Comprehensive Analysis of Stroke Contributing Factors and Demographics

Group B

Prepared By:

Parker Williams, Alex Ball, Braum Russell, Tré Gonzales, Skarlett Espinoza-Melgar, and Ashley Meza

Dr. Heshan-Sun

Price College of Business, *The* University *of* Oklahoma

MIT 5742 Data Science & Analytics

March 6, 2024

Introduction

Stroke is a critical health concern worldwide, with significant implications for individuals, families, and healthcare systems. Understanding the factors that contribute to stroke risk is essential for prevention. This report delves into a comprehensive Stroke Prediction Dataset obtained from Kaggle, curated by user “federsoriano”. Our analysis aims to uncover insights into the factors associated with an increased risk of stroke. Leveraging numerical and categorical data, we explore patterns, correlations, and potential predictors of stroke occurrence. To facilitate our analysis, we utilized data visualization tools and statistical techniques. Additionally, we address data cleaning procedures to ensure the integrity and reliability of our findings.

Through this report, we seek to contribute to the knowledge surrounding stroke risk assessment and prevention. By identifying key determinants and risk factors, we aim to provide insight into proactive measures to mitigate stroke risk and promote better health outcomes.

Problem Statement

In 2021, one out of every six deaths related to cardiovascular disease were due to strokes, underlining the significant toll strokes continue to take on individuals and healthcare systems globally (Center for Disease Control and Prevention, 2023). Additionally, every 40 seconds, someone in the United States has a stroke, further highlighting the urgency of addressing this pressing public health issue. Accordingly, strokes remain a significant challenge for individuals, families, and healthcare systems worldwide, necessitating a thorough understanding of the multifaceted factors that contribute to their occurrences.

Despite considerable research efforts, gaps persist in our understanding of the intricate interplay between demographic, lifestyle, and clinical variables in determining stroke risk. Notably, nearly one in three U.S. adults has high blood pressure, yet nearly one-third of these individuals are unaware of their condition due to the lack of symptoms. High blood pressure stands as the number one modifiable risk factor for stroke, contributing not only to strokes but also to heart attacks, heart failure, kidney failure, and atherosclerosis (American Stroke Association).

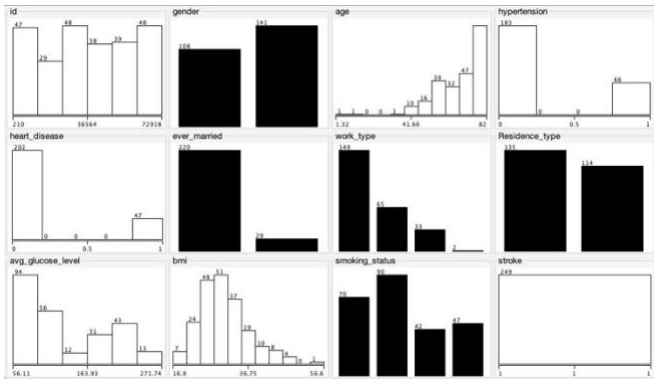
The problem at hand lies in the complexity of stroke risk assessment, intensified by the heterogeneous nature of risk factors and the limited generalizability of existing studies across diverse populations. Moreover, while large-scale datasets such as the Stroke Prediction Dataset offer a wealth of information, extracting meaningful insights requires overcoming methodological challenges related to data processing and predictive modeling.

Therefore, the overarching problem addressed in this study is to bridge these gaps by conducting an analysis of the Stroke Prediction Dataset. This involves delving deep into the dataset to uncover hidden patterns, exploring potential interactions among predictor variables, and developing robust predictive models for stroke occurrence. By addressing these challenges, this research aims to contribute to the refinement of stroke risk assessment methodologies and the development of targeted prevention strategies. Ultimately, the goal is to empower healthcare practitioners, educators, and individuals with actionable insights to mitigate the burden of stroke and improve health outcomes.

Description of Data

Our dataset is a Stroke Prediction Dataset pulled from Kaggle, was published by user "fedesoriano". This dataset contains 5110 rows of patient data, within these 5110 rows 249 of the patients suffered from stroke. It contains 12 different columns consisting of ratio & interval numerical data, as well as some categorical data. The columns consist of different parameters including gender, age, hypertension, heart disease, average glucose level, BMI, and smoking status. We can take this data, analyze it, and see what factors might lead to an increased risk of stroke.

We had to clean up some data before visualizing the dataset in Weka, as shown below. Since there were over 5000 entries of patient data, but so few had recorded a stroke, we wanted to see solely the data of patients who had a stroke. We can see things like a histogram of the BMI, average glucose level, the number of males and females, etc. When doing some of our analysis, however, we wanted to use the whole data set, as opposed to just patients who suffered a stroke.



Procedure

Before diving into analyzing our data, we first had to comb through our data to better understand it for analysis. Based on what columns the data contained and how they were related, we conducted our report based on those facts. We cleaned the data so that it would be easier to

analyze because any issues with data integrity would skew or nullify our results. The data was mostly fine, except we had to create multiple columns with ones or zeroes that described certain characteristics about the patients. These categories that warranted a one included if they had smoked before or currently smokes, if their age was above or equal to 68, if their BMI is greater than or equal to 30, etc. Otherwise, these categories were given a zero if the contrary were true. Hypertension or a history of heart disease were other categories that were interpreted using zeroes and ones. This data cleaning allowed us to present the data in clusters for better interpretation. We performed two different analyses using Association Rules and Cluster Analysis, which we learned during class. This analysis would provide us with results that would give a conclusion based on the data categories and what relationships were proven. Cluster analysis would give us clusters where we can see group patterns of patients together to find similar categories. Finally, we would determine conclusions based on the data given and what relationships correlated based on our analysis.

Results

Our first method to assess the relation of stroke risk to our attributes involved performing a clustering analysis. Clustering is a technique in data mining used to partition data points into groups or clusters based on their similarities. We employed the SimpleKMeans algorithm using WEKA software to identify distinct groups within our dataset for our clustering analysis. We considered attributes such as age, hypertension, heart disease, ever married status, average glucose level, BMI, smoking status, and residence type to determine the centralized points associated with stroke occurrences. After running the clustering algorithm, we obtained five clusters, each representing a unique group of individuals with different attribute profiles. To interpret the results, we examined the centroid values for each cluster, focusing particularly on

the centroid values for the "stroke" attribute. These centroid values provided insights into the likelihood of stroke occurrence within each cluster. Our analysis revealed that Cluster 2 exhibited the highest centroid value for the "stroke" attribute, indicating that individuals within this cluster are most likely to have experienced a stroke compared to individuals in other clusters. Attributes such as older age, high hypertension, high heart disease, being married, and elevated glucose levels were prevalent in Cluster 2, suggesting potential correlations with stroke occurrence. Conversely, other clusters exhibited lower centroid values for the "stroke" attribute, indicating lower likelihoods of stroke occurrence within those groups.

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(3372.0)	(1320.0)	(569.0)	(509.0)	(442.0)	(532.0)
gender	0.4214	0.4061	0.4218	0.442	0.4231	0.438
age	0.1696	0	0.0861	0.9882	0.0294	0.0132
hypertension	0.0988	0.0485	0.1054	0.279	0.0181	0.1109
heart_disease	0.0558	0.0159	0.0334	0.2417	0.0068	0.0414
ever_married	0.6625	0.672	0.7979	0.9607	0	0.7594
Residence_type	0.4958	0.6606	0	0.5265	0	1
avg_glucose_level	0.1803	0.125	0.1916	0.3517	0.0792	0.2256
bmi	0.3719	0	1	0.3006	0	1
smoking_status	0.1518	0.1629	0.181	0.1022	0.1018	0.1823
stroke	0.0445	0.0189	0.0299	0.1749	0.0113	0.0263

For our second method of analysis, we used association rules. Association rules are a method of data mining where you can determine relationships between certain variables. We used association rules to determine the relationships between hypertension, heart disease, if a patient was ever married, and stroke. We used the python libraries NLTK and Pandas to analyze the variables and determine association. When first running the python code we were given a list of 21 different possible relationships, however, not every result given back to us was a significant relationship. The three main statistics we need to look at when analyzing this data are support, confidence, and lift. From the list of 21 relationships, we filtered out the relationships that give us a lift greater than 1.0, and a confidence level of greater than .6. We can see from the

filtered relationships that the greatest relationship is between hypertension and that a patient was married. This shows that approximately 89% of the patients who suffered from hypertension were also married. The lift of this relationship is 1.36, showing the degree of association between the two. In simple terms, patients who suffered from hypertension were about 1.3 times more likely to have been married.

```
resultsDF=rules[(rules['lift']>=1.0)&(rules['confidence']>=0.6)]
```

```
resultsDF
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
3	(hypertension)	(ever_married)	0.097456	0.656164	0.087084	0.893574	1.361815	0.023137	3.230757	0.294374
7	(heart_disease)	(ever_married)	0.054012	0.656164	0.047750	0.884058	1.347312	0.012309	2.965582	0.272499
8	(stroke)	(ever_married)	0.048728	0.656164	0.043053	0.883534	1.346513	0.011079	2.952244	0.270523
12	(hypertension, heart_disease)	(ever_married)	0.012524	0.656164	0.010763	0.859375	1.309695	0.002545	2.445053	0.239462
17	(stroke, hypertension)	(ever_married)	0.012916	0.656164	0.010763	0.833333	1.270007	0.002288	2.063014	0.215385

Conclusions

The dataset we chose to analyze provides valuable insights into various factors that are associated with strokes. Using the analysis techniques learned in our classes, we identified significant risk factors that can trigger a stroke. These are relationship status, health issues, and older age. However, it is important to recognize the limitations of our dataset, such as potential data biases, missing variables, and limited sample sizes. Continued research can help further expand our results and improve our understanding of strokes. Overall, our findings from the data showcase the importance of early detection and lifestyle adaptations.

References

Centers for Disease Control and Prevention. (2023). Stroke Facts. Retrieved from https://www.cdc.gov/stroke/facts.htm

Fedesoriano. (2021). Stroke Prediction Dataset. Retrieved from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

University of Maryland Medical System. (n.d.). Stroke Facts. Retrieved from https://www.umms.org/sjmc/health-services/stroke/facts