JOIV : Int. J. Inform. Visualization, 6(4) - December 2022 798-808

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Classification of Student Graduation by Naïve Bayes Method by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection

Dony Fahrudy [a,*], Shofwatul 'Uyun [a]

[a] *Informatics Magister, Universitas Islam Negeri Sunan Kalijaga, Yogyakarta, 55281, Indonesia*

Corresponding author: [*]donnyfahrudi@gmail.com

***Abstract*— Class-imbalanced data with high attribute dimensions in datasets frequently contribute to issues in a classification process that can affect algorithms' performance in the computing process because there are imbalanced numbers of data in each class and irrelevant attributes that must be processed. Therefore, some techniques need to overcome the class-imbalanced data and feature selection to reduce data complexity and irrelevant features. Therefore, this study applied random oversampling (ROs) method to overcome the class-imbalanced data and two feature selections (information gain and forward selection) compared to determine which feature selection is superior, more effective, and more appropriate to apply. The feature selection results were then used to classify the student graduation by creating a classification model of Naïve Bayes algorithm. This study indicated an increase in the average accuracy of the Naïve Bayes method without the ROs pre-processing and the feature selection (81.83%), with the ROs (83.84%), with information gain with three selected features (86.03%) and forward selection with two selected features (86.42%); consequently, these led to increased accuracy of 4.2% from no pre-processing to information gain and 4.59% from no pre-processing to forward selection. Therefore, the best feature selection was the forward selection with two selected features (GPA of the 8th semester and the overall GPA), and the ROs, and both feature selections were proven to improve the performance of the Naïve Bayes method.**

***Keywords*— Forward selection; information gain; student graduation; naïve bayes; ROs.**

Manuscript received 24 Jun. 2022; revised 16 Jul. 2022; accepted 30 Oct. 2022. Date of publication 31 Dec. 2022.

I. INTRODUCTION

Student graduation is a success factor for a university and becomes one of the accreditation assessments for the university. If the students graduate on time in completing their studies, they can support the accreditation assessment [1]. One of the success criteria for students to obtain their bachelor's degree at a university is graduating on time or less than or equal to four years. In fact, they cannot always complete their studies in less than or equal to four years [2].

Based on datasets obtained at some universities, students cannot always complete their studies on time, so universities must anticipate this problem. Education authorities, academic administrators, and parents are concerned about this problem. The universities attempt to increase their on-time graduates; they need classification ability to accurately anticipate this problem to establish strategic programs to assist and improve the students' performance to graduate on time [3]. The universities need to improve their academic quality and optimize their resources to help students complete their studies on time [4]. This needs a model to classify the student graduation to determine the quality of the students [3]. The availability of training and testing data for each class is one of the criteria that determines the model's success. Several issues are often found in computing, the imbalanced data between classes and data with high attribute dimensions.

The problem of class imbalance can occur when the number of instances of one class exceeds the number of instances of the majority class and minority class, leading to the misclassification of the minority class that affects the classification results of the majority class. Generally, there are three resampling techniques: random undersampling, random oversampling, and hybrid methods [5]. To overcome the imbalance in the minority class, the random oversampling technique can be used by randomly replicating instances in the minority class [6].

Usually, classification algorithms develop models applying the features/attributes in the dataset, but all of these attributes are not included in the classification process. If applied on

data with extremely large sizes and dimensions, the algorithms' performance may be ineffective due to some irrelevant features that must be processed. Implementing feature selection is one approach to solving this issue. Feature selection is one of the stages in pre-processing the classification, by selecting attributes related to information that affects the classification results. It is also used to improve the effectiveness and efficiency of the classification algorithms' performance and to reduce dimensions of irrelevant data and features [7]. Its purpose is to reduce the complexity of a classification algorithm, increase the classification algorithms' accuracy, and identify attributes that affect the algorithms' performance [8]. A basic principle of feature selection is to search for all possible combinations of features to identify optimal features for classification [9].

There are three feature selection approaches, including *filter*, *wrapper,* and *embedded* approach [7]. The filter approach is a technique to select features based on ratings and to remove features below thresholds [10]. Simple ranking criteria are used in this approach to generate relevant features. Also, the filter approach applies a separate evaluation technique from the learning algorithm [11]. Then this approach is more independent, scalable for large data sets, fast, and easy to use [12]. However, it can ignore feature dependencies and lack the involvement of classifiers because it is applied independently [13].

Next, the wrapper approach utilizes a set of feature combinations from the search technique, trains a predictive model on the subset of features, and then determines performance accuracy by assessing the subset using supervised learning techniques. Each combination of features is compared, and the model of the algorithm is used to evaluate the set of existing combinations [14]. This approach solves problems in the filter method. This technique can communicate with the classifier learning model and focus on how attributes relate to one another [12]. It is optimized in conjunction with a classification learning algorithm; as a result, compared to the filter approach, it generally results in superior performance accuracy. However, compared to the filter approach, this method is computationally more expensive, more sophisticated, and longer in processing time [15].

Then the embedded approach focuses on the ideal feature subset for a particular classification algorithm by creating a classifier [16]. Feature selection in this method is influenced by the classifier's hypothesis and is not compatible with other certain classifiers because it depends on the classifier that makes appropriate decisions [15]. Therefore, this study compares the feature selection methods (filter and wrapper approach) and algorithms of information gain & forward selection.

Information gain is a machine learning approach commonly used as an attribute selection criterion. It specifically assigns a ranking to each existing feature (attribute) and removes features (attributes) that do not meet criteria arranged from the highest to the lowest value [17]. Features with high information gain values are better than other features, indicating that more attribute information is related to class [18].

Forward selection is the simplest search algorithm to reduce dataset dimensions by eliminating irrelevant or redundant attributes [19]. This method is a model that begins with zero variables (empty model) or no variables in the model, and then the variables are inserted one by one. The performance will be evaluated for each added variable, and only attributes with the highest performance are added to the selection for object functions until certain criteria are fulfilled [20]. The feature selection results are used to create a classification model of the Naïve Bayes algorithm to determine which feature selection method is better, more efficient, and more appropriate.

According to Vanaja and Kumar [7], data mining is collecting large amounts of data, and then the data are extracted into certain knowledge that can be applied. Classification is one of the techniques used in data mining to determine unidentified classes used to predict a particular class or label.

Classification is a type of analyzed data that can help to determine the class of the sample data that will be classified and to identify relationships between input features and target features (classes) [18]. It is widely used to predict a certain class by classifying data by creating models based on training data and to predict new unidentified classes from datasets by using models from classification to predict new data [21]. Ashari et al. [22] study on the performance of Decision Tree, Naïve Bayes, and K-Nearest Neighbor demonstrated that the Naïve Bayes algorithm had the best performance based on Precision, Recall, F measure, Accuracy, and AUC. The Naïve Bayes had a better Decision Tree and K-Nearest Neighbor on all parameters except precision, so the Naïve Bayes algorithm was used for prediction.

The Naïve Bayes algorithm is a machine learning technique that predicts the probability of class membership by using probability calculations and statistics [23]. Thomas Bayes, a British scientist, invented the Naïve Bayes method. Bayes' theorem states that the method can predict future probabilities based on historical data [24]. When used with large databases, it has high accuracy and speed [23].

In a study entitled Breast Cancer Mining-Based Feature Selection for Mammography, data mining from feature extraction from mammographic images was used in feature selection-based research to focus on the feature selection process. Decision Tree and rule induction were two algorithms used in the mining. Then the following classification algorithms, K-nearest Neighbors, Decision Tree, and Naïve Bayes, with a 10-fold cross-validation scheme and a stratified sampling strategy, were used to evaluate the features of the selected algorithms. The five descriptors selected were the best characteristics contributing to classifying benign and malignant tumors. The Decision Tree, generated using five features, obtained the best classification results with an accuracy of 93.18%, a sensitivity of 87.5%, a specificity of 3.89%, FPR of 6.33%, and TPR of 92.11% [25].

Another study used bootstrap and SMOTE methods in selecting features that were applied to the water quality status to deal with class imbalance. Meanwhile, experiments were conducted to remove noise on attributes by combining the filter method with some feature selection algorithms (information collection, correlation, rules, derivation, and chi-square). Based on the test results, the SMOTE bootstrap technique could increase the accuracy from 83.3% to 98.8%, according to 10-fold cross-validation. Meanwhile,

eliminating noise on data attributes could increase accuracy to 99.5% (using a subset of features generated by the Decision Tree method and information acquisition techniques) [26].

A study on graduation classification on student study results indicated that the pre-processing improved the results of the classification accuracy of the kNN algorithm. The data without pre-processing methods produced an accuracy of 72.28%, the pre-processed data using the K-means and Euclidean methods produced an accuracy of 98.42% (increased by 26.14%), and the K-means and Manhattan methods produced an accuracy of 97.76% (25.48% increase) [27]. Another study using Naïve Bayes on student graduation pointed out that the Naïve Bayes algorithm could predict student graduation with an accuracy of 94.92% on algorithm testing [28]. Another study applying Naïve Bayes with information gain on student performance in a national exam showed that the Naïve Bayes algorithm with information gain obtained an accuracy of 82.1%. [29]. Another study applying Naïve Bayes with forward selection on student academic performance showed that by applying some features, the predictive model of students' academic grades could perform better, so the accuracy of 94.43% was obtained with three selected features rather than only using Naïve Bayes (85.56%) [30].

Based on previous studies with the same case studies, the classification of student graduation only applied the Naïve Bayes algorithm without feature selections, while other previous studies only applied one feature selection method to each case study, such as information gain or forward selection without pre-processing methods to overcome class imbalance, but with a different case study, namely student performance on national exams and student academic scores. Therefore, this current study is more emphasizes computing. It applied three pre-processing methods for comparison, specifically Random Oversampling (ROs) to overcome class imbalances, feature selections with filter and wrapper approaches with algorithms of information gain and forward selection to determine the appropriate feature selection method, and Naïve Bayes classification algorithm to classify student graduation.

This study showed that most students, especially from departments of Industrial Engineering, Electrical Engineering, and Information Systems and Mathematics, could not complete their studies on time. Therefore, it affects the accreditation assessment; therefore, the university must pay attention to the quality of its institution that becomes an accreditation assessment to prevent the student academic failure. Based on these problems, the researchers conducted a study entitled "Classification of Student Graduation by Naïve Bayes Method by Comparing between Random Oversampling and Selection Features of Information Gain and Forward Selection". This study is to classify the student graduation to determine student academic performance, to analyze various features to determine relevant features for classification data analysis, to identify feature selections which are the best in classifying the student graduation, and to understand the performance of the Naïve Bayes algorithm before and after applying random oversampling and feature selections.

II. Material and Method

In general, this research aims to reduce the complexity of data and irrelevant features, improve the performance of classification algorithms, and develop classification models using experimental-based data mining carried out with feature selection. The stages in the study are shown in Fig. 1.

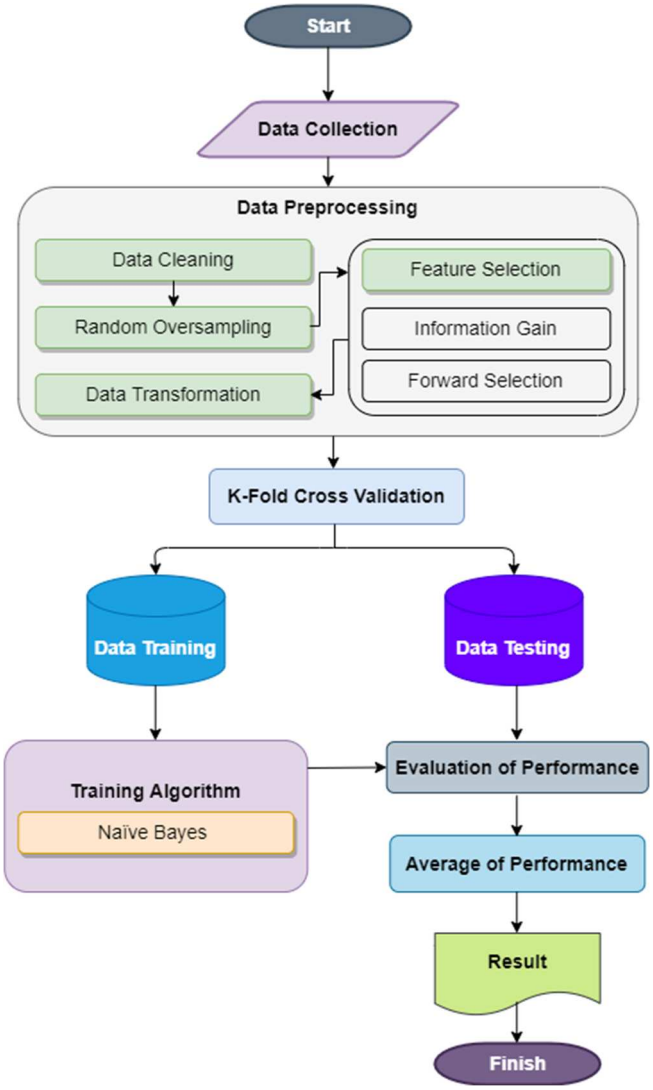

Fig. 1 Research Method

The stages of this study, Fig. 1, are explained in the following sub-sections.

A. Data Collection

The datasets in this study, student academic data, were obtained from the Center for Information Technology and Database (PTIPD) of Universitas Islam Negeri Sultan Syarif Kasim Riau. The datasets were student academic data from 2016–2019 who had graduated, and there were 1420 student data consisting of 272 student data who graduated on time and 1148 student data who did not graduate on time. The datasets were used as training data by creating a model of classification algorithms after applying random oversampling and feature selections with selected attributes and then to produce knowledge patterns used to classify the student graduation.

B. Data Preprocessing

The pre-processing stage was significant in the data analysis to improve the data quality and overcome issues in the data. Several stages of pre-processing on datasets were data cleaning, resampling, feature selections, and data transformation, as elaborated in the following [31].

1) Data Cleaning: Data cleaning was the first stage in the pre-processing stage to make improvements to the data of student graduation because raw data tend to be not ready to be used, such as missing values of the data originating from data with attributes without value or information, so these need to be addressed. The missing values could be solved by replacing the missing data based on the previous data on each attribute of identified attribute values [32].

2) Resampling: Resampling was a technique used to tackle class-imbalance data. The class-imbalanced data might occur when the number of instances of one class exceeded the number of instances (majority class) of another class (minority class), leading to misclassification in the minority class contributing to biased data on the results of the classification of the majority class [6].

The resampling techniques could be categorized into three methods: random undersampling, random oversampling and hybrid method. The random undersampling created a subset of the original data set by eliminating instances of majority class; the random oversampling created a superset of the original data set by replicating multiple instances or creating new instances from existing ones (minority class instances); and the hybrid method combined both sampling methods [5].

The random oversampling could be applied to deal with the imbalance of the minority class. It tried to balance the distribution of classes randomly replicating instances of the minority class [6].

The steps of random oversampling were as follows [33]:

- Insert datasets.
- Count the number of majority class and minority class.
- Calculate the difference (deviation) with Eq. (1):

$$\text{Deviation} = \text{the number of majority class} - \text{the number of minority class} \quad (1)$$

- Initialize i = 1 as a looping index.
- Check conditions. If i was <= difference, duplicate the minority class randomly i = i + 1. Otherwise, combine the remaining majority class with the minority class as balanced datasets.

3) Feature Selections: In the pre-processing stage, feature selections were performed by applying information gain and forward selection algorithms that will be compared then. The feature selections were in the following.

- Information Gain

Information gain is a criterion for selecting attributes to determine the limits of the roles of an attribute. It specifically ranked each feature and removed features that did not meet the criteria arranged from the highest to the lowest value. It had a value obtained from the total value of the entropy for all criteria on the feature diminished by the entropy of each criterion. Entropy was diversity; the higher the entropy value, the better the diversity in the data. The value measurement was to identify whether certain attributes would be used or not. The next step was to use the attributes that meet the weighting requirements [34].

The feature selection with information gain consists 3 stages, including [35] :

1. Calculate the information gain value for each attribute in the original dataset.
2. Determine the expected threshold that allows attributes with a weight equal to the limit or greater the limit.
3. Improve the datasets by reducing the attributes.

The feature selection steps with information gain are in the following [36].

1. Separate each attribute according to its class or label on the datasets.
2. Calculate the total entropy for all criteria of each attribute with the Eq. (2) :

$$\text{Entropy}(S) = \sum_{i=1}^{m} - p(I) \log_2 p(I) \quad (2)$$

Entropy(S) was the total entropy for all criteria on an attribute; S was the set of all cases (datasets); m was the number of criteria in S and; and p(I) was the ratio of the number of samples in class I to the total sample in the datasets.

3. Calculate the information gained after determining the entropy of each record. The criterion for selecting the attributes was information gain. The calculation was with the Eq. (3) :

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{v \in A} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (3)$$

Gain (S,A) was the Information gain for attribute A; Entropy(S) was the total entropy value for all criteria on an attribute; S was the set of all cases (datasets); A was the Attribute; v was the possible value for attribute A; |S| was the sum of all data samples; |Sv| was the number of data samples for the attribute criteria of v; and Entropy(Sv) was the Entropy for each criterion value of v.

4. Sort the attributes by value in descending order (high to low).
5. Take the needed top-ranking attribute with the determined threshold.

- Forward Selection

Forward selection is a technique to reduce dataset dimensions by eliminating irrelevant or redundant attributes [19]. It was a model that started with zero variables (empty model) or no variables in the model, and then the variables were inserted one by one. Performance would be evaluated for each added variable, and only the highest-performing attributes were added to the selection for object functions until certain criteria were met [20].

Features or variables with a large number of datasets having irrelevant information could reduce the performance of the classification algorithms in predicting a certain class. The forward selection was used to select relevant features for data influencing the classification results, reducing data dimensions, and increasing the effectiveness and efficiency of the classification algorithm performance [30].

The feature selection with forward selection, generally, is in the following [37] :

1. The iterative forward selection method started with having no features/variables in the model or zero variables (empty model).
2. In each iteration, the addition of the most optimal or significant features continued and improved the model's performance until the addition of new features did not improve the model's performance.

The steps of feature selection with forward selection on Naïve Bayes are in the following [37].

1. In the datasets, use only one variable to train the model with the Naïve Bayes algorithm. Each variable was used to train the model separately by using target variables (predictor).
2. Perform the testing process for each variable after the training stage to identify the performance of Naïve Bayes algorithm.
3. A variable with the highest performance was a variable with optimal features after testing, so the variable was added to the selected feature.
4. Then add one more variable besides the previously selected variable and train the model separately by using the target variables so that the test accuracy was obtained.
5. A variable with the highest performance would be retained.
6. If the process did not stop at the stopping criterion, repeat the step of adding variables until there was no significant increase in performance (adding features) so that a model with optimal or significant selected features would be obtained.
7. After obtaining the optimal features in the testing process on the variables, the final accuracy would be obtained from the feature selection by using the forward selection - Naïve Bayes algorithm.

4) Data Transformation

Data transformation is changing the original data's measurement scale so that the data can meet appropriate analytical assumptions for processing. In this case, data transformation was performed by data conversion and normalization as follows.

- Data Conversion

Data conversion is a technique of converting string data into numbers or called encoding. After the data was cleaned, the data with nominal types were converted to data with numeric type to accelerate the data normalization process [38].

- Data Normalization

The data were mapped into ranges by using a normalization approach. The normalized data were all datasets converted to numeric data. The data were normalized according to the scale range. The data normalization process was performed to balance the data values if the range of data values had a significant difference and to accelerate the next process. The normalization applied the min-max normalization calculation in Eq. (4).

According to Albarak et al. [39], data normalization have some purposes as follows:

1. To reduce data duplication
2. To reduce data redundancy
3. To reduce complexity
4. To facilitate data modification

The data were normalized by changing the data values into a data range value of 0 to 1. The min-max normalization calculation is as follows [40] :

$$x_n = \frac{x_0 - x_{min}}{x_{max} - x_{min}} \quad (4)$$

Xn was the new value for the variable X; X0 was the old value; Xmin was the lowest value in the datasets; and Xmax was the highest value in the datasets.

C. K-Fold Cross Validation

A statistical method of testing learning algorithms with data divided into two parts is known as cross-validation. The 1st data in the first subset of 10-fold cross-validation were used to validate the model, and the 2nd to 10th data were used to study the model by training it. Each training and validation set should be cross-linked sequentially so that each data has a chance to be validated. The most common type of cross-validation is k-fold cross-validation. Sampling in the cross-validation was performed in various ways so that no two sets of tests overlap. The available learning sets were partitioned into k separate subsets of approximately equal size. The default value of k was 10. The number of sets was part of the fold. The model was trained by using the k-1 subset as the training set. The model was applied to the available subset as a validation set, and then the performance was measured to determine its accuracy. The steps would be repeated until every k-part set functioned as a validation set. The cross-validation performance could be seen from the average of k performance measurements on k validation sets [41].

The datasets were divided into k subsets, each with the same data points. The 10-fold cross validation method is depicted in Fig. 2 with k equal to 10. The validation set was the first subset of the first fold, while the training set was the second to the tenth subset. The second subset in the second fold was the validation set, and the other subset was the training set, and so forth [42].

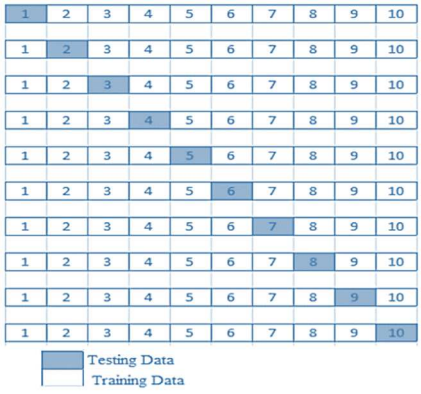

Fig. 2 Schema of 10-fold cross-validation

D. Training Data

In the training process, the stage performed was creating a model on the Naïve Bayes algorithm for the student graduation datasets. The Naïve Bayes algorithm was to predict the probability of class membership by using probability and statistical calculations [23].

Below are the training stages on the Naïve Bayes algorithm, as shown in Fig. 3.

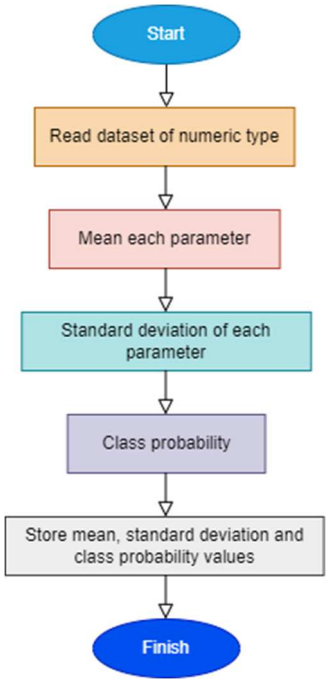

Fig. 3 Training on Naïve Bayes

The following is the training algorithm of this study.

1) Naïve Bayes: Thomas Bayes, a British scientist, invented the Naïve Bayes method. Bayes theorem states that the Naïve Bayes method can predict future probabilities based on historical data [24]. The Naïve Bayes algorithm for numeric/continuous type data was calculated by the Gaussian Distribution/Gaussian Density in Eq. (5) [24].

$$P(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2_{ij}}} \quad (5)$$

P was the probability; Xi was the i attribute; xi was the value of the i attribute; Y was the sought class, yj was the sought sub class Y; μ was the average of all attributes (mean), σ was the variant of all attributes (standard deviation).

The flow of training process of the Gaussian distribution on the Naïve Bayes algorithm can be seen in the following [43] :

1. Read the datasets.
2. Calculate the mean and standard deviation (attribute variant) for numeric/continuous data types:
   a. Calculate the mean and standard deviation (attribute variant) of each parameter which was numeric data.
   The formula used to calculate the mean was by the Eq. (6) as follows:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6)$$

   μ was the mean; xi was the value of the –i sample; n was the number of samples.
   The formula for calculating the standard deviation was by the Eq. (7) as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}} \quad (7)$$

   σ was standard deviation; xi was the value of x to – I; μ was the calculated mean; n was the number of samples.
   b. Calculate the class probability value by calculating the number of appropriate data from the same category divided by the number of data in all categories.
3. Save the mean, standard deviation and class probability values.

E. Data Testing

In the testing process, testing on the model obtained from the training results was performed. The class obtained in the test results was compared with the actual result class at the classification stage so that the accuracy of the test was obtained.

The following are testing stages on the Naïve Bayes, as shown in Fig. 4 [43].

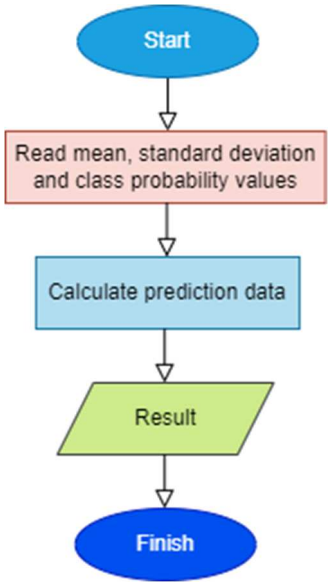

Fig. 4 Testing on Naïve Bayes

The flow of a testing process of the Gaussian distribution on the Naïve Bayes algorithm is in the following [43] :

1. Read the training stage's mean, standard deviation and class probability values.
2. Calculate the testing data by calculating the probability of the Gaussian distribution based on the values in the table of the mean, standard deviation and probability of each attribute, and then multiply all the calculated probability values on the attributes based on their class.
3. The greatest value in the class/label was the result.

The following is testing on the performance of this study.

1) Evaluation of Performance: Performance evaluation on the data mining method was conducted using a confusion matrix. The matrix was used to assess the classification model and to determine whether an object was true or false [44]. It illustrated the accuracy of the solution for the classification. It provided information on the actual and expected values of

the classification and was intended to be compared with the initial input class. Data from the matrix were used to assess the performance of the algorithm [45]. The confusion matrix for binary classification is shown in Fig. 5 below [46].

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Fig. 5 Confusion Matrix for Binary Classification

The confusion matrix for the binary classification had a dimension of 2 x 2 (Fig. 5), and one label was considered positive and the other label was negative. The matrix elements were classified into true positive (TP), true negative (TN), false positive (FP), and false negative (FN) based on the prediction label (positive, negative) and the comparison of the prediction with the actual class (true, false). In this case, the confusion matrix for binary classification was used to evaluate the performance of the algorithms. The following is how to calculate accuracy (8), precision (9) and recall (10) in the metric specified for the confusion matrix for binary classification [46].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

2) Average of Performance: Performance testing produced accuracy, and the accuracy was obtained from the confusion matrix on the k-fold cross-validation, a model validation method on the learning algorithm. The accuracy was obtained for each set of validation algorithms for each fold. For validation in fold-1 to fold-10, the average accuracy of all the folds was the final accuracy of the algorithm performance measurement as the average performance [41].

III. Results and Discussion

Based on the stages in the research methods above, the results achieved in the study include data collection, data pre-processing, k-fold cross-validation, and performance evaluation charts that are discussed in the following section.

A. Data Collection

The data collection was a stage of collecting data obtained from PTIPD Universitas Islam Negeri Sultan Syarif Kasim Riau in the form of student graduation datasets. The datasets were 1420 student graduation academic data from 2016–2019 consisting of 272 students who graduated on time and 1148 students who did not graduate on time. The datasets had 22 attributes (features), namely NIM (student number), Gender, Place of Birth, Place of Residence, Transportation Means, Year of Class, Major, Father's Income, Father's Occupation, Father's Education, Mother's Income, Mother's Occupation, Mother's Education, GPA of Semester 1, GPA of Semester 2, GPA of Semester 3, GPA of Semester 4, GPA Semester 5, GPA of Semester 6, GPA of Semester 7, GPA of Semester 8, and overall GPA; and they had class/target in the form of student graduation status consisting of 2 classes, graduating on time and graduating not on time. The datasets are illustrated in Table I.

TABLE I
DATASETS

No	Student Number	Gender	Place of Birth	Class
1	10952005554	M	Pekanbaru	Not on time
2	10952006782	M	pulau sarak	Not on time
3	10952006799	M	Siak	Not on time
4	10952006926	M	bangun rejo	Not on time
5	10952008054	M	sungai pinang	Not on time
...	...	...	...	...
1420	11555202644	F	Duri	On time

B. Data Preprocessing

The pre-processing stage was significant in the data analysis to improve quality and overcome issues in the data. Steps in pre-processing datasets include data cleaning, resampling, feature selection and data transformation.

1) Data Cleaning: The data cleaning was to improve the student graduation data with missing values. Missing values were data that had no value or information and needed to be addressed. These could be solved by replacing the missing data (empty) based on the previous data on each attribute of the identified attribute values. The following are the cleaning results on the datasets, as shown in Table II.

TABLE II
TABLE OF THE CLEANING RESULTS

No	Student Number	Gender	Place of Birth	Class
1	10952005554	L	Pekanbaru	Not on time
2	10952006782	L	pulau sarak	Not on time
3	10952006799	L	Siak	Not on time
4	10952006926	L	bangun rejo	Not on time
5	10952008054	L	sungai pinang	Not on time
...	...	...	...	...
1420	11555202644	P	Duri	On-time

2) Resampling: Resampling was to overcome class imbalances that occurred when instances from one class exceeded the number of instances (majority class) from another class (minority class), thus leading to misclassification in the minority class, influencing the results of the majority class classification. To deal with the minority class imbalance, random oversampling technique was applied. Random oversampling was to balance the distribution of classes that randomly replicate instances of the minority class. The following are the results of resampling with the random oversampling, as shown in Table III.

TABLE III
RESULTS OF RANDOM OVERSAMPLING

No	Student Number	Gender	Place of Birth	Class
1	10952005554	L	Pekanbaru	Not on time
2	10952006782	L	Pulau sarak	Not on time
3	10952006799	L	Siak	Not on time
4	10952006926	L	Bangun Rejo	Not on time
5	10952008054	L	Sungai Pinang	Not on time
...	...	...	...	...
2296	11652103418	L	Pulau Tengah	On time

The following is the data distribution for each class before and after applying random oversampling as shown in Table IV and Table V.

TABLE IV
THE DATA NUMBER OF EACH CLASS

No	Class	Data Number
1	On time	272
2	Not on time	1148
Total		1420

TABLE V
THE DATA NUMBER OF EACH CLASS–RANDOM OVERSAMPLING

No	Class	Data Number
1	On time	1148
2	Not on time	1148
Total		2296

3) Feature Selections: At this pre-processing stage, feature selection was performed by applying algorithms of information gain and forward selection that were compared to identify their comparison.

- Information Gain

Information gain was an attribute selection method to rank each feature from the highest to the lowest value. This value measurement was to determine the attributes to be used or not. Attributes that met the weighting criteria based on the threshold were used for the next process. The following is the parameter of the selection results as shown in Table VI.

TABLE VI
THE PARAMETER OF SELECTION RESULTS WITH INFORMATION GAIN

Threshold	Number of Attributes	Attribute
0.439	3	GPA of Semester 8, GPA Semester of 7, dan Overall GPA

The following are the results of attribute selection on datasets with the information gain as shown in Table VI.

TABLE VI
SELECTION RESULTS WITH INFORMATION GAIN

No	Student Number	GPA of Semester 8	GPA of Semester 7	Class
1	10952005554	0.90	2.01	Not on time
2	10952006782	2.02	1.78	Not on time
3	10952006799	1.36	3.36	Not on time
4	10952006926	1.26	2.96	Not on time
5	10952008054	1.45	3.26	Not on time
...	...	...	...	...
2296	11652103418	3.70	4.00	On time

- Forward Selection

Forward selection was to reduce the dimensions of the dataset by eliminating irrelevant or redundant attributes. It was a model that started with zero variables (empty model) or no variables in the model, and then the variables were inserted one by one. Performance was evaluated for each added variable. Only the highest-performing attributes were added to the selection and retained for object functions until certain criteria were fulfilled. The following is the parameter of the selection results, as shown in Table VII.

TABLE VII
THE PARAMETER OF SELECTION RESULTS WITH FORWARD SELECTION

Number of Attribute	Attribute
2	GPA of Semester 8, Overall GPA

The following are the results of attribute selection on datasets with the forward selection, as shown in Table VIII.

TABLE VIII
SELECTION RESULTS WITH FORWARD SELECTION

No	Student Number	GPA of Semester 8	Overall GPA	Class
1	10952005554	0.90	2.88	Not on time
2	10952006782	2.02	2.83	Not on time
3	10952006799	1.36	3.49	Not on time
4	10952006926	1.26	3.38	Not on time
5	10952008054	1.45	3.10	Not on time
...	...	...		...
2296	11652103418	3.70	2.69	On time

4) Data Transformation: Data transformation was changing the measurement scale of the original data so that the data could meet the appropriate analytical assumptions for processing. In this case, the data transformation was processed by conversion and normalization.

- Data Conversion

Data conversion was to convert the string data into numbers (encoding). After the data were cleaned, the data with nominal type were converted to numeric type to facilitate the data normalization process. The following are the data conversion results, as shown in Table IX.

TABLE IX
TABLE OF CONVERSION RESULTS

No	Student Number	Gender	Place of Birth	Class
1	10952005554	0	332	Not on time
2	10952006782	0	636	Not on time
3	10952006799	0	448	Not on time
4	10952006926	0	564	Not on time
5	10952008054	0	655	Not on time
...	...	...	...	...
2296	11652103418	0	299	On time

- Data Normalization

The normalized data were all datasets converted to numeric data. The data were normalized according to a scale range from 0 to 1, to balance the data values and to facilitate the next process, the Min-max Normalization calculation. The following are the results of data normalization as shown in Table X.

TABLE X
TABLE OF NORMALIZATION RESULTS

No	Student Number	Gender	Place of Birth	Class
1	10952005554	0.0	0.4977511244377811	Not on time
2	10952006782	0.0	0.9535232383808095	Not on time
3	10952006799	0.0	0.671664167916042	Not on time
4	10952006926	0.0	0.8455772113943029	Not on time
5	10952008054	0.0	0.9820089955022488	Not on time
...	...	...	...	...
2296	11652103418	0.0	0.4482758620689655	On time

C. K-Fold Cross Validation

The data distribution was processed on the datasets contained in Table X after the normalization process using the k-fold cross-validation technique. The number of used data was 2296 student graduation data. The data were divided into training data and testing data (validation). Training data were student graduation data used for the classification model development process applied as a match with testing data. Meanwhile, testing data were student graduation data used as a test to evaluate the model applied as a match to the training data. Of the 2296 data, 2290 were used randomly, consisting of 1145 student data who graduated on time and 1145 student data who did not graduate on time divided into ten subsets using the k-fold cross-validation. Therefore, the data used as training data were 2061 data, and the data used as test data were 229 data in each subset. The following is the distribution of test data using the 10-fold cross-validation.

TABLE XI
DISTRIBUTION OF TRAINING DATA AND TESTING DATA

229									
	229								
		229							
			229						
				229					
					229				
						229			
							229		
								229	
									229

Notes:

	=	*Training Data*
	=	*Testing Data*

Dividing the data into two parts, k-fold cross-validation was to evaluate the learning method, the 1st data in the first subset of 10-fold cross-validation was to validate the model, and the 2nd to10th was to study the model by training the data by crossing each other sequentially, thereby allowing each data to be validated with 10-fold cross-validation for 10 times. The average of k performance on the k validation sets was cross-validated performance.

The following is a comparison of the performance of the Naïve Bayes algorithm before and after applying random oversampling (ROs) and feature selections with Information Gain (IG) and Forward Selection (FS) as shown in Table XII.

TABLE XII
PERFORMANCE OF NAÏVE BAYES

Fold	Accuracy (%)	Precision (%)	Recall (%)	Not
1	88.03	99.01	86.21	Without
2	87.32	98.02	86.09	ROs
3	83.10	95.79	81.98	
4	85.92	95.96	85.59	
5	82.39	97.85	79.82	
6	77.46	93.26	76.15	
7	80.99	93.07	82.46	
8	77.46	94.79	77.12	
9	77.46	93.94	78.15	
10	78.17	100.00	74.38	
Performance Average	**81.83**	**96.17**	**80.80**	
1	83.41	90.63	75.00	ROs
2	82.97	86.81	74.53	
3	83.41	91.00	75.83	
4	79.91	91.58	69.60	
5	87.34	90.22	80.58	
6	83.41	88.07	79.34	
7	86.03	95.51	75.22	
8	81.66	89.00	74.17	
9	84.72	88.76	75.96	
10	85.59	91.18	79.49	
Performance Average	**83.84**	**90.28**	**75.97**	
1	88.21	88.03	88.79	Feature
2	82.97	78.63	86.79	IG
3	85.59	82.71	91.67	
4	85.15	86.40	86.40	
5	87.34	84.26	88.35	
6	86.03	85.04	89.26	
7	88.65	89.19	87.61	
8	86.03	85.48	88.33	
9	83.84	78.63	88.46	
10	86.46	82.58	93.16	
Performance Average	**86.03**	**84.10**	**88.88**	
1	88.21	88.03	88.79	Feature
2	82.97	78.63	86.79	FS
3	85.59	83.21	90.83	
4	84.72	86.29	85.60	
Fold	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	
5	87.34	84.26	88.35	
6	86.90	85.27	90.91	
7	89.52	89.38	89.38	
8	86.03	84.38	90.00	
9	86.03	80.51	91.35	
10	86.90	82.71	94.02	
Performance Average	**86.42**	**84.27**	**89.60**	

D. Performance Evaluation Chart

Performance evaluation on the data mining method applied a confusion matrix to evaluate the classification model. Performance testing produced accuracy obtained from the confusion matrix on the k-fold cross-validation method, which was a model validation method on the learning algorithms. In each validation set in the algorithms, the accuracy of each fold was obtained, for validation in fold-1 to fold-10; the average accuracy in all folds was the final accuracy of the algorithms' performance measurement as the average performance. Based on the average accuracy obtained, Table IX is visualized by using a line graph, as shown in Fig. 6.

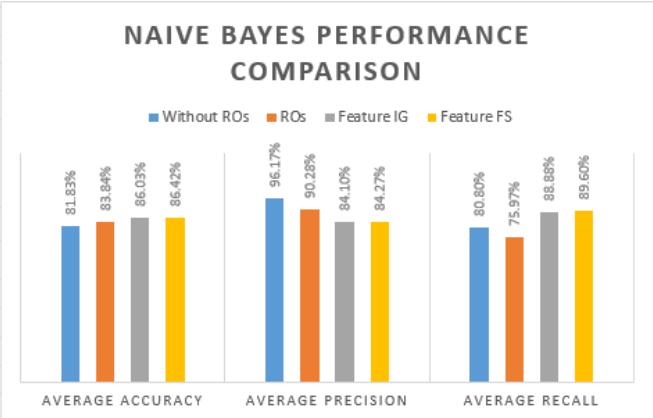

Fig. 6 Performance Comparison Chart of Naïve Bayes

IV. CONCLUSION

The results of this study indicated that this study could classify students' graduation to determine their academic performance, analyze various features to identify relevant features, determine which feature selection is the best in classifying the students' graduation and compare the performance of the Naïve Bayes algorithm before and after applying the Random Oversampling and the feature selections.

The Naïve Bayes method could be used to classify the student graduation. Testing on the Naïve Bayes algorithm was to compare classification performance before and after applying random oversampling and feature selections of information gain and forward selection. Based on the performance testing, it was found that the classification algorithms had a good performance on the student graduation datasets. The Naïve Bayes algorithm had improved performance after applying random oversampling and feature selections. The forward selection had a better performance in classifying data, so it could reduce data complexity and irrelevant features and improve the performance of the Naïve Bayes algorithm.

The Naïve Bayes algorithm indicated 81.83% of accuracy without random oversampling and feature selection with k-fold cross validation, 83.84% of accuracy with random oversampling, 86.03% of accuracy with information gain with 3 selected features (GPA of semester 8, GPA of semester 7 and Overall GPA), 86.42% of accuracy with forward selection with 2 selected features (GPA of semester 8 and overall GPA), thereby leading to increased accuracy of 4.2% from no pre-processing to information gain and 4.59% from no pre-processing to forward selection. Therefore, the pre-processing stages of random oversampling and feature selection influenced the Naïve Bayes algorithm, and the forward selection had a better performance with 2 selected attributes in classifying data adequately, thereby improving the performance of the Naïve Bayes algorithm.

The suggestion for further research is to apply other pre-processing methods such as outlier detection and feature selection, with other approaches (except filter and wrapper approach), such as embedded approach to determine the most influential feature selection on the classification. In addition, it is also expected to test classification methods other than Naïve Bayes to determine which classification method has the best performance.

REFERENCES

[1] ACCJC/WASC, "Guide To Evaluating Institutions," p. 56, 2012, [Online]. Available: www.g-fras.org.

[2] Nuffic, "Education system Indonesia," 2017, [Online]. Available: www.nuffic.nl/en/home/copyright.

[3] J. S. Bassi, E. G. Dada, A. A. Hamidu, and M. D. Elijah, "Students Graduation on Time Prediction Model Using Artificial Neural Network," *IOSR J. Comput. Eng.*, vol. 21, no. 3, pp. 28–35, 2019, doi: 10.9790/0661-2103012835.

[4] C. Lei and K. F. Li, "Academic Performance Predictors," *Proc. - IEEE 29th Int. Conf. Adv. Inf. Netw. Appl. Work. WAINA 2015*, pp. 577–581, 2015, doi: 10.1109/WAINA.2015.114.

[5] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A Novel Ensemble Method for Imbalanced Data Learning," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–11, 2017.

[6] S. Naganjaneyulu and M. R. Kuppa, "A novel framework for class imbalance learning using intelligent under-sampling," *Prog. Artif. Intell.*, vol. 2, no. 1, pp. 73–84, 2013, doi: 10.1007/s13748-012-0038-2.

[7] S. Vanaja and K. Ramesh Kumar, "Analysis of Feature Selection Algorithms on Classification: A Survey," *Int. J. Comput. Appl.*, vol. 96, no. 17, pp. 29–35, 2014, doi: 10.5120/16888-6910.

[8] C.Deisy, B.Subbulakshmi, D. S.Baskar, and Dr.N.Ramaraj, "Efficient Dimensionality Reduction Approaches for Feature Selection," *Proc. - Int. Conf. Comput. Intell. Multimed. Appl. ICCIMA 2007*, vol. 4, pp. 270–272, 2007, doi: 10.1109/ICCIMA.2007.288.

[9] S. L. Ting, W. H. Ip, and A. H. C. Tsang, "Is Naïve bayes a good classifier for document classification?," *Int. J. Softw. Eng. its Appl.*, vol. 5, no. 3, pp. 37–46, 2011.

[10] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[11] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," *2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014*, no. 1997, 2015, doi: 10.1109/ICCIC.2014.7238499.

[12] M. W. Mwadulo, "A Review on Feature Selection Methods For Classification Tasks," *Int. J. Comput. Appl. Technol. Res.*, vol. 5, no. 6, pp. 395–402, 2016, doi: 10.1109/ICISC.2017.8068746.

[13] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 971–989, 2016, doi: 10.1109/TCBB.2015.2478454.

[14] D. A. A. Gnana, Singh, S. A. Balamurugan, and E. J. Leavline, "Literature Review on Feature Selection Methods for High-Dimensional Data," *Int. J. Comput. Appl.*, vol. 136, no. 1, pp. 9–17, 2016.

[15] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egypt. Informatics J.*, vol. 19, no. 3, pp. 179–189, 2018, doi: 10.1016/j.eij.2018.03.002.

[16] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2019, doi: 10.1016/j.jksuci.2019.06.012.

[17] A. Khemphila and V. Boonjing, "Heart disease classification using neural network and feature selection," *Proc. - ICSEng 2011 Int. Conf. Syst. Eng.*, no. 2007, pp. 406–409, 2011, doi: 10.1109/ICSEng.2011.80.

[18] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," *Pertanika J. Sci. Technol.*, vol. 26, no. 1, pp. 329–340, 2018.

[19] R. Panthong and A. Srivihok, "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm," *Procedia Comput. Sci.*, vol. 72, pp. 162–169, 2015, doi: 10.1016/j.procs.2015.12.117.

[20] P. Kalyani and D. M. Karnan, "Attribute Reduction using Forward Selection and Relative Reduct Algorithm," *Int. J. Comput. Appl.*, vol. 11, no. 3, pp. 8–12, 2010, doi: 10.5120/1564-1499.

[21] A. H. Seh, "A Review on Heart Disease Prediction Using Machine Learning Techniques A Review on Heart Disease Prediction Using Machine Learning Techniques," vol. 9, no. April, pp. 208–224, 2019.

[22] A. Ashari, I. Paryudi, and A. M. Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 11, pp. 33–39, 2013.

[23] K. Artaye, "International Conference On Information Technology And Business ISSN 2460-7223 Implementation of Naïve Bayes Classification Method to Predict Graduation Time of IBI Darmajaya Scholar Z . A . Pagar Alam Street No . 93 Bandar Lampung," no. August, pp. 284–290, 2015.

[24] A. Ali, A. Khairan, F. Tempola, and A. Fuad, "Application Of Naïve Bayes to Predict the Potential of Rain in Ternate City," *E3S Web Conf.*, vol. 328, p. 04011, 2021, doi: 10.1051/e3sconf/202132804011.

[25] S. Uyun and L. Choridah, "Feature selection mammogram based on breast cancer mining," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 1, pp. 60–69, 2018, doi: 10.11591/ijece.v8i1.pp60-69.

[26] S. Uyun and E. Sulistyowati, "Feature selection for multiple water quality status: Integrated bootstrapping and SMOTE approach in imbalance classes," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 4, pp. 4331–4339, 2020, doi: 10.11591/ijece.v10i4.pp4331-4339.

[27] S. Sugriyono and M. U. Siregar, "Preprocessing kNN algorithm classification using K-means and distance matrix with students' academic performance dataset," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 4, pp. 311–316, 2020, doi: 10.14710/jtsiskom.2020.13874.

[28] C. Mulyadi and . Sukron, "Prediction of Timeliness of Graduating with Naïve Bayes Algorithm," no. Icri 2018, pp. 3043–3050, 2020, doi: 10.5220/0009946430433050.

[29] H. Z. Hashemi, P. Parvasideh, Z. H. Larijani, and F. Moradi, "Analyze students performance of a national exam using feature selection methods," *2018 8th Int. Conf. Comput. Knowl. Eng. ICCKE 2018*, no. Iccke, pp. 7–11, 2018, doi: 10.1109/ICCKE.2018.8566671.

[30] A. Saifudin, Ekawati, Yulianti, and T. Desyani, "Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes," *J. Phys. Conf. Ser.*, vol. 1477, no. 2, 2020, doi: 10.1088/1742-6596/1477/3/032007.

[31] G. Shivali, Joni Birla, "Knowledge Discovery in Data-Mining," *Int. J. Eng. Res. Technol.*, vol. 3, no. 10, pp. 1–5, 2015, [Online]. Available: https://www.ijert.org/research/knowledge-discovery-in-data-mining-IJERTCONV3IS10051.pdf.

[32] J. W. Grzymala-Busse and W. J. Grzymala-Busse, "Handling Missing Attribute Values," *Data Min. Knowl. Discov. Handb.*, pp. 33–51, 2010, doi: 10.1007/978-0-387-09823-4_3.

[33] O. Heranova, "Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring," vol. 1, no. 10, pp. 10–12, 2021.

[34] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 612–619, 2020, doi: 10.14569/ijacsa.2020.0110277.

[35] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00472-4.

[36] Y. Wang, Y. Li, Y. Song, X. Rong, and S. Zhang, "Improvement of ID3 algorithm based on simplified information entropy and coordination degree," *Algorithms*, vol. 10, no. 4, pp. 1–18, 2017, doi: 10.3390/a10040124.

[37] P. Bermejo, J. A. Gámez, and J. M. Puerta, "Speeding up incremental wrapper feature subset selection with Naive Bayes classifier," *Knowledge-Based Syst.*, vol. 55, pp. 140–147, 2014, doi: 10.1016/j.knosys.2013.10.016.

[38] J. W. van Lith and J. Vanschoren, "From Strings to Data Science: a Practical Framework for Automated String Handling," pp. 1–19, 2021, [Online]. Available: https://arxiv.org/abs/2111.01868v1.

[39] M. Albarak, M. Alrazgan, and R. Bahsoon, "Identifying and Managing Technical Debt in Database Normalization Using Machine Learning and Trade-off Analysis," 2017, [Online]. Available: http://arxiv.org/abs/1711.06109.

[40] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 4, no. 1, pp. 13–20, 2021, doi: 10.47738/ijiis.v4i1.73.

[41] D. Berrar, "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[42] D. Normawati and D. P. Ismi, "K-Fold Cross Validation for Selection of Cardiovascular Disease Diagnosis Features by Applying Rule-Based Datamining," *Signal Image Process. Lett.*, vol. 1, no. 2, pp. 23–35, 2019, doi: 10.31763/simple.v1i2.3.

[43] F. Harahap, A. Y. N. Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, and C. B. Harahap, "Implementation of Naïve Bayes Classification Method for Predicting Purchase," *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citsm, pp. 1–5, 2019, doi: 10.1109/CITSM.2018.8674324.

[44] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2011.

[45] T. R. Patil and M. S. S. Sherekar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, 2019, doi: 10.18201/ijisae.2019252786.

[46] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem," *Technologies*, vol. 9, no. 4, p. 81, 2021, doi: 10.3390/technologies9040081.