**An In-Depth Analysis of the Stroke Dataset**

Peter La

Regis University

MSDS 670: Data Visualization

Professor John Koenig

March 12, 2025

**An In-Depth Analysis of the Stroke Dataset**

According to the World Health Organization (WHO), stroke is the second leading cause of death worldwide, accounting for approximately 11% of all deaths. This report examines key factors associated with stroke risk, using comprehensive visual techniques and statistical methods to identify significant predictors of stroke occurrence. Through various visualization techniques, this analysis offers clear insights into how demographic attributes and health indicators - such as age, glucose levels, and other lifestyle factors - contribute to the likelihood of experiencing a stroke.

**Correlation Heatmap for the Stroke Dataset**

A graph with different colored squares

AI-generated content may be incorrect.

The correlation heatmap above shows the relationships between variables in the dataset. Age, hypertension, heart disease, and elevated glucose levels all positively correlate with stroke risk, indicating that older individuals and those with these conditions are more prone to strokes. BMI and smoking status also have mild correlations with stroke risk. In contrast, factors like gender and work type show minimal correlation. However, it is important to keep in mind that correlation does not imply causation. Strokes typically result from various interacting factors, making moderate correlations common. Further statistical analyses can provide deeper insights into these relationships.

**Distribution of Age by Stroke Status**

A graph of a graph

AI-generated content may be incorrect.

This histogram illustrates the age distribution of individuals, differentiating between those who experienced a stroke (represented by orange bars) and those who did not (represented by blue bars). The visualization clearly indicates that strokes predominantly occur among older individuals, as shown by the higher concentration of orange bars on the right side of the plot. Younger individuals rarely experience strokes, while the frequency significantly increases with advancing age. This information highlights how age affects the risk of having a stroke. As people get older, the chances of a stroke usually increase. The overlap in some age groups shows that while age is very important, it is not the only factor. Other elements like high blood pressure, heart disease, and lifestyle choices should be considered also.

**Age versus Glucose Level by Stroke Status**

A screen shot of a graph

AI-generated content may be incorrect.

The scatter plot illustrates the relationship between age and average glucose levels in relation to stroke occurrence. Red dots represent individuals who had a stroke, while blue dots indicate those who did not. The plot shows a trend where strokes are more common in older adults and is linked to higher average glucose levels. There is a notable overlap between stroke and non-stroke groups, suggesting a hypothesis that both age and elevated glucose levels are associated with an increased risk of stroke. Further analysis is needed to understand the combined impact of these variables on stroke likelihood.

**Predicting Stroke with Decision Tree Analysis**

A diagram of a network

AI-generated content may be incorrect.

The decision tree is a visualization tool that helps identify the most important factors for predicting stroke risk. Each box represents a decision point, starting with the most critical factor, which branches into secondary factors based on answers. The color coding indicates stroke outcomes: blue for no stroke and orange for stroke. The depth of the tree is limited for clarity, focusing on key predictors like age and glucose levels. Although these factors are strong predictors on their own, stroke risk is influenced by multiple interacting variables, pointing to its complex nature.

**Conclusion**

The stroke dataset analysis identified several key factors influencing stroke risk. Age was the primary determinant, with older individuals at a significantly higher risk. Elevated glucose levels, hypertension, heart disease, and body mass index (BMI) also correlate positively with stroke incidence. These findings highlight the importance of comprehensive assessments and model predictions for identifying individuals at high risk and developing targeted preventive strategies.