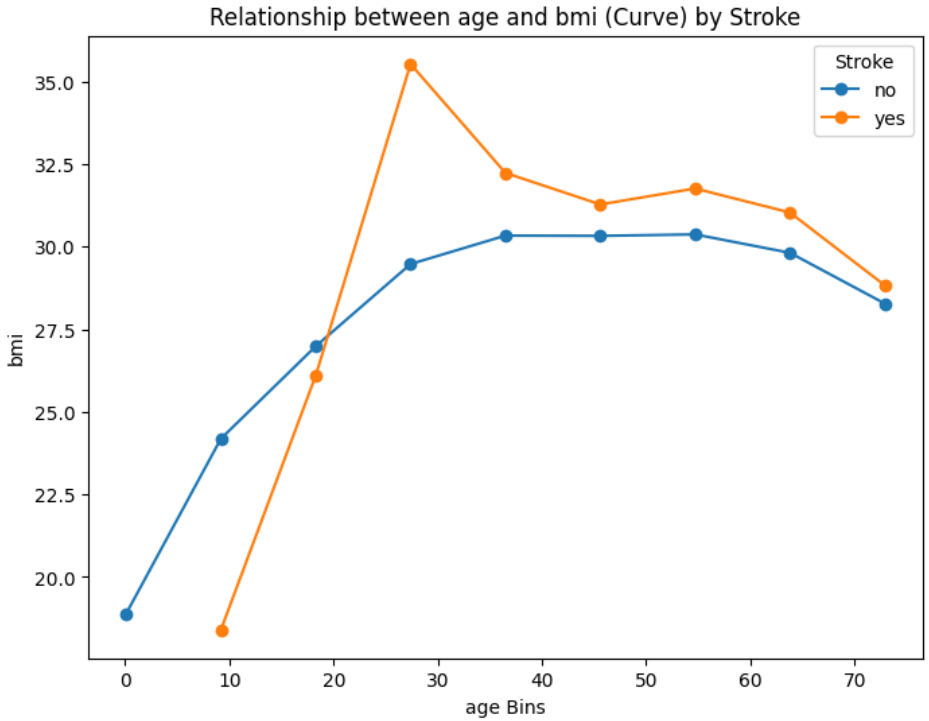
# Correlation Matrix of Numerical Features and Target




- **Key Insights:** There is a high correlation between age and BMI (0.39), so plot this relationship to analyze more patterns.

# Age vs. BMI Relationship to Stroke Impact

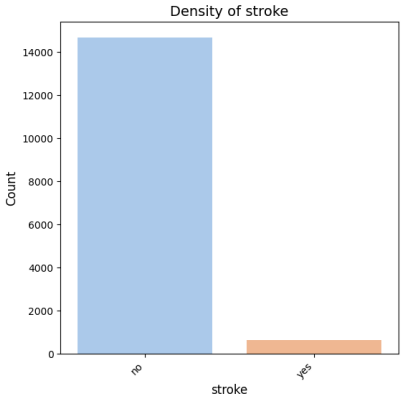


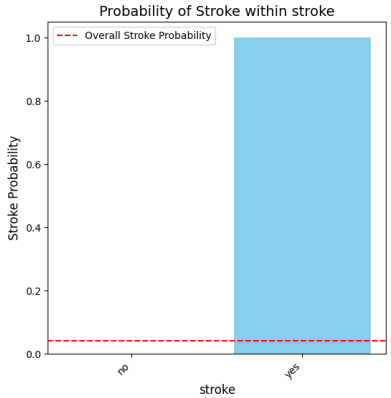



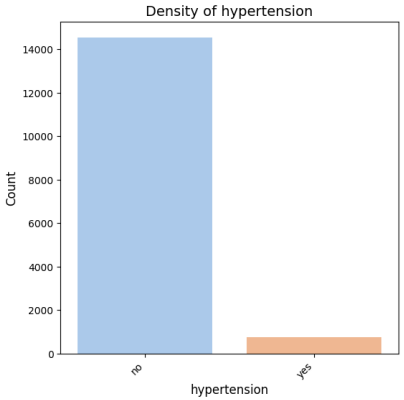

- **Conclusion:** At the same age, people with higher BMI are more likely to have a stroke.

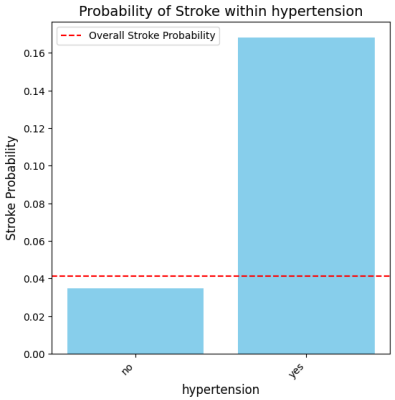
# Categorical Feature Analysis

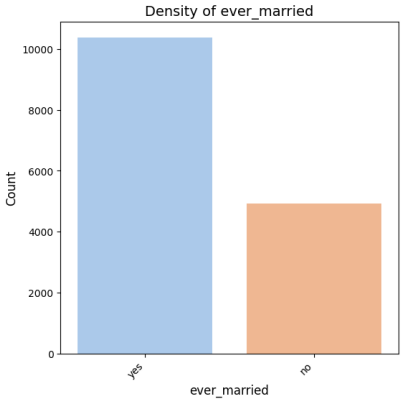
# Stroke Analysis Visualization

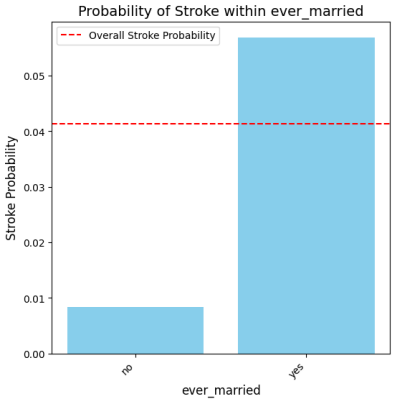


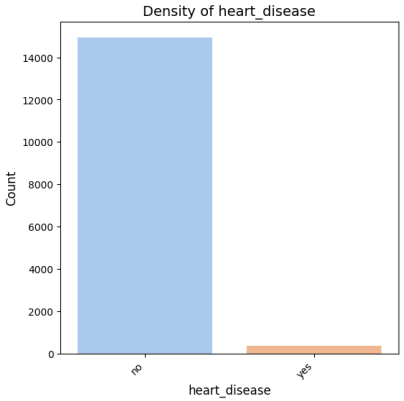



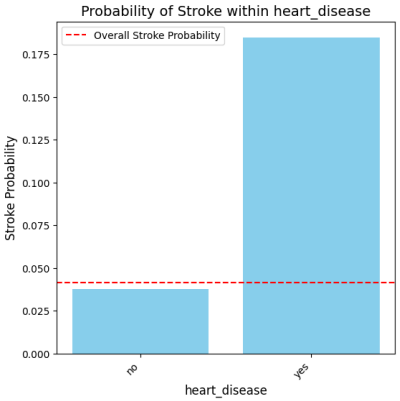



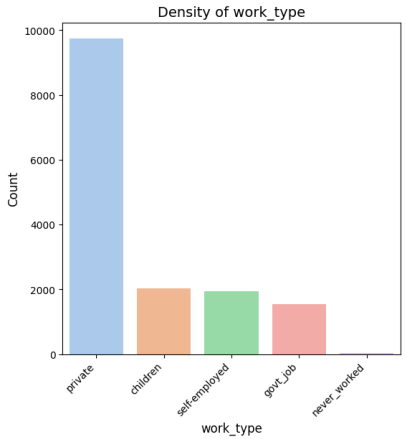



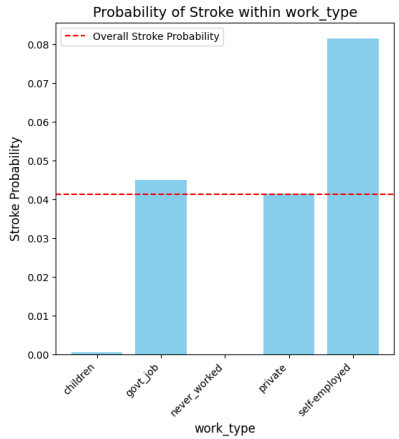



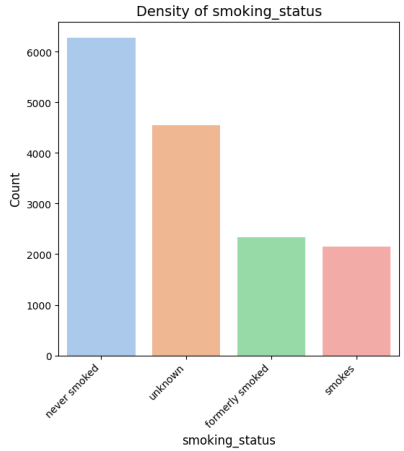



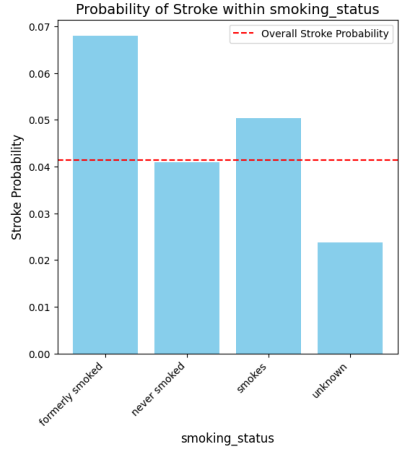



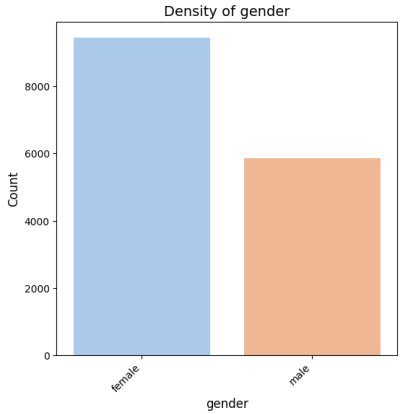
# Stroke Analysis Visualization

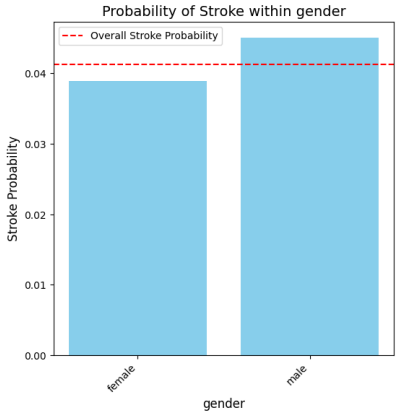


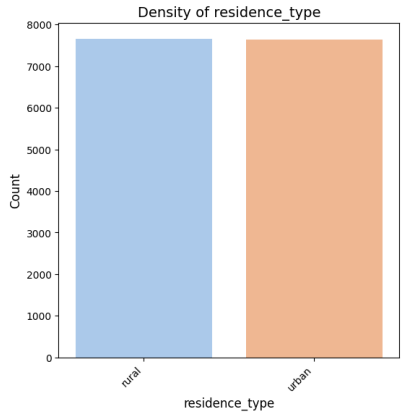



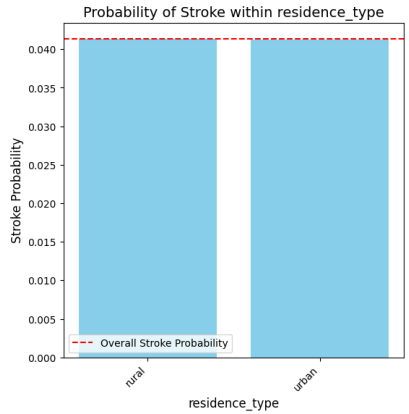



# Conclusions on Stroke Risk Factors

- **Conclusions:**
  - Patients with a history of **hypertension** or **heart disease** face a stroke risk **4-6 times higher** compared to healthy individuals.
  - Individuals who **currently smoke** or have **smoked in the past** are at a higher risk of stroke compared to **non-smokers**.
  - **Men** generally show a greater risk of stroke compared to **women**.
  - **Residence type** has minimal impact on stroke occurrence.
  - Being **unmarried** appears to reduce the likelihood of stroke. *[Why?]*
  - **Children** and individuals who have never been employed rarely experience strokes.
  - The **self-employed** group shows a higher stroke rate compared to those in **private** and **government** sectors. *[Verify this?]*

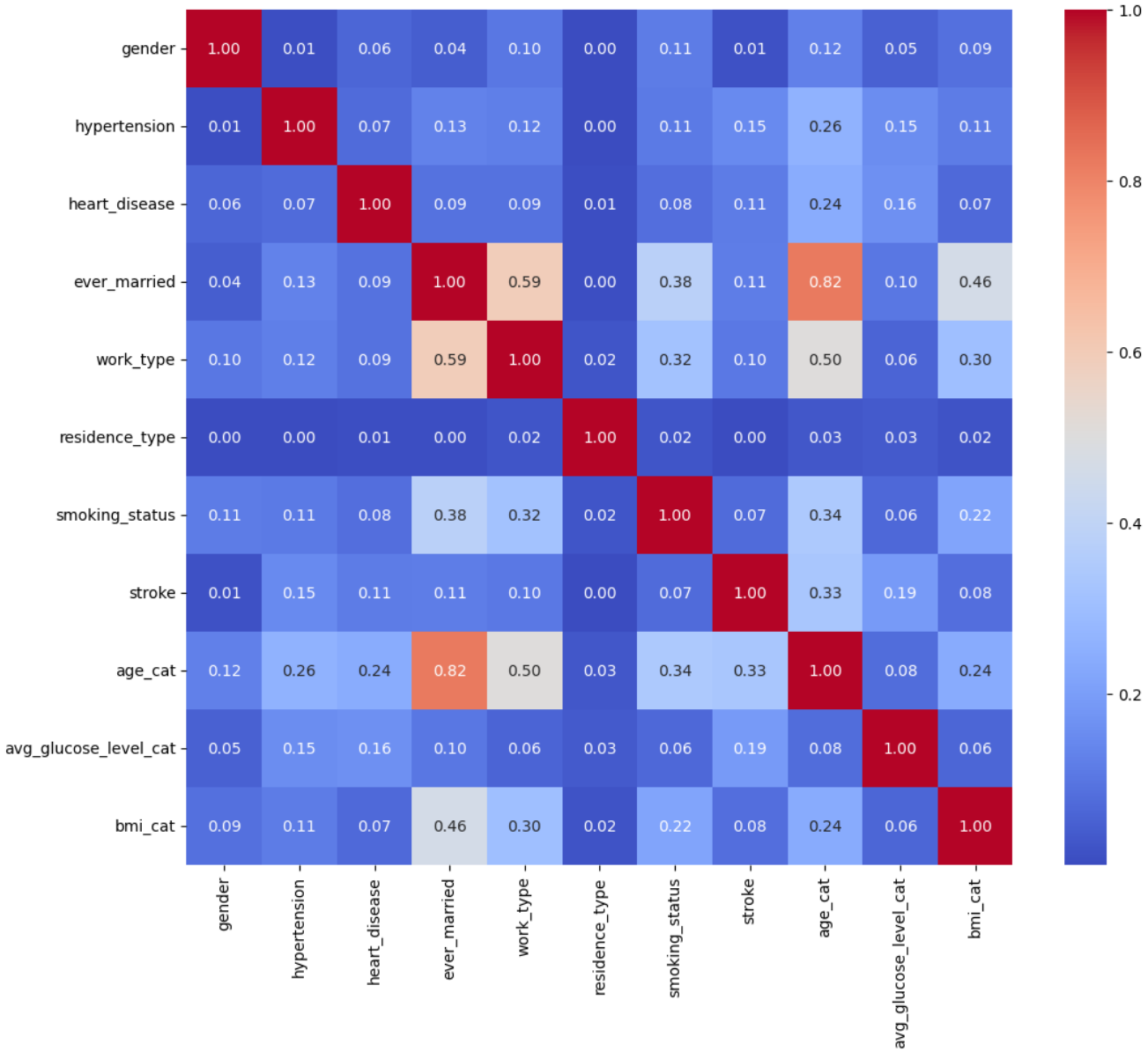
# Verification of Conclusions by Cramér's V

- **Categorizing**: Binning transforms Numerical features into categorical features.
- **Cramér's V**: Cramér's V is computed based on the chi-squared statistic:

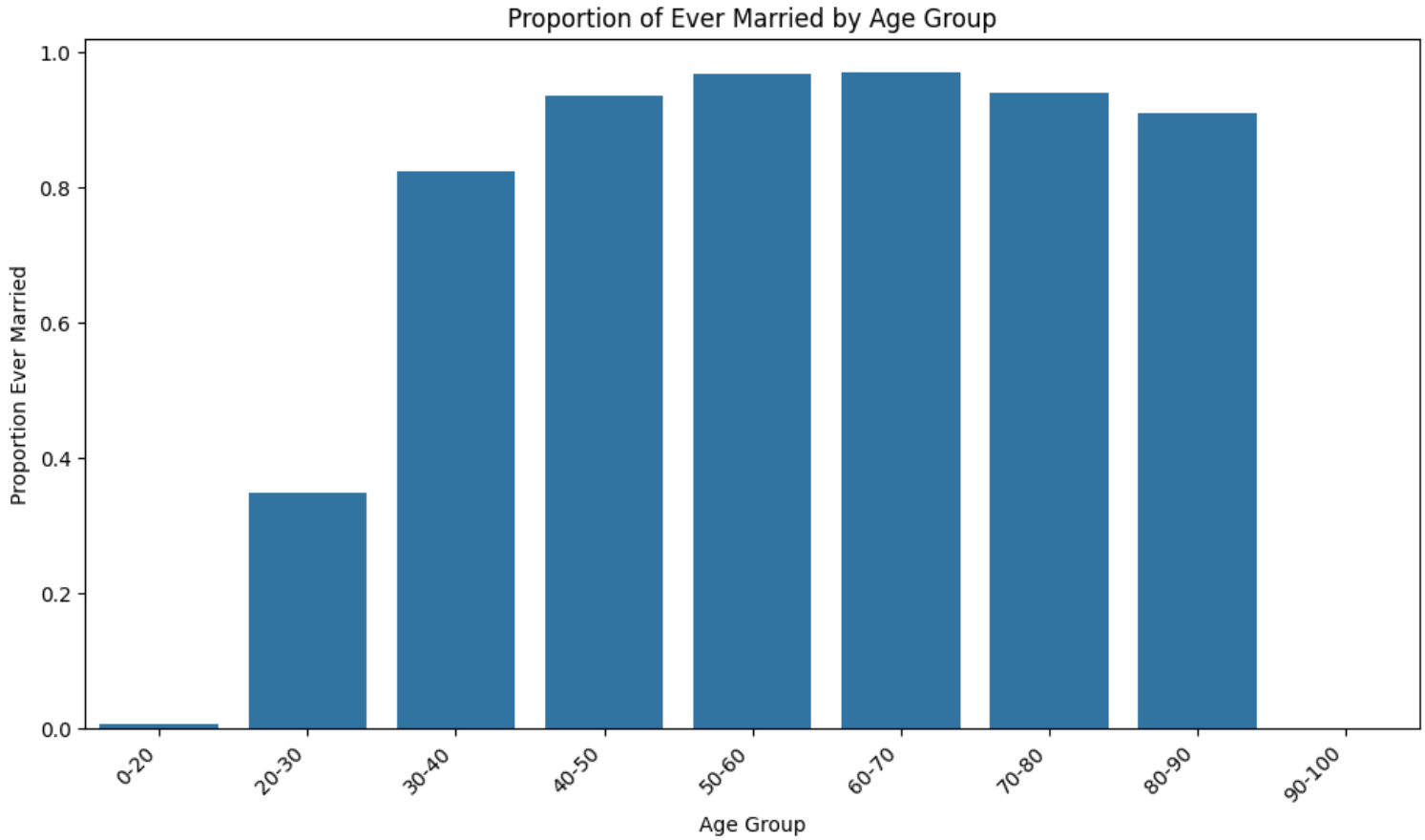
$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$$

- **Key Insights:**
  - The highest correlation is observed between age and marital status (0.82).
  - Significant correlations are noted between work type and marital status (0.59) and between work type and age (0.50).

# Chi-squared Correlation Matrix

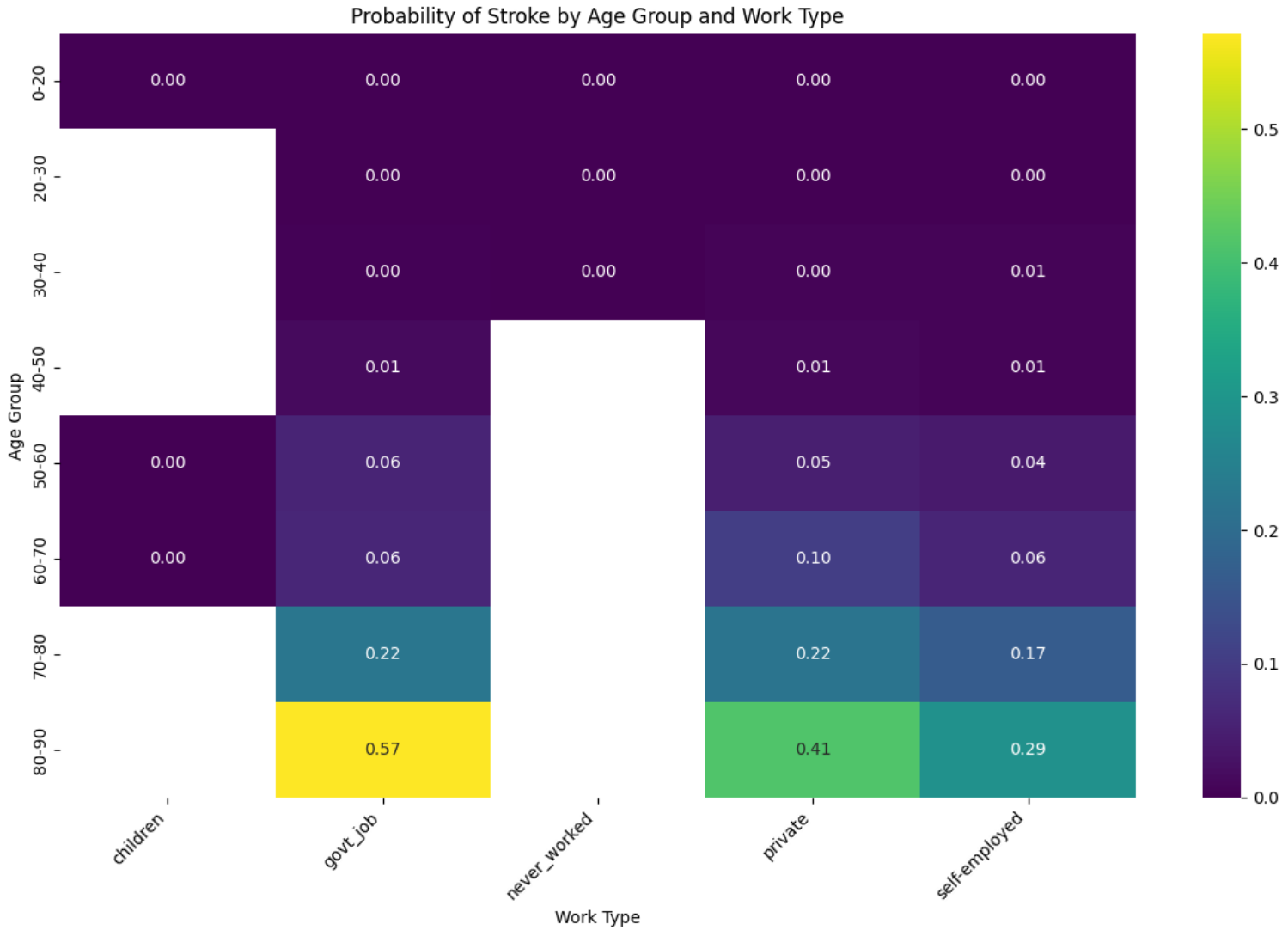


# Proportion of Ever Married by Age Group




- **Conclusion:** Marriage rates rise with age, significantly impacting stroke risk. Thus, being **single** does not necessarily reduce it.

# Probability of Stroke by Age Group and Work Type



# Probability of Stroke by Age Group and Work Type

- **Key Insights:**
  - This heatmap highlights the probability of stroke across various age groups and work types.
  - Individuals aged 70-82, especially those in government and private jobs have the highest stroke probability.
  - **Interestingly**, within the same age group, self-employed individuals have a lower stroke rate compared to those in private and government jobs, challenging our initial conclusion.

# Key Findings from EDA

- ⋆ **Age** and **Glucose Levels** are primary stroke predictors.
- ⋆ **Hypertension**, **Heart Disease**, **BMI**, and **Smoking** are significant risk factors.
- ⋆ Demographic variables such as **Gender**, **Residence Type**, and **Marital Status** have less pronounced effects.
- ⋆ **Employment Type** may influence stroke risk, requiring further investigation.

## Summary

*These findings emphasize the critical role of lifestyle, medical conditions, and specific demographic factors in stroke risk, aligning with existing research and highlighting areas for early intervention.*

# Model Training and Evaluation

# Model Training

- **Data Splitting:**
  - Split the dataset into two parts:
    - 12,242 samples for training
    - 3,061 samples for testing
  - Use stratification to ensure the target variable's distribution remains consistent across both sets.
- **Pipeline Construction:**
  - Build a pipeline that includes both preprocessing and the ensemble model.
  - Ensemble models using the Soft Voting method help increase generality and limit the overfitting of individual models.

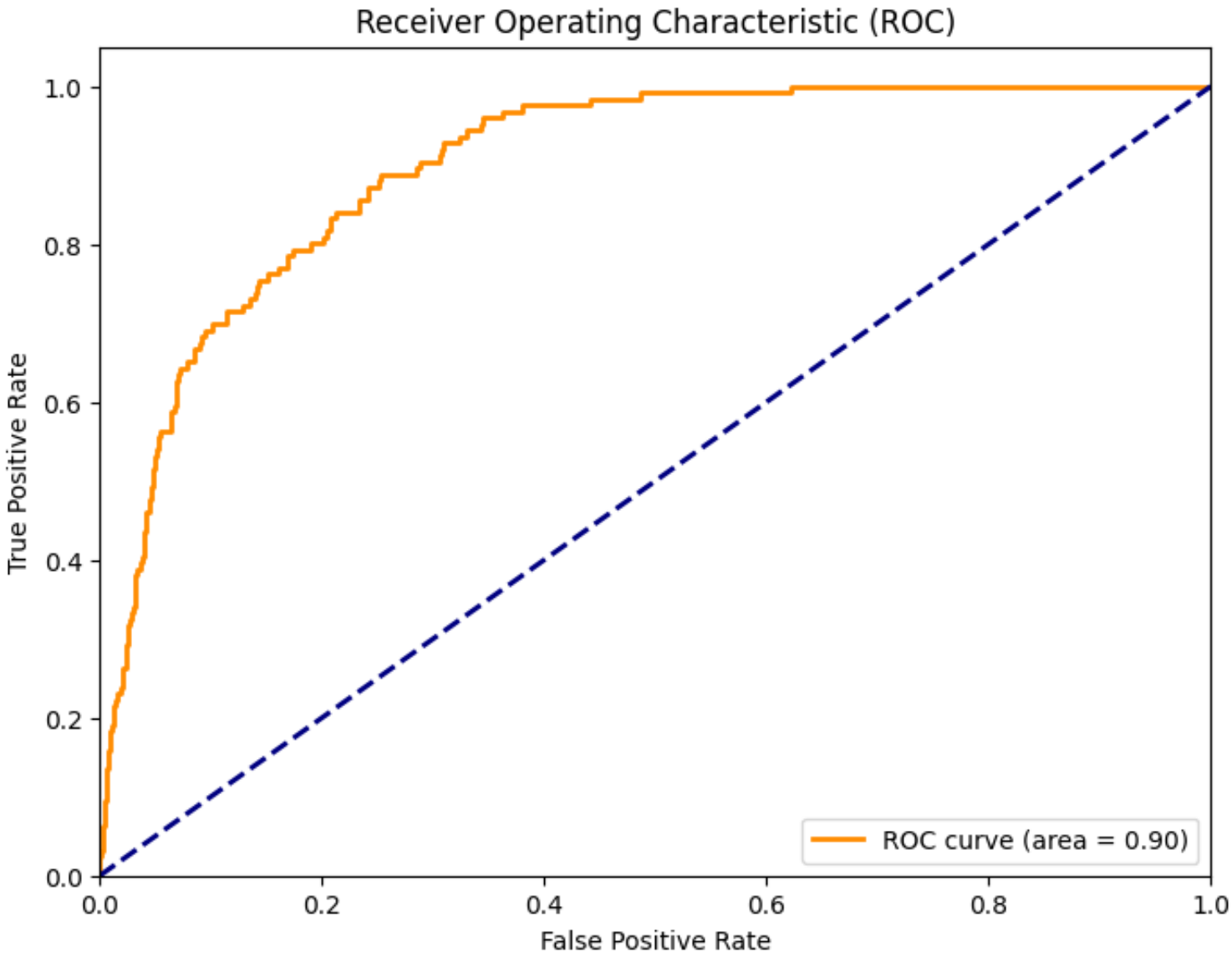
# Model Training

- **Preprocessing Steps:**
  - Convert categorical data to numeric format.
  - Apply one-hot encoding for specific features.
  - Use an imputer to fill in missing values as needed.
- **Ensemble Model Selection:**
  - Utilize three popular models for categorical data:
    - XGBoost
    - CatBoost
    - Random Forest
  - Instead of using sampling methods like SMOTE for data imbalance, apply:
    - Scale positive weight
    - Class weight set to 'balanced' to minimize disruption to data distribution.

# Model Evaluation Metrics

- **Overall Model Performance:**
  - Accuracy: 0.9579 (3061 samples)
  - Macro Average:
    - Precision: 0.7129
    - Recall: 0.5679
    - F1-Score: 0.5983
  - Weighted Average:
    - Precision: 0.9436
    - Recall: 0.9579
    - F1-Score: 0.9471
- **ROC AUC:** 0.9003

# ROC Curve


Receiver Operating Characteristic (ROC)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.90)

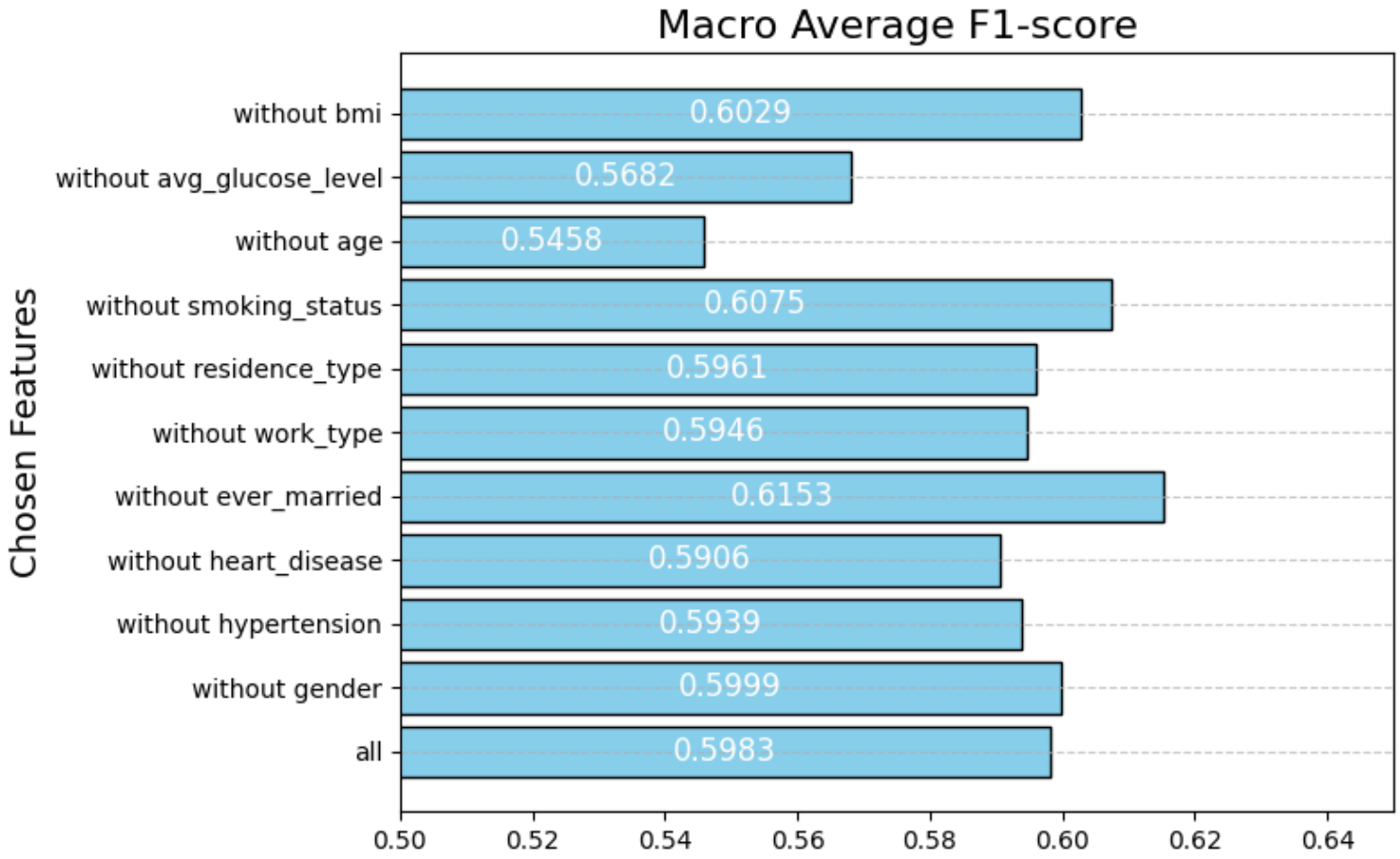
# ROC Curve

- **Overview:**
  - The ROC curve visualizes the model's performance across different classification thresholds.
  - It plots the true positive rate (sensitivity) against the false positive rate.
- **Key Insights:**
  - The curve illustrates a high true positive rate with a low false positive rate, indicating effective classification.
  - The area under the curve (AUC) is 0.90, suggesting excellent model performance.
- **Interpretation:**
  - An AUC of 0.90 indicates that the model can distinguish between classes effectively, making it a strong candidate for predictive analysis.

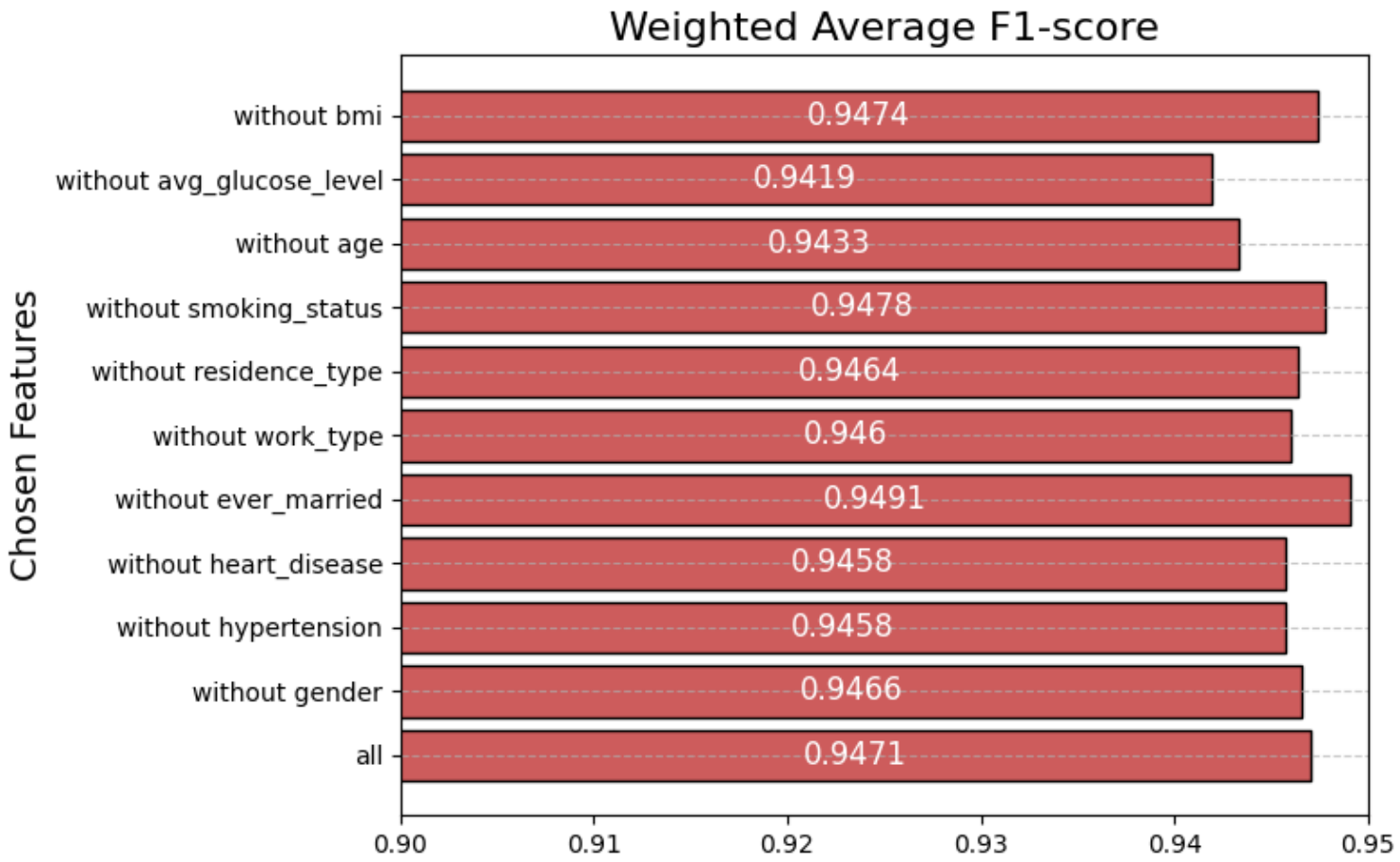
# Conclusion Validation Process

- **A Priori Analysis:**
  - Using exploratory data analysis (EDA) to make some prior conclusions and verify them.
- **A Posteriori Analysis:**
  - Utilize the trained model with high confidence to confirm EDA conclusions.
  - **Factor Evaluation:**
    - Sequentially drop each feature and retrain the model.
    - Evaluate model performance to identify important and redundant factors affecting stroke risk.
  - **Probability Validation:**
    - Use stroke prediction probabilities from the trained model and Partial Dependence Plot (PDP) to re-plot probability graphs for each feature.
    - Validate **prior analysis** and illustrate how prediction probabilities change with each feature.

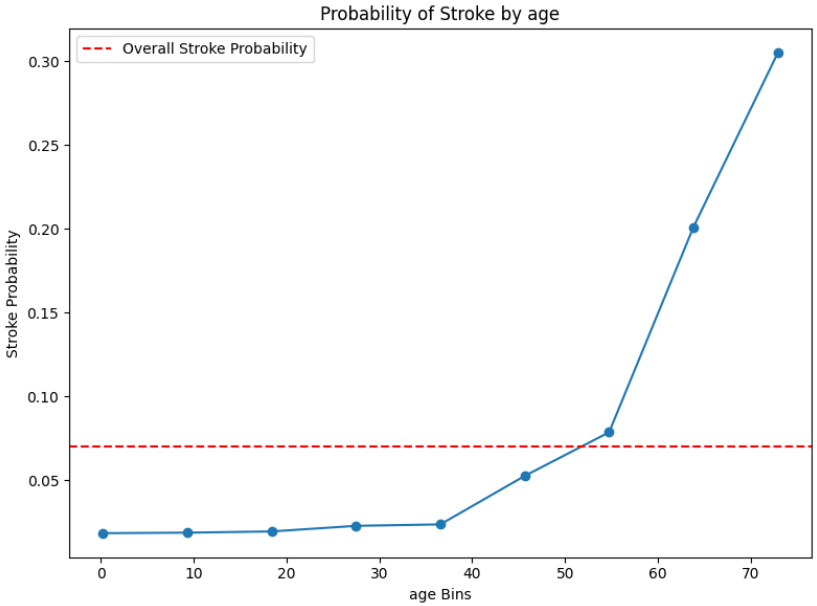
# Factor Evaluation



# Factor Evaluation



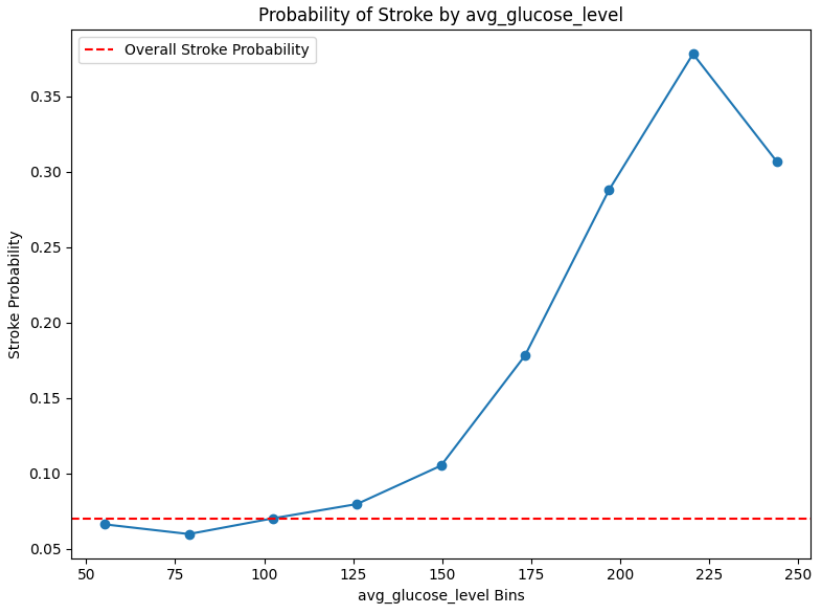
# Probability Validation

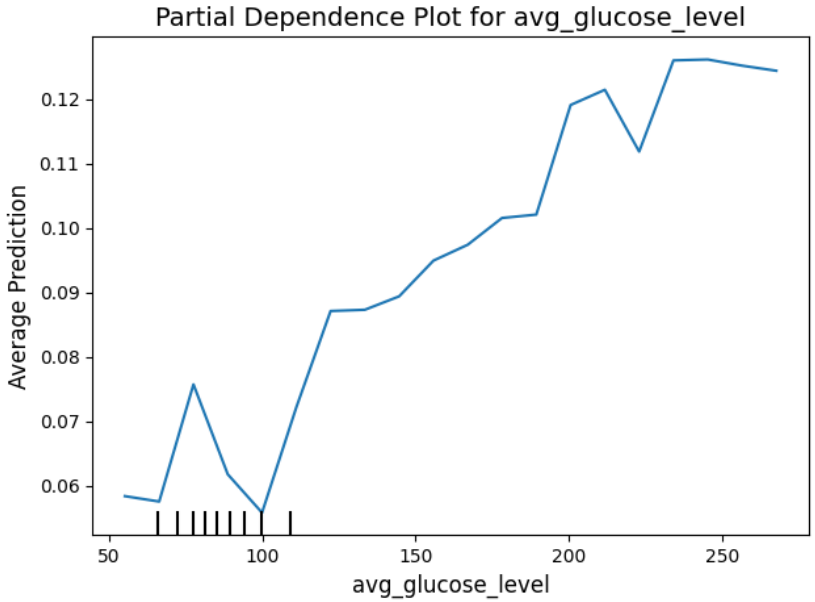




- **Key Insights:**
  - The two lines are roughly similar, with the predicted probability being low before the age of 40, then starting to grow significantly.

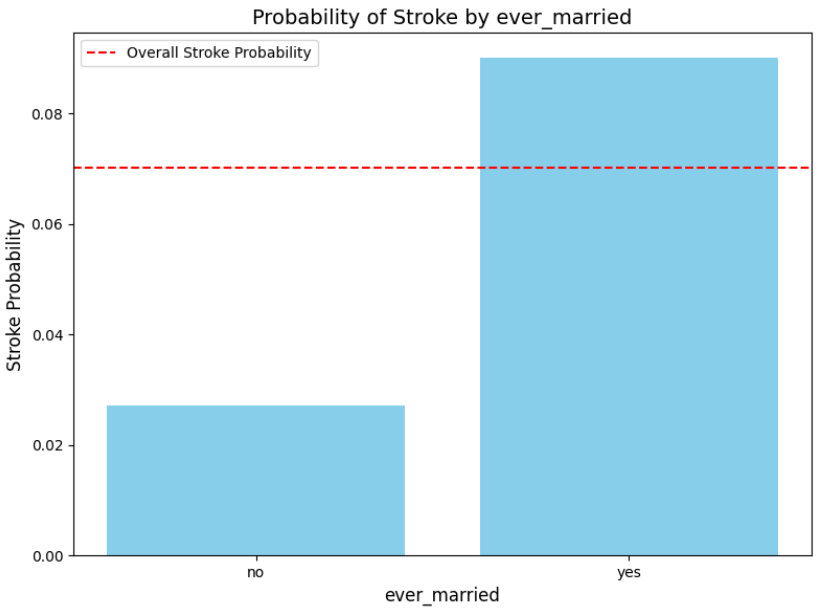
# Probability Validation

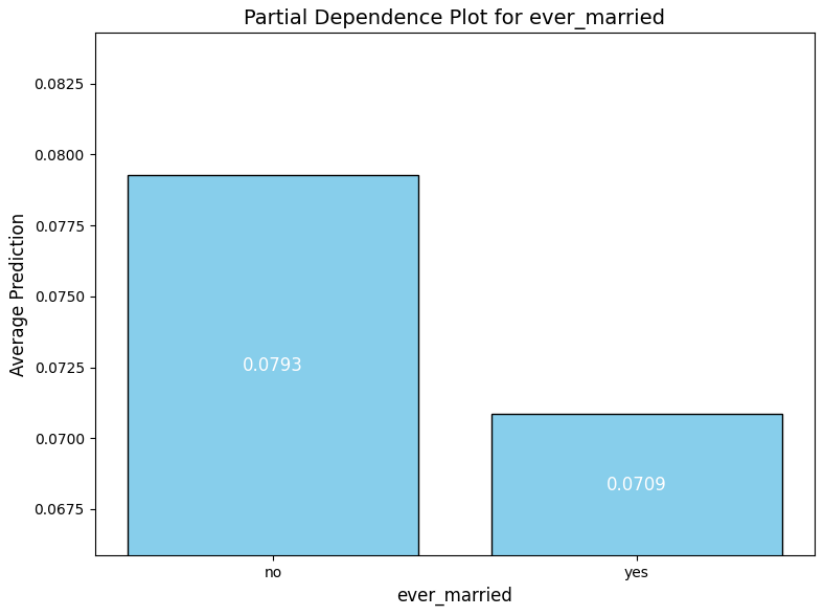



- **Key Insights:**
  - The lines shown in the two figures are not very similar, but still both show an upward trend for glucose levels higher than 100 mg/dL.

# Probability Validation




- **Key Insights:**
  - The two figures show different trends about stroke probability, suggesting that the difference in stroke probability between unmarried and married groups is not determined by the marital status itself.

# Conclusion Validation Process

- In conclusion, we identified age and glucose levels as the two most influential factors in classifying stroke risk. The variable "ever married" was deemed redundant and influenced by age, so it can be removed from consideration.
- Our posterior analysis aligned well with the conclusions drawn from the prior analyses, reinforcing the reliability of our findings.

# Project Summary

- **Data Storytelling:**
  - Formulated questions and hypotheses to guide exploration and decision-making.
- **Exploratory Data Analysis (EDA):**
  - Drew initial conclusions, verified with statistical tests (e.g., Chi-Squared).
  - Analyzed relationships between highly similar variables.
- **Posteriori Analysis:**
  - Leveraged the trained model to validate the reliability of our findings.
- **Outcome:**
  - Conclusions will serve as a knowledge base for our minimum viable product (MVP).
  - The trained model will facilitate early warning predictions and personalized user recommendations with explainable ML.

# Demonstration System

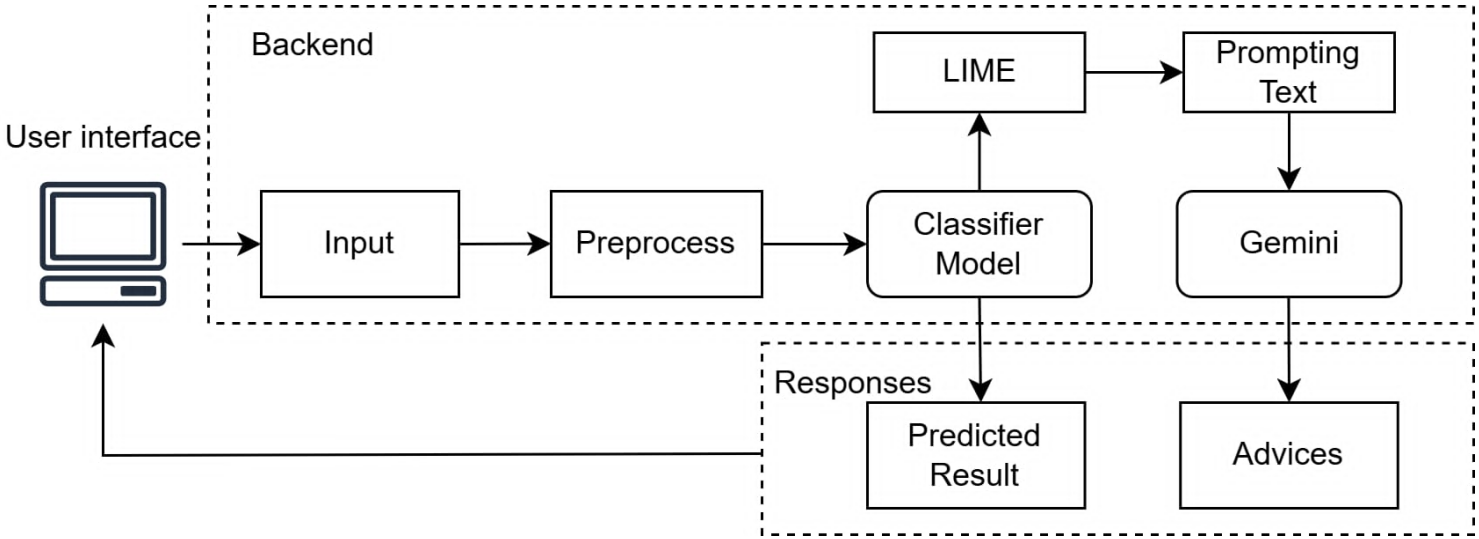



Figure: System Flow for Stroke Risk Prediction

- Use LIME to identify **key features** that contribute to stroke risk.
- These features are used to prompt Gemini, which generates personalized advice to help reduce the risk effectively.

# Questions?

Thank you for your attention!

Feel free to ask any questions or share your feedback.

# Dataset Reference

- **Dataset Title:** Stroke Prediction Dataset
- **Description:** The dataset contains information on factors affecting stroke risk, including demographic, medical, and lifestyle features.
- **Source:** Kaggle
- **Access Link:**
  - Click here to access the dataset
- **Citation:** Fedesoriano (2020). Stroke Prediction Dataset. Retrieved from Kaggle.

# References


- **Is It a Stroke or a Heart Attack?** - Healthline
- **Heart Disease and Stroke** - WebMD
- **What is Stroke?** - Heart and Stroke Foundation
- **Stroke After a Heart Attack: What's the Risk?** - Harvard Health