TEKNOKOM : Jurnal Teknologi dan Rekayasa Sistem Komputer
Vol. 4, No. 2, September 2021, pp. 48 - 52
ISSN: 2686-3219, DOI: 10.31943/teknokom.v4i2.64

HOW MACHINE LEARNING METHOD PERFORMANCE FOR IMBALANCED DATA (Case Study: Classification of Working Status of Banten Province)

Pardomuan Robinson Sihombing[1*]
[1]Directorate National Account, BPS-Statistics, Indonesia

ARTICLEINFO
History of the article:
Received July 1, 2021
Revised July 11, 2021
Accepted July 12, 2021
Published July 13, 2021

Keywords:
Machine learning
Predictive
Resample
Sensitivity
Specificity

ABSTRACT

This study will examine the application of several classification methods to machine learning models by taking into account the case of imbalanced data. The research was conducted on a case study of classification modeling for working status in Banten Province in 2020. The data used comes from the National Labor Force Survey, Statistics Indonesia. The machine learning methods used are Classification and Regression Tree (CART), Naïve Bayes, Random Forest, Rotation Forest, Support Vector Machine (SVM), Neural Network Analysis, One Rule (OneR), and Boosting. Classification modeling using resample techniques in cases of imbalanced data and large data sets is proven to improve classification accuracy, especially for minority classes, which can be seen from the sensitivity and specificity values that are more balanced than the original data (without treatment). Furthermore, the eight classification models tested shows that the Boost model provides the best performance based on the highest sensitivity, specificity, G-mean, and kappa coefficient values. The most important/most influential variables in the classification of working status are marital status, education, and age.

Correspondece:
Pardomuan Robinson Sihombing,
Directorate National Account,
BPS-Statistics, Indonesia,
Email : robinson@bps.go.id

INTRODUCTION

Predictive analytics is one of the analytical methods often used in addition to descriptive and prescriptive analytics. Predictive models study the relationships between variables and then create a statistical model to predict the value of new events and future events. One of the predictive models is a classification technique other than the regression model. Along with the development of science and technology, classification models are growing. Several programs, especially in R software, continue to be developed to produce classification methods with good performance.

Some of the classification techniques that are often used include Classification and Regression Tree (CART), Naïve Bayes, Random Forest, Rotation Forest, Support Vector Machine (SVM), Analysis Neural Network (ANN), OneR, and Boosting. Each method has advantages and disadvantages. ANN has the advantage of acquiring knowledge even though there is no certainty, having fault tolerance, and the ability to calculate in parallel so that the process is shorter than others. The disadvantage of ANN is that it cannot perform numerical operations with high precision and less able to perform arithmetic algorithm operations, logical operations, and symbolic operations. One of the packages in R for the ANN method is the “nnet” package developed by Venables and Ripley [1].

Random Forest has the advantage of not being sensitive to data; there is no overfitting problem and can sort variables that contribute to predicting. The disadvantage of this method is that tree predictions must be uncorrelated and often appear as black boxes (error messages). One of the packages in R for the Random Forest method is the “randomForest” package developed by Liaw and Wiene [2]. The oneR method has the advantage of producing an accurate model to establish a good baseline, efficient in processing big data. The drawback of the one R model is that it is less efficient for complex models. One of the One R packages uses the “OneR” package developed by Jouanne [3].

Rotation Forest has the advantage of improving the predictive ability of the decision tree by utilizing the principal component principle and maintaining data diversity. One of the packages in R for the Rotation Forest method is the “rotationForest” package developed by Balling and Poel [4].

The CART method has the advantage of not requiring normalization or data scaling, data handling missing values, easy visualization. The drawbacks of the CART method are that it tends to be overfitting, sensitive to outliers, and less efficient for extensive data. One of the packages in R for the CART method is the “rpart” package developed by Therneau and Atkinson [5]. The SVM method has the advantage of performing well in classifying variables with high dimensions, such as image data, gene data, medical data. In addition, the SVM method is also not sensitive to outlier data. The disadvantage of the SVM method is that it is less efficient for larger data sets, so it takes much time. In addition, it requires expertise in selecting the appropriate hyper parameters and kernel functions so that the model's performance is good.

The Naïve Bayes method has advantages in time efficiency, which is very fast in data processing, can be scaled with large data sets, and can be used for multi-class predictions. The disadvantage of the Naive Bayes method is that feature independence does not apply: The basic assumption of Naive Bayes is that there is an assumption that the independent contribution between variables and training data must represent the population well. One of the packages in R for SVM and nave Bayes methods is the "e1071" package developed by Meyer et al. [6]. The Boosting method has the advantages of feature engineering, which is less required (no need for scaling, data normalization, can also handle missing values well), easy to interpret, suitable for large data, and efficient. Disadvantages of the Boosting method are difficult interpretation, complex visualization, and sometimes overfitting. One of the packages in R for the Boosting method is the “xgbost” package developed by Chen et al. [7].

In general, the assumption for classification method is based on that the data used has a balanced proportion. According to Maalouf and Siddiqi [8], one of the problems in data classification is a rare event or imbalanced data, namely the amount of data that is not balanced between different classes. One of the consequences of imbalanced data is that the classification results tend to eliminate opportunities from the minority class because the predicted value will tend to be in the majority category. The accuracy of the resulting classification is not good [9].

This study aims to compare the various existing classification methods. The case studies used are factors that affect a person's working status in Banten Province. In this study, a resampling technique was applied to overcome imbalanced data to improve the performance of the classification model used.

RESEARCH METHOD

Data Sources and Research Variables

The data used in this study came from the National Labor Force Survey of Banten Province for August 2020, which the BPS-Statistics Indonesia conducted [10]. The total sample used is 11,469 respondents, of which 10.2 percent are unemployed, the remaining 89.8 percent are working. The variables used in the study can be seen in Table 1.

Table 1. Research variable

Variable Name	Information	Scale
Working status	0 Working 1 Not Working	Nominal
Area type	0 Urban 1 Rural	Nominal
Gender	0 Women 1 Male	Nominal
Marital status	0 Not yet/ Not Married 1 Married 2 Divorce	Nominal
Age	0 15-25 years 1 26-50 years old 2 > 50 years	Nominal
Education	Not completed school Primary School Junior School High School University	ordinal
Course certificate	0 No 1 Yes	Nominal
Visual impairment	0 No 1 Yes	Nominal
Hearing impairment	0 No 1 Yes	Nominal
Walking impairment	0 No 1 Yes	Nominal
Holding impairment	0 No 1 Yes	Nominal
Speech impairment	0 No 1 Yes	Nominal
Other impairment	0 No 1 Yes	Nominal

Classification modeling

In this study, the data is divided into two: training data for model building and data testing to test model performance. The distribution of data is based on a deterministic/holdout method, namely by determining the ratio of the division of the two datasets, in this study using a ratio of 70 percent for training data and 30 percent for testing data. The resample technique was then carried out, using both/combine sampling methods from the existing testing data so that the model's performance could be compared on the treated and untreated data.

Classification Performance Evaluation

Evaluation is done by using data testing both on models that use treatment or those that do not. The confusion matrix is used as a classification performance measure. According to Han et al. [11], a confusion matrix is a valuable tool for analyzing how well or how accurately the classification method can recognize objects of observation from different classes. Table 2 is a confusion matrix for binary classification. The column section shows the actual label for each class, while the row section shows the class label based on the predicted results.

Table 2. Confusion matrix

Confusion Matrix		**Actual Class**		**Total**
		Yes	***No***	
Prediction Class	***Yes***	TP	FP	P'
	No	FN	TN	N'
Total		P	N	

Some of the classification performance measures that can be obtained from the confusion matrix are as in Equations (1) - (4).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

$$G - mean = \sqrt{Sensitivity\ x\ Specificity} \quad (4)$$

Besides using the confusion matrix, according to Landis and Koch [12], the model's goodness can also be seen by the value of the Kappa coefficient. The value of this test is also used to determine the strength of agreement/reliability. In this case, the higher the Kappa value, the better the model used.

Data Analysis Stages

The stages of data analysis in this study are as follows:

a. Divide the data into training data and testing data with proportions of 0.7 and 0.3 respectively
b. Resample the training data using the combine/both sampling technique
c. Perform classification modelling with 8 (eight) methods, namely Naive Bayes, ANN, SVM,
d. Random Forest, CART, OneR, Rotation Forest, and Boost
e. Conduct descriptive analysis of research variables
f. Evaluating the classification performance of the 8 (eight) machine learning methods used
g. Choose the best machine learning method to predict working status based on the balanced and highest specificity, sensitivity, G-mean, and Kappa coefficient criteria
h. Displays the predictor variables that have the most influence on the model
i. Data processing using R software version 4.1.0 and using the machine learning package mentioned in the introduction

RESULTS AND DISCUSSION

The first step before carrying out the classification method is to present descriptive statistics regarding the research variables. The research sample shows that as many as 10.2 percent of the population do not work in Banten Province. The unemployed population is dominated by people living in urban areas, unmarried/unmarried, male, and young (15-25 years old). Descriptive statistics for research variables in a thorough manner can be seen in Table 3.

Table 3. Descriptive Analysis of Research Variables

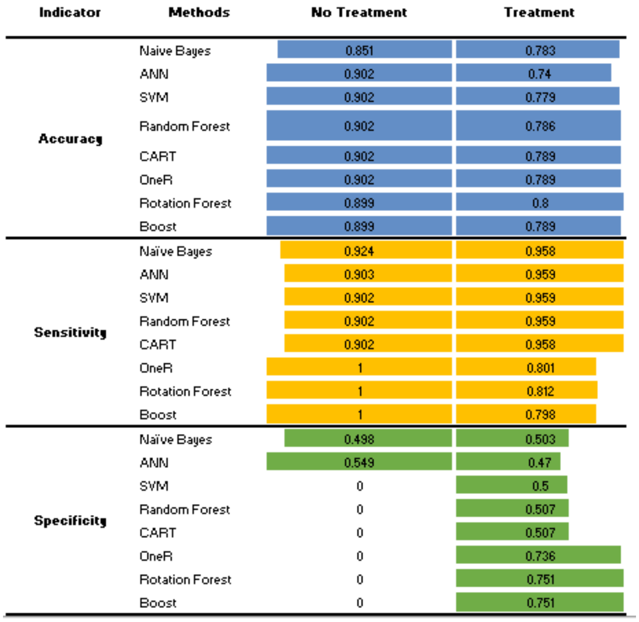
Variable	**Indicator**	**Work**	**Does not work**	**Total**
Area	rural	90.40	9.60	100
	urban	89.52	10.48	100
Marital status	Not yet/Not Married	71.85	28.15	100
	Marry	95.82	4.18	100
	Divorce	92.97	7.03	100
Gender	Male	89.2	10.8	100
	Woman	91.0	9.0	100
Education	Not completed	94.5	5.5	100
	Primary	94.6	5.4	100
	Junior High School	88.6	11.4	100
	Senior High School	84.4	15.6	100
	University	93.3	6.7	100
Age	15-25	69.9	30.1	100
	26-50	94.2	5.8	100
	>50	95.6	4.4	100
Course certificate	Not	89.9	10.1	100
	Yes	89.6	10.4	100

Next, there is a discussion about the performance of the classification method with machine learning techniques. Based on the information in the descriptive analysis, there is a relatively significant difference in the proportion between those who do not work and those who work. If you pay attention to the value of accuracy and specificity, then the modelling for imbalanced data, without treatment, the value is higher than the data that has used resample combine/both samplings. However, the performance of other classifications on classification modelling without treatment, such as specificity, G-mean, and Kappa coefficient, has a value of 0 in the SVM, Random Forest, CART, OneR, Rotation Forest, and Boost techniques. Meanwhile, for the Naïve Bayes and ANN models, the sensitivity and G-mean values are below 0.6, and Kappa values are below 0.3.

Without treatment on the eight models, classification modeling showed almost the same classification performance values, namely accuracy and sensitivity around 0.9. This result is that the predicted value tends to be classified in the majority class (the class that is not considered) compared to the minority class (the class that is considered in this case the population does not work). So the level of accuracy in the classification modeling without treatment on the eight models gives poor results. The existence of misclassification will result in inaccurate errors in planning or government policymaking in handling unemployed residents.

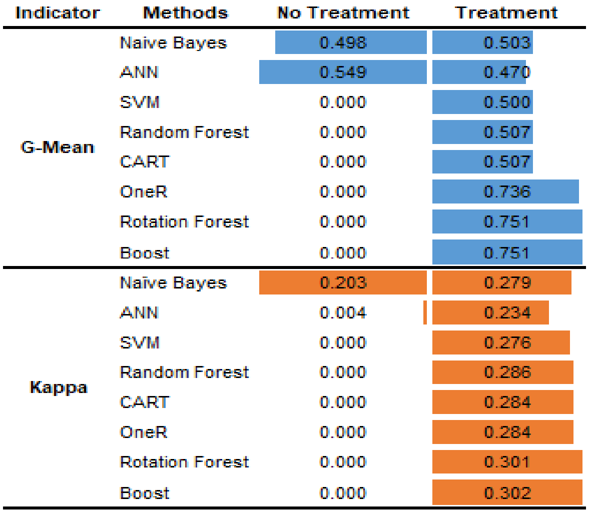
To improve the classification accuracy, especially for the minority class, this study applies both sampling methods in handling cases of imbalanced data, where the proportion of training data for both categories is balanced. The results obtained show that in the eight models for the specificity value, which shows a measure of classification accuracy in the minority class that is correctly predicted by the model, ranging from Sensitivity to 0.751. In addition, it also increases the G-mean value, which ranges from 0.5 to 0.75, and increases the Kappa value, which ranges from 0.234 to 0.302. On the other hand, a decrease in the value of accuracy to be in the range of 0.532 to 0.674. In other words, the handling of imbalanced data cases results in more balanced specificity and sensitivity values resulting in lower accuracy values, ranging from 0.740 to 0.800.

Classification modelling to predict the proportion of the population does not work by considering the values of accuracy, sensitivity, specificity, G-mean, and Kappa coefficient; the best model is the boost model with a combined/both sampling scheme. This result is because the model has the most significant and balanced classification performance value compared to other classification models. The model has an accuracy value of 0.789, sensitivity of 0.798, specificity of 0.751, G-mean of 0.750, and a kappa coefficient of 0.302. Because the classification performance measure in the best model is above the cut-off (0.5), the model can be said to be good. This result shows that the best classification model can correctly classify the working status of the population in Banten Province.



Indicator	Methods	No Treatment	Treatment
Accuracy	Naive Bayes	0.851	0.783
	ANN	0.902	0.74
	SVM	0.902	0.779
	Random Forest	0.902	0.786
	CART	0.902	0.789
	OneR	0.902	0.789
	Rotation Forest	0.899	0.8
	Boost	0.899	0.789
Sensitivity	Naïve Bayes	0.924	0.958
	ANN	0.903	0.959
	SVM	0.902	0.959
	Random Forest	0.902	0.959
	CART	0.902	0.958
	OneR	1	0.801
	Rotation Forest	1	0.812
	Boost	1	0.798
Specificity	Naïve Bayes	0.498	0.503
	ANN	0.549	0.47
	SVM	0	0.5
	Random Forest	0	0.507
	CART	0	0.507
	OneR	0	0.736
	Rotation Forest	0	0.751
	Boost	0	0.751

Figure 1a Comparison of Machine Learning's Performance



Indicator	Methods	No Treatment	Treatment
G-Mean	Naive Bayes	0.498	0.503
	ANN	0.549	0.470
	SVM	0.000	0.500
	Random Forest	0.000	0.507
	CART	0.000	0.507
	OneR	0.000	0.736
	Rotation Forest	0.000	0.751
	Boost	0.000	0.751
Kappa	Naive Bayes	0.203	0.279
	ANN	0.004	0.234
	SVM	0.000	0.276
	Random Forest	0.000	0.286
	CART	0.000	0.284
	OneR	0.000	0.284
	Rotation Forest	0.000	0.301
	Boost	0.000	0.302

Figure 1b Comparison of Machine Learning's Performance

Table 4 presents the mean decrease in *Modeling Classification Performance.* Gini from the best boost classification model. This result shows that the 4 (four) most important/most influential variables in the classification of working status are marital status, education, age, and hearing loss. The relationship between marital status and working status is quite close, as Yulianti et al. [13]. This result is related to a person's marital status related to the responsibility in meeting family needs. The relationship between education and work status is quite close, as in the research of Mutiadanu et al. [14].

Table 4. The Most Influential Predictor Variables

Features	Gain	Cover	Frequency
Marital status	0.894	0.426	0.333
Education	0.046	0.191	0.333
Age	0.032	0.224	0.167
hearing impairment	0.027	0.159	0.167

This result relates to education being considered as an investment in employment opportunities. The relationship between age and working status is quite close, as in Dhanani's research [15]. This case relates to the level of establishment and experience that a person has in getting a job as he gets older than others.

CONCLUSION

In general, classification modeling using resample techniques in imbalanced data and large data sets is proven to improve classification accuracy, especially for minority classes which can be seen from the value specificity, which is higher than the original data (without treatment). The Naïve Bayes and ANN models can produce specificity values even though the data used are imbalanced, while the other six models produce zero specificity values. Using the resampling technique, the model's accuracy, sensitivity, and specificity values become more balanced than others. The Bost model is the best model with accuracy, sensitivity, specificity, a more balanced GMean, and the most significant Kappa coefficient.

REFERENCES

[1] W. Venables and B. Ripley, Modern Applied Statistics with S, Fourth ed., New York: Springer, 2021.

[2] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News,* vol. 2, no. 3, pp. 18-22, 2018.

[3] H. v. Jouanne-Diedrich, "OneR: One Rule Machine Learning Classification Algorithm with Enhancements," 2017.

[4] M. Ballings and D. V. d. Poel, "RotationForest: Fit and Deploy Rotation Forest Models," 2017.

[5] T. Therneau and B. Atkinson, "rpart: Recursive Partitioning and Regression Trees," 2019.

[6] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch, "e1071: Misc Functions of the Department of Statistics, Probability Theory Group," 2021.

[7] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng and Y. Li, "xgboost: Extreme Gradient Boosting," 2021.

[8] M. Maalouf and Siddiqi, "Wieghted Logistic Regression for Large-Scale Imbalanced and Rare Events Data," *Journal of Knowledge-Based Systems,* vol. 59, pp. 142-148, 2014.

[9] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Journal of Political Analysis,* vol. 9, no. 2, pp. 137-163, 2001.

[10] Badan Pusat Statistik, "Labor Market Indicators Indonesia August 2020," Badan Pusat Statistik, Jakarta, 2021.

[11] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techhiques, Third Edition ed., Waltham: Elsevier Inc, 2012.

[12] J. Landis and G. Koch, The Measurment of Observer Agreement for Categorical Data, 2013.

[13] Yuliatin, T. Huseno and Febriani, "Pengaruh Karakteristik Kependudukan Terhadap Pengangguran di Sumatera Barat," *Jurnal Manajemen dan Kewirausahaan,* vol. 2, no. 2, 2011.

[14] S. Mutiadanu, M. R. Adry and D. Z. Putri, "Analisis Sosial Ekonomi Terhadap Pengangguran Muda.," *Ecosains,* vol. 7, no. 2, pp. 89-98, 2018.

[15] Dhanani, "Unemployment and Underemployment in Indonesia," International Labour Office, Switzeland:, 2004.