THE 1st INTERNATIONAL CONFERENCE ON DATA SCIENCE AND OFFICIAL STATISTICS
ICDSOS
"Strengthening the Contribution of Data Science and Official Statistics to the Society in the Distruption Era"
2021

SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data

Case Study: IFLS 5

A R B Alamsyah[1], S Rahma[2], N S Belinda[3], A Setiawan[4]

[1]Department of Statistics, STIS Polytechnic of Statistics, Jakarta, Indonesia
[2]Department of Statistics, Muhammadiyah of Semarang University, Semarang, Indonesia
[3]Department of Statistics, IPB University, Bogor, Indonesia
[4]Department of Statistics, Yogyakarta State University, Yogyakarta, Indonesia

*Corresponding author's e-mail: anasbudi98@gmail.com

Abstract. Unbalanced data are often encountered in practice. They complicate the search for a model suitable for classification. This is because the number of individuals who have a history of a disease is less than the number of individuals who do not. We analyse the IFLS 5 data on medical history of a set of patients. We split the dataset in the proportion 80:20 to training and test subsets. Of course, both datasets are unbalanced, with only a small minority of patients who had a stroke. We apply the SMOTE and Nearmiss methods and evaluate the rate of correct classification. After being treated using the two methods, the training data was transformed into balanced data. The classification process is carried out to test the comparison of the effectiveness of the two methods in solving the problem of unbalanced data. Based on the results obtained, it can be concluded that the Nearmiss method is better than SMOTE in balancing the data. It was obtained by comparing several measures such as accuracy, F-score, Kappa, sensitivity, and specificity on the SMOTE and Nearmiss methods.

1. Introduction

Unbalanced data or what is commonly called data unbalance is one of the main problems that will arise in the detection of anomalies in real-time datasets. The dataset is considered unbalanced if in the training data one class has a very large dominance compared to the other classes [1]. Even in some cases of multi-class classifiers, this data unbalance results in a low representation of the data, and ultimately this data tends to be ignored altogether [2]. For the most part, classifier algorithms tend to implicitly assume that the processed data has a balanced distribution, therefore the standard classifier is more inclined towards data with a dominant number of classes.

In some cases, real data is rarely balanced. The problem of unbalanced class data is often caused by one class outnumbering the other classes in the dataset. Examples include oil spill detection [3], remote sensing [4], and text classification [5], so this is an important issue for researchers in the field of data mining [6]. In a health-care study of a population, the number of sufferers from a medical condition is often much smaller than the number of heatlhy subjects. Methods that rely on balance, or perform well only for balanced data are not very useful in this setting, unless they are modified.

In general, some solutions can be used to deal with unbalanced datasets, i.e. at the algorithmic level or the data level. The approach at the algorithmic level is when machine learning algorithms are modified to accommodate data unbalances. Commonly modified algorithms are C4.5, Naïve Bayes, Random Forest, Neural Network K-Means, and so on. While the approach at the data level involves re-sampling to reduce the class unbalance. The two basic sampling techniques used at the data level are random oversampling (ROS) and random undersampling (RUS). ROS randomly duplicates data from a minority class. ROS can be a good choice when there is not much data available, but it may cause overfitting because this method creates exact duplicates of data from minority classes. Synthetic Minority Over-Sampling Technique (SMOTE) is a method of ROS, SMOTE is a technique that creates a new sample of minority data to balance the data by resampling the minority class [7]. Meanwhile, to modify the class distribution, RUS will discard data (from the majority class) randomly. The disadvantage of RUS is that it can cause underfitting because it can delete information that may be valuable. One of the well-known RUS methods is the Nearmiss method. This method can balance data by eliminating data points from a larger class when there are two very close points of different classes.

Several studies have tried to implement several methods including SMOTE and Nearmiss. Hairani's research used SMOTE to deal with class unbalances in the classification of diabetes with a total of 268 datasets from the positive class (minority class) and 500 data from the negative class (majority class) shows that the SVM classification method has accuracy by 82% and the best sensitivity, which is 77% [8]. The research conducted by Johariyah was used to assess the quality of obstetric services in health facilities and identify Nearmiss indicators as the cause of maternal morbidity. The results obtained showed that patients with Nearmiss in RSUD Cilacap and RSI Fatimah were the most in the healthy category, namely 98.6% and 979% with Nearmiss patients who died in RSUD Cilacap as many as 0.9% and there were no Nearmiss patients who died in RSI. Fatimah [9].

In this study, we compare the Random Oversampling (ROS) and Random Undersampling (RUS) data balancing methods in cases of classification of stroke history that occurred in a person. The technique used this time is an algorithmic approach, namely SMOTE as a candidate for Random Oversampling and Nearmiss as a candidate for Random Undersampling. In this study, a validation model with binary logistic classification was applied to assess the algorithm's accuracy performance so that it has effective and good performance.

2. Method of Research

This study aims to compare the balanced data method in the case of stroke classification. This involves four steps, data collection, balancing data, classification, and interpretation, illustrated in ‘figure 1’.

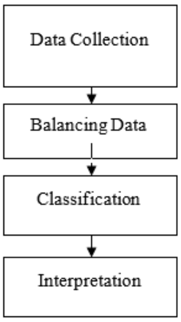

Figure 1. Research process.

2.1. Data Collection

The data used in this research is data from Indonesia Family Live Survey 5 (IFLS 5). Indonesian Family Live Survey 5 (IFLS 5) data is the latest data that includes social, economic, and health variables in Indonesia. The variables used in this study do not include all the variables in the 2014 IFLS data, but only a few variables as a result of Data Cleaning which will be explained in ‘table 1’.

Table 1. Research variable.

Variable	Variable Name	Data Type	Description
Y	Stroke Desease History Status	Nominal	Yes: 1 No: 0
X_1	Age	Ratio	In range 14 – 103
X_2	Body Mass Index	Ratio	In range 10.85 – 68.76

2.2. Balancing Data

Balancing data is a process that is carried out to overcome unbalanced data which is commonly known as unbalanced data. Unbalanced data is one of the main problems that arise in the detection of anomalies in real-time datasets. A dataset is considered unbalanced data when one of its classes has a very large dominance compared to other classes [2]. There are two methods used to overcome the unbalance data in this case, the two methods are Synthetic Minority Over-Sampling Technique (SMOTE), and Nearmiss, each of which is performed using R software in ‘figure 2’.

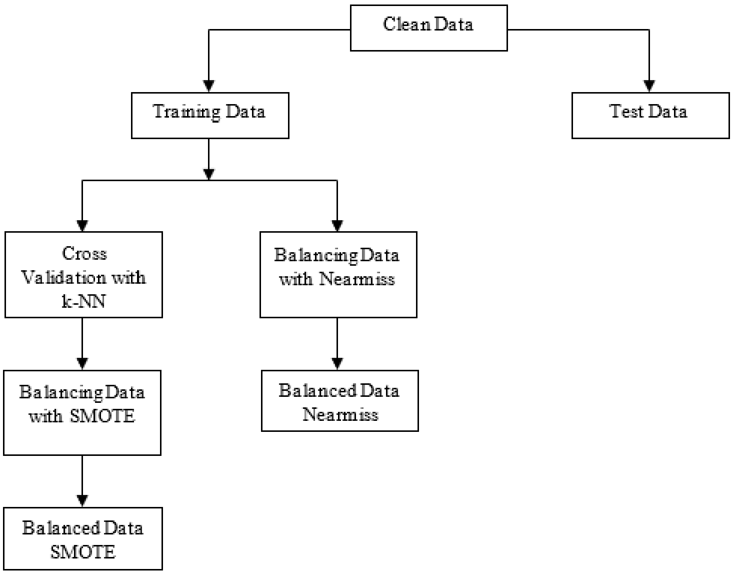
Clean Data
Training Data
Test Data
Cross Validation with k-NN
Balancing Data with Nearmiss
Balancing Data with SMOTE
Balanced Data Nearmiss
Balanced Data SMOTE

Figure 2. Dataset balancing process.

The way data balancing works using SMOTE begins by calculating the distance between data on minority data, determining the percentage of SMOTE, then determining the number of closest k, and finally generating synthetic data with the following equation [10].

$$x_{syn} = x_i + (x_{nn} - x_i)\delta \tag{1}$$

where
x_{syn} = Synthetic data
x_i = Data to be replicated
x_{knn} = Data that has the closest distance to the data to be replicated
δ = Random value from 0 to 1

The determination of the closest k value in the SMOTE process is carried out in the Cross-Validation process using K-Nearest Neighbor (k-NN). The k-NN method is one of the supervised learning algorithms used for the classification process. In practice, this method classifies an object based on the distance between the object and other objects to predict the new class [7]. The number of neighboring objects used is denoted by k. After determining the value of the k nearest neighbors of the object, then calculating how much data follows each class in the k neighbors. The class with the most followers will be the winner given as the class label on the related object. Nearmiss method used in this study is Nearmiss-1. The workings of this method are to select a sample from the majority class which is close to several samples from the minority class. The criteria used in the selection of samples from the majority class is the sample that has the smallest average distance to the three closest samples from the minority class [11].

After being treated with both methods, the previously unbalanced data was transformed into balanced data. After obtaining balanced data from each method, the two methods will be compared their effectiveness in classifying.

2.3. Classification

The classification process used in this study is classification with binary logistic regression. Binary logistic regression models the probability of the success of two classes of criteria. The linear combination of the predictor variables is used to adjust the logit transformation of the probability of success of each subject (π_i) with the following equation [12].

$$Ln\left[\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right] = w_0 + w_1 X_{1i} + \cdots w_n X_{ni} \tag{2}$$

The regression coefficient is estimated using eq. (2), where the above equation is the result of the transformation of the following equation.

$$\hat{\pi}_i = \frac{\mathrm{e}^{w_0 + w_1 X_{1i} + \cdots w_n X_{ni}}}{1 + \mathrm{e}^{w_0 + w_1 X_{1i} + \cdots w_n X_{ni}}} \tag{3}$$

If the eq. (3) for each i have a probability ($\hat{\pi}_i$) more than 0.5, then the subject i classified into the successful group, and vice versa if the probability obtained is less than 0.5, then the subject i classified into the non-success group.

The purpose of the classification process here is to see how effective the two data balancing methods are in producing classification results. The complete classification process using binary logistics is described in ‘figure 3’.

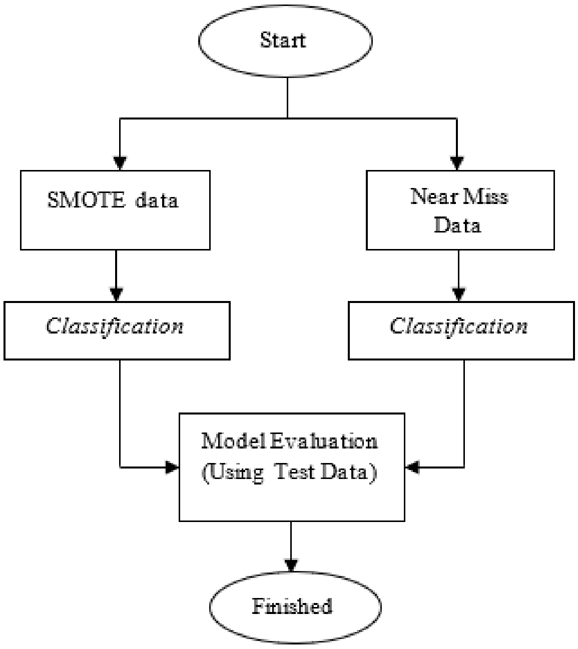

Figure 3. Classification process.

The classification process begins by preparing two balanced data consisting of data from the SMOTE method and data from the Nearmiss method. Classification modeling is carried out using binary logistic classification to obtain a model of each dataset balancing method. After each model is formed, the next step is to evaluate each model using the test data obtained during the dataset balancing process. The results obtained from the results of the model evaluation include the confusion matrix, the total accuracy value (accuracy), specificity, and sensitivity.

2.4. Interpretation of Classification Result

After the classification results from each method are obtained, it is continued with the interpretation of the results, which aims to determine the best dataset balancing method in classifying stroke history. Determining which method is the best in balancing datasets is done by comparing the Confusion Matrix SMOTE and Near Miss results. The results of model evaluation include the confusion matrix, 'table 2' below shows the Confusion Matrix table.

Table 2. Confusion matrix table.

Class	*Predictive Positive*	*Predictive Negative*
Actual Positive	TP	FN
Actual Negative	FP	TN

True Positive (TP) and True Negative (TN) are the numbers of classes of positive and negative that are correctly classified. While False Positive (FP) and False Negative (FN) are the numbers of positive and negative classes that are not classified properly [7]. From the confusion matrix table above, various measures such as accuracy, precision, sensitivity, and specificity are obtained, which can be used to compare the classification results between dataset balancing methods.

3. Result and Discussion

3.1. Cross-Validation Descriptive Statistics

The description of the independent variable data is shown in 'table 3'.

Table 3. Descriptive statistics of independent variable.

Variable Name	Minimum	Q1	Q2	Q3	Maximum	Mean
Age	14	26	35	48	103	38.14
Body Mass Index	10.85	19.86	22.57	25.92	68.76	23.22

The data is divided into 80% for training data and 20% for test data. The training data obtained 25923 observations, with 179 observations stating that they had had a stroke, while the test data obtained 6481 observations, with 45 observations stating that they had had a stroke. ‘Figure 4’ shows a visualization of the classification between the Yes class for blue point and the No class for red point. It can be seen in ‘figure 4’ that the non-stroke class is more dominant than the stroke class.

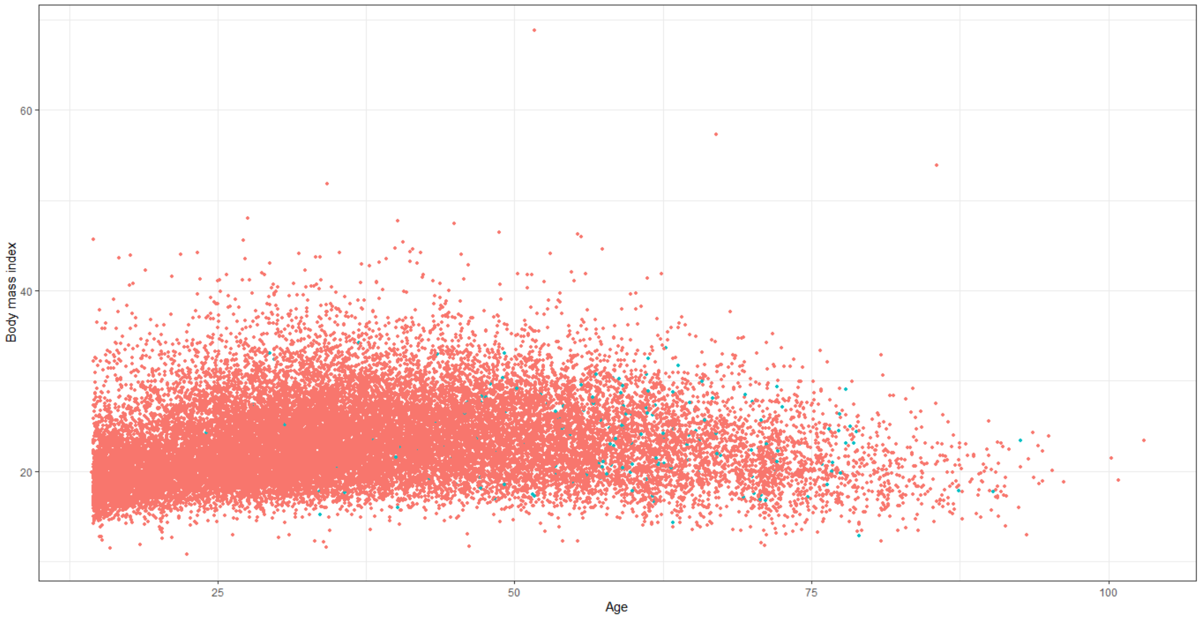

Figure 4. Visualization of classification on training data.

In the training data, the optimal K value is determined using K-Nearest Neighbor. The training data was tested by cross-validation with 10 folds and 3 repetitions obtained as shown in ‘table 4’. Based on table 3, the optimal K value is K = 5, 7, and 9. We obtain accuracy of 99.3% and Kappa coefficient equal to zero.

3.2. SMOTE

After doing the SMOTE method with K=5 on the training data, the data for the non-stroke class was 25744 (50.14%) and 25597 (49.86%), the visualization of the classification can be seen in ‘figure 5’.

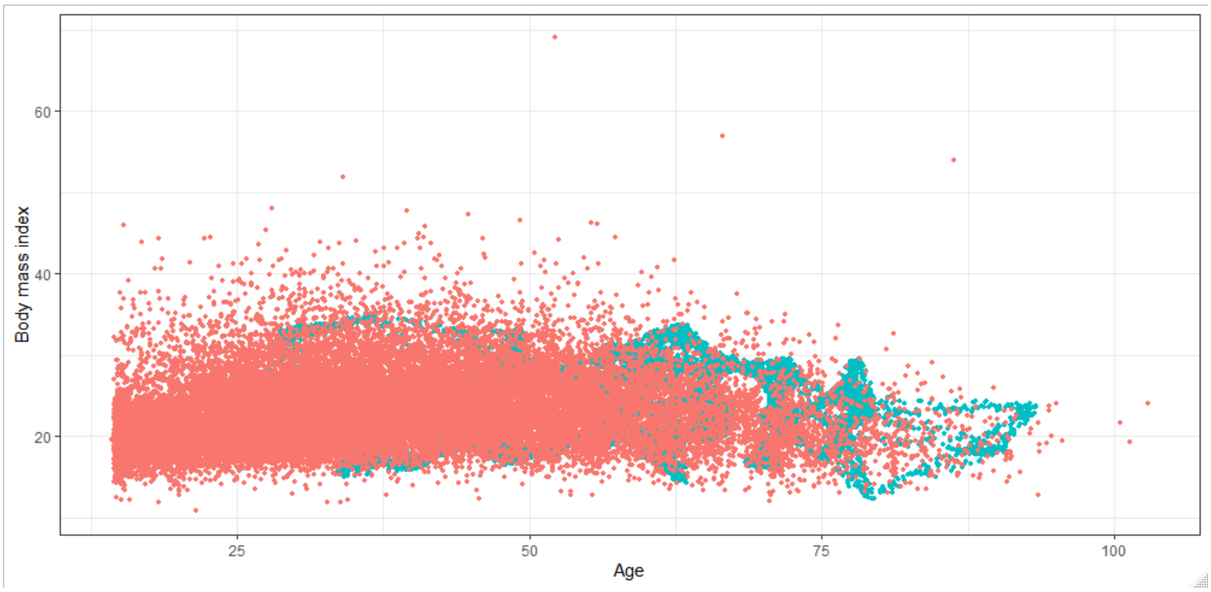

Figure 5. Visualization of classification on SMOTE training data.

After obtaining the train data from the SMOTE method, a classification analysis was performed using Binary Logistics Regression. Predictions are made with the test data and the results of the comparison with the original test data are shown in ‘table 4’ and the visualization is in ‘figure 6’.

Table 4. Confusion matrix.

		Actual Values	
		No	Yes
Predicted	No	4807	8
Values	Yes	1629	37

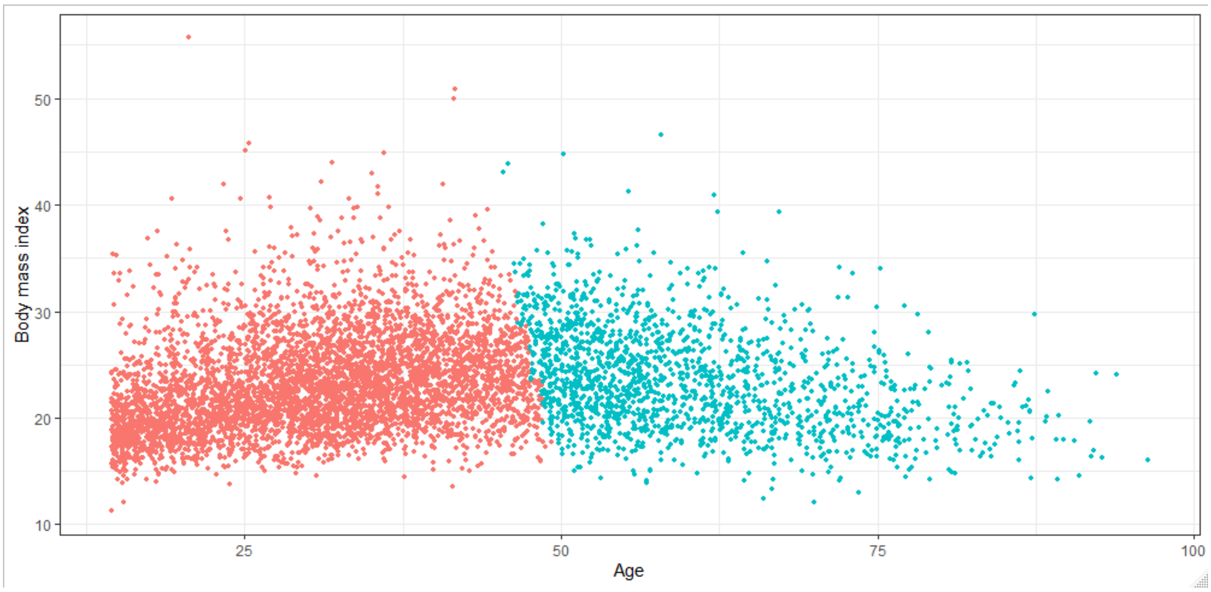

Figure 6. Visualization of prediction results classification on SMOTE training data.

3.3. Nearmiss

After the Nearmiss-1 method was applied to the training data, data for the non-stroke class was 179 (50%) and the stroke class was 179 (50%). Classification visualization after the Nearmiss method can be seen in 'figure 7'.

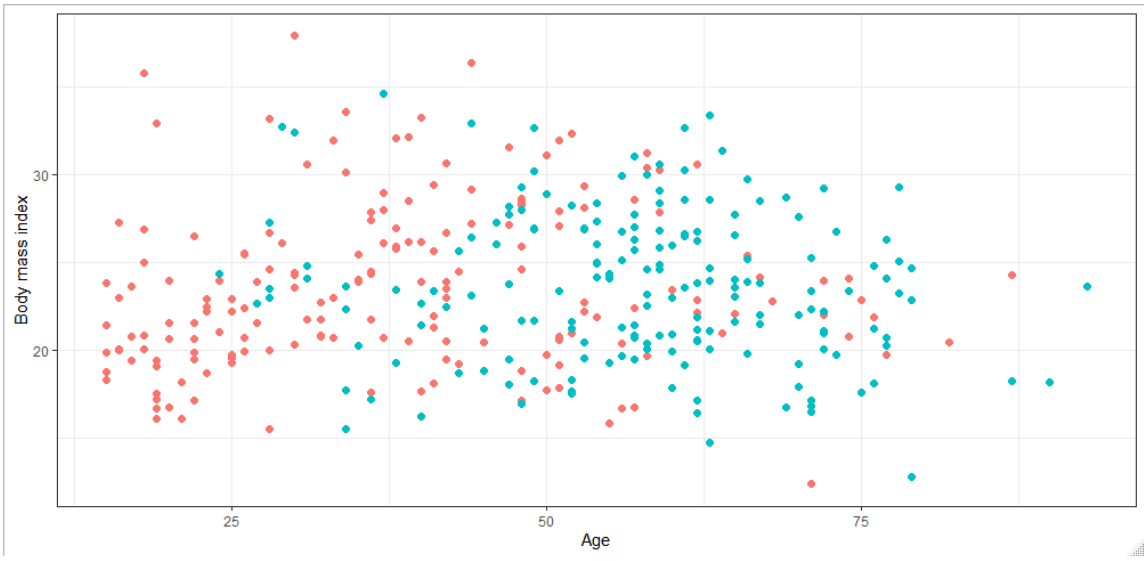

Figure 7. Visualization of training data for the Nearmiss method.

After obtaining the training data from the Nearmiss method, a classification analysis was performed using Binary Logistics Regression. After that, predictions are made with the test data and the results of the comparison with the original test data are shown in 'table 5' and the visualization is in 'figure 8'.

Table 5. Confusion matrix.

		Actual Values	
		No	Yes
Predicted	No	4884	8
Values	Yes	1552	37

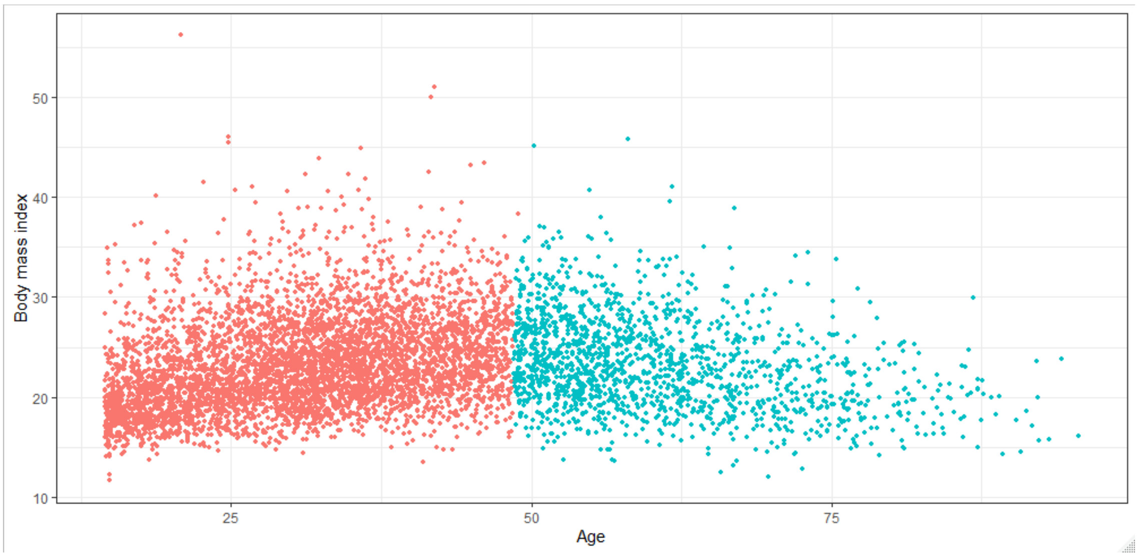

Figure 8. Visualization of prediction results classification on Nearmiss training data.

3.4. Comparison of the Two Methods

Comparison between methods in classification cases can be done by looking at the values of accuracy, sensitivity, specificity, F-score, and Kappa. The five values can measure how well the model is used. This is indicated by the higher the value of the goodness of the model, the better the model. The size of the goodness of the model for the SMOTE method with the Nearmiss method can be seen in 'table 6'.

Table 6. Confusion matrix.

Goodness of fit	*SMOTE*	*Nearmiss*
Accuracy	0.7474	0.7593
Fscore	0.8545	0.8623
Kappa	0.0300	0.0322
Sensitivity	0.7469	0.7589
Specificity	0.8222	0.8222

Based on table 6, the value of the goodness of the Nearmiss method is higher than the SMOTE method. So it can be concluded that in the data used in this study, the Nearmiss method is better used than the SMOTE method.

4. Conclusion

From the results of this study, we can conclude that the two methods used, namely SMOTE and Nearmiss, can overcome unbalanced data where the data train consists of 25923 observations, with the proportion of Yes and No classes being 99.3% and 0.7%, respectively after resampling with SMOTE and Nearmiss in the train data, the proportions of Yes: No classes are 50.14%:49.86% and 50%:50%, respectively. In addition, by using binary logistic regression analysis, that can measure the goodness of the fit. These values are the values of Accuracy, Sensitivity, Specificity, Kappa, and F-score. The measure of goodness of the Nearmiss method is higher than the SMOTE method. Based on this, it is concluded that the Nearmiss method is better than the SMOTE method for handling Stroke data in IFLS 5.

References

[1] Sastrawan AS, Studi P, Informatika T, Studi P, Komputasi I, Sains F, et al. Analisis Pengaruh Metode Combine Sampling Dalam Churn Prediction untuk Perusahaan Telekomunikasi. 2010;2010(semnasIF):14–22.

[2] Arifiyanti AA, Wahyuni ED. Smote: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining. SCAN **-** *J Teknol Inf dan Komun*. 2020;**15**(1):34–9.

[3] Kubat, Miroslav;Matwin S. Addressing the Course of Imbalanced Training Sets: One Sided Selection. 148:148–62.

[4] Bruzzone L, Serpico SB. Classification of imbalanced remote-sensing data by neural networks. *Pattern Recognit Lett*. 1997;**18**(11–13):1323–8.

[5] Kubat M, Holte RC, Matwin S. Machine learning for the detection of oil spills in satellite radar images. *Mach Learn*. 1998;**30**(2–3):195–215.

[6] Wu X, Kumar V, Ross QJ, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. *Knowl Inf Syst*. 2008;**14**(1):1–37.

[7] Siringoringo R. Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor. *J ISD*. 2018;**3**(1):44–9.

[8] Hairani H, Saputro KE, Fadli S. K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes. *J Teknol dan Sist Komput*. 2020;**8**(2):89–93.

[9] Johariyah. Analisis perbandingan kejadian near miss pada pasien obstetri sebagai penyebab morbiditas ibu. *J Kesehat Al-Irsyad (JKA)*, Vol **IX**, No 1. 2016;IX(1):59–69.

[10] Khaulasari H. Combine Sampling Least Square Support Vector Machine Untuk Klasifikasi

Multi Class Imbalanced Data. *Widya Loka IKIP Widya Darma*. 2018;**5**(1):82–93.

[11] Jianping Z, Mani I. kNN Approach to Unbalanced Data Distribution: A Case Study involving Information Extraction. 148:148–62.

[12] Elhassan T, Aljurf M, Al-Mohanna F, Shoukri M. Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. *Glob J Technol Optim*. 2016;**01**(S1).