Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 12, No. 3, Juli 2024

ISSN 0216 – 0544
e-ISSN 2301– 6914

THE INFLUENCE OF DATA CATEGORIZATION AND ATTRIBUTE INSTANCES REDUCTION USING THE GINI INDEX ON THE ACCURACY OF THE CLASSIFICATION ALGORITHM MODEL

[a]Willy Fernando, [b]Deny Jollyta, [c]Dadang Priyanto, [d]Dwi Oktarina

[a,b,d]Institut Bisnis dan Teknologi Pelita Indonesia
[c]Universitas Bumigora
E-mail: willyfernando530.bkls@gmail.com, deny.jollyta@lecturer.pelitaindonesia.ac.id, dadang.priyanto@universitasbumigora.ac.id, dwi.oktarina@lecturer.pelitaindonesia.ac.id

Abstract

Numerical data problems are typically caused by a failure to comprehend the data and the outcomes of its processing. In order to give richer context and a deeper understanding of the facts, numerical data must be transformed into categories. On the other hand, changes in data have a significant impact on the analysis's outcomes. The purpose of this study is to see how transforming numerical data into categories affects the model produced by the classification algorithms. The dataset used in this study is the Maternal Health Risk. Categorization refers to formal arrangements. Categorization is also accomplished by using the Gini Index to limit the number of instances of an attribute. The classification is carried out using the Random Forest (RF), K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM) algorithms to produce a model. The influence of data modifications to model can be observed in the confusion matrix with 5 different data splitting. The study results suggested that changing numerical data to categories data significantly improved the performance of the SVM model from 76.92% to 80.77% at a data splitting percentage of 95/5.

Key words: Categorization, Classification Algorithms, Confusion Matrix, Numerical Data.

INTRODUCTION

Data is a critical component in producing information. The veracity and correctness of data have an impact on the desired outcomes. According to study [1], data is something that is not yet obvious (row fact), and it must be processed in order to have significance. Data can take the shape of numbers, text, or other symbols, all of which can be processed or understood. Data can be collected as a consequence of observations, measurements, or from a variety of sources, including studies, surveys, recordings, and sensors. Data might be raw or processed [2].

Data categorization strategies employ both quantitative and qualitative data. Quantitative data is numerical or ordered data, whereas qualitative data is non-numerical data [3]. These two data formats are used in tandem. Classification approaches need objective qualities, which are often in the form of non-numerical data.

Problems that often arise in data processing are when users want to change data, for example changing numeric data to non-numerical or vice versa. This is occasionally necessary for data processing purposes. However, it does not offer information about changes in outcomes caused by changing data. This leads to discrepancies in analysis, which can either improve or degrade the results.

In the ever-changing world of data analysis, the need to analyze and handle data in a more contextual manner is becoming increasingly essential. Converting numerical data into categories is one important method that has evolved.

A variety of research investigate the effect of converting numerical data into categories or vice versa on analytical outcomes. The Association Rule General Analytic System (ARGAS) technique was employed to evaluate non-numerical social media data [4]. The ARGAS technique was successful in constructing important analysis. The normalized correlation coefficient is used in research [5] to transform non-numerical data series to numerical data series. Dynamic Time Warping (DTW) metrics are employed to examine the relationship. This method has a significant impact on enhancing DTW accuracy. Apart from that, categorical data has been successful in producing data analysis in a several of industries such as education, health, and telecommunications [6]. Data classification aids in the production of the appropriate analysis.

Based on the description provided, the purpose of this study is to examine the influence of data categorization on the information supplied. The challenge employs Maternal Health Risk data, with all characteristics being numerical. The influence may be observed in the classification results obtained using three classification algorithms namely RF, K-NN, and SVM by splitting the data five times. These three methods frequently combine data of diverse types. However, there are currently few that examine the effect of data changes on the model generated by the methods. In research [7], it was explained that to provide a fair outcome and eliminate bias in model evaluation, the subset data utilized in training and testing must be distinct. This is why these three methods were selected to support the idea of this research.

This study's contribution is in the data categorization method, which employs criteria linked to Maternal Health Risk data arrangements and use the Gini Index technique to minimize the number of instances. Here are the relevant arrangements:

Table 1. Arrangements of Categorized Data

Attributes	References
Age	Regulation of the Minister of Health of the Republic of Indonesia Number 25 of 2016
Systolic BP	Peralta, et. al,2014
Diastolic BP	Peralta, et. al, 2014
Blood Sugar	Wahyuni, et. al, 2022
Body Temp	Geneva, et. al, 2019
Heart Rate	Tangirala, et, al, 2020 (Gini Index)
Risk Level	-

Maternal Health Risk Data is a UCI Machine Learning dataset with a variety of variables. This dataset is data that classifies the risk of maternal death into 3 levels, namely High Risk (HR), Mid Risk (MR) and Low Risk (LR). This information has been thoroughly analyzed in several researches [8][9][10]. Aside from that, a number of additional research, employ the same attributes as the Maternal Health Risk dataset but are categorized using different algorithms [11][12][13][14]. The previous research did not categorize data, thus it is believed that this new research would give a fresh perspective on data analysis by categorizing the data.

MATERIAL AND METHODS

Classification Algorithm

Random Forest (RF)

RF is an algorithm that mixes numerous randomly generated decision trees to provide more accurate predictions [15]. In a number of studies, RF consistently outperforms alternative classification algorithms such as SVM, Neural Network and Decision Tree [16][17][18]. Equation (1) depicts the RF algorithm form.

$$Regression: T(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \qquad (1)$$

Classification: with $\widehat{C}_b(x)$ as prediction class from b^th Random Forest trees, then $\widehat{C}_{rf}^{B}(x) = majority\ vote\left\{\widehat{C}_b(x)\right\}_1^B$

K-Nearest Neighbor (K-NN)

Classification and regression problems can be solved by K-NN with an input of k closest to the data set [19][20]. To solve classification problems with the K-NN algorithm, it is necessary to understand the following basic steps:

1) Determine k positive integers based on the availability of learning data.
2) Select k nearest neighbors from the new data.
3) Determine the most common classification in step 2, using the highest frequency.
4) Classification results from new data.

The Euclidean Distance formula is used to calculate nearest neighbors. Data can be in one or more than one dimension in its implementation [21]. The following is the Euclidean Distance equation:

$$dis(x_1,x_2)=\sqrt{\sum_{i=0}^{n}(x_{1i}-x_{2i})^2} \qquad (2)$$

Where:

= Distance of data i to cluster center j

= The i data in the k data attribute

= The jth center point on the kth attribute

Support Vector Machine (SVM)

SMV is a linear classifier at its core, but it was designed to work on nonlinear data with a kernel concept in a high-dimensional workspace [22][23]. The SVM algorithm is executed in the following stages:

1) Margin

$$\text{Equation for margin} = \frac{1}{w} \qquad (3)$$

where:

w = margin weight

Equation for largest margin:

$$d(\boldsymbol{X}^T) = \sum_{i=1}^{l} y_i \alpha_i \boldsymbol{X}_i \boldsymbol{X}^T + b_0,$$

Where:

$d(X^T)$ = maximum margin

yi = label class

XT = training data

α_i = weight value of each data point

b_0 = SVM parameters/biases

2) Hyperplane

The basic hyperplane equation is formed by paying attention to the margins on both sides:

$$\frac{1}{2}||w||^2 = \frac{1}{2}({w_1}^2 + {w_2}^2) \qquad (4)$$

Then:

$$wx_i + b = 0 \qquad (5)$$

Or

$$w_1x_1 + w_2x_2 + b = 0 \qquad (6)$$

Where:

w = margin weight

x = data point

b = biases

3) Class

Class boundaries are formed from the outermost data points (support vector) to form the desired hyperplane. Take a look at the margins in equations (4) and (6) with conditions.:

$$y_i(wx_1 + b) \geq 1 \quad \text{i=1, 2, 3, …N} \qquad (7)$$

Then the combined negative and positive hyperplane equations are:

$$y_i(w_1x_1 + w_2x_2 + b) \geq 1 \qquad (8)$$

Hyperplane for positive class:

$$w_1x_1 + w_2x_2 + b \geq 1 \; for \; y_i = +1 \qquad (9)$$

Hyperplane for negative class:

$$w_1x_1 + w_2x_2 + b \leq -1 \; for \; y_i = -1 \qquad (10)$$

Dimana:

b = biases

x = data point

y = class

Gini Index

Classification errors due to random selection can be measured using the Gini index. The lower the Gini Index, the less likely misclassification [24]. The calculation formula is as follows:

$$I_G(i) = 1 - \sum_{j=1}^{m} f^2(i,j) \qquad (11)$$

Where :

i = variable value

f = number of samples belonging to class j in data set i

m = number of classes in the data set

If all samples in the data set belong to the same class, the Gini Index reaches zero, indicating a low level of impurity. In contrast, if the samples are distributed evenly across all classes, the maximum value 1 is obtained, indicating a high level of impurity. The decision-making process in a decision tree using the Gini splitting index is as follows:

1) Calculate the Gini Index for each split based on certain attributes.
2) Choose the slice that has the lowest Gini Index as the choice for dividing the data.

3) Repeat the process at each node in the decision tree.

Choosing the attributes and cut values that result in the lowest Gini splitting index aids in the construction of an optimal decision tree for data classification. The better the separation performed by the attribute, the lower the Gini splitting index value. The Gini splitting index is calculated using the formula below.

$$GINI_{split} = \sum_{i=1}^{p} \frac{n_i}{n} GINI(i) \quad (12)$$

Where :
n_i = total sample included in class i
n = total of all data
$GINI_{(i)}$ = Gini index value

Principal Component Analysis (PCA)

PCA is a technique for decreasing the complexity of high-dimensional data while maintaining trends and patterns. It achieves this by compressing the data into fewer components (attributes), which may be thought of as feature analysis [25]. In this study, PCA simplified the attributes, resulting in a total of six.

Data Category

The concept of converting numerical data to non-numerical data in the form of categorization entails a number of provisions pertaining to the maternal health risk data used in this study. Age, Systolic BP, Diastolic BP, Blood Sugar, Body Temp, Heart Rate, and Risk Level are among the attributes chosen for processing. Except for the Risk Level, all data in attributes is numerical. The data is classified according to the arrangements, as shown in Table 1. The attributes of the maternal health risk dataset are classified as follows:

Age Attribute

The age attribute data is classified in accordance with the Regulation of the Minister of Health of the Republic of Indonesia No. 25 of 2016 [26]. Age ranges are regulated in this regulation, so that age data is classified as follows:

Table 2. Age Category

Interval (years)	Category
5-11	Children
12-17	Teenager
18-59	Adult
>60	Elder

Systolic BP Attribute

Data on the Systolic Blood Pressure (Systolic BP) attribute is the blood pressure that occurs when the heart muscle contracts to pump blood through the arteries throughout the body [27], so systolic blood pressure for pregnant women is categorized as follows:

Table 3. Systolic BP Category

Interval	Category
Systolic<120	Under
Systolic=120	Normal
120<Systolic<129	Elevated
129<Systolic<139	High Blood Pressure Stage 1
139<Systolic<180	High Blood Pressure Stage 2
Systolic>180	Hypertensive Crisis

Diastolic BP Attribute

Data on the Diastolic Blood Pressure (Diastolic BP) attribute is the blood pressure in the arteries when the heart is resting or relaxing [27], so diastolic blood pressure for pregnant women is categorized as follows:

Table 4. Diastolic BP Category

Interval	Category
Diastolic<80	Under
Diastolic=80	Normal
80<Diastolic<89	High Blood Pressure Stage 1
89<Diastolic<120	High Blood Pressure Stage 2
Diastolic>120	Hypertensive Crisis

Blood Sugar Attribute

Blood Sugar is the amount of glucose a person has in their blood at a certain time. According to [28], a person has high blood sugar if the blood sugar at any time is more than 200 mg/dL, or 11 millimoles per liter (mmol/L) and has low blood sugar if the level drops drastically below 70 mg/dL or 3.9 millimoles per liter (mmol/L).

Body Temp Attribute

Body Temp is the temperature of the human body during activities. In research [29], human body temperature is categorized as normal if it is at 98 °F and categorized as high if it exceeds 98 °F.

Heart Rate Attribute

Heart rate is the number of times a person's heart beats per minute. The normal heart rate varies from person to person, but for adults, the normal range is 60 to 100 beats per minute. A normal heart rate, on the other hand, is dependent on the individual, age, body size, heart condition, activity level, use of certain medications, and even air temperature [30]. Because the dataset contains heart rates ranging from 60 to 100, it was divided using the Gini Index.

The following calculations are performed using equation (11). For example, for the heart rate attribute with a value of 60, there are 39 data points, with 19 in the LR category, 10 in the MR category, and 10 in the HR category. There are 461 data for the heart rate attribute with a value of >60, with 206 data in the LR category, 113 data in the MR category, and 142 data in the HR category. The Gini index value is calculated in the following manner.

For <=60 :
I_G(≤60)=1-((19/39)^2+(10/39)^2+(10/39)^2)
I_G(≤60) =1-(0,2373+0,0657+0,0657)
I_G(≤60) =1-0,3687
I_G(≤60) =0,6313

For >60 :
I_G(>60)=
1-((206/461)^2+(113/461)^2+(142/461)^2)
I_G(>60)=1-(0,1996+0,0601+0,0948)
I_G(>60)=1-0,3545
I_G(>60)=0,6455

After determining the Gini index value, the Gini Splitting Index value is calculated. The Gini Splitting Index value can be calculated using equation (12):

GINI_split=(39/500×0,6313)+(461/500×0,6455)
GINI_split=(39/500×0,6313)+(461/500×0,6455)
GINI_split=0,0492+0,5951
GINI_plit=0,0492+0,5951
GINI_split=0,6443

The Gini Splitting Index for the heart rate attribute with a value range of =60 to >60 is 0.6443. The same process is used to calculate the other value ranges for the heart rate attribute. Table 5 shows the results of the Gini Index and Gini Splitting Index calculation processes.

Table 5. Heart Rate Category

Heart Rate	Risk Level			Total	Gini Indeks	Splitting Index
	Low Risk	Mid Risk	High Risk			
<=60	19	10	10	39	0,6312	0,6443
>60	206	113	142	461	0,6454	
<=65	21	12	10	43	0,6295	0,6438
>65	204	111	142	457	0,6452	
<=66	46	21	25	92	0,6241	0,6437
>66	179	102	127	408	0,6481	
<=67	46	23	28	97	0,6356	0,6444
>67	179	100	124	403	0,6465	
<=68	46	24	28	98	0,6381	0,6444
>68	179	99	124	402	0,6459	
<=70	111	61	51	223	0,6251	0,6376
>70	114	62	101	277	0,6476	
<=75	115	61	55	231	0,6257	0,6386
>75	110	62	97	269	0,6496	
<=76	148	81	67	296	0,6239	0,6311
>76	77	42	85	204	0,6415	
<=77	178	86	80	344	0,6157	0,6232
>77	47	37	72	156	0,6399	
<=78	180	99	85	364	0,6270	0,6242
>78	45	24	67	136	0,6167	
<=80	203	111	112	426	0,6359	0,6296
>80	22	12	40	74	0,5931	
<=82	209	113	112	434	0,6337	0,6228
>82	16	10	40	66	0,5509	
<=86	211	122	125	458	0,6423	0,6283
>86	14	1	27	42	0,4751	
<=88	225	123	141	489	0,6419	0,6278
>88	0	0	11	11	0	
<=90	225	123	152	500	0,6446	0,6446
>90	0	0	0	0	1	

According to Table 5, the Heart Rate attribute has the smallest Gini Index value, namely 0.6228, with a value range between =82 and >82, so this range is used as an example for the Heart Rate attribute.

RESULT AND DISCUSSION

The Python programming language is used to process data in this study. The RF, K-NN, and SVM algorithms are used to classify the dataset. To determine whether data categorization has an effect on the classification model, the processing is repeated twice, once with the initial dataset and once with a dataset that has been categorized using the arrangement in Table 1 and the Gini Index.

Classification Results of Initial Data

Each classification algorithm's classification process refers to the previously presented equation. Equation (1) refers to RF classification, equation (2) refers to K-NN, and equations (3)-(10) refer to the SVM algorithm process. Table 6 shows the initial 500 data points

Table 6. Initial Dataset Points

Age	Systolic BP	Diastolic BP	Blood Sugar	Body Temp	Heart Rate	Risk Level
25	130	80	15	98	86	**HR**
35	140	90	13	98	70	**HR**
29	90	70	8	100	80	**HR**
30	140	85	7	98	70	**HR**
35	120	60	6.1	98	76	**LR**
23	140	80	7.01	98	70	**HR**
23	130	70	7.01	98	78	**MR**
35	85	60	11	102	86	**HR**
32	120	90	6.9	98	70	**MR**
42	130	80	18	98	70	**HR**
….	….	….	….	….	….	**….**
16	120	75	7.9	98	7	**LR**

Figure 1-3 depicts the classification results of the three algorithms.

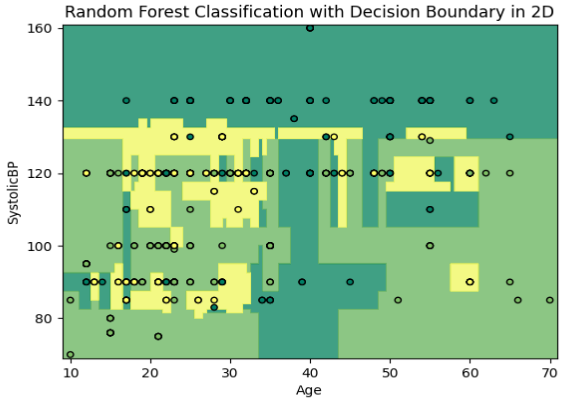


Fig 1. The RF **v**isualization of **i**nitial **d**ataset

The RF algorithm is depicted in Figure 1. 15 trees were used in the classification process. The process of aggregating all trees yields classification rules or models. The experiment was repeated five times, with the data divided into two categories: training data and testing data. The confusion matrix was used to test all models, with the highest accuracy of 88.5%.

The K-NN algorithm classification is performed by selecting the 5 nearest neighbors as determined by equation (2). Figure 2 depicts a visualization of the K-NN model. The K-NN model was developed through five experiments in which the data was divided into two categories: test data and test data. The highest model accuracy is obtained by dividing 95% of training data by 5% of testing data, yielding a value of 85%.

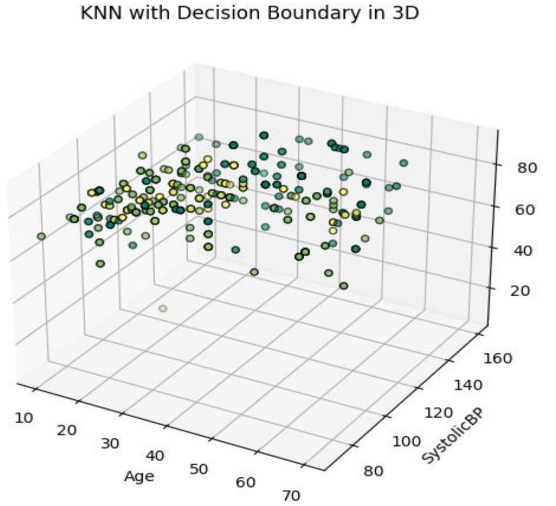


Fig 2. The K-NN visualization of initial dataset

The SVM algorithm is then used for classification. Figure 3 depicts a visualization of the hyperplane constructed from the first 500 datasets. High-dimensional data is included in Maternal Health Risk data. Aside from the large amount of data, this dataset consists of numerical data with extremely close instance proximity. The hyperplane visualization is shown in equations (3)-(10) as follows:

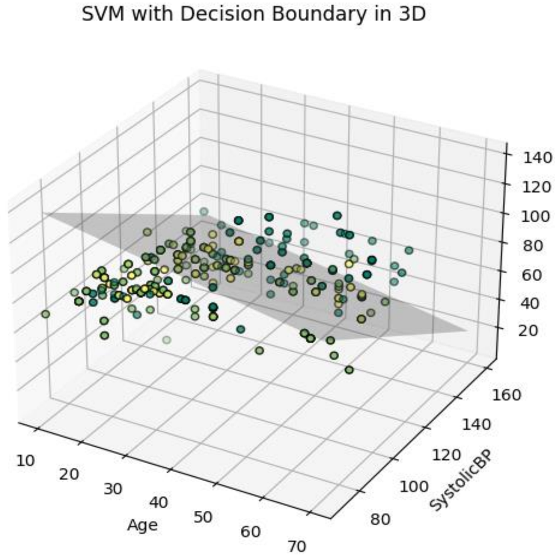


Fig 3. The SVM visualization of initial dataset

Figure 3 depicts a well-formed hyperplane. The same test was performed on all SVM models. With an accuracy value of 76.9%, the model with the highest accuracy is divided into 95% training data and 5% testing data.

Classification Results of Categorized Data

In this classification, the initial dataset in Table 6 is categorized following the arrangements in Table 1, with the following results in Table 7 and 8. The attributes of Age, Systolic BP, Distolic BP, Blood Sugar, and Body Temperature are the only attributes in numerical form and categorized by arrangements. Attribute of Heart Rate is categorized by the Gini Index. The dataset is then classified using the visualization, show in Figure 4.

Table 7. Dataset Categorized

No	Age	Systolic BP	Diastolic BP	Blood Sugar
1	Adult	High Blood Pressure Stage 1	Normal	High
2	Adult	High Blood Pressure Stage 2	High Blood Pressure Stage 2	High
3	Adult	Under	Under	Normal
4	Adult	High Blood Pressure Stage 2	High Blood Pressure Stage 1	Normal
5	Adult	Normal	Under	Normal
6	Adult	High Blood Pressure Stage 2	Normal	Normal
7	Adult	High Blood Pressure Stage 1	Under	Normal
8	Adult	Under	Under	Normal
9	Adult	Normal	High Blood Pressure Stage 2	Normal
10	Adult	High Blood Pressure Stage 1	Normal	High
….	….	….	….	….
500	Teenager	Normal	Under	Normal

Table 8. Dataset Categorized Continue

No	Age	Body Temp	Heart Rate	Risk Level
1	Adult	Normal	>82	HR
2	Adult	Normal	<= 82	HR
3	Adult	High	<= 82	HR
4	Adult	Normal	<= 82	HR
5	Adult	Normal	<= 82	LR
6	Adult	Normal	<= 82	HR
7	Adult	Normal	<= 82	MR
8	Adult	High	>82	HR
9	Adult	Normal	<= 82	MR
10	Adult	Normal	<= 82	HR
….	….	….	….	….
500	Teenager	Normal	<= 82	LR

Figure 4 shows the RF algorithm using categorized data. The classification process employs 15 trees and the same iterations as the initial dataset classification. Because the categorized dataset is high-dimensional, Python uses Principal Component Analysis (PCA) to create the model. Because determining the data hyperplane is difficult, PCA is required. The highest RF accuracy is found at 85% and a data split of 15% for a total of 78.9%.

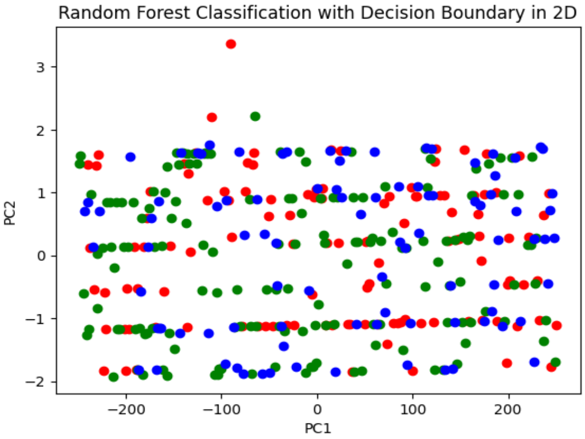


Fig 4. The RF visualization of categorized dataset

The K-NN algorithm, as shown in Figure 5, creates a data group with five classes. PCA is used in processing to produce a consistent distribution of data in each class. The visualization, on the other hand, displays a wide range of data. On a split of 95% training data and 5% testing data, visualization shows the best classification with an accuracy of 76.9%.

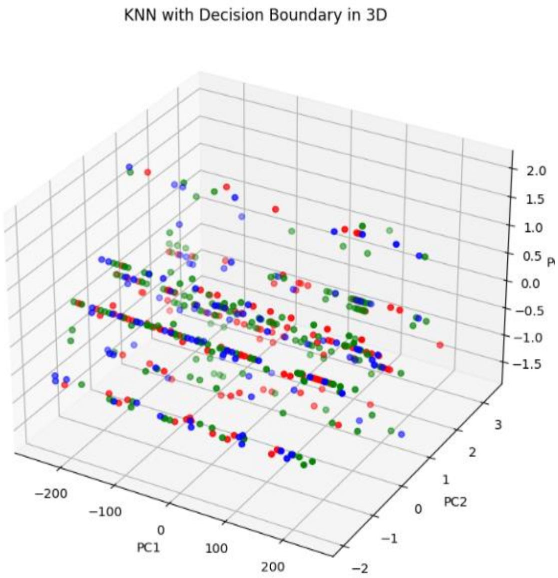


Fig 5. The K-NN visualization of categorized dataset

Figure 6 shows a visualization of the results of the SVM algorithm classification. The categorized dataset generates a hyperplane at boundary 0, so the data class boundary is determined by the data closest to the boundary that can be formed. Changes in the shape of the dataset have a significant impact on the hyperplane's formation, resulting in strict data boundaries for high-dimensional data.

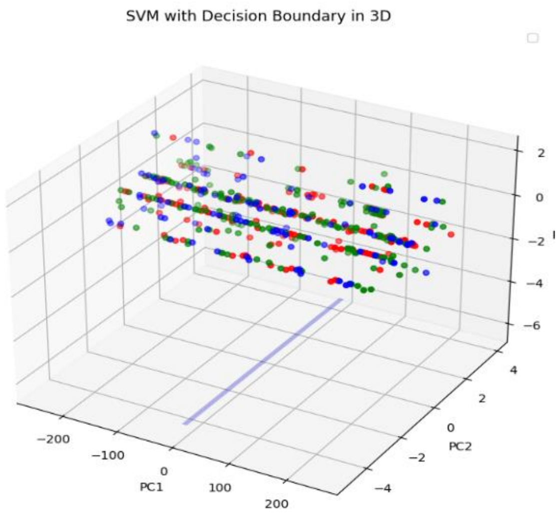
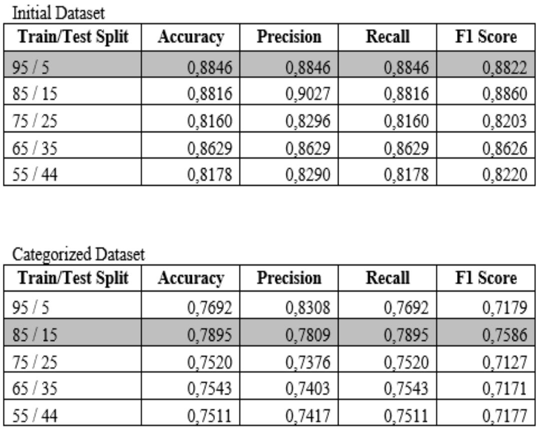


Fig 6. The SVM visualization of categorized dataset

Classification Model Testing

Confusion Matrix is used to test the model produced by the classification algorithm. This is to determine the model's feasibility, or whether it is acceptable. The highest classification model accuracy for each data split was presented in the previous explanation. In the three algorithms for the initial dataset and categorization dataset, the confusion matrices used are accuracy, precision, recall, and f1-score. Figure 7-9 displays the complete confusion matrix values in question.

The confusion matrix for the RF model is shown in Figure 7. The highest accuracy is found in different splits of training and testing data for each dataset, namely 95/5 percentage splitting for the initial dataset and 85/15 percentage splitting for the categorized dataset.



Initial Dataset

Train/Test Split	Accuracy	Precision	Recall	F1 Score
95 / 5	0,8846	0,8846	0,8846	0,8822
85 / 15	0,8816	0,9027	0,8816	0,8860
75 / 25	0,8160	0,8296	0,8160	0,8203
65 / 35	0,8629	0,8629	0,8629	0,8626
55 / 44	0,8178	0,8290	0,8178	0,8220

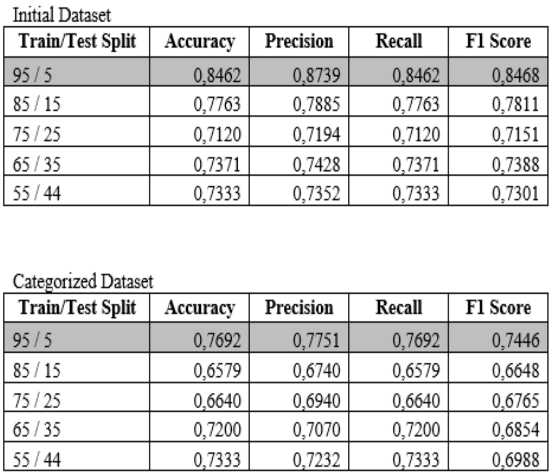
Categorized Dataset

Train/Test Split	Accuracy	Precision	Recall	F1 Score
95 / 5	0,7692	0,8308	0,7692	0,7179
85 / 15	0,7895	0,7809	0,7895	0,7586
75 / 25	0,7520	0,7376	0,7520	0,7127
65 / 35	0,7543	0,7403	0,7543	0,7171
55 / 44	0,7511	0,7417	0,7511	0,7177

Fig 7. RF confusion matrix

The categorized dataset has a lower confusion matrix than the initial dataset. The accuracy, precision, recall, and f1-score values are balanced at 85% and 15% splits. The RF model precision is 83.1% at 95% and 5% split. This demonstrates that the RF model from the categorized dataset performs well.

Figure 8 shows that the categorized dataset has a greater influence on K-NN. The 95/15 percentage splitting has the highest confusion matrix, while the 85/15 and 75/25 percentage splitting have the lowest. However, the K-NN model test results improved in the 65/35 and 55/45 percentage splitting. This is inversely proportional to the confusion matrix initial data, where the confusion matrix decreases in splits other than 95/5 percentage. This means that the more balanced the training and testing data are, the better the K-NN model test results with a categorized dataset will be.



Initial Dataset

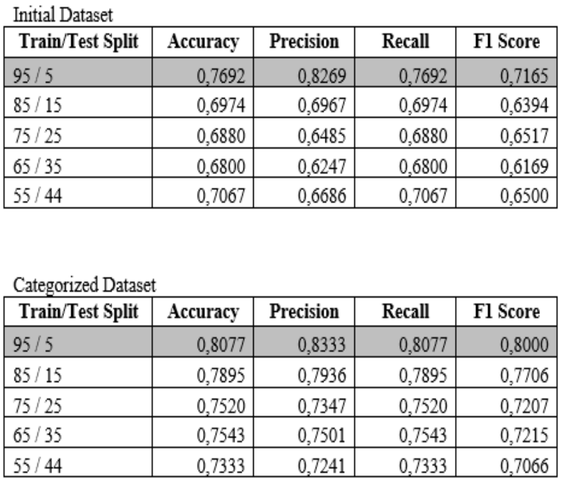
Train/Test Split	Accuracy	Precision	Recall	F1 Score
95 / 5	0,8462	0,8739	0,8462	0,8468
85 / 15	0,7763	0,7885	0,7763	0,7811
75 / 25	0,7120	0,7194	0,7120	0,7151
65 / 35	0,7371	0,7428	0,7371	0,7388
55 / 44	0,7333	0,7352	0,7333	0,7301

Categorized Dataset

Train/Test Split	Accuracy	Precision	Recall	F1 Score
95 / 5	0,7692	0,7751	0,7692	0,7446
85 / 15	0,6579	0,6740	0,6579	0,6648
75 / 25	0,6640	0,6940	0,6640	0,6765
65 / 35	0,7200	0,7070	0,7200	0,6854
55 / 44	0,7333	0,7232	0,7333	0,6988

Fig 8. K-NN confusion matrix

The final examination focuses on SVM classification. The results of the SVM model test are shown in Figure 9, where the highest accuracy, precision, recall, and f1-score values are in the 95/5 percentage splitting for both types of dataset. The confusion matrix values for all splits for categorized datasets in this study are higher than the initial dataset, with a decreasing tendency for balanced data splits. This demonstrates that SVM performance improves on categorized datasets.



Initial Dataset

Train/Test Split	Accuracy	Precision	Recall	F1 Score
95 / 5	0,7692	0,8269	0,7692	0,7165
85 / 15	0,6974	0,6967	0,6974	0,6394
75 / 25	0,6880	0,6485	0,6880	0,6517
65 / 35	0,6800	0,6247	0,6800	0,6169
55 / 44	0,7067	0,6686	0,7067	0,6500

Categorized Dataset

Train/Test Split	Accuracy	Precision	Recall	F1 Score
95 / 5	0,8077	0,8333	0,8077	0,8000
85 / 15	0,7895	0,7936	0,7895	0,7706
75 / 25	0,7520	0,7347	0,7520	0,7207
65 / 35	0,7543	0,7501	0,7543	0,7215
55 / 44	0,7333	0,7241	0,7333	0,7066

Fig 9. SVM confusion Matrix

The accuracy of the classification models of the three algorithms tested using categorized datasets varies. The accuracy of the three algorithms decreases in models classified with RF and K-NN. However, the decrease in RF test results is still greater than 70%, whereas for K-NN it is less than 70%, implying that the K-NN model has the lowest decrease in RF accuracy. In the SVM model, various conditions occur. The results of testing the model with the categorized dataset are actually higher than the results of testing the model with the initial dataset, with an average of more than 70%. The findings of this study suggest that data categorization influences whether or not the classification model produced by the three classification algorithms is good.

CONCLUSION

Changing numerical data to non-numerical data in the form of categorization has a significant impact on the classification model's accuracy. Categorized data refers to rules, and using the Gini Index to reduce or increase the number of instances can affect model testing results. The ability of the algorithm to read data determines whether a model's performance improves or degrades. The study's findings revealed that only the accuracy of the SVM algorithm increased when numerical data was replaced with non-numerical data, from 76.92% to 80.77%. This also demonstrated that the SVM algorithm performs better when the data changes from numerical to non-numerical.

This study mentioned several other factors that influence classification model formation, such as data splitting during testing, randomization during the bootstrapping process in the RF algorithm, and determining cluster center points in the K-NN algorithm, as well as other processes that accompany the classification process, such as CPA when reading data with a high number of dimensions.

This research can still be developed for scientific purposes by testing categorized data on other classification algorithms such as Neural Network and AdaBoost. For appropriate datasets, the categorization process can also be carried out using Python. It is hoped that the findings of this study will shed light on the performance of classification algorithms and increase understanding of data utilization.

REFERENCES

[1] O. Dammann, "Data, Information, Evidence, and Knowledge: A Proposal for Health Informatics and Data Science," *Online J. Public Health Inform.*, vol. 10, no. 3, p. 9, 2019, doi: 10.5210/ojphi.v10i3.9631.

[2] M. Islam, "Data Analysis: Types, Process, Methods, Techniques and Tools," *Int. J. Data Sci. Technol.*, vol. 6, no. 1, pp. 10–15, 2020, doi: 10.11648/j.ijdst.20200601.12.

[3] J. Sanders, "Defining terms: Data, information and knowledge," in *SAI Computing Conference 2016*, 2016, no. July, pp. 1–6, doi: 10.1109/SAI.2016.7555986.

[4] P. Frederick, J. C. Finley, and C. Magalis, "A Quantitative Analysis for Non-Numeric Data," *Int. J. Quant. Qual. Res. Methods*, vol. 11, no. 1, pp. 1–11, 2023, doi: 10.37745/ijqqrm13/vol11n1111.

[5] H. J. Park, "A method to convert non-numeric characters into numerical values in dynamic time warping for string matching," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, pp. 2660–2665, 2021, doi: 10.11591/ijece.v11i3.pp2660-2665.

[6] A. Ardiansyahroni, A. Tjalla, and M. Mahdiyah, "Data Kategorik dalam Penelitian: Review Bibliometrik," *J. Ilm. Mandala Educ.*, vol. 9, no. 1, pp. 796–802, 2023, doi: 10.58258/jime.v9i1.4814.

[7] K. Khadijah, N. Sabilly, and F. A. Nugroho, "Sentiment Analysis of League of Legends: Wild Rift Reviews on Google Play Using Naã• Ve Bayes Classifier," *J. Ilm. Kursor*, vol. 12, no. 1, pp. 23–30, 2023, doi: 10.21107/kursor.v12i01.328.

[8] T. O. Togunwa, A. O. Babatunde, and K. U. R. Abdullah, "Deep hybrid model for maternal health risk classification in pregnancy: synergy of ANN and random forest," *Front. Artif. Intell.*, vol. 6, no. July, pp. 1–11, 2023, doi: 10.3389/frai.2023.1213436.

[9] D. Mennickent *et al.*, "Machine learning applied in maternal and fetal health: a narrative review focused on pregnancy diseases and complications," *Front. Endocrinol. (Lausanne).*, vol. 14, no. May, pp. 1–22, 2023, doi: 10.3389/fendo.2023.1130139.

[10] Rekha S Kambli and Nirmala, "Model for Predicting Risk Levels in Maternal Healthcare," *Int. J. Adv. Res. Innov. Ideas Educ.*, vol. 8, no. 6, pp. 1633–1637, 2022.

[11] T. Ibrahim and A. Ridwan, "Determinan Penyebab Kematian Ibu dan Neonatal di Indonesia," *J. Kedokt. Nanggroe Med.*, vol. 5, no. 2, pp. 43–48, 2020.

[12] M. D. A. Rosyid and S. Subektiningsih, "Klasifikasi Tingkat Risiko Kesehatan Ibu Hamil Menggunakan Algoritma Support Vectore Machine," *Indones. J. Comput. Sci.*, vol. 12, no. 5, pp. 2798–2807, 2023, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135.

[13] T. Triana, E. Utami, and A. D. Hartanto, "Implementasi Algoritma Nearest Neighbor Pada Aplikasi Deteksi Resiko Tinggi Pada Kehamilan," *INFOKES J. Ilm. Rekam Medis dan Inform. Kesehat. Vol*, vol. 13, no. 2, pp. 64–71, 2023.

[14] D. M. U. Atmaja, A. R. Hakim, A. Basri, and A. Ariyanto, "Klasifikasi Metode Persalinan pada Ibu Hamil Menggunakan Algoritma Random Forest Berbasis Mobile," *JRST (Jurnal Ris. Sains dan Teknol.*, vol. 7, no. 2, pp. 167–174, 2023, doi: 10.30595/jrst.v7i2.16705.

[15] M. Savargiv, B. Masoumi, and M. R. Keyvanpour, "A new random forest algorithm based on learning automata," *Comput. Intell. Neurosci.*, vol. 2021, no., pp. 1–19, 2021, doi:

10.1155/2021/5572781.

[16] X. Peng *et al.*, "A Comparison of Random Forest Algorithm-Based Forest Extraction with GF-1 WFV, Landsat 8 and Sentinel-2 Images," *Remote Sens.*, vol. 14, no. 5296, pp. 1–16, 2022, doi: 10.3390/rs14215296.

[17] B. Zagajewski, M. Kluczek, E. Raczko, A. Njegovec, A. Dabija, and M. Kycko, "Comparison of random forest, support vector machines, and neural networks for post-disaster forest species mapping of the krkonoše/karkonosze transboundary biosphere reserve," *Remote Sens.*, vol. 13, no. 2581, pp. 1–23, 2021, doi: 10.3390/rs13132581.

[18] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *Inform. J. Ilmu Komput.*, vol. 18, no. 3, pp. 239–248, 2022, doi: 10.52958/iftk.v18i3.4694.

[19] H. A. Roysid, A. Maulana, and U. Pujianto, "Can K-Nearest Neighbor Method Be Used To Predict Success in Indonesia State University Student Selection," *Kursor*, vol. 9, no. 4, pp. 137–144, 2018, doi: 10.28961/kursor.v9i4.186.

[20] Q. Zheng, L. Wang, J. He, and T. Li, "KNN-Based Consensus Algorithm for Better Service Level Agreement in Blockchain as a Service (BaaS) Systems," *Electronics*, vol. 12, no. 1429, pp. 1–21, 2023, doi: 10.3390/electronics12061429.

[21] I. M. S. Bimantara and I. M. Widiartha, "Optimization of K-Means Clustering Using Particle Swarm Optimization Algorithm for Grouping Traveler Reviews Data on Tripadvisor Sites," *J. Ilm. Kursor*, vol. 12, no. 1, pp. 1–10, 2023, doi: 10.21107/kursor.v12i01.269.

[22] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 6, no. 2, pp. 129–136, 2021, doi: 10.30998/string.v6i2.10133.

[23] D. Jollyta, P. Prihandoko, A. Hajjah, E. Haerani, and M. Siddik, *Algoritma Klasifikasi Untuk Pemula Solusi Pyhton dan RapidMiner*. Yogyakarta: Deepublish, 2023.

[24] 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.

[25] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, pp. 1–16, 2016, doi: 10.1098/rsta.2015.0202.

[26] M. Kesehatan, *PERATURAN MENTERI KESEHATAN REPUBLIK INDONESIA NOMOR 25 TAHUN 2016 TENTANG RENCANA AKSI NASIONAL KESEHATAN LANJUT USIA TAHUN 2016-2019*, vol., no. 2016, p. 97.

[27] C. A. Peralta, R. Katz, A. B. Newman, B. M. Psaty, and M. C. Odden, "Systolic and diastolic blood pressure, incident cardiovascular events, and death in elderly persons: The role of functional limitation in the cardiovascular health study," *Hypertension*, vol. 64, no. 3, pp. 472–480, 2014, doi: 10.1161/HYPERTENSIONAHA.114.03831.

[28] Y. Wahyuni, C. Zaddana, A. Maesya, and A. Izzuddin, "Early detection model of normal and abnormal blood flow using pulse Oximetry non-invasive of pregnant heart rate," *Sink. J. dan Penelit. Tek. Inform.*, vol. 7, no. 3, pp. 2125–2133, 2022.

[29] I. I. Geneva, B. Cuzzo, T. Fazili, and W. Javaid, "Normal body temperature: A systematic review," *Open Forum Infect.*

Dis., vol. 6, no. 4, pp. 1–7, 2019, doi: 10.1093/ofid/ofz032.

[30] P. Kansara, R. Dhar, R. Shah, D. Mehta, and P. Raut, "Heart Rate Measurement," in *Journal of Physics: Conference Series*, 2021, vol. 1831, no. 1, p. 12, doi: 10.1088/1742-6596/1831/1/012020.

.