Journal of Information Systems and Informatics
Vol. 6, No. 1, March 2024 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v6i1.664 Published By DRPM-UBD

Comparative Analysis of KNN and Decision Tree Classification Algorithms for Early Stroke Prediction: A Machine Learning Approach

Karin Eldora[1], Erick Fernando[2*], Winanti[3]

[1,2]Information Systems, Faculty of Engineering and informatics, Universitas Multimedia Nusantara, Indonesia
[3]Information Systems, Faculty of Technology and Information, Universitas Insan Pembangunan, Tangerang, Indonesia
Email: erick.fernando@umn.ac.id

Abstract

Stroke is the second most deadly disease in the world and the third leading cause of disability. However, most deaths due to stroke can be prevented by recognizing the symptoms of stroke and taking preventive measures using information technology. Therefore, this research utilizes the role of information technology using a machine learning approach to predict stroke in a person using the K-Nearest Neighbor and Decision Tree classification methods. The two algorithms were compared to determine which algorithm was more effective in predicting stroke. Data analysis using the CRISP-DM approach was carried out using a dataset containing 5110 observations with 12 relevant attributes. Implementation of Exploratory Data Analysis (EDA) was also carried out for preprocessing, and oversampling techniques were applied to overcome the problem of unbalanced classes. The research results show that the predictive model with the highest level of accuracy was obtained at around 97.1845% using the K-Nearest Neighbor algorithm. This research makes a significant contribution to stroke prevention efforts through the use of information technology and machine learning algorithms for early identification of stroke risk.

Keywords: Early Stroke, Prediction, Machine Learning, Comparison Algorithms, Classification, KNN, Decision Tree

1. INTRODUCTION

A stroke is a medical condition characterized by the interruption of blood flow to a part of the brain due to either a blockage or the rupture of a blood vessel in the brain [1], [2]. According to the World Health Organization (WHO), stroke is a manifestation of nerve function impairment resulting from cerebrovascular diseases, primarily stemming from non-traumatic cerebral circulation disorders[3], [4], [5]. This grim reality underscores why stroke is ranked as the second-leading cause of mortality and the third-leading cause of disability globally. In 2017, the

Ministry of Health of the Republic of Indonesia highlighted that stroke stands as the foremost cause of both disability and death in Indonesia [6]. These statistics underscore the grave threat posed by stroke.

Stroke risk factors commonly include conditions like chronic hypertension, diabetes, elevated blood sugar levels (hyperglycemia), high cholesterol levels (hyperlipidemia), obesity, and high blood pressure, among others [2], [7]. To reduce the likelihood of experiencing a stroke, individuals are advised to adopt a healthy lifestyle, incorporating regular physical activity and a balanced diet rich in fruits and vegetables.

Medical consensus emphasizes the critical role of early detection and immediate, appropriate treatment in stroke prevention, as these measures can significantly reduce the extent of brain damage and prevent potential complications [2], [5], [8]. Timely intervention is paramount, as untreated strokes can lead to prolonged brain damage, long-term disabilities, or even fatal outcomes [9]. The sooner an individual receives medical attention following a stroke, the lower the likelihood of severe damage, thereby minimizing the overall impact of stroke-related fatalities. This underscores the paramount importance of our research, which leverages machine learning algorithms for early stroke prediction, aiming to facilitate timely interventions that can ultimately reduce the incidence of stroke-related deaths and disabilities, thus improving overall public health outcomes [9], [10], [11], [12].

Hence, there is an urgent and significant need for substantial contributions in the realm of early stroke prevention and treatment. Information technology plays a pivotal role in developing predictive models for stroke, capitalizing on the advancements in the medical field [9], [10], [11], [13], [14]. Utilizing Machine Learning techniques, particularly through the classification method [4], [15], [16], [17], offers a promising avenue for predicting stroke occurrences. This underscores the growing synergy between healthcare and technology, offering innovative solutions to enhance stroke risk assessment, early intervention, and ultimately, the reduction of stroke-related morbidity and mortality.

The primary objective of this research is to harness the power of information technology to construct a highly accurate predictive model for stroke outcomes using Machine Learning (ML) techniques, specifically employing the K-Nearest Neighbor and Decision Tree algorithms for data classification. The ultimate aim of this study is to furnish invaluable insights and actionable information for healthcare professionals, enabling them to deliver timely and essential care to individuals at risk of stroke. By leveraging technology and data-driven approaches, this research seeks to contribute significantly to the realm of stroke prevention, ultimately reducing the incidence and severity of stroke cases and enhancing overall public health outcomes.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2. METHODS

The research methodology employed in this study adheres to the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, which has been recognized and utilized since 1996 in Europe, according to the Ministry of Finance of the Republic of Indonesia[18], [19], [20]. The CRISP-DM framework comprises six distinct stages, with the initial three stages being potentially non-mutually exclusive, as per the author's experience. These six stages of CRISP-DM provide a structured and systematic approach to data analysis, facilitating the orderly progression of tasks and activities throughout the data mining process, as depicted in Figure 2.

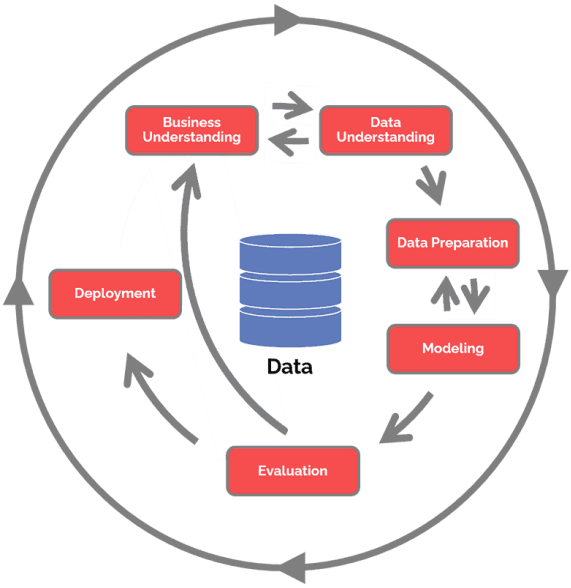

Figure 2. CRISP-DM Process Diagram

In this study, the steps that must be done in each stage of CRISP-DM to predict stroke are as follows:

2.1. Business Understanding

The first phase of the CRISP-DM methodology, involves the clear and concise articulation of business problems in a straightforward manner. In the context of this research, the focal issue pertains to the identification of factors contributing to stroke, a condition that can lead to an elevated risk of human mortality if effective preventive measures and early treatment strategies are not implemented.

2.2. Data Understanding

The second stage of the CRISP-DM process, involves the data analyst gaining a comprehensive comprehension of the dataset intended for use in the analysis. This critical step is pivotal in facilitating the research process and ultimately arriving at solutions to the identified problems. A thorough understanding of the data is essential for effective data analysis and the subsequent development of strategies to address the research objectives.

2.3. Data Preparation

The third step in the CRISP-DM framework, involves the meticulous processing of the designated dataset. During this stage, several essential tasks are undertaken, including data visualization, the assessment and removal of null values, addressing data outliers, conducting data formatting and normalization[21], as well as data encoding. These critical data preprocessing steps are carried out to ensure the dataset's quality, integrity, and suitability for subsequent analysis. This meticulous preparation is essential for facilitating accurate and effective data analysis as part of the research process.

2.4. Modeling

The fourth step in the CRISP-DM framework, is the stage where a combination of statistical methods and machine learning techniques is employed to construct predictive models. To ensure optimal results, a thorough understanding of the chosen algorithms is crucial. In this specific research, supervised learning techniques are utilized, employing a classification approach that incorporates both the K-Nearest Neighbor and Decision Tree algorithms. These carefully selected algorithms are leveraged to develop predictive models, enabling the research to attain its objectives effectively.

1) K-Nearest Neighbor

K-Nearest Neighbor (KNN) represents a supervised classification technique, where the classification of a newly introduced test sample is determined by the majority category among its K-nearest neighbors from the dataset[22], [23], [24]. In simpler terms, the K-Nearest Neighbor Algorithm is employed to classify objects by comparing them to the closest neighbors or those with minimal dissimilarity values within the training data [22]. The fundamental objective of this algorithm is to classify novel objects based on their attributes in relation to both the test data and the established training data. KNN serves

as a valuable tool in machine learning, aiding in tasks that involve categorizing data points based on their similarity to existing examples, making it a versatile approach for various classification problems.

2) Decision Tree

The Decision Tree, a machine learning algorithm, utilizes a tree-like structure for classification tasks. It represents information as a tree with nodes corresponding to attributes, branches indicating test outcomes, and leaves denoting class groups[4], [10], [11]. The root node, usually representing the most influential attribute, initiates a top-down search strategy for decision-making. When classifying new data, the algorithm traverses from the root node to a terminal node, predicting the data's class. Decision Trees are widely employed in machine learning for diverse classification tasks. Entropy is a crucial criterion in decision tree classifiers, measuring the purity of a node; lower entropy signifies greater class disparity [22]. The mathematical formula involves probabilities of data points belonging to specific classes. Another crucial parameter is the maximum depth of the tree, determining its extension limits. Balancing tree depth is vital for optimal accuracy, as excessively deep or shallow trees can impact performance. Finding the right depth is essential for assessing the accuracy and effectiveness of a decision tree classifier.

$$E(S) = \sum_{i=1}^{c} - P_i \, log_2 \, P_i \quad (1)$$

2.5. Evaluation

The fifth step of the CRISP-DM methodology, involves a comprehensive review of the results obtained in alignment with the initial business understanding and the effectiveness of the achieved outcomes. In this specific research, two distinct algorithms are employed, necessitating a comparative analysis of the models based on the utilized algorithms. This evaluation process is crucial for assessing the performance and reliability of the predictive models generated through the application of machine learning techniques, ensuring that they align with the research objectives and provide effective solutions.

Commonly used in the assessment of prediction model performance, F1 score, recall, and precision constitute essential evaluation metrics[16]. Recall, also referred to as sensitivity, serves as a crucial measure that assesses the accuracy of the true positive rate in a model's predictions. It is quantified using a specific formula for measurement. These evaluation metrics play a pivotal role in quantifying the effectiveness of predictive models, providing valuable insights into

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

their ability to correctly identify positive cases, minimize false negatives, and optimize overall prediction accuracy[16].

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

Precision, often referred to as confidence, is a crucial evaluation metric in machine learning that assesses the accuracy of correctly identified positive instances, emphasizing the model's ability to avoid false positives. It represents true positive accuracy and is computed using a specific formula. When used alongside recall and the F1 score, precision offers a comprehensive assessment of a predictive model's performance, providing insights into its capacity to accurately identify positive cases while minimizing false positives. These metrics collectively contribute to the assessment and refinement of machine learning algorithms, enhancing their precision and reliability across diverse applications [16].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

The F1 score, also known as F1, is a crucial metric in assessing model performance by combining both recall and precision measurements, providing an overall evaluation of test accuracy [16]. Calculated using a specific formula, it helps strike a balance between a model's ability to accurately identify positive instances (recall) and its capability to minimize false positives (precision). The F1 score offers a comprehensive assessment of a predictive model's accuracy and reliability by considering both aspects of its performance. Together with recall and precision, the F1 score plays a pivotal role in enhancing and optimizing machine learning algorithms, making it an indispensable tool for evaluating their effectiveness across various applications[16].

$$F1\ score = \frac{2\ x\ precision\ x\ recall}{precision + recall} \quad (3)$$

2.6. Deployment

The final step within the CRISP-DM framework, involves the practical implementation and utilization of the model, making it accessible for end-users.

In the context of this research, the predictive model developed serves as a valuable tool for medical staff, enabling them to engage in proactive measures for stroke prevention and early intervention. The deployment of this model facilitates its integration into the medical field, empowering healthcare professionals to effectively address the risk of stroke and provide timely c, thereby enhancing patient care and well-being.

2.6.1 Object of Research

This research is centered on the prediction of stroke occurrence. The research employs a dataset containing various parameters, including age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), and smoking status, for performing predictive modeling. The dataset comprises 5110 data points and 12 variables.

2.6.2 FlowMethods of Research

The methodology for this research involved several key steps. Firstly, the dataset was carefully selected from the Kaggle website and downloaded for analysis. Subsequently, the dataset was imported into Jupyter Notebook for further examination. The data cleaning and manipulation process followed, where irrelevant or inconsistent data was removed or refined to enhance its quality and facilitate analysis. Visualizations were created to provide comprehensive insights into the dataset, with each coding step meticulously executed to support the research's objectives. To classify the data effectively, the K-Nearest Neighbor (KNN) and Decision Tree algorithms were employed, guided by the insights gained from data visualization. A thorough comparison of the algorithms and their correlation with the coded data was conducted, leading to the formulation of conclusions. These conclusions represent the research's ultimate findings and serve as the study's response to the research question. Finally, the journal's conclusion section provides a closing summary of the research, encapsulating the key outcomes and insights obtained throughout the study.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

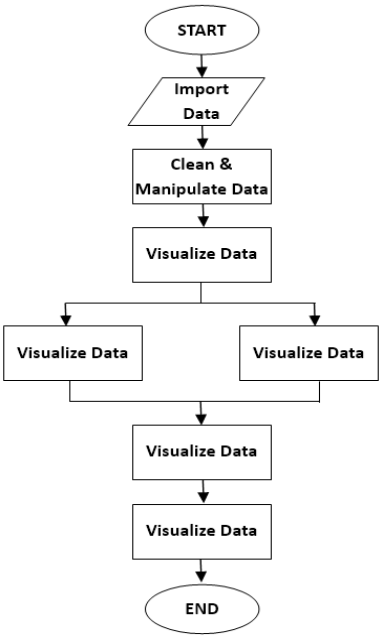

Figure 1. Flow Methods of Research

3. RESULTS AND DISCUSSION

This chapter will examine the development of a stroke prediction model utilizing the K-Nearest Neighbor and Decision Tree approach, building upon the stages identified in the preceding chapter.

3.1 Dataset Description

In this research, data was sourced from Kaggle, a cloud-based platform known for its collection and sharing of datasets tailored for data science applications. Specifically, the dataset utilized for this study is referred to as "healthcare-dataset-stroke-data," in file *.csv which can be accessed and downloaded through the following link: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Kaggle serves as a valuable resource for researchers and data scientists, providing access to diverse datasets that are instrumental in conducting comprehensive and data-driven investigations, as exemplified in this study focused on stroke prediction.

This dataset, last updated in 2020, comprises a total of 12 attributes and 5110 observations. The primary objective of incorporating this dataset into the research is to facilitate the prediction of whether a patient is at risk of experiencing a stroke,

predicated on various parameters including gender, age, hypertension, heart disease, marital status, occupation, residential type, average glucose level, body mass index, and smoking habits. Detailed descriptions of these attributes are provided in Table 1, offering a comprehensive foundation for the study's investigation into stroke prediction.

Table 1. Dataset Information

Attribute	Data Type	Description
id	Int	Unique id to identify each patients
gender	Categorical	Patient's gender: "male" / "female" / "other"
age	Int	Patient age
hypertension	Int	0 → doesn't have hypertension 1 → has hypertension
heart_ disease	Int	0 → don't have heart disease 1 → have a heart disease
ever_ married	Categorical	"Yes" → married "No" →married
work_type	Categorical	Patient's type of work
Residence_ type	Categorical	The patient's area of residence. (Urban/Rural)
avg_glucose_level	int	glucose level of patients at average
bmi	int	body mass index of patients
smoking_ status	categorical	Patient's smoking status. (formerly_smoked never_smoked smoked)
stroke	int	0 → stroke 1 → stroke

3.2 Data Analysis

Demographic Analysis from data the gender-based distribution of patients with and without strokes. The blue data points represent individuals without strokes, while the orange data points denote patients who have experienced strokes. Upon examining the graph in figure 2, it becomes apparent that women constitute the largest demographic of both stroke-free individuals and those who have suffered strokes. This observation underscores the importance of gender as a potential factor in stroke occurrence, warranting further analysis and consideration within the research.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

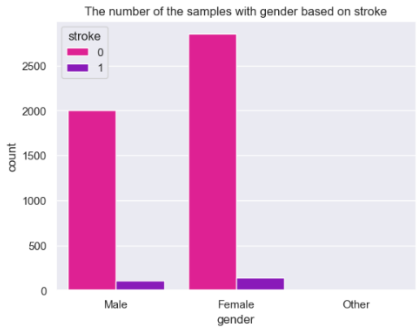

Figure 2. Countplot of Stroke and Non-Stroke Patients by Gender

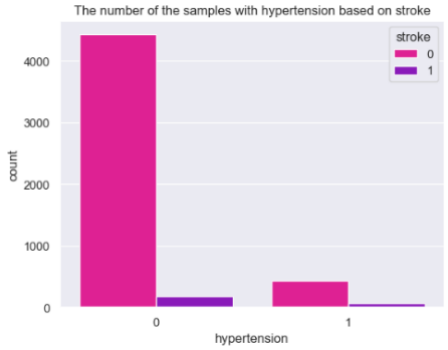

Figure 3. Countplot of Stroke and Non-Stroke Patients based on Hypertension

Figure 3 provides an insight into the distribution of patients with and without strokes based on the presence of hypertension. Within the graph, blue data points represent individuals without strokes, while orange data points represent those who have experienced strokes. Notably, the numerical values 0 and 1 on the graph signify the absence or presence of hypertension, respectively. A discernible pattern emerges, indicating that patients without hypertension constitute the largest group of individuals who have not suffered strokes, but paradoxically, they also comprise the highest number of patients who have experienced strokes. This complex relationship between hypertension and stroke occurrence calls for a more in-depth examination within the research.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

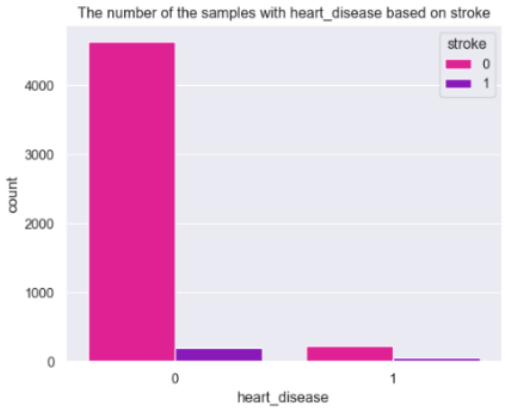

Figure 4. Countplot of Stroke and Non-Stroke Patients based on Heart Disease

Figure 4 offers a visual representation of the distribution of patients with and without strokes, categorized by the presence or absence of heart disease. Within the graph, blue data points signify individuals without strokes, while orange data points represent those who have experienced strokes. The numerical values 0 and 1 on the graph indicate the absence or presence of heart disease, respectively. A distinct pattern emerges, revealing that patients without heart disease form the largest group of individuals who have not suffered strokes, yet paradoxically, they also constitute the majority of patients who have experienced strokes. This intriguing relationship between heart disease and stroke incidence warrants further exploration within the research.

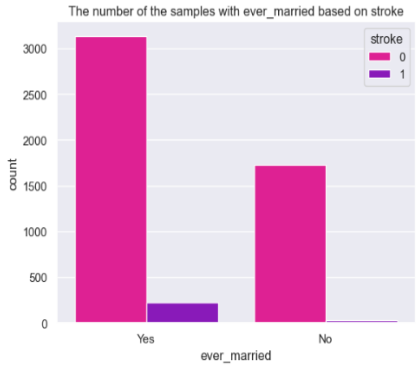

Figure 5. Countplot of Stroke and Non-Stroke Patients based on Marriage Status

Figure 5 provides an illustrative breakdown of stroke and non-stroke patients categorized by marital status. In the graph, blue data points represent individuals without strokes, while orange data points denote patients who have experienced

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

strokes. The terms "Yes" and "No" on the graph indicate the marital status of patients, where "Yes" signifies married individuals, and "No" indicates those who are not married. Notably, the graph reveals that patients who are married comprise both the largest group of individuals who have not suffered strokes and the majority of patients who have experienced strokes. This correlation between marital status and stroke incidence suggests a potential area of interest for further investigation within the research.

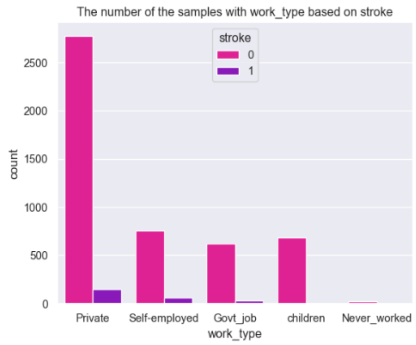

Figure 6. Countplot of Stroke and Non-Stroke Patients based on Work Type

Figure 6 offers an informative depiction of the distribution of stroke and non-stroke patients categorized by their work type. Within the graph, blue data points represent individuals without strokes, while orange data points indicate patients who have experienced strokes. The graph encompasses five distinct work types as parameters. Notably, the data reveals that patients with the work type "Private" constitute both the largest group of individuals who have not suffered strokes and the majority of patients who have experienced strokes. This association between work type and stroke incidence highlights the significance of occupational factors and warrants further examination within the research.

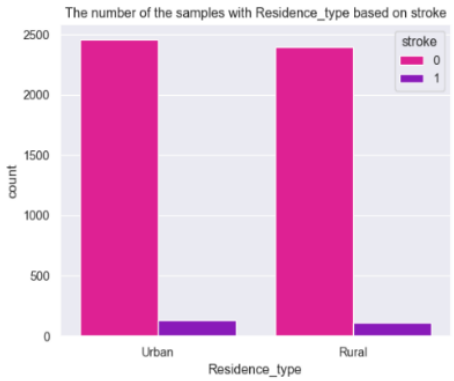

Figure 7. Countplot of Stroke and Non-Stroke Patients based on Residence Type

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 7 presents a visual representation of the distribution of stroke and non-stroke patients, categorized by their residence type. Blue data points signify individuals without strokes, while orange data points represent patients who have experienced strokes. In this context, the graph indicates two residence types: "Urban" and "Rural." Interestingly, the data shows that patients with the residence type "Urban" constitute both the largest group of individuals who have not suffered strokes and the majority of patients who have experienced strokes. This observation suggests a potential connection between residence type and stroke incidence, which merits further exploration within the research.

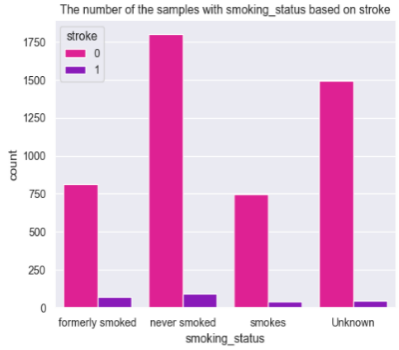

Figure 8. Countplot of Stroke and Non-Stroke Patients based on Smoking Status

Figure 8 provides an illustrative representation of the distribution of stroke and non-stroke patients, categorized by their smoking status. Blue data points denote individuals without strokes, while orange data points indicate patients who have experienced strokes. The graph distinguishes between two smoking status categories: "Never smoked" and "Smokes." Remarkably, the data shows that patients who have never smoked and do not have stroke disease constitute both the largest group of individuals who have not suffered strokes and the majority of patients who have experienced strokes. Conversely, patients who smoke and suffer from strokes exhibit the lowest distribution. This finding suggests a noteworthy association between smoking status and stroke incidence, underscoring the relevance of smoking habits in the context of stroke research.

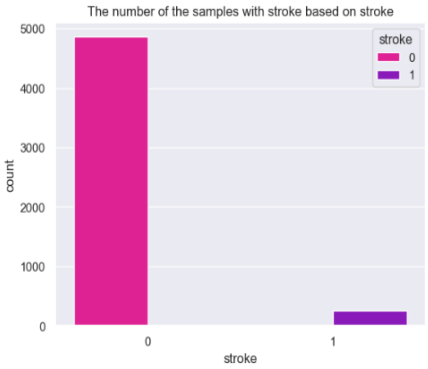

Figure 9. Countplot of Stroke and Non-Stroke Patients

Figure 9 visually presents the distribution of patients, distinguishing between those who have experienced strokes (represented by orange data points) and those who have not (represented by blue data points). One notable observation is the evident class imbalance in the stroke data classification, with a significantly larger number of patients labeled as not having suffered strokes. This class imbalance could potentially pose challenges in the development and evaluation of predictive models for stroke outcomes, as it may affect the model's ability to accurately predict the minority class, i.e., patients who have had strokes. Addressing class imbalance is a critical consideration in machine learning approaches to ensure the model's effectiveness and the reliability of stroke predictions[25], [26].

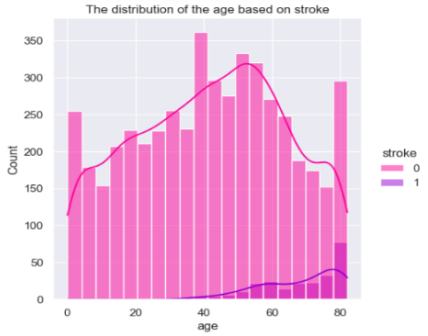

Figure 10. Displot of Distribution of Age based on Stroke

Figure 11 illustrates the distribution of stroke and non-stroke patients with respect to age. The graph reveals a notable trend, indicating an increase in the number of patients suffering from strokes as they age, with a noticeable rise starting from around 40 years of age and continuing up to approximately 80 years of age. This age-related pattern suggests that stroke incidence tends to be more prevalent

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

among older individuals, highlighting the importance of age as a significant factor in stroke risk assessment and prediction.

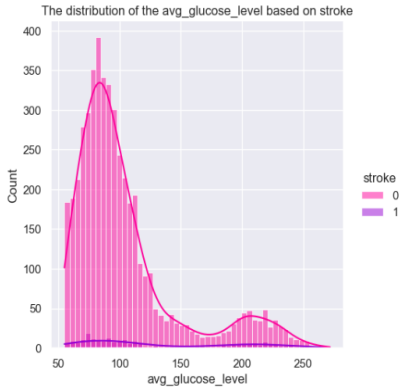

Figure 12. Displot of Distribution of Average Glucose Level based on Stroke

Figure 12 depicts the distribution of stroke and non-stroke patients based on the average glucose level of patients. The graph does not exhibit a substantial increase or decrease in stroke patients' average glucose levels. However, it does reveal a notable decrease in the number of patients who do not have a stroke when their average glucose level falls within the range of 100 to 150 and between 150-200. This observation suggests that there might be some correlation between average glucose levels and stroke risk, particularly within these specified ranges. Further analysis is needed to explore this potential relationship and its significance in stroke prediction.

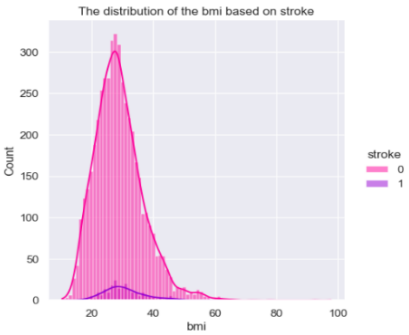

Figure 13. Displot of Distribution of BMI based on Stroke

Figure 13 illustrates the distribution of stroke and non-stroke patients based on a patient's BMI (Body Mass Index). The graph exhibits a distribution pattern that resembles a normal distribution, with values centered around a certain range. This suggests that BMI values among the patients in the dataset follow a typical

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

distribution pattern, which is essential information for understanding the relationship between BMI and stroke risk. Further analysis can explore whether specific BMI ranges are associated with a higher or lower likelihood of experiencing a stroke, contributing to the development of a predictive model.

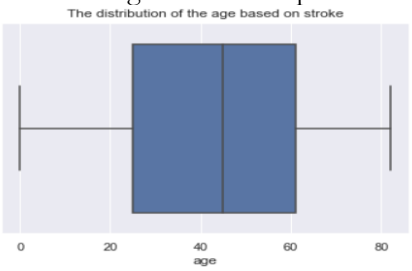

Figure 14. Boxplot of Distribution of Age based on Stroke

Figure 14 displays the distribution of stroke and non-stroke patients according to age. The graph indicates that there are no outliers within the age data, suggesting that the age values in the dataset are within a reasonable and expected range. This absence of outliers is a crucial observation as it ensures the data's reliability and consistency, allowing for a more accurate analysis of the relationship between age and stroke risk.

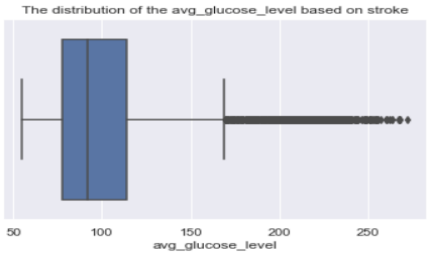

Figure 15. Boxplot of Distribution of Average Glucose Level based on Stroke

Figure 15 illustrates the distribution of stroke and non-stroke patients with respect to the average glucose level of each patient. Notably, this dataset contains outliers within the average glucose level data, indicating the presence of values that significantly deviate from the norm or exhibit extreme values. The identification of outliers is an essential step in data analysis as it can impact the overall accuracy and reliability of predictive models. Therefore, addressing these outliers appropriately is crucial to ensure the integrity of the analysis and the subsequent predictive model.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

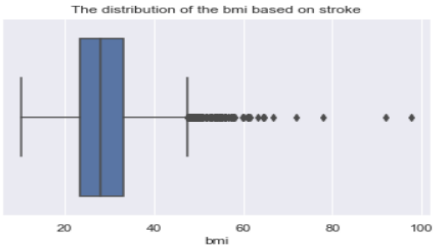

Figure 16. Boxplot of Distribution of BMI based on Stroke

Figure 16 presents the distribution of stroke and non-stroke patients concerning the body mass index (BMI) of each patient. In this distribution, it is evident that outliers are present within the data, signifying values that deviate significantly from the typical range or exhibit extreme values. The identification and handling of outliers in BMI data are critical aspects of data preprocessing, as they can impact the accuracy and reliability of predictive models. Addressing these outliers appropriately is essential to ensure the integrity of the analysis and the subsequent development of an effective predictive model.

3.3 Data Preprocessing

During the data preprocessing phase, the initial step involves data cleansing, which entails the removal of all null or missing values from the dataset. This essential data cleaning procedure ensures that the dataset is free from any incomplete or missing information[27], which can adversely affect the accuracy and reliability of subsequent analyses and predictive modeling. By eliminating these null values, the dataset becomes more suitable for robust and meaningful analysis, leading to more accurate stroke prediction models.

```
id                 0
gender             0
age                0
hypertension       0
heart_disease      0
ever_married       0
work_type          0
Residence_type     0
avg_glucose_level  0
bmi                0
smoking_status     0
stroke             0
dtype: int64
```

Figure 17. All Nulls Value Dropped

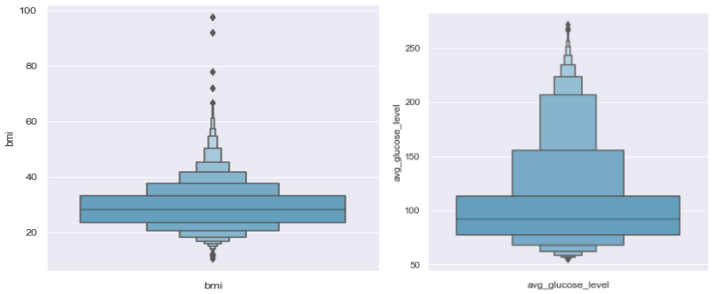

Figure 18. Outliers in the Distribution of bmi and avg_glucose_level

As depicted in Figure 18, it is evident that outliers are present in the distribution of both BMI and average glucose level data. Consequently, a preprocessing step is deemed necessary to address this issue, primarily involving the removal of these outliers. The method employed for outlier removal in this dataset is the Interquartile Range (IQR) method implemented in Python. Following the removal of outliers using this method, the data distributions for both BMI and average glucose level are modified, resulting in the distributions displayed in Figure 19. This preprocessing step helps enhance the quality and reliability of the data, ensuring that the subsequent analysis and predictive modeling are not unduly influenced by outliers.

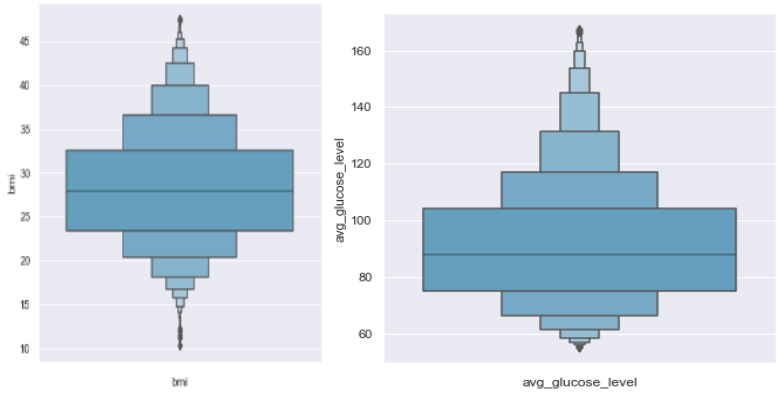

Figure 19. bmi and avg_glucose_level Distribution After Outliers are Removed Pre-Modeling

3.3.1 Encode data

Moving forward to the next stage of data preprocessing, the dataset is subjected to encoding to transform it into a format suitable for analysis, as demonstrated in Figure 22. The specific method applied in this stage is One Hot Encoding, primarily applied to the independent variables within the dataset. One Hot

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Encoding is employed to convert categorical variables into a numerical format that can be utilized effectively in machine learning algorithms[28]. This transformation ensures that the categorical attributes are appropriately represented for subsequent modeling and analysis, facilitating the development of predictive models for stroke prediction.

	age	hypertension	heart_disease	avg_glucose_level	bmi	gender_Female	gender_Male	gender_Other	ever_married_No	ever_married_Yes	...	work_typ
2836	14.0	0	0	83.42	28.7	1.0	0.0	0.0	1.0	0.0	...	
13	48.0	0	0	84.20	29.7	0.0	1.0	0.0	1.0	0.0	...	
5930	77.0	0	0	105.22	31.0	1.0	0.0	0.0	0.0	1.0	...	
7619	78.0	0	0	133.19	23.6	0.0	1.0	0.0	0.0	1.0	...	
7089	57.0	0	0	110.52	28.5	1.0	0.0	0.0	0.0	1.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
4829	50.0	1	0	73.18	30.3	1.0	0.0	0.0	0.0	1.0	...	
7291	60.0	0	0	89.22	37.8	1.0	0.0	0.0	1.0	0.0	...	
1344	50.0	0	0	85.92	37.3	1.0	0.0	0.0	0.0	1.0	...	
7293	82.0	0	1	144.90	26.4	0.0	1.0	0.0	0.0	1.0	...	
1289	53.0	0	0	113.40	35.1	0.0	1.0	0.0	0.0	1.0	...	

Figure 20. Hot Encode Data

3.3.2 Splitting the Data

To evaluate the performance of a machine learning model and ascertain whether it meets the expected criteria and desired level of accuracy, a crucial step is to conduct training and testing, necessitating the segregation of data into appropriate subsets. This data separation is achieved through the utilization of the train_test_split method, with specific parameters set to facilitate the process. Notably, a test_size of 0.5 is specified, indicating that half of the data will be allocated for testing purposes, while the random_state parameter is set to 10 to ensure reproducibility and consistency in the data splitting process. This division of data is essential for model validation, enabling the assessment of its predictive capabilities and performance against the expected benchmarks.

3.3.3 Standardize Data

Subsequently, an essential preprocessing step involves standardizing specific features, namely age, avg_glucose_level, and bmi, to ensure they adhere to a normal distribution pattern. This standardization process is accomplished through the application of the MinMaxScaler technique, applied to both the training and testing datasets. By performing this standardization, the data is rescaled to fall within a consistent range, preserving the integrity of the features and facilitating accurate model training and evaluation. Standardizing these critical attributes

contributes to the model's robustness and enhances its performance in predictive tasks.

3.3.4 Oversampling Data

Given the presence of imbalanced data classes in the distribution of stroke patients and those without strokes, addressing this issue is crucial at this stage. To rectify the class imbalance, oversampling techniques are employed, specifically utilizing the RandomOverSampler method[17]. Through this oversampling approach, the dataset is adjusted to achieve a balanced representation of both stroke and non-stroke cases, as illustrated in Figure 24. This rebalancing of data classes is imperative for enhancing the effectiveness and fairness of machine learning models, ensuring that the predictive model performs optimally across various classes and provides accurate insights for stroke prediction.

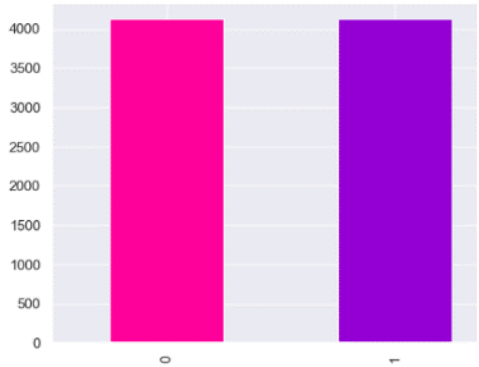

Figure 21. Balanced Data Class

3.4 Machine Learning Model Classification Result Analysis

3.4.1 KNN

In the initial phase of constructing the prediction model, the K-Nearest Neighbor (KNN) algorithm is employed. The KNN model is systematically evaluated across a range of neighbor values from 1 to 30. The objective is to identify the neighbor value that yields the lowest error rate, which will be chosen as the optimal parameter for the stroke prediction model. This thorough evaluation ensures that the KNN model is fine-tuned to provide accurate and reliable predictions for stroke outcomes.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

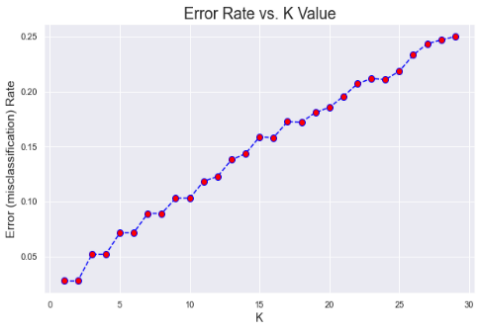

Figure 22. Error Rate vs K Value Between 1-30 Neighbor in KNN Model

Figure 25 displays a graph illustrating the error rates associated with various neighbor values in the K-Nearest Neighbor (KNN) algorithm. Upon examination, it becomes evident that the smallest error rate is observed at index 1. Consequently, index 1, corresponding to a single neighbor, is selected as the optimal parameter for the KNN algorithm in building the stroke prediction model. Subsequently, the outcomes of the KNN prediction model are presented in Figure 26, reflecting the model's performance using the selected neighbor value of 1. This meticulous evaluation process ensures the accuracy and effectiveness of the KNN-based stroke prediction model. The results of the KNN prediction model analysis have an accuracy value of 97.1% and a recall score of 100% and an F-Score value of 97.2%.

3.4.2 Decision Tree

The prediction model is constructed using the Decision Tree algorithm, with a specified maximum depth of 15. The choice of a maximum depth of 15 is a careful balance to ensure efficient model performance without unnecessary computational overhead. A larger maximum depth could lead to slower model execution. The Decision Tree model's visual representation is presented in Figure 23, providing a clear depiction of its structure and decision-making process. Subsequently, the prediction outcomes generated by the Decision Tree model are offering insights into its effectiveness in predicting stroke occurrences. This approach ensures that the Decision Tree-based prediction model operates efficiently while maintaining accuracy in its predictions[15]. The results of the decision tree algoritma prediction model analysis have an accuracy value of 90.4% and a recall score of 87.9% and an F-Score value of 90.2%.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

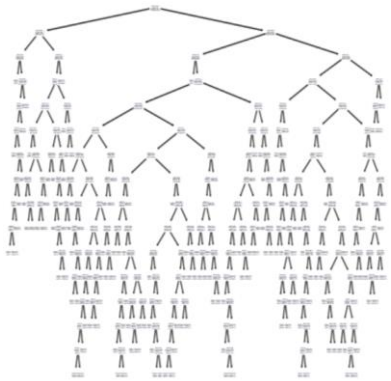

Figure 23. Decision Tree Visualization

3.5 Model Evaluation

Following the comprehensive evaluation of model performance using the two algorithms investigated in this study, it is evident that the K-Nearest Neighbor (KNN) algorithm stands out as the machine learning algorithm that yields the most accurate predictive model for addressing the specific problem at hand. The Decision Tree algorithm also demonstrates respectable model accuracy, although it falls slightly behind the KNN algorithm in terms of performance. It's worth noting that the data preprocessing step involved oversampling to balance the classification of data classes, which likely contributes to the overall robustness of the model evaluation metrics. The Confusion Matrix displayed in Figure 24 provides a visual summary of the KNN model's prediction outcomes, offering valuable insights into its predictive capabilities.

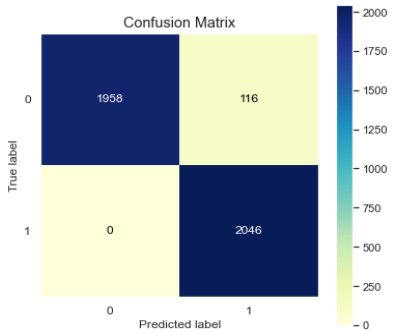

Figure 24. Confusion Matrix of KNN Model

The confusion matrix reveals that the model's predictions encompass 0 false negatives and 116 false positives, while correctly identifying 1958 true positives

and 2046 true negatives. This information provides a detailed breakdown of the model's performance in classifying data instances, particularly highlighting its ability to minimize false negatives and accurately detect true positives.

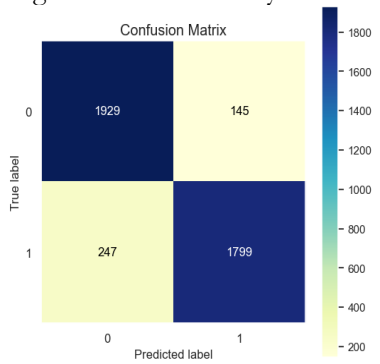

Figure 25. Confusion Matrix of Decision Tree Model

The confusion matrix illustrates the model's predictions, which include a total of 247 false negatives and 145 false positives, alongside correctly identifying 1929 true positives and 1799 true negatives. This comprehensive breakdown of the model's performance aids in assessing its ability to correctly classify instances, with a focus on minimizing false negatives and maximizing true positives.

3.6 Discussion of Comparison of KNN Model and Decision Tree Model

The graphical representation in Figure 26 indicates that the KNN model has performed exceptionally well in predicting stroke outcomes based on the dataset's parameters, achieving an impressive accuracy score of 0.971845 or 97.1845%. In contrast, the Decision Tree model exhibits a lower accuracy score of 0. 904854 or 90.4854%. Notably, the KNN prediction model demonstrates an exceptionally low false negative prediction rate of 0, underscoring its strong performance in correctly identifying cases of stroke. These results suggest that the KNN algorithm is highly effective in this predictive task.

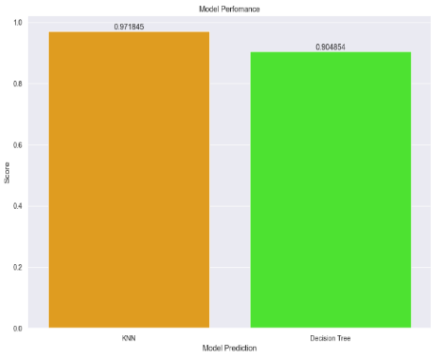

Figure 26. Comparison of KNN Model's Accuracy and Decision Tree Model's Accuracy

5. CONCLUSION

This research highlights the significant role of information technology, particularly the implementation of machine learning, within the healthcare sector. By leveraging data science and employing machine learning algorithms, this research has successfully developed a predictive model with the capability to forecast stroke outcomes based on a patient's health information. The potential application of this model in real-world scenarios, particularly in the healthcare industry, holds great promise for enabling medical professionals and doctors to enhance their stroke prevention and early treatment efforts. It underscores the transformative impact of data-driven approaches and machine learning in the medical field, offering a valuable tool to improve patient care and reduce the incidence of strokes through proactive interventions and timely treatments.

Furthermore, this research emphasizes the critical importance of selecting the most appropriate machine learning algorithm for early stroke prediction. The study's results clearly indicate that, among the tested classification algorithms, the K-Nearest Neighbor (KNN) algorithm outperformed the Decision Tree algorithm in terms of predictive accuracy for stroke outcomes. With an impressive accuracy score of 97.1845% for the KNN model compared to 90.4854% for the Decision Tree model, it is evident that the KNN algorithm excels in accurately identifying patients at risk of stroke. This research underscores the potential of data-driven approaches and machine learning in healthcare, particularly for early stroke prediction.

ACKNOWLEDGEMENTS

This research was supported/partially supported by Universitas Multimedia Nusantara.

REFERENCES

[1] B. W. Lee, J. Yoon, and S. J. Lee, "A ripple effect in prehospital stroke patient care," *Int. J. Prod. Res.*, vol. 59, no. 1, pp. 168–187, 2021, doi: 10.1080/00207543.2020.1825862.

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

[3] "World Stroke Day 2022." Accessed: Jan. 24, 2024. [Online]. Available: https://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-day-2022

[4] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 539–545, 2021, doi: 10.14569/IJACSA.2021.0120662.

[5] A. Mishra *et al.*, "Stroke genetics informs drug discovery and risk prediction

across ancestries," *Nature*, vol. 611, no. 7934, pp. 115–123, Nov. 2022, doi: 10.1038/s41586-022-05165-3.

[6] "Stroke Dapat Dicegah, Kenali Faktor Risiko dan Gejalanya - Penyakit Tidak Menular Indonesia." Accessed: Apr. 16, 2024. [Online]. Available: https://p2ptm.kemkes.go.id/post/stroke-dapat-dicegah-kenali-faktor-risiko-dan-gejalanya

[7] S. J. Mendelson and S. Prabhakaran, "Diagnosis and Management of Transient Ischemic Attack and Acute Ischemic Stroke," *JAMA*, vol. 325, no. 11, p. 1088, Mar. 2021, doi: 10.1001/jama.2020.26867.

[8] M. S. V. Elkind, A. K. Boehme, C. J. Smith, A. Meisel, and M. S. Buckwalter, "Infection as a Stroke Risk Factor and Determinant of Outcome After Stroke," *Stroke*, vol. 51, no. 10, pp. 3156–3168, Oct. 2020, doi: 10.1161/STROKEAHA.120.030429.

[9] Y. Wu and Y. Fang, "Stroke prediction with machine learning methods among older chinese," *Int. J. Environ. Res. Public Health*, vol. 17, no. 6, pp. 1–11, 2020, doi: 10.3390/ijerph17061828.

[10] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine Learning for Brain Stroke: A Review," *J. Stroke Cerebrovasc. Dis.*, vol. 29, no. 10, 2020, doi: 10.1016/j.jstrokecerebrovasdis.2020.105162.

[11] M. Daidone, S. Ferrantelli, A. Tuttolomondo, M. Daidone, and M. Daidone, "Machine learning applications in stroke medicine: Advancements, challenges, and future prospective," *Neural Regen. Res.*, vol. 19, no. 4, pp. 769–773, 2024, doi: 10.4103/1673-5374.382228.

[12] S. Castaneda-Vega *et al.*, "Machine learning identifies stroke features between species," *Theranostics*, vol. 11, no. 6, pp. 3017–3034, 2021, doi: 10.7150/THNO.51887.

[13] X. Feng *et al.*, "Intelligible Models for HealthCare: Predicting the Probability of 6-Month Unfavorable Outcome in Patients with Ischemic Stroke," *Neuroinformatics*, vol. 20, no. 3, pp. 575–585, 2022, doi: 10.1007/s12021-021-09535-6.

[14] M. L. Scholz, H. Collatz-Christensen, S. N. F. Blomberg, S. Boebel, J. Verhoeven, and T. Krafft, "Artificial intelligence in Emergency Medical Services dispatching: assessing the potential impact of an automatic speech recognition software on stroke detection taking the Capital Region of Denmark as case in point," *Scand. J. Trauma. Resusc. Emerg. Med.*, vol. 30, no. 1, pp. 1–17, 2022, doi: 10.1186/s13049-022-01020-6.

[15] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[16] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.

[17] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data

imbalance in classification: Experimental evaluation," *Inf. Sci. (Ny).*, vol. 513, pp. 429–441, Mar. 2020, doi: 10.1016/j.ins.2019.11.004.

[18] V. Plotnikova, M. Dumas, and F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Comput. Sci.*, vol. 6, pp. 1–43, 2020, doi: 10.7717/PEERJ-CS.267.

[19] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.

[20] J. S. Saltz and I. Krasteva, "Current approaches for executing big data science projects—a systematic literature review," *PeerJ Comput. Sci.*, vol. 8, pp. 1–24, 2022, doi: 10.7717/PEERJ-CS.862.

[21] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[22] M. Kubat, *An Introduction to Machine Learning*. 2021. doi: 10.1007/978-3-030-81935-4.

[23] D. Chopra and R. Khurana, *Introduction to Machine Learning with Python*. Singapore: Bentham Science, 2023.

[24] X.-S. Yang, *Introduction to Algorithms for Data Mining and Machine Learning*. Candice Janco, 2019.

[25] A. Viloria, O. B. Pineda Lezama, and N. Mercado-Caruzo, "Unbalanced data processing using oversampling: Machine Learning," *Procedia Comput. Sci.*, vol. 175, pp. 108–113, 2020, doi: 10.1016/j.procs.2020.07.018.

[26] C. Fernandez-Lozano *et al.*, "Random forest-based prediction of stroke outcome," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021, doi: 10.1038/s41598-021-89434-7.

[27] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean J. Anesthesiol.*, vol. 70, no. 4, p. 407, 2017, doi: 10.4097/kjae.2017.70.4.407.

[28] T. Al-Shehari and R. A. Alsowail, "An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques," *Entropy*, vol. 23, no. 10, p. 1258, Sep. 2021, doi: 10.3390/e23101258.