JKT Jurnal Komputer Terapan
ISSN : 2443-4159 (Print) ISSN: 2460-5255 (Online)
Politeknik Caltex Riau
Empower You to Global Competition

CRITICAL EXPLORATORY DATA ANALYSIS OF THE STROKE PREDICTION DATASET

Muhammad 'Ariful Furqon[1*], Nina Fadilah Najwa[2], Mohamad Zarkasi [3], Priza Pandunata[4] dan Gama Wisnu Fajariyanto[5]

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], mohammad.zarkasi@unej.ac.id [3], priza@unej.ac.id [4], gamawisnuf@unej.ac.id [5]

Penulis Koresponden

ABSTRAK

Stroke merupakan masalah kesehatan global yang signifikan serta membutuhkan pemahaman yang mendalam tentang faktor-faktor kompleks yang berkontribusi pada terjadinya stroke. Usia, indeks massa tubuh (IMT), dan rata-rata tingkat glukosa telah diidentifikasi sebagai faktor kunci dalam etiologi stroke. Penelitian ini menggunakan teknik analisis data eksploratori, mencakup analisis statistik deskriptif seperti analisis univariat dan multivariat untuk mengeksplorasi hubungan antar variabel dalam dataset prediksi stroke. Melalui analisis deskriptif statistik, diperoleh wawasan tentang komposisi dataset dan variabilitasnya. Berdasarkan analisis data eksploratori, ditemukan hubungan yang signifikan antara usia, hipertensi, penyakit jantung, rata-rata tingkat glukosa, dan stroke. Namun, peran IMT terhadap stroke menunjukkan tingkat signifikansi yang lebih rendah. Temuan ini memberikan kontribusi penting untuk pemahaman tentang faktor-faktor yang berkontribusi pada risiko stroke. Perbandingan temuan dengan penelitian terdahulu menunjukkan konsistensi dengan temuan sebelumnya tentang hubungan antara usia, hipertensi, dan penyakit jantung dengan risiko stroke.

Kata kunci: *Anlisis 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

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.

Table 1. Stroke prediction dataset

id	gender	age	hypertension	heart_disease	…	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	…	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	...	202.21	NaN	never smoked	1
31112	Male	80.0	0	1	...	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	...	171.23	34.4	smokes	1
1665	Female	79.0	1	0	...	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...
18234	Female	80.0	1	0	...	83.75	NaN	never smoked	0
44873	Female	81.0	0	0	...	125.20	40.0	never smoked	0
19723	Female	35.0	0	0	...	82.99	30.6	never smoked	0
37544	Male	51.0	0	0	...	166.29	25.6	formerly smoked	0
44679	Female	44.0	0	0	...	85.28	26.2	Unknown	0

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 2. Stroke prediction dataset description

No.	Feature	Description
1.	id	Identification number assigned to each individual patient
2.	gender	Gender classification of the patient
3.	age	Age of the patient
4.	hypertension	Binary indicator denoting whether the patient has been diagnosed with hypertension
5.	heart_disease	Binary indicator indicating whether the patient has been diagnosed with heart disease
6.	ever_married	Binary attribute signifying whether the patient is married or not
7.	work_type	Categorization of the patient's occupation or employment status
8.	residence_type	Categorization of the patient's residence as urban or rural
9.	avg_glucose_level	Numerical representation of the patient's average glucose level in the blood
10.	bmi	Body mass index of the patient
11.	smoking_status	Categorization of the patient's smoking habits
12.	stroke	Binary target variable indicating whether the patient has experienced a stroke event

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 and categorical data is described in Table 3 and Table 4.

Table 3. Statistical descriptive analysis of numerical data

Statistical Analysis	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.00	5110.00	5110.0	5110.00	5110.00	4909.00	5110.00
mean	36517.83	43.23	0.1	0.05	106.15	28.89	0.05
std	21161.72	22.61	0.3	0.23	45.28	7.85	0.22
min	67.00	0.08	0.0	0.00	55.12	10.30	0.00
25%	17741.25	25.00	0.0	0.00	77.24	23.50	0.00
50%	36932.00	45.00	0.0	0.00	91.88	28.10	0.00
75%	54682.00	61.00	0.0	0.00	114.09	33.10	0.00
max	72940.00	82.00	1.0	1.00	271.74	97.60	1.00

Table 4. Statistical descriptive analysis of categorical data

Statistical Analysis	gender	ever_married	work_type	residence_type	smoking_status
count	5110	5110	5110	5110	5110
unique	3	2	5	2	4
top	Female	Yes	Private	Urban	never smoked
freq	2994	3353	2925	2596	1892

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.

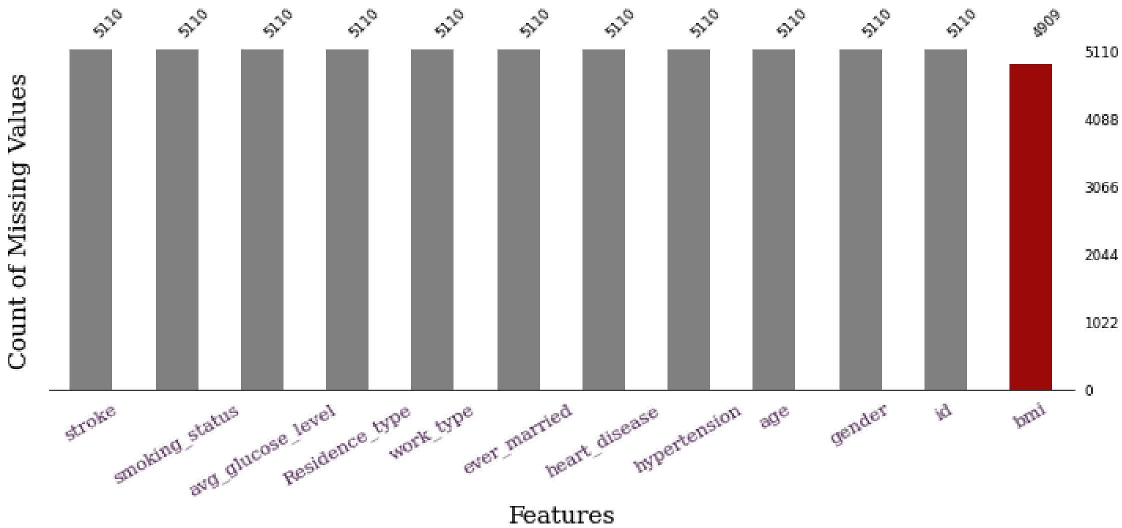


Figure 1. Visualization of missing values in the dataset

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

[22]. 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.

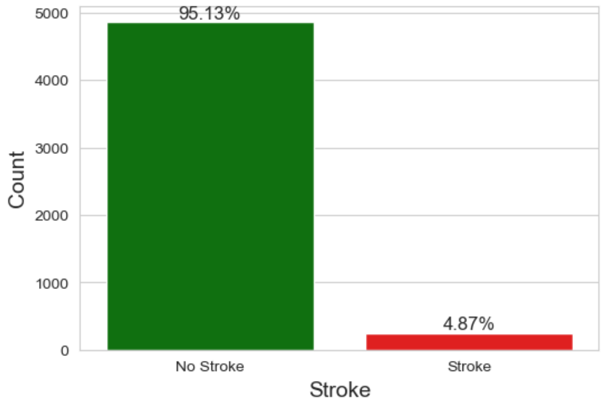
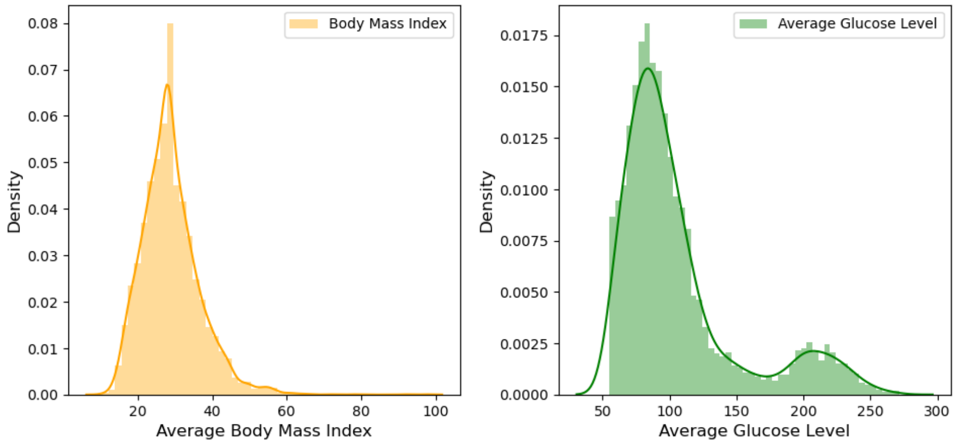


Figure 2. Univariate analysis on stroke target class

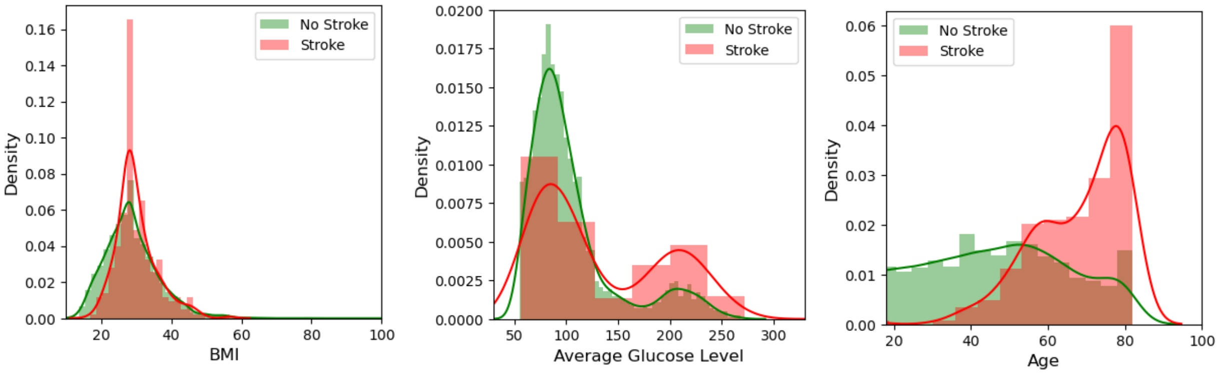
The bar plot in Figure 2, depicting the distribution of stroke incidence, illustrates a notable class imbalance, with a higher frequency of non-stroke cases than stroke cases. Specifically, the plot indicates that the dataset contains a relatively minor number of instances where the stroke occurred compared 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.



(a) **(b)**

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.



(a) **(b)** **(c)**

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

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.

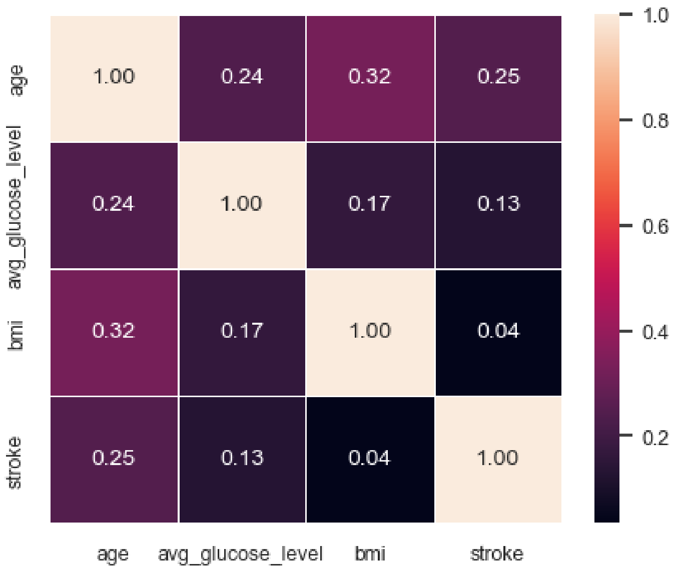


Figure 5. Correlation heatmap of key features

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.

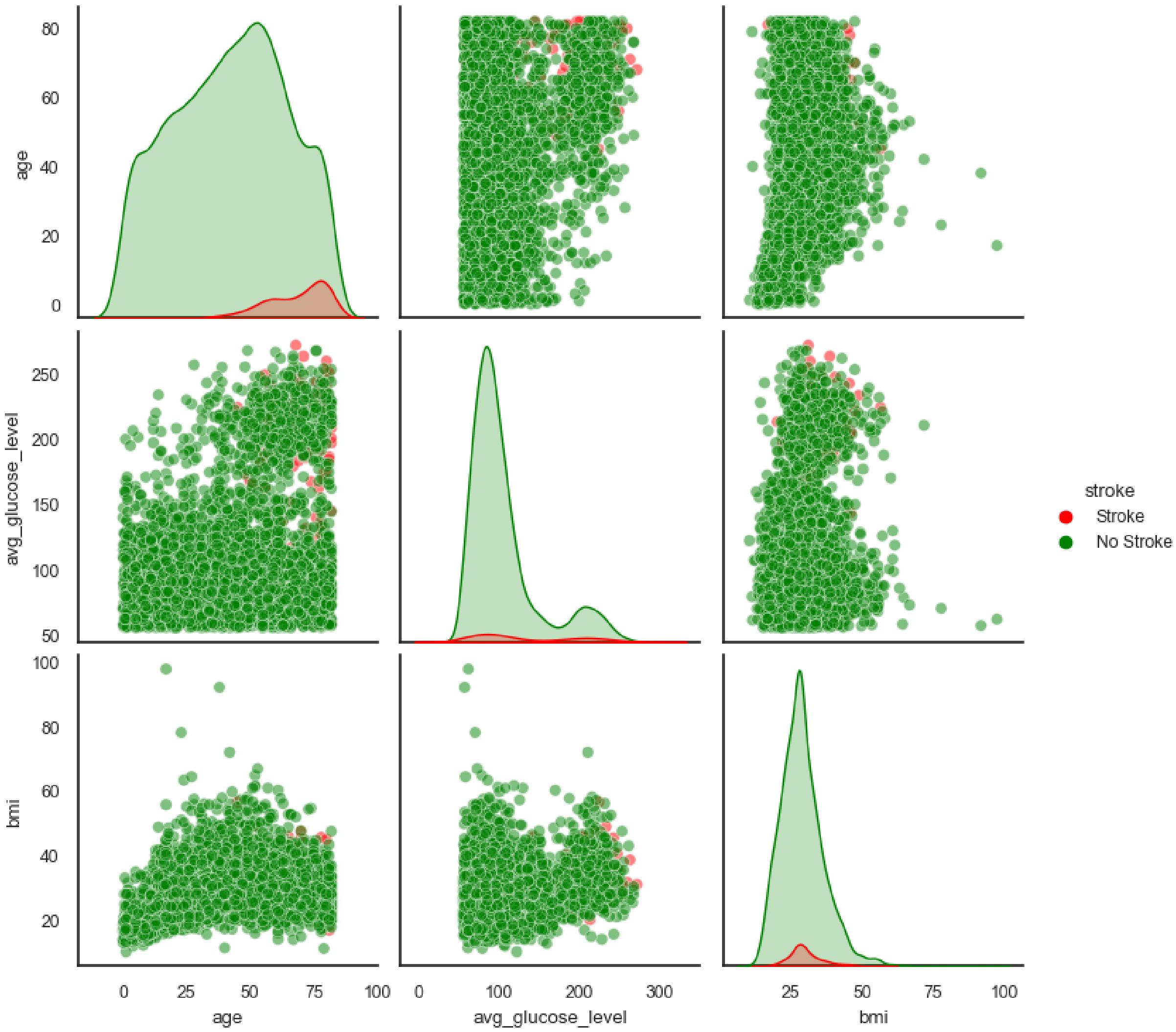


Figure 6. Correlation heatmap of key features

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.

5. REFERENCES

[1] E. de Robertis, O. Piazza, and G. Servillo, "The role of ventricular stroke work in daily clinical practice," *Minerva Anestesiol*, vol. 76, no. 11, 2010, [Online]. Available: https://www.researchgate.net/publication/47384756

[2] G. J. Del Zoppo and J. M. Hallenbeck, "Advances in the Vascular Pathophysiology of Ischemic Stroke," *Thromb Res*, vol. 98, no. 3, pp. 73–81, May 2000, doi: 10.1016/S0049-3848(00)00218-8.

[3] W. Johnson, O. Onuma, M. Owolabi, and S. Sachdev, "Stroke: A global response is needed," *Bulletin of the World Health Organization*, vol. 94, no. 9. World Health Organization, pp. 634A-635A, Sep. 01, 2016. doi: 10.2471/BLT.16.181636.

[4] A. K. A. Unnithan, J. M Das, and P. Mehta, "Hemorrhagic Stroke," *StatPearls*, Jul. 2020, Accessed: Feb. 08, 2023. [Online]. Available: http://europepmc.org/books/NBK559173

[5] S. D. Smith and C. J. Eskey, "Hemorrhagic Stroke," *Radiologic Clinics*, vol. 49, no. 1, pp. 27–45, Jan. 2011, doi: 10.1016/J.RCL.2010.07.011.

[6] S. K. Feske, "Ischemic Stroke," *Am J Med*, vol. 134, no. 12, pp. 1457–1464, Dec. 2021, doi: 10.1016/J.AMJMED.2021.07.027.

[7] S. A. Randolph, "Ischemic Stroke," *Workplace Health Saf*, vol. 64, no. 9, p. 444, Sep. 2016, doi: 10.1177/2165079916665400.

[8] W. Riyadina, J. Pradono, D. Kristanti, and Y. Turana, "Stroke in Indonesia: Risk factors and predispositions in young adults," *J Cardiovasc Dis Res*, vol. 11, no. 2, pp. 178–183, 2020, doi: 10.31838/jcdr.2020.11.02.30.

[9] N. R. Wati, E. Husna, S. Prima, and N. Bukittinggi, "Analisis Faktor Yang Berhubungan Dengan Kejadian Stroke Pada Penderita Stroke di Ruang Rawat Inap C Lantai 1 dan 2 RSSN Bukittinggi Tahun 2016," 2018.

[10] I. Setyopranoto, H. F. Bayuangga, A. S. Panggabean, S. Alifaningdyah, L. Lazuardi, F. S. T. Dewi, and R. G. Malueka, "Prevalence of stroke and associated risk factors in sleman district of Yogyakarta Special Region, Indonesia," *Stroke Res Treat*, vol. 2019, 2019, doi: 10.1155/2019/2642458.

[11] A. Guzik and C. Bushnell, "Stroke Epidemiology and Risk Factor Management," *CONTINUUM Lifelong Learning in Neurology*, vol. 23, no. 1, pp. 15–39, Feb. 2017, doi: 10.1212/CON.0000000000000416.

[12] A. K. Boehme, C. Esenwa, M. S. V Elkind, M. Fisher, C. Iadecola, and R. Sacco, "Stroke Risk Factors, Genetics, and Prevention," *Circ Res*, vol. 120, no. 3, pp. 472–495, Feb. 2017, doi: 10.1161/CIRCRESAHA.116.308398.

[13] A. Alloubani, A. Saleh, and I. Abdelhafiz, "Hypertension and diabetes mellitus as a predictive risk factors for stroke," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 12, no. 4, pp. 577–584, Jul. 2018, doi: 10.1016/J.DSX.2018.03.009.

[14] S. Juvela, J. Siironen, and J. Kuhmonen, "Hyperglycemia, excess weight, and history of hypertension as risk factors for poor outcome and cerebral infarction after aneurysmal

subarachnoid hemorrhage," *J Neurosurg*, vol. 102, no. 6, pp. 998–1003, Jun. 2005, doi: 10.3171/JNS.2005.102.6.0998.

[15] J. Z. Willey, Y. P. Moon, E. Kahn, C. J. Rodriguez, T. Rundek, K. Cheung, R. L. Sacco, and M. S. V. Elkind, "Population attributable risks of hypertension and diabetes for cardiovascular disease and stroke in the Northern Manhattan study," *J Am Heart Assoc*, vol. 3, no. 5, Sep. 2014, doi: 10.1161/JAHA.114.001106.

[16] W. L. Martinez, A. R. Martinez, and J. Solka, *Exploratory data analysis with MATLAB*. Chapman and Hall/CRC, 2017.

[17] U. Singh, M. Hur, K. Dorman, and E. S. Wurtele, "MetaOmGraph: a workbench for interactive exploratory data analysis of large expression datasets," *Nucleic Acids Res*, vol. 48, no. 4, pp. e23–e23, Feb. 2020, doi: 10.1093/NAR/GKZ1209.

[18] L. Kharb, A. Tyagi, and D. Chahal, "Meta-analysis Review International Journal of Current Research and Review Exploratory Data Analysis on the Epidemiology of Coronavirus (COVID-19) Pandemic Outbreak," *Int J Cur Res Rev* |, vol. 13, p. 12, 2021, doi: 10.31782/IJCRR.2021.SP170.

[19] R. Kaur, K. Hambarde, R. George, A. Hussain, C. Gomkar, and S. Sonawani, "Stroke Prediction using Optimization and Exploratory Data Analysis," *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2022*, 2022, doi: 10.1109/IATMSI56455.2022.10119295.

[20] M. Furqon, S. M. S. Nugroho, R. F. Rachmadi, A. Kurniawan, I. K. E. Purnama, and M. H. S. B. Aji, "Arrhythmia Classification Using EFFICIENTNET-V2 with 2-D Scalogram Image Representation," in *2021 TRON Symposium (TRONSHOW)*, 2021, pp. 1–9.

[21] E. S. Sintiya, A. Kusumawardana, M. A. Furqon, N. F. Najwa, A. C. Puspitaningrum, and A. S. Afrah, "SARIMA and Holt-Winters Seasonal Methods for Time Series Forecasting in Tuberculosis Case," in *2020 4th International Conference on Vocational Education and Training (ICOVET)*, 2020, pp. 1–5.

[22] S. E. Kemp, M. Ng, T. Hollowood, and J. Hort, "Introduction to Descriptive Analysis," *Descriptive Analysis in Sensory Evaluation*, pp. 1–39, Dec. 2017, doi: 10.1002/9781118991657.CH1.

[23] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data 2021 8:1*, vol. 8, no. 1, pp. 1–37, Oct. 2021, doi: 10.1186/S40537-021-00516-9.

[24] N. Yadav and N. Badal, "Data preprocessing based on missing value and discretisation," *International Journal of Forensic Software Engineering*, vol. 1, no. 2/3, p. 193, 2020, doi: 10.1504/IJFSE.2020.110584.

[25] L. Wilkinson, "Visualizing Big Data Outliers Through Distributed Aggregation," *IEEE Trans Vis Comput Graph*, vol. 24, no. 1, pp. 256–266, Jan. 2018, doi: 10.1109/TVCG.2017.2744685.

[26] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, "Statistical explorations and univariate timeseries analysis on COVID-19 datasets to understand the trend of disease spreading and death," *Sensors*, vol. 20, no. 11, p. 3089, 2020.

[27] M. Abzalov and M. Abzalov, “Exploratory data analysis,” *Applied Mining Geology*, pp. 207–219, 2016.

[28] M. Valera, R. K. Walter, B. A. Bailey, and J. E. Castillo, “Machine learning based predictions of dissolved oxygen in a small coastal embayment,” *J Mar Sci Eng*, vol. 8, no. 12, p. 1007, 2020.

[29] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, “Data imbalance in classification: Experimental evaluation,” *Inf Sci (N Y)*, vol. 513, pp. 429–441, 2020.

[30] Y. Zan, W. Xiong, X. Zhang, Y. Han, C. Cao, H. Hu, Y. Wang, and H. Ou, “Body mass index has a non-linear association with three-month outcomes in men with acute ischemic stroke: An analysis based on data from a prospective cohort study,” *Front Endocrinol (Lausanne)*, vol. 13, p. 1041379, Dec. 2022, doi: 10.3389/FENDO.2022.1041379/BIBTEX..