Comparison of Naïve Bayes Algorithm and Decision Tree C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement

Utomo Pujianto a, 1, *, Asa Luki Setiawan a, 2, Harits Ar Rosyid a, 3, Ali M. Mohammad Salah b, 4

a Department of Electrical Engineering, Universitas Negeri Malang
Jl. Semarang No.5, Malang 65145, Indonesia
b Dept. Of Computer information systems, Al Quds Open University
Beit Jalla-The Main road-Khallat Al Badd, Bethlehem, Palestine

1 utomo.pujianto.ft@um.ac.id; 2 asahkiiasa@gmail.com; 3 harits.ar.ft@um.ac.id; 4 salah@qou.edu
* corresponding author

I. Introduction

Special care for diabetic patients is important for their survival. The HbA1c examination is useful for controlling patients for diabetes. Therefore, diabetes is a metabolic disorder because the body cannot use the insulin that is produced effectively [1]. The hormone that regulates the balance of blood sugar levels is a function of insulin, so that if there is an increase in the concentration of glucose in the blood it causes an abnormality called hyperglycemia [2]. International Diabetes Federation (IDF) states that the prevalence of Diabetes Mellitus in the world is 1.9 % and has made Diabetes Mellitus the seventh leading cause of death in the world while in 2012 the incidence of diabetes mellitus in the world is 371 million [3]. The high prevalence of Diabetes Mellitus is caused by risk factors that cannot change such as heredity and changeable risk factors such as smoking habits, education level, occupation, physical activity, alcohol consumption, body mass index, waist circumference and age [4].

HbA1c (glycated hemoglobin) is hemoglobin that binds to glucose. Ordinarily, glucose binds to each other with hemoglobin in the red blood cells. Consequently, the amount of HbA1c in the human body is balanced with blood sugar levels. The higher the blood sugar level impact the higher the HbA1c level. Nevertheless, HbA1c can measure the average blood sugar level for three months [5]. Hospital readmission is a medical term to take action to re-treat patients who have previously received inpatient services in hospitals [6]. The readmission process relates to calculating the quality of patient
handling by the hospital [7]. Several attributes of a diabetic patient dataset are influential on the quality of treatment which refers to the resistance of glycemic serum in the body. Consequently, the better quality of treatment for the hospital identified by the longer the glycemic serum is at a healthy level. But the differences in attributes associated with diabetic patients result in the calculation of quality, tend to be complicated [8]. The readmission process is very important to anticipate diabetic patients who are late in re-treating their disease.

Pattern recognizing data in the field of informatics is often known as classification [9]. In a study of the classification of Hospital Readmission Diabetes Patients, some methods that have been used are Logistic Regression [10]. The advantage of Logistic Regression is the output of logistic regression is more informative than other classification algorithms. Like any regression approach, it expresses the relationship between an outcome variable (label) and each of its predictors (features) [11]. The disadvantages of Logistic Regression include vulnerability to underfitting in the imbalance data set and, consequently the value of accuracy is uncertain [12]. Other studies of the classification of Hospital Readmission Diabetes Patients, are compared to Decision Tree algorithms, K-Nearest Neighbor (k-NN), and Naïve Bayes with various parameters [8], resulted in the Naïve Bayes classification model having better statistics than other algorithm models such as Decision Tree and k-NN with an accuracy value of 57.52 %, MAE of 0.512, and the kappa statistic of 0.182. There is another study by implementing the C4.5 algorithm to classify the readmissions of diabetic patients, tested the C4.5 algorithm with several different experiments. The results of this study, the C4.5 algorithm can classify readmissions of diabetic patients with an accuracy rate of 74.5 % with preprocessing data treatment using two label classes. Nevertheless, the highest accuracy in the classification of the three label classes has an accuracy rate of only 57 % using the C4.5 algorithm as a classification method [13].

Based on the consideration of the algorithm discussed earlier, this study uses the Naïve Bayes algorithm and gives a comparison of the Decision Tree C4.5 algorithm which has the advantage of being able to process a numerical data (continuous), category (discrete), handle missing attribute values and generate rules which is easily interpreted [14]. Both algorithms are used to determine the performance of the preprocessing stage, which is done as an improvement in the accuracy of the classification, such as comparing the performance of the two methods by testing the dataset before and after changing the imbalance class dataset using SMOTE (Synthetic Minority Over-Sampling Technique). Accordingly, SMOTE is one of the supervised learning preprocessing methods to overcome imbalance classes [15], and in this case, SMOTE is used for oversampling minority classes so that the data in the class is balanced. The next comparison is by using the feature selection to simplify the number of attributes. The wrapper is used because this method can perform a feature selection optimally which can be adjusted with the desired algorithm [16].

In this study, Naïve Bayes and Decision Tree C4.5 methods were tested to classify hospital readmissions of diabetic patients using input test results from laboratory tests and other variables in diabetic patients. The results of this study are the best performance results in the classification of hospital readmissions from several trial scenarios that have been carried out. Consequently, they can be developed into further research in making recommendations for diabetic patients needing re-treatment in less than 30 days of previous treatment, more than 30 days of previous treatment and do not require treatment. The purpose of this study is to find out the best algorithm in classifying hospital readmissions of diabetic patients, and the best combination of preprocessing methods.

II. Materials and Methods

Machine Learning is a field of science about how a machine can manage data as desired [17]. Machine Learning is a part of Artificial Intelligence that focuses on developing a system that is able to learn its own patterns based on a training test and determined without human intervention. The application of Machine Learning is found in several fields, such as the field of education [18], the field of games [19], and in this research applying machine learning in the medical field. Machine Learning has three types of learning methods, namely Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Supervised Learning is a structured learning method that the purpose is to group test data into the label class based on the model that has been found through learning in the training data. While Unsupervised Learning is an unstructured learning method so there is no class of labels, but only data that will be grouped into groups or new label classes. Meanwhile, Reinforcement
Learning is a learning method without any knowledge so that in learning something, the system will do a certain action and see the results of the action [20].

Basically, the way machine learning works is learning like humans by using examples and after that, it can answer a related question. This learning process uses data called the dataset train. Unlike static programs, Machine Learning was created to form programs that can learn on their own. Problems that can be solved by Machine Learning include regression, clustering, and classification. Classification is a method of grouping data that has been determined by its class. In this research classification process use algorithms such as Naïve Bayes and Decision Tree C4.5 to classify a problem that is combined with the SMOTE and feature selection.

A. Dataset

The data used in the study are data obtained from the UCI Machine Learning Repository about diabetic patients. Data on these diabetic patients represent 10 years (1999 to 2008) patient data at diabetes care clinics in 130 US hospitals that are interconnected with other networks. This dataset consists of 50 attributes and 101,776 instances. Table 1 is a table of the dataset metadata used in this study.

B. Preprocessing Data

Comparing the Naïve Bayes and C4.5 algorithms require preprocessing data before the process is done [21]. Preprocessing data applies process types that process raw data to prepare the next data processing [22]. The purpose of this preprocessing is to transform data into a format that is easier and more effective for user needs, with more accurate indicators of results, reduction of computational time for large scale problems, making data values smaller without changing the contained information.

The first preprocessing stage is trimming the data used by using only patient data that have an HbA1c examination. Consequently, the attribute data A1c test result deletes the value of the "none" variable which amounts to 84,748 instances with the intention of data on patients who do not take the HbA1c examination. The results of the data after trimming only amounts to 17,018 instances. This is advantageous for this research with a smaller amount of data can improve processing time. Several preprocessing stages are compared to eight different preprocessing scenarios (See Table 2). This scenario compares the effect of SMOTE and feature selection in processing data before entering the classification phase. Data in all scenarios only applies the data cleaning method as the initial preprocessing stage. The first scenario without using the SMOTE preprocessing method and the feature selection only uses initial data with label classes totaling three classes "No", ">30", and "<30".

The second scenario in this study applies the SMOTE method for minority class data so that the distribution of label classes is balanced, moreover the number of label classes consists of the same three classes with scenario one. The third scenario in this study applies the feature selection method using a wrapper for feature selection. The features that are omitted are features that have an unbalanced data distribution or one of the empty data distribution values (zero). In the fourth scenario, apply both the preprocessing method of balancing three label classes using SMOTE then using the feature selection to simplify the number of attributes. After that in the fifth to eight scenarios apply the same method in a row with the first to the fourth method, but only use two label classes ">30" and "<30" for the next classification data.

Several scenarios test are useful to find out the combination of preprocessing techniques that produce high accuracy values in the next process. The scenarios arranged are several combinations of SMOTE preprocessing techniques and Feature selection. This research is tested by the 10-fold cross validation method by comparing Naïve Bayes and C4.5 algorithms.

1) Data Cleaning

The process of detecting and repairing datasets that have missing value, noise, and other imperfections can be detected by the data cleaning process. Data cleaning is useful for identifying data that is incomplete, incorrect and noise. Consequently, the data will be replaced, modified or deleted. This data cleaning process is quite important in conducting modeling of Machine Learning algorithms because at this stage data cleaning can prevent duplicate data, missing value data, ambiguous data and naming conflicts. There are several focus areas in the data cleaning like missing values, outliers, inconsistent codes, schema integration, and duplicates [23]. One of the frequently used data cleaning techniques is handling data missing. According to Twisk 2002, a method that is able to handle the
The working principle of this method is to detect each instance that has empty data. And then take the average value of the data attribute that has missing and fill in the average value of the attribute to the data that has empty data. This is useful as a substitute value for the empty data so that it is expected to increase accuracy in the subsequent modeling.

The concept of data cleaning applied in this study is by removing attribute values that have very high missing values such as the attribute "payer code" with a missing value of 52 % which has the potential to have no correlation with this study, and the "weight" and "Medical specialty" attribute that should be removed because it has very large missing values. This attribute causes data ineffectiveness on processing with a 97 % and 53 % missing value. In addition, these three attributes, attributes that have a missing value will use the Replace with value method in the missing value by giving the results in the attributes found in Table 3.

Table 1. List of attributes in the dataset

<table>
<thead>
<tr>
<th>Attributes Name</th>
<th>Data Type</th>
<th>Attributes Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encounter ID</td>
<td>Numerical</td>
<td>Visit Number as ID.</td>
</tr>
<tr>
<td>Patient number</td>
<td>Numerical</td>
<td>Number of patients.</td>
</tr>
<tr>
<td>Race</td>
<td>Nominal</td>
<td>Values: Caucasian, Asian, African American, Hispanic, and others.</td>
</tr>
<tr>
<td>Age</td>
<td>Nominal</td>
<td>Value: Grouped with 10-year intervals (0 to 10, 10 to 20, &amp; 90 to 100).</td>
</tr>
<tr>
<td>Gender</td>
<td>Nominal</td>
<td>Values: Male, Female, and Unknown.</td>
</tr>
<tr>
<td>Weight</td>
<td>Nominal</td>
<td>Weight in pounds.</td>
</tr>
<tr>
<td>Admission Type</td>
<td>Nominal</td>
<td>Value: Emergency, urgent, elective, newborn, and not available.</td>
</tr>
<tr>
<td>Discharge Disposition</td>
<td>Nominal</td>
<td>Value: discharge to home, expired, not available.</td>
</tr>
<tr>
<td>Admission Source</td>
<td>Nominal</td>
<td>Value: referral physician, emergency room, and transfer from hospital.</td>
</tr>
<tr>
<td>Time in Hospital</td>
<td>Numerical</td>
<td>The number or duration of patients treated at home from enrollment to discharge.</td>
</tr>
<tr>
<td>Payer Code</td>
<td>Nominal</td>
<td>Payment Code.</td>
</tr>
<tr>
<td>Medical specialty</td>
<td>Nominal</td>
<td>Special handling such as cardiology, internal medicine, etc.</td>
</tr>
<tr>
<td>Number of lab procedures</td>
<td>Numerical</td>
<td>The number of lab tests carried out at one visit.</td>
</tr>
<tr>
<td>Number of procedures</td>
<td>Numerical</td>
<td>Number of procedures at one visit.</td>
</tr>
<tr>
<td>Number of medications</td>
<td>Numerical</td>
<td>Number of medicines for patients given in one visit.</td>
</tr>
<tr>
<td>Number of outpatients visits</td>
<td>Numerical</td>
<td>Number of outpatient visits in the treatment process.</td>
</tr>
<tr>
<td>Number of emergency visits</td>
<td>Numerical</td>
<td>Number of emergency visits while in the maintenance phase.</td>
</tr>
<tr>
<td>Number of inpatients visits</td>
<td>Numerical</td>
<td>Number of inpatient visits that are in the care stage.</td>
</tr>
<tr>
<td>Diagnosis 1</td>
<td>Nominal</td>
<td>Main diagnosis; there are 848 different values.</td>
</tr>
<tr>
<td>Diagnosis 2</td>
<td>Nominal</td>
<td>Second diagnosis; there are 923 different values.</td>
</tr>
<tr>
<td>Diagnosis 3</td>
<td>Nominal</td>
<td>Additional diagnosis that supports a second diagnosis; there are as many as 954 different values.</td>
</tr>
<tr>
<td>Number of diagnoses</td>
<td>Numerical</td>
<td>The number of diagnoses that are input into the system.</td>
</tr>
<tr>
<td>Glucose serum test result</td>
<td>Nominal</td>
<td>Indicates the range of results; value, &gt; 200, &gt; 300, normal, and none (none).</td>
</tr>
<tr>
<td>A1c test result</td>
<td>Nominal</td>
<td>Vulnerable indications of the HbA1c test with a value of “&gt; 8” if the test results are more than 8 %, “&gt; 7” if the test results are more than 7 % but less than 8 %, “norm” if the test results are less than 7 % and “none” if do not test.</td>
</tr>
<tr>
<td>Change of medication</td>
<td>Nominal</td>
<td>Indicates if there are changes in treatment (either the dose of the drug or drug used), the value: No (if there is no change) or Change (if change)</td>
</tr>
<tr>
<td>Diabetes Medication</td>
<td>Nominal</td>
<td>Indicate if there is another diabetes treatment prescribed. Value: Yes and No</td>
</tr>
<tr>
<td>24 features for medications</td>
<td>Nominal</td>
<td>Information about changes in medication dosage during treatment with a value of “up” indicates an increased dose, “down” the drug dose is lowered, “Steady” remains. 24 types of drugs such as: metformin, repaglidel, nateglindide, chlorpropamide, and others.</td>
</tr>
<tr>
<td>Readmitted</td>
<td>Nominal</td>
<td>A value of “&gt; 30” for patients readmissions for more than 30 days, Value “&lt;30” for patients readmissions for less than 30 days, and “No” for those who are not readmissions.</td>
</tr>
</tbody>
</table>
SMOTE

Addressing data imbalance problems need to pay attention to unbalanced data distribution from each class. SMOTE is one of the supervised learning preprocessing methods to overcome the imbalance class [15]. And in this case, SMOTE is used for oversampling minority classes so that the data in the class is balanced. The label class data in this dataset show the imbalance of the data shown in Table 4.

There is a second scenario in this study, which is found in the Felix Tamin 2017 study by eliminating the class label "No" and assumed to be the same as the label class "<30" because the label "No" does not have a history of readmissions [13]. The elimination of the class label "No" is also based on that diabetes cannot be cured [25], with this statement the class value label "No" becomes irrelevant, because basically when a person has diabetes, they have readmission to the hospital with a certain period of time to control the patient's blood sugar level.

When a person has diabetes, the cure that can be attempted by medical personnel is to control the blood sugar of the patient so that the patient's blood sugar remains in the normal position. The comparison of the data before and after preprocessing is using two class labels as can be found in Table 5.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Preprocessing</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>No SMOTE &amp; Feature selection</td>
<td>3 Classes</td>
</tr>
<tr>
<td>2.</td>
<td>SMOTE</td>
<td>3 Classes</td>
</tr>
<tr>
<td>3.</td>
<td>Feature selection</td>
<td>3 Classes</td>
</tr>
<tr>
<td>4.</td>
<td>SMOTE + Feature selection</td>
<td>3 Classes</td>
</tr>
<tr>
<td>5.</td>
<td>Tanpa SMOTE &amp; Feature selection</td>
<td>2 Classes</td>
</tr>
<tr>
<td>6.</td>
<td>SMOTE</td>
<td>2 Classes</td>
</tr>
<tr>
<td>7.</td>
<td>Feature selection</td>
<td>2 Classes</td>
</tr>
<tr>
<td>8.</td>
<td>SMOTE + Feature selection</td>
<td>2 Classes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Data Type</th>
<th>% Missing Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>Nominal</td>
<td>2 %</td>
</tr>
<tr>
<td>Diagnosis 3</td>
<td>Nominal</td>
<td>1 %</td>
</tr>
</tbody>
</table>

2) SMOTE

Addressing data imbalance problems need to pay attention to unbalanced data distribution from each class. SMOTE is one of the supervised learning preprocessing methods to overcome the imbalance class [15]. And in this case, SMOTE is used for oversampling minority classes so that the data in the class is balanced. The label class data in this dataset show the imbalance of the data shown in Table 4.

<table>
<thead>
<tr>
<th>Label Class</th>
<th>Data Distribution</th>
<th>After SMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before SMOTE</td>
<td>After SMOTE</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Percentage</td>
</tr>
<tr>
<td>No</td>
<td>9542</td>
<td>56 %</td>
</tr>
<tr>
<td>&gt;30</td>
<td>5800</td>
<td>34 %</td>
</tr>
<tr>
<td>&lt;30</td>
<td>1676</td>
<td>10 %</td>
</tr>
<tr>
<td>Total</td>
<td>17018</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label Class</th>
<th>Data Distribution</th>
<th>After SMOTE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before SMOTE</td>
<td>After SMOTE</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Percentage</td>
</tr>
<tr>
<td>&gt;30</td>
<td>5800</td>
<td>77 %</td>
</tr>
<tr>
<td>&lt;30</td>
<td>1676</td>
<td>23 %</td>
</tr>
<tr>
<td>Total</td>
<td>7476</td>
<td></td>
</tr>
</tbody>
</table>
3) Feature selection

Optimizing the performance of the classification algorithm model by feature selection is an important part. Feature selection can be based on a large reduction in feature space. For example by eliminating less relevant attributes. Using the right feature selection algorithm can improve the performance of the algorithm. The feature selection can be divided into filters and wrappers. Examples of filter types are information gain (IG), chi-square, and log likelihood ratio. Examples of wrapper types are forward selection, wrapper subset evaluation, and backward elimination. The results of the precision using wrapper are higher than the filter method, but these results are achieved with a large degree of complexity. Consequently, high complexity can cause problems [26]. One feature selection method that can be used to make feature selection is Wrapper Subset Evaluation. Wrapper Subset Evaluation used to evaluate the set of attributes using the learning scheme and to estimate the accuracy of the learning scheme for several attributes is by using cross validation [27].

This study uses the wrapper subset evaluation with the greedy stepwise method in selecting features for several data processing scenarios. In the data scenario with three label classes, the application of feature selection used for scenario 3 and 4. The attributes used before feature selection is 47 attributes. In the feature selection of the Naïve Bayes algorithm, the features used only 18 attributes on scenario 3 and 18 attributes on scenario 4. And in the C4.5 algorithm classification for scenario 3 and 4, the attribute used after feature selection is 7 attributes. In the scenario using two label classes, the application of feature selection used for scenario 7 and 8. The Naïve Bayes algorithm feature selection test uses 25 attributes and in the C4.5 algorithm uses 9 attributes.

C. Classification

The process to find a model that is able to distinguish data classes based on rules in order to predict the class of an unknown data label called classification. Classification is also a field of research in the acquisition of information that develops methods to determine or categorize data into one or more groups that have been previously known automatically based on the contents of the data. Classification aims to group unstructured data into groups that describe the contents of the dataset [28]. Classification is useful for finding models from training data that distinguish records into appropriate categories or classes, the model is then used to classify records whose classes have not been previously known in testing data. Classification can also make decisions by predicting a case based on the classification results obtained [29]. The data classification in this study is used to test two classification algorithms, Naïve Bayes and Decision Tree C4.5 in classifying readmission diabetes patients.

1) Naïve Bayes

The Naïve Bayes algorithm is a simple classification method that calculates probabilities by calculating the frequency of combination values on a given dataset [30]. Using the Naïve Bayes algorithm assumes that all attributes become independent considering the value of the class variable has conditional properties. The Naïve Bayes algorithm predicts future opportunities based on prior experience so that it is known as the Bayes Theorem. The main feature of Naïve Bayes is a very naive assumption of independence from each condition or event. This algorithm is so popular in machine learning applications because Naïve Bayes has a simple algorithm that allows each attribute to contribute to the final decision. This simplicity is similar to computational efficiency, which makes the Naïve Bayes algorithm interesting and suitable for many domains [31]. This algorithm performs pattern recognition and several approaches to get the desired results [32]. Naïve Bayes works very well compared to other classifier models. This is evidenced in the journal Xhemali 2009 that Naïve Bayes has a better level of accuracy than other classifier models [31].

The use of the Naïve Bayes algorithm has several important benefits, one of which is that this method only requires a relatively small amount of training data in determining the estimated parameters needed for the classification process. Because what is assumed to be an independent variable, only the variance of a variable in a class is needed to determine the classification, not the whole of the covariance matrix [33]. The stages of the Naïve Bayes algorithm process are quite simple, including:

1. Calculate the total number of classes.
2. Calculate the probability of each class.
3. Apply the Bayes formula (1) by multiplying all class variables.
4. Compare the results of each class.
To describe the Bayes theorem there are bayes formula as in (1)

\[
P(c|x) = \frac{P(x|c)P(c)}{P(x)}
\]  

(1)

Where \(x\) is data with an unknown class, \(c\) is the data hypothesis of a specific class, \(P(c|x)\) is probability of hypothesis based on condition, \(P(c)\) is Probability of hypothesis (prior probability), \(P(x|c)\) is probability based on conditions in the hypothesis, and \(P(x)\) is Probability c.

2) Decision Tree C4.5

Decision Tree C4.5 algorithm is an algorithm that has the advantage of being able to process numerical data (continuous), categories (discrete), handles missing values and produce rules that are easily interpreted [14]. This C4.5 algorithm is the development of the ID3 algorithm. The working principle of algorithm ID3 and C4.5 algorithm is similar, but there are some differences that make the C4.5 algorithm have better results than the ID3 algorithm. The C4.5 algorithm is able to handle attributes with discrete or continuous types. The selection of attributes in this algorithm uses entropy size, known as information gain, as a heuristic for selecting attributes that are the best part of the example in the class. All attributes are discrete value categories where attributes with continuous values must be discounted. Attribute discretization aims to facilitate the grouping of values based on predetermined criteria, and also to simplify the problems and improve the learning process accuracy [34].

The selection of attributes in the C4.5 algorithm using gain replaces the information gain value. The selection of a good attribute is an attribute that makes it possible to get the smallest decision tree size or attributes that can separate objects according to their class. Heuristically the attribute chosen is the attribute that produces the cleanest node. The cleanest size is expressed with the level of impurity, and to calculate it, can be done using the concept of entropy, entropy expresses the impurity of a collection of objects [35]. Based on Hansun 2017, there are four stages in carrying out the classification step using C4.5 algorithm [36], including:

1. Select attributes as roots.
2. Make a branch for each value.
3. Divide each case in a branch.
4. Repeat the process in each branch so that all cases in the branch have the same class.

Calculations start from counting the number of attributes and determining which attributes will be used as the root of the decision tree. Subsequently, Entropy and gain calculation will be carried out to form leaf from the decision tree. After calculations completed, a decision tree can be formed based on the previously calculated gain value. The attribute with the highest gain value will be located at a higher priority and has a higher position also in the decision tree. The formula for finding Entropy is as follows:

a) Entropy

Equation (2) shows the formulay on Entropy

\[
Entropy(S) = \sum_{j=1}^{k} -p_j \log_2 p_j
\]  

(2)

where \(S\) is dataset, \(K\) is number of \(S\) partitions, and \(p_j\) is the probability obtained from sum is divided by total cases.

b) Gain Ratio

Gain ratio can be found using (3)

\[
gain\ ratio(a) = \frac{gain(a)}{split(a)}
\]  

(3)

where \(a\) is the attribute, \(gain(a)\) is information gain in attribute \(a\), and \(split(a)\) is split information on attributes \(a\).

c) SplitInfo

SplitInfo on (3) can be calculated using (4)
\[
\text{SplitInf}o(S, A) = - \sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S}
\]

where \( S \) is the sample room used for training, \( A \) is the attribute, and \( S_i \) is the number of samples for attributes \( i \).

d) Gain

Finally, the Gain can be achieved using (5)

\[
\text{Gain}(A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)
\]

where \( S \) is the set of cases, \( A \) is the number of partition attributes \( A \), \(|S_i|\) is the number of samples for attribute \( i \), \(|S|\) is the number of all data samples, and \( \text{Entropy}(S_i) \) represent the entropy for samples that have values \( i \).

D. Output & Evaluation

The evaluation phase of the classification results in this study uses Confusion Matrix. Confusion Matrix is an evaluation method in the form of a matrix table that shows the performance of the classification model being tested. Confusion Matrix gives results in the form of numbers that show the amount of data that is successfully predicted correctly and the data that is not. This model is useful to know the accuracy, precision, recall of the algorithm model being tested. The Confusion Matrix model in the dataset has two label classes in Table 6.

The results of confusion matrix are useful for calculating the accuracy, precision, and recall of algorithm performance using the following formula:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \%
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \times 100 \%
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \times 100 \%
\]

Based on the results of the evaluation of the confusion matrix, the best classification results are obtained based on the highest value of accuracy, precision, and recall. Accuracy is used to calculate effectiveness and evaluate the performance of classification methods. Precision is used to calculate the level of accuracy between the information requested by the user and the answer given by the system. Whereas recall is the success rate of the system in rediscovering information.

Data classification sometimes does not only have two label classes, so it is different in determining positive classes and negative classes. There are several data that have more than two label classes. This case can use the confusion matrix multiclass classification evaluation method as shown in Table 7. In the confusion matrix multiclass classification there is an evaluation metrics formula that is different from confusion matrix binary classification. The accuracy formula, precision, and recall algorithm performance with the confusion matrix multiclass classification are as follows:

\[
\text{Akurasi} = \frac{\sum_{i=1}^{l} TP_i + TN_i}{TP + TN + FP + FN} \times 100 \%
\]

\[
\text{Precision} = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_i + FP_i)} \times 100 \%
\]

\[
\text{Recall} = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_i + FN_i)} \times 100 \%
\]

where \( TP_i \) is True Positive, which is the amount of positive data that is correctly classified by the system for class \( i \), \( TN_i \) is True Negative, which is the amount of negative data that are correctly classified by the system for class \( i \), \( FN_i \) is False Negative, which is the number of negative data but incorrectly classified by the system for class \( i \), \( FP_i \) is False Positive, that is the number of positive data but is incorrectly classified by the system for class \( i \), and \( l \) is the number of classes.
III. Results and Discussion

A. Research Results

This research gets results from the final stages of evaluation. In this evaluation process, compared the performance of Naïve Bayes classification algorithms and Decision Tree C4.5 with several preprocessing combinations performed. So, the best scenario combination can be found in the preprocessing SMOTE method and feature selection. This evaluation process determines the best algorithm between Naïve Bayes and Decision Tree C4.5 based on the value of accuracy to classify hospital readmissions of a diabetic patient. A comparison of accuracy can be seen in Table 8.

The results of Table 8 show that the best accuracy is in scenario 8 with the preprocessing method using combination of SMOTE and feature selection which classifies the two label classes. Decision Tree C4.5 algorithm is also a better algorithm for classifying hospital readmissions of diabetic patients with an accuracy of 82.73%. The results of the confusion matrix from each stage of the scenario are in Table 9 with the positives class uses for scenario with 3 class label is “>30” class.

The confusion matrix of the best results is in Scenario 8 C4.5 on Table 9, it shows the detail data that is successfully classified correctly and the amount of data that is incorrectly classified. From the results of the confusion, the matrix can also be calculated the values of evaluation metrics using (6) to (8) for binary class classification (9) to (11) for multiclass classification. The results of the evaluation metrics of Scenario 8 C4.5 on Table 9 as the best results show an accuracy of 82.74%, a precision of 87.1% and a recall of 82.7%. In more detail, the results of each trial are compared based on the evaluation values of the metrics. A comparison of the performance of all classification trial scenarios is shown in Figure 1 to Figure 3.

Based on the results shown in Figure 1, the comparison of the performance of the Naïve Bayes algorithm and the Decision Tree C4.5 based on the accuracy of each scenario has insignificant differences, but it can be seen that the accuracy value of the C4.5 algorithm is always better than the

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Preprocessing</th>
<th>Label</th>
<th>Naïve Bayes Accuracy</th>
<th>C4.5 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>No SMOTE &amp; Feature selection</td>
<td>3 Classes</td>
<td>59.47 %</td>
<td>59.68 %</td>
</tr>
<tr>
<td>2.</td>
<td>SMOTE</td>
<td>3 Classes</td>
<td>59.85 %</td>
<td>62.30 %</td>
</tr>
<tr>
<td>3.</td>
<td>Feature selection</td>
<td>3 Classes</td>
<td>59.28 %</td>
<td>60.85 %</td>
</tr>
<tr>
<td>4.</td>
<td>SMOTE + Feature selection</td>
<td>3 Classes</td>
<td>60.22 %</td>
<td>61.32 %</td>
</tr>
<tr>
<td>5.</td>
<td>Tanpa SMOTE &amp; Feature selection</td>
<td>2 Classes</td>
<td>75.61 %</td>
<td>77.58 %</td>
</tr>
<tr>
<td>6.</td>
<td>SMOTE</td>
<td>2 Classes</td>
<td>77.69 %</td>
<td>78.88 %</td>
</tr>
<tr>
<td>7.</td>
<td>Feature selection</td>
<td>2 Classes</td>
<td>76.39 %</td>
<td>77.58 %</td>
</tr>
<tr>
<td>8.</td>
<td>SMOTE + Feature selection</td>
<td>2 Classes</td>
<td>79.39 %</td>
<td>82.74 %</td>
</tr>
</tbody>
</table>
Naïve Bayes algorithm in each scenario. Significant differences in the value of accuracy are found in the performance of preprocessing applied to each scenario. Accuracy values look significantly different in the scenario 4 with the scenario 5, this is because in the scenario 1 to the scenario 4, the label class in the four scenarios uses three classes, thus increasing the data complexity and influencing the accuracy value of the Naïve Bayes and C4.5. Whereas in the scenario 5 to the scenario 8, all four scenarios use two label classes so that the low level of complexity makes it easier for the algorithm to classify the data.

Based on the results shown in Figure 2, the lowest precision results were obtained by Naïve Bayes classification in the scenario 3 with 55.4%, and the highest precision is obtained by classification C4.5 in the scenario 8 with 87.1%. Precision shows the results of the accuracy between the information requested and the results so that in the classification results C4.5 scenario 8, the accuracy of predictions with true classes gets the best results compared to other scenarios.

Based on the results shown in Figure 3, the comparison chart of recall values gives the best results in the scenario 8 trial using the C4.5 algorithm method. The recall value generated by the C4.5 algorithm when classifying the scenario 8 data is 82.7%. The recall is the result of data that can be recovered by the system. In C4.5 classification the scenario 8 can recover the desired data well compared to other scenarios.
B. Discussion

The comparative results of the SMOTE and feature selection show that combining the two preprocessing methods produce better performance than applying this method independently. Table 8 shows that the application of the SMOTE method independently shows better results than the feature selection method. While the feature selection method applies to data on diabetic patients tends not to increase accuracy significantly because the label class on the dataset is still imbalance. This shows that the imbalance of data has a negative effect on the performance of the classification in the case of diabetes patient data. However, the feature selection combined with SMOTE can produce excellent accuracy values.

SMOTE can overcome the imbalance of the data by adding new data to the minority class based on the value of the nearest neighbor so that it has properties similar to the minority class. New data were added at the SMOTE stage amounts to a majority of classes, so the label class is balanced. After the label class is balanced, the combination of feature selection methods will eliminate the attributes that are less relevant. Thus, the imbalance distribution of data and does not affect the performance of the algorithm or actually decreases accuracy. In the case of diabetes patients, the feature selection method is very useful, because the number of initial attributes is 47 attributes. With the selection, the feature can reduce complexity by eliminating some irrelevant attributes. Feature selection is also useful for anticipating the curse of dimensionality which can cause the classification accuracy at a certain point to decrease if the number of attributes is too much while the number of sample data is limited.

From the results of the experiments found in several tables above, it can be seen that the Decision Tree C4.5 algorithm has better results than the Naïve Bayes algorithm. The best results are found in...
scenario 8 with preprocessing treatment combining SMOTE and feature selection. In the trial scenario 8 using the C4.5 algorithm, the results obtained were the best results from another scenario trials with an accuracy of 82.74%, precision of 87.1% and recall of 82.7%.

The best results of scenario 8, shows that at the stage of applying SMOTE and feature selection in this scenario using 9 attributes from 47 attributes. Selected attributes in building the C4.5 model in scenario 8 are Admission Source, Time in Hospital, Number of emergency visits, Glucose serum test result, Replaginide, Glipzide, Glyburide, Rosiglitazone, and Readmitted.

The attributes selected using the feature selection can make the best decision tree because it contains high gain values and includes attributes that do not cause outliers. The highest gain value is the “time in hospital” attribute in the form of numerical data, then it is used as the root of the decision tree C4.5 and other attributes as branches of the specified value. The attribute “time in hospital” is considered relevant in this study because it provides enough information about whether diabetic patients need hospital readmissions with the total length of time patients to stay in the hospital. The attribute “admission source” is also an attribute that is considered relevant in classifying readmissions of diabetic patients because this data is useful for knowing the source of acceptance of these patients. Some drug dosage information attribute that have good data distribution on this dataset are replaginide, glipzide, glyburide, and rosiglitazone, so it can produce decision trees that have high accuracy.

IV. Conclusion

Based on the results of the discussion of this study it can be concluded that the application of several pre-processing methods can improve the performance of the tested algorithm so as to obtain maximum evaluation values. Combining several pre-processing methods are also recommended to improve accuracy and close weaknesses found in the data to be tested. The results of the application of the preprocessing method and without using preprocessing show very significant results, by using the preprocessing method the results have better accuracy. This study also shows better results than previous studies using the Naïve Bayes algorithm and also than studies using the Decision Tree C4.5 algorithm.

Acknowledgement

This research was supported by Universitas Negeri Malang and Al Quds Open University. We thank our colleagues from both institutions who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. We thank Dr. Aji P. Wibawa for assistance with suggestion in methodology and for comments that greatly improved the manuscript.

Declarations

A. Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

B. Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

C. Conflict of interest

The authors declare no conflict of interest.

D. Additional information

No additional information is available for this paper.

References


A. Riezka, Analisis dan Implementasi Data-Cleaning dengan Menggunakan Metode Multi-Pass Neighborhood (MPN), Universitas Telkom, 2011.


