CRITICAL EXPLORATORY DATA ANALYSIS OF THE STROKE PREDICTION DATASET

Muhammad ‘Ariful Furqon1*, Nina Fadilah Najwa2, Mohamad Zarkasi3, Priza Pandunata4 dan Gama Wisnu Fajariyanto5

Informatika (Universitas Jember, Jember, Indonesia) 1,5
Sistem Informasi (Politeknik Caltex Riau, Pekanbaru, 28265, Indonesia) 2
Teknologi Informasi (Universitas Jember, Jember, Indonesia) 3,4

ariful.furqon@unej.ac.id 1, nina@pcr.ac.id 2, mohamad.zarkasi@unej.ac.id 3, priza@unej.ac.id 4, gamawisnuf@unej.ac.id 5

*Penulis Koresponden

ABSTRAK


Kata kunci: Analisis Data Eksploratori, Stroke, Analisis Statistik Deskriptif, Faktor Risiko

ABSTRACT

Stroke is a significant global health issue, requiring a profound understanding of the complex factors contributing to its occurrence. Age, body mass index (BMI), and average glucose levels have been identified as key factors in stroke etiology. This study employs exploratory data analysis techniques, including descriptive statistical analysis such as univariate and multivariate analysis, to explore the relationships among variables in a stroke prediction dataset. Through descriptive statistical analysis, insights into the composition and variability of the dataset are obtained. Based on exploratory data analysis, significant relationships between age, hypertension, heart disease, average glucose levels, and stroke are found. However, the role of BMI in stroke shows a lower significance level. These findings make a significant contribution to the understanding of factors contributing to stroke risk. Comparison of findings with previous research indicates consistency with previous findings regarding the relationship between age, hypertension, and heart disease with stroke risk.

Keywords: Exploratory Data Analysis, Stroke, Statistical Descriptive Analysis, Risk Factor

Histori Artikel:
Diserahkan: 2 Mei 2024 Diterima setelah Revisi: 9 Mei 2024 Diterbitkan: 14 Juni 2024

https://doi.org/10.35143/jkt.v10i1.6307
1. INTRODUCTION

Stroke is characterized by a sudden loss of neurological function due to decreased blood supply to the brain [1]. Blockage, constriction, or rupture of the blood arteries that supply the brain can reduce blood supply to the brain [2]. According to the World Health Organization, a stroke is characterized by the fast development of focal and global neurologic impairments that can be severe and endure for at least 24 hours or cause death without an apparent cause other than vascular [3]. Two varieties of stroke, namely: (1) hemorrhagic stroke, which is caused by bleeding [4], [5], and (2) ischemic stroke, which is caused by an obstruction of blood supply to the brain [2], [6], [7].

In Indonesia, stroke is one of the non-infectious diseases that causes the highest mortality after heart disease and cancer [8]. According to the Indonesian Stroke Foundation, the number of stroke patients in Indonesia in 2012 was 200 out of a total population of one million, with 2.5% dying and the remaining patients suffering from severe or minor disability [9]. According to a study by [10], the risk of stroke impacted 4,884 of 13,605 individuals over 20 years, and 69 (1.4%) experienced strokes. In addition, 18.37% of the research participants had a history of hypertension, and 5.68% had a stroke. A person suffering from a stroke cannot engage in social activities freely. The multifactorial nature of stroke makes effective and efficient treatment unavailable.

Understanding the complex interplay of factors contributing to stroke incidence is paramount for effective prevention and management strategies [11]. While the underlying mechanisms of stroke are multifactorial, certain risk factors have been consistently identified in the literature [12]. These include hypertension, elevated body mass index (BMI), cardiovascular disease, and average glucose level, each contributing to the pathological processes that culminate in stroke onset [12], [13], [14], [15]. However, the relative impact of these risk factors on stroke incidence within the population still needs to be better understood, necessitating focused investigation. Exploratory data analysis offers a method to scrutinize the relationship among stroke risk factors.

Exploratory data analysis is a statistical approach used to examine, summarize, and visualize data sets to understand their underlying structure, patterns, and relationships [16], [17]. Exploratory data analysis offers a powerful means of uncovering patterns and relationships within complex datasets, providing valuable insights into disease etiology and epidemiology [18]. Exploratory data analysis systematically explores variables such as hypertension, BMI, cardiovascular disease, and average glucose levels to uncover patterns and relationships contributing to stroke onset [19]. Visualization techniques like histograms, scatter plots, and box plots facilitate the identification of distributions and trends within the data [20]. Moreover, correlation analysis enables quantification of the relationships between different risk factors and stroke incidence [19].

To bridge this gap, this study employs exploratory data analysis on a comprehensive dataset provided by the World Health Organization (WHO). The aim is to uncover patterns and relationships relevant to the incidence of stroke, particularly focusing on the predictive variables of age, BMI, and average glucose level in influencing stroke outcomes. By conducting thorough analysis and visualization techniques, this study aims to provide valuable insights that can inform the development of targeted interventions to reduce the burden of stroke.

2. METHODS

In this study, Exploratory data analysis is a powerful tool for uncovering insights into stroke risk factors using the Stroke Prediction Dataset. The dataset, which encompasses 5,110 rows and 12 columns, as shown in Table 1, offers a rich source of information for investigating factors associated with stroke. The stroke dataset was structured into a data frame using the Pandas library in Python to facilitate comprehensive analysis.
The dataset utilized in this study comprises 11 features and one binary target variable, sourced from a comprehensive survey conducted among individuals to investigate stroke incidence. Each feature provides valuable insights into demographic, clinical, and lifestyle-related factors potentially associated with stroke risk. A brief overview of the dataset features is provided in Table 2.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Identification number assigned to each individual patient</td>
</tr>
<tr>
<td>gender</td>
<td>Gender classification of the patient</td>
</tr>
<tr>
<td>age</td>
<td>Age of the patient</td>
</tr>
<tr>
<td>hypertension</td>
<td>Binary indicator denoting whether the patient has been diagnosed with hypertension</td>
</tr>
<tr>
<td>heart_disease</td>
<td>Binary indicator indicating whether the patient has been diagnosed with heart disease</td>
</tr>
<tr>
<td>ever_married</td>
<td>Binary attribute signifying whether the patient is married or not</td>
</tr>
<tr>
<td>work_type</td>
<td>Categorization of the patient's occupation or employment status</td>
</tr>
<tr>
<td>residence_type</td>
<td>Categorization of the patient's residence as urban or rural</td>
</tr>
<tr>
<td>avg_glucose_level</td>
<td>Numerical representation of the patient's average glucose level in the blood</td>
</tr>
<tr>
<td>bmi</td>
<td>Body mass index of the patient</td>
</tr>
<tr>
<td>smoking_status</td>
<td>Categorization of the patient's smoking habits</td>
</tr>
<tr>
<td>stroke</td>
<td>Binary target variable indicating whether the patient has experienced a stroke event</td>
</tr>
</tbody>
</table>

This dataset offers a comprehensive perspective on factors influencing stroke incidence, encompassing individual characteristics and health-related parameters. Including demographic, clinical, and lifestyle attributes facilitates a holistic analysis of stroke risk factors, informing evidence-based interventions and predictive modeling strategies [21]. The dataset was subjected to a comprehensive statistical description analysis to gain insights into the characteristics and distributions of both categorical and numerical variables. The descriptive analysis aimed to provide a foundational understanding of the dataset's composition and variability, guiding subsequent data preprocessing and analysis [22]. The statistical descriptive analysis for numerical data is provided in Table 3 and Table 4.

<table>
<thead>
<tr>
<th>Statistical Analysis</th>
<th>id</th>
<th>age</th>
<th>hypertension</th>
<th>heart_disease</th>
<th>avg_glucose_level</th>
<th>bmi</th>
<th>stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>5110.0</td>
<td>5110.0</td>
<td>5110.0</td>
<td>5110.0</td>
<td>5110.0</td>
<td>4909.0</td>
<td>5110.0</td>
</tr>
<tr>
<td>mean</td>
<td>36517.83</td>
<td>43.23</td>
<td>1.0</td>
<td>0.05</td>
<td>5110.0</td>
<td>28.89</td>
<td>0.05</td>
</tr>
<tr>
<td>std</td>
<td>21161.72</td>
<td>22.61</td>
<td>0.23</td>
<td>0.23</td>
<td>105.18</td>
<td>7.85</td>
<td>0.22</td>
</tr>
<tr>
<td>min</td>
<td>67.00</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>10.30</td>
<td>10.30</td>
<td>0.00</td>
</tr>
<tr>
<td>25%</td>
<td>17741.25</td>
<td>25.00</td>
<td>0.00</td>
<td>0.00</td>
<td>77.24</td>
<td>23.50</td>
<td>0.00</td>
</tr>
<tr>
<td>50%</td>
<td>36932.00</td>
<td>45.00</td>
<td>0.00</td>
<td>0.00</td>
<td>91.88</td>
<td>28.10</td>
<td>0.00</td>
</tr>
<tr>
<td>75%</td>
<td>54682.00</td>
<td>61.00</td>
<td>0.00</td>
<td>0.00</td>
<td>114.09</td>
<td>33.10</td>
<td>0.00</td>
</tr>
<tr>
<td>max</td>
<td>72940.00</td>
<td>82.00</td>
<td>1.00</td>
<td>1.00</td>
<td>271.74</td>
<td>97.60</td>
<td>1.00</td>
</tr>
</tbody>
</table>

https://doi.org/10.35143/jkt.v10i1.6307
The dataset under scrutiny exhibits various numerical and categorical variables, each offering distinct insights into the population's characteristics. Analysis of numerical variables reveals a diverse demographic, with participants averaging approximately 43 years of age. There were notable standard deviations across variables such as average glucose level and BMI, indicative of considerable variability within the population. Binary variables such as hypertension and heart disease showcase prevalence rates of approximately 10% and 5%, respectively, while stroke incidence appears at around 5%. Categorically, each feature encompasses 5,110 observations, with varying degrees of uniqueness and frequency across categories. The predominance of females, married individuals, and those employed in private sectors underscores the dataset's demographic composition. Urban residency and a prevalence of non-smokers further characterize the population.

Visualizing the nullity of the dataset provides a comprehensive overview of missing data patterns within the dataset, which is presented in Figure 1.

From the visualization of missing values, as depicted in Figure 1, it is evident that a considerable portion of missing values is observed in the "bmi" feature. This observation underscores the importance of handling missing data in the "bmi" feature during the data preprocessing stage. A median imputation technique was employed to address this issue. Median imputation involves replacing missing values with the median value of the non-missing entries in the same feature [23]. Additionally, binning was applied to all continuous values for feature extraction. Binning involves grouping continuous numerical data into discrete intervals or bins, simplifying the data, and capturing underlying patterns [24]. By applying binning to features such as age, average glucose level, and BMI, the dataset is structured into meaningful categories that facilitate analysis and interpretation.

The Exploratory data analysis on the Stroke Prediction Dataset employed a multifaceted approach to uncover insights into stroke risk factors and their potential predictive value [19]. Initially, descriptive statistics were computed to summarize the dataset's numerical and categorical variables, providing a foundational understanding of their distributions and central tendencies.

### Table 4. Statistical descriptive analysis of categorical data

<table>
<thead>
<tr>
<th>Statistical Analysis</th>
<th>gender</th>
<th>ever_married</th>
<th>work_type</th>
<th>residence_type</th>
<th>smoking_status</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>5110</td>
<td>5110</td>
<td>5110</td>
<td>5110</td>
<td>5110</td>
</tr>
<tr>
<td>unique</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>top</td>
<td>Female</td>
<td>Yes</td>
<td>Private</td>
<td>Urban</td>
<td>never smoked</td>
</tr>
<tr>
<td>freq</td>
<td>2994</td>
<td>3353</td>
<td>2925</td>
<td>2596</td>
<td>1892</td>
</tr>
</tbody>
</table>

From the visualization of missing values, as depicted in Figure 1, it is evident that a considerable portion of missing values is observed in the "bmi" feature. This observation underscores the importance of handling missing data in the "bmi" feature during the data preprocessing stage. A median imputation technique was employed to address this issue. Median imputation involves replacing missing values with the median value of the non-missing entries in the same feature [23]. Additionally, binning was applied to all continuous values for feature extraction. Binning involves grouping continuous numerical data into discrete intervals or bins, simplifying the data, and capturing underlying patterns [24]. By applying binning to features such as age, average glucose level, and BMI, the dataset is structured into meaningful categories that facilitate analysis and interpretation.

The Exploratory data analysis on the Stroke Prediction Dataset employed a multifaceted approach to uncover insights into stroke risk factors and their potential predictive value [19]. Initially, descriptive statistics were computed to summarize the dataset's numerical and categorical variables, providing a foundational understanding of their distributions and central tendencies.
Visualization techniques, including histograms, box plots, and scatter plots, were then utilized to visually explore relationships, identify patterns, and detect outliers within the data [25]. Correlation analysis was employed to quantify the strength and direction of associations between stroke incidence and risk factors such as age, average glucose levels, and BMI. Interpreting these findings involved understanding the magnitude and direction of these relationships and considering potential confounding variables and biases. Additionally, the relative importance of each risk factor in predicting stroke incidence may have been assessed, providing insights into the critical drivers of stroke within the dataset [12].

3. RESULTS AND DISCUSSIONS

The exploratory data analysis on the Stroke Prediction Dataset employed a comprehensive approach to unveil insights into potential stroke risk factors and their predictive significance [19]. The initial step in this analysis involved univariate analysis, which focused on examining individual variables within the dataset to understand their distributions and inherent characteristics [26]. By scrutinizing each variable independently, we gained valuable insights into their prevalence, variability, and potential relevance to stroke. Univariate analysis of the target variable of stroke is presented in Figure 2.

![Figure 2](image-url)

Figure 2. Univariate analysis on stroke target class

The bar plot in Figure 2, depicting the distribution of stroke incidence, illustrates a notable class imbalance, with a higher frequency of non-stroke cases than stroke cases. Specifically, the plot indicates that the dataset contains a relatively minor number of instances where the stroke occurred compared to instances where it did not. After conducting univariate analysis on the target variable, the next step is to examine the distribution of various features within the stroke dataset [27]. This analysis involves exploring the distributions of individual variables among individuals who have experienced a stroke and those who have not. Figure 3 visualizes the distribution plot of two key factors in the dataset: BMI and average glucose level.

![Figure 3](image-url)

Figure 3. Distribution plot of (a) BMI and (b) average glucose level
These distribution plots provide a visual representation of the spread and central tendency of BMI and glucose level values within the dataset, allowing for a quick understanding of their distribution characteristics and potential insights into their relationships with stroke incidence. The analysis of BMI, average glucose level, and age revealed notable differences between individuals with and without stroke are presented in Figure 4.

![Figure 4. Distribution analysis of key factors: (a) BMI, (b) average glucose level, and (c) age categorized by patients with and without stroke.](image)

The distribution analysis shown in Figure 4 reveals compelling insights into the relationship between various factors and stroke incidence. Firstly, a higher density of overweight individuals among those who have experienced a stroke suggests a potential correlation between elevated BMI and increased stroke risk. Secondly, a notable density of individuals with a glucose level below 100 is observed among stroke cases, hinting at a potential association between lower average glucose levels and heightened stroke susceptibility. Lastly, it indicates a higher density of individuals aged above 50 among stroke cases, indicating advancing age as a potential risk factor for stroke occurrence.

![Figure 5. Correlation heatmap of key features.](image)

Based on the correlation heatmap shown in Figure 5 presents a comprehensive overview of the pairwise Pearson correlation coefficients among age, average glucose level, BMI, and stroke variables. Firstly, age demonstrates a moderate positive correlation with both BMI and stroke, indicating that older individuals tend to have higher BMI values and may be more prone to
experiencing a stroke. Additionally, average glucose level exhibits weak positive correlations with age and stroke, suggesting a mild association with these variables. Conversely, BMI displays only a weak positive correlation with age and a weak correlation with stroke, indicating a less pronounced relationship. Furthermore, the target variable stroke shows weak positive correlations with age and average glucose level but a weak correlation with BMI.

The next stage was to visualize the pairwise relationships between age, average glucose level, and BMI in patients with and without stroke. This visualization technique allows for the simultaneous exploration of multiple variables, providing insights into potential patterns and differences between the two groups [28]. A pairplot shown in Figure 6 visualizes the pairwise relationships between age, average glucose level, BMI, and stroke variables.

![Figure 6. Correlation heatmap of key features](image)

The pairplot offers a detailed visual exploration of the relationships and distributions of age, average glucose level, and BMI, categorized by stroke status. Upon examination, several notable observations emerge. Firstly, in the comparison of age versus average glucose level, while no clear relationship is evident overall, stroke patients appear to skew towards older age groups with higher glucose levels. Similarly, the comparison between age and BMI reveals no distinct correlation, yet stroke patients tend to exhibit older age brackets without significant differences in BMI compared to non-stroke individuals. Furthermore, the comparison of average glucose
level versus BMI displays no apparent relationship, although stroke patients consistently show higher glucose levels, regardless of their BMI. The diagonal plots further reinforce these findings, indicating that stroke patients generally trend towards older ages and higher glucose levels, while their BMI distribution aligns closely with non-stroke individuals.

The analysis of the stroke prediction dataset revealed several significant findings regarding the predictive factors associated with stroke incidence. Firstly, it was noted that the target variable, stroke, exhibited a substantial class imbalance, with most instances indicating no stroke compared to those indicating stroke. This imbalance necessitated careful consideration when selecting machine learning models and evaluation metrics [29]. Furthermore, categorical variables such as gender, hypertension, and heart disease displayed varied distributions, with hypertension and heart disease being more prevalent among patients who had a stroke. Conversely, continuous variables such as age and average glucose level were found to be significantly higher in stroke patients, suggesting their potential as strong predictors of stroke incidence. However, the analysis did not reveal a significant difference in BMI between stroke and non-stroke patients. This lack of significance suggests that BMI may not be a robust predictor for stroke incidence in this dataset [30].

These findings have important implications for clinical practice and stroke prevention guidelines. Incorporating age, hypertension, heart disease, and average glucose level into risk assessment models may enhance their predictive accuracy and assist healthcare professionals in identifying individuals at higher risk of stroke. Conversely, the lack of significance of BMI suggests that it may not be a reliable indicator for stroke risk assessment in isolation. Linking these results to the objectives stated earlier in the paper reinforces the coherence and relevance of the study. The aim of the study was to identify and evaluate predictive factors for stroke incidence. The findings provide valuable insights that can inform the development of more accurate predictive models for stroke risk assessment and prevention strategies.

In summary, the analysis highlights the importance of considering factors such as age, hypertension, heart disease, and average glucose level in stroke prediction models while emphasizing the limited predictive value of BMI. By addressing the imbalance in the dataset and incorporating these findings into clinical practice, healthcare professionals can potentially improve stroke risk assessment and prevention efforts.

4. CONCLUSIONS

The exploratory data analysis conducted on the stroke prediction dataset has provided valuable insights into stroke-related factors. Age, hypertension, heart disease, and average glucose level have emerged as significant predictors, indicating their potential importance in stroke risk assessment. The weaker correlation of BMI with stroke incidence raises questions about its statistical significance and implications for stroke risk prediction. While BMI may still contribute to stroke risk, its relative weakness compared to other factors like age, hypertension, and heart disease suggests that prioritizing these stronger predictors could enhance the accuracy of predictive models. Future longitudinal studies and clinical trials should delve into various stroke risk factors to deepen the understanding and inform personalized preventive strategies. When addressing class imbalance in datasets, oversampling and under-sampling techniques offer distinct advantages and disadvantages. Oversampling preserves all original data points but may lead to overfitting, while under-sampling reduces computational complexity but may result in information loss. Preferred techniques depend on dataset characteristics and analysis goals, with a balanced approach involving a combination of methods often yielding the best results, guided by rigorous evaluation through techniques like cross-validation.

https://doi.org/10.35143/jkt.v10i1.6307
5. REFERENCES


https://doi.org/10.35143/jkt.v10i1.6307

