RESEARCH ARTICLE | MAY 09 2023

The use of resampling techniques to overcome imbalance of data on the classification algorithm

Riska Aryanti ✉; Yoseph Tajul Arifin; Sayyid Khairunas; ... et. al

AIP Conference Proceedings 2714, 020053 (2023)
https://doi.org/10.1063/5.0128424

The Use of Resampling Techniques to Overcome Imbalance of Data on the Classification Algorithm

Riska Aryanti,[1, a)] Yoseph Tajul Arifin,[1, b)] Sayyid Khairunas,[1, c)] Titik Misriati,[1, d)] Sopiyan Dalis,[1, e)] Taufik Baidawi,[1, f)] Rizky Ade Safitri,[2, g)] and Siti Marlina[2, h)]

[1)] *Universitas Bina Sarana Informatika, Jakarta, Indonesia.*
[2)] *Universitas Nusa Mandiri, Jakarta, Indonesia*

[a)] *Corresponding author: riska.rts@bsi.ac.id*
[b)] *Electronic mail: yoseph.ypa@bsi.ac.id*
[c)] *Electronic mail: sayyid.skh@bsi.ac.id*
[d)] *Electronic mail: titik.tmi@bsi.ac.id*
[e)] *Electronic mail: sopiyan.spd@bsi.ac.id*
[f)] *Electronic mail: taufiq.tfb@bsi.ac.id*
[g)] *Electronic mail: rizki.rzs@nusamandiri.ac.id*
[h)] *Electronic mail: siti.smr@nusamandiri.ac.id*

Abstract. Imbalance of dataset in the accuracy testing process can lead to biased results. It occurs due to insufficient data in the training phase where unbalanced data causes problems in Machine Learning. Classification and predicting results become difficult when there is insufficient data to study. To overcome this, it takes steps in balancing the data, one of which is the random over sampling technique. The basic principle of using this technique is to rebalance an unbalanced data set with a concrete strategy. The use of sampling technique in the case of data imbalance is proven to be able to improve the performance of the algorithm. The results of testing the KNN, Naive Bayes, SVM, J.48 and Random Forest algorithms using 10 fold cross validation on the public dataset of early stage diabetes risk prediction from the Hospital in Sylhet, Bangladesh after going through the re-sample stage proved to be able to improve measurement results with a high level of accuracy. The highest using KNN with an accuracy of 99.4231%, while the results of the J.48 algorithm test with an accuracy rate of 99.2308%, Random Forest with an accuracy rate of 98.0769%, SVM with an accuracy rate of 96.5385% and Naïve Bayes with an accuracy rate of 90.3846%.

INTRODUCTION

Diabetes is a serious threat to global health that does not respect socioeconomic status or national boundaries. Professor Rhys Williams, as Chair of the IDF Diabetes Atlas Committee[1]. Diabetes is a chronic metabolic disease that occurs when the pancreas is no longer able to make insulin or use the insulin produced by the body properly, this can cause an increase in blood glucose levels (otherwise known as hyperglycaemia), which in the long run causes damage to the body and failure of various organs. serious organs and tissues such as the heart, blood vessels, eyes, kidneys, and nerves [2]. The International Diabetes Federation states that 1 in 11 people live with diabetes. Other information was also conveyed that in 2019 there were around 463 million adults aged (20-79 years) living with diabetes worldwide. And it is predicted that in 2045 it will increase up to 700 million. About 79% of adults with diabetes live in low- and middle-income countries. The region with the highest estimated number of diabetes-related adult deaths is the Asian Region. In 2019, Bangladesh was recorded as being ranked as the 10th country with the highest number of people with diabetes (8.4 million) [3]. Research has been carried out by Faniqul Islam by analyzing datasets taken directly from questionnaires distributed to Diabetes Patients Hospital in Sylhet, Bangladesh. The results of the questionnaire distribution data will be used to model disease risk prediction with the possibility of having new diabetes or will become diabetic patients with the aim of developing an accessible and easy-to-use tool for end users to check the risk of suffering from diabetes [4]. The use of Naive Bayes Algorithm, Logistic Regression Algorithm, and Random Forest Algorithm as well as the application of the Ten Fold Cross-Validation and Percentage Split evaluation techniques found that the best accuracy results from testing the algorithm on the dataset were using the Random Forest Algorithm, with an accuracy level of 10 Fold technique Cross-Validation is 97.4% and Percentage Split is 99%. Through further observations regarding the use of the early stage diabetes risk prediction dataset containing data on signs and symptoms of new diabetics or those who will become diabetic patients taken from the UCI Machine Learning Repository website, it was found that the total data was 520 instances, 320 positive and 200 negative. From these data, it can be found that there is an imbalance in the amount of data between positive class data and negative class

2nd International Conference on Advanced Information Scientific Development (ICAISD) 2021
AIP Conf. Proc. 2714, 020053-1–020053-11; https://doi.org/10.1063/5.0128424
Published by AIP Publishing. 978-0-7354-4520-8/$30.00

data. In other words there is an imbalance of datasets where unbalanced data causes problems in Machine Learning classification and predicting results becomes difficult when there is not enough data to study [5]. To overcome this, the dataset must first go through the preprocessing stage to balance the data, where the most widely used technique is the sampling model from the many sampling-based preprocessing methods that have been proposed to solve the problem of unbalanced dataset classification. The basic principle of this method is to rebalance an unbalanced data set with a concrete strategy[6]. The use of sampling techniques in cases of data imbalance is proven to be able to improve the performance of learning algorithms [7]. Class imbalance occurs when the minority class is much smaller or less frequent than the majority class [8], where models made using unbalanced data will produce low minority prediction accuracy. Information from the majority class dominates the minority class, causing biased decision boundaries in the classification system [9]. To overcome the class imbalance problem, this study uses a data level approach by applying the resampling method, namely random oversampling according to the dataset criteria owned so as to improve accuracy performance in algorithm testing.

METHODS

The method of resampling unbalanced datasets randomly or resampling is a dataset processing technique involving the creation of a new transformed version of the training dataset in which the selected samples have different class distributions. This is a simple and effective strategy for unbalanced classification problems [10]. By applying a resampling strategy to get a more balanced distribution of data, it is an effective solution to the imbalance problem [11]. There are two main approaches to random sampling techniques in the classification of data imbalances, namely oversampling and undersampling, where random oversampling works to duplicate samples randomly in the minority class while random undersampling works to remove random samples in the majority class. Random oversampling involves selecting samples at random from the minority class dataset, performing substitutions, and adding them to the training dataset. For random undersampling, it involves randomly selecting samples from the majority class and removing them from the training dataset by applying the two techniques above, a balance of majority or minority class which are selected and added randomly [12]. However, the resampling technique can be readjusted to the criteria of the dataset and the model used. In this study, the random sampling technique used to overcome the data imbalance is the random over sampling technique. Using this method will determine the number of instances to fetch for a given class where data.numInstances() will return the total number of instances in the dataset, numInstancesPerClass[i] and store the number of instances in class i and numActualClasses according to the actual number of classes present in the dataset. With this technique, the majority class undersample dataset will have the same number of instances. The following is a resampling technique using an expression to determine the number of instances to sample for certain class i:

```
int sampleSize = ( int ) (( m_SampleSizePercent / 100,0 ) * (( 1
- m_BiasToUniformClass ) * numInstancesPerClass [ i ] +
m_BiasToUniformClass * Data . numInstances () / numActualClasses ));
```

FIGURE 1. Expression model of random sampling (re-sampling)

Figure 2 describes the research method and the dataset processing flow of this research:

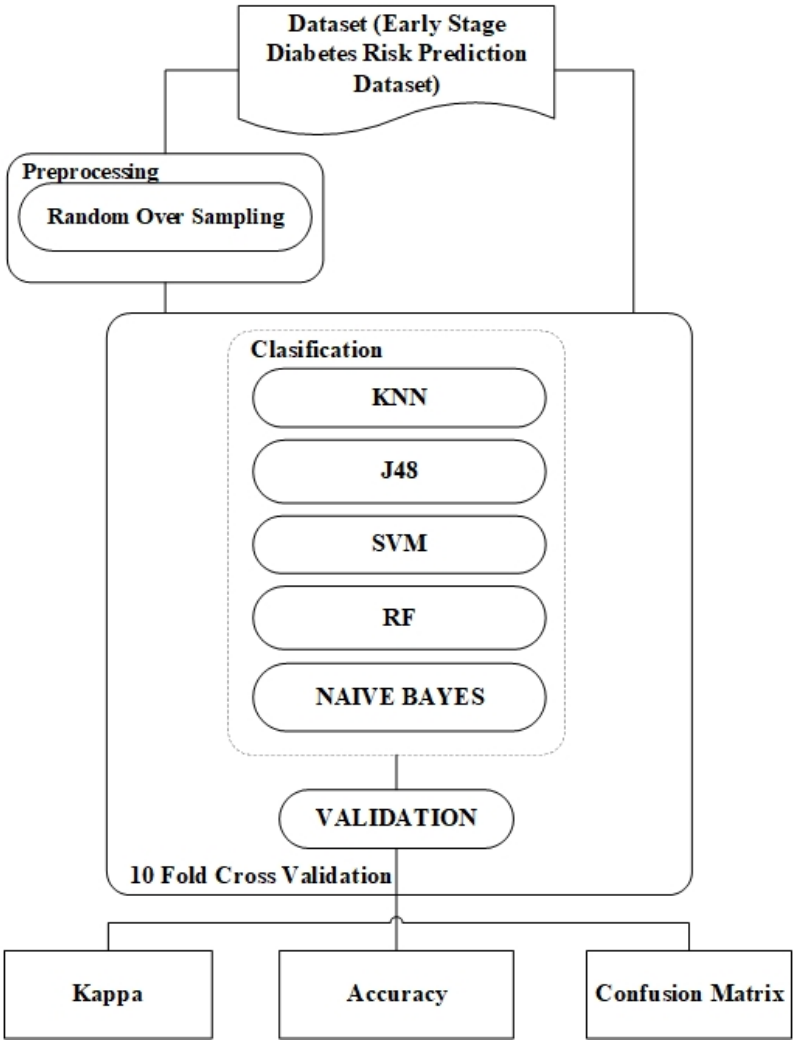

FIGURE 2. The proposed research method

Data Collection

The first step is collecting datasets. The dataset in this study was retrieved from the UCI Machine Learning Repository It is an Early-stage diabetes risk prediction dataset consisting of 16 attributes and 520 data. This dataset is derived from a collection of questionnaires for prospective patients with early-stage diabetes at Sylhet Hospital, Bangladesh.

TABLE I. Description of Dataset

Description	Number of Attributes	Number of Case
Diabetes symptom data collection	16	520

TABLE II. Attribute Description

Name of Attribute	Value
Age	1.20–35, 2.36–45, 3.46–55,4.56–65, 6.above 65
Sex	1.Male, 2.Female
Polyuria	1.Yes, 2.No
Polydipsia	1.Yes, 2.No
Sudden weight loss	1.Yes, 2.No
Weakness	1.Yes, 2.No
Polyphagia	1.Yes, 2.No
Genital thrush	1.Yes, 2.No
Visual blurring	1.Yes, 2.No
Itching	1.Yes, 2.No
Irritability	1.Yes, 2.No
Delayed healing	1.Yes, 2.No
Partial paresis	1.Yes, 2.No
Muscle stiffness	1.Yes, 2.No
Alopecia	1.Yes, 2.No
Obesity	1.Yes, 2.No
Class	1.Positive, 2.Negative

The total data is 520 which consist of 320 data with positive values and 200 data are negative values. Detailed descriptions of the dataset and its attributes are shown in Table 1 and 2. Two class variables were used to find out whether the patient had diabetes risk (Positive) or not (Negative).

Processing

Before testing the dataset, it is having preprocessed phase first. By using a resampling technique, namely the undersampling and oversampling sampling technique model for the KNN, Naive Bayes, SVM, J.48 and Random Forest algorithms. This resample technique is used therefore the data will be balance. In addition, it is to make the parameters that are changed within the To Uniform Class bias parameter becomes 1.0. This parameter is changed because sampling only applies if there is a class variable. When fully biased towards the input distribution (B=0), each subsample replicates the class distribution of the complete dataset. B=1 is equivalent to unsupervised resampling where points are drawn uniformly from the entire population regardless of class.

Classification

Next phase, the dataset is tested with several classification algorithms, namely: KNN algorithm, Naive Bayes, SVM, J.48 and Random Forest using 10-fold cross validation. As it completes, the confusion matrix, accuracy, and kappa values were obtained using the weka tools.

a. KNN Algorithm
   KNN is used for both classification and regression problems, but in industries it is widely known for solving classification problems rather than regression[13].

$$Q(d_i, C_m) = \sum_{j=1}^{K} sim(d_i, d_j)\delta(d_i, C_m) \tag{1}$$

$$\delta(d_i, C_m) = \begin{Bmatrix} 1, if\, d_i \varepsilon C_m \\ 0, if\, d_i \varepsilon C_m \end{Bmatrix} \tag{2}$$

b. Naïve Bayes

Naive Bayes uses a probabilistic algorithm. The algorithm assumes the features and variables provided are independent to one another. It is carried out by using a probabilistic approach, which determines class probabilities and predicts most probable classes. The following equation from 3 to 5 represent the classification formula, where Pos and Neg represent a person with diabetes risk and without diabetes risk, which are the values of the class attribute for this dataset. X is the instances of the dataset as well as person.

$$P(Pos|X) = P(X_1|pos) * P(X_2|pos) * \ldots * (X_n|pos) * P(Pos) \tag{3}$$

$$P(Neg|X) = P(X_1|neg) * P(X_2|neg) * \ldots * (X_n|neg) * P(Neg) \tag{4}$$

$$P(X_i|pos) = \frac{(Total pos|x_i)}{Total pos} \tag{5}$$

c. SVM

SVM can solve the classification problem, but SVM has a weakness in the difficulty of selecting appropriate and optimal features in the weight of the attributes used, causing the classification accuracy to be low[15]. SVM is a supervised machine learning algorithm. It is believed to be the best "off the-shelf" algorithm, especially for high-dimensional spaces. SVM uses decision boundary, also called as separating hyperplane, to distinguish between various classes of the target variable. The hypothesis function (h) for an SVM is given by

$$h_{w,b}(x) = g(w^T x + b) \tag{6}$$

d. J.48

J48 algorithm is a kind of decision tree which belongs to the supervised learning algorithm. It is one of the most important classifiers as it is easy and simple to implement[4]. J.48 is an algorithm derived from C4.5. This algorithm produces a binary tree where in the classification process, the tree will be built and each tuple of the tree will be applied to the database and the classification results of the tuples. The J.48 algorithm will ignore incomplete values in the tree creation process. The basis of this algorithm is to divide the data into several parts based on the attribute values of the items in the training dataset. The J.48 algorithm can classify either through the decision tree or the rules obtained from the tree. The steps in the J.48 algorithm are as follows:

1) Define training dataset.
2) Determination of the roots of the decision tree.
   Calculation of Gain value using the equation

$$Entropy(S) = \sum_{i=1}^{n} \breve{} p_i * log_2 pi \tag{7}$$

   Restart step 2 until all tuples are divided by using equation

$$Gain(S,A) = S - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * S_i \tag{8}$$

3) The division process will stop when all the tuples in point N have obtained the same class and or there are no attributes in the subdivided tuples or there are no tuples in the empty branch.

e. Random Forest

Random forest uses bagging method to train the dataset[4]. For a training set of X = $x_1, \ldots, x_n$ and Y = $y_1, \ldots, y_n$, it selects random sample B times with replacement of the training set and fits trees to these samples. After training, it predicts unseen samples x by averaging the predictions from all the individual regression trees on x and also by taking the majority vote in the case of classification trees.

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{9}$$

Performance Evaluation

To evaluate the result of our model, the standard performance measures of pattern classification including accuracy, precision, recall, ROC, KAPPA, and MAE score are calculated. These evaluation metrics are defined based on the following four basic test statistics in classification task[14]:

1 True positive (TP): the cases in which the model predicts positive and the actual label is positive

2 True negative (TN): the cases in which the model predicts negative and the actual label is negative

3 False positive (FP): the cases in which the model predicts positive but the actual label is negative

4 False negative (FN): the cases in which the model predicts negative but the actual label is positive

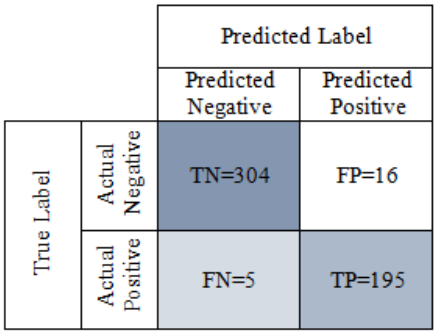

FIGURE 3. Plot of the un-normalized confusion matrix for the proposed mode

Thus, it can be calculated as:
Accurasy (ACC)

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

Precision(P)

$$P = \frac{TP}{TP+FP} \tag{11}$$

Recall/Sensitivity

$$R = \frac{TP}{TP+FN} \tag{12}$$

MAE (Mean Absolute Error)

MAE is the average absolute difference between the actual (actual) value and the predicted (forecast) value. MAE is used to measure the accuracy of a statistical model in making predictions or forecasting.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |A_i \breve{} F_i| \tag{13}$$

where:
n is the sample size
A_i is the actual data value up to-i
F_i is the forecasting data value up to-i

KAPPA

$$K = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+} * x_{+i})} \tag{14}$$

where r is the number of rows in the matrix, $x_i i$ is the number of observations in row i and column i, x_i+ and $x_+ i$ are the marginal totals of row i and column i, respectively, and N is the total number of observations.

Receiver Operating Characteristic (ROC)

The ROC curve is based on the value obtained in the calculation with the confusion matrix, namely between the False Positive Rate and the True Positive Rate. Where:

1. False Positive Rate (FPR) = False Positive / (False Positive + True Negative)
2. True Positive Rate (TPR) = True Positive / (True Positive + False Negative)

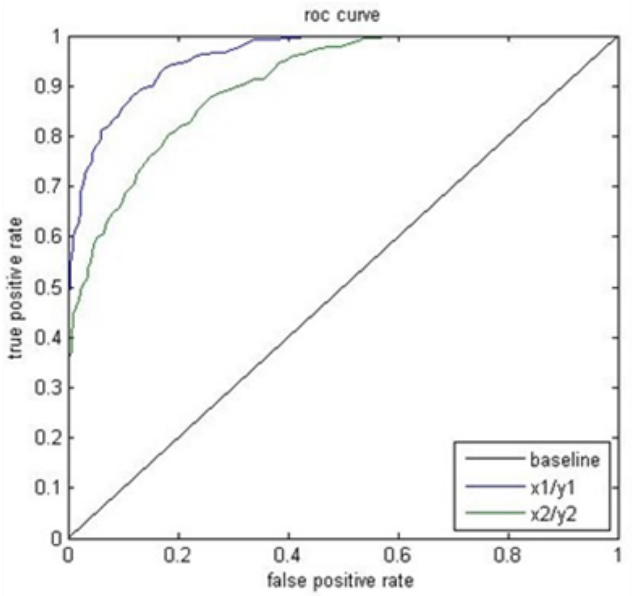

FIGURE 4. ROC Curve

RESULT AND DISCUSSION

Initial testing was carried out through the stages of dataset collection, validation and accuracy directly using the Decision Tree J.48, KNN, SVM, Naïve Bayes and Random Forest classification algorithm models. The visualization of the class label dataset can be seen in Figure 5, the visualization of all the attributes in the dataset can be seen in Figure 6 and the results of the comparison of the algorithm model testing can be seen in Table 3.

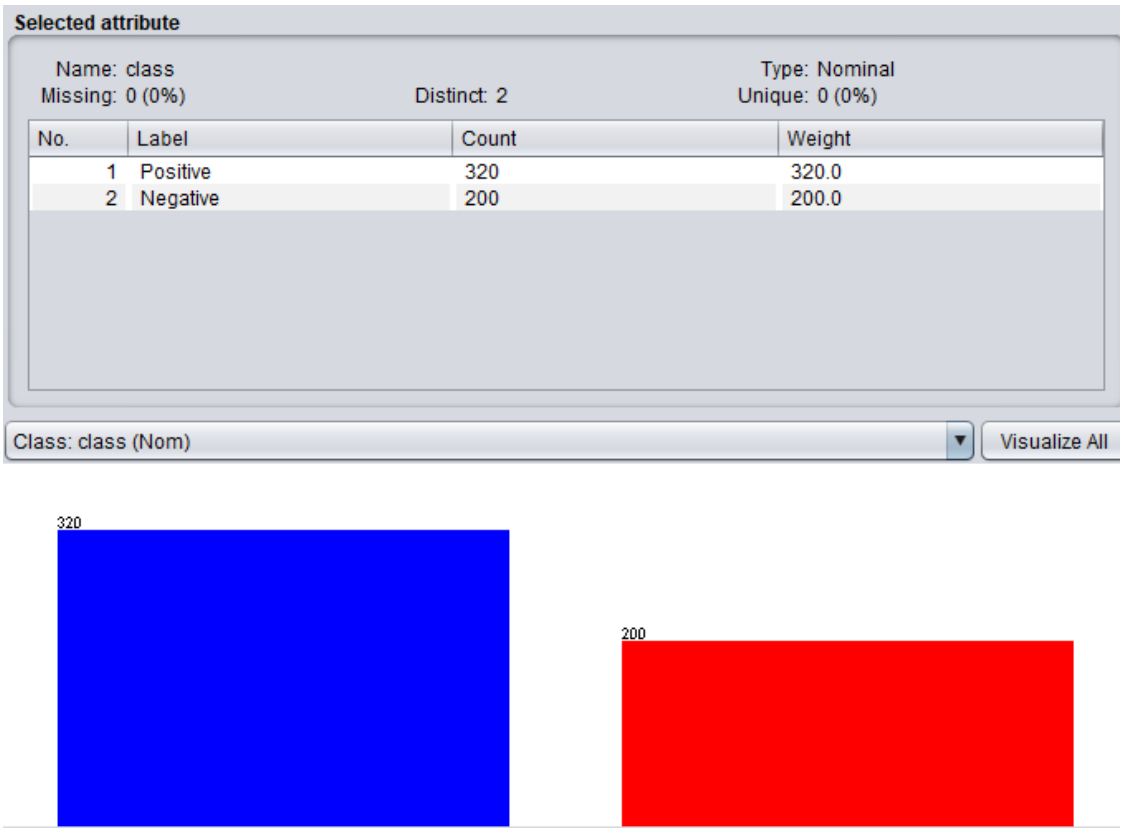

FIGURE 5. Dataset Visualization (Imbalance Data)

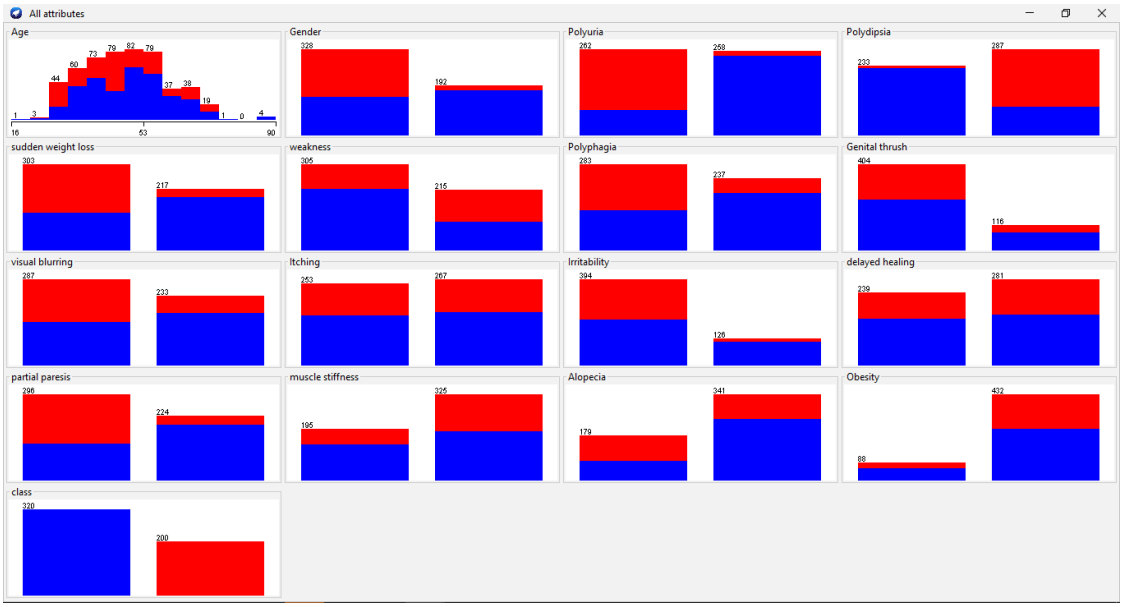

FIGURE 6. Visualization of all attributes on the dataset

TABLE III. Classification Test Results from Algorithm C.45, KNN, SVM, Nave Bayes and Random Forest

Algorithm	Accuracy	Kappa Statistic	MAE	Precision	Recall	ROC
J.48	95.9615%	0.9156	0.0549	0,961	0,960	0,966
KNN	98.0769%	0.9596	0.0207	0,981	0,981	0,984
SVM	92.1154%	0.8339	0.0788	0,921	0,921	0,918
Naïve Bayes	87.1154%	0.734	0.149	0,878	0,871	0,945
Random Forest	97.5%	0.9472	0.0566	0,975	0,975	0,998

According to data in Table 3, it can be concluded that the J.48 classification algorithm is 95.9615%, the SVM classification algorithm is 92.1154%, the Naïve Bayes classification algorithm is 87.1154%, while the Random Forest classification algorithm is 97.5%. The classification algorithm with the best level of accuracy is KNN with an accuracy rate of 98.0769%. Furthermore, testing is carried out with a pre-processing process to balance the data (resampling) using random over sampling technique followed by an algorithm testing process in the form of validation and accuracy using the Decision Tree J.48, KNN, SVM, Naïve Bayes and Random Forest classification algorithm models. The visualization of the class label dataset can be seen in Figure 7 and the results of the comparison of the algorithm model testing can be seen in Table 4 and visualized in Figure 8.

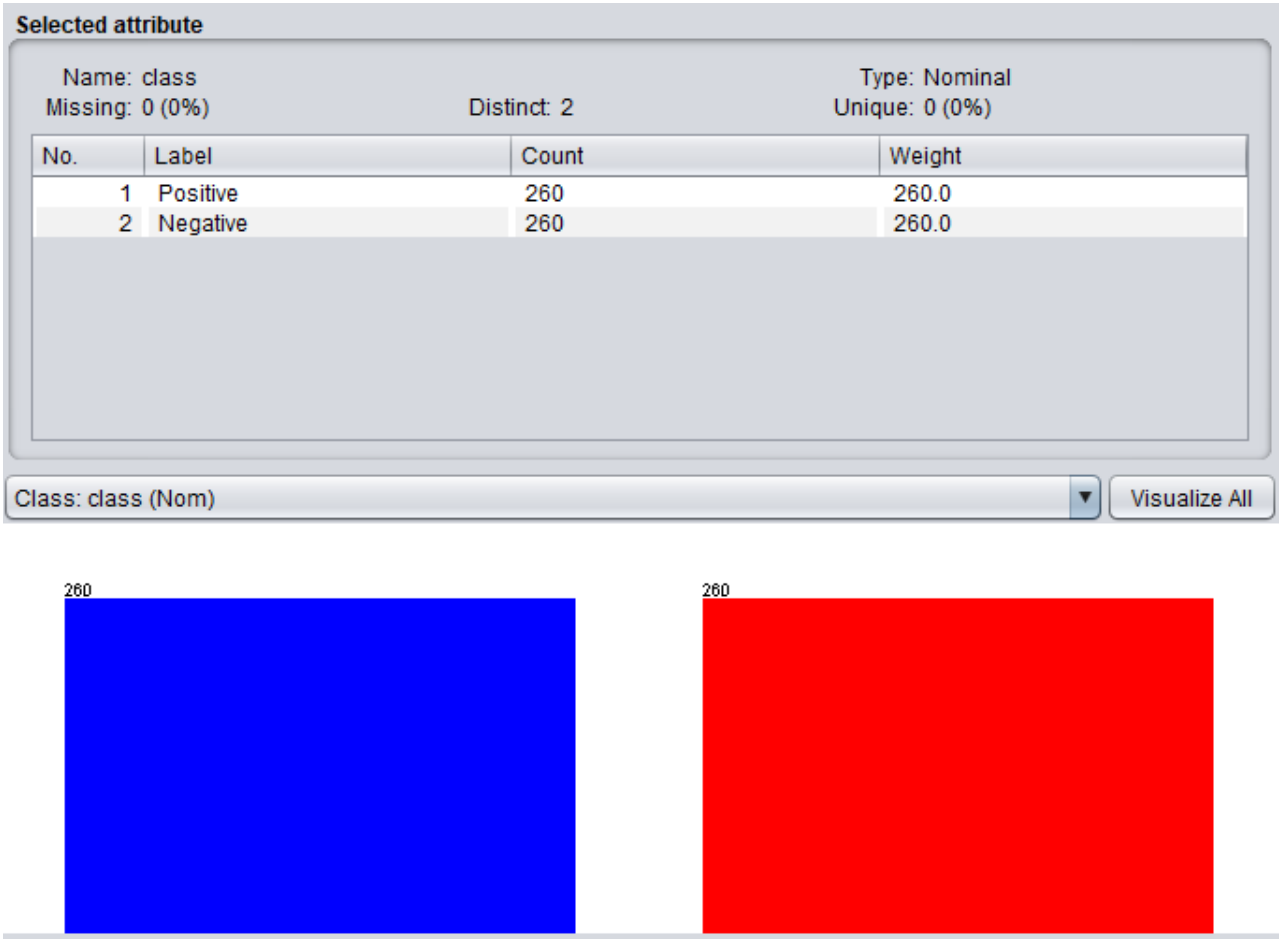

FIGURE 7. Dataset visualization after resampling (Random Over Sampling)

TABLE IV. Classification Test Results with Random Over Sampling Technique

Algorithm	Accuracy	Kappa Statistic	MAE	Precision	Recall	ROC
J.48	99.2308%	0.9846	0.0093	0,992	0,992	0,994
KNN	99.4231%	0.9885	0.0065	0,994	0,994	0,996
SVM	96.5385%	0.9308	0.0346	0,966	0,965	0,965
Naïve Bayes	90,3846%	0.8077	0.0944	0,904	0,904	0,969
Random Forest	98.0769%	0.9615	0.0418	0,981	0,981	0,999

Table 4 explains that after passing the data balancing process using the random over sampling technique, there was an increase in the accuracy of all classification algorithms. Having tested these five algorithms, the best algorithm with the highest accuracy is acquired by using KNN algorithm with accuracy rate of 99.4231%. Followed by the results of the J.48 algorithm with an accuracy rate of 99.2308%, Random Forest with an accuracy rate of 98.0769%, SVM with an accuracy rate of 96.5385% and Naïve Bayes with an accuracy rate of 90.3846%.

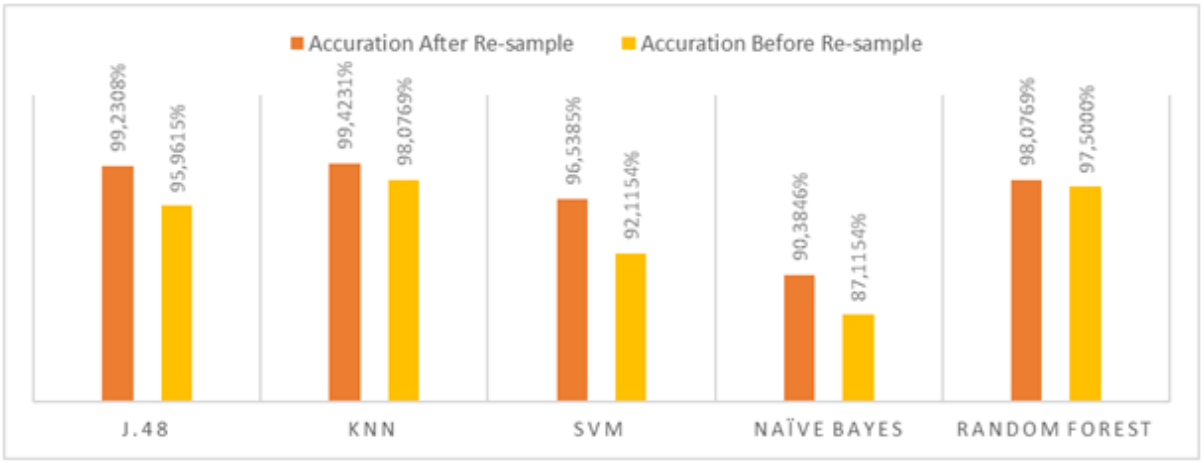

FIGURE 8. Visualization of the comparison of algorithm accuracy results

CONCLUSION

This study uses a resampling technique (random over sampling) on preprocessing data by testing datasets from several classification algorithms including J.48, Support Vector Machine, Naive Bayes, K-Nearest Neighbor and Random Forest. According to the result after testing and comparing all of the technique towards the diabetes dataset, the accuracy of the J.48 algorithm is 95.9615%, the SVM algorithm is 92.1154%, the Naïve Bayes algorithm of 87.1154%, the Random Forest algorithm is 97.5% and the classification algorithm with the best accuracy level is KNN. It acquired 98.0769% of accuracy. Meanwhile, from the results of classification testing using resample techniques on diabetes datasets by using the Decision Tree J.48, Random Forest, Naive Bayes, KNN and SVM algorithms, it is obvious that there is an escalation in the accuracy value of each algorithm, where the best accuracy result is the KNN algorithm with an accuracy value of 99.4231%, the kappa value is 0.9885, the MAE value is 0.0065, the Precision value is 0.994, the Recall value is 0.994 and the ROC value is 0.996.

REFERENCES

1. International Diabetes Federation, 2019, IDF Diabetes Atlas 9th edition.
2. International Diabetes Federation, 2021, What is diabetes. [Online]. Available: https://idf.org/aboutdiabetes/what-is-diabetes.html. [Accessed: 29-Jun-2021].
3. International Diabetes Federation, 2021, Demographic and geographic outline. [Online]. Available: https://www.diabetesatlas.org/en/sections/demographic-and-geographic-outline.html. [Accessed: 29-Jun-2021].
4. Islam M M F Ferdousi R Rahman S and Bushra H Y, 2020 Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques Adv. Intell. Syst. Comput. 992 p. 113–125.
5. Padurariu C and Breaban M E, 2019 Dealing with data imbalance in text classification Procedia Comput. Sci. 159 p. 736–745.
6. Li M Xiong A Wang L Deng S and Ye J, 2020 ACO Resampling: Enhancing the performance of oversampling methods for class imbalance classification Knowledge-Based Syst. 196, xxxx p. 105818.
7. Yildirim P, 2016 Pattern Classification with Imbalanced and Multiclass Data for the Prediction of Albendazole Adverse Event Outcomes Procedia Comput. Sci. 83, Dmdms p. 1013–1018.
8. Ren F Cao P Li W Zhao D and Zaiane O, 2017 Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm Comput. Med. Imaging Graph. 55 p. 54–67.
9. Jian C Gao J and Ao Y, 2016 A new sampling method for classifying imbalanced data based on support vector machine ensemble Neurocomputing 193 p. 115–122.
10. Brownlee J, 2020, Random Oversampling and Undersampling for Imbalanced Classification, Machine Learning Mastery. [Online]. Available: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/. [Accessed: 29-Jun-2021].
11. Branco P Torgo L and Ribeiro R, 2015, A Survey of Predictive Modelling under Imbalanced Distributions. .
12. He, Haibo; Ma Y, 2013 Imbalanced Learning: Foundations, Algorithms, and Applications 1st Edition 1st ed. Wiley-IEEE Press.
13. Shah K Patel H Sanghvi D and Shah M, 2020 A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification Augment. Hum. Res. 5, 1.
14. Zhao Y Otto S K Brandt N Selzer M and Nestler B, 2020 Application of random forests in TOF-SIMS data Procedia Comput. Sci. 176 p. 410–419.
15. R. Aryanti, A. Saryoko, A. Junaidi, S. Marlina, Wahyudin, and L. Nurmalia, "Comparing Classification Algorithm with Genetic Algorithm in Public Transport Analysis," J. Phys. Conf. Ser., vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012017.