Comparison of Data Mining Classification Algorithms for Stroke Disease Prediction Using the SMOTE Upsampling Method

Ronald Sebastian[1*], Christina Juliane[2]

[1,2] *Program Studi Pascasarjana, Magister Sistem Informasi Bisnis STMIK LIKMI, Indonesia*

*corr_author: ronaldsebastian58@gmail.com

Abstract - Stroke is a circulation disorder in the brain that can cause symptoms and signs related to the affected part of the brain and is the leading cause of death and disability in Indonesia. Everyone is at risk of experiencing a stroke, and it is important to recognize and manage risk factors. Data Mining techniques can help in the extraction and prediction of information, as well as finding hidden patterns in stroke medical data. The dataset used in this research comes from Kaggle and is imbalanced, so the SMOTE Upsampling technique is used to address this imbalance issue. The results of the study conclude that the use of SMOTE technique in the C4.5, NB, and KNN algorithms can increase precision, recall, and AUC. The C4.5 algorithm and SMOTE technique as the best performing algorithm were selected for testing new data, and the results show that the model created can predict stroke risk more accurately than the C4.5 model without SMOTE. However, it should be noted that based on the author's interview with one of the medical practitioners, the model cannot be directly used in medical practice because the observations in the medical field to determine factors related to stroke are highly complex. Thus, a new understanding revealed that predicting stroke in a practical setting is highly complex. While data mining can be used as a predictive tool in the initial stage for predictions in the general population, it is strongly recommended to undergo direct examination by doctors in a hospital to obtain more accurate and comprehensive medical evaluations.

Keywords: SMOTE upsampling, K-Nearest Neighbour, Naïve Bayes, C4.5, stroke

I. INTRODUCTION

Stroke is the leading cause of death and disability in Indonesia. Everyone, regardless of age, is at risk of experiencing a stroke [1]. According to the World Health Organization (WHO): Task Force in Stroke and other Cerebrovascular Disease, stroke is an acute neurological dysfunction caused by a rapid (within seconds) or, at the latest, within hours, abnormality in blood circulation, with symptoms and signs related to the specific part of the brain affected [2]. Meanwhile, Riskesdas (2018) defines stroke as sudden, gradual, and rapid brain damage due to non-traumatic disturbances in brain blood circulation. This condition causes rapid symptoms, such as facial or limb paralysis, unclear or slurred speech, changes in consciousness, visual disturbances, and others. Globally, stroke has become more common in the last 20 years, as reported by Mukherjee in the article "Dominant Risk Factors for Stroke Patients in Indonesia" [3]. WHO estimates that by 2025, the annual number of Europeans affected by stroke will increase from 1.1 million in 2000 to 1.5 million. Indonesia is no exception, as Riskesdas found in 2018 that 10.9% of all deaths in Indonesia were due to stroke.

Stroke incidence continues to increase in Indonesia, and the risk factors for this disease are crucial to be recognized and managed as soon as possible to prevent further damage and death. Unfortunately, the number of specialist doctors in Indonesia is still limited, and only 20% of Indonesians know the signs and symptoms, so many people wait too long before bringing stroke patients to the hospital. According to an article by Dr. Nanda L Prasetya (2020) cited on the website https://sippn.menpan.go.id/, if patients receive proper care, the effects of mild strokes can usually be managed in less than 10 minutes, and 90% can be reduced within less than four hours. Based on the above explanation, it can be concluded that by knowing and understanding the factors that cause stroke, support can be provided to take effective preventive measures to prevent stroke in the future [3].

The medical industry urgently needs a reliable and fast automated computerized system to provide a diagnosis of the causes and patterns of stroke. This is why it is crucial to maintain a record of data for each patient. The collected data can be used as a source to predict the likelihood of stroke in the future. Therefore, Data Mining techniques play a crucial role in extracting and predicting information and discovering hidden patterns in stroke medical data.

Data Mining is a method that utilizes statistics, mathematics, artificial intelligence, and machine learning techniques to extract and uncover information and knowledge that can be used from large databases. The practice of Data Mining refers to a set of processes used to extract previously unknown knowledge from a dataset [4]. The Data Mining methods that the writer will use are the Decision Tree (C4.5) algorithm, the Naïve Bayes (NB) algorithm, and the K-Nearest Neighbor (KNN) algorithm.

The writer chose KNN, C4.5, and NB because these methods have reliable advantages in classifying data. As explained by [5] in the journal "Comparing Different Supervised Machine Learning Algorithms for Disease Prediction" published by the National Library of Medicine, Naive Bayes has advantages such as the ability to handle discrete and continuous data, can make probabilistic predictions, and requires less training data. KNN has advantages such as the ability to classify instance data quickly and can handle instance data with noise or missing attribute values. Meanwhile, C4.5 has advantages in the classification tree, which is easier to understand and interpret, and supports multiple data types such as numeric, nominal, and categorical.

The problem of imbalanced class distribution data often occurs in medical data [6]. This can occur when the number of data in the majority and minority classes is unbalanced, which can cause errors in classification. In the case of stroke patient data, an imbalance in data between classes can cause misdiagnosis and inappropriate treatment. Therefore, it is necessary to study how to handle the problem of imbalanced class data in the medical world, especially in stroke patient cases discussed in this research.

One technique that can be used to handle imbalanced class data problems is Synthetic Minority Over-sampling Technique (SMOTE). This technique is an oversampling technique that synthetically adds new samples to the minority class so that the number of samples in both classes becomes balanced. In this sense, SMOTE can increase classification accuracy and improve the results of patient diagnosis and treatment [7].

Based on the above background, the writer will conduct research entitled "Comparison of Data Mining Classification Algorithms for Stroke Disease Prediction Using the SMOTE Upsampling Method". The problem statement in this study is as follows: 1). What is the comparison of accuracy between C4.5, NB, and KNN algorithms using SMOTE Upsampling technique and without it?, 2).What is the accuracy of C4.5, NB, and KNN algorithms using SMOTE Upsampling technique in predicting stroke?. The objectives of this study are as follows: 1). To compare the classification performance of datasets using SMOTE Upsampling technique and without it, 2). To determine the best performance of the three Data Mining classification algorithms (C4.5, NB, and KNN) in predicting the likelihood of stroke using SMOTE Upsampling technique.

II. METHOD

A. Data Mining

Data mining is a process of identifying and extracting relevant information from large datasets using statistical, mathematical, artificial intelligence, and machine learning methods. It helps in discovering new and substantial information from databases and assists in decision-making for the future by finding important patterns in large databases. According to [8], data mining is the most crucial stage because it can reveal hidden patterns in data. Data mining has several uses, such as description, estimation, prediction, classification, clustering, and association. It can be included in a problem-solving strategy called CRIPS-DM (The Cross-Industry Standard Process for Data Mining), as explained by [9]. CRIPS-DM has a life cycle consisting of six phases, where each phase depends on the results of the previous phase. Figure 1 shows the CRIPS-DM cycle, with adaptive arrows indicating the relationship between each stage. Based on the above Fig. 1, the CRISP-DM cycle consists of 6 phases [9], namely: business/research understanding phase, data understanding phase, data preparation phase, modeling phase, evaluation phase, and deployment phase.

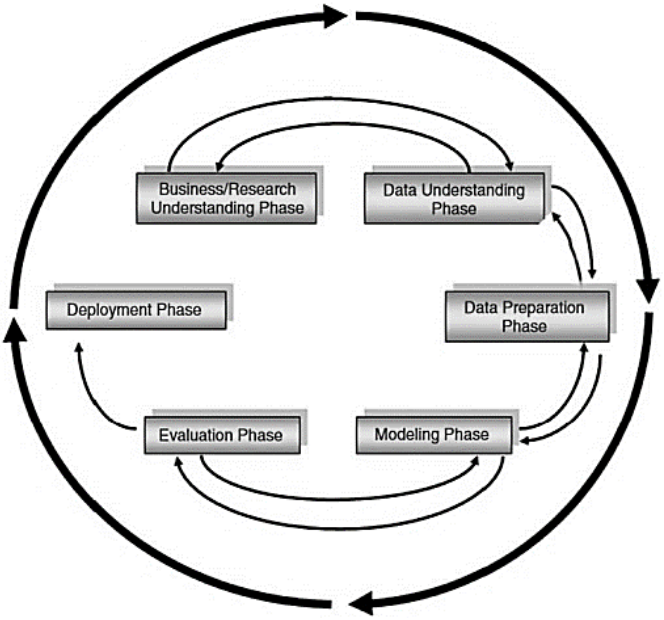

Fig. 1 CRISP-DM cycle [9]

B. Naïve Bayes (NB)

Naive Bayes is one of the popular classification methods used in machine learning. This algorithm is based on the Bayes theorem developed by the English mathematician, Thomas Bayes. This algorithm is used to determine the probability of a class or category based on the given data. The advantage of this method is its low complexity and requiring little data for training [10]. The Bayes theorem has a general form like (1).

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad (1)$$

Explanation:
y : data with an unknown class
x : hypothesis that data y belongs to a specific class
P(x|y) : probability of hypothesis x based on condition y (posterior probability)
P(x) : probability of hypothesis x (prior probability)
P(y|x) : probability of y based on hypothesis x
P(y) : probability of y

Naive Bayes is a simplification of Bayes' Theorem. The following is the simplified formula of Naive Bayes as (2).

$$P(x|y) = P(y|x)\,P(x) \quad (2)$$

C. Decision Tree (C4.5)

One of the most common methods for representing classification is C4.5. C4.5 is an algorithm used in machine learning to build decision tree models based on available data. C4.5 is a variant of the Decision Tree algorithm developed by computer expert J. R. Quinlan. C4.5 has several advantages over other algorithms, including the ability to handle data with continuous (numeric) and discrete (categorical) attributes, handle data with missing attribute values, create models that are more accurate and consistent than other Decision Tree algorithms, and handle heterogeneous and non-scaled data. According to [11], many branches of science have conducted in-depth research on the problem of decision tree construction from available data, including statistics, machine learning, pattern recognition, and data mining. There are several steps that can be taken to build a decision tree, one of which is to use the C.5 algorithm [9]. The steps are as follows:

1. First, collect training data. Training data usually comes from historical data, also known as previous data, which has been categorized into specific classifications.
2. Calculate the root of the tree. The root will be taken from the attribute that will be selected, and the initial root will be determined by summing the gain values of all attributes. First, calculate the entropy value, then the gain value of the attribute. The (3) formula is used to calculate the entropy value:

$$\text{Entropy (S)} - \sum_{i=1}^{n} - pi * \log_2 pi \quad (3)$$

Explanation:
S : set of cases
n : number of partitions in S
pi : proportion of Si with respect to S

3. Then calculate the gain value using the (4) formula.

$$\text{Gain (S, A)} = \text{Entropy (S)} - \sum_{i=1}^{n} \frac{|\text{Si}|}{|\text{S}|} * \text{Entropy (Si)} \quad (4)$$

Explanation:
S: set of cases
A: attribute
n: number of partitions in attribute A
|Si|: number of cases in partition i
|S|: number of cases in S

4. Continue using steps 2 and 3 until all records have been partitioned.
5. The decision tree partitioning process will stop when:
   a. all records in node N have the same class.
   b. There are no more attributes left to partition in the records.
   c. There are no records in the empty branch.

D. K-Nearest Neighbor (KNN)

According to [12], the KNN algorithm is a classification method that considers the attributes of training data to make predictions on new data. KNN stores all training data and compares the attributes of new data with the records in the training data to determine the class of the new data. This is an example of instance-based learning and includes case-based reasoning that handles symbolic data. This algorithm is also an example of lazy learning techniques, which wait until a question is asked before processing the training data. The formula for calculating the distance between two cases is:

$$similarity(T,S) = \frac{\sum_{i=1}^{n}(T_i,S_i)\; x\; w_i}{w_i} \quad (5)$$

Explanation:
T: new case
S: case in storage
n: number of attributes in each case

i: individual attribute between 1 and n
f: attribute similarity function between case T and S
w: weight given to the i-th attribute

E. SMOTE Upsampling

The common issue in medical data is the uneven distribution of data between classes [6]. Misclassification can occur if there is an imbalance between the major and minor classes. When there is an imbalance, the classifier will default to the major class, which can result in misdiagnosis and mistreatment of patients. Therefore, understanding the issue of imbalanced data is crucial in the medical field.

To address the imbalance in the number of objects in two data classes, the Synthetic Minority Oversampling Technique (SMOTE) can be applied. The major class is the data class with the most objects, while the minor class is the other class. The results of models built using imbalanced data can have a significant negative impact on processing outcomes. Imbalanced data is often overlooked by processing algorithms, which can cause the major class to dominate the minor class.

According to [13], the SMOTE approach is an alternative to oversampling strategies previously used to address imbalanced data problems. The SMOTE method differs from traditional oversampling methods in that it generates synthetic data by linking data from the minor class with neighboring data from the major class. In this way, the number of data from the minor class can be increased to match the number of data from the major class, achieving data balance. The SMOTE approach is very useful in cases where imbalanced data causes poor model performance.

The KNN method is used to create synthetic or fabricated data. The number of KNN is determined for ease of implementation. Synthetic data is generated with different numerical scales from categorical ones. Euclidean distance is used as a benchmark when working with numerical data, while mode is a simpler metric when working with categorical data. The Value Difference Metric (VDM) formula (6) is used to determine the distance between subclass samples where the variable is on a categorical scale [13].

$$\triangle\ (\mathrm{X},\mathrm{Y}) = \mathrm{w}_x \mathrm{w}_y \sum_{i=1}^{N} \delta = \mathrm{x}_i\ , \mathrm{y}_i)^r \qquad (6)$$

dengan:
$\triangle\ (\mathrm{X},\mathrm{Y})$: distance between observation X and Y
$w_x\ w_y$: observation weights (can be ignored)
N : number of predictor variables
R : takes a value of 1 (Manhattan distance) or 2 (Euclidean distance)
$\delta\ (\mathrm{x}_i\ ,\ \mathrm{y}_i)^{\ r}$: distance between categories, with the (7) formula.

$$\delta\ (V_1, V_2) = \sum_{i=1}^{n} \left|\frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2}\right|^k \qquad (7)$$

with:
$\delta\ (V_1, V_2)$: distance between values of V1 and V2
C_{1i} : number of V1 values in class i
C_{2i} : number of V2 values in class i
i : number of classes; i = 1, 2,...,m
C_1 : number of occurrences of value 1
C_2 : number of occurrences of value 2
N : number of categories
k : constant (usually 1)

Data generation procedure for:

1. Numeric Data
   - Calculate the differences between the main vector and its k nearest neighbors.
   - Multiply the differences with a randomly generated number between 0 and 1.
   - Add the differences to the original main vector value to obtain a new main vector.
2. Categorical Data
   a. Choose the majority value among the main vector and its k-nearest neighbors for the nominal value. If there is a tie, choose randomly.
   b. Make the chosen value as the new artificial class example.

F. Object Research

The data used in this study was obtained from Kaggle through the link https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. Although the original source of the dataset is not listed on the website, several research journals indicate that the dataset was originally derived from Electronic Health Records (EHR) owned by McKinsey & Company, a global management consulting firm that advises various businesses, governments, and institutions. The company is based in the United States and has branches in various countries [14].

Previously, this dataset was used as part of a health hackathon organized by McKinsey & Company in 2018. The total available data is 29,072, divided into 30% public data and 70% private data [15]-[16]. Other studies indicate that the data used on the website comes from the medical records of 5,110 individuals in the city of Bangladesh, but has undergone preprocessing from the original dataset originating from Electronic Health Records (EHR) managed by McKinsey & Company [17].

G. Research Steps

This section will explain the method used by the author in completing this thesis report. The research methodology used can assist the author in conducting research from start to finish, so that the thesis report can be organized neatly and systematically. In summary, the research methodology used by the author is described in Fig. 2.

H. Data Understanding

The data understanding phase aims to analyze the collected data. In this study, the data collection was obtained from the Kaggle website, as explained in subsection F. The dataset used is titled "Stroke Prediction Dataset" and consists of 5110 data instances with 12 attributes (Table I).

The 11 attributes mentioned in Table I are supported by several medical journals, which explain that these attributes can be factors in the occurrence of stroke. One medical journal that supports the 11 attributes is a journal written by [18]., The journal was published on the National Center for Biotechnology Information (NCBI) website and explains that age, gender, hypertension, history of heart disease, high blood glucose levels, and body mass index (BMI) are important risk factors in the occurrence of stroke.

According to a medical journal written by [19], it shows that unmarried, divorced, and widowed individuals have lower death rates within 1 week and 1 month after a stroke compared to married individuals. The study was conducted on 60,507 stroke patients in Denmark during the period of 2003-2012. The "mortality displacement" factor associated with shorter life expectancy in unmarried, divorced, and widowed individuals may explain the research findings. The study explains that the attribute "ever_married" can have an influence on stroke incidence.

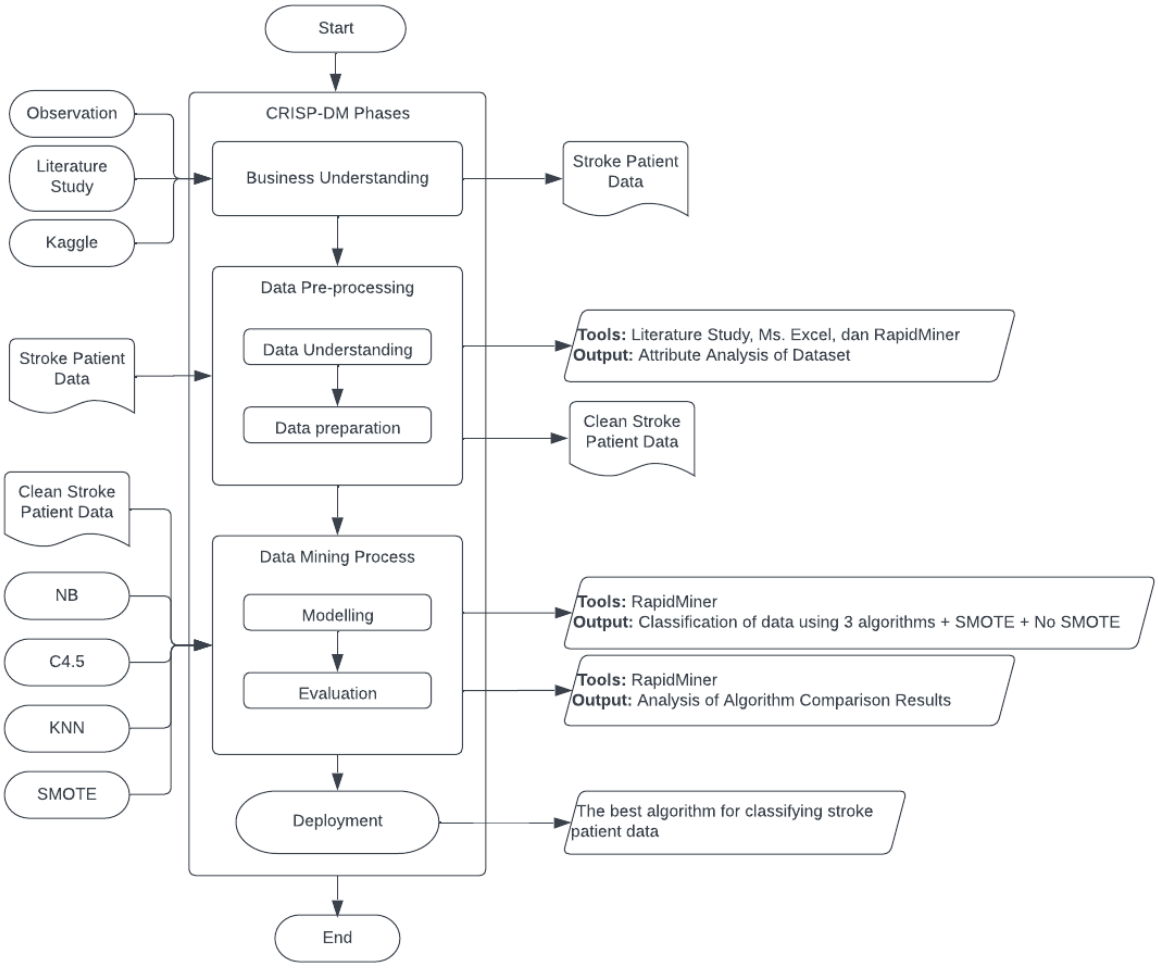

Fig. 2 Research steps

TABLE I
DATA OF STROKE PATIENTS BEFORE PREPROCESSING

No	Id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	90460	*M*	67	.0	1.0	*Yes*	*Private*	*Urban*	228.7	36.6	*formerly smoked*	1
2	51676	*F*	61	.0	.0	*Yes*	*Self-employed*	*Rural*	202.2	N/A	*never smoked*	1
3	31112	*M*	80	.0	1.0	*Yes*	*Private*	*Rural*	105.9	32.5	*never smoked*	1
4	60182	*F*	49	.0	.0	*Yes*	*Private*	*Urban*	171.2	34.4	*smokes*	1
5	16650	*F*	79	1.0	.0	*Yes*	*Self-employed*	*Rural*	174.1	24	*never smoked*	1
6	56669	*M*	81	.0	.0	*Yes*	*Private*	*Urban*	186.2	29	*formerly smoked*	1
7	53882	*M*	74	1.0	1.0	*Yes*	*Private*	*Rural*	70.1	27.4	*never smoked*	1
8	10434	*F*	69	.0	.0	*No*	*Private*	*Urban*	94.4	22.8	*never smoked*	1
9	27419	*F*	59	.0	.0	*Yes*	*Private*	*Rural*	76.2	N/A	*Unknown*	1
10	90460	*M*	67	.0	1.0	*Yes*	*Private*	*Urban*	228.7	36.6	*formerly smoked*	1
…		...	…	…	…	...	...	...	…	…	...	…
511	44679	*F*	44	.0	.0	*Yes*	*Govt_job*	*Urban*	85.3	26.2	*Unknown*	0

Here is the description of the attributes mentioned in the Table I.

- Stroke is the target attribute that indicates whether a patient has suffered a stroke or not. It has two categories: 1 (Stroke) and 0 (No Stroke).
- Gender represents the gender of the patients. It has three categories: Male, Female, and Other.
- Age represents the age of the patients in the dataset. The available age data ranges from 0 years to 82 years.
- Hypertension is a medical condition where a patient's blood pressure is consistently high, which can lead to damage to blood vessels and other organs. This attribute is used to identify whether the patients have a history of hypertension or not.
- Heart_disease is used to determine the presence of a history of heart disease in the patients. It indicates whether a patient has a history of heart disease or not.
- Ever_married indicates whether the patients have ever been married or not.
- Work_type indicates the type of work the patients are engaged in. It includes categories such as Children, Private, Self Employed, and Govt_job.
- Residence_type indicates the type of residence of the patients, which can be classified as Urban or Rural. It is important to consider as it can affect different health factors in each category of residence.
- Avg_glucose_level is an attribute used in health analysis, which measures the average glucose level in the patients' blood. It is important in determining the patients' health condition and taking necessary actions to control abnormal blood sugar levels. This attribute is crucial in assessing the risk of diabetes and taking preventive measures to reduce the risk of related complications.
- BMI (Body Mass Index) is an index used to determine the overall healthiness of an individual's weight. It is calculated by multiplying the weight in kilograms by the square of the height in meters. The result indicates the level of overweight or underweight of an individual and can be used as a health indicator.
- Smoking_status is used to identify the smoking status of the patients. It has categories indicating whether the patient is a former smoker, current smoker, or never smoked.

A medical journal written by [20], published on the American Heart Association (AHA) website, explains that based on available data, the rate of stroke and stroke-related deaths is higher in rural populations compared to urban populations. Vascular risk factors such as hypertension, diabetes mellitus, smoking, and atrial fibrillation are more frequently found in rural populations and are less controlled. Additionally, other

factors such as obesity, sedentary lifestyle, alcohol consumption, dietary patterns, and social deprivation also influence the stroke rate in rural areas. Therefore, it can be concluded that "resident_type" (type of residence) can be a factor in stroke occurrence, and better management of vascular risk factors is needed in rural areas.

A medical journal written by [21], published in the Neurology Journal, explains that occupation can contribute to an increased risk of stroke. Jobs with high levels of stress (high strain jobs) are associated with an increased risk of stroke, especially ischemic stroke. The results are more significant in women than in men. However, active or passive job characteristics are not associated with an increased risk of stroke compared to jobs with low stress levels.

Furthermore, before entering the data preparation stage, the author conducted an interview with a doctor from a hospital in Bandung City. The interview was conducted to validate the attributes taken from the Kaggle website with the attributes commonly used in the field. As a result, the doctor provided 7 recommended attributes commonly used in the field and also found in the Kaggle dataset. These seven recommended attributes include gender, age, hypertension, heart_disease, avg_glucose_level, bmi, and smoking_status.

These seven recommended attributes are based on the latest data from the CDC, which serves as a guideline. The CDC [22] is a data-driven and science-based service organization in the United States that protects public health. CDC has been in operation for over 70 years and has put science into action to help children stay healthy so they can grow and learn, assist families, businesses, and communities in fighting diseases and staying strong, and protect public health.

The latest CDC data [23] includes 13 attributes that can be risk factors for stroke, including 1) previous stroke or transient ischemic attack (TIA), 2) high blood pressure, 3) high cholesterol, 4) heart disease, 5) diabetes, 6) obesity, 7) sickle cell disease, 8) genetics and family history, 9) age, 10) sex, 11) race or ethnicity, 12) not getting enough physical activity, and 13) lifestyle (tobacco use, not getting enough physical activity, alcohol, eating a diet high in saturated fats, trans fat, and cholesterol). This data is also in line with the stroke implementation guidelines written by [24], which explain the 11 attributes that can be potential risk factors for stroke, including high blood pressure, diabetes, coronary heart disease, alcohol consumption, high cholesterol, smoking habits, obesity, blood clotting disorders, stress, lack of physical activity, and unchangeable risk factors such as advanced age (>60 years) and genetics. The explanations above explain that these seven recommended attributes are included in the 13 and 11 attributes previously described as potential risk factors for stroke. The results of the interview with the doctor are based on the recommended journals provided as data references to strengthen some of the recommended attributes, namely: gender [25], age [23], hypertension [26], heart_desease [27], avg_glucose_level [28], bmi [29], and smoking_status [30].

III. RESULTS AND DISCUSSION

A. Data Preprocessing

Data preprocessing is performed to be used in the next steps. This preprocessing involves selecting patient data that has complete information regarding smoking status. The ID attribute is not used in this process because it is not relevant as a determinant of stroke disease. Thus, the purpose of this step is to generate a dataset consisting of patient data that has the necessary information for stroke prediction analysis. After preprocessing, the stroke patient data is reduced to 4024 records, with a difference of 3843 records for suffered_stroke and 181 records for no_stroke. The data to be used consists of 11 attributes that determine the stroke disease, with 10 predictor attributes and 1 target attribute. This is done to ensure that the data used in the next stage is of good quality and ready to be used for building an accurate stroke disease prediction model. After going through data preprocessing, the dataset is divided into two for testing in RapidMiner. The first dataset will be modeled without SMOTE, while the second dataset will apply the SMOTE method to generate balanced fabricated data between suffered_stroke and no_stroke. The output result of the dataset after using the SMOTE Upsampling method amounts to 7686 records, with 3843 records for suffered_stroke and 3843 records for no_stroke (Table II).

stroke	Suffered Stroke
gender	Male
age	65-74
hypertension	1
heart_disease	0
ever_married	Yes
work_type	Private
residence_type	Urban
avg_glucose_level	<180 MG/DL
bmi	<30 KG/M
smoking_status	smokes

Fig. 3 Testing new stroke patient data

TABLE II
ATTRIBUTES AND CATEGORICAL VALUES IN PREDICTING STROKE DISEASE [31]

No	Attribute	Type	Category
1	*gender*	*binominal*	*Male; female*
2	*age*	*polynominal*	<15; 15-24; 25-34; 35-44; 55-64; 65-75; >75
3	*hypertension*	*binominal*	0; 1
4	*heart_disease*	*binominal*	0; 1
5	*ever_married*	*binominal*	*No; yes*
6	*work_type*	*polynominal*	*Children; govt_job; never_worked; private; self-employed*
7	*residence_type*	*binominal*	*Rural; urban*
8	*avg_glucose_level (after meal)*	*binominal*	>180 mg/dl; <180 mg/dl
9	*bmi*	*polynominal*	>30 kg/m; 30 kg/m; <30 kg/m
10	*smoking_status*	*polynominal*	*formely smoked; smoked; never smoked*
11	stroke	*binominal*	*suffered_stoke; no_stroke*

In the development of models in the medical field, accuracy, precision, recall, F1 score, and AUC value are some important metrics to evaluate the performance of the model. However, having high accuracy does not always indicate that the model built has good performance in predicting results in medical data. Therefore, it is important to consider other metrics such as precision, recall, F1 score and AUC value to evaluate the overall performance of the model.

The differences between NB, C4.5, and KNN algorithms in testing using SMOTE technique and without SMOTE technique indicate that this technique can improve the model's performance in handling imbalanced medical data. The test results show that using the SMOTE technique can yield significant differences in several evaluations, such as class precision, recall, F1 score, and AUC values for each algorithm. The NB algorithm has a difference of 53.48% for class precision, 63.13% for recall, 58.28% for F1 score, and a difference of 0.02 for AUC value when using the SMOTE technique compared to testing without the SMOTE technique. The C4.5 algorithm has a difference of 80.93% for class precision, 79.34% for recall, 80.11% for F1 score, and 0.37 for AUC value when using the SMOTE technique. On the other hand, the KNN algorithm has a relatively large difference, which is 83.53% for class precision, 59.14% for recall, 70.31% for F1 score, and 0.28 for AUC value when using the SMOTE technique.

In the development of models in the medical field, SMOTE technique can be used to handle imbalanced data and improve the performance of the model in predicting results. In the journal "Stroke Risk Prediction with Machine Learning Techniques" written by [32] and published in the National Library of Medicine, it is explained that class balance is very important in designing effective methods for predicting stroke, one of which is by using the Synthetic Minority Over-Sampling Technique (SMOTE). The journal also explains that when dealing with imbalanced data, metrics such as precision and recall are more suitable for identifying model errors. Precision measures how many of the patients who actually had a stroke are included in this class, while recall measures how many of the patients who had a stroke are correctly predicted.

The two metrics, precision, and recall can also affect the AUC value on the ROC curve. The closer the AUC value is to one, the better the performance of the machine learning model in distinguishing between patients who have had a stroke and those who have not. Therefore, in addition to considering accuracy, it is also important to pay attention to other metrics such as precision, recall, and AUC value, as well as using techniques like SMOTE to improve model performance so that prediction results can be relied upon and useful in medical decision-making in cases of data imbalance.

Based on the comparison results of the data mining algorithm testing to predict stroke disease in Table III, the Decision Tree C4.5 algorithm showed the best accuracy, recall, precision, F1 Score, and AUC value compared to Naïve Bayes and K-Nearest Neighbor. Therefore, new data testing needs to be done to predict stroke disease to maximize testing. The new data used in this analysis testing is taken from one of the hospitals in Bandung city different different from the hospital where the authors conducted interviews with medical practitioners. The C4.5 algorithm can be chosen to obtain accurate and reliable prediction results as the best algorithm among the three selected algorithms. Fig. 3 shows the new data of stroke patients to be tested for prediction, where the stroke patients have suffered_stroke disease.

In the testing conducted using the C4.5 algorithm with the SMOTE method and without it on the new data from one of the hospitals in Bandung, the results have been documented in the Table IV. From the testing results, it is proven that the use of the SMOTE method in the C4.5 algorithm can produce accurate predictions of

stroke disease compared to the C4.5 algorithm without SMOTE.

Furthermore, the author also validated the results by conducting follow-up interviews with doctors who provided recommendations regarding the attributes. The findings in this context can be summarized as follows: although the developed model successfully predicts stroke accurately, it should be noted that the model cannot be directly used in medical practice because the observations in the medical field to determine factors related to stroke are highly complex. Thus, a new understanding revealed that predicting stroke in a practical setting is highly complex. While data mining can be used as a predictive tool in the initial stage to make predictions for the general population, it is strongly recommended to seek direct examination by doctors in a hospital to obtain more accurate medical evaluations.

IV. CONCLUSION

This study compared the performance of three data mining classification algorithms (Naïve Bayes, Decision Tree C4.5, and K-Nearest Neighbor) in predicting stroke disease. Two different datasets were used to test the algorithms' performance, one with the application of the SMOTE technique and one without it. The results showed that the use of the SMOTE technique improved the precision, recall, F1 score, and AUC for all three algorithms, although the accuracy was slightly lower compared to the data without SMOTE. The C4.5 algorithm with the SMOTE technique demonstrated the best performance. Therefore, C4.5 was chosen to predict the data of new stroke patients obtained from one of the hospitals in Bandung. The use of C4.5 with the SMOTE technique proved to predict with higher accuracy than C4.5 without SMOTE. However, it should be noted that based on the author's interview with one of the medical practitioners, the model cannot be directly used in medical practice because the observations in the medical field to determine factors related to stroke are highly complex. Thus, a new understanding revealed that predicting stroke in a practical setting is highly complex. While data mining can be used as a predictive tool in the initial stage for predictions in the general population, it is strongly recommended to undergo direct examination by doctors in a hospital to obtain more accurate and comprehensive medical evaluations.

TABLE III
COMPARISON OF NB, C4.5, AND KNN ALGORITHM TESTING RESULTS USING SMOTE AND WITHOUT SMOTE

No	Algorithm	Without SMOTE					With SMOTE				
		Accuracy	*Precision*	*Recall*	AUC	F1 Score	Accuracy	*Precision*	*Recall*	AUC	F1 Score
1	NB	92,22	20,02	23,31	0,832	21,15	77,61	73,50	86,44	0,856	79,43
2	C4.5	95,50	0,00	0,00	0,500	0,00	80,31	80,93	79,34	0,875	80,11
3	KNN	95,08	5,26	0,53	0,595	1,00	76,03	88,79	59,67	0,875	71,31

TABLE IV
TESTING NEW STROKE PATIENT DATA USING C4.5 ALGORITHM WITH SMOTE METHOD

Category	Decision Tree (C4.5) without SMOTE	Decision Tree (C4.5) + SMOTE
prediction(stroke)	No Stroke	Suffered Stroke
confidence (no stroke)	0.957	0.336
confidence (suffered stroke)	0.043	0.664
gender	Male	Male
age	65-74	65-74
hypertension	1	1
heart_disease	0	0
ever_married	Yes	Yes
work_type	Private	Private
residence_type	Urban	Urban
avg_glucose_level	<180 MG/DL	<180 MG/DL
bmi	<30 KG/M	<30 KG/M
smoking_status	smokes	smokes

REFERENCES

[1] P2PTM Kemenkes RI, "Germas Cegah Stroke - Direktorat P2PTM," *P2PTM Kemenkes RI*, 2017, Diakses: 11 Oktober 2022. [Daring]. Tersedia pada: http://p2ptm.kemkes.go.id/tag/germas-cegah-stroke

[2] A. Yonata dan A. S. P. Pratama, "Hipertensi sebagai Faktor Pencetus Terjadinya Stroke," 2016.

[3] L. (Lannywati) Ghani, D. (Delima) Delima, dan L. K. (Laurentia) Mihardja, "Faktor Risiko Dominan Penderita Stroke di Indonesia," *Indones. Bull. Heal. Res.*, vol. 44, no. 1, hal. 20146, Mei 2016, doi: 10.22435/BPK.V44I1.4949.49-58.

[4] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, hal. 213–219, Feb 2017, doi: 10.22202/EI.2016.V2I2.1465.

[5] S. Uddin, A. Khan, M. E. Hossain, dan M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, Des 2019, doi: 10.1186/S12911-019-1004-8.

[6] S. Mutmainah, "PENANGANAN IMBALANCE DATA PADA KLASIFIKASI KEMUNGKINAN PENYAKIT STROKE," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 1, no. 1, Agu 2021, Diakses: 10 November 2022. [Daring]. Tersedia pada: https://journal.uii.ac.id/jurnalsnati/article/view/20060

[7] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, dan G. T. Wang, "LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM," *Knowledge-Based Syst.*, vol. 196, hal. 105845, Mei 2020, doi: 10.1016/J.KNOSYS.2020.105845.

[8] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*. Elsevier Inc., 2012. doi: 10.1016/C2009-0-61819-5.

[9] D. T. Larose, "Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition," *Discov. Knowl. Data An Introd. to Data Min. Second Ed.*, vol. 9780470908, hal. 1–316, Jul 2014, doi: 10.1002/9781118874059.

[10] A. Samosir, M. S. Hasibuan, W. E. Justino, dan T. Hariyono, "Komparasi Algoritma Random Forest, Naïve Bayes dan K- Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung," *Pros. Semin. Nas. Darmajaya*, vol. 1, no. 0, hal. 214–222, Sep 2021, Diakses: 19 September 2022. [Daring]. Tersedia pada: https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/2955

[11] H. Dahan, S. Cohen, L. Rokach, dan O. Maimon, "Proactive Data Mining with Decision Trees," 2014, doi: 10.1007/978-1-4939-0539-3.

[12] H. LEIDIYANA, "Komparasi Algoritma Klasifikasi Data Mining Dalam Penentuan Resiko Kredit Kepemilikan Kendaraan Bemotor," 2013. doi: 10.31294/P.V15I2.6349.

[13] R. A. Barro, I. D. Sulvianti, dan F. M. Afendi, "PENERAPAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) TERHADAP DATA TIDAK SEIMBANG PADA PEMBUATAN MODEL KOMPOSISI JAMU," *Xplore J. Stat.*, vol. 1, no. 1, Apr 2013, doi: 10.29244/XPLORE.V1I1.12424.

[14] "Mckinsey & Company Intelligence Benefits Society A Conversation – Management," 2020. https://bbs.binus.ac.id/management/2020/03/mckinsey-company-intelligence-benefits-society-a-conversation/ (diakses 3 April 2023).

[15] M. S. Pathan, Z. Jianbiao, D. John, A. Nag, dan S. Dev, "Identifying Stroke Indicators Using Rough Sets," *IEEE Access*, vol. 8, hal. 210318–210327, 2020, doi: 10.1109/ACCESS.2020.3039439.

[16] "McKinsey Analytics Online Hackathon - Healthcare Analytics," 2018. https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon/#ProblemStatement (diakses 3 April 2023).

[17] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, dan M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," *Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2020*, hal. 1464–1469, Nov 2020, doi: 10.1109/ICECA49313.2020.9297525.

[18] S. J. Murphy dan D. J. Werring, "Stroke: causes and clinical features," *Medicine (Abingdon).*, vol. 48, no. 9, hal. 561, Sep 2020, doi: 10.1016/J.MPMED.2020.06.002.

[19] K. K. Andersen dan T. S. Olsen, "Stroke case-fatality and marital status," *Acta Neurol. Scand.*, vol. 138, no. 4, hal. 377–383, Okt 2018, doi: 10.1111/ANE.12975.

[20] M. K. Kapral *dkk.*, "Rural-urban differences in stroke risk factors, incidence, and mortality in people with and without prior stroke: The CANHEART stroke study," *Circ. Cardiovasc. Qual. Outcomes*, vol. 12, no. 2, Feb 2019, doi: 10.1161/CIRCOUTCOMES.118.004973.

[21] Y. Huang *dkk.*, "Association between job strain and risk of incident stroke," *Neurology*, vol. 85, no. 19, hal. 1648–1654, Nov 2015, doi: 10.1212/WNL.0000000000002098.

[22] "Centers for Disease Control and Prevention," 21 Februari 2023. https://www.cdc.gov/about/organization/cio.htm (diakses 5 April 2023).

[23] D. Mozaffarian *dkk.*, "Heart disease and stroke statistics-2016 update a report from the American Heart Association," *Circulation*, vol. 133, no. 4, hal. e38–e48, Jan 2023, doi: 10.1161/CIR.0000000000000350.

[24] S. K. Handayani, Dr. Fitria, SKp., MKep., S. K. K. Widyastuti, Rita Hadi SKp., MKep., dan M. E. Eridani,

Dania ST., "Buku Panduan Penatalaksanaan stroke," hal. 1–66, 2019.

[25] K. M. Rexrode, T. E. Madsen, A. Y. X. Yu, C. Carcel, J. H. Lichtman, dan E. C. Miller, "The Impact of Sex and Gender on Stroke," *Circ. Res.*, vol. 130, no. 4, hal. 512–528, Feb 2022, doi: 10.1161/CIRCRESAHA.121.319915.

[26] M. Wajngarten dan G. Sampaio Silva, "Hypertension and Stroke: Update on Treatment," *Eur. Cardiol. Rev.*, vol. 14, no. 2, hal. 111, 2019, doi: 10.15420/ECR.2019.11.1.

[27] W. Kim dan E. J. Kim, "Heart Failure as a Risk Factor for Stroke," *J. Stroke*, vol. 20, no. 1, hal. 33, Jan 2018, doi: 10.5853/JOS.2017.02810.

[28] R. Chen, B. Ovbiagele, dan W. Feng, "Diabetes and Stroke: Epidemiology, Pathophysiology, Pharmaceuticals and Outcomes," *Am. J. Med. Sci.*, vol. 351, no. 4, hal. 380, Apr 2016, doi: 10.1016/J.AMJMS.2016.01.011.

[29] A. Bardugo *dkk.*, "Body Mass Index in 1.9 Million Adolescents and Stroke in Young Adulthood," *Stroke*, vol. 52, no. 6, hal. 2043–2052, Jun 2021, doi: 10.1161/STROKEAHA.120.033595.

[30] R. S. Shah dan J. W. Cole, "Smoking and stroke: the more you smoke the more you stroke," *Expert Rev. Cardiovasc. Ther.*, vol. 8, no. 7, hal. 917, 2010, doi: 10.1586/ERC.10.56.

[31] R. S. Rohman, R. A. saputra, dan D. A. Firmansaha, "Komparasi Algoritma C4.5 Berbasis PSO Dan GA Untuk Diagnosa Penyakit Stroke," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 1, hal. 155–161, Jan 2020, Diakses: 11 November 2022. [Daring]. Tersedia pada: https://jurnal.unimed.ac.id/2012/index.php/cess/article/view/15225

[32] E. Dritsas dan M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors*, vol. 22, no. 13, Jul 2022, doi: 10.3390/s22134670.