ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

A COMPARATIVE STUDY ON THE IMPACT OF FEATURE SELECTION AND DATASET RESAMPLING ON THE PERFORMANCE OF THE K-NEAREST NEIGHBORS (KNN) CLASSIFICATION ALGORITHM

I Gede Aris Gunadi[1], Dewi Oktofa Rachmawati[2]

[1,2]Universitas Pendidikan Ganesha

email: igedearisgunadi@undiksha.ac.id[1], dewioktofa.r@undiksha.ac.id[2]

Abstract

This study aims to evaluate the impact of dataset balancing and feature selection on the performance of the K-Nearest Neighbors (KNN) classification algorithm. The primary objective is to determine the effect of different training data balance ratios on classification performance. Additionally, the study analyzes the contribution of feature selection methods and data balancing to the overall performance of the classification algorithm. Three datasets (Titanic, Wine Quality, and Heart Diseases) sourced from Kaggle, were utilized in this research. Following the preprocessing stage, the datasets were subjected to three resampling scenarios with balance ratios of 0.3, 0.6, and 0.9. Feature selection was performed by combining correlation test values and information gain values, each weighted at 50%. The selected features were those with positive combined values of summation, correlation, and information gain. The KNN classification algorithm was then applied to datasets with and without feature selection. The results indicate that achieving a perfectly balanced ratio (ratio = 1) is not essential for improving classification performance. A balance ratio of 0.6 yielded results comparable to those of a perfect balance ratio. Furthermore, the findings demonstrate that feature selection has a more significant impact on classification performance than data balancing. Specifically, data with a balance ratio of 0.3 and feature selection outperformed data with a balance ratio of 0.6 but without feature selection.

Keywords : balanced data, feature selection, classification, knn

Received: 05-07-2024 | **Revised:** 23-07-2024 | **Accepted:** 27-07-2024
DOI: https://doi.org/10.23887/janapati.v13i2.82174

INTRODUCTION

Data balance problems are associated with the unbalanced distribution of data in each class in a dataset. The problem with a dataset with an unbalanced data distribution in each class is that the learning process is not optimal. For example, on the skin classification to determine sick skin or healthy skin, However, the distribution of training data is unbalanced. Data for healthy skin only contains 10% of the data, while sick data contains 90% of the data. The training process will not be perfect. Training on the recognition of healthy skin data will be very limited. In the end, when testing, if given data that is actually healthy skin data, it is likely to misidentify healthy skin recognized as diseased skin.

A study[1] conducted a study on the detection of transaction fraud on credit card data. It was found on a large number of datasets, and the data was unbalanced. In this study, 24 test scenarios were carried out with two test models, the Adabost model and the LGBM classifier model. The ratio of data imbalance testing is 0.25, 0.5, and 1. From this study, it was found that the first few facts about the influence of accuracy on unbalanced data did not show improvement, from unbalanced data to balanced data. In the performance of the recall, there is an improvement, but it is not significant. In the study, it is more important to select good features, rather than balance the data to get better classification performance.

In the article [2], it is stated that the condition of unbalanced data is actually not too problematic for some conditions. One of the conditions in question is the condition of a dataset with a large enough size. This condition is in accordance with the study [1], In this study, the IEEE CIS Fraud detection dataset with a size of 148,896 rows was used. In this study, it was found that the influence of data balance was not so significant.

Another paper states that the influence of the cassification algorithm becomes more

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

important than the effect of data balancing, as stated in the study [3]. In this study, it was stated that the CNN LSTM algorithm is strong enough to overcome the problem of data imbalance.

In addition to the conditions mentioned above, the data imbalance does not have a very significant effect if the right evaluation metrics are used. The AUPRC evaluation metrix is a good alternative to the evaluation metrix used to measure the performance of the casifier[4].

There are two questions as the background of this research. Firstly, if the data is naturally unbalanced, Statistically, we can do resampling to change the imbalance in the data, of course, the new data obtained does not reflect the population. At least whether a balance ratio of 100% is really needed. The second thing is that, naturally, there are strong features that can differentiate data classes. Under such conditions, the effect of data balance remains very significant. In this research, first we will analyze the extent of the influence of the training data balance ratio in determining the performance of the classification algorithm. Does the absolute maximum classification performance occur when the balance ratio is perfect? . Second, this research will also examine the comparison of the contribution of the influence of data balancing with the influence of selecting strong features on the performance of classification algorithms.

METHOD

The research procedures in the study were carried out as follows:

Research's Procedure

The research procedure was carried out in the process of determining the dataset, data pre-processing, resampling scenarios, feature selection, and classification testing.The flow of this research procedure is shown in Figure 1.

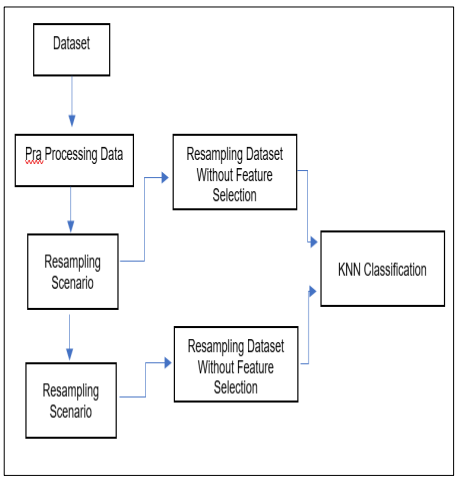


Figure 1. Research Procedure

Dataset

In several studies, several things related to data sets are stated based on the size of the data dimensions. In the article in[5], it is stated that datasets based on the number of rows are small datasets with a size or number of rows of less than 10,000 rows, medium datasets of less than one million rows, and large datasets of more than one million rows. The following are several things that can be summarized regarding the differences between small, medium, and large datasets. In the research process, three small test datasets found on Kagle were used, namely:

Tictanic Dataset : 889 rows, 2 classes
Wine Dataset : 1443 rows, 6 classes
Heart Diseases Dataset : 1025 rows, 2 classes

Titanic, Wine, and Heart deseases dataset information is as follows:

Table 1. Titanic information

RangeIndek : 889 enteries, 0 to 888		
Data columns (total 9 columns)		
column	**Non Null Count**	**Dytpe**
passegerId	889 non-null	Int64
Survived	889 non-null	Int64
Pclass	889 non-null	Int64
Sex	889 non-null	Object
Age	889 non-null	Float64
SibSp	889 non-null	Int64
Parch	889 non-null	Int64
Fare	889 non-null	Int64
Embarked	889 non-null	Int64

In the Titanic dataset, there is one feature (sex) of non-numerical data type, so before processing, this feature must be transformed into numerical data: male: 1, female: 0.

Table 2. Wine Dataset information

RangeIndek : 1143 enteries, 0 to 1142		
Data columns (total 13 columns)		
column	**Non Null Count**	**Dtype**
fixed acidity	1143 non-null	float64
Volatile acidity	1143 non-null	float64
citric Acid	1143 non-null	float64
risidual Sugar	1143 non-null	float64
chlorides	1143 non-null	float64
free Sulfur D	1143non-null	float64
total Sulfur D	1143 non-null	float64
density	1143 non-null	float64
ph	1143 non-null	float64
sulphates	1143 non-null	float64
alcohol	1143 non-null	float64
quality	1143 non-null	Int64
Id	1143 non-null	Int64

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

Table 3. Heart Deseases Dataset information

RangeIndek : 1025 enteries, 0 to 1024		
Data columns (total 14 columns)		
column	Non Null Count	int64
Age	1025 non-null	Int64
Sex	1025 non-null	Int64
Cp	1025 non-null	Int64
trestbps	1025 non-null	Int64
chol	1025 non-null	Int64
fbs	1025 non-null	Int64
restecg	1025 non-null	Int64
thalach	1025 non-null	Int64
exang	1025 non-null	Int64
oldpeak	1025 non-null	float64
slope	1025 non-null	Int64
ca	1025 non-null	Int64
thal	1025 non-null	Int64
target	1025 non-null	Int64

Resampling Scenario

Data resampling is intended to overcome the condition of unbalanced training data. Unbalanced training data conditions cause the training process to be less than optimal. In the minority data class, the amount of data is very minimal, so the minority class data is difficult to recognize. This recognition error is due to very minimal recognition training. In ideal conditions, the training data must be balanced, or at least close to balanced conditions. To overcome unbalanced training data, a resampling process is carried out.

In several studies [6], [7], [8], and [9], it is stated that the resampling process can be carried out in two ways. The first is oversampling, namely generating new data in the minority class, so that data balance occurs. Second, resampling with undersampling The underside process is deleting data in the majority class so that data balance occurs.

The oversampling process is illustrated in Figure 3. One of the oversampling methods [10], [11], and [12], uses the SMOTE (Synthetic Minority Oversampling Technique) method. An illustration of SMOTE is shown in Figure 2. In this research, it was stated that SMOTE made a good contribution to improving classification performance. New data is generated using the following equation 3.

$$X\ new = Xi + (RN) * (Xzi - Xi) \quad (3)$$

Xnew : The New Data is generated
Xi : Old data selected
RN : random Number (0 -1)
Xzi : Nears Neigbors Xi

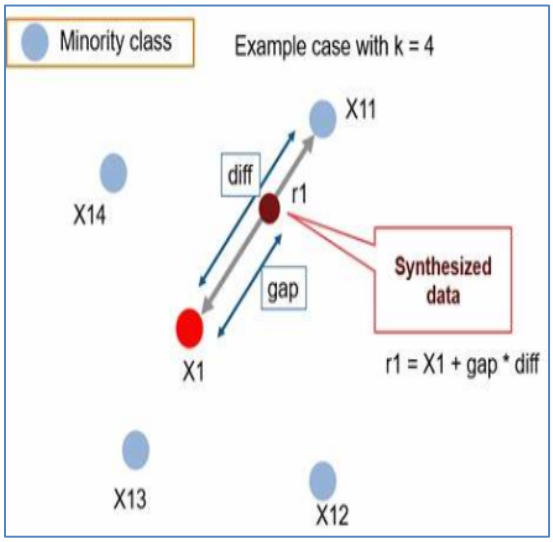


Figure 2. Ilustration of SMOTE ([13]

Meanwhile, the undersampling process can be carried out using several methods, including Random EnderSampling (RUS), Tomek Links (TL), UnderSampling based on Clustering (BSC), Evolutionary Undersampling. One illustration of the most commonly used undersampling method, the RUS method, is shown in Figure 3.

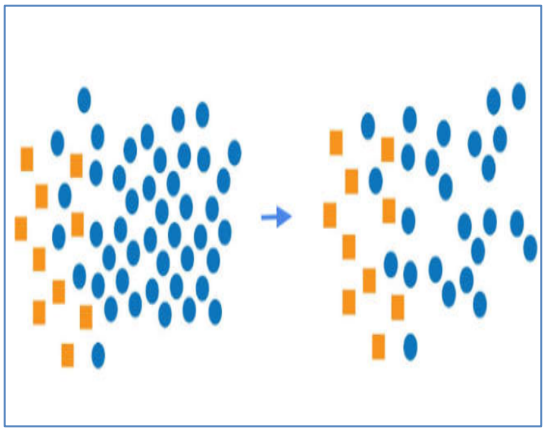


Figure 3. Ilustration of UnderRandomSampling (RUS) ([13])

The RUS method is quite widely used because it is easy to implement, but there are actually fundamental weaknesses. Because the process is random, it does not differentiate the quality of data membership in the majority class. There is a possibility that the data that is deleted is data that is at the center of the cluster. In principle, the quality of data deletion will be good if the data being deleted is data that is in the border area between clusters or classes.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

The resampling scenario for each dataset is carried out with three conditions based on the ratio of the amount of data from the minority class to the majority class. The dataset will be tested in the classification process with the following conditions:: ratios of 0.3, 0.6, and 0.9. To obtain these conditions, the RandomOverSampling algorithms are used. By using the imblearn library in Python, we can access both resampling algorithms.

Features Selection

Feature selection in this research was carried out by combining correlation test methods and feature information gain.

Gain Information

One method to improve the performance of classification algorithms is to carry out good pre-processing. The selection of strong data features is important. Gain analysis is a fairly popular method for feature selection, as used in research [14], [15], and [6].

Feature gain, often called the gain information Information Ratio, is the value of a feature. This value describes how well the feature differentiates one class from other classes in a dataset. To determine the feature gain value, you must first determine the entropy of the dataset with n classes. Entropy is determined by the formula in Equation 1.

$$S = -\sum_{i=1}^{n} -P\,i.log2(Pi) \quad (1)$$

Pi is the probability of the appearance of class i in the dataset. The meaning of entropy is the distribution of class i in the data set. There are three possible entropy values.

Table 4. Perception of Entropy Value

Entropy	Deskripsi
0	It shows that the dataset is homogeneous, containing only one class.
0 - 1	It shows the distribution of each class is diverse (unbalanced data)
1	It shows the distribution of each class balance

Based on the entropy value, the gain value of feature A can be determined, using Equation 2, as follows:

$$(S,A) = E - \sum_{j=1}^{j} \frac{|Sj|}{|S|}\, Entropy\,(Sj) \quad (2)$$

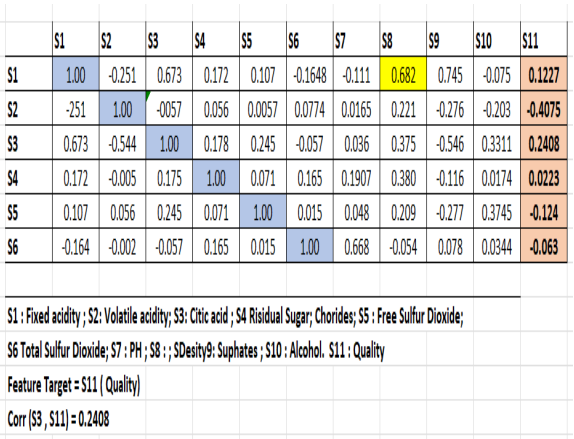
The interpretation of the gain value is: (1) A high gain value indicates that the feature with that gain value has a very large contribution in dividing data into existing classes. This means that features with high gain are strong features. They must be retained in the dataset. (2) Features with low gain values have the opposite meaning, meaning that these features are weak features and can be removed from the dataset.

Correlation Test

In the book [7], it is stated that data correlation tests are important to carry out at the pre-processing stage. The goal is the same to determine strong features in the dataset.

There are several correlation test methods, including (1) Pearson correlation test, (2) ANOVA test. (3) Spearman correlation test, (4) Kendal correlation test, (5) Phi coefficient correlation.

Basically, correlation tests are carried out to see the relationship between one variable and other variables in a dataset. The correlation test process is carried out in two ways. First, look at the correlation between the feature variable and the target variable. Second, perform correlation between feature variables. Figure 4 shows a snapshot of the dataset.



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
S1	1.00	-0.251	0.673	0.172	0.107	-0.1648	-0.111	0.682	0.745	-0.075	0.1227
S2	-251	1.00	-0057	0.056	0.0057	0.0774	0.0165	0.221	-0.276	-0.203	-0.4075
S3	0.673	-0.544	1.00	0.178	0.245	-0.057	0.036	0.375	-0.546	0.3311	0.2408
S4	0.172	-0.005	0.175	1.00	0.071	0.165	0.1907	0.380	-0.116	0.0174	0.0223
S5	0.107	0.056	0.245	0.071	1.00	0.015	0.048	0.209	-0.277	0.3745	-0.124
S6	-0.164	-0.002	-0.057	0.165	0.015	1.00	0.668	-0.054	0.078	0.0344	-0.063

S1 : Fixed acidity ; S2: Volatile acidity; S3: Citic acid ; S4 Risidual Sugar; Chorides; S5 : Free Sulfur Dioxide;
S6 Total Sulfur Dioxide; S7 : PH ; S8 : ; SDesity9: Suphates ; S10 : Alcohol. S11 : Quality
Feature Target = S11 (Quality)
Corr (S3 , S11) = 0.2408

Figure 4. Correlation Between Featur With Target in Wine Quality Dataset

In Figure 4, the target variable is quality, and the other variables (fixed acidity, volatile acidity, citric acid, and residual sugar) are feature variables. The interpretation of the correlation between the feature variable and the target variable is that the greater the correlation value, that means the feature variable is an important feature and must be maintained. Conversely, if the correlation value is low, then the feature variable is a weak feature and can be removed.

The second interpretation of correlation is correlation between features. The meaning is

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

that if the correlation value between features is small, then both features are meaningful, mutually independent features. On the other hand, if the correation value between features is high, it shows that the two features actually depend on each other, both features have the same meaning, and feature redundancy occurs. Then one of the features can be deleted. To clarify the meaning of the correlation between features, it is shown in the illustration in Figure 5.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	1.00	-0.251	0.673	0.172	0.107	-0.1648	-0.111	0.682	0.745	-0.075
S2	-251	1.00	-0057	0.056	0.0057	0.0774	0.0165	0.221	-0.276	-0.203
S3	0.673	-0.544	1.00	0.178	0.245	-0.057	0.036	0.375	-0.546	0.3311
S4	0.172	-0.005	0.175	1.00	0.071	0.165	0.1907	0.380	-0.116	0.0174
S5	0.107	0.056	0.245	0.071	1.00	0.015	0.048	0.209	-0.277	0.3745
S6	-0.164	-0.002	-0.057	0.165	0.015	1.00	0.668	-0.054	0.078	0.0344

S1 : Fixed acidity ; S2: Volatile acidity; S3: Citic acid ; S4 Risidual Sugar; Chorides; S5 : Free Sulfur Dioxide;
S6 Total Sulfur Dioxide; S7 : PH ; S8 : ; SDesity9: Suphates ; S10 : Alcohol.

Corr (S1 , S8) = 0.682

Figure 5. Correlation Between Features in Wine Quality Dataset

Test the correlation between feature variables and the target variable; for example, corr (citric acid, quality) = 0.2408, while the correlation between corr (fixed acidity, quality) = 0.1219. This shows that the citric acid feature is stronger than the fixed acidity feature.Furthermore, for the correlation between feature variables, corr (fixed acidity, density) = 0.682. This value is relatively high. This shows that in the dataset, the fixed acidity feature and the density feature are similar features. So that in the feature selection process, one of them can be removed. In this research we used correlation method base on Spearman correlation.

Meaningfulness of Features

Base on gain information value and correlation value is determined new variable. A new variable is stated as feature meaningfulness, which is the sum of the correlation values and gain information. Meaningfulness of features is calculated with the equation 3.

$$Meaningfulness\ of\ features = 0.5 * Corr + 0.5 * Gain \quad (3)$$

The selected features are features with a positive meaningful value. Furthermore, from the 3 datasets used after feature selection, the following are the features used in each dataset.

Table 5 . Selected Fature

Dataset	Selected Features
Titanic	Fitur : 'Sex','Age','Parch','Fare', Target : 'Survived'
Heart Deseases	Fitur : 'cp', 'chol', 'restecg', 'thalach', 'slope' Target : 'quality'
Wine	Fitur : 'fixed acidity','citric acid','residual sugar','pH','sulphates','alcohol', Target : 'quality'

KNN (K Nearest Neighboors) Classification

The classifier used in this research is KNN (K Nearest Neighbors). In this study, classification was carried out with 3 K values, namely K = 5, K = 9, and K = 15. The final performance uses the average test performance value of the 3 K values.

The reason for using 3 variations of K values (K = 5, K = 9, and K = 15) is that in many tests, generally the tests are in the range K = 5 to K = 15. The range of K = 13 or K = 15 is considered sufficient to provide optimal classification performance. In a study [16], an attempt was made to determine the best K by testing from K = 1 to K = 49, and the best K was obtained at K = 13. Thus, the selection of variations K = 5, 9, and 15 is believed to provide optimal classification performance.

Evalution of Performance Classifer

Data quality, at the end, will be tested for classification performance (for classification purposes). In general, there are four confusion matrix-based parameters that can be used to measure classifier performance: accuracy, precision, recall, and F1 score.

$$Accuration = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

$$Precesion = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ Score = 2.\frac{Precesion * Recall}{Precesion + Recall} \quad (7)$$

The TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative) values refer to a matrix comparing the real conditions of

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

the data class with the predicted results of the data class, expressed as a confusion matrix.

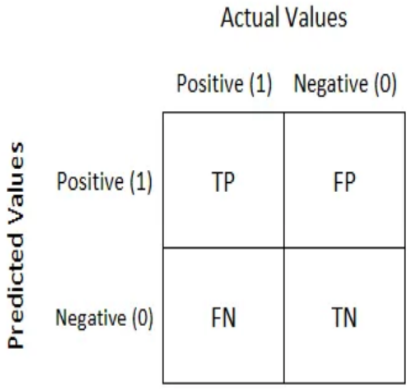


Figure 6. Confusion Matrix

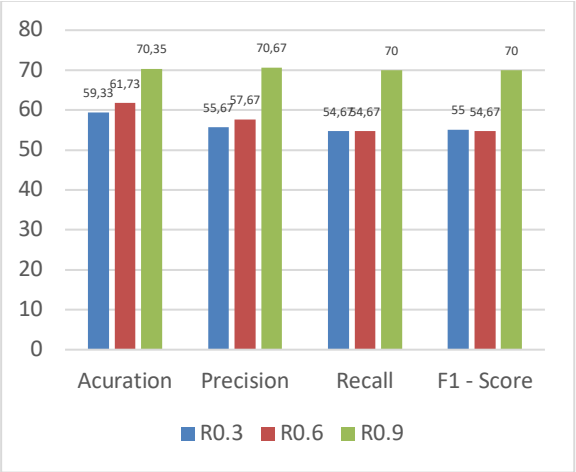


Figure 7. Visualisation Classification,Performance on D.Set Titanic Without Feature Selection

RESULTS AND DISCUSSION

The results of this research are described to analyze the influence of the resampling process on classification performance as well as the influence of feature selection on the performance of classification algorithms.

Results on the Titanic Dataset

The research results on three datasets were obtained as follows. The following is an evaluation of the Titanic dataset:

Tabel 6. Classification Perfermance on D.Set Titanic Without Feature Selection

	Accuracy	Precision	Recall	F1 - Score
R-0.3	59.33	55.67	54.67	55.00
R-0.6	61.73	57.67	54.67	54.67
R-0.9	70.35	70.67	70.00	70.00

Table 6 , explains the performance with feature selection.

Tabel 7. Classification Perfermance on D.Set Titanic WithFeature Selection

	Accuracy	Precision	Recall	F1 - Score
R- 0.3	78.15	67.33	62.33	63,67
R -0.6	78.03	76.67	77.33	33.00
R-0.9	70.35	70.67	70.00	70.00

R 0.3, R 0.6, and R 0.9 represent the minority-to-majority ratio. Data visualization showing classification comparisons with selection features and without selection features for three resampling scenarios.

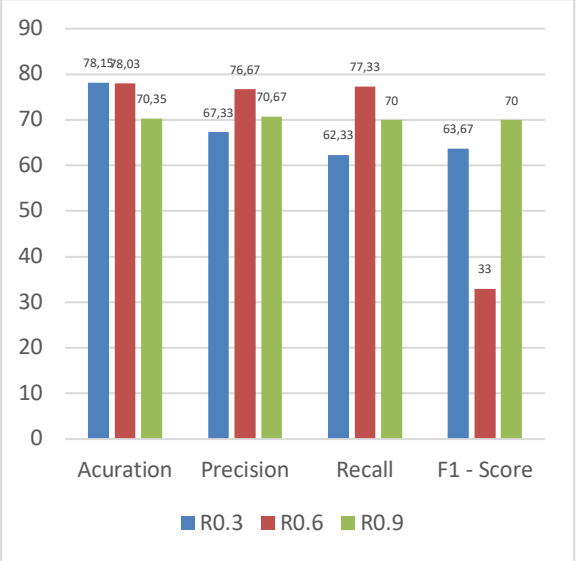


Figure 8. Visualisation Classification Perfermance on D.Set Titanic With Feature Selection

Results on the Wine Dataset

The following is an evaluation of the Wine dataset:

Table 8. Classification Perfermance on D.Set Wine Without Feature Selection

	Accuracy	Precision	Recall	F1 - Score
R0.3	47.56	26.27	25.00	25.00
R0.6	51.18	59.00	65.33	60.33
R0.9	63.54	62.67	68.00	63.67

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

Table 9. Classification Perfermance on D.Set Wine Without Feature Selection

	Acuration	Precession	Recall	F1 - Score
R0.3	56.03	33	31	31
R0.6	64.23	64.67	69.67	65.33
R0.9	68.65	68	71.67	68.67

Data visualization comparing classification performance without feature selection with feature selection.

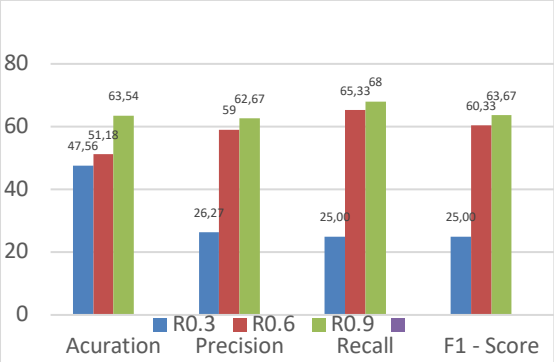


Figure 9. Visualization Classification , Performance on D.Set Wine Without Feature Selection

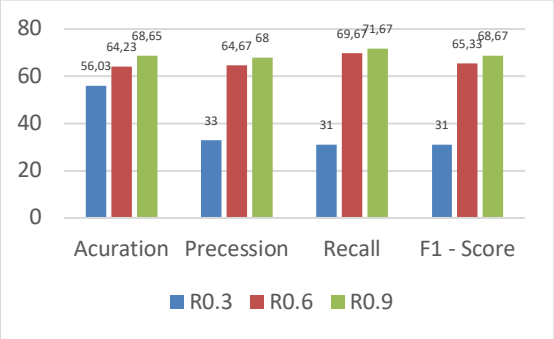


Figure 10. Visualization Classification, Performance on D.Set Wine With Feature Selection

Results on the Heart Deseases Dataset
The following is an evaluation of the heart deseases dataset:

Table 10. Classification Perfermance on D.Set Heart Deseases Without Feature Selection

	Acuracy	Precision	Recall	F1 - Score
R0.3	72.63	70.67	64.33	64.67
R0.6	74.14	75.00	74.33	74.33
R0.9	75.05	75.33	75.33	75.00

Table 11. Classification Perfermance on D.Set Heart Deseases Without Feature Selection

	Accuracy	Precision	Recall	F1 - Score
R0.3	74.81	74.67	71.00	69.33
R0.6	75.40	75.33	74.33	74.33
R0.9	75.69	76.00	75.33	74.00

The data visualization representation is in Figure 11 and Figure 12.

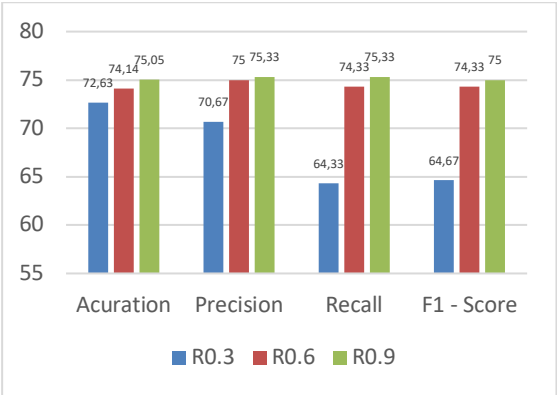


Figure 11. Visualization Classification, Performance on D.Set Heart Deseases Without Feature Selection

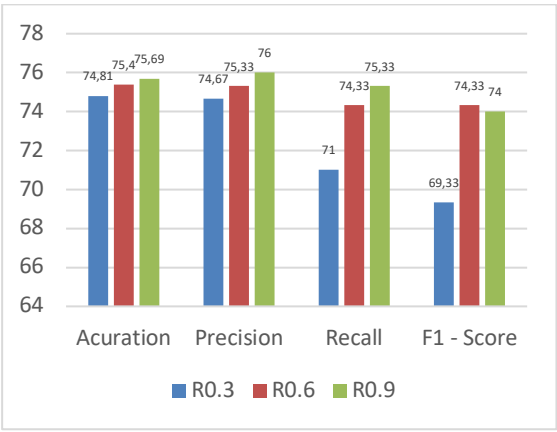


Figure 12. Visualisation Classification, Performance on D.Set Heart Deseases With Feature Selection

Disccusion
Data was obtained from three datasets: Titanic, Wine, and Heart Desases. Testing was carried out by resampling the minority to majority ratio of 0.3, 0.6, and 0.9. Other treatments are also given without selection features and with selection features. Based on this data, we try to see data patterns to answer several problems.The first problem is related to how data resampling is affectedon the performance of classification algorithms. The same pattern was found to show

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 2, July 2024

a significant increase in performance when resampling was increased from a ratio of 0.3 to a ratio of 0.6. However, there are other findings that when the ratio is increased from 0.6 to 0.9, the increase in performance is not as high as in the resampling condition of 0.3 to 0.6. The data pattern is shown in Table 10.
The wine and heart disease datasets show a very significant increase in accuracy performance of 22% and 15%, respectively, when the resampling ratio is carried out from 0.3 to 0.6. However, when the resampling ratio was increased from 0.6 to 0.9, there was still an increase, but the increase was no longer significant.
Titanic dataset with the selection feature when resampling from 0.3 to 0.6, there was an increase in precision and accuracy of 6% and 1%, respectively, but after resampling to 0.9, there was a decrease in performance.

Table 12. Precision Accuracy Performance for each Resampling increment.

		Without Feature Selection		With Feature Selection	
		Precision (%)	Accuracy(%)	Precision (%)	Accuracy%)
Titanic	R.OverSampling 0.3 to 0.6	4	4	14	1
	R.OverSampling 0.6 to 0.9	23	14	-8	-10
Wine	R.OverSampling 0.3 to 0.6	121	22	96	15
	R. OverSampling 0.6 to 0.9	6	9	5	7
Heart D	R.OverSampling 0.3 to 0.6	6	2	1	1
	R.OverSampling 0.6 to 0.9	0	1	1	0

In theory, the more balanced the data for each class, the ratio is closer to 1, the better pattern recognition during training will be. The better the training, the better the classification performance. However, based on data analysis in this study, this is not entirely true. The composition of the minority to majority class ratio in the range of 0.6 is good enough for class introduction. Increasing the ratio to near perfect balance conditions will not have a significant effect on classification performance.
The second problem is related to comparing the influence of feature selection with the influence of data resampling on the performance of the classification algorithm. To find out this, we try to compare the performance of the algorithm with feature selection at a ratio of 0.3 with the performance of the algorithm without feature selection at a ratio of 0.6 and 0.9. The results obtained are shown in Table 11.

Table 11. Comparison of Feature Selection Accuracy Against Resampling

	Titanic		Wine		Heart D	
	FS %	RS %	FS %	RS %	FS %	RS%
RS.0.3	78		56		75	
RS.0.6	78	61	64	58	76	74
RS.0.9		70		63		75

Table12. Comparison of Feature Selection Precession Against Resampling

	Titanic		Wine		Heart D	
	FS %	RS %	FS %	RS %	FS %	RS%
RS.0.3	67		33		75	
RS.0.6	76	57	64	59	75	75
RS.0.9		70		62		75

Based on tables 11 and 12, it can be stated that the influence of feature selection is stronger than data resampling.The data resampling process to get balanced training data is important, but the feature selection process should be prioritized. Ideal conditions to improve the classification performance of both are carried out. However, based on this research, it can be stated that the priority of feature selection is more important than data resampling for data balancing.

CONCLUSIONS

There are several things that can be concluded in this research, including:The first method that can be used to improve the performance of the classification algorithm is balancing the training data. However, the balance ratio does not absolutely have to be 1, the amount of training data for each class is the same. The condition of the balance ratio of the

minority class to the majority at a value of around 0.6 already provides a value that is relatively the same as a ratio of 1.

Both effects of feature selection are stronger compared to training data balancing. In priority to improve the performance of the feature selection classification algorithm compared to balancing the training data, In this study, it was found that training data with a balance ratio of 0.3 with feature selection had relatively the same performance as data with a ratio of 0.6 without feature selection.However, further testing needs to be carried out involving more datasets to ensure the correctness of the conclusions of this research.

REFFRENCES

[1] I. W. Dharmana, I. G. A. Gunadi, and L. J. E. Dewi, "Deteksi Transaksi *Fraud* Kartu Kredit Menggunankan *Oversampling* ADASYN dan Seleksi Fitur SVM-RFECV," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 125–134, 2024, doi: 10.25126/jtiik.20241117640.

[2] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, 2018, doi: 10.1186/s40537-018-0151-6.

[3] A. Balla, M. H. Habaebi, E. A. A. Elsheikh, M. R. Islam, and F. M. Suliman, "The Effect of Dataset Imbalance on the Performance of SCADA Intrusion Detection Systems," *Sensors*, vol. 23, no. 2, 2023, doi: 10.3390/s23020758.

[4] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big Data," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00724-5.

[5] A. C. Muller and S. Guido, *Introduction to Machine Learning with Python*. 2023. doi: 10.2174/9789815124422123010l.

[6] L. D. Utami *et al.*, "Integrasi Metode Information Gain untuk Seleksi Fitur dan AdaBoost untuk Mengurangi Bias pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naive Bayes," *J. Intell. Syst.*, vol. 1, no. 2, pp. 120–126, 2015.

[7] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Secxond., vol. 23, no. 12. Newyork: Elsevier, 2022. doi: 10.1016/b978-0-08-100741-9.00012-7.

[8] A. Indrawati, "Penerapan Teknik Kombinasi Oversampling Dan Undersampling Untuk Mengatasi Permasalahan Imbalanced Dataset," *JIKO (Jurnal Inform. dan Komputer)*, vol. 4, no. 1, pp. 38–43, 2021, doi: 10.33387/jiko.v4i1.2561.

[9] W. Ustyannie and S. Suprapto, "Oversampling Method To Handling Imbalanced Datasets Problem In Binary Logistic Regression Algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 14, no. 1, p. 1, 2020, doi: 10.22146/ijccs.37415.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. February, pp. 321–357, 2002, doi: 10.1613/jair.953.

[11] S. Sofyan and A. Prasetyo, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Tingkat Pendapatan Pekerja Informal Di Provinsi D.I. Yogyakarta Tahun 2019," *Semin. Nas. Off. Stat.*, vol. 2021, no. 1, pp. 868–877, 2021, doi: 10.34123/semnasoffstat.v2021i1.1081.

[12] A. Y. Triyanto and R. Kusumaningrum, "Implementation of Sampling Techniques for Solving Imbalanced Data Problem in Determination of Toddler Nutritional Status using Learning Vector Quantization," *Jur. Ilmu Komputer/Informatika Univ. Diponegoro*, vol. 19, no. 12, pp. 39–50, 2017.

[13] *Hairani, M.Eng. i*. Mataram: Universitas Bumigora, 2023.

[14] I. M. Arya, A. Dwija, I. M. Gede, and I. G. Aris, "JTIM : Jurnal Teknologi Informasi dan Multimedia https://journal.sekawan-org.id/index.php/jtim/ Perbandingan Algoritma Naive Bayes Berbasis Feature Selection Gain Ratio dengan Naive Bayes Kovensional dalam Prediksi Komplikasi Hipertensi I Made Arya Adinat," vol. 6, no. 1, pp. 37–49, 2024, [Online]. Available: https://doi.org/10.35746/jtim.v6i1.488

[15] F. Septianingrum and A. S. Y. Irawan, "Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes: Sebuah Literature Review," *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 799, 2021, doi: 10.30865/mib.v5i3.2983.

[16] Indrayanti, D. Sugianti, and A. Al Karomi, "OPTIMASI PARAMETER K PADA ALGORITMA K-NEAREST NEIGHBOUR UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS," in *SNATIF 4*, 2017, pp. 823–829.