IT Journal Research and Development (ITJRD)
Vol.8, No.1, August 2023, E-ISSN : 2528-4053 | P-ISSN : 2528-4061
DOI : 10.25299/itjrd.2023.12412

Evaluate of Random Undersampling Method and Majority Weighted Minority Oversampling Technique in Resolve Imbalanced Dataset

Muhammad Asyroful Nur Maulana Yusuf[1], Meida Cahyo Untoro[2*]
Terknik Informatika, Institut Teknologi Sumatera, Lampung, Indonesia
muhammad.119140026@student.itera.ac.id[1], cahyo.untoro@if.itera.ac.id[2]

Article Info

Article history:

Received Mar 08, 2023
Revised May 22, 2023
Accepted Aug 18, 2023

Keyword:

Imbalanced Data
Random Undersampling
MWMOTE
Confusion Matrix

ABSTRACT

Classification is a model for making predictions based on existing data. Unbalanced data leads to misclassification or modeling errors where the data is irrelevant and results in poor classification modeling. The poor classification model is caused by an imbalance in the data on the classification label, so it is necessary to balance the data as a solution to overcome this problem. The methods used to deal with data imbalance are Random Undersampling and MWMOTE. The aim is to see the implementation of Random Undersampling and MWMOTE work well in dealing with unbalanced datasets and to know the performance and accuracy in modeling. The dataset used is an open source dataset from Kaggle which consists of Diabetes data, Bank Turnover data, Stroke data, and Credit Card data with various data ratios, with the aim of overcoming the problem of data imbalance. Model evaluation was carried out using confusion matrix and decision tree algorithms by looking at the values of precision, recall, f-measure, and accuracy of the original data, the Random Undersampling method, and MWMOTE. In the original data with 48.86% precision, 54.90% recall, 51.73% f-measure, and 85.30% accuracy. Random Undersampling can overcome data imbalance problems with 76.28% precision, 76.74% recall, 76.48% f-measure, and 76.21% accuracy. MWMOTE can solve data imbalance problems with 86.04% precision, 87.30% recall, 86.66% f-measure, and 86.61% accuracy. It can be concluded that the MWMOTE method is better than the Random Undersampling method because the evaluation average of the Confusion Matrix Random Undersampling method is smaller than the MWMOTE method.

Corresponding Author:

Muhammad Asyroful Nur Maulana Yusuf and Meida Cahyo Untoro
Teknik Informatika
Institut Teknologi Sumatera
Jati Agung, Lampung Selatan, Indonesia
Email: muhammad.119140026@student.itera.ac.id and cahyo.untoro@if.itera.ac.id

1. INTRODUCTION

In machine learning, classification is a model for making predictions based on existing data. The classification performed in making predictions is designed with a large and balanced data set in order to maximize classification accuracy. Data processing with a large data set and with various classes often results in an imbalanced class [1]. Unequal distribution of samples in different

classes is a problem in class imbalance data, where most of the samples have several classes and the rest belong to other classes. If the sample only has two classes, then the class that has the most of the sample is called the majority class and the others are the minority class [2]. Imbalanced data or Imbalanced dataset is a condition in which the classification label on a dataset experiences a large discrepancy between the majority class data and minority class data.

Classification of data with unbalanced classes results in lower classification accuracy [3]. Classification with unbalanced data will tend to the majority class data compared to the minority class data. Data imbalance results in misclassification or modeling errors where data is irrelevant and results in poor classification modeling [4]. The imbalance of data greatly affects several classifications such as credit data [5], stroke data , [6]online [7]news data , biomedical data [8], diarrhea case data for toddlers [9], poor household classification data [10], and other data.

There needs to be special handling of imbalanced datasets prior to data analysis. A bad classification model is caused by an imbalance in the data on the classification label, it is necessary to balance the data as a solution to solving this problem. The undersampling and oversampling methods are solutions for data imbalance [11].

Undersampling method is a method which reduces the majority class data and stores all minority class data [12]. One method that is often used to overcome imbalanced datasets by reducing the majority class is the random undersampling method [13]. This method is used by randomly selecting and deleting majority class data until the number of majority class data and minority data is the same. The advantage of this method is to do it randomly without certain conditions in eliminating or reducing the value of the majority class. The weakness of the undersampling method can eliminate important parts of the majority class data so that it affects classification performance [14].

Oversampling method is a method by making data replication on minority class data until the number of minority class data and majority data is the same [15]. The weakness of the oversampling method is that there is overfitting because the synthesis data is too racing on the training data so that it cannot make accurate predictions [16]. Overfitting can be overcome by using the MWMOTE method. The stages of creating synthetic data in MWMOTE consist of three stages, namely identifying the majority class and minority class, measuring the minority class, and grouping data using the clustering method . The creation of minority class data replication depends on synthetic data. Misclassification becomes a problem in creating synthetic data when unbalanced data causes certain classes [18].

Undersampling and oversampling methods in dealing with imbalanced datasets have advantages and disadvantages between the two. Therefore, this research wants to compare and analyze two imbalanced dataset methods, namely random undersampling and MWMOTE by looking at the accuracy, precision, recall, and f-measure values . The purpose of this research is to find the best method for handling imbalanced datasets between random undersampling and MWMOTE methods.

2. RESEARCH METHOD

2.1. Data Mining

Data mining is often referred to as knowledge discovery in database is a process of taking , processing , and data analysis with purpose make it information for retrieval decisions in the form of classification, clustering, association, and prediction [21]. There is various stages in carrying out the process of data mining namely data selection, data cleaning, data transformation, data integration, pattern evaluation, and knowledge.

Classification in the data can be measurement, categorical, signal, or image based on the class on the data label. Machine learning is one method of analysis in data mining. Machine learning programs learn patterns in some interrelated data linkage . So, simply machine learning is learning machine to achieve appropriate and desired results based on a set of existing data [22]. Analysis results from data mining processing can be in the form of description, association, clustering, prediction and classification [10].

2.2. Imbalanced

Imbalanced data is data that has ratio differs from first class data to other class data. Where by general Imbalanced dataset is the ratio of the number of unbalanced majority data with minority data. Application machine learning in analyzing or Data processing is often a problem in imbalanced datasets [23].

Unbalanced data can result classification error which has an impact on the accuracy value decreases as well as allows the minority class considered as outliers [24]. Methods of classification and application agortima being one way to deal with unbalanced data . Reducing the amount of majority data using the undersampling method and using synthetic data in adding minority classes using the oversampling method can be a solution to overcoming unbalanced data.

2.3. Random Undersampling

Undersampling is the simplest method to deal with unbalanced data. Random undersampling calculate the difference between majority and minority class data so that later the majority data class is selected and deleted in a manner random to the sum of the minority class equal to the number of the majority class [5]. In figure 1 the minority class marked with a dotcolored orange. Point colored blue represents the majority class data, period blue is the blocked sample in a manner random until the number of majority data classes and minority data balanced. Deleting data will reduce storage and upgrade processing time. However, by deleting in a manner random majority class data then it can cause loss important and influential information accuracy results in classification [13]. Here's a representation graphic undersampling using random undersampling method in the image below following:

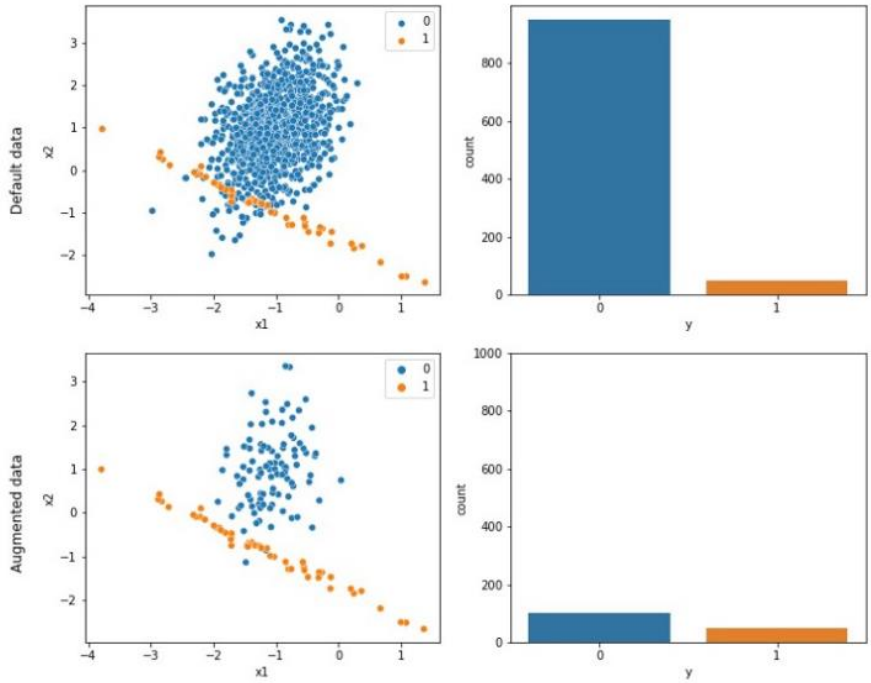

Figure 1. Representation Graphical Random Undersampling

2.4. MWMOTE

MWMOTE or Majority Weighted Minority Oversampling Technique is a method of creating synthetic data based on the minority class on a data label [25]. MWMOTE is a repair method from SMOTE via synthetic data generation. Stages synthetic data creation on MWMOTE consists from three stages that is identify the majority class and the minority class, measure the minority class, and group the data using the clustering method [17]. Stages MWMOTE synthetic

data creation starts with a selection sample from the majority class and minority class, then carry out the measurement process or weighting on the minority class in order to know the position of the minority class adjacent to the borderline, then every minority data given weight as needed as well as data interests, and the last one is to carry out the clustering process aim to _ results from synthetic data are in a group or cluster [18]. Illustration MWMOTE stages in the image below following:

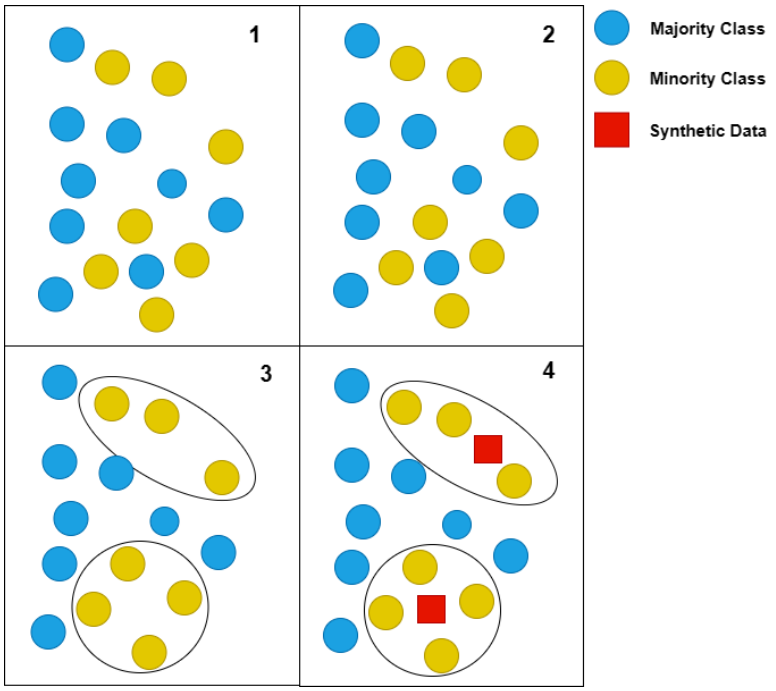

Figure 2. MWMOTE

Making synthetic data using the MWMOTE method in dealing with imbalanced datasets consists of 4 phases, namely:

Phase 1: Identifying minority class samples.

1. Separate the minority data sample (Smin) and the majority data sample (Smaj).
2. Create a set of Sminf by finding the nearest neighbor node in every xi in Smin. Search for nodes using KNN in predicting noise in the minority class and looking for the minority class in the majority class.

$$S_{minf} = S_{min} - \{xi \in S_{min}: N\ N\ (xi\ )\} \quad (1)$$

3. Determine the boundary line for the majority class which will later be useful in sharing information about minority data.

$$Sbmaj = \cup xi \in Sminf\ Nmaj(xi\ ) \quad (2)$$

4. Form an informative minority class data sample.

Phase 2: Weighting of minority class data

1. Perform weight calculations on the Simin sample (Iw), each xi that is in Simin will be given a sample weight (Sw)

$$Sw(xi\ ) = \sum Iw(yixi\ ) \quad (3)$$

2. Convert each Sw(xi) into a probability sample (Sp)

$$Sp\ (xi\ ) = Sw(xi\ ) / \sum Sw(Zi) \quad (4)$$

Phase 3: Making synthetic data on the minority class using the clustering method.

Phase 4: Balanced data with the addition of synthetic data from the minority class.

2.5. Classification

Classification as a performance evaluator from method comparison is a process carried out to identify and compare the effectiveness of various methods used in measuring or evaluating the performance of a particular system or process. Classification refers to the division or grouping of these methods based on certain characteristics or attributes. By classifying, we can organize these methods into relevant groups, making it easier to analyze and evaluate performance[12].

As a "performance evaluator", the role of this classifier is to assess and compare these methods in terms of their ability to measure or evaluate the performance of the system or process being observed. In this case, these methods can be algorithms, statistical models, or other approaches used to collect data, make measurements, and analyze performance results. In making comparisons, the classification takes into account various factors, such as accuracy, precision, recall, computation time, efficiency, reliability, and practicality of the method[14]. Taking these factors into account, the classification will assist in determining which method is most suitable and can provide the most accurate and reliable performance evaluation results. In addition, classification as a performance evaluator can also help in understanding the strengths and weaknesses of each method, as well as gaining better insight into situations where certain methods are more effective than others. As such, classifications can provide valuable guidance in the selection and use of the optimal performance evaluation method for specific needs.

2.6. Confusion Matrix

Confussion matrix is a matrix used in measurement performance from machine learning classification. There are 4 terms in the confusion matrix, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [18]. The confusion matrix table can be seen in the table below following :

Table 1. Confusion Matrix

	Predicted Positive	***Predicted Negatives***
Actual Positive	TP	FN
Actual Negatives	FP	TN

Result of grouping based on the number of positive data or negative data then do a comparison between mark actually with the predicted value based on the evaluation matrix. The evaluation matrix consists from accuracy, precision, recall, and f-measure [26]values. This matrix provides helpful information about model performance how well the model classifies the correct data.

Accuracy measures how many correct predictions from the overall prediction of the model. Accuracy is calculated by dividing the number of correct predictions by the total number of predictions. However, accuracy is not the best evaluation matrix when the data is unbalanced, therefore it is necessary to evaluate precision, recall and f-measure to be more informative. The accuracy value is obtained by the following formula:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{5}$$

Precision measures how many positive predictions are true of all positive predictions. Precision is very important in classifying data because where positive prediction error own bad consequences. The precision value is obtained by the following formula:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{6}$$

Recall measures how much many positive predictions are correct from the total number of true positive classes. Recall is very important in classifying data because where positive prediction error own bad consequences. The recall value is obtained by the following formula:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{7}$$

F-Measure is the average of precision and recall, where is the range from F-Measure is 0 to 1. The f-measure value is obtained by the following formula:

$$\boldsymbol{F\text{-}Measure} = \mathbf{2}.\frac{\boldsymbol{precision}\,.\,\boldsymbol{recall}}{\boldsymbol{precision+recall}} \tag{8}$$

2.7. Research Workflow

In this study, the flow used in the final evaluation task research was random undersampling and Majority Weighted Minority Oversampling Technique in overcoming Imbalanced dataset can be seen in Figure 3:

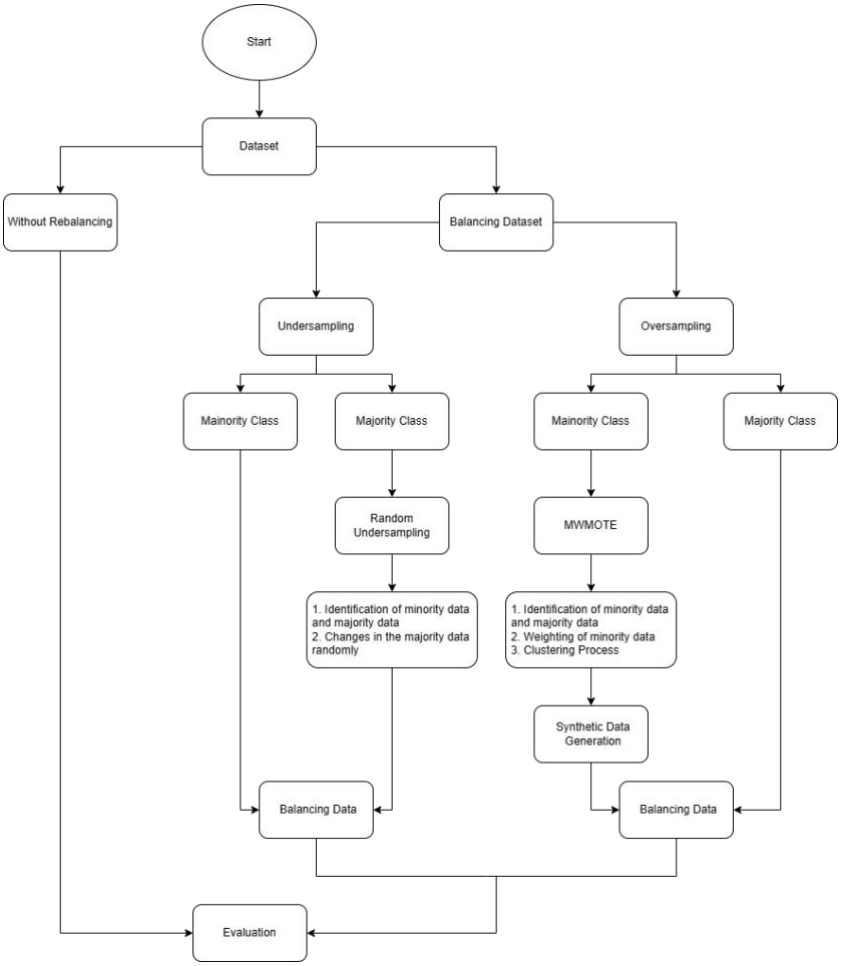

Figure 3. Modeling Design

At this stage is a modeling design that will be used to test the model on several unbalanced datasets. In this modeling design is divided into several processes, which can be seen in Figure 3 The modeling design process as following. After the four datasets perform the data preprocessing stages The next step is to check Imbalanced dataset based on the results of the classification of the

Bank Turnover, Diabetes, Stroke, and Credit Fraud datasets as shown in the table and plotting below following:

Table 2. Description of Datasets

Datasets	Amount of data	Attribute	Minority Class	Majority Class	Ratio
Diabetes	768	9	268	500	34% : 66%
Bank Turnovers	10,000	13	2037	7,963	20% : 80%
Strokes	5.110	19	249	4,861	4% : 96%
Credit card	284,807	30	492	284.315	1% : 99%

Next, check the distribution of the data by using scatter plotting from the 4 datasets used. The Diabetes dataset is checked for data distribution by looking at the correlation of values between the Age and BMI attributes, which can be seen in Figure 4. The Bank Turnover dataset is checked for data distribution by looking at the correlation between the Age and CreditScore attributes, which can be seen in Figure 5. The Stroke dataset is distributed by looking at the correlation between the avg_glucose_level and BMI attributes, which can be seen in Figure 6. The Credit Card dataset is distributed by looking at the correlation of values between attributes V1 and V28 which can be seen in Figure 7.

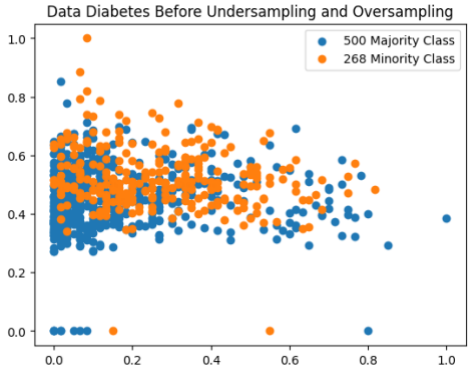

Figure 4. Distribution of Diabetes Data

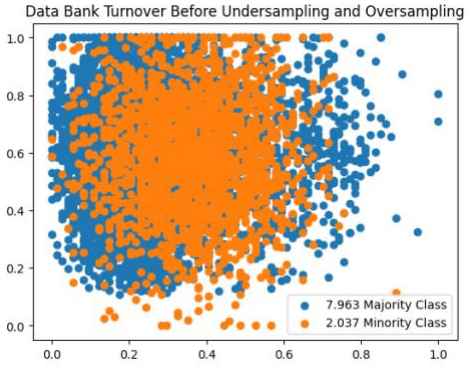

Figure 5. Distribution of Bank Turnover Data

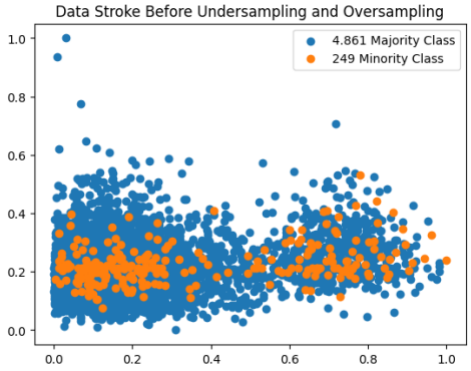

Figure 6. Distribution of Stroke Data

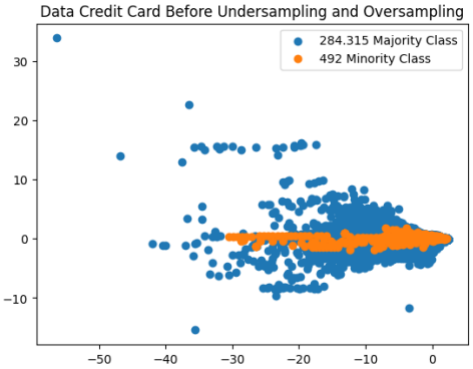

Figure 7. Distribution of Credit Card Data

3. RESULTS AND ANALYSIS

3.1. Random Undersampling

The data used to train the Random Undersampling model uses data after preprocessing where later the amount of data from the majority class will be the same as the amount of data from the minority class. Furthermore, balancing the dataset using the Random Undersampling method where by changing the amount of data from the majority class will be equal to the amount of data

from the minority class. There are 2 stages in balancing data with the Random Undersampling method, which is to separate the minority and majority class data, then resample the majority data class randomly as much as the minority class data. The results of balancing data using the Random Undersampling method can be seen in Table 3:

Table 3. 1Distribution of Balanced Data Random Undersampling Method

Datasets	Amount of data	Attribute	Minority Class	Majority Class	Ratio
Diabetes	536	9	268	268	50% : 50%
Bank Turnovers	4,074	13	2037	2037	50% : 50%
Strokes	498	19	249	249	50% : 50%
Credit card	984	30	492	492	50% : 50%

Next, check the distribution of the data by using scatter plotting from the 4 datasets used. The Diabetes dataset is checked for data distribution by looking at the correlation of values between the Age and BMI attributes, which can be seen in Figure 8. The Bank Turnover dataset is checked for data distribution by looking at the correlation between the Age and CreditScore attributes, which can be seen in Figure 9. The Stroke dataset is distributed by looking at the correlation between the avg_glucose_level and BMI attributes, which can be seen in Figure 10. The Credit Card dataset is distributed by looking at the correlation of values between attributes V1 and V28 which can be seen in Figure 11.

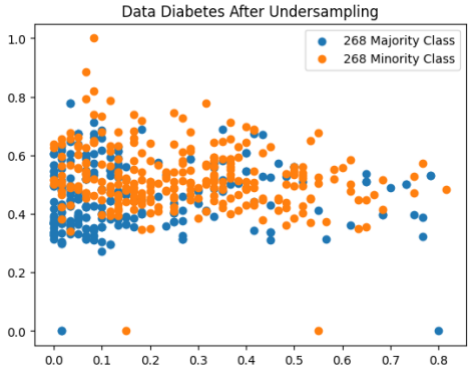

Figure 8. Distribution of Diabetes Data

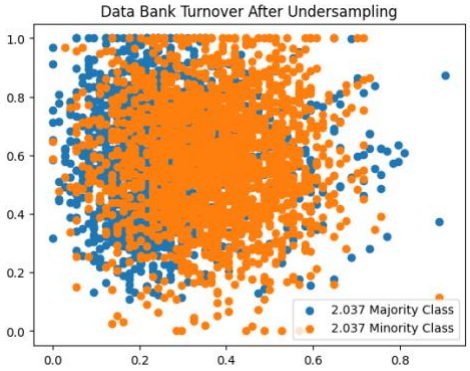

Figure 9. Distribution of Bank Turnover Data

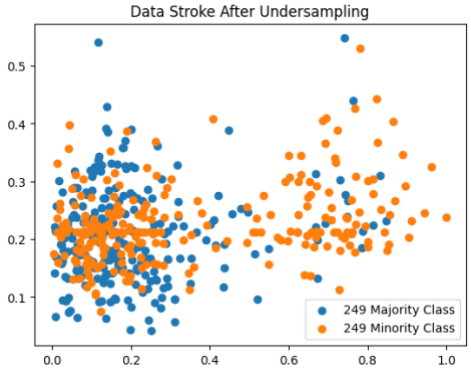

Figure 10. Distribution of Stroke Data

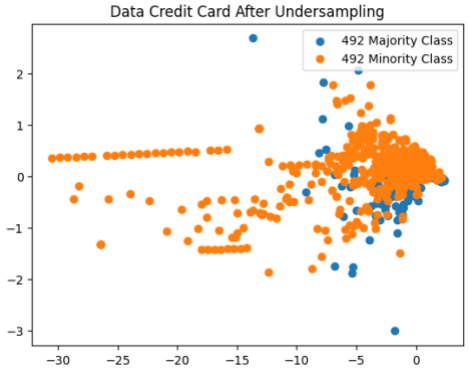

Figure 11. Distribution of Credit Card Data

3.2. MWMOTE Oversampling

The data used to train the MWMOTE Oversampling modeling uses data after preprocessing where the amount of minority class data will be the same as the amount of majority class data. Furthermore, balancing the dataset using the MWMOTE method where by changing the

amount of minority class data will be the same as the amount of majority class data with the growth of synthetic data. There are 3 stages in balancing data with the MWMOTE method, namely by separating minority data into the majority data class, doing weighting, and doing clustering in making synthetic data. The results of balancing data using the Random Undersampling method can be seen in Table 4.

Table 4.2Balanced Data Distribution of the MWMOTE Method

Datasets	Amount of data	Attribute	Minority Class	Majority Class	Ratio
Diabetes	1,000	9	500	500	50% : 50%
Bank Turnovers	15,926	13	7,963	7,963	50% : 50%
Strokes	9,722	19	4,861	4,861	50% : 50%
Credit card	568,630	30	284,315	284,315	50% : 50%

Next, check the distribution of the data by using scatter plotting from the 4 datasets used. The Diabetes dataset is checked for data distribution by looking at the correlation of values between the Age and BMI attributes, which can be seen in Figure 12. The Bank Turnover dataset is checked for data distribution by looking at the correlation between the Age and CreditScore attributes, which can be seen in Figure 13. The Stroke dataset is distributed by looking at the correlation between the avg_glucose_level and BMI attributes, which can be seen in Figure 14. The Credit Card dataset is distributed by looking at the correlation of values between attributes V1 and V28 which can be seen in Figure 15.

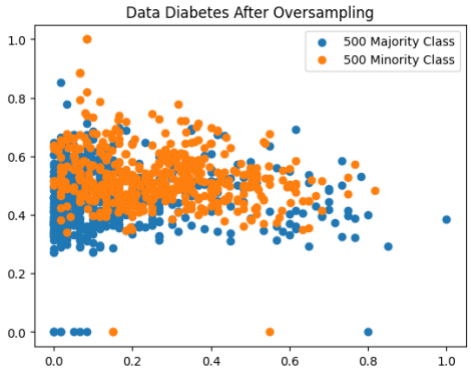

Figure 12. (a)Distribution of Diabetes Data

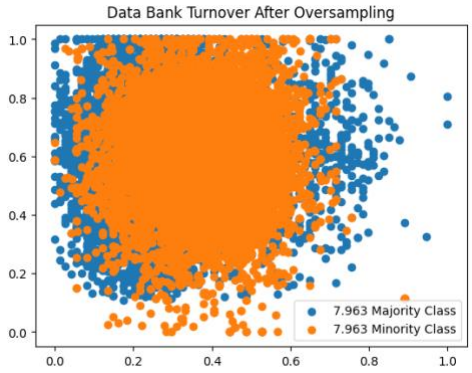

Figure 13. (b)Distribution of Bank Turnover Data

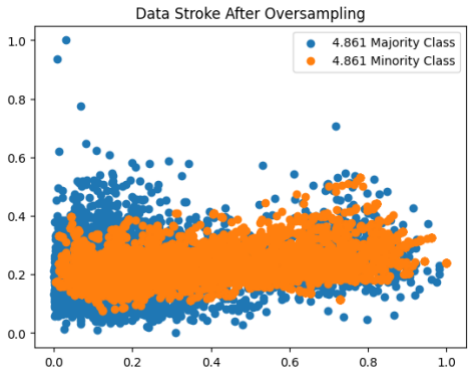

Figure 14. (c)Distribution of Stroke Data

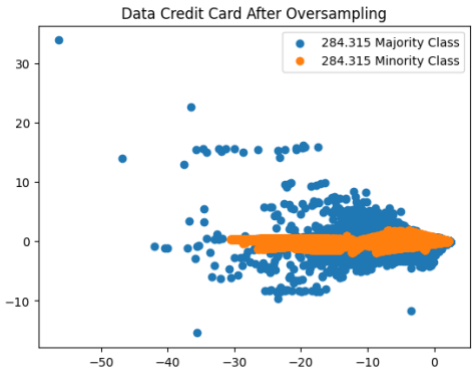

Figure 15. (d)Distribution of Credit Card Data

3.3. Model Evaluation

The following is a comparison table of the evaluation results of the method in balancing the dataset.

Table 5. Comparison of Confusion Matrix Evaluation No Rebalancing and Random Undersampling

Datasets	*No Rebalancing*				*Random Undersampling*			
	Precision	*Recall*	*F-Measure*	*Accuracy*	*Precision*	*Recall*	*F-Measure*	*Accuracy*
Diabetes	56,25%	65,00%	60,46%	70,56%	70,37%	75,00%	72,61%	73,29%
Bank Turnover	49,75%	52,39%	51,04%	80,43%	74,83%	72,03%	73,40%	72,69%
Stroke	19,38%	21,34%	20,32%	90,28%	69,73%	67,94%	68,83%	68,00%
Credit Fraud	70,06%	80,88%	75,08%	99,91%	90,19%	92,00%	91,08%	90,87%
Avg	48,86%	54,90%	51,73%	85,30%	76,28%	76,74%	76,48%	76,21%

Table 6. Comparison of Confusion Matrix Evaluation No Rebalancing and MWMOTE

Datasets	*No Rebalancing*				MWMOTE			
	Precision	*Recall*	*F-Measure*	*Accuracy*	*Precision*	*Recall*	*F-Measure*	*Accuracy*
Diabetes	56,25%	65,00%	60,46%	70,56%	73,20%	74,66%	73,92%	73,66%
Bank Turnover	49,75%	52,39%	51,04%	80,43%	83,86%	83,61%	83,73%	83,82%
Stroke	19,38%	21,34%	20,32%	90,28%	87,65%	91,33%	89,45%	89,44%
Credit Fraud	70,06%	80,88%	75,08%	99,91%	99,44%	99,59%	99,52%	99,52%
Avg	48,86%	54,90%	51,73%	85,30%	86,04%	87,30%	86,66%	86,61%

Imbalanced data consists of majority class and minority class. Using the Random Undersampling method, the process is carried out by changing the amount of the majority data to be equal to the number of minority data randomly. Using the MWMOTE method is done by changing the amount of minority data to be equal to the amount of majority data by adding synthetic data. Making synthetic data using the MWMOTE method goes through 3 stages, namely by separating minority data into the majority data class, doing weighting, and doing clustering in making synthetic data. In Table 5 and Table 6 the evaluation results use the decision tree algorithm by testing 3 types of data, namely without rebalancing, Random Undersampling, and MWMOTE with the aim of evaluating and comparing the performance of the imbalanced dataset method. Random Undersampling can overcome the problem of unbalanced data with a precision value of 76.28%, 76.74% recall, 76.48% f-measure, and 76.21% accuracy. MWMOTE can overcome the problem of unbalanced data with a precision value of 86.04%, 87.30% recall, 86.66% f-measure, and 86.61% accuracy.

In the Diabetes dataset, the results of the model evaluation test using the Random Undersampling and MWMOTE methods show that the values for precision, recall, f-measure, and accuracy are relatively the same, but different from the other 3 datasets which produce relatively different values. This shows that the ratio in a dataset has an influence on the process of balancing the dataset. The diabetes dataset has a ratio of minority data and majority data of 34%: 66% where the dataset has an almost balanced ratio of differences between minority data and majority data. The significant difference to the other 3 datasets regarding the ratio of minority data and majority data causes the balancing of datasets using the Random Undersampling and MWMOTE methods to have significantly different results.

For the precision value of 4 datasets used in the research, it has increased when you have done data balancing. Such a significant change occurred in the Storke dataset with a change value of 68.27% from the original data to the MWMOTE oversampling data. The change in the optimal precision value in the Diabetes dataset is 16.95% from the original data to the MWMOTE oversampling data. The change in the optimal precision value in the Bank Turnover dataset is 34.11% from the original data to the MWMOTE oversampling data. The change in the optimal precision value in the Credit Card dataset is 29.38% from the original data to the MWMOTE oversampling data.

For the recall value of the 4 datasets used in the research, it has increased when balancing the data. Such a significant change occurred in the Storke dataset with a change value of 69.99% from the original data to the MWMOTE oversampling data. The change in the optimal recall value in the Diabetes dataset is 10% from the original data to Random Undersampling data. The change in the optimal recall value in the Bank Turnover dataset is 31.22% from the original data to the MWMOTE oversampling data . Changes in the optimal recall value on the Credit Card dataset of 18.71% from the original data to the MWMOTE oversampling data.

For the value of f-measure 4, the dataset used in the research has increased when balancing the data. Such a significant change occurred in the Storke dataset with a change value of 69.13% from the original data to the MWMOTE oversampling data. The change in the optimal f-measure value in the Diabetes dataset is 13.46% from the original data to the MWMOTE oversampling data. The change in the optimal f-measure value in the Bank Turnover dataset is 32.69% from the original data to the MWMOTE oversampling data. The change in the optimal f-measure value in the Credit Card dataset is 24.44% from the original data to the MWMOTE oversampling data.

Accuracy value of 2 datasets used in the research has increased and the other 2 datasets have decreased when balancing the data. Such a significant change occurred in the Stroke dataset by experiencing a change in value of -22.28% from the original data to Random Undersampling data. Changes in the accuracy value on the Credit Card dataset experience a change value of -9.04% from the original data to Random Undersampling data. The change in the optimal accuracy value for the Diabetes dataset is 3.10% from the original data to the MWMOTE oversampling data. The change in the optimal accuracy value in the Bank Turnover dataset is 3.39% from the original data to the MWMOTE oversampling data.

4. CONCLUSION

Based on the results of the research and discussion in the previous chapter, the following conclusions can be drawn:

1. Random Undersampling Method and the Majority Weighted Minority Oversampling Technique (MWMOTE) can overcome imbalanced dataset problems. Where is the Random Undersampling method changing the number of majority data will be equal to the number of minority class data. There are 2 stages in balancing data with the Random Undersampling method, which is to separate the minority and majority class data, then resample the majority data class randomly as much as the minority class data. The MWMOTE method where by changing the number of minority class data will be equal to the amount of majority class data with the growth of synthetic data. There are 3 stages in balancing data with the MWMOTE method, namely by separating minority data into the majority data class, doing weighting, and doing clustering in making synthetic data.

2. The evaluation results of the confusion matrix using the decision tree algorithm by looking for precision, recall, f-measure, and accuracy values for the Random Undersampling and MWMOTE methods have increased over the stages without dataset rebalancing . The small amount of data and attributes does not really affect the process of balancing the dataset because the results of testing the two methods have significant average values of precision, recall, f-measure, and accuracy . However, differences in data ratios affect the performance results of the Random Undersampling and MWMOTE methods where the datasets with minority data ratios and majority data are not as significant as in the diabetes dataset, the performance results or model evaluation of the Random Undersampling method are relatively the same as the MWMOTE method. The results are different when the dataset with the ratio of minority data and majority data is so significant that the evaluation results or performance of the MWMOTE method are better than the Random Undersampling method.

3. In the original data with 48.86% precision, 54.90% recall, 51.73% f-measure, and 85.30% accuracy. Random Undersampling can overcome data imbalance problems with 76.28% precision, 76.74% recall, 76.48% f-measure, and 76.21% accuracy. MWMOTE can solve data imbalance problems with 86.04% precision, 87.30% recall, 86.66% f-measure, and 86.61% accuracy. It can be concluded that the MWMOTE method is better than the Random Undersampling method because the evaluation average of the Confusion Matrix Random Undersampling method is smaller than the MWMOTE method.

ACKNOWLEDGEMENTS

The author expresses his deepest gratitude to Teknik Infromatika, Institut Teknologi Sumatera who has guided and provided input on this research. Hopefully this will be a spirit in moving forward to develop research that is much better and can be useful for others.

REFERENCES

[1] P. Agustia Rahayuningsih and P. Studi Sistem Informasi Akuntansi Kampus Kota Pontianak, "Penerapan Teknik Sampling Untuk Mengatasi Imbalance Class Pada Klasifikasi Online Shoppers Intention," *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 4, no. 1, 2020.

[2] Hudori, "Resampling Neural Network untuk Penanganan Class Imbalance pada Prediksi Klaim Asuransi A. PENDAHULUAN," vol. 10, no. 1, pp. 57–64, 2020, doi: 10.36350/jbs.v10i1.

[3] D. Chen, X. J. Wang, C. Zhou, and B. Wang, "The Distance-Based Balancing Ensemble Method for Data With a High Imbalance Ratio," *IEEE Access*, vol. 7, pp. 68940–68956, 2019, doi: 10.1109/ACCESS.2019.2917920.

[4] I. Pratama, A. Y. Chandra, and P. T. Presetyaningrum, "Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN," *Jurnal Eksplora Informatika*, vol. 11, no. 1, pp. 38–49, Jan. 2022, doi: 10.30864/eksplora.v11i1.578.

[5] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit," *JURNAL INFORMATIKA*, vol. 5, no. 2, 2018.

[6] S. Mutmainah, "Penanganan Imbalance Data Pada Klasifikasi Kemungkinan Penyakit Stroke," 2021. [Online]. Available: https://library.uii.ac.id/osr

[7] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, A. Nikmatul Kasanah, U. Pujianto, T. Elektro, F. Teknik, and U. Negeri Malang, "Terakreditasi SINTA Peringkat 2 Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *masa berlaku mulai*, vol. 1, no. 3, pp. 196–201, 2017.

[8] H. Ali, N. A. Samat, and H. M. Ashgher, "Adaptive Semi-Unsupervised Weighted Oversampling with Sparsity Factor for Imbalanced Biomedical Data," *Journal of Soft Computing and Data Mining*, vol. 01, no. 01, Mar. 2020, doi: 10.30880/jscdm.2020.01.01.003.

[9] P. Statistika STIS, P. M. Statistika STIS Alfa Rizki, and P. Statistika STIS Rani Nooraeni, "Penerapan Metode Resampling dalam Mengatasi Imbalanced Data Pada Determinan Kasus Diare Pada Balita di Indonesia Andriansyah Muqiit WS Intan Putri Ananda Zahrotin Dwi Hapsari."

[10] T. Purwa, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017)," *Jurnal Matematika, Statistika dan Komputasi*, vol. 16, no. 1, p. 58, Jun. 2019, doi: 10.20956/jmsk.v16i1.6494.

[11] S. Bagui and K. Li, "Resampling imbalanced data for network intrusion detection datasets," *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-020-00390-x.

[12] M. Bach, A. Werner, and M. Palt, "The proposal of undersampling method for learning from imbalanced datasets," in *Procedia Computer Science*, 2019, vol. 159, pp. 125–134. doi: 10.1016/j.procs.2019.09.167.
[13] S. Mishra, "Handling Imbalanced Data: SMOTE vs. Random Undersampling," *International Research Journal of Engineering and Technology*, 2017, [Online]. Available: www.irjet.net
[14] A. Fauzi, "Komparasi Algoritma Dengan Pendekatan Random Undersampling Untuk Menangani Ketidakseimbangan Kelas Pada Prediksi Cacat Software," *Maret*, vol. 15, no. 1, p. 27, 2019, [Online]. Available: www.nusamandiri.ac.id
[15] I. Nekooeimehr and S. K. Lai-Yuen, "Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets," *Expert Syst Appl*, vol. 46, pp. 405–416, Mar. 2016, doi: 10.1016/j.eswa.2015.10.031.
[16] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE - Majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Trans Knowl Data Eng*, vol. 26, no. 2, pp. 405–425, Feb. 2014, doi: 10.1109/TKDE.2012.232.
[17] M. C. Untoro and J. L. Buliali, "Penanganan imbalance class data laboratorium kesehatan dengan majority weighted minority oversampling technique," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 4, no. 1, pp. 23–29, Jan. 2018, doi: 10.26594/register.v4i1.1184.
[18] M. C. Untoro, M. Praseptiawan, M. Widianingsih, I. F. Ashari, A. Afriansyah, and Oktafianto, "Evaluation of Decision Tree, K-NN, Naive Bayes and SVM with MWMOTE on UCI Dataset," in *Journal of Physics: Conference Series*, 2020, vol. 1477, no. 3. doi: 10.1088/1742-6596/1477/3/032005.
[19] P. Y. Saputra, M. Z. Abdullah, and A. P. Kirana, "Improvisasi Teknik Oversampling MWMOTE Untuk Penanganan Data Tidak Seimbang," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 398, Apr. 2021, doi: 10.30865/mib.v5i2.2811.
[20] M. C. Untoro, "MWMOTE optimization for imbalanced data using complete linkage," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 2, pp. 77–82, Apr. 2021, doi: 10.14710/jtsiskom.2021.13748.
[21] M. Iqbal Ramadhan, "Penerapan Data Mining Untuk Analisis Data Bencana Milik Bnpb Menggunakan Algoritma K-Means Dan Linear Regression," 2017.
[22] Y. Pristyanto, "Penerapan Metode Ensemble Untuk Meningkatkan Kinerja Algoritme Klasifikasi Pada Imbalanced Dataset," 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/User+Knowledge
[23] C. E. Puspita, O. N. Pratiwi, and E. Sutoyo, "Perbandingan Algoritma Klasifikasi Support Vector Machine Dan Naive Bayes Pada Imbalance Data," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 8, no. 1, pp. 11–18, Dec. 2021, doi: 10.33330/jurteksi.v8i1.1185.
[24] M. Koziarski, "Radial-Based Undersampling for imbalanced data classification," *Pattern Recognit*, vol. 102, Jun. 2020, doi: 10.1016/j.patcog.2020.107262.
[25] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "NI-MWMOTE: An improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems," *Expert Syst Appl*, vol. 158, Nov. 2020, doi: 10.1016/j.eswa.2020.113504.
[26] T. Purwa, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017)," *Jurnal Matematika, Statistika dan Komputasi*, vol. 16, no. 1, p. 58, Jun. 2019, doi: 10.20956/jmsk.v16i1.6494.