Data-driven Exploration of Stroke Risk Factors Using a Public Dataset

Introduction

Stroke is a major cause of global mortality and disability, making the identification of key risk factors a priority for preventative healthcare. This study uses a public healthcare dataset to explore the structural relationships between demographic, clinical, and lifestyle variables and stroke occurrence. The primary focus of this project is on the interpretability of these relationships and the quantification of relative risk across major factors.

Data and Preprocessing

The analysis was based on a public healthcare dataset containing approximately 5,000 individuals, covering 12 categories of demographic, metabolic, and lifestyle metrics. The binary outcome variable is the presence or absence of stroke.

Variables examined included:

- Demographic: Age, Gender, Ever Married
- Clinical: Hypertension, Heart Disease, BMI
- Metabolic: Average Glucose Levels
- Lifestyle/Socioeconomic: Work Type, Residence Type

To ensure the integrity of the results, several preprocessing steps were implemented:

- The 'other' gender category was excluded to prevent statistical instability due to its low count.
- Missing BMI values were filled using median imputation, which maintains robustness against outliers.
- Categorical variables were converted 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 levels and BMI. Stroke prevalence was compared across categorical variables using normalized cross-tabulations.

To assess the robustness of 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 x Glucose) was constructed to evaluate combined effects.

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

In addition, a 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. Since 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. The tool is intended solely for educational purposes and is not a clinical diagnostic system.

Results

Age and stroke probability:

The analysis revealed substantial age differences between stroke and non-stroke groups. Stroke patients were older on average and repeated random splits demonstrated a consistent monotonic increase in stroke probability with age, confirming robustness (figure 1).

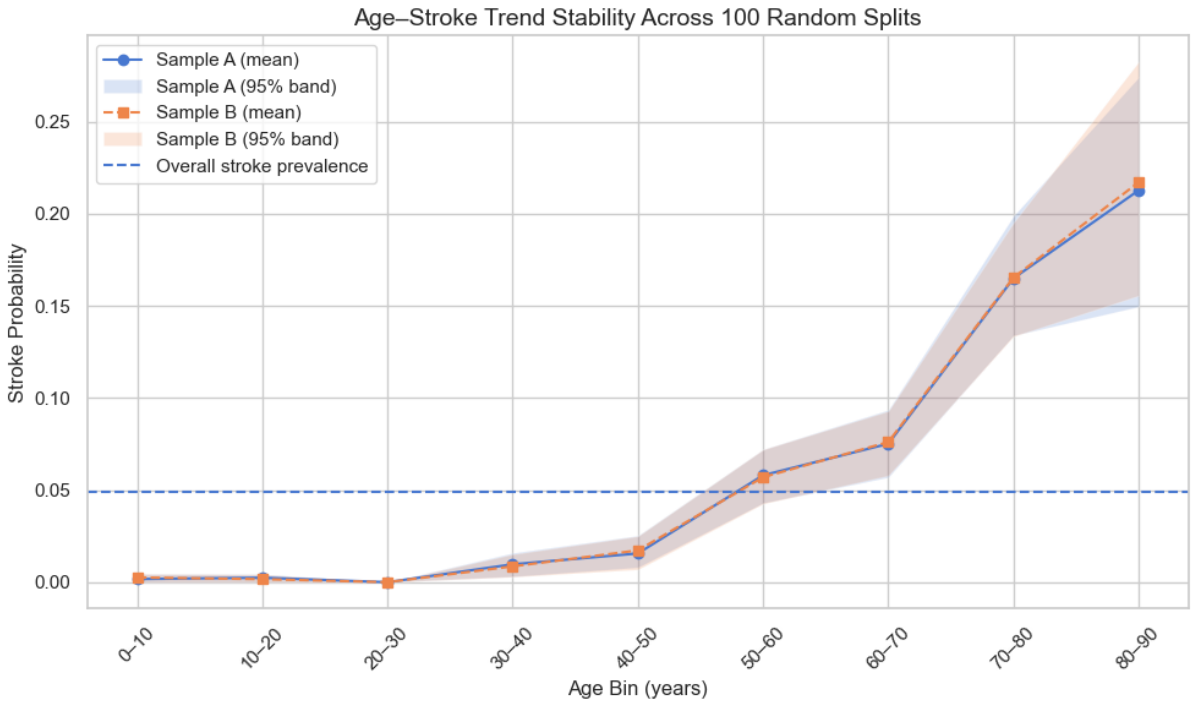

Figure 1: Robustness of Age-Stroke Association.
The line represents the average stroke probability as age increases, calculated across 100 random half-sample splits. The shaded region around the line indicates the 95% variability range, showing the stability of the trend across different data subsets.

Metabolic Trends - Glucose and BMI:

Average glucose levels were higher among stroke cases. Binned analysis suggested a “threshold-like” increase in stroke risk at higher ranges (figure 2). BMI showed a more modest association, with only a slight increase in mean values among the stroke group.

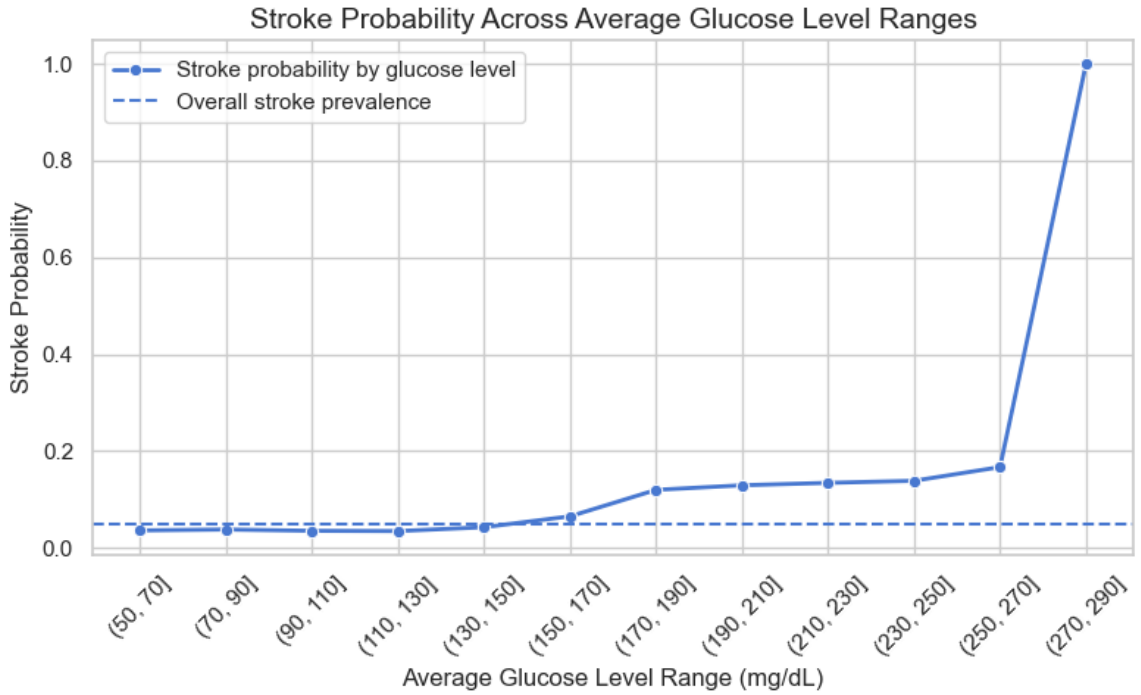

Figure 2: Stroke Risk by Glucose Levels
This chart displays the probability of stroke occurrences across binned intervals of average glucose levels. The height of the line indicates the relative frequency of stroke cases within each specific glucose range.

Cumulative Vascular Impact:
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 (figure 3).

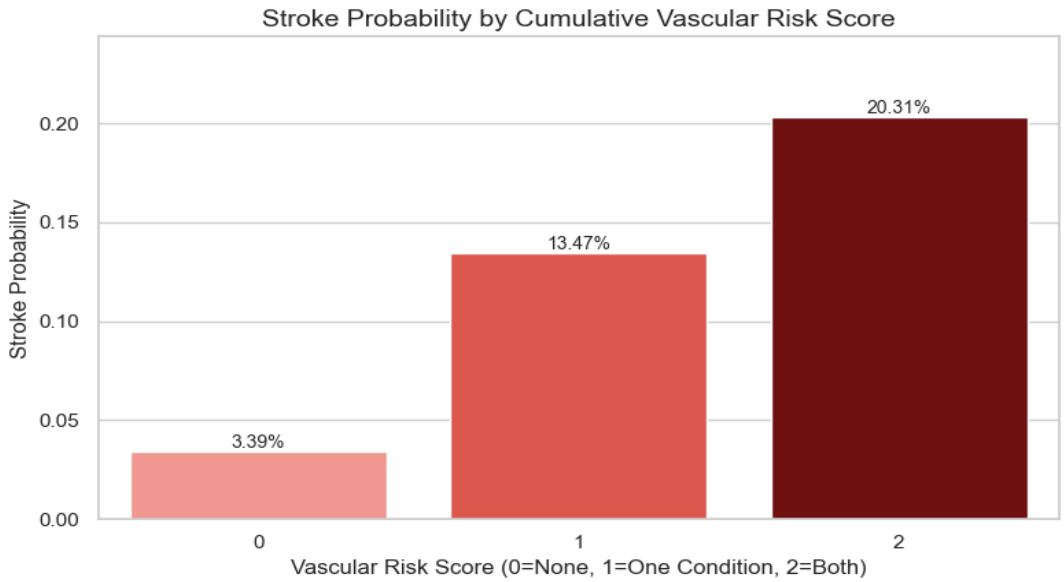

Figure 3:
The bars categorize individuals based on the presence of hypertension, heart disease, or both. The Y axis represents the stroke rate, showing how the frequency of stroke changes as these markers are combined.

Socioeconomic and Occupational Factors:
Residence type showed minimal association, while work type displayed a clearer variation in stroke rates across categories. This suggest that occupational or socioeconomic factors may be more relevant than residential settings. (figure 4).

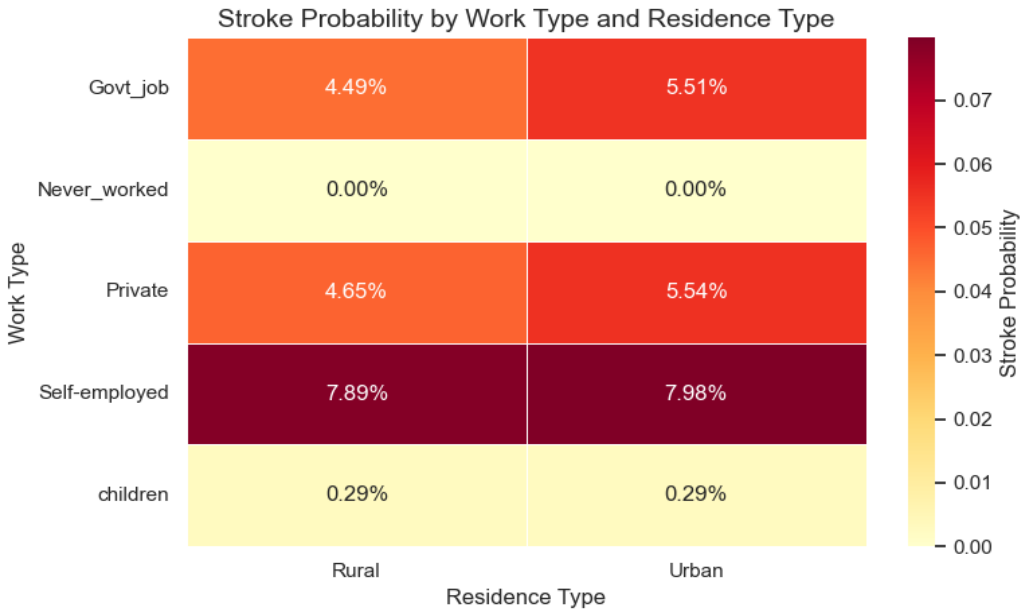

Figure 4: Occupational vs. Residential Factors
The bars show the percentage of stroke cases across different occupational categories and geographic settings. The chart allows for a side-by-side comparison of stroke distribution across these variables. The different colours allow for an easier read and comparison of the different combinations of settings.

Confounding Variables:
Initial stroke rate associations with marital status and smoking were largely explained by age confounding. Married individuals were substantially older on average (figure 5), and smoking categories exhibited distinct age distributions (figure 6). Once age was considered, these factors were found to be markers of demographic composition rather than independent causal drivers of stroke risk.

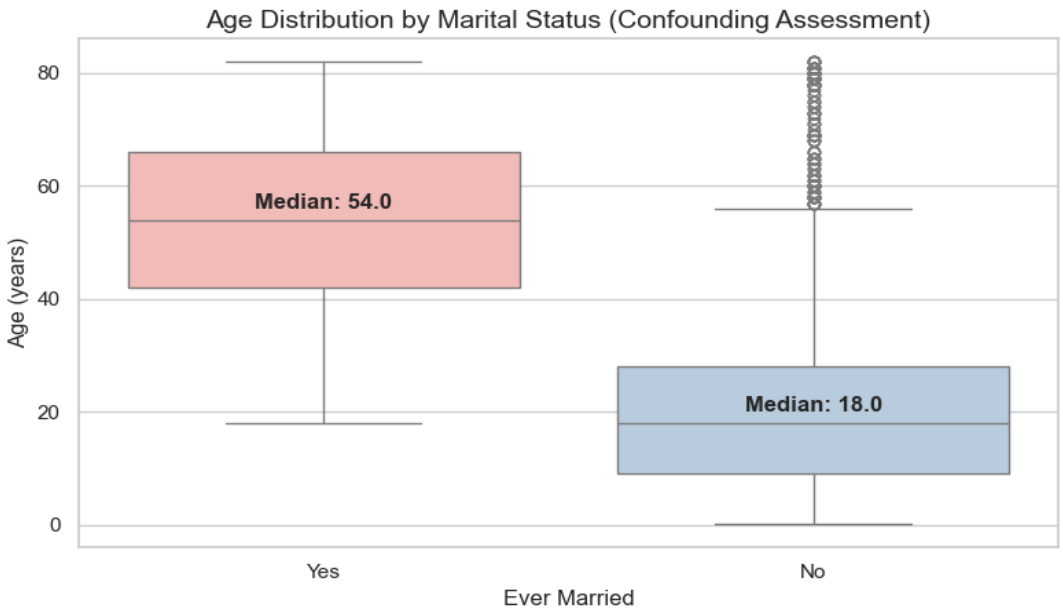

Figure 5: Marital status and Age confounding
This box plot displays the median, quartiles, and range of the ages within the “ever married” and the “married” groups. It compares the two to illustrate the demographics of each group.

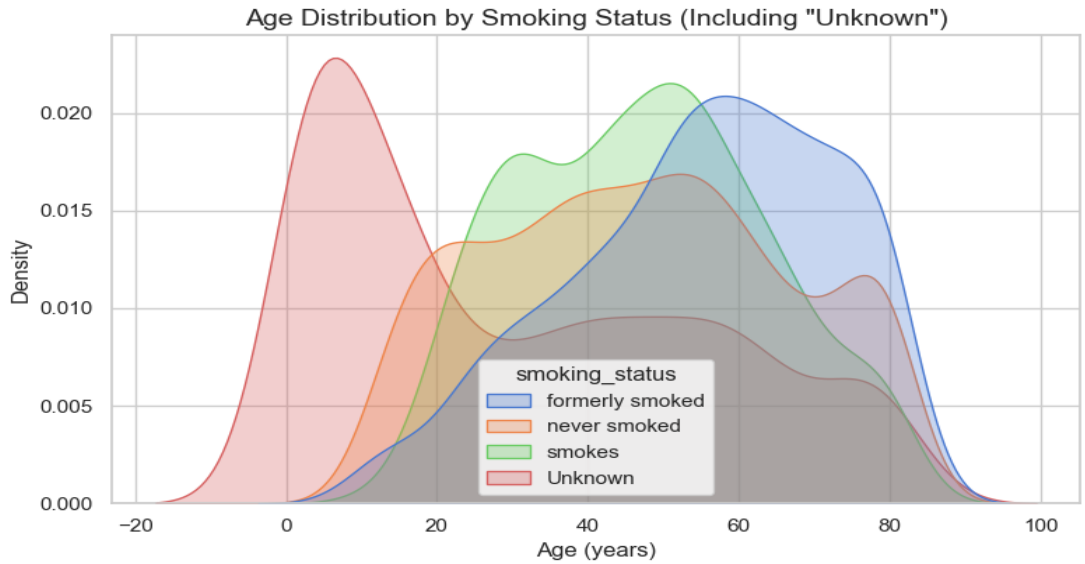

Figure 6: Smoking Status and Age Confounding
This visualization shows the age ranges for each smoking status category. Specifically, it highlights the age distribution of the “formerly smoked” group and the “unknown” group, showing the demographic age gaps between the groups.

Discussion

The findings indicate that stroke risk in this dataset is primarily driven by age and specific 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 behaviour at higher ranges. BMI demonstrated weaker association with stroke. However, the slightly higher mean BMI among stroke cases suggests a slight effect.

A key distinction emerged between residence type and work type. Residence showed a 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 given data and the need to distinguish correlation from structural drivers.

The logistic regression demonstration illustrates how multiple variables can join to influence predicted stroke probability. Age and vascular conditions contribute most strongly to estimated risk, while other factors provide some additional adjustments. It serves as an educational demonstration rather than a diagnostic system.

Conclusion

This analysis concludes that age and vascular health (hypertension and heart disease) are 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 aging, metabolic imbalance and vascular pathology.