The Performance of Equal-Width and Equal-Frequency Discretization Methods on Data Features in Classification Process

Pramaishella Ardiani Regita Putri[1)]*, Sri Suryani Prasetiyowati[2)], Yuliant Sibaroni[3]
[1,2,3)]School of Computing, Telkom University, Bandung Indonesia
[1)]shellaarp@student.telkomuniversity.ac.id, [2)]srisuryani@telkomuniversity.ac.id,
[3)]yuliant@telkomuniversity.ac.id

Submitted : Jul 19,2023 | **Accepted** : Jul 25, 2023 | **Published** : Oct 1, 2023

Abstract: The classification process often needs help with suboptimal accuracy values, which can be attributed to various factors, including the dataset's wide range of attribute values. Discretization methods offer a solution to address these issues. This study aims to compare the effectiveness of Equal-Width and Equal-Frequency discretization methods in enhancing accuracy during the classification process using datasets with varying sizes. The research employs Naïve Bayes, Decision Tree, and Support Vector Machine as classification models, with three datasets utilized: Bandung City Traffic data (3804 records), Bandung City COVID-19 cases data (2718 records), and Bandung City Dengue Fever Disease Index data (150 records). Three experimental scenarios are executed to assess the impact of the two discretization methods on accuracy. The first scenario involves no discretization, the second employs Equal-Width, and the third applies Equal-Frequency discretization. Experimental results indicate significant accuracy improvements post-discretization. The Naïve Bayes model achieved 94% accuracy for the Traffic dataset, while the Decision Tree achieved 71% accuracy for the COVID-19 dataset and an impressive 98% for the Dengue Fever Disease dataset. These outcomes demonstrate that applying Equal-Width and Equal-Frequency discretization methods addresses the challenge of wide attribute value ranges in the classification process.

Keywords: Accuracy, Discretization, Equal-width, Equal-Frequency, Classification

INTRODUCTION

Data preparation is one of the stages necessary in data mining (DM) before implementing data mining classification algorithms (DMCA) to the data. Data preprocessing is a necessary phase in data mining encompassing techniques such as data transformation, cleansing, reduction, and discretization (Hacibeyoglu & Ibrahim, 2018). The variation of data values processed in data mining is often found in one attribute between one value and another having a range or gap that is too far. In the past few decades, researchers have studied one of the discretization algorithms that is now one of the most frequently used preprocessing techniques in data mining (Xiong et al., 2018)

The working principle of discretization is to change or separate continuous properties into categories or nominal. Variables with continuous values are converted to discrete variables, built with several non-overlapping intervals in this process. The primary benefits of discretization are: (1) data is reduced and takes up less storage space. (2) Discrete data is easier to grasp, utilize, and explain since it is closer to the level of knowledge than continuous data. (3) With discrete data, DMCA can work faster and achieve higher classification accuracy. Data mining classification algorithms such as Naïve Bayes, Decision

*name of corresponding author

Tree, and Support Vector Machine are ten frequently used classification models (Tsai & Chen et al., 2019). However, these algorithms can only process or process numerical data. It aims to ensure that by using discretization, performance would be maximized by creating more effective and efficient models (Stańczyk et al., 2020).

The research that has been studied by (Xiong et al., 2018) compared the discretization algorithm in decision tree classification. Moreover, research (Yamasari et al., 2020) implemented unsupervised discretization in naive Bayes classification. In the study conducted by (Surono et al., 2020), using equal-width discretization with the naive Bayes classification model increased the accuracy value in TB patients to 81%. The three studies prove that discretization in data preprocessing can improve the accuracy value in a classification model.

Based on the description of discretization unsupervised equal-width and equal-frequency methods in previous studies. This research will then be distinguished into three scenarios: without data preprocessing or discretization, implementing equal-width discretization, and implementing equal frequency discretization. This study aims to apply the unsupervised equal-width and equal-frequency discretization methods on data features with a range that is too far can improve the classification model. The contribution of this study is to identify the influence of the amount of data on the performance of unsupervised equal-width and equal-frequency discretization.

LITERATURE REVIEW

Discretization is a data pre-processing technique that transforms continuous data into discrete values. This method can be helpful for various tasks, such as classification, clustering, and regression. Many different discretization methods are available, each with its advantages and disadvantages. Two of the most common discretization methods are equal-width and equal-frequency. Equal-width discretization divides the range of continuous data into a fixed number of intervals of equal width. Equal-frequency discretization divides the range of continuous data into a fixed number of intervals with equal frequency.

Numerous researchers have researched how discretization algorithms perform on classification problems (Xiong et al., 2018) Conducted research that compared the discretization algorithm in decision tree classification. Furthermore, researchers (Yamasari et al., 2020) used unsupervised discretization in Naïve Bayes classification. In the study conducted by (Surono et al., 2020) using equal-width discretization with the Naïve Bayes classification model, the accuracy value in TB patients increased to 81%. The three studies prove that discretization in data pre-processing can improve the accuracy value in a classification model.

Moreover, equal-width discretization is generally more effective than equal-frequency discretization for some classification algorithms, such as decision trees and support vector machines. However, equal frequency discretization is more effective for other classification algorithms, such as naive Bayes.

The choice of discretization method can also depend on the characteristics of the data set. For example, if the data set is highly skewed, equal frequency discretization may be more effective than equal width discretization.

Discretization

The process of reducing continuous data into discrete values is known as discretization. This method can be done for various reasons, such as simplifying the data, making it more compatible with a particular algorithm, or improving the performance of a machine learning model.

There are many different discretization methods available. Some of the most common methods include:

- Equal width discretization: This method divides the range of continuous data into a fixed number of intervals of equal width.
- Equal frequency discretization: This method divides the range of continuous data into a fixed number of intervals with equal frequency.
- Chi-squared discretization: This method uses the chi-squared statistic to determine the optimal number of intervals for discretization.
- Information gain discretization: This method uses information gain to determine the optimal number of intervals for discretization.

*name of corresponding author

The choice of discretization method depends on the specific application. However, in general, equal-width discretization is a good choice for simple data sets, while equal-frequency discretization is a good choice for complex data sets.

Discretization Equal-Width

Equal-width discretization is a simple and effective discretization method. It divides the range of continuous data into a fixed number of intervals of equal width. The width of each interval is determined by the total range of the data and the number of intervals.

The following formula can be used to calculate the width of each interval in equal-width discretization:

$$width: (max - min) / n$$

Where: width is the width of each interval, max is the maximum value in the data set, min is the minimum value in the data set, and n is the number of intervals.

Discretization Equal-Frequency

Equal frequency discretization is another simple and effective discretization method. It divides the range of continuous data into a fixed number of intervals with equal-frequency. The number of values in each interval is determined by the total number of values in the data set and the number of intervals.

The following formula can calculate the number of values in each interval in equal frequency discretization:

$$n\_values = total\_values / n$$

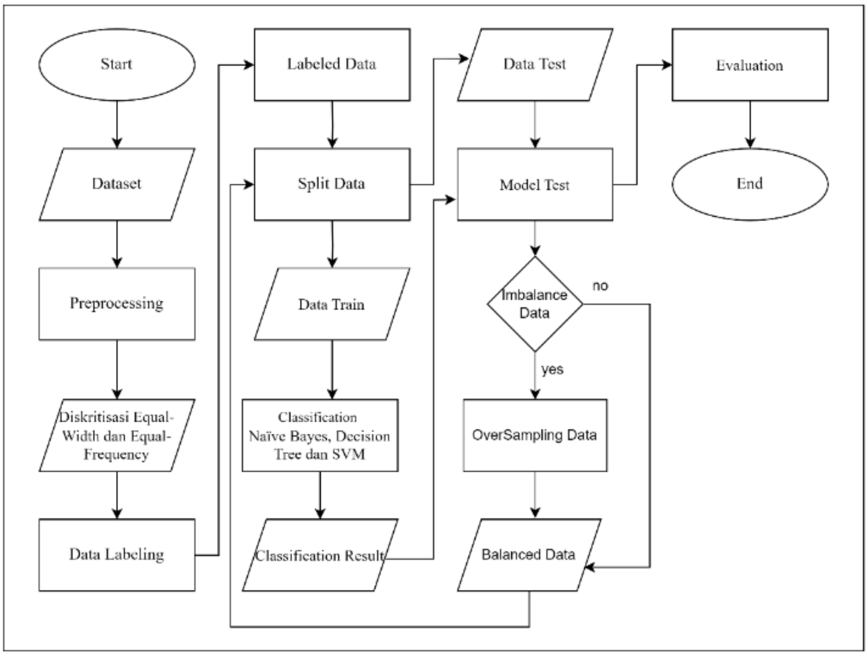
Where: n_values are the number of values in each interval, total_values are the total number of values in the dataset, and n is the number of intervals

METHOD

System Design

The system built in this study was to compare the performance of equal-width and equal-frequency discretization using Naïve Bayes, Decision Tree, and Support Vector Machine classification models. The study also implemented three scenarios: without, equal-width, and equal-frequency discretization. The following is a flowchart of the system design that was built:

Figure 1 Flowchart design system



Dataset

*name of corresponding author

This study employed three separate datasets, each with a different amount of data points. The small dataset was the Dengue Fever Disease dataset from Bandung Regency. The medium-sized dataset was the COVID-19 dataset from Bandung Regency, with 2718 data points. The large-sized dataset was the Traffic Dataset from Bandung Regency, with 3804 data points. This study was done to determine whether the number of data points affects the performance of the discretization method.

1. Traffic Congestion Dataset
   The traffic measurement data contains 3804 rows from April 1st to April 30th, 2022. The traffic measurement data contains street name, lane, time, day, date, number of motorcycles, number of cars, number of trucks, headway, GAP, 85th percentile speed, AVG Speed, road width, and queue length. Only the attributes with an extensive range of gaps between values are discretized in this dataset, as shown in the following table:

Table 1 Traffic Congestion Dataset

Fitur	Description	Range
X1	Number of Motorcycles	(-501) - 4232
X2	Number of Cars	1 - 2647
X3	Number of trucks	0 - 1280
X4	Total number of vehicles	2 - 5011
X5	Number of GAP(s)	0.2 - 503.57
X6	Number of Queue lengths (m)	0 - 850

2. COVID-19 Dataset
   The COVID-19 dataset contains 2716 rows from December 2020 to April 2022. The data contains Month, Village, Male, Female, Rainfall, Sunlight, Average Temperature, Maximum Temperature, Minimum Temperature, No/Not in School, Not Graduated from Elementary School, Graduated from Elementary School, Junior High School, Senior High School, Vocational I and II, Vocational III, Bachelor Degree / Vocational IV, Magister Degree, Doctoral Degree, Dose 1, Dose 2, Dose 3, Compliance with Mask Wearing, Compliance with Social Distancing, Confirmed Cases, Recovered Cases, and Death Cases. Only the attributes with an extensive range of gaps between values are discretized in this dataset, as shown in the following table:

Table 2 COVID-19 Dataset

Fitur	Description	Range
X1	Number of Male	1009 - 20302
X2	Number of Female	1082 - 20451
X3	Number of No/Not yet in School	225 - 7828
X4	Number of Not yet graduated elementary school	144 - 3754
X5	Number of Graduated from elementary School	63 - 6569
X6	Number of Graduated from Junior High School	145 - 5698
X7	Number of Graduated from Senior High School	587 - 13844
X8	Number of Graduated from Vocational III	49 - 2272
X9	Number of Graduated with bachelor's degree / Vocational IV	149 - 5790
X10	Number of Vaccine dose 1	0 - 17772

X11	Number of Vaccine dose 2	0 - 15648
X12	Number of Vaccine dose 3	0 - 3400
X13	Number of Discipline in wearing masks	755 - 37566
X14	Number of Discipline in social distancing	958 - 36494
X15	Number of Total Cases of infection	2 - 8051

3. Dengue Fever Disease Dataset
The dengue fever disease dataset contains 150 rows from 2017 to 2021. The data contains district, population, population proportion, rainfall, temperature, humidity, blood group A, blood group B, blood group AB, blood group O, elementary school graduate, junior high school graduate, senior high school graduate, college graduate, and incidence rate of disease. Only the attributes with an extensive range of gaps between values are discretized in this dataset, as shown in the following table:

Table 3 Dengue Fever Disease Dataset

Fitur	Description	Range
X1	Number of Population	24145 - 142528
X2	Number of blood group A	1276 - 19330
X3	Number of blood group B	1379 - 16489
X4	Number of blood group AB	749 - 7036
X5	Number of blood group O	2537 - 23738
X6	Number of Graduated from elementary School	2126 - 29951
X7	Number of Graduated from Junior High School	2848 - 24062
X8	Number of Graduated from Senior High School	7897 - 46043
X9	Number of Graduated with bachelor’s degree	11612 - 64669

Preprocessing

In this preprocessing, the author divides it into three different scenarios for each dataset:

a. Scenario 1
In the first scenario, the author does not implement any preprocessing process without discretization. The final values obtained from this process will serve as a benchmark for performance comparison in the following discretization scenarios.

b. Scenario 2
In scenario 2, the author implements equal-width discretization, where attributes in the dataset have a wide or long range of values. They are then discretized using different values of k for each dataset, with the size of k determined based on the final classification results in each model.

c. Scenario 3
In scenario 3, the author applies equal-frequency discretization, where attributes in the dataset have a wide or long range of values. They are then discretized using different bin values for each dataset, with the k size determined based on each model's final classification results.

Discretization equal width

The equal-width discretization method is the simplest and easiest method to apply for data processing. It works by dividing the range of values into equal k bins, where the value of k is calculated according to the predetermined binning (Hacibeyoglu et al., 2018). Equal-width calculates the interval of the bin and the limits of attribute A, limited by the values $\boldsymbol{a_{min}}$ and $\boldsymbol{a_{max}}$ based on the following equation:

$$\textbf{Width bin} = (\boldsymbol{a_{max}} - \boldsymbol{a_{min}})/\boldsymbol{k} \tag{1}$$

$$\textbf{Limit} = \boldsymbol{a_{min}} + (\mathbf{i} \times \textbf{width bin})$$
$$\textbf{Where i} = 1, 2, \ldots\ldots\ldots, \mathbf{k} - 1 \tag{2}$$

Table 4 Algorithm equal-width

Input: The value of a continuous attribute (Example A = $\{\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots\ldots\ldots, \boldsymbol{a_{n-1}}, \boldsymbol{a_n}\}$) and the value of interval $\boldsymbol{k}$, where $\boldsymbol{k} > 0$.

Proses 1: Sorting the number of $\boldsymbol{A}$ in ascending order and calculating the values of $\boldsymbol{a_{max}}$ dan $\boldsymbol{a_{min}}$,

Proses 2: Calculating the bin width based on equation (1),

Process 3: Creating bins based on the bin width,

Process 4: Creating intervals and limits based on equation (2),

Process 5: Converting the continuous values of A into discrete values by calculating the range values,

Output: A with discrete values.

Although equal-width is quite simple, it is challenging to determine the appropriate number of intervals, k. This method may provide unbalanced or empty intervals if the property has outliers or extreme values (Hacibeyoglu et al., 2018).

Discretization equal frequency

The equal-frequency discretization method proceeds by dividing the data into k intervals. Each interval has n/k values, where n is the total number of possible values. The most frequent approach is equal-frequency discretization. It reduces the impact of outliers and groups comparable data into a single period. However, similar to equal width, the equal-frequency discretization approach requires assistance in determining the ideal number of intervals, k. This approach can place identical values in two or more neighboring intervals.

Table 5 Algorithm Equal-Frequency

Input: The values of continuous attributes (Example A = $\{\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots\ldots\ldots, \boldsymbol{a_{n-1}}, \boldsymbol{a_n}\}$) and the value of interval $\boldsymbol{k}$, where $\boldsymbol{k} > 0$.
Process 1: Sorting $\boldsymbol{all\ valued\ of\ A}$ in ascending
Process 2: Dividing $\boldsymbol{A}$ into $\boldsymbol{k}$ intervals,
Process 3: Creating bins based on the number of elements in each range,
Process 4: Determining the average value of the most significant bin to determine the limits of each interval,
Process 5: Converting the continuous values of A into discrete values by identifying the range from the selected bin's value and the smallest value from the next bin,
Output: A with discrete values.

The most common approach is equal-frequency discretization. It reduces the impact of outliers and groups comparable data into a single period. However, the equal-frequency discretization approach, like equal width, requires assistance in determining the ideal number of intervals k. This approach can place identical values in two or more neighboring intervals(Hacibeyoglu et al., 2018).

Labeling Data

Labeling data is done to categorize it based on existing rules. For example, the occupancy dataset is grouped into traffic congestion levels regarding the Indonesian Road Capacity Manual (1997). The levels are classified into four groups: label 0 for free flow, label 1 for stable flow, label 2 for stable and controlled flow, and label 3 for unstable flow.

Table 6 Labeling Traffic Congestion Dataset

Class	Class Label	Range
Free Flow	0	Occupancy ≤ 60%
Stabil Flow	1	60% < occupancy ≤ 70%
Controlled Flow	2	70% < occupancy ≤ 80%
Unstable	3	80% <occupancy ≤ 90%

Table 7 Any rules have not established the grouping or labeling in the COVID-19 dataset; therefore, the author categorized it into three levels based on the number of cases: 0 for low, 1 for medium, and 2 for high.

Table 7 Labeling COVID-19 Dataset

Class	Class Label	Range
Low	0	Cases < 218
Medium	1	218 ≤ Cases < 419
High	2	Cases ≥ 419

Table 8 The labeling in the Dengue Fever Disease dataset is based on the number of cases that occur per 100,000 population and is grouped into three levels based on the number of cases: 0 for low, 1 for medium, and 2 for high.

Table 8 Labeling Dengue Fever Disease Dataset

Class	Class Label	Range
Low	0	Cases < 55
Medium	1	55 ≤ Cases < 100
High	2	Cases ≥ 100

Oversampling Data

Oversampling is the process of distributing the class with a more significant number of instances to the class with a smaller number of instances. This method allows the smaller classes to be noticed during classification (Thabtah et al., 2020). However, oversampling, used to address imbalanced data and maximize classification results, is only sometimes efficient in the same way in this research case. In some scenarios of this study, oversampling leads to suboptimal classification results, which several factors can cause. One of the reasons is the significant gap or range between data points in each class.

Implementation

a. Naïve Bayes

A naive Bayes classifier is an example of a Bayesian network utilized for categorizing issues. This simple probabilistic classification model computes the likelihood of a given target variable or group of variables. For example, feature or attribute variables can be utilized to predict the target variable's class properly. The computational simplicity of the Nave Bayes technique and method allows it to be trained more rapidly than other machine learning models. The typical Naive Bayes method has one target variable and several feature variables. Assume T is the state or class of the target variable, and X = (X1, X2,…, Xn) represents the state of n features (Saleh et al., 2023). The maximum-a-posteriori (MAP) rule determines the final classification result. The formula of Bayes' theorem is as follows:

$$P(H|X) = \frac{P(X|H)\ .\ P(H)}{P(X)} \quad (3)$$

Where $P(H)$ and $P(X)$ are constants that can be derived directly from the data, while $P(X|H)$: is left to solve. The Naïve Bayes approach can be developed in the following equation:

$$P(C|F1....F_n) = \frac{P(C)\ \ P(F1...F_n|C)}{P(F1....F_n)} \quad (4)$$

Where variable C represents the class and variables F1...Fn indicates the classification research method features. The formula may then be expressed more simply as follows (Fajriati et al., 2023)

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \quad (5)$$

b. Decision Tree

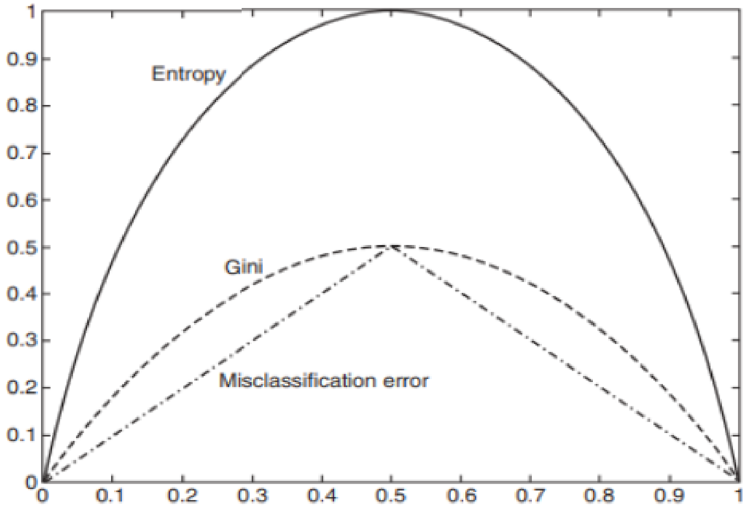
The Decision tree model is built like a tree, with each node representing a test on an attribute and each branch indicating the test outcome. Each leaf node corresponds to a class label (Setyawan et al., 2020). The Decision tree is built top-down, from nodes to the root, and recursively partitions the data processing until each partition ends with a leaf node. Furthermore, the decision tree can be transformed into several rules that assist in this research to be clearly understood. The chosen measurement techniques for attribute test division are entropy, Gini index, and gain ratio.

- Entropy Gain and Gain Ration

  Entropy is a measure of a dataset's unpredictability or uncertainty. Entropy values are always between 0 and 1. It is better when the value is closer to 0; when it is closer to 1, it is worse. If the goal is G with varied attribute values, as illustrated in "Figure 3," the classification entropy of set S associated with the variable c is determined. This is seen in the "equation(6)."

*name of corresponding author

Figure 2 Value of Entropy (Charbuty et al., 2021)



$$Entropy = \sum_{i=i}^{c} P_i \log 2^{P_i} \ (6)$$

P_i is the ratio of the subset of sample numbers to the value of attribute i.
Information Gain, also known as mutual information, is one of the measures used for segmentation. This measure indicates how much information is gathered about the random variable. It is the inverse of entropy, with greater levels indicating better performance. Data Gain(S, A) is defined in the notion of entropy as follows, as illustrated in "equation (6)".

$$Gain(S,\ A) = \sum_V \in V(A) \frac{|S_v|}{|S|}\ Entropy(S_v) \ (7)$$

V(A) represents the interval of attribute A, and S_v is a subset of S that equals the attribute value v (Charbuty et al., 2021)

c. Support Vector Machine (SVM)
Support Vector Machine (SVM) is an approach that uses supervised learning used to examine data and uncover patterns in classification and regression analysis. (Syahputra et al., 2022). The main idea behind this methodology is to discover the best separator space among many classifications in a dataset. This strategy involves searching for a hyperplane or dividing line that separates one class from another. It can also help with multi-domain applications in large data settings. SVM, on the other hand, is mathematically hard and computationally costly. SVM's linear function may be represented as follows:

$$f(x) = (\omega, x) + b \qquad (7)$$

Where ω the weight values as coefficients for each feature, modified throughout processing. The algorithm determines the weight values to produce a hyperplane with the greatest significant margin (Nedumaran et al., 2020). Numerous kernel functions are often employed in SVM, such as RBF, polynomial, and sigmoid, and they are denoted as follows:

a. Radial Basis Function (RBF)

$$K(xi, xj\ ) = ex\ p(-y \parallel xi, xj \parallel) \ \ (8)$$

b. Polynomia

$$K(xi\ , xj\ ) = (yXi\ T\ xj + r)\ p\ , y > 0$$

c. Sigmoid

$$K(x_i, x_j) = tanh\ (x^T\ x_j\ + r) \qquad (9)$$

*name of corresponding author

Evaluation

The confusion matrix is a technique for evaluating the results or correctness of the classification proces (Luque et al., 2019). It uses a confusion matrix to evaluate how well the classifier identifies input from distinct classifications. This is the confusion matrix employed in Table:

Actual Label	Positive	TP	FP
	Negative	FN	TN
		Positive	Negative
		Prediction Label	

Regarding on the available Table above, the performance of the constructed classification can be calculated, including:

a. Accuracy

Accuracy can be described as the ratio of correct predictions per document to total predictions for those categories. The following is an example of accuracy equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

b. Precision

It is the true positive data divided by the total data classified correctly

$$Precision = \frac{TP}{TP + FP}$$

c. Recall

It represents true positive data classified correctly by the system.

$$Recall = \frac{TP}{TP + FN}$$

d. F1-Score

The F1-Score is the harmonic mean of accuracy and recall, and it is directly proportional to both. The F1-score may be calculated as follows:

$$F1 - Score = \frac{2(recall\ x\ precision)}{recall + precision}$$

RESULT

This section will present the results of data processing from each dataset using three different scenarios. The results of the classification modeling can be seen in Table 9, Table 11, Table 12, and Table 13. WD refers to without discretization, EW refers to Equal-Width and EF refers to Equal-Frequency

*name of corresponding author

Table 9 Result of Accuracy

Dataset	Accuracy								
	Naïve Bayes			Decision Tree			SVM		
	WD	EW	EF	WD	EW	EF	WD	EW	EF
Traffic congestion	61%	68%	94%	92%	93%	82%	92%	93%	86%
COVID-19	61%	62%	64%	66%	71%	68%	66%	71%	68%
Dengue Fever Disease	36%	47%	89%	49%	47%	98%	49%	47%	91%

Table 10 Result accuracy of previous research (Nugroho et al., 2022)

No	**Methods**	**Accuracy**
1	Naïve Bayes	96.68%
2	Naïve Bayes + Discretization Unsupervised (equal-width)	97.66%

In the Accuracy table 9, there is a significant improvement in the values for each dataset. The highest increase occurred in the Dengue Fever Disease dataset, where the accuracy value increased to 98% from the previous value of 36%. In the Traffic dataset, the accuracy value increased to 94% from the previous value of 61%. And in the COVID-19 dataset, the accuracy value improved to 71% from the previous value of 61%.

In Table 10 of the study conducted by (Nugroho et al., 2022), the authors applied the unsupervised discretization method known as "equal-width" to classify Study Programs for Prospective New Students. The result obtained from this experiment showed an accuracy of 97.66% when tested on a dataset of 161 records.

This research further explored and compared various discretization methods on three different datasets. Through this investigation, the authors discovered that discretization methods play a significant role in improving the accuracy of classification models. This finding indicates that discretization, specifically the "equal-width" and "equal-frequency" approaches, can effectively enhance the accuracy of the classification process. This method can lead to more precise and reliable classifications by dividing continuous attribute values into equal-width intervals. The study highlights the importance of considering suitable discretization methods based on the specific characteristics and complexity of the dataset. It also emphasizes the potential impact of using different discretization techniques to achieve better classification results in various scenarios.

Table 11 Result of Precision

Dataset	Precision								
	Naïve Bayes			Decision Tree			SVM		
	WD	EW	EF	WD	EW	EF	WD	EW	EF
Traffic congestion	42%	79%	48%	70%	79%	48%	70%	79%	56%
COVID-19	60%	61%	62%	66%	70%	66%	60%	70%	62%
Dengue Fever Disease	65%	51%	88%	57%	51%	98%	57%	51%	92%

In the Precision table where WD refers to without discretization, EW refers to Equal-Width, and EF refers to Equal-Frequency, there is a significant improvement in the values for each dataset. The highest increase occurred in the Dengue Fever Disease dataset, where the precision value increased to 98% from the previous value of 65%. In the Traffic dataset, the precision value increased to 79%

from the previous value of 42%. Moreover, in the COVID-19 dataset, the precision value improved to 70% from the previous value of 60%.

Table 12 Result of Recall

Dataset	Recall								
	Naïve Bayes			Decision Tree			SVM		
	WD	EW	EF	WD	EW	EF	WD	EW	EF
Traffic congestion	72%	74%	47%	72%	74%	47%	72%	74%	50%
COVID-19	59%	60%	62%	65%	70%	66%	65%	70%	66%
Dengue Fever Disease	62%	43%	88%	52%	43%	98%	52%	43%	92%

The Recall table where WD refers to without discretization, EW refers to Equal-Width, and EF refers to Equal-Frequency shows a significant improvement in the values for each dataset. The highest increase occurred in the Dengue Fever Disease dataset, where the recall value increased to 98% from the previous value of 62%. In the Traffic dataset, the recall value increased to 74% from the previous value of 72%. Moreover, in the COVID-19 dataset, the recall value improved to 70% from the previous value of 59%.

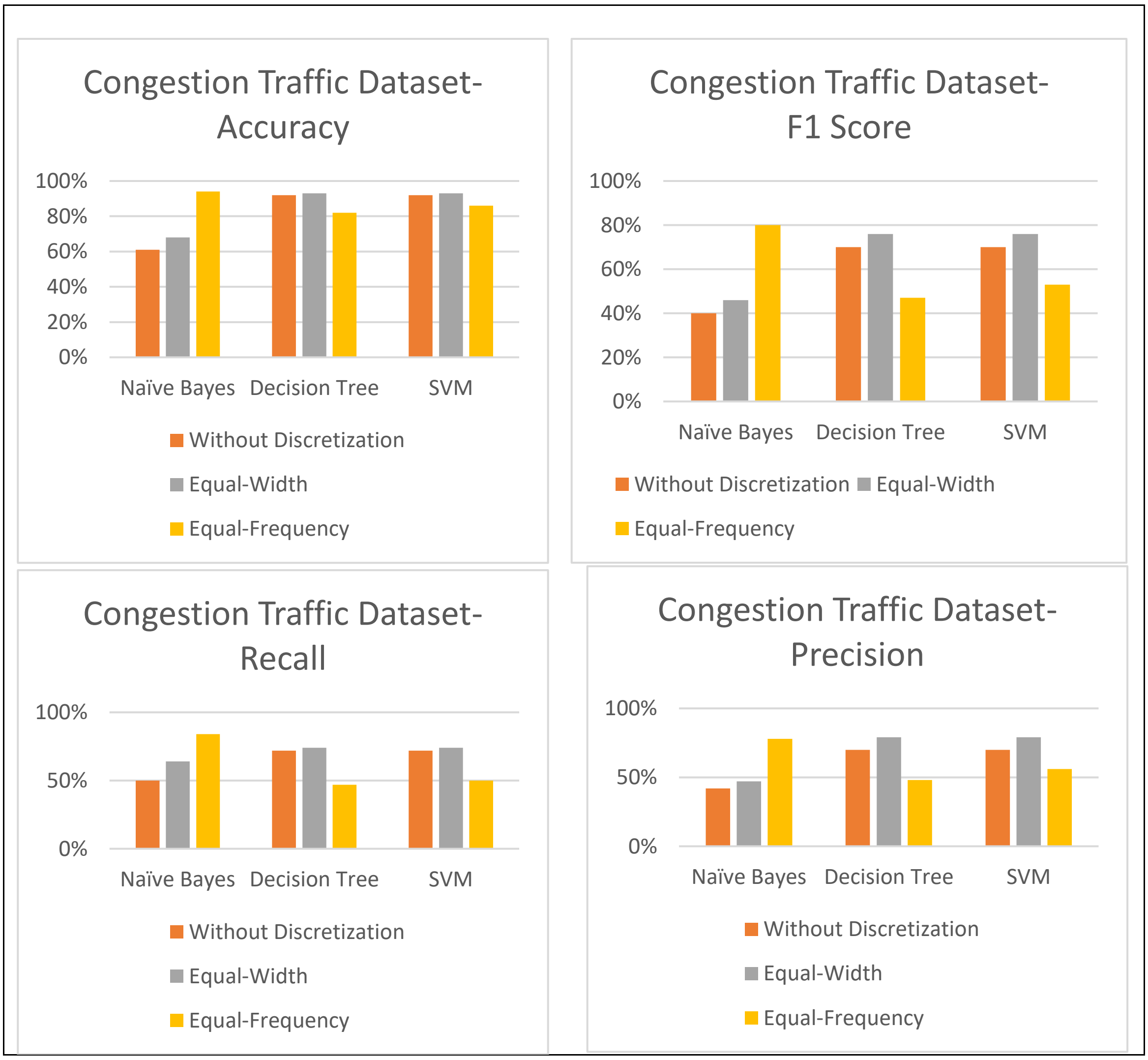
Table 13 Result of F1-Score

Dataset	F1 Score								
	Naïve Bayes			Decision Tree			SVM		
	WD	EW	EF	WD	EW	EF	WD	EW	EF
Traffic congestion	40%	76%	53%	70%	76%	47%	70%	76%	53%
COVID-19	59%	60%	61%	65%	69%	66%	65%	69%	66%
Dengue Fever Disease	61%	42%	88%	52%	45%	98%	49%	45%	91%

The F1 Score table, where WD refers to without discretization, EW refers to Equal-Width, and EF refers to Equal-Frequency, shows a significant improvement in the values for each dataset. The highest increase occurred in the Dengue Fever Disease dataset, where the F1-Score increased to 98% from the previous value of 61%. In the Traffic dataset, the F1 Score increased to 76% from the previous value of 40%. Moreover, in the COVID-19 dataset, the F1 Score improved to 69% from the previous value of 59%.

Congestion Traffic Dataset

Figure 3 Classification results graph of the Congestion Traffic Dataset



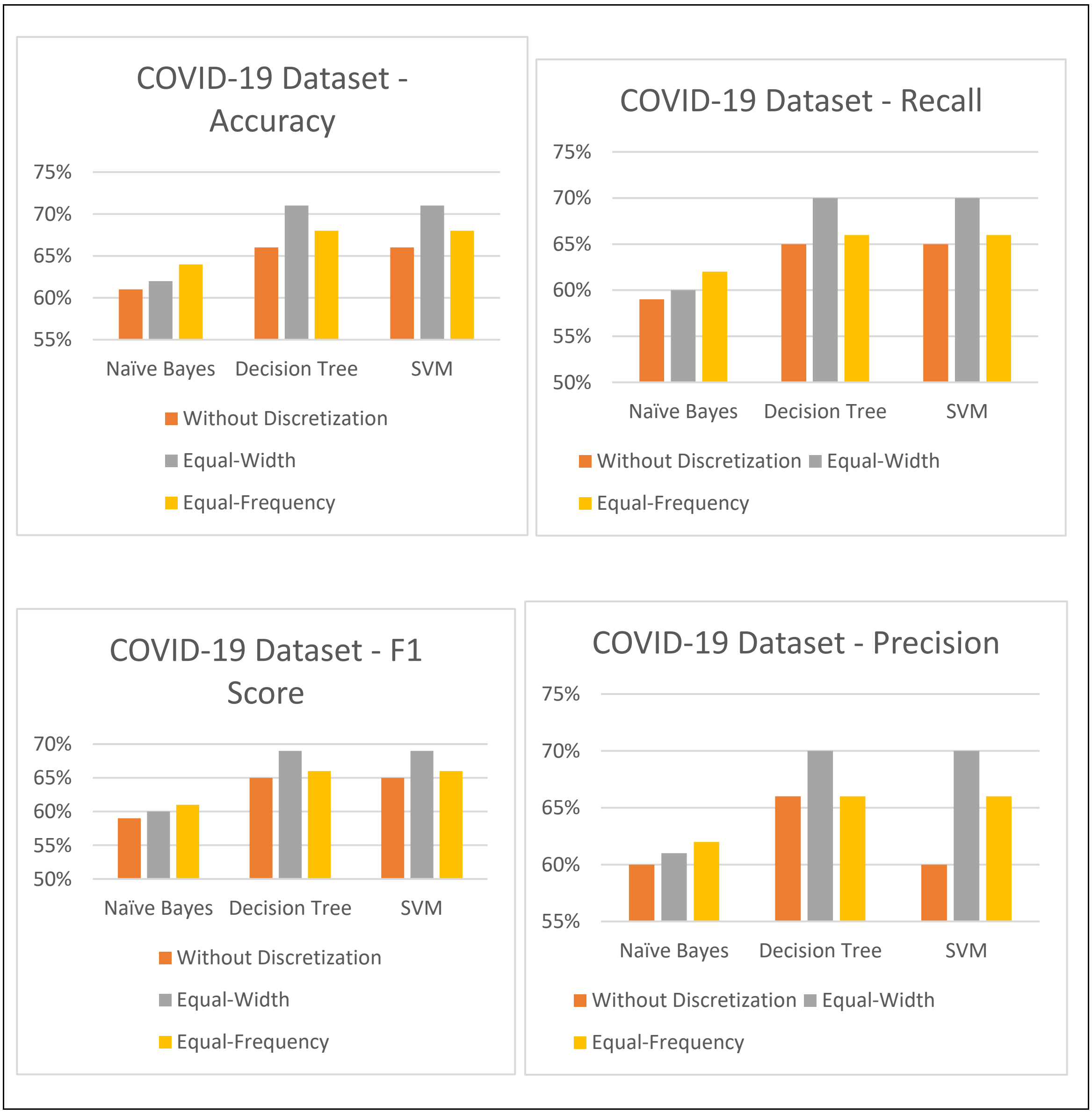
In the above figure, it can be observed that the classification model results from the three scenarios are different. In the first scenario without discretization, Naïve Bayes achieved an accuracy of 61%, precision of 42%, recall of 50%, and F1 Score of 40%. Decision Tree achieved an accuracy of 92%, precision of 70%, recall of 70%, and F1 Score of 70%. Lastly, Support Vector Machine achieved an accuracy of 92%, precision of 70%, recall of 70%, and F1 Score of 70%.

In the second scenario, applying Equal-Width discretization using k = 9, all three models had different outcomes. Naïve Bayes achieved an accuracy of 68%, precision of 47%, recall of 64%, and F1 Score of 46%. Decision Tree achieved an accuracy of 93%, precision of 79%, recall of 74%, and F1 Score of 76%. Support Vector Machine achieved an accuracy of 93%, precision of 79%, recall of 74%, and F1 Score of 76%.

In the last scenario, using equal-frequency discretization with k = 5, all three models had different outcomes. Naïve Bayes achieved an accuracy of 94%, precision of 78%, recall of 84%, and F1 Score of 80%. Decision Tree achieved an accuracy of 82%, precision of 48%, recall of 47%, and F1 Score of 47%. Support Vector Machine achieved an accuracy of 86%, precision of 56%, recall of 50%, and F1 Score of 53%.

COVID-19 Dataset

Figure 4 Classification results graph of the COVID-19 Dataset



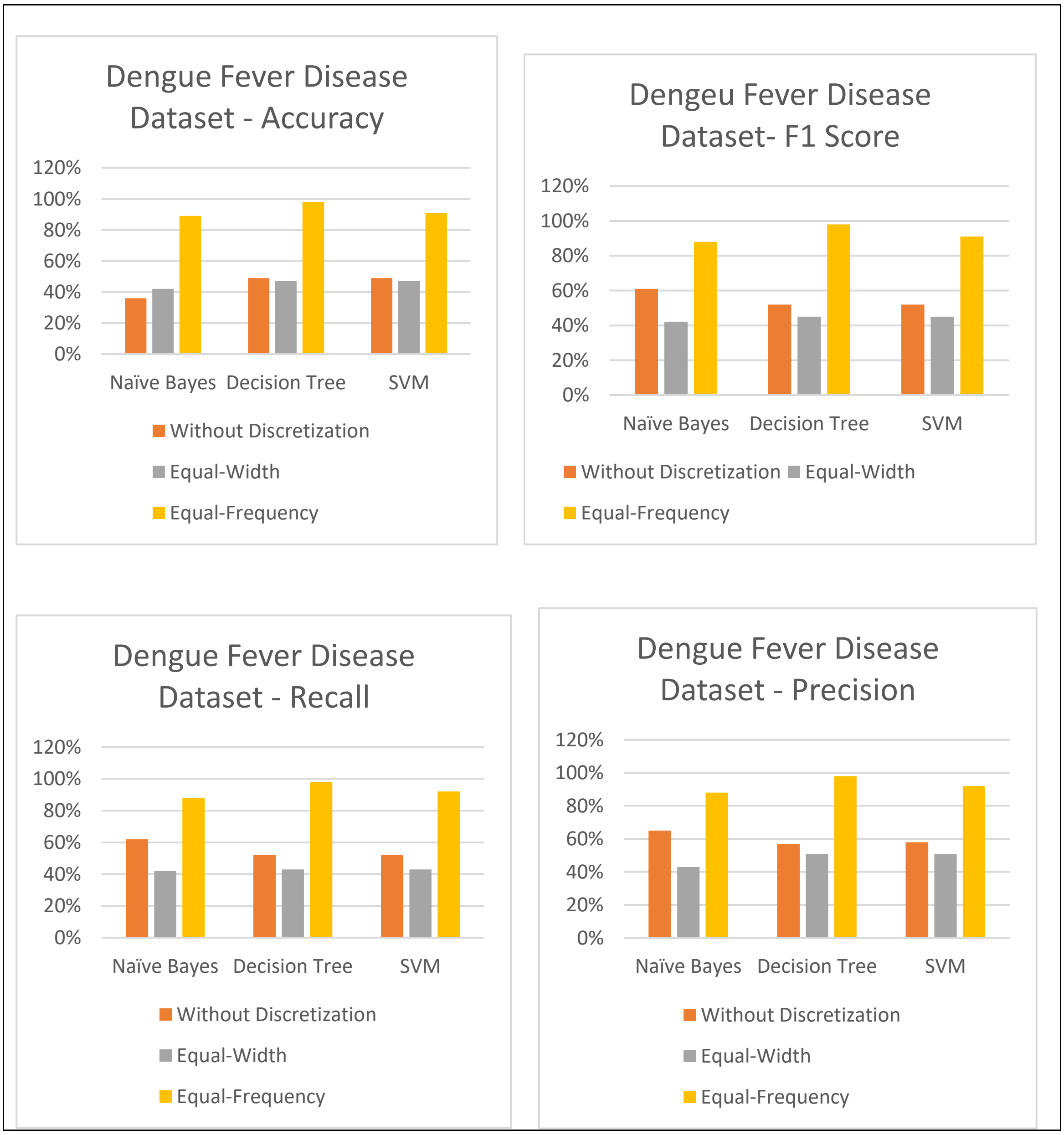
In the above figure, it can be observed that the classification model results from the three scenarios are different. In the first scenario without discretization, Naïve Bayes achieved an accuracy of 60%, precision of 60%, recall of 59%, and F1 Score of 59%. Decision Tree achieved an accuracy of 66%, precision of 66%, recall of 65%, and F1 Score of 65%. Lastly, Support Vector Machine achieved an accuracy of 66%, precision of 66%, recall of 65%, and F1 Score of 65%.

In the second scenario, applying Equal-Width discretization using k = 12, all three models had different outcomes. Naïve Bayes achieved an accuracy of 62%, precision of 61%, recall of 62%, and F1 Score of 60%. Decision Tree achieved an accuracy of 71%, precision of 70%, recall of 70%, and F1 Score of 70%. Support Vector Machine achieved an accuracy of 71%, precision of 70%, recall of 70%, and F1 Score of 70%.

In the last scenario, using equal-frequency discretization with k = 6, all three models had different outcomes. Naïve Bayes achieved an accuracy of 64%, precision of 62%, recall of 62%, and F1 Score of 61%. Decision Tree achieved an accuracy of 68%, precision of 66%, recall of 66%, and F1 Score of 66%. Support Vector Machine achieved an accuracy of 68%, precision of 66%, recall of 66%, and F1 Score of 66%.

Dengue Fever Disease Dataset

Figure 5 Classification results graph of the Dengue Fever Disease Dataset



In the above figure, it can be observed that the classification model results from the three scenarios are different. In the first scenario without discretization, Naïve Bayes achieved an accuracy of 36%, precision of 65%, recall of 62%, and F1 Score of 61%. Decision Tree achieved an accuracy of 49%, precision of 57%, recall of 52%, and F1 Score of 52%. Lastly, Support Vector Machine (SVM) achieved an accuracy of 49%, precision of 57%, recall of 52%, and F1 Score of 52%.

In the second scenario, applying Equal-Width discretization using k = 12, all three models had different outcomes. Naïve Bayes achieved an accuracy of 42%, precision of 43%, recall of 42%, and F1 Score of 42%. Decision Tree achieved an accuracy of 47%, precision of 51%, recall of 43%, and F1 Score of 45%. Support Vector Machine achieved an accuracy of 47%, precision of 51%, recall of 43%, and F1 Score of 45%.

In the last scenario, using equal-frequency discretization with k = 6, all three models had different outcomes. Naïve Bayes achieved an accuracy of 89%, precision of 88%, recall of 88%, and F1 Score of 88%. Decision Tree achieved an accuracy of 98%, precision of 98%, recall of 98%, and F1 Score of 98%. Support Vector Machine achieved an accuracy of 91%, precision of 92%, recall of 92%, and F1 Score of 91%.

DISCUSSIONS

In this research, the author compared the performance of unsupervised equal-width and equal-frequency discretization on three different datasets with varying sizes. We used three classification models, Naïve Bayes, Decision Tree, and Support Vector Machine, to examine the effect of data size on the discretization performance. The study's results showed that the data size influences the performance of the discretization methods. For the small-sized dataset (dengue fever dataset), equal-frequency discretization provided the best performance, with an accuracy improvement ranging from 49% to 98%. For the medium-sized dataset (COVID-19 dataset), equal-width discretization yielded the best performance, with an accuracy improvement ranging from 66% to 71%.

Meanwhile, for the large-sized dataset (congestion traffic dataset), equal-frequency discretization also gave the best performance, with an accuracy improvement ranging from 61% to 94%. Significant accuracy improvements were also observed in recall, precision, and F1 score values. The table indicates that the average improvement occurred in processing the small-sized dataset, the dengue fever dataset, with 150 data points. A difference from previous studies is that this research used three datasets with different complexities of data features. These datasets were combined with two unsupervised discretization methods in the three commonly used classification models. (Tsai et al., 2019).

Several previous experiments on optimizing Naïve Bayes algorithms by discretization have been undertaken. (Saleh et al., 2020) used the Naïve Bayes algorithm for classifying student majors and applied the equal-width interval discretization method to improve the classification accuracy using the Naïve Bayes algorithm. The results showed that implementing discretization increased the accuracy of the Naïve Bayes algorithm from 90% to 92.8%. Another study by (Nugroho et al., 2022) also used the Naïve Bayes algorithm for classifying study programs for prospective new students and applied the equal-width interval discretization method to improve the classification accuracy using the Naïve Bayes algorithm. The results showed that the classification using the Naïve Bayes algorithm with discretization resulted in higher accuracy, namely 97.66%, compared to without discretization, which resulted in lower accuracy of 96.68%.

ACKNOWLEDGMENT

The author expresses heartfelt gratitude to Telkom University and the supervisors for their support and guidance, which contributed to the completion of this research. Moreover, to all my friends, thank you for your encouragement and assistance.

CONCLUSION

This study evaluated the performance of unsupervised equal-width and equal-frequency discretization on datasets of varying sizes compared to previous research papers. The three tested scenarios allowed us to observe how these methods interpret and classify datasets with different complexities. Our findings align with existing research, indicating that data complexity plays a crucial role in the effectiveness of discretization methods. Notably, equal-frequency discretization demonstrated significantly improved accuracy across various classification models in both datasets (Congestion Traffic and Dengue Fever Disease). Equal-frequency discretization was observed to perform optimally on datasets with lower data complexity or fewer features. In comparison, equal-width discretization showed better outcomes for datasets with higher complexity and more features. This research can aid in selecting the appropriate discretization method based on the complexity of data features. For future studies, other discretization methods can be explored to compare their performance, address data imbalances, and utilize larger datasets to enhance accuracy.

REFERENCES

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Fajriati, N., Prasetiyo, B., & Korespondensi, P. (2023). *OPTIMASI ALGORITMA NAÏVE BAYES DENGAN DISKRITISASI K-MEANS PADA DIAGNOSIS PENYAKIT JANTUNG*. *10*(3), 503–512. https://doi.org/10.25126/jtiik.2023106510

Hacibeyoglu, M., & Ibrahim, M. H. (2018). EF_Unique: An Improved Version of Unsupervised Equal Frequency Discretization Method. *Arabian Journal for Science and Engineering*, *43*(12), 7695–7704. https://doi.org/10.1007/s13369-018-3144-z

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Nedumaran, A., Ganesh Babu, R., Kassa, M. M., & Karthika, P. (2020). Machine level classification using support vector machine. *AIP Conference Proceedings*, *2207*. https://doi.org/10.1063/5.0000041

Nugroho, W. E., Prihandoyo, T., & Somantri, O. (2022). Optimalisasi Metode Naive Bayes untuk Menentukan Program Studi bagi Calon Mahasiswa Baru dengan Pendekatan Unsupervised Discretization. *Infotekmesin*, *13*(1), 161–167. https://doi.org/10.35970/infotekmesin.v13i1.1048

Saleh, A., Dharshinni, N., Perangin-Angin, D., Azmi, F., & Sarif, M. I. (2023). Implementation of Recommendation Systems in Determining Learning Strategies Using the Naïve Bayes Classifier Algorithm. *Sinkron*, *8*(1), 256–267. https://doi.org/10.33395/sinkron.v8i1.11954

Saleh, A., & Nasari, F. (n.d.). *PENERAPAN EQUAL-WIDTH INTERVAL DISCRETIZATION DALAM METODE NAIVE BAYES UNTUK MENINGKATKAN AKURASI PREDIKSI PEMILIHAN JURUSAN SISWA (STUDI KASUS: MAS PAB 2 HELVETIA,MEDAN) IMPLEMENTATION OF EQUAL-WIDTH INTERVAL DISCRETIZATION IN NAIVE BAYES METHOD FOR INCREASING ACCURACY OF STUDENTS' MAJORS PREDICTION (CASE STUDY : MAS PAB 2 HELVETIA,MEDAN)*.

Setyawan, D. A., & Fatichah, C. (2020). ENHANCEMENT OF DECISION TREE METHOD BASED ON HIERARCHICAL CLUSTERING AND DISPERSION RATIO. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *18*(2), 179. https://doi.org/10.12962/j24068535.v18i2.a1005

Stańczyk, U., Zielosko, B., & Baron, G. (2020). Discretisation of conditions in decision rules induced for continuous data. *PLoS ONE*, *15*(4). https://doi.org/10.1371/journal.pone.0231788

Surono Program Studi Matematika FAST UAD Jl Ringroad Selatan, S. (n.d.). *DISKRITISASI EQUAL-WIDTH INTERVAL PADA NAÏVE BAYES (STUDI KASUS: KLASIFIKASI PASIEN TBC) EQUAL-WIDTH INTERVAL DISCRETIZATION IN NAÏVE BAYES (CASE STUDY: CLASSIFICATION TBC PATIENTS)*.

Syahputra, R., Yanris, G. J., & Irmayani, D. (2022). SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter. *Sinkron*, *7*(2), 671–678. https://doi.org/10.33395/sinkron.v7i2.11430

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Tsai, C. F., & Chen, Y. C. (2019). The optimal combination of feature selection and data discretization: An empirical study. *Information Sciences*, *505*, 282–293. https://doi.org/10.1016/j.ins.2019.07.091

Xiong, W., IEEE Computer Society, International Association for Computer & Information Science, Pattern Recognition and Machine Intelligence Association., & Institute of Electrical and Electronics Engineers. (n.d.). *17th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2018) : proceedings : June 6-8, 2018, Singapore*.

Yamasari, Y., Qoiriah, A., Rochmawati, N., Yustanti, W., Tjahyaningtijas, H. P. A., & Rusimamto, P. W. (2020, October 3). Combining the Unsupervised Discretization Method and the Statistical Machine Learning on the Students' Performance. *Proceeding - 2020 3rd International Conference on Vocational Education and Electrical Engineering: Strengthening the Framework of Society 5.0 through Innovations in Education, Electrical, Engineering and Informatics Engineering, ICVEE 2020*. https://doi.org/10.1109/ICVEE50212.2020.9243273