Stroke Analysis: A Data-Driven Exploration

Group 9, Data Mining, UIT

December 22, 2024

Abstract

Stroke is a leading global cause of mortality and disability, emphasizing the urgent need for predictive analytics to mitigate its impact. This study investigates stroke risk factors through detailed data analysis and machine learning models, aiming to deliver actionable insights for healthcare. Using a structured methodology, we explored stroke causes, validated findings, and developed a Minimal Viable Product (MVP) to assist individuals in reducing stroke risk while promoting community awareness.

1 Introduction

Stroke is among the foremost causes of global disability and mortality, highlighting the urgent need for early detection and prevention. Understanding the factors contributing to stroke can help healthcare providers develop targeted interventions. Our study aims to:

- Analyze predictors of stroke risk using exploratory data analysis (EDA).
- Develop and evaluate machine learning models to accurately predict stroke occurrences and validate our findings from data analysis.
- Provide evidence-based recommendations to mitigate stroke risks in at-risk populations.

Combining *statistical analysis* with *advanced machine learning techniques*, this project illuminates critical drivers of stroke and their implications for healthcare strategies.

2 Dataset Overview

The dataset comprises features spanning three categories:

1. **Demographic Information:** Age, gender, marital status, work type, and residence type.
2. **Medical History:** Conditions like hypertension and heart disease.
3. **Lifestyle Factors:** Smoking status, average glucose level, and BMI.

The target variable is stroke occurrence, categorized as a binary classification (Yes/No). The dataset includes 15,303 samples, with 4% positive cases and 96% negative cases.

Key Statistics

- **Numerical Features:**
  - **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.

- **Categorical Features:** Gender, work type, residence type, marital status, smoking status, hypertension, and heart disease.

This diverse dataset provides a comprehensive foundation for analyzing relationships between predictors and stroke occurrences.

3 Proposal Process

To achieve our research objectives, we adopted a structured, rigorous approach:

1. **Data Storytelling:** Formulated research questions and hypotheses to guide exploration.
2. **Exploratory Data Analysis (EDA):** Used statistical methods (e.g., Chi-squared tests) and visualizations to uncover patterns and correlations, which derive *qualitative insights* and validate with *quantitative methods.*
3. **Model Training and Evaluation:** Developed machine learning models to validate findings, employing techniques such as:
   - **Factor Evaluation:** Sequentially removed individual features to assess variable importance.
   - **Probability Validation:** Used predicted probabilities and Partial Dependence Plot (PDP) to confirm EDA trends.

This combination of **a priori** (EDA) and **a posteriori** (model validation) analyses ensured reliable and actionable conclusions.

3.1 Priori Analyses

The EDA identified key patterns and validated assumptions about stroke risk factors. As shown in Figures 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, the key areas include:

- **Age:** Stroke prevalence rises significantly with age, particularly for those over 50.
- **BMI and Glucose Levels:** Elevated BMI and glucose levels correlate strongly with higher stroke risk.
- **Smoking:** Current and former smokers face greater risks than non-smokers.
- **Medical Conditions:** Hypertension and heart disease increase stroke risk by 4–6 times.
- **Demographics:** Gender differences are minimal; marital status and geographic location have limited impact.
- **Occupational Insights:** Self-employed individuals exhibit lower stroke rates than private or government workers within similar age groups.

Conclusion: These findings highlight the importance of age, glucose levels, and lifestyle factors, as well as suggest areas for further investigation to mitigate stroke risks.

3.2 Model Development and Evaluation

We employ robust machine-learning models to validate our findings. Preprocessing included:

- Splitting the dataset into training and testing sets (80:20) using stratification.
- Encoding categorical variables using one-hot encoding.
- Handling missing values through statistical imputation.
- Balancing class weights to address data imbalance.

Models Used

- **XGBoost:** Handles complex patterns effectively.
- **CatBoost:** Optimized for categorical features.
- **Random Forest:** Ensures stability and generalization.

These models are combined using the Soft Voting method to enhance performance and mitigate overfitting.

Performance Metrics

- **Accuracy:** 95.79 %
- **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

These metrics underscore the model's robustness and reliability in predicting stroke risks.

3.3 Findings Validation by Trained Model

The trained model is used to validate the findings from exploratory data analysis. Two methods, **Factor Evaluation** and **Probability Validation**, are applied. As shown in Figures 19, 20, 21, 22, 23, 24, 25, 26, key results include:

- **Key Predictors:** Age and glucose levels are the most influential factors.
- **Redundancy:** The ever-married variable is deemed redundant, as its impact was predominantly captured by age.

The results from the posterior analysis are consistent with the findings of the prior analyses, providing evidence of the reliability and robustness of the conclusions drawn from the dataset.

4 Conclusion

4.1 Summary of Findings

This study identifies key factors that influence stroke risk, including age and glucose levels, while emphasizing the importance of medical conditions such as hypertension and heart disease. In addition, we explore lifestyle modifications that can help reduce stroke risk. Using machine learning models, we validated our findings and provided actionable insights for healthcare interventions aimed at mitigating stroke risk.

4.2 Our MVP

Based on our findings, we developed an MVP with the hope to reduce stroke risk for personnel and enhance public awareness, following features:

- **Explainable Risk Prediction:** Users input health parameters, and the system predicts stroke risk with **explanations**.
- **Recommendations:** Tailored suggestions to mitigate risk and improve health.

- **Educational Platform:** A website with visualizations and analyses for healthcare professionals and the public.

The MVP shown in Figure 1, bridges advanced AI techniques with user-friendly design, supporting individual and community health initiatives.

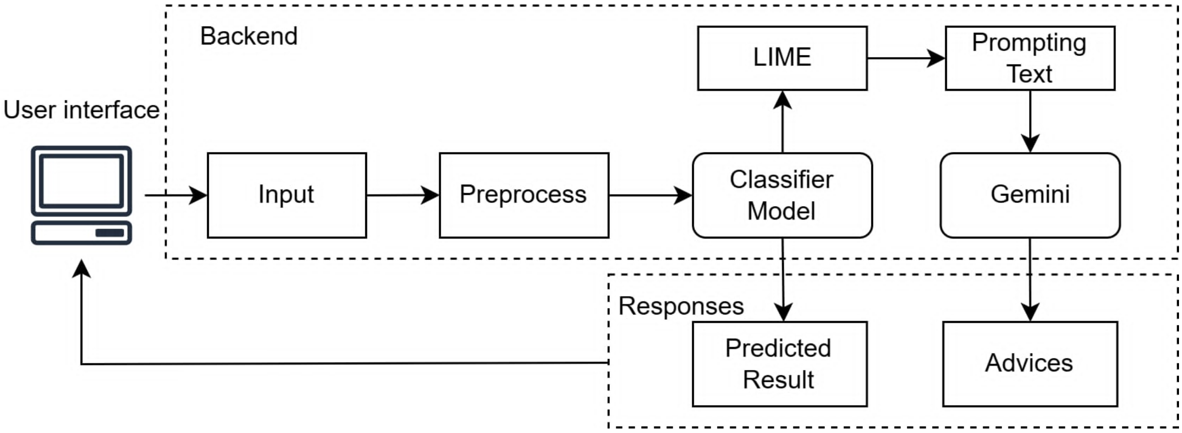

Figure 1: System architecture of our MVP

References

GitHub Repository:

Our project story

Related Articles

1. Healthline. Is It a Stroke or a Heart Attack?
2. WebMD. Heart Disease and Stroke
3. Heart and Stroke Foundation. What is Stroke?
4. Harvard Health. Stroke After a Heart Attack: What's the Risk?

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.

Figures

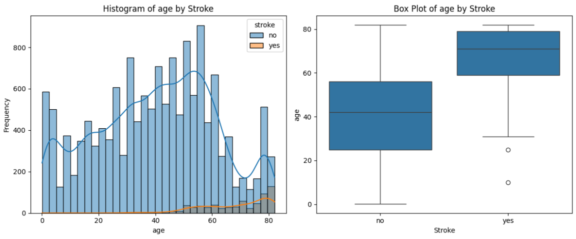

Figure 2: Age Distribution and Stroke

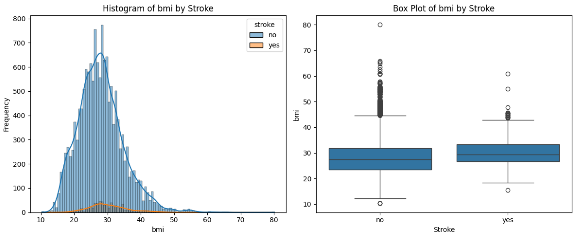

Figure 3: Glucose Distribution and Stroke

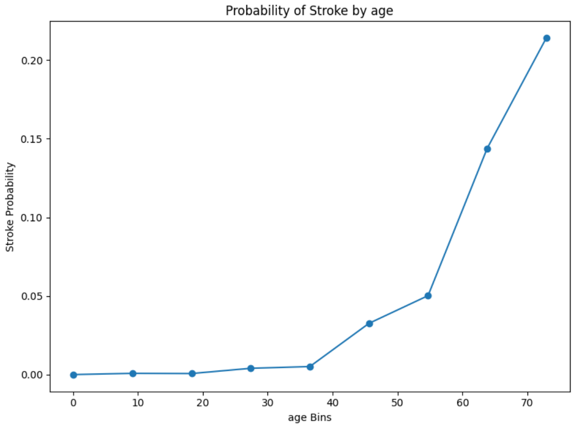

Figure 4: Probability of Stroke by Age

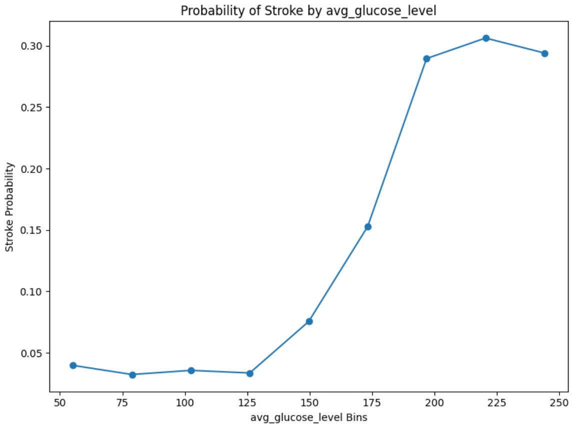

Figure 5: Probability of Stroke by Glucose

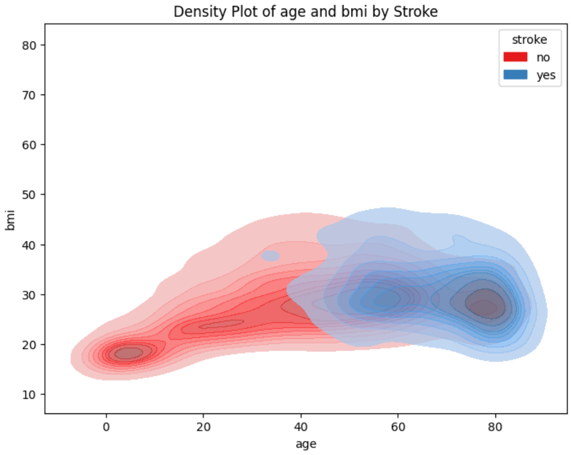

Figure 6: Age vs. BMI Relationship.1

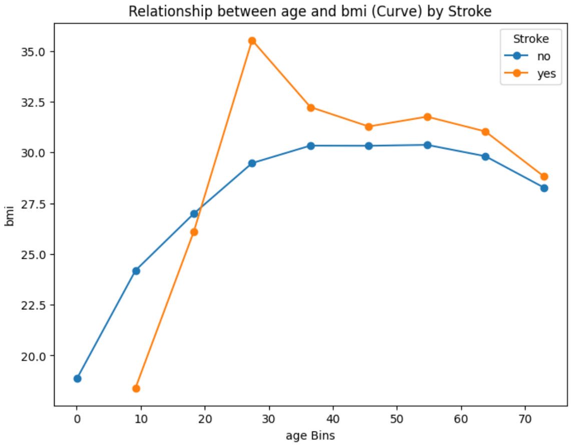

Figure 7: Age vs. BMI Relationship.2

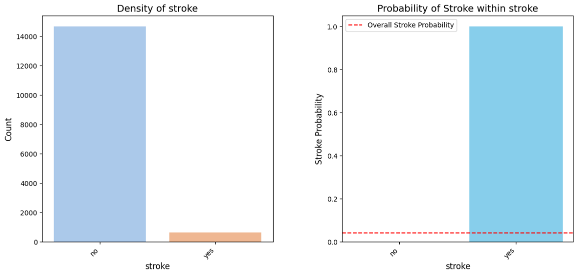

Figure 8: Stroke Analysis Visualization.1

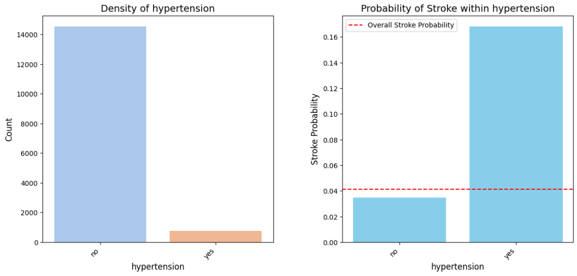

Figure 9: Stroke Analysis Visualization.2

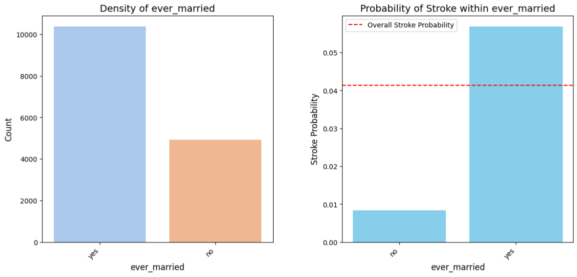

Figure 10: Stroke Analysis Visualization.3

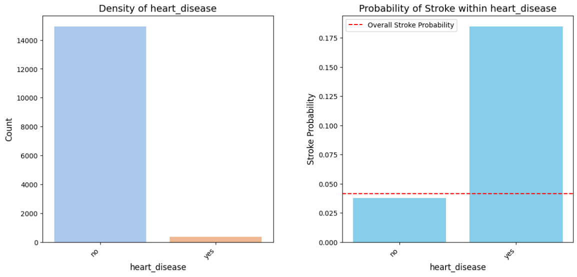

Figure 11: Stroke Analysis Visualization.4

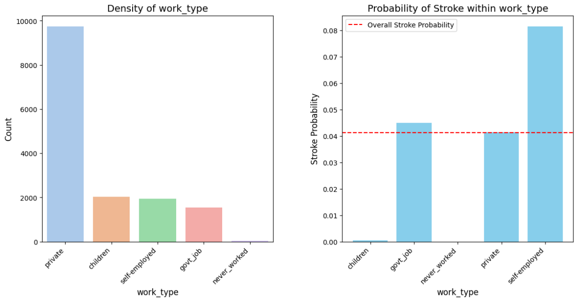

Figure 12: Stroke Analysis Visualization.5

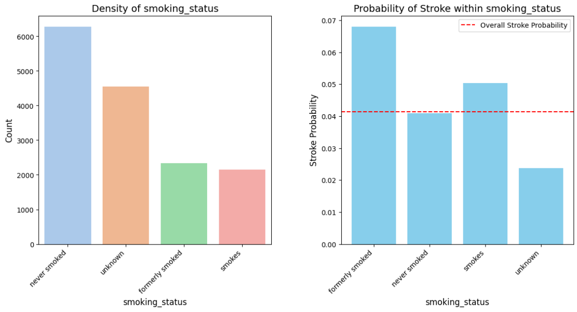

Figure 13: Stroke Analysis Visualization.6

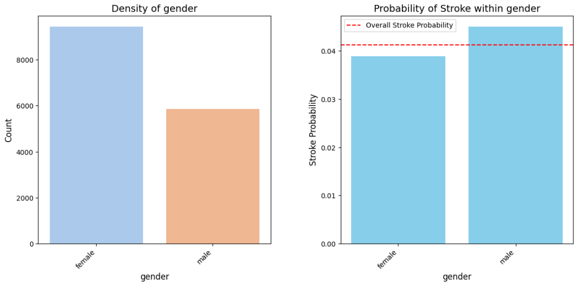

Figure 14: Stroke Analysis Visualization.7

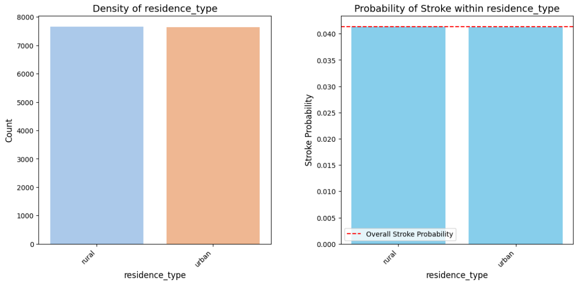

Figure 15: Stroke Analysis Visualization.8

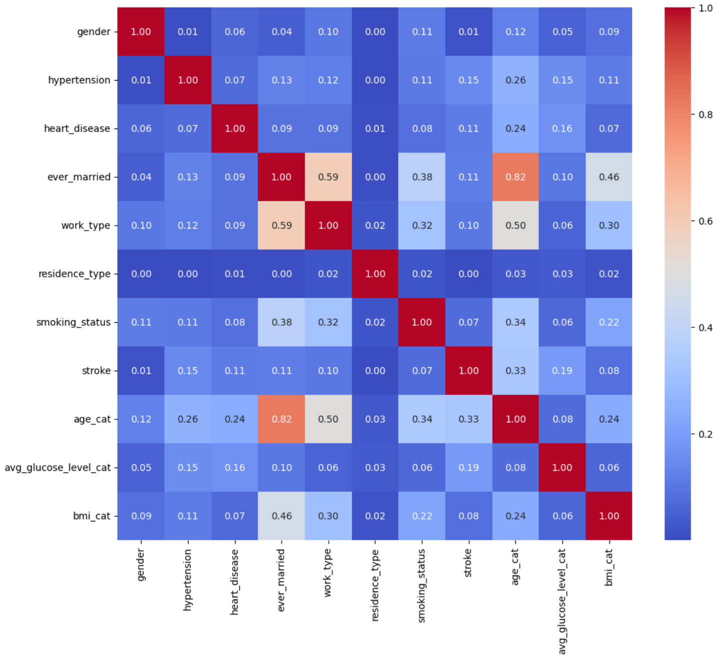

Figure 16: Chi-squared Correlation Matrix

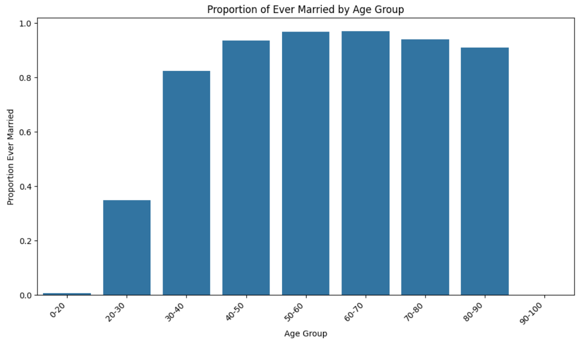

Figure 17: Proportion of Ever Married by Age Group

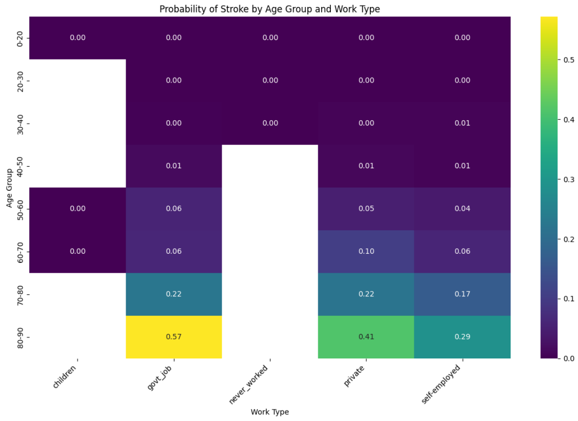

Figure 18: Probability of Stroke by Age Group and Work Type

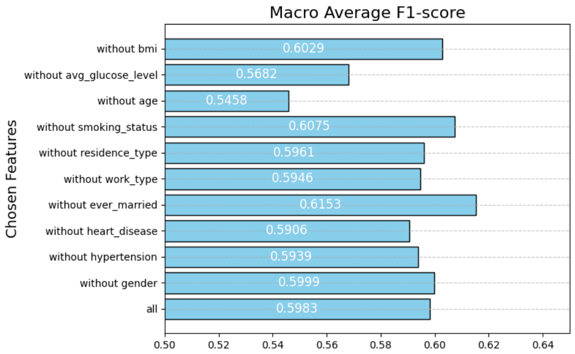

Figure 19: Macro Average F1-score When Dropping Each Feature

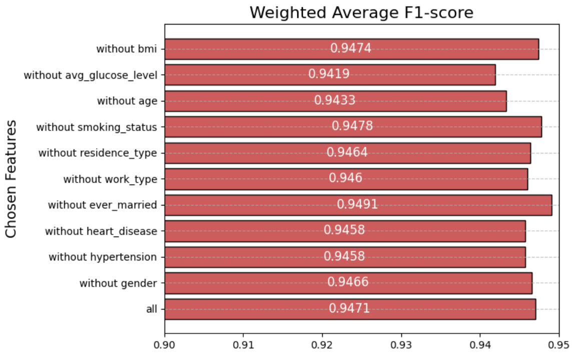

Figure 20: Weighted Average F1-score When Dropping Each Feature

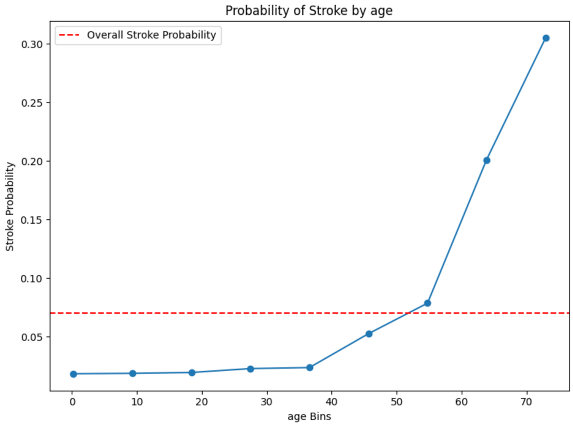

Figure 21: Original Predicted Stroke Probability by Age

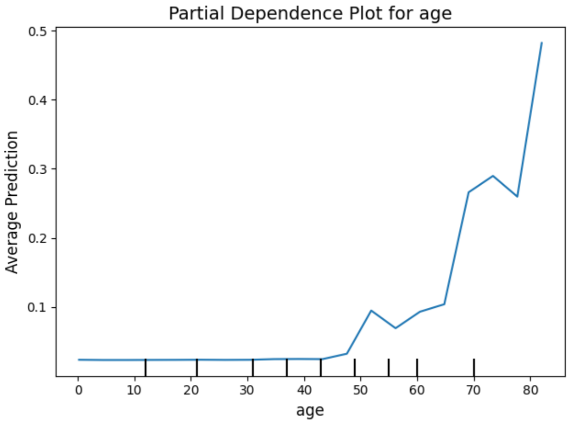

Figure 22: Partial Dependence Predicted Stroke Probability by Age

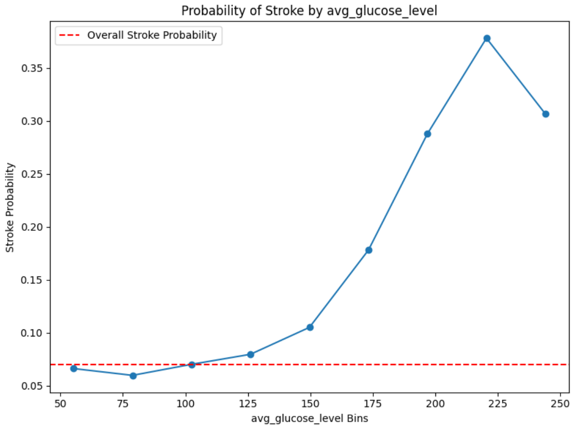

Figure 23: Original Predicted Stroke Probability by Glucose Level

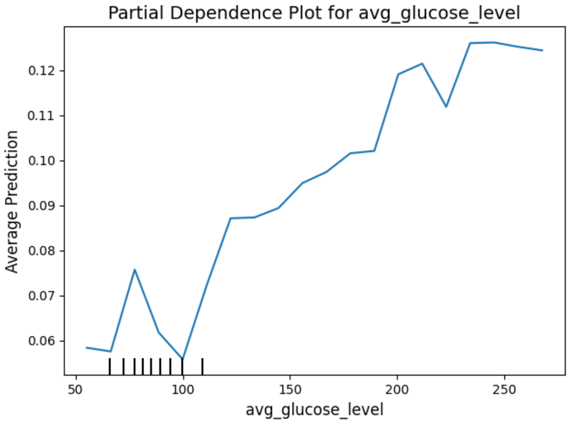

Figure 24: Partial Dependence Predicted Stroke Probability by Glucose Level

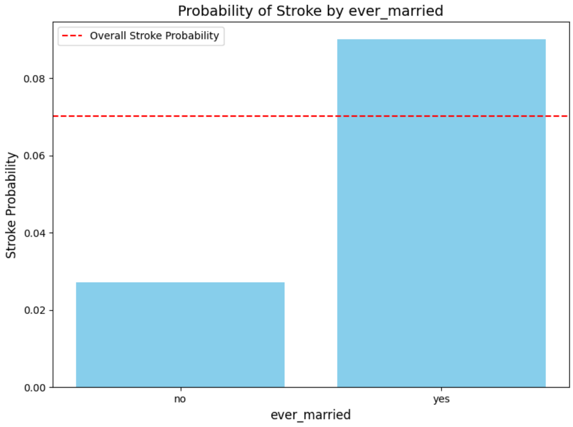

Figure 25: Original Predicted Stroke Probability by Marital Status

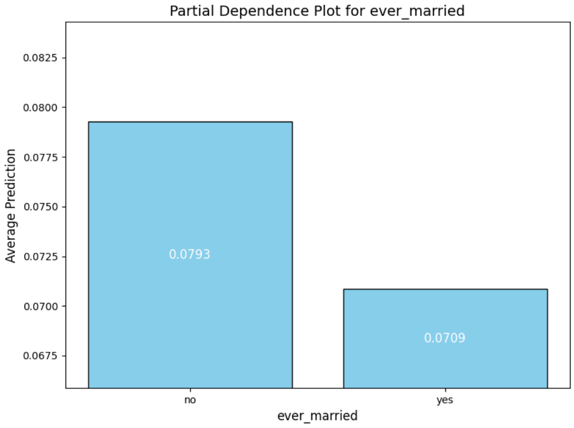

Figure 26: Partial Dependence Predicted Stroke Probability by Marital Status