Combination of ADASYN-N and Random Forest in Predicting of Obesity Status in Indonesia: A Case Study of Indonesian Basic Health Research 2013

To cite this article: M Aqsha *et al* 2021 *J. Phys.: Conf. Ser.* **2123** 012039

Combination of ADASYN-N and Random Forest in Predicting of Obesity Status in Indonesia: A Case Study of Indonesian Basic Health Research 2013

M Aqsha[1], SA Thamrin[1*], and Armin Lawi[2]
[1] Department of Statistics, Universitas Hasanuddin, Makassar, South Sulawesi, Indonesia
[2] Department of Mathematics, Universitas Hasanuddin, Makassar, South Sulawesi, Indonesia

[*]Email: tuti@unhas.ac.id

Abstract. Obesity is a pathological condition due to the accumulation of excessive fat needed for body functions. The risk factors for obesity are related to their obesity status. Various machine learning approaches are an alternative in predicting obesity status. However, in most cases, the available datasets are not sufficiently balanced in their data classes. The existence of data imbalances can cause the prediction results to be inaccurate. The purpose of this paper is to overcome the problem of data class imbalance and predict obesity status using the 2013 Indonesian Basic Health Research (RISKESDAS) data. Adaptive Synthetic Nominal (ADASYN-N) can be used to balance obesity status data. The balanced obesity status data is then predicted using one of the machine learning approaches, namely Random Forest. The results obtained show that through ADASYN-N with a balance level parameter of 1 ($\beta = 100\%$) after synthetic data generation and Random Forest with a tree number of 200 and involving 7 variables as risk factors, giving the results of the classification of obesity status which is good. This can be seen from the AUC value of 84.41%.

1. Introduction

Obesity is an imbalance in the amount of food intake compared to energy expenditure made by the body [1]. The World Health Organization (WHO) declares obesity as a global epidemic with a prevalence of more than 1 billion adults who are overweight and up to 300 million are clinically obese [2]. The prevalence of obesity in developed countries ranges from 2.4% in South Korea to 32.2% in the United States while in developing countries it ranges from 2.4% in Indonesia to 35.6% in Saudi Arabia [3]. Therefore, a model is needed to determine a person's obesity status.

Along with the development of computing technology and machine learning (ML) algorithms, it has implicated for the rapid development of technology in various fields, including the field of biogenetic research. Machine Learning (ML) technology utilizes computers to perform the learning process from data and generate predictions. Decision tree (DT) is one of the methods in ML that is able to extract information from data sets into intuitive and easy-to-understand knowledge [4].

An ensemble is a way to overcome Decision Tree Constraints in the availability of training data with weak predictive values. In this ensemble method, several classifiers or prediction models are built with learning algorithms. In another study, it was concluded that classification and prediction with well-handled ensemble algorithms can generally produce higher accuracy and stability than using one algorithm alone [5]. Two common ways to do the ensemble method are boosting and bagging.

Combining many estimated values into one estimated value can be done by applying the concept of ensemble with bagging. One method of DT with bagging that can be used is Random Forest (RF). In this RF method, a random candidate predictor in each tree is used for the training process. Furthermore, after the training process, the most class labels are selected from the results of the entire tree formed [6]. However, in the case of classification, sometimes there is an imbalance problem in the data where the information in one class is not comparable or less (minority data) from the other class (majority data). The Adaptive Synthetic (ADASYN) method is one of technique to deal with problems with imbalanced data. ADASYN is a method for sampling approach that uses data distribution weights in minority classes based on the level of learning difficulty, so that synthesis data generated from minority classes has a higher level of learning difficulty compared to minority data itself [7]. Therefore, in this study, we will combine ADASYN-N and Random Forest in predicting obesity status in Indonesia using the 2013 Indonesian Basic Health Research data.

2. Material and methods

2.1. Data source

The data used in this study is secondary data, namely Indonesian obesity data from the 2013 Indonesian Basic Health Research survey. The type of variables used in this study are nominal and ordinal types. Variables with nominal type consist of obesity status (X1), gender (X2), age (X3), smoking (X4), strenuous activity (X4), moderate activity (X5), fruit consumption (X6), vegetables consumption (X7), sweet foods (X8), salty foods (X9), fatty foods (X10) and stress (X11). The proportion of training data used in obesity data is 80% (564,101), consisting of 28,677 obesity class data and 535,424 non-obese class data. Then, the proportion of test data used was 20% (140,876) consisting of 70,438 obesity class data and 70,438 non-obese class data.

2.2. Filtering and imbalanced data

Filtering is the process of selecting a subset of available data for analysis purposes. The amount of data usually causes an unknown value. Filtering is done to select data that has a meaningful value or select data according to needs.

Data imbalance is a condition where the distribution of data classes is not balanced, namely the number of data classes that dominates more than the number of other data classes. Generally, imbalanced data will result in poor classification prediction accuracy in the minority class. It is difficult to get a good and meaningful predictive model for minor classes due to insufficient information [8].

2.3. Adaptive synthetic

Adaptive Synthetic (ADASYN) is an approach method for sampling in the case of unbalanced data [7]. The use of distribution weights in minority classes based on the level of difficulty of understanding is the main idea of ADASYN. Through ADSYN, synthesis data from minority classes can be generated which is easier to understand. ADASYN can improve data understanding in minority classes by reducing bias caused by class imbalance and shifting classification decision boundaries to adaptive data difficulties.

Adaptive Synthetic Nominal (ADASYN-N) is an improved form of ADASYN which is only for numeric data [9]. In this study, we used categorical variables. Therefore, the Modified Value Difference Metric (MVDM) technique is used to calculate the distance between the minor class samples [10].

2.4. Random forest

In Random Forest (RF), several classification methods are combined, known as the ensemble method. This ensemble method aims to improve accuracy [6]. The basic method of data mining from RF is a decision tree. In this decision tree, the bottom (leaf) serves to determine what class the data belongs to and the top (root) is the place to enter the input. Then RF performs a classification in the form of a structured collection of tree classifiers in which each tree defines a class, then selects the most popular class [11].

Random Forest Algorithm for classification [12] as follow:

Input:

1. Data $\boldsymbol{D}$ with n sample $\{\boldsymbol{X}_i, y_i\}, i = 1,...,n$, where n is the total number of samples, $\boldsymbol{X}_i$ is a sample on the dimensional feature space p, $y_i \in \boldsymbol{Y} = \{1,-1\}$ is the identity class label bound to $\boldsymbol{X}_i$ and p is the number of features in the data.
2. Determine the values of B and m, where B is the number of trees and m is the number of variables that will be randomly selected when creating the tree.

Procedure:

1. **For** $b = 1$ **until** $\boldsymbol{B}$
   a. Random sampling bootstrap $\boldsymbol{Z}^* \in \boldsymbol{D}$ with size n from training data.
   b. Make a tree T_b from the bootstrap sample data by performing the steps below recursively for each node of the tree, until at least the minimum node size is reached or until the data can be no longer be divided.
   c. Choose m variable randomly
   d. Take the variable with the best partition among m variables.
   e. Specifies the end node
2. **Output of Random Forest model is $\{T_b\}_1^B$, where $\{T_b\}_1^B$ is a collection of trees that have been built starting from the first tree to $\boldsymbol{B}$ trees.**

2.5 Classification performance

The evaluation matrix has an important role in ML. One of the matrices used in evaluating machine learning algorithms is the confusion matrix. In this matrix, the number of observations in the prediction class is expressed in columns. Meanwhile, the actual number of observations in class is expressed in rows [13]. The way to evaluate the classification results based on the value in the confusion matrix is to calculate the values of accuracy, sensitivity, and specificity. Accuracy describes the level of accuracy of the classification as a whole. Sensitivity describes the accuracy of the data in the class, while the specificity describes the accuracy in the class. The Area Under Curve (AUC) value range is between 0.5 to 1. The value range is classified into five parts, namely incorrect accuracy (0.5-0.6), weak accuracy (0.6-0.7), medium accuracy (0.7-0.8), high accuracy (0.8-0.9) and very high accuracy (0.9-1.0) [13-14].

3. Result

The Out-of-Bag (OOB) error value for each Random Forest model can be seen in Figure 1. The optimal of trees (B) and variable (m) are 100 classification trees with 1 random variable selection with the smallest error value (5.06%). The confusion matrix of the Random Forest model in Table 1 is used to calculate the performance of the model. The performance of this model can be described in Table 2. The ROC and AUC curves can be seen in Figure 2. From Table 2 it can be seen that the AUC value is 50.60%, indicating that the accuracy of the model is not good, moreover the sensitivity is very small. This causes the prediction of the model only leads to the major (non-obese) class, because the information from the minor (obese) class is very little so that it tends to be ignored (considered as noise).

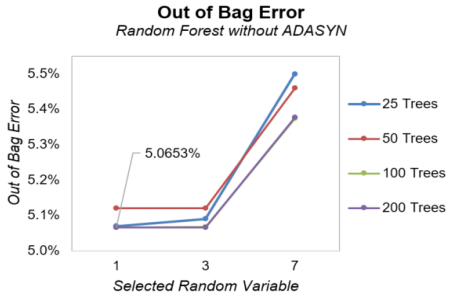

Figure 1. OOB of random forest model.

Table 1. Confusion matrix model random forest model for obesity data.

Actual	Predicted	
	Obese	Non-obese
Obese	5	70,433
Non-obese	0	70,438

Table 2. Performance of the random forest without ADASYN-N model

Metric	Performance value
Accuracy	50.00%
Sensitivity	0.00%
Specificity	100.00%
AUC	50.60%

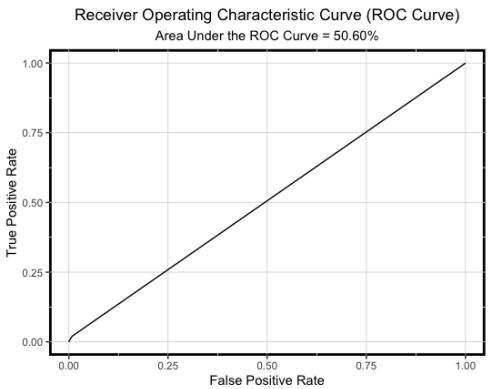

Figure 2. ROC dan AUC of random forest model for obesity data.

Table 3. Proportion of data on the combination of random forest and ADASYN-N with β =25%

Obesity Status	Without ADASYN-N (%)	With ADASYN-N (%)
Obese	99,115 (14.05)	225,802 (27.15)
Non-obese	605,862 (85.95)	605,862 (72.85)
Total	704,977 (100.00)	831,664 (100.00)

In the combination of Random Forest and ADASYN models with β= 25%, synthetic data was generated by 25% from the difference between minor data (obesity) and major data (non-obesity), which is 126,687. This new data is added to the original data so that the total data is 831,664 (Table 3).

The total data synthesized by the ADASYN-N technique in the training data used is 80% (664,986), consisting of 142,463 obesity class and 522,523 non-obese class. Then, the proportion of test data used was 20% (166,678), consisting of 83,339 obesity class and 83,339 non-obese class. The OOB error value for each Random Forest model can be seen in Figure 3a. The optimal of trees (B) and variable (m) are 200 classification trees with 7 random variables selected with the smallest error value of 18.58%. The value of the confusion matrix from the combination of Random Forest and ADASYN-N models with β=25% can be seen in Table 4. Based on Table 5, ADASYN-N with β=25% improves classification performance. This is in line with the ROC and AUC curves in Figure 4a.

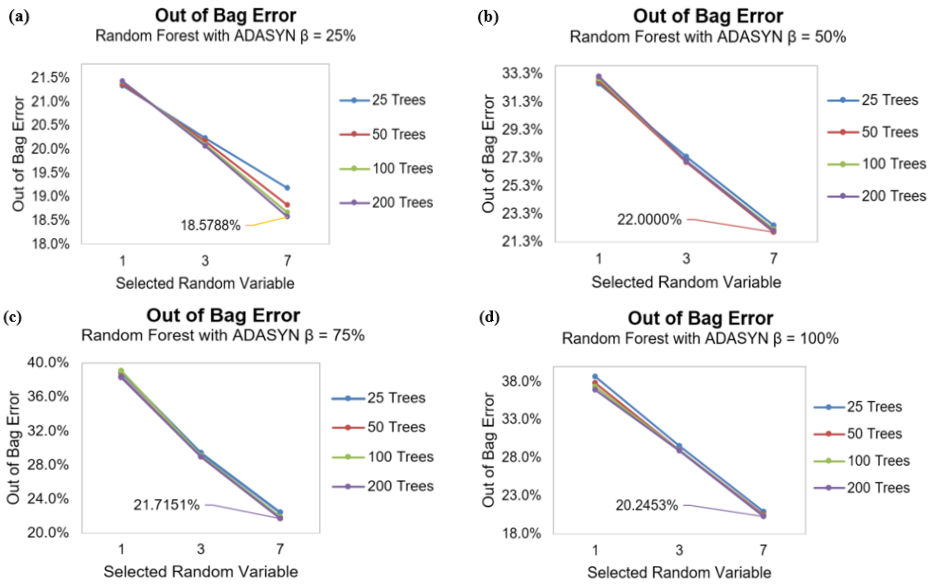

Figure 3. OOB of the combination of random forest and ADASYN-N models with different β parameter values (a) β = 25%, (b) β = 50%, (c) β = 75%, and (d) β = 100%.

Table 4. Confusion matrix of the combination of random forest and ADASYN-N models with different β parameter values

Actual	Predicted							
	β = 25%		β = 50%		β = 75%		β = 100%	
	Obese	Non-obese	Obese	Non-obese	Obese	Non-obese	Obese	Non-obese
Obese	26,946	56,393	58,767	37,159	80,572	28,297	100,521	20,384
Non-obese	4,317	79,022	12,665	83,261	19,787	89,082	28,419	92,486

Table 5. Performance of the combination of random forest and ADASYN-N models with different β parameter values

Metric	β = 25%	β = 50%	β = 75%	β = 100%
Accuracy	65.35%	74.26%	77.70%	79.89%
Sensitivity	36.97%	61.67%	73.44%	83.16%
Specificity	93,73%	88,39%	81.95%	76.61%
AUC	71,36%	78,52%	81.83%	84.41%

Furthermore, the combined Random Forest and ADASYN methods with β= 50% resulted in synthetic data generated by 50% of the difference between minor (obese) data and major (non-obese) data, amounting to 253,373. This new data is added to the original data so that the total is 958,350. The value of the synthetic data from the combination of the Random Forest and ADASYN methods with β= 50% can be seen in Table 6.

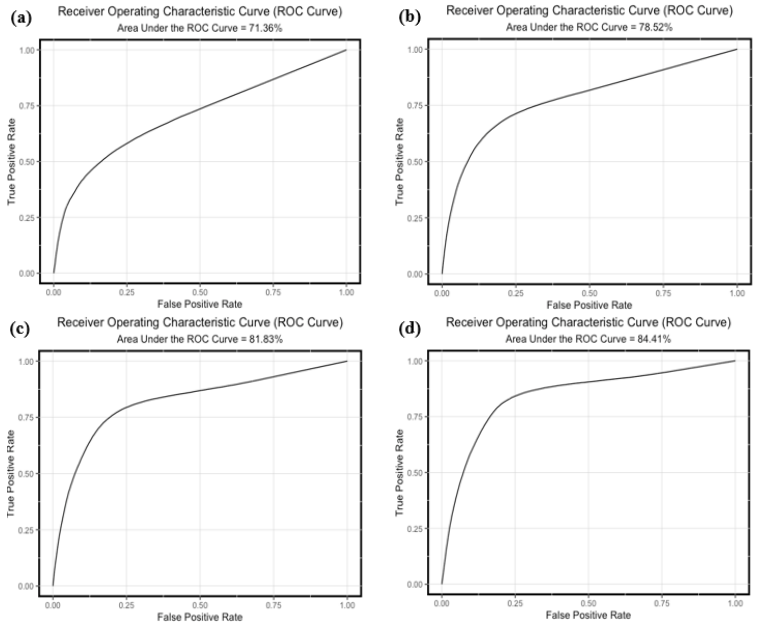

Figure 4. ROC and AUC model for the combination of random forest and ADASYN-N with different level of β values (a) β = 25%, (b) β = 50%, (c) β = 75%, and (d) β = 100%.

Table 6. Proportion of data on the combination of random forest and ADASYN-N with β =50%

Obesity Status	Without ADASYN-N (%)	With ADASYN-N (%)
Obese	99,115 (14.05)	352,488 (36.78)
Non-obese	605,862 (85.95)	605,862 (63.22)
Total	704,977 (100.00)	958,350 (100.00)

In this combination, the proportion of training data used is 80% (766,499), consisting of 256,563 obesity class and 509,936 non-obese class. Then, the proportion of test data used was 20% (191,852) consisting of 95,926 obesity class and 95,926 non-obese class. The OOB error value for β= 50% for each Random Forest model can be seen in Figure 3b. The optimal of trees and variable are 50 classification trees and 7 random variable selections with the smallest error value of 22%. Next, the confusion matrix combination of Random Forest and ADASYN-N is calculated with = 50% (Table 4). Based on this confusion matrix, the performance of the combination of the Random Forest and ADASYN-N models can be obtained with β= 50% (Table 5). With this combination of Random Forest and ADASYN-N, the performance classification shows better. This can be seen also in the ROC graph in Figure 4b.

Furthermore, the value of β was increased to 75% in the combination of Random Forest and ADASYN-N model. Synthetic data is generated by 75% from the difference between minor (obese) data and major (non-obese) data and it obtained by 380,060 and this new data is added to the original data, so that it is 1,085,037 as shown in Table 7.

From Table 7, the proportion of training data used is 80%, which is 867,299, which consists of 370,306 obesity class and 496,993 non-obese class. Then the proportion of test data used is 20%, which is 217,738 data, which consists of 108,869 obesity class and 108,869 non-obese class. The value of OOB error at β =75% for each Random Forest model can be seen in Figure 3c. The optimal of trees and variables values are 200 classification trees and 7 random variable selections with the smallest error value of 21.71%. Furthermore, to obtain the performance value (Table 5), the combination model of

Random Forest and ADASYN-N with =75% is calculated first by calculating the confusion matrix (Table 4). With ADASYN-N and β=75%, the classification performance is even better. The ROC graph shown in Figure 4c. In the combination of Random Forest and ADASYN-N methods with β= 100%, synthetic data is generated by 100% from the difference between minor (obese) and major (non-obese) data. The resulting synthesis data is 506,747 and this new data is added to the original data so that it is 1,211,724. The results can be seen in Table 8.

Table 7. Proportion of data on the combination of random forest and ADASYN-N with β =75%

Obesity Status	Without ADASYN-N (%)	With ADASYN-N (%)
Obese	99,115 (14.05)	479,175 (44.16)
Non-obese	605,862 (85.95)	605,862 (55.84)
Total	704,977 (100.00)	1,085,037 (100.00)

Table 8. Proportion of data on the combination of random forest and ADASYN-N with β =100%

Obesity Status	Without ADASYN-N (%)	With ADASYN-N (%)
Obese	99,115 (14,05)	605,862 (50,00)
Non-obese	605,862 (85,95)	605,862 (50,00)
Total	704,977 (100,00)	1,211,724 (100,00)

Table 9. Comparison of classification performance of random forest model with ADASYN-N and without ADASYN-N

Model	*B* Tree	Accuracy	Sensitivity	Specificity	AUC
Random Forest	100 Trees 1 Rand. Var.	50.00%	0.00%	100.00%	50.60%
Random Forest with ADASYN 25%	200 Trees 7 Rand. Var.	65.35%	36.97%	93.73%	71.36%
Random Forest with ADASYN 50%	50 Trees 7 Rand. Var.	74.26%	61.67%	88.39%	78.52%
Random Forest with ADASYN 75%	200 Trees 7 Rand. Var.	77.70%	73.44%	81.95%	81.83%
Random Forest with ADASYN 100%	200 Trees 7 Rand. Var.	79.89%	83.16%	76.61%	84.41%

In Table 8, the proportion of training data used is 80%, which is 969,914, which consists of 484,957 obesity class and 484,957 non-obese class. The proportion of test data used was 20%, which is 241,810 data, which consisted of 120,905 obesity class and 120,905 non-obese class. The value of OOB Error in the combination of Random Forest and ADASYN-N with β= 100% for each Random Forest model can be seen in Figure 3d. The optimal of trees and variable are 200 classification trees and 7 random variable selections with the smallest error value of 20.24%. The confusion matrix from the results of this combination and the performance can be seen in Tables 4 and 5. With ADASYN-N and β=100%, the classification performance becomes better. The ROC graph shown in Figure 4d. From Table 9 it can be seen that the AUC value and the sensitivity value of the Random Forest model without ADASYN-N are lower than the combination of the Random Forest model with ADASYN-N. The highest AUC value (84.41%) was obtained in the combination model of Random Forest and ADASYN-N with β=100%. This shows that the performance of the combination of Random Forest and ADASYN-N is better than the Random Forest without ADASYN-N models (Table 9).

4. Conclusion

The ADASYN-N method with β =100% can overcome the imbalance of obesity data in Indonesia. This can be seen in the level of sensitivity that continues to increase along with the increase in the value of the ADASYN-N method. The combination of ADASYN-N and the Random Forest model as many as 200 classification trees with 7 variables from obesity data that were randomly selected gave a fairly high accuracy value (79.89%), sensitivity value of 83.16% and AUC value of 84.41%.

5. Acknowledgement

The second author and correspondence author would like to thank to the Ministry of Education and Culture, Research and Technology which has funded this research through the PDUPT Grant of Universitas Hasanuddin with contract No. 752/ UN4.22/PT.01.03/2021. We thank also to the Health Research and Development Agency, Ministry of Health, for providing access to RISKESDAS data.

6. References

[1] RISKESDAS. Kementerian Kesehatan RI Riset Kesehatan Dasar 2013. https://pusdatin.kemkes.go.id/resources/download/general/Hasil%20Riskesdas%202013.pdf.
[2] Soegih R R and Wiramihardja KK 2009 Obesitas: Permasalahan dan Terapi Praktis, Sagung Seto.
[3] Sugianti E, Hardinsyah, and Afriansyah N 2009 Faktor Risiko Obesitas Sentral pada Orang Dewasa di DKI Jakarta: Analisis Lanjut Data RISKESDAS 2007, 32(2), 105–116.
[4] Baros R C, Basgalupp M P, Carvalho A C P L F, and Freitas A A 2011 Towards the automatic design of decision tree induction algorithms, Proceedings of The 13th Annual Conference Companion on Genetic and Evolutionary Computation, 567–574.
[5] Yang P, Yang Y H, Zhou B B, and Zomaya A Y 2010 A Review of Ensemble Methods in Bioinformatics: Including Stability of Feature Selection and Ensemble Feature Selection Methods, Current Bioinformatics, 5(4), 296–308.
[6] Han J, Kamber M, and Pei J 2012 Data Mining: Concept and Techniques Third Edition, Elsevier Inc.
[7] He H, Bai Y, Gracia E A, and Li S 2008 ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, IEEE , 1322–1328.
[8] Yap B W, Rani K A, Rahman H A A, Fong S, Khairudin Z, and Abdullah N N 2014 An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets, Proceedings of The First International Conference on Advanced Data and Information Engineering, 13–22.
[9] Fithriasari K, Hariastuti I, and Wening K S 2020 Handling Imbalance Data in Classification Model with Nominal Predictors, Internasional Journal of Computing Science and Applied Mathematics, 6(1), 33–37.
[10] Cost S and Salzberg S 1993 A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features, Machine Learning, Kluwer Academic.
[11] Breiman L 2001 Random Forests, Machine Learning, Kluwer Academic.
[12] Hastie T, Tibshirani R, and Friedman J H 2008 The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer.
[13] Thamrin S A, Arsyad D S, Kuswanto H, Lawi A, and Nasir S 2021 Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018, Frontiers in Nutrition, 8, 669155.
[14] Gorunescu F 2011 Data Mining: Concept, Models and Techniques, Springer-Verlag Berlin Heidelberg.