Implementation of Winsorizing and random oversampling on data containing outliers and unbalanced data with the random forest classification method

Article *in* Jurnal Natural · June 2022

DOI: 10.24815/jn.v22i2.25499

CITATIONS
3

READS
171

3 authors:

Fahrezal Zubedi
Universitas Negeri Gorontalo
13 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Bagus Sartono
Bogor Agricultural University
134 PUBLICATIONS **433** CITATIONS
SEE PROFILE

Khairil Anwar Notodiputro
Bogor Agricultural University
144 PUBLICATIONS **355** CITATIONS
SEE PROFILE

Jurnal Natural
Vol. 22, (2) 2022
DOI 10.24815/jn.v22i2.25499
pISSN 1411-8513
eISSN 2541-4062
ORIGINAL RESEARCH

Implementation of winsorizing and random oversampling on data containing outliers and unbalanced data with the random forest classification method

FAHREZAL ZUBEDI[1,2*], BAGUS SARTONO[1], KHAIRIL ANWAR NOTODIPUTRO[1]

[1]Department of Statistics, IPB University, Bogor, Indonesia
[2]Statistics Study Program, Universitas Negeri Gorontalo, Gorontalo, Indonesia

Abstract. Many researchers conduct research using the classification method to find out the best method for predicting the class of an observation. Some of these studies explain that random forest is the best method. However, classifying data containing outliers and unbalanced data is a complicated problem. Many researchers are also conducting research to deal with these problems. In this study, we propose a winsorizing to deal with outliers by replacing the outlier values with the upper and lower limit values obtained from the interquartile range method and random oversampling to balance the data. It is also known that cases of the Human Development Index (HDI) in regencies/cities in eastern Indonesia vary widely, so cases of HDI in these areas can be used as case studies of data containing outliers and unbalanced data. This study aimed to compare the performance of the random forest before and after the data were applied to the winsorizing and random oversampling to predict HDI in districts/cities in eastern Indonesia. Classification method random forest after handling data containing outliers and unbalanced data has better performance in terms of accuracy and kappa values, 96.43% and 93.41%, respectively. The variables of expenditure per capita and the mean years of schooling are the most important.

Keywords: Winsorizing, Random oversampling, Interquartile range, Random forest, HDI

INTRODUCTION

Classification is a multivariate technique used to separate and allocate observations into pre-defined groups [1]. Classifying data that contains outliers is a problem in Statistics Machine Learning. Outliers occur because some data samples may have very different characteristics from others belonging to the same class and thus be far from the mass of data in that class. Unbalanced data is also a problem in Statistics Machine Learning. For unbalanced data, the majority class is represented by a large percentage of all observations. On the other hand, the minority class has relatively rare observations. Unbalanced data can cause misclassification in minority classes [2].

*Corresponding Author:
zubedifahrezal@apps.ipb.ac.id

Received: March 2022| Revised: May 2022 | Accepted: June 2022

Various techniques have been proposed to deal with outliers and unbalanced data sets. There are three ways to treat outliers: keep the outlier and treat it like any other data point, modify its value to be closer to the other sample values (winsorizing), and eliminate it from the sample (trimming) [3]. Data that contains outliers cause low accuracy, so data outliers must be handled. For example, to deal with outliers, some researchers identify and remove them completely [4]. Zhu, C. et al. conducted a study to classify Diabetes Mellitus data containing Outliers using the K-means Clustering algorithm to detect Outliers by grouping the data. Observations that are incorrectly clustered are considered outliers and are deleted [5]. Data Outliers are still a research subject that should still be used, so the outliers need to be handled, for example, by transforming the data [6]. Winsorizing is a statistical transformation by replacing outlier values in the data to reduce the effect of possible outliers. The advantage of winsorizing is that it retains the information that is the highest (or lowest) value in the distribution but protects against some of the

harmful effects of outliers. Winsorizing is known as robust statistics [7]. To address imperfections such as class ratio imbalances this is most often achieved by undersampling or oversampling as needed [8].

This study provides a major contribution, namely a combination of techniques for handling data containing outliers and unbalanced data to improve classification accuracy. The proposed approach is to identify outliers in each explanatory variable using the Interquartile Range (IQR) and then replace the data containing outliers with the technique Winsorizing before the data is divided into training and testing. The interquartile range is an important tool for describing the spread given by a data set. This is usually used when the shape of the distribution is skewed, or there are outliers. The interquartile range is easily calculated for a sample of size n, as is the population's interquartile range for a continuous probability distribution [9]. After that, apply Random Oversampling to get a balanced dataset on the training. Random oversampling is a non-heuristic algorithm. The main goal is to balance the distribution of classes through random repetition of the minority target class. In [10], a study on the comparison of undersampling and oversampling methods in the case of unbalanced data to predict and identify which customers will make transactions in the future regardless of the amount of money transacted. Technique random oversampling was better than undersampling in predicting the case based on evaluation metrics. In [11], a comparative study of undersampling and oversampling methods on empirical data, the results show that oversampling methods give excellent results in practice, especially Random Oversampling can improve classification accuracy in the minority class by increasing the sensitivity value and AUC value. In [12], a comparative study of the methods of oversampling in the data of Poor Households in Yogyakarta with the KNN classifier, the results show that the random oversampling method provides the highest accuracy value compared to other oversampling methods in handling unbalanced data. In the classification stage, Random Forest is used. Random Forest is an ensemble method, where the ensemble method is a way to improve the accuracy of the classification method by combining classification methods [13]. The advantage of the random forest compared to other classification methods, namely neural networks, is in terms of computation time. The random forest has fast computation time [14].

A comparative measure of the success of human development which includes three dimensions of human life is the definition of the Human Development Index (HDI). The three dimensions are life opportunity, knowledge, and a decent life. The HDI in eastern Indonesia is still far behind when compared to the HDI in western Indonesia. The government's determination of development programs must be right on target and by regional priorities based on the HDI category owned by the region because it affects the high and low HDI scores. To assist the effectiveness of the government's performance in analyzing the HDI category for each Regency/City in Eastern Indonesia, a decision system is needed that can determine the classification of HDI categories quickly and accurately [15]. This study aims to compare the performance of the random forest before and after handling data containing outliers and unbalanced data to predict HDI in districts/cities in Eastern Indonesia.

METHODOLOGY

Data

The data used is from Badan Pusat Statistik Publication Data at www.bps.go.id/indikator/. The data is Human Development Index (HDI) for districts/cities in Eastern Indonesia in 2020. The data has 7 explanatory variables with a total of 185 observations. The data used in this study contains outliers. This outliers data will be handled by the winsorizing technique, replacing the outlier data with the upper and lower limit values obtained by the Interquartile Range method. The variables used in this study can be seen in Table 1.

Table 1. Response variable and explanatory variables in the study

Variables	Type
Category HDI (**Y**)	Categoric
Expenditure per capita ($\mathbf{X_1}$)	Numeric
Expected years of schooling ($\mathbf{X_2}$)	Numeric
Life expectancy ($\mathbf{X_3}$)	Numeric
Mean years of schooling ($\mathbf{X_4}$)	Numeric
Number of Poor Population ($\mathbf{X_5}$)	Numeric
Open Unemployment rate ($\mathbf{X_6}$)	Numeric
Labor force participation rate ($\mathbf{X_7}$)	Numeric

Model

Classification modeling process using random forest. Random forest is a development of the CART (Classification and Regression Tree) method by applying the bootstrap aggregating (bagging) and random feature selection methods. The following are the stages of the random forest algorithm [16].

i. Random sampling is performed with the recovery of size n from the data cluster. This stage is called the bootstrap stage.
ii. The tree is built to its maximum size (without pruning) using the bootstrap example. The tree construction is done by applying random feature selection to each selection process, where m explanatory variables are selected randomly with $m < q$, and the best sorter is selected based on these m explanatory variables. This stage is known as the random sub-setting stage.
iii. Steps (i) and (ii) were carried out L times to obtain L decision trees.

Random forest predicts the observation response by using all the prediction results of the decision tree. In the case of classification, the majority vote technique is used to determine the prediction results, namely the category that most often appears as a result of predictions from the classification tree. The problem with CART is that it uses an algorithm greedy, which minimizes the error in choosing which variable to use as a separator. This causes a lot of similarity in the resulting tree structure, leading to a high correlation between the explanatory variables. Combining predictions from various models in an ensemble will work well if the predictions from the submodels are unrelated or at least very weakly related. The random forest changes the algorithm so that the predicted results from all subtrees have a small correlation [17].

Random forest modeling also produces important variables derived from the Mean Decrease Gini (MDG). The definition of MDG is the average value of reduced impurities that occur during the sorting process in the formation of a single classification tree in the random forest algorithm. As for determining the MDG, the number of observations of the terminal node in the classification tree from a random forest is limited to 50 observations. If a node has reached 50 observations, the sorting will stop [16]. The calculation of the MDG value is:

$$MDG = \frac{1}{m}\sum_{t}[\Delta i(s,t)I(s,t)] \qquad (1)$$

where m is the number of trees formed, $\Delta i(s,t)$ is the reduced impurity value for the explanatory variable X_s at node t and $I(s,t)$ is an indicator function that has a value of 1 when X_s selects node t and 0 other [18].

In comparing classification methods, performance evaluation is carried out as a measuring tool to determine the best classification method. The evaluation process is carried out by comparing the prediction results on the test data to the original data in a confusion matrix. The results of the classification in 4 classes or categories will form a confusion matrix as follows:

Table 2. Confusion matrix for 4 classes or categories

		Actual			
		A	**B**	**C**	**D**
Prediction	**A**	AA			
	B	Ab	BB		
	C	Ac		CC	
	D	Ad			DD

Accuracy and Kappa are generally used to measure the performance of classification methods. Accuracy assesses how effective a classification method is in predicting the true class of each observation. The accuracy value can be calculated as follows [19]:

$$\text{Accuracy} = \frac{\text{AA} + \text{BB} + \text{CC} + \text{DD}}{\text{the number of data}} \qquad (2)$$

kappa or kappa statistics is a measure that compares the observed accuracy with the expected accuracy (random probability). kappa statistics usually range in value from 0 to 1 [5].

Performance measures based on accuracy and kappa actually describe the classification accuracy. As an alternative, it is necessary to determine the value of sensitivity, specificity, and accuracy balanced for each class. Sensitivity is the ability of the model to correctly identify observations that have a positive class (true positive) to all observations that are part of the positive class [20]. The sensitivity value of class A can be calculated as follows

$$\text{Sensitivity} = \frac{\text{AA}}{(\text{AA} + \text{Ab} + \text{Ac} + \text{Ad})} \qquad (3)$$

Specificity is the ability of the model to correctly identify observations that have a negative class (true negatives) against all observations that are part of the negative class [21]. Accuracy balanced is the average accuracy obtained from the class or can be defined as the average of sensitivity and specificity [22].

Data Analysis Procedure

1. Doing data exploration
   Data exploration is used to obtain an overview of the variables to be analyzed.
2. Doing random forest classification modeling with data handling

a) Detecting outliers in the dataset in each explanatory variable (X) using the Interquartile Range (IQR). The interquartile range is a data preprocessing technique used

to find outliers and extreme values. It measures dispersion by dividing a data set sorted by rank into four equal parts, called quartiles. The values dividing each section are denoted by Q1, Q2, and Q3, where Q1 and Q3 are the median values in the first and second half of the data set ordered by rank, respectively, and Q2 is the median value across the entire set. The IQR value is Q3 - Q1. Outliers here are data that are below Q1-1.5 x IQR and above Q3+1.5 x IQR [2].

b) Handling Outlier data with Winsorizing. Winsorizing is a statistical transformation by replacing outlier values in the data to reduce the effect of possible outlier data [7]. We replace the outlier values above the upper limit with the upper limit value and the outlier values below the lower limit with the lower limit value. Determination of the upper and lower limits using IQR.

c) Divide data into training data and testing data using k-fold cross-validation. The data used in the model formation process is referred to as training data, and the data used to validate the model is referred to as testing data. A dataset is divided into k subsets with the same number of k-fold cross-validation. Each subset will take turns being testing. The calculation of the average of all accuracy results for all folds is a cross-validation assessment of the overall model accuracy [23]. Meanwhile, in this study, the data is divided into 5 folds or 5 subsets. The model in the classification will be tested 5 times using 5 tests obtained (50 replications).

d) Handling unbalanced data by random oversampling on training data.

e) Performing classification modeling based on the stages of the random forest algorithm described previously and then generating the important variables obtained from the calculation of the average reduced impurity of each explanatory variable or called the Mean Decrease Gini (MDG) when used as a sorter in each single classification tree that is formed so that the variable that is most often sorted can be seen with the highest MDG value.

f) Calculating the value of accuracy, kappa, sensitivity, specificity, and accuracy balanced the model on testing data after handling data containing outliers and unbalanced by applying a 5-fold cross-validation which was repeated 50 times.

3. Doing random forest classification modeling without handling the data.

a) Divide data into training data and testing data using k-fold cross-validation. The data is divided into 5 folds or 5 subsets. The model in the classification will be tested 5 times using 5 tests obtained (50 replications).

b) Performing classification modeling based on the stages of the random forest algorithm described previously and then generating. The important variables are obtained from the calculation of the average reduced impurity of each explanatory variable or called the Mean Decrease Gini (MDG).

c) Calculating the value of accuracy, kappa, sensitivity, specificity, and accuracy balanced the model on testing data by applying 5-fold cross-validation, which was repeated 50 times

4. Comparing the performance of the random forest classification method before and after handling data containing outliers and unbalanced data based on the values of accuracy, kappa, sensitivity, specificity, and accuracy balanced.

5. Interpreting the comparison results to state the best model based on the performance indicated by the classification method.

RESULTS AND DISCUSSION

Exploration Data

Data on HDI in regencies/cities in Eastern Indonesia has different amounts, so categorizing response classes into low, medium, high, and very high HDI is determined based on the United Nations Development Program (UNDP) criteria. The four categories of the HDI scale are as follows [15].

Low : HDI < 60,0
Medium : 60 ≤ HDI <70
High : 70 ≤ HDI <80
Very high : HDI ≥80

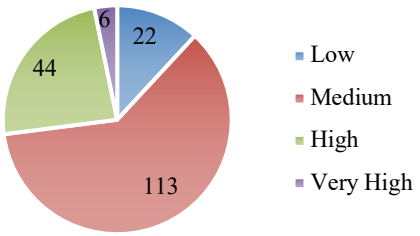

Figure 1. Classification of HDI Data

Based on Figure 1, the HDI in the medium category is the largest number, namely 113 regencies/cities in eastern Indonesia in 2020, and the HDI in the very high category is the least, namely 6 regencies/cities in Eastern Indonesia in 2020, the HDI in the high category

as many as 44 regencies/cities and low category HDI as many as 22 regencies/cities in Eastern Indonesia in 2020.

Data exploration of the explanatory variable is shown in Table 3. Each variable has different units, as shown in table 3. Each explanatory variable has very different maximum and minimum values. Based on the statement, this data may have outliers. The mean and median of the expenditure per capita variables and the number of poor population have different values. Based on this statement, there is also the possibility of outliers in the data, which can be investigated further with a boxplot.

Figure 2 presents a boxplot for each explanatory variable. The flat lines above and below the box indicate the upper and lower bounds, and the circles indicate the outliers.

Random Forest Classification Modeling by handling data containing outliers and unbalanced

For handling outlier data, using the winsorizing technique, which means replacing the outlier data with a value, the upper limit if the outlier is above the upper limit and the lower limit value if the outlier is below the lower limit.

Determination of the upper and lower limits using the Interquartile Range (IQR) method. The following is an illustration of the detection and handling of outliers on the expenditure per capita variable. The value of Q1, Q2, and Q3 are 12109, 13625, and 15301, respectively. The value of IQR is 3192, obtained from Q3 - Q1 = 3192. Lower limit = Q1-1.5× IQR = 12109 – 4788= 7321. Upper limit = Q3 + 1.5 × IQR = 10513+ 4788 = 15301. Identified that 16873, 17503, and 19723, which are above the upper limit, are outliers, so these values are replaced with upper limits to eliminate outliers. In the same way, obtained outliers on other variables and how to handle them.

Random Oversampling (ROs) was carried out in three stages because there were three categories of response variables which were minor classes. The three minor categories are low, high, and very high HDI. The proportion of data before and after random oversampling is shown in Table 4.

Table 3. Exploration of data on the explanatory variables

Explanatory Variables	Mean	Median	Minimal	Maximum
Expenditure per capita ($\mathbf{X_1}$)	9047	8837	3975	19723
Expected years of schooling ($\mathbf{X_2}$)	12.69	12.81	3.61	16.62
Life expectancy ($\mathbf{X_3}$)	67.57	67.58	55.27	75.10
Mean years of schooling ($\mathbf{X_4}$)	7.978	803	1.13	12.20
Number of Poor Population ($\mathbf{X_5}$)	30.08	23.03	3.89	183.84
Open Unemployment rate ($\mathbf{X_6}$)	4.614	4.17	0.21	15.92
Labor force participation rate ($\mathbf{X_7}$)	69.93	69.62	36.65	96.25

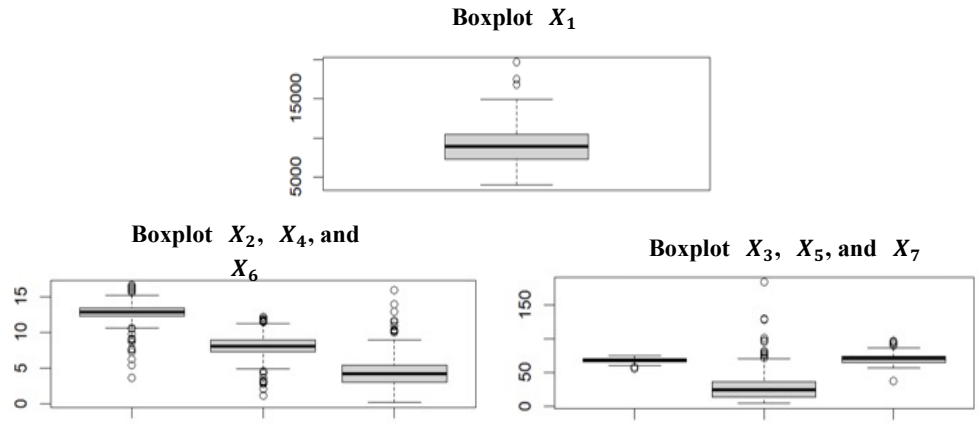

Figure 2. Boxplot of each explanatory variables

Table 4. Percentage of training data before and after ROs

Category HDI	Before ROs	After ROs
Low	11.62%	24.75%
Medium	61.25%	25.74%
High	24.03%	25.08%
Very high	3.10%	24.43%

Based on Table 4, the percentage of data before and after ROs showed a significant difference Data that has been processed with ROs shows a more balanced proportion when compared to before being processed with ROs.

After handling the unbalanced data with random oversampling, the random forest classification modeling is then applied. The number of trees and a subset of variables used remains the same, namely, as many 500 trees and two subsets of variables. The model that is formed is then tested on the testing data. The test results on the testing data form a confusion matrix, as presented in Table 5.

Table 5. Confusion matrix of random forest with the handling of Data

		Actual			
		L	M	H	VH
Prediction	L	7	0	0	0
	M	0	34	1	0
	H	0	0	12	1
	VH	0	0	0	1

Note: L: Low; M:Medium; H:High; VH:Very high

The random forest classification method correctly predicted 7 districts/cities with low categories and 34 districts/cities with medium categories. However, districts/cities with high and very high categories still do not predict correctly. It can be seen that 1 of each category is the result of an incorrect prediction.

The most important explanatory variables in modeling random forest with the handling of data based on the level of importance of the variable (variable importance) are shown in Figure 4.

Based on Figure 4, the highest value, Mean Decrease Gini (MDG), was obtained by the expenditure per capita variable. Several other variables have the highest MDG values after expenditure per capita, namely the mean years of schooling, life expectancy, expected years of schooling, open unemployment rate, labor force participation rate, and the number of the poor population. Expenditure per capita is the most important variable in this study.

Random Forest Classification Modeling without Handling of Data

The process classification modeling of random forest used 500 classification trees and two subsets of variables in the training data. The model that is formed is then tested on the testing data. The test results on the testing data form a confusion matrix, as presented in Table 6.

Table 6. Confusion matrix of random forest with the handling of Data

		Actual			
		L	M	H	VH
Prediction	L	6	0	0	0
	M	1	31	2	0
	H	0	3	11	2
	VH	0	0	0	0

Note: L: Low; M:Medium; H:High; VH:Very high

The random forest method classification without applying the winsorizing technique and the random oversampling method has not

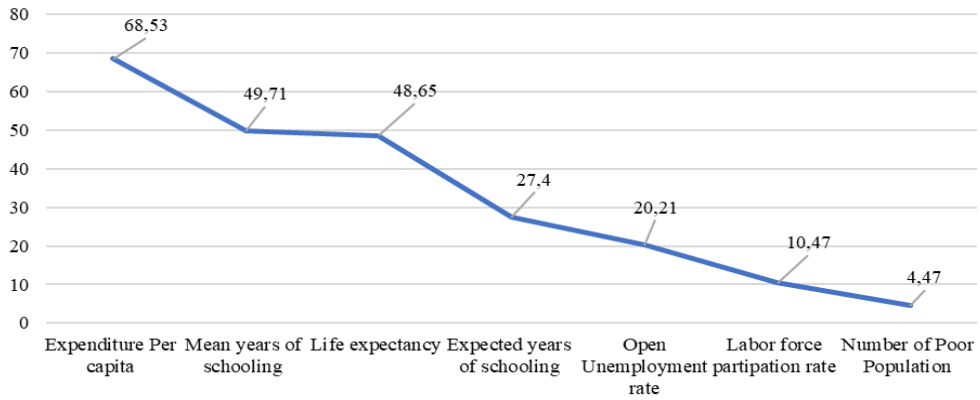

Figure 3. The level of importance of the variables on the Random Forest method with the handling of the Data

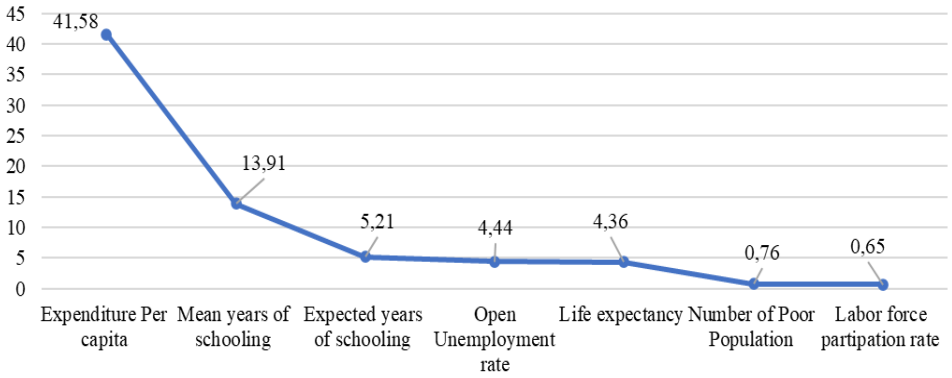

Figure 4. The level of importance of the variables on Random Forest without handling the data

shown good performance in predicting. It can be seen that only 6 of 7 regencies/cities in the low category, 31 of 34 regencies/cities in the medium category, 11 of 13 regencies/cities in the high category, and 0 of 2 regencies/cities in the very high category were successfully predicted correctly by the random forest method without applying the winsorizing technique and the random oversampling method. The most important explanatory variables in modeling random forest without applying the winsorizing technique and the random oversampling method based on the level of importance of the variable (variable importance) are shown in Figure 4.

Based on Figure 4, the highest mean decrease Gini (MDG) was obtained by variable expenditure per capita. Several other variables have the highest MDG values after expenditure per capita, namely mean years of schooling, expected years of schooling, open unemployment rate, life expectancy, number of poor population, and labor force participation rate.

Comparison of Classification Performance

The classification method formed can be evaluated for its performance by calculating the accuracy and kappa values. In addition, each class can be evaluated with sensitivity, specificity, and accuracy balanced. These values can compare how well the classification method predicts the class of observations on the testing data. Figure 5 shows the accuracy and kappa values of the classification. Accuracy and kappa values in the random forest after handling the data with winsorizing and random oversampling showed an increasing number. This shows that the performance of the classification method that has been applied to winsorizing and random oversampling is better before handling the data. In general, the accuracy value shows a very high number above 80%. This performance measure is not enough to describe the accuracy of the classification. As an alternative, the sensitivity, specificity, and accuracy balanced values for each class are presented in Table 7 below by applying a 5-fold cross-validation repeated 50 times.

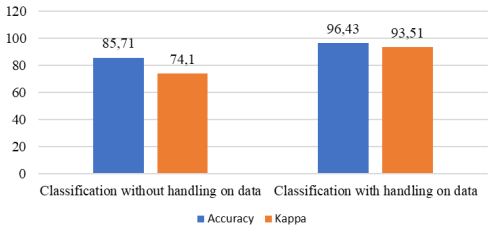

Figure 5. Comparison of classification performance based on the value of accuracy and kappa

The sensitivity value shows an increasing number in each class or category in the random forest method after handling the data with winsorizing and random oversampling. This shows the performance of the classification method after handling the data is better than before handling. However, the ability of the random forest method classification after being applied to data handling with winsorizing and random oversampling techniques in predicting HDI in districts/cities in very high class is only 50%. The specificity value showed an increasing number in the medium and high categories. The accuracy balanced of the random forest method after handling the data using the winsorizing and random oversampling also increases with the increased sensitivity value.

Table 7. Values of sensitivity, specificity, and accuracy balanced for each class

	Classification before handling data (%)				Classification after handling data (%)			
	L	M	H	VH	L	M	H	VH
Sensitivity	85.71	91.18	86.42	0	100	100	92,31	50
Specificity	100	86.36	88.37	100	100	95.45	97.67	100
Accuracy balanced	92.86	88.77	86.49	50	100	97.73	95	75

Note: L: Low; M:Medium; H:High; VH:Very high

CONCLUSION

The value of accuracy and kappa in the random forest after data handling with winsorizing and random oversampling showed an increasing number of 96.43% and 93.51%, respectively. The Value of sensitivity and accuracy balanced show an increasing number in each class or category in the random forest after handling the data with winsorizing and random oversampling. The specificity value showed an increasing number in the medium and high categories. This shows that the classification method random forest after handling outliers and unbalanced data with winsorizing and random oversampling is better for predicting Data Human Development Index in districts/cities in eastern Indonesia. The two models formed of them show expenditure per capita and mean years of schooling as the most important variables.

ACKNOWLEDGMENT

The authors would like to thank the Editor and Reviewers for helpful suggestions to improve the quality of this manuscript.

REFERENCES

[1] Johnson, R. A.; Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis 6th ed*. (London : Pearson Education)

[2] Nnamoko, N.; Korkontzelos, I. 2020. Efficient treatment of outliers and class imbalance for diabetes prediction. *Artif. Intell. Med.* **104** 1–12.

[3] Ghosh, D.; Vogt, A. 2012. Outliers: An Evaluation of Methodologies. *Proc. Am. Stat. Assoc.* **2012** 3455-3460.

[4] Ferdowsi, H.; Jagannathan, S.; Zawodniok, M. 2014. An online outlier identification and removal scheme for improving fault detection performance. *IEEE Trans. Neural Netw. Learn. Syst.* **25** 908-919 DOI : 10.1109/TNNLS.2013.2283456.

[5] Zhu, C.; Idemudia, C. U.; Feng, W. 2019. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Inform. Med. Unlocked* **17** 100179 DOI : 10.1016/j.imu.2019.100179.

[6] Xu, X.; Liu, H.; Li, L.; Yao, M. 2018. A comparison of outlier detection techniques for high-dimensional data. *Int. J. Comput. Intell. Syst.* **11** 652–662.

[7] Reifman, A.; Garrett, K. 2010. Winsorize. In: N J Salkind, editors.. *Encyclopedia of Research Design*. Thousand Oaks (CA): Sage Publishing. p. 1636–1638.

[8] Skryjomski, P.; Krawczyk, B. 2017. Influence of minority class instance types on SMOTE imbalanced data oversampling. *Proc. Mach. Learn. Res*. **74** 7–21.

[9] Whaley, D. L. 2005. *The Interquartile Range: Theory and Estimation* [Master's Thesis, East Tennessee State University]. Electronic Theses and Dissertations. https://dc.etsu.edu/etd/1030

[10] Mohammed, R.; Rawashdeh, J.; Abdullah, M. 2020. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *Int. Conf. Inf. Commun. Syst*. **2020** 243-248 DOI : 10.1109/ICICS49469.2020.239556.

[11] Batista, G. E. A. P. A.; Pati, R. C.; Monard, M. C. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. News*. **6** 20-29.

[12] Santoso, B.; Wijayanto, H.; Notodiputro, K. A.; Sartono, B. 2018. A comparative study of synthetic over-sampling method to improve the classification of poor households in yogyakarta province. *IOP Conf. Ser. Earth Environ. Sci.* **187** 012048.

[13] Han, J.; Micheline, K.; Pei, J. 2014. *Data mining concepts and techniques 3rd ed*. (USA : Morgan Kaufmann)

[14] Pratiwi, A.; Notodiputro, K. A.; Wijayanto, H. 2018. Pemodelan Loyalitas Konsumen Susu Pertumbuhan dalam Mengikuti Program Rewards Menggunakan Metode Random Forest dan Neural Network. *Xplore: J. Stat*. **2** 41-48 DOI : 10.29244/xplore.v2i2.104.

[15] BPS. 2014. *Indeks Pembangunan Manusia 2013*. (Jakarta : Badan Pusat Statistik RI)

[16] Breiman, L. 2001. Random Forests. *Mach. Learn.* **45** 5–32 DOI : 10.1023/A:1010950718922.

[17] Triscowati, D. W.; Sartono, B.; Kurnia, A.; Domiri, D. D.; Wijayanto, A. W. 2019. Classification of Rice-Plant Growth Phase using Supervised Random Forest Method Based On Landsat-8 Multitemporal Data. *Int. J. Remote Sens. Earth Sci.* **16** 1–11.

[18] Sandri, M.; Zuccolotto, P. 2006. Variable Selection Using Random Forests. In: Zani S. Cerioli A. Riani M. Vichi M, editors. *Data Analysis, Classification and the Forward Search. Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin: Springer Publishing. p. 263–270.

[19] Belouafa, S.; Habti, F.; Benhar, S.; Belafkih, B.; Tayane, S.; Hamdouch, S.; Bennamara, A.; Abourriche, A. 2017. Statistical tools and approaches to validate analytical methods: Methodology and practical examples. *Int. J. Metrol. Qual. Eng.* **8** 1-10 DOI : 10.1051/ijmqe/2016030.

[20] Sokolova, M.; Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag*. **45** 427-437 DOI : 10.1016/j.ipm.2009.03.002.

[21] Trevethan, R. 2017. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Front. Public Health* **5** 1-7.

[22] Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. 2010. The balanced accuracy and its posterior distribution. *Proc. Int. Conf. Pattern Recognit.* **2010** 3121-3124 DOI : 10.1109/ICPR.2010.764.

[23] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. Int. Joint Conf. Artif. Intell*. **2** 1137–1143.