SOLUTION OF UNBALANCED DATA CLASSIFICATION WITH A BASED APPROACH COMBINATION OF OVERSAMPLING AND UNDERSAMPLING

Riza Susanto[1], Irwan Budiman[2], Dodon T. Nugrahadi[3], M. Reza Faisal[4], Friska Abadi[5]
[1,2,3,4,5]Computer Science Study Program FMIPA ULM
Jl. A. Yani Km 36 Banjarbaru, South Kalimantan
Email: rizasusanto.tube@gmail.com

Abstract

This study applies the Combination of Oversampling and Undersampling method to deal with class imbalances. Researchers do Preprocessing to normalize the attributes used for prediction, then divide the training data and testing data. Researchers resampled unbalanced data using Oversampling, Undersampling and a combination of Oversampling and Undersampling. The results of the classification with the experimental data class balancing approach, the best classification performance is the combination of Oversampling and Undersampling classified by the k-Nearest Neighbor (KNN) method with an accuracy of 0.8672; sensitivity of 0.9000; specificity of 0.3750; and AUC of 0.6651042. Classification with Oversampling has performance results, namely accuracy of 0.875; sensitivity of 0.9250; specificity of 0.1250; and AUC of 0.6078125, while Undersampling classification has classification performance, namely accuracy of 0.3438; sensitivity of 0.33333; specificity of 0.50000; and AUC of 0.3645833.

Keywords *: Oversampling, Undersampling, k-Nearest Neighbor (KNN).*

1. INTRODUCTION

Data mining is the process of looking for interesting patterns or information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining are very diverse. The selection of the right technique or method really depends on the objectives and the overall KDD (Knowledge Discovery in Database) process. Data mining has an important function to help obtain useful information and increase knowledge for users.

Classification is the process of finding a model or function in describing a class or label for training, generally data in the real world. The growing availability of data in large-scale and complex systems is an important reason to advance the understanding of knowledge discovery and analysis of raw data to support decision-making processes. Although knowledge discovery techniques and data discovery techniques have shown great success in real-world applications, the problem in machine learning that is class imbalance is still a challenge in the case of classification. Class imbalance occurs when there are classes that have a much more dominating amount of data between one another.

Sampling is a part of statistical science that focuses on research on the selection of data generated from a collection of data populations. Sampling method or better known as resample is a common method used in solving data imbalance problems. With the application of **sampling** on imbalanced data, the level of imbalance is getting smaller and classification can be done correctly. In

addition, there are other methods that can be used to balance the data, namely the oversampling and undersampling methods.

From the problems above, this study proposes to compare classification algorithms to handle unbalanced datasets, so that a model that fits the data used is obtained. Researchers want to use a combination of oversampling and undersampling techniques to carry out the data balancing process. The combination of oversampling and undersampling is a method to balance data by increasing and decreasing data classes in a balanced way.

Based on the description above, the authors are interested in conducting research by applying a combination of oversampling and undersampling to handle data that is not, then classified using the k-Nearest Neighbor (k-NN) method.

2. RESEARCH METHODS

2.1. Research procedure

The research procedures carried out in this study are as follows:

a. Data Collection, Dataset in the form of Blood Transfusion Service Center originating from UCI Machine Learning Repository. The Blood Transfusion Service Center data has a total of 598 data, has 5 attributes, has 2 data classes and unbalanced data classes where the majority class = 0, has 570 data and the minority class = 1, has 28 data.

b. The research flow, schematic and research modeling are presented in Figure 1 as follows.

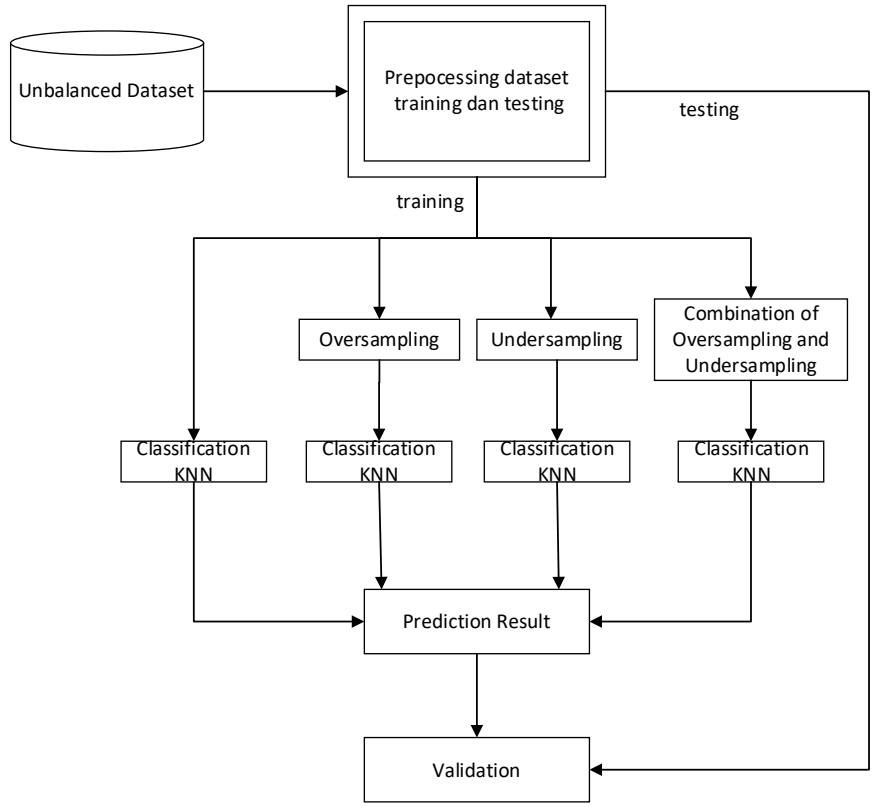

Figure 1. Flow of the proposed research scheme

c. Preprocessing, in KDD data cleaning, data integration, data selection, and data transformation are also known in one unit as data preprocessing (Han,

2012). Data preprocessing is the stage where the process determines the training data and testing data, then distributes the majority and minority classes on unbalanced data classes.

d. Data Mining, a process that aims to find interesting patterns or information in selected data using certain techniques or methods. The technique used in this data mining process is the oversampling and undersampling approaches to balance the unbalanced data classes. Then the classification using the K-Nearest Neighbor (KNN) method so that the prediction results are obtained.

e. Validation, at this stage testing the quality of the knowledge obtained. The results of. data mining are evaluated and the validity of the data is tested with training data. The validity test is carried out by testing the Accuracy level, where the accuracy test will be carried out after the data is balanced with the AUC and Confusion Matrix.

3. RESULTS AND DISCUSSION

3.1 Precosessing Data

In this section, normalization of the class attribute dataset is carried out into 2 factor levels and then the numerical variables tend to have varying ranges. In some data mining algorithms, the difference in the range can have an undue influence on the classification results because the variable with a larger range will dominate. So that min-max normalization is performed on the dataset.

3.2 Resampling

Handling unbalanced class data by doing 3 modeling can be seen below :

a. *Oversampling*

In this study, the researcher tried to balance the data class with the Oversampling approach. The changes that occur can be seen in Table 1.

Table 1. Data is balanced by Oversampling.

Training Data	Information	Class (No)	Class (Yes)
Before	Amount of data 470	450	20
After	Amount of data 470	450	434

b. *Undersampling*

In this study, the researcher tried to balance the class of data with the undersampling approach. The changes that occur can be seen in Table 2.

Table 2. Data is balanced by undersampling.

Training Data	Information	Class (No)	Class (Yes)
Before	Amount of data 470	450	20
After	Amount of data 470	19	20

c. Combination of Oversampling and Undersampling

In this study, the researchers tried to balance the data classes with a combination of Oversampling and Undersampling approaches. The changes that occur can be seen in Table 3.

Table 3. Data is balanced by a combination of Oversampling and Undersampling.

Training Data	Information	Class (No)	Class (Yes)
Before	Amount of data 470	450	20
After	Amount of data 470	253	217

3.3 Classification

a. k-*Nearest Neighbor* (KNN)

Classification in this study uses the value of k = 1-10. Where the optimal k value in the KNN classification for this unbalanced dataset is 10. The accuracy of each k value can be seen in Table 4.

Table 4. Unbalanced training data accuracy table from the value of k = 1 – 10.

k	Accuracy
1	0.9234043
2	0.9319149
3	0.9446809
4	0.9531915
5	0.9531915
6	0.9553191
7	0.9574468
8	0.9574468
9	0.9574468
10	0.9574468

b. *Oversampling*

Classification in this study uses the value of k = 1-10. Where the optimal k value in the KNN classification against the dataset through this Oversampling approach is 1. The accuracy of each k value can be seen in Table 5.

Table 5. Oversampling approach accuracy table from the value of k = 1 – 10.

k	Accuracy
1	0.9094995
2	0.8981869
3	0.8800434
4	0.8653473
5	0.8551966
6	0.8439096
7	0.8246936
8	0.8088355
9	0.7861721
10	0.7782686

c. *Undersampling*

Classification in this study uses the value of k = 1-10. Where the optimal k value in the KNN classification against the dataset through this undersampling approach is 6. The accuracy of each k value can be seen in Table 6.

Table 6. Undersampling approach accuracy table from the value of k = 1 - 10.

k	Accuracy
1	0.3333333
2	0.4166667
3	0.3583333
4	0.4583333
5	0.4250000
6	0.4833333
7	0.4333333
8	0.3750000
9	0.2750000
10	0.3583333

d. Combination of Oversampling and Undersampling

Classification in this study uses the value of k = 1-10. Where the optimal k value in the KNN classification against the dataset through the combination approach of Oversampling and Undersampling is 1. The accuracy of each k value can be seen in Table 7.

Table 7. Table of accuracy of the combination of Oversampling and Undersampling approaches from the value of k = 1 - 10.

k	Accuracy
1	0.9107636
2	0.8616886
3	0.8362010
4	0.8084470
5	0.7870799
6	0.7767133
7	0.7574237
8	0.7383653
9	0.7341100
10	0.7149129

3.4 Validation

Confusion Matrix and AUC performance results after classifying using k-Nearest Neighbor (KNN) can be seen in the table and bar graph below :

a. k-*Nearest Neighbor* (KNN)

Table 8. Confusion Matrix and AUC on KNN.

k	Accuracy	Sensitivity	Specificity	AUC
1	0,9141	0,9750	0,0000	0,5390625

2	0,9062	0,9667	0,0000	0,5286458
3	0,9375	1,0000	0,0000	0,5703125
4	0,9375	1,0000	0,0000	0,6260417
5	0,9375	1,0000	0,0000	0,6151042
6	0,9375	1,0000	0,0000	0,5932292
7	0,9375	1,0000	0,0000	0,5713542
8	0,9375	1,0000	0,0000	0,5604167
9	0,9375	1,0000	0,0000	0,5583333
10	0,9375	1,0000	0,0000	0,5494792

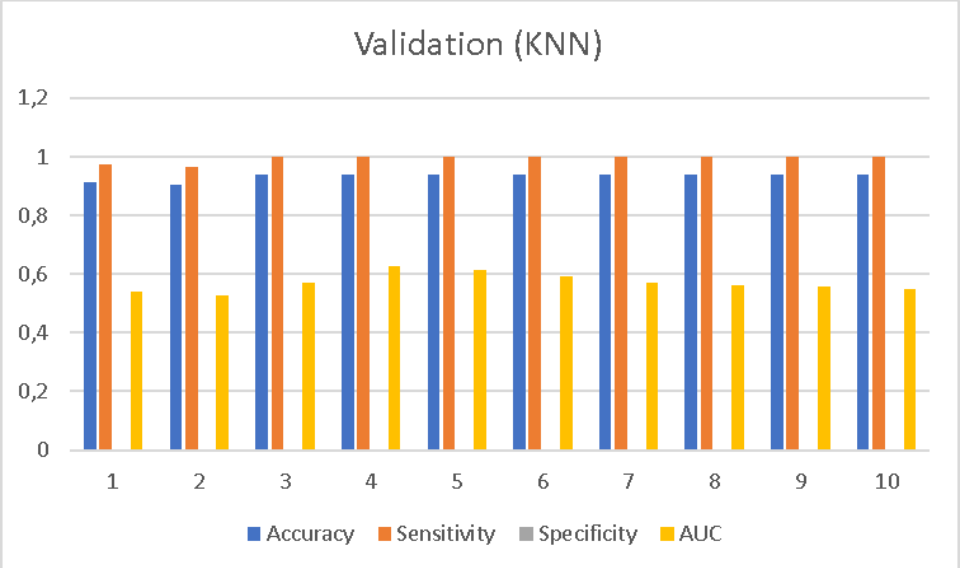

Figure 2. Confusion Matrix and AUC on KNN.

b. *Oversampling*

Table 9. Confusion Matrix and AUC on Oversampling+KNN.

k	Accuracy	Sensitivity	Specificity	AUC
1	0,8750	0,92500	0,94070	0,6078125
2	0,8594	0,90833	0,93966	0,5963542
3	0,8359	0,88333	0,12500	0,5812500
4	0,8359	0,87500	0,25000	0,6380208
5	0,8203	0,85830	0,25000	0,6296875
6	0,7812	0,81667	0,25000	0,6036458
7	0,7422	0,77500	0,25000	0,5765625
8	0,6953	0,72500	0,25000	0,5515625
9	0,6875	0,71667	0,25000	0,5473958
10	0,6562	0,68330	0,25000	0,5307292

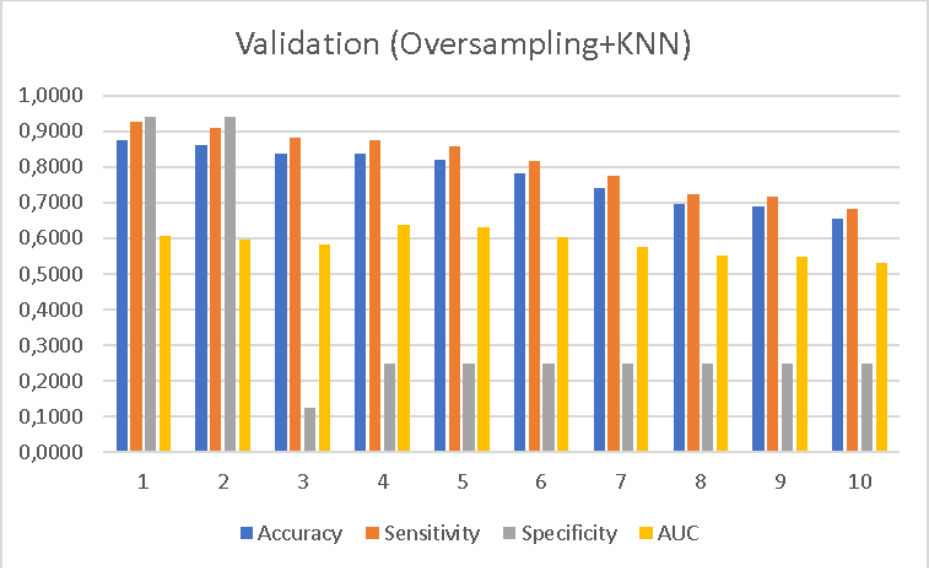

Figure 3. Confusion Matrix and AUC on Oversampling+KNN.

c. *Undersampling*

Table 10. Confusion Matrix and AUC on Undersampling+KNN.

k	Accuracy	Sensitivity	Specificity	AUC
1	0,5156	0,52500	0,37500	0,4656250
2	0,3828	0,38333	0,37500	0,3718750
3	0,3438	0,32500	0,62500	0,4380208
4	0,3047	0,27500	0,75000	0,5229167
5	0,2812	0,25833	0,62500	0,3848958
6	0,3825	0,39167	0,25000	0,3645833
7	0,3047	0,29167	0,50000	0,4307292
8	0,3984	0,39167	0,50000	0,3322917
9	0,3828	0,38333	0,37500	0,3114583
10	0,4297	0,458300	0,00000	0,2479167

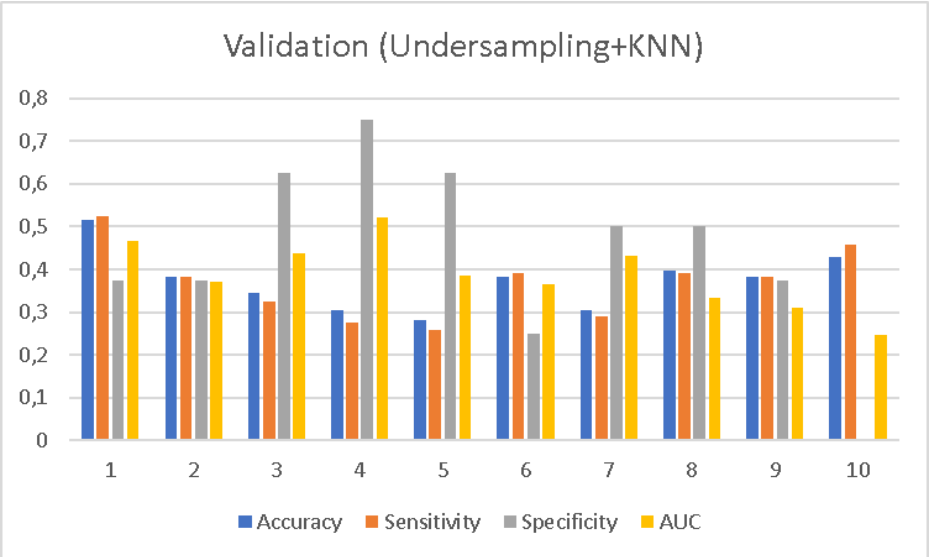

Figure 4. Confusion Matrix and AUC on Undersampling+KNN.

d. Combination of Oversampling and Undersampling

Table 11. Confusion Matrix and AUC on the Combination of Oversampling and Undersampling+KNN.

k	Accuracy	Sensitivity	Specificity	AUC
1	0,8672	0,90000	0,37500	0,6651042
2	0,8516	0,88330	0,37500	0,6463542
3	0,7812	0,80830	0,37500	0,6067708
4	0,7344	0,75833	0,37500	0,5817708
5	0,6875	0,70833	0,37500	0,5515625
6	0,6641	0,68333	0,37500	0,5359375
7	0,5781	0,59167	0,37500	0,4932292
8	0,5625	0,57500	0,37500	0,4848958
9	0,5312	0,54167	0,37500	0,4682292
10	0,5078	0,51667	0,37500	0,4255208

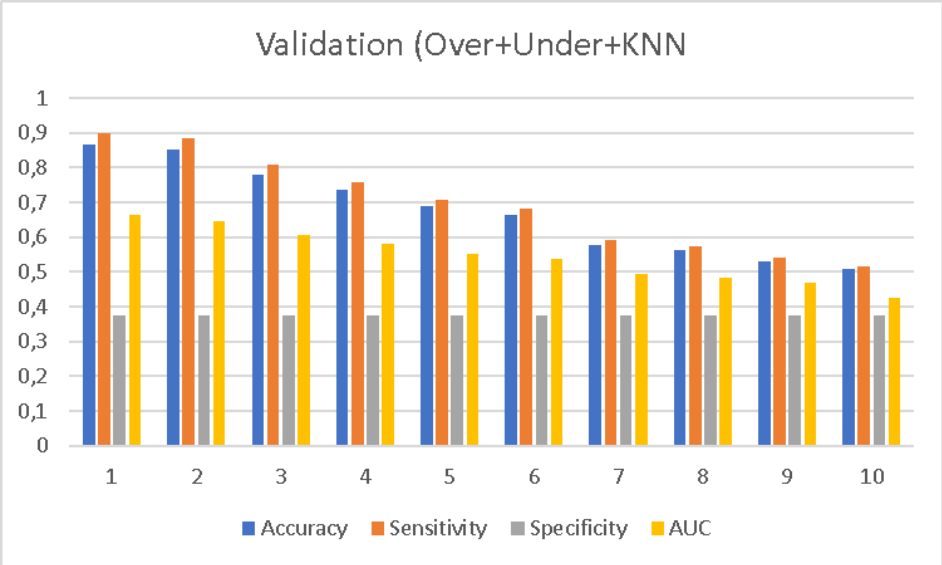

Figure 5. Confusion Matrix and AUC in Over+Under+KNN combination.

3.5 Discussion

This study aims to determine the performance of the k-Nearest Neighbor (KNN) classification on data that has an unbalanced class, and to find out the number of k closest neighbors that can produce the best work. Before entering the data mining stage. the dataset obtained first through the preprocessing stage. The dataset obtained has a numeric class which will be used as a classification result, then a data transformation will be carried out where class "0" is changed to "No" and class "1" is changed to "Yes". Because the distance calculation will be carried out, the predicted attributes must be normalized first using Min-Max normalization (data transformation). Furthermore, the distribution of training and testing data is carried out, where the training data is 80% and the testing data is 20%.

The researcher then balances the data on each training data before classifying it using KNN. The methods used are Oversampling, Undersampling and Combination of Oversampling + Undersampling. For the results of the KNN classification on unbalanced data having high accuracy results, it can be seen in Table 12. Undersampling and Combination of Oversampling + Undersampling. In Figure 6 it

can be seen that Oversampling + KNN has a fairly stable value of accuracy, sensitivity, specificity and AUC compared to the classification without a balanced dataset. Then the method used by the researcher is Undersampling + KNN which has stable accuracy, sensitivity, specificity and AUC values, but has the lowest value compared to other methods, can be seen in Figure 6. After that, the researchers used a combination method of Oversampling + Undersampling + KNN which has stable accuracy, sensitivity, specificity and AUC values and has superior AUC values compared to other methods.

Table 12. Comparison of the classification performance of each method with optimal k.

Method	Accuracy	Sensitivity	Specificity	AUC
KNN	0,9375	1,0000	0,0000	0,5494792
Over+KNN	0,875	0,9250	0,1250	0,6078125
Under+KNN	0,3438	0,33333	0,50000	0,3645833
Over+Under+KNN	0,8672	0,9000	0,3750	0,6651042

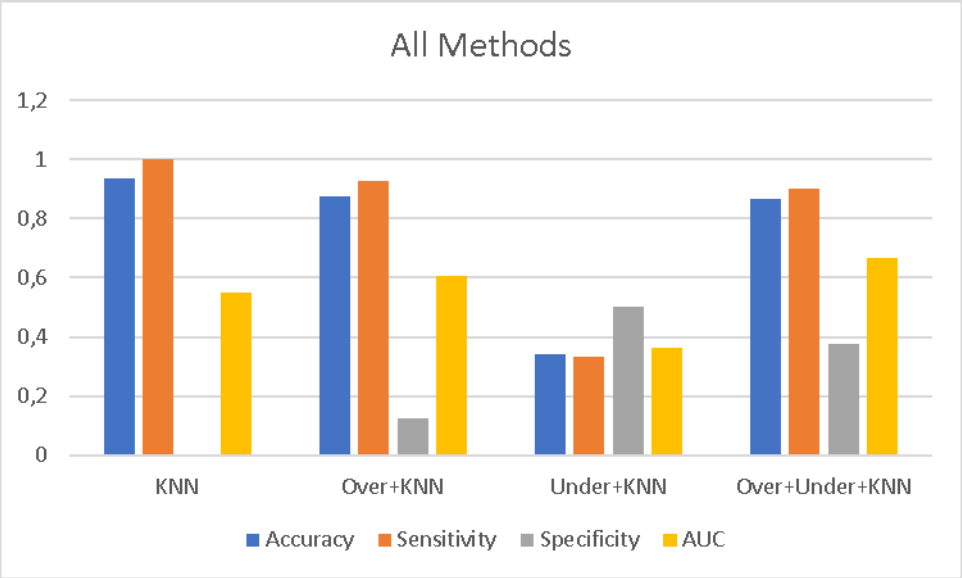

Figure 6. Performance results of all classifications.

4. CONCLUSION

Based on the results of research and discussions that have been carried out, it can be concluded that:

a. The combination of Oversampling and Undersampling is the best classification performance to handle unbalanced data classification solutions compared to no combination.

b. The classification performance using k-Nearest Neighbor in the Undersampling approach has the lowest performance results, namely the accuracy of 0.3438; sensitivity of 0.33333; specificity of 0.50000; and AUC of 0.3645833, Oversampling has performance results, namely accuracy of 0.8750; sensitivity of 0.9250; specificity of 0.1250; and AUC of 0.6078125; and the combination of Oversampling and Undersampling has performance

results, namely accuracy of 0.8672; sensitivity of 0.9000; specificity of 0.3750; and AUC of 0.6651042.

REFERENCES

[1] Abdullah, Dahlan, dkk. 2015. **Implementasi Metode Rough Set Untuk Menentukan Data Nasabah Potensial Mendapatkan Pinjaman**. ISSN : 2460 – 4690. Vol. 1. Universitas Putra Indonesia YPTK Padang.

[2] Fauzi Abdul Aziz. 2016. **Penerapan K-Means Pada Imbalanced Data Untuk Klasifikasi Metagenom. Fakultas Matematika dan Ilmu Pengetahuan Alam.** Institut Pertanian Bogor. Bogor.

[3] Han, Jiawei & Kamber, Micheline. 2006. **Data Mining: Concept and Techniques Second Edition.** Morgan Kaufmann Publishers.

[4] Turban Efraim, Jay E. Aronson, Ting-Peng Liang. 2005. **Decision Support Systems and Intelligent Systems**. Edisi 7 Jilid 1, Penerbit Andi.