Data-Driven Exploration of Stroke Risk Factors Using a Public Healthcare Dataset

Introduction

Stroke is a major cause of mortality and long-term disability worldwide. Understanding the factors associated with stroke risk is essential for prevention and early intervention. Data-driven analysis of population health datasets enables the systematic exploration of demographic, metabolic, lifestyle, and vascular variables and their relationship to stroke occurrence.

In this project, we analyze a public healthcare dataset to identify stable and interpretable patterns associated with stroke. Rather than constructing a complex predictive system, the primary objective is to explore structural relationships in the data, assess robustness, examine interaction and confounding effects, and quantify relative risk across major factors. Emphasis is placed on interpretability, reproducibility, and statistical reasoning.

Data and Preprocessing

The analysis is based on a public healthcare dataset containing approximately 5,000 individuals. Each row represents a single individual with demographic, clinical, and lifestyle information. The binary outcome variable is stroke (0 = no stroke, 1 = stroke).

Variables examined include:
Demographic: Age, Gender, Ever Married, Residence Type
Clinical/Vascular: Hypertension, Heart Disease, BMI
Metabolic: Average Glucose Level
Lifestyle/Socioeconomic: Work Type, Smoking Status

Preprocessing steps included removing the rare 'Other' gender category to avoid instability due to extremely low counts, imputing missing BMI values using the median (robust to outliers), and converting categorical variables to appropriate data types. The cleaned dataset was saved to ensure reproducibility of all analyses.

Methodology

Exploratory data analysis (EDA) was conducted using histograms, boxplots, and summary statistics for age, average glucose level, and BMI. Stroke prevalence was compared across categorical variables using normalized cross-tabulations.

To assess robustness of the age–stroke relationship, 100 repeated random half-sample splits were performed. For each split, stroke probability was computed across age bins and summarized using mean trends and 95% variability bands.

To investigate non-linear effects, glucose levels were grouped into intervals and stroke probability per bin was examined. An interaction feature (Age × Glucose) was constructed

to evaluate combined effects. Potential confounding between marital status and age, as well as smoking status and age, was assessed. A cumulative vascular risk score (hypertension + heart disease) was created to examine additive effects. Relative Risk (RR) was computed for selected factors.

In addition, a logistic regression model was implemented to demonstrate how multiple variables can be combined into a probabilistic stroke risk score. Categorical variables were encoded using one-hot encoding. The dataset was split into training (75%) and testing (25%) sets using stratified sampling. Because stroke cases are relatively rare, class weighting was applied to mitigate imbalance. The model outputs a probability score (0–100 scale) presented via an interactive demonstration interface. This tool is intended solely for educational purposes and is not a clinical diagnostic system.

Results

Exploratory analysis revealed substantial age differences between stroke and non-stroke groups. Stroke patients were markedly older on average, and age distributions showed a clear rightward shift among stroke cases. Repeated random splits demonstrated a consistent monotonic increase in stroke probability with age, confirming robustness.

Average glucose levels were higher among stroke cases. Binned analysis suggested a threshold-like increase in stroke probability at higher glucose ranges. BMI distributions showed substantial overlap between groups; however, summary statistics indicated that mean BMI in the stroke group was slightly higher, suggesting a modest contribution despite weaker visual separation.

Hypertension and heart disease exhibited strong associations with stroke. Individuals with both elevated blood pressure and diagnosed heart disease showed the highest stroke probability, demonstrating cumulative vascular risk. Residence type showed minimal association, whereas work type displayed clearer variation in stroke rates, suggesting that occupational or socioeconomic factors may be more relevant than residential setting.

Initial differences observed for marital status and smoking status were largely explained by age confounding. Married individuals were substantially older on average, and smoking categories exhibited distinct age distributions. Once age was accounted for, these factors appeared to reflect demographic structure rather than independent causal effects.

Discussion

The findings indicate that stroke risk in this dataset is primarily driven by age and explicitly defined vascular conditions (hypertension and heart disease). The robustness analysis confirms that the age–stroke association is stable across repeated sampling, reflecting a structural pattern rather than random variation.

Metabolic factors, particularly elevated glucose levels, contribute meaningfully and may exhibit non-linear behavior at higher ranges. BMI demonstrated weaker standalone

association; however, the slightly higher mean BMI among stroke cases suggests a modest effect.

A key distinction emerged between residence type and work type. Residence showed little variation in stroke probability, while work categories demonstrated clearer variability, indicating that occupational or socioeconomic conditions may play a more meaningful role.

Observed associations for marital status and smoking were largely explained by age-related confounding. This emphasizes the importance of careful interpretation of observational data and the need to distinguish correlation from structural drivers.

The logistic regression demonstration illustrates how multiple variables jointly influence predicted stroke probability. Age and vascular conditions contribute most strongly to estimated risk, while other factors provide incremental adjustments. The dataset is imbalanced and the model is not clinically validated; it serves as an educational demonstration rather than a diagnostic system.

Conclusion

This project demonstrates how interpretable statistical analysis combined with simple predictive modeling can uncover stable and meaningful patterns in healthcare data. Age and vascular health (specifically hypertension and heart disease) emerged as the strongest factors associated with stroke. Glucose levels contributed notably at higher ranges, while BMI showed modest association. Overall, stroke risk appears to arise from the interaction of demographic aging, metabolic imbalance, and vascular pathology.