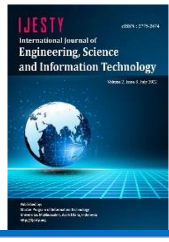
International Journal of Engineering, Science & InformationTechnology (IJESTY)
Volume 2, No. 3 (2022) pp. 34-39
ISSN 2775-2674 (online)
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v1i4.280
Research Paper, Short Communication, Review, Technical Paper



Supervised Models to Predict the Stunting in East Aceh

Eva Darnila[1]*, Maryana[2], Khalid Mawardi[3], Marzuki Sinambela[4], Iwan Pahendra[5]

[1]Department of Informatics, Universitas Malikussaleh, Aceh, Indonesia
[2]Department of Mathematics, Universitas Malikussaleh, Aceh, Indonesia
[3]Department of Nutrition and Public Health, Dinas Kesehatan, East Aceh, Indonesia.
[4]Indonesia Meteorology, Climatology, And Geophisics Agency, Medan, Indonesia
[5]Department of Power Enginering, Universitas Sriwijaya, Palembang, Indonesia
**Corresponding author E-mail: eva.darnila@unimal.ac.id*

Manuscript received 15 April 2022; revised 1 May 2022; accepted 15 June 2022. Date of publication 25 July 2022

Abstract

Nowadays, Undernutrition is the leading cause of child death in developing countries. Many people and organizations are trying to mitigate or minimize child death cases. Thus, this paper aims to have an excellent method to handle undernutrition cases by exploring the efficacy of machine learning (ML) approaches to predict Stunting in East Aceh administrative zones of Indonesia and to identify the most important predictors. The study employed ML techniques using retrospective cross-sectional survey data from East Aceh, a national-representative data collected from the government using 2019 about stunting data. We explored Random forest commonly used ML algorithms. Random Forest (RF) is an extension of bagging that takes random samples of data and uses a random subset of features that mitigates overfitting. Our results showed that the considered machine learning classification algorithms by random forest could effectively predict the stunting status in East Aceh administrative zones. Persistent stunting status was found in the east part of Aceh. Identifying high-risk zones can provide more helpful information and data to decision-makers in reducing child undernutrition.

Keywords: *Stunting, Random Forest, Machine Learning.*

1. Introduction

Children's growth is an essential focus for every parent because at this age, food intake must be considered for brain development and memory. When feeding children do not meet their nutritional needs, they can be at risk of experiencing malnutrition. Therefore, information about nutrition knowledge is needed in children who can inform and meet the knowledge needs of the community, especially in this case, parents. Thus, to have the ability to have this knowledge, one must have knowledge and understanding of nutrition to be able to provide initial action to be taken when their child is experiencing symptoms of malnutrition.

Aceh is in the third-highest national ranking for the number of children under five with stunting, behind East Nusa Tenggara (NTT) and West Sulawesi [1]. The leading case of this study is how to predict the stunting in East Aceh as a national priority to evaluate the growth of children. We used the random forest method's supervised model to predict the East Aceh stunting [2] [3].

Proper nutrition is so crucial to leading a healthy lifestyle. Malnutrition, particularly undernutrition, is a global concern for children's health conditions and survival [2] [3]. Almost half of the deaths of children in developing countries were directly or indirectly linked to malnutrition. Malnourished children are more vulnerable to different illnesses compathanr counterparts [5] [6]. A considerable number of studies investigating the issue targeting under-five children malnutrition and the risk factors associated with this age group. These studies employed classical models such as generalized linear (mixed) models [7]. The finding from the investigations, among others, showed that the nutritional status of children of this age group has gradually improved over the last decades in East Aceh. Particularly, it has been found that the prevalence of under-five children under-weight in East Aceh was 20.56% in 2019, while the majority of stunting was 36.9% in 2019. Similarly, 10.7% of under-five children were wasted in 7% in 2019. The prevalence of having at least one of the under-nutrition indicators was measured in terms of the composite index for anthropometric failure (CIAF) 42.4 in 2019. Moreover, the CIAF is computed by grouping different forms of anthropometric failure as such: B-wasting only, C-wasting and underweight, D-wasting, stunting and underweight, E-stunting and underweight, F-stunting only, and Y-underweight only. The CIAF was calculated by aggregating these six (B–Y) categories [7] [8]. Most of such studin in this country depicted the effects of socio-economic and demographic covariates ere associated with under-five children's undernutrition status using the classical regression models [5] [9]. Those traditional models are widely used for causal inferences and with the selection of built-in features, with a relatively small number of covariates. Correlations between covariates (multicollinearity) and a large number of factors are the common analytical challenges in traditional modeling [8] [9]. Moreover compared to those classical models, the machine learning (ML) methods have the qualities of using a lmore significant number of predictors, requiring fewer assumptions, incorporating “multi-dimensional correlations,” and producing a more flexible relationship among the predictor variables and the outcome variables. In addition, the ML models can create models for prediction purposes that show superiority in handling classification problems when compared with the classical approaches [10] [11].

2. Methods

To study stunting in East Aceh, we used the stunting survey data set selected to be used in this study [14]. This data set was collected in 2019 by surveying 143 Somerville residents about their stunting and satisfaction with city services. There are six attributes in Table 1, X1 to X6, with values 1 to 5 and a binary decision attribute.
D = decision attribute (D) with values 0 (unhappy) and 1 (happy)
• X1 = the weight
• X2 = the Height
• X3 = the head circumference
• X4 = the Upper arm circumference
• X5 = the fathom length
• X6 = the Knee Height
No missing values have been observed in this data set, so no further action is needed to deal with them. Table 2 describing with method statistic, and show the stunting survey data in East Aceh [15].
Your paper must be in two-column format with a space of 0.5cm between columns.

Table 1. Stunting Survey Data

	D	X1	X2	X3	X4	X5	X6
0	0	2	5	3	2	1	4
1	0	4	1	3	3	4	3
2	1	3	2	4	3	4	4
3	0	5	3	4	5	4	5
4	0	5	1	4	3	4	5

Based on the described methods in Table 2, we can calculate the statistic for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame. It analyzes both numeric and object series and also the DataFrame column sets of mixed data types.

Table 2. Describe Methods Statistic

	D	X1	X2	X3	X4	X5	X6
count	143	143	143	143	143	143	143
mean	0.538462	4.307692	2.447552	3.244755	3.657343	3.615385	4.223776
std	0.500271	0.849447	1.085619	1.001525	0.864845	1.106467	0.825812
min	0	1	1	1	1	1	1
25%	0	4	2	3	3	3	4
50%	1	5	2	3	4	4	4
75%	1	5	3	4	4	4	5
max	1	5	5	5	5	5	5

In this study Random Forest (RF) method were used for training classification model to predict normal and abnormal [16]. Model building the ML models have shown superiority in taking care of classification problems when compared with the traditional models (like generalized linear mixed models) [17]. The raw data are usually not found in the form and shape that is required for optimal performance of the machine learning algorithms. The algorithms that would be implemented in ML are only numerical values and therefore it is important to transform the categorical variables into numerical values. Hence, the preprocessing step is the most important aspect in the ML model applications [14] [16] [17]. The categorical features of the dataset are encoded to transform these features into numerical values and the continuous data in this study were normalized. For ML approaches, the dataset is randomly split into two: a training dataset which trains the model, and a test dataset where we predict the response variable and check whether the predicted outcome is similar to the actual outcomes, and the validation dataset is considered for the parameter estimates to be incorporated in the training models [20].

3. Result and Discussion

This analysis consisted of data from 143 children of age 0–59 months. Of these, 15,281 (52.09%) had at least one form of the undernutrition indicators (stunting, wasting, and underweight) measured in terms of CIAF. We examined the prevalence of CIAF of U5C experience

across different child and mother-household level covariates. The prevalence of CIAF was more common among parents with no formal education compared to parents with secondary and post-secondary levels of educations. Most of the undernourished children were from rural areas. Also, the prevalence of undernourished children was reported from the lower wealth index of households, from mothers having no media exposure, from unimproved toilets and sanitation compared with their counterparts. Covariates that were significant in the statistics were used to develop the ML algorithms on the training dataset (Table 2).
Based on data analysis by using the person correlation we can devide the

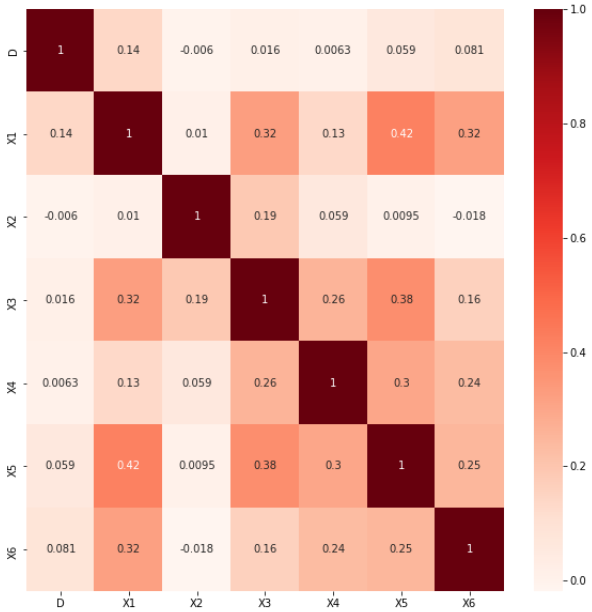


Fig 1. Data analysis using person correlation

By looking at results, we can say that attribute X1 in Fig 1 has the most correlation with the target among other features.

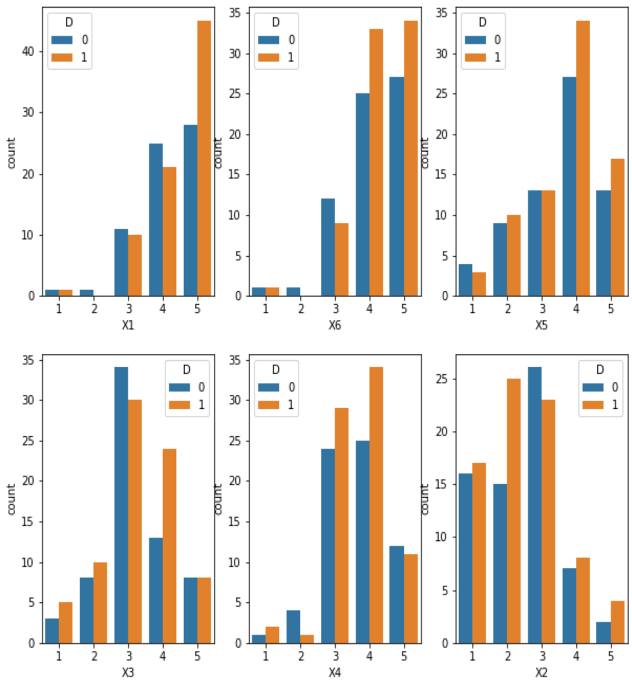


Fig 2. Multiple Count Plot of Correlation with Output Varible

Table 3. Dataset into Training and Testing Sets

	X1	X2	X3	X4	X5	X6
0	2	5	3	2	1	4
1	4	1	3	3	4	3
2	3	2	4	3	4	4
3	5	3	4	5	4	5
4	5	1	4	3	4	5

Based on the split in Table 3, we got the x_train.shape is (114, 6) x_test.shape is (29, 6), y_train.shape is (114,), (y_test.shape is (29,). The continuous data in this study were normalized and the categorical variables were encoded. The machine learning models are known as advanced approaches and techniques for quick and accurate prediction of real world problems. In this paper, the ML techniques are analyzed by investigating the influence of training/testing ratio on the performance of the six popular ML models to predict the undernutrition of under-five children. The performance of the ML models was slightly changed under the two different ratios. The result revealed that the ratio 70/30 was the most suitable ratio for the training and validating ML models. This study is in line with previously published studies [18, 23, 30–44, 83–86]. The ML tool can offer insight into the identification of novel factors associated with under-five undernutrition that can serve as targets for intervention. Among the six predictive models built using these techniques, the Random Forest (RF) model reveals a higher predictive power as compared to other ML models including the logistic regression. The RF model reveals that urban rural settlement ratio, the literacy level of parents, under five populations, BMI of mothers, locations (zones, place of residence), and rainfall distributions were the top important predictors of under-five undernutrition in East Aceh.

According to the Random Forest we have the accuracy Random Forest's Accuracy 58.62 % and the confusion matrix as show in fig 3. In Fig 3 shows the confusion matrix results of Random Forest algorithm. It is a summary of the prediction results of this algorithm in classifying stunted and non-stunted children. Both correct and incorrect classes are represented in the table below.

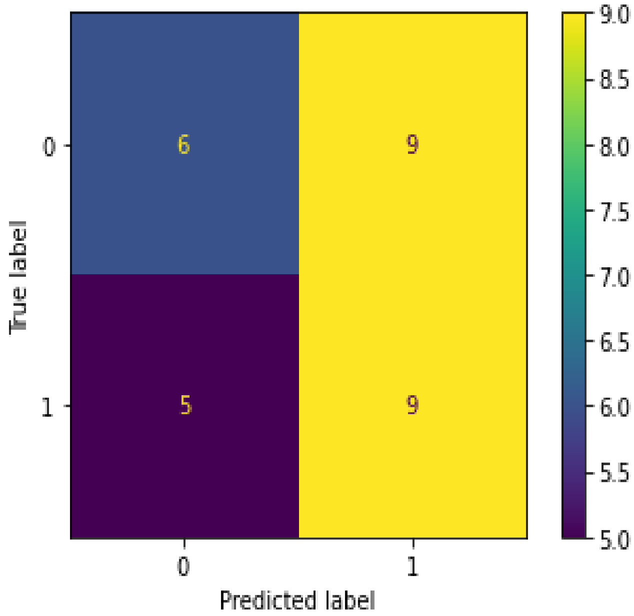


Fig 3. Confusion Matrix

For the graphical view based on the random forest we can get the visual as show in fig 4.

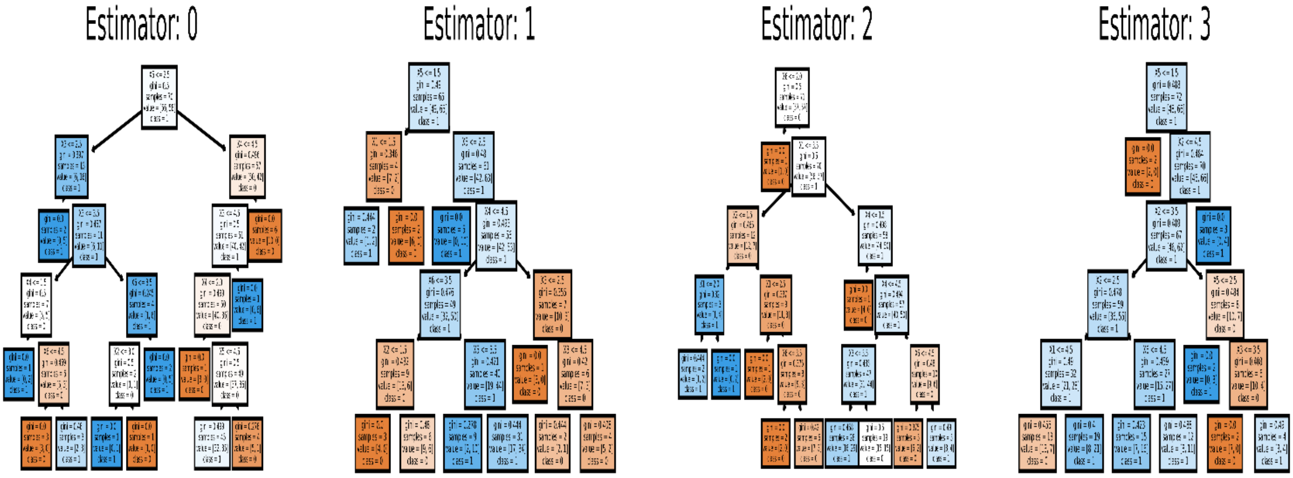


Fig 4. Graphical view based on random forest model

In Fig 4, showing a random forest, every tree will be built differently. We used these images to display the reasoning behind a decision tree (and subsequently a random forest) rather than for specific details. It's helpful to limit maximum depth in your trees when you have a lot of features. Otherwise, the end up with massive trees, which look impressive, but cannot be interpreted at all. To access the single decision tree from fig 4 the random forest in scikit-learn use estimators_ attribute. In Fig 4 covered how to visualize decision trees using Graphviz and Matplotlib. The way to visualize decision trees using Matplotlib is a newer method so it might change or be improved upon in the future. Graphviz in sunting dataset is currently more flexible as you can always modify your dot files to make them more visually appealing like did using the dot language or even just alter the orientation of your decision tree

4. Conclusion

The main objective of this study was to predict and evaluate the performance of Random forest machine learning (ML) algorithms considering the influence of two train-test splits ratios in predicting the stunting classification. Popular statistical indicators, such as accuracy and area under the curve were employed to evaluate the predictive power of the ML models under different testing and training ratios. The accuracy the model had, the better was the performance of the model. Our results confirm that ML models can effectively predict the stunting status and hence may be useful for concerned body decision tools. The best model of the RF, with accuracy of 58.6% respectively. The findings from this paper showed that considerable zonal disparities in the stunting status persist in the east part of Aceh.

Acknowledgement

We are grateful to all the scientists and scientific personal. We would like thank to support from The Universitas Malikussaleh, and Dinas Kesehatan Kabupaten Aceh Timur for providing the data in this study.

References

[1] N. Hartaty and R. Mastura, "On overview of family knowledge on fish consumption in avoiding stunting in Meuraxa Sub-District of Banda Aceh municipality," *J. Syiah Kuala Dent. Soc.*, vol. 6, no. 1, pp. 18–23, 2021, doi: 10.24815/jds.v6i1.21889.

[2] F. O. Aridiyah, N. Rohmawati, and M. Ririanty, "Faktor-faktor yang Mempengaruhi Kejadian Stunting pada Anak Balita di Wilayah Pedesaan dan Perkotaan (The Factors Affecting Stunting on Toddlers in Rural and Urban Areas)," *e-Jurnal Pustaka Kesehat.*, 2015.

[3] R. K. Phalkey, C. Aranda-Jan, S. Marx, B. Höfle, and R. Sauerborn, "Systematic review of current efforts to quantify the impacts of climate change on undernutrition," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 33, pp. E4522–E4529, 2015, doi: 10.1073/pnas.1409769112.

[4] M. T. Niles, B. F. Emery, S. Wiltshire, M. E. Brown, B. Fisher, and T. H. Ricketts, "Climate impacts associated with reduced diet diversity in children across nineteen countries," *Environ. Res. Lett.*, vol. 16, no. 1, 2021, doi: 10.1088/1748-9326/abd0ab.

[5] A. R. El-Ghannam, "The global problems of child malnutrition and mortality in different world regions," *J. Heal. Soc. Policy*, vol. 16, no. 4, pp. 1–26, 2003, doi: 10.1300/J045v16n04_01.

[6] D. L. Pelletier and E. A. Frongillo, "Changes in child survival are strongly associated with changes in malnutrition in developing countries," *J. Nutr.*, vol. 133, no. 1, pp. 107–119, 2003, doi: 10.1093/jn/133.1.107.

[7] F. Habyarimana, T. Zewotir, and S. Ramroop, "A proportional odds model with complex sampling design to identify key determinants of malnutrition of children under five years in Rwanda," *Mediterr. J. Soc. Sci.*, vol. 5, no. 23, pp. 1642–1648, 2014, doi: 10.5901/mjss.2014.v5n23p1642.

[8] W. Rasheed and A. Jeyakumar, "Magnitude and severity of anthropometric failure among children under two years using Composite Index of Anthropometric Failure (CIAF) and WHO standards," *Int. J. Pediatr. Adolesc. Med.*, vol. 5, no. 1, pp. 24–27, 2018, doi: 10.1016/j.ijpam.2017.12.003.

[9] R. J. Biellik and W. A. Orenstein, "Strengthening routine immunization through measles-rubella elimination," *Vaccine*, 2018, doi: 10.1016/j.vaccine.2018.07.029.

[10] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges," *Eur. Heart J.*, vol. 38, no. 23, pp. 1805–1814, 2017, doi: 10.1093/eurheartj/ehw302.

[11] B. Ambale-Venkatesh *et al.*, "Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis," *Circ. Res.*, vol. 121, no. 9, pp. 1092–1101, 2017, doi: 10.1161/CIRCRESAHA.117.311312.

[12] J. R. Khan, J. H. Tomal, and E. Raheem, "Model and variable selection using machine learning methods with applications to childhood stunting in Bangladesh," *Informatics Heal. Soc. Care*, vol. 00, no. 00, pp. 1–18, 2021, doi: 10.1080/17538157.2021.1904938.

[13] S. Hanieh *et al.*, "The Stunting Tool for Early Prevention: Development and external validation of a novel tool to predict risk of stunting in children at 3 years of age," *BMJ Glob. Heal.*, vol. 4, no. 6, pp. 1–12, 2019, doi: 10.1136/bmjgh-2019-001801.

[14] R. Kusumaningrum, T. A. Indihatmoko, S. R. Juwita, A. F. Hanifah, K. Khadijah, and B. Surarso, "Benchmarking of multi-class algorithms for classifying documents related to stunting," *Appl. Sci.*, vol. 10, no. 23, pp. 1–13, 2020, doi: 10.3390/app10238621.

[15] A. Hadi, "The Internalization of Local Wisdom Value in Dayah Educational Institution," *J. Ilm. Peuradeun*, 2017, doi: 10.26811/peuradeun.v5i2.128.

[16] D. Ruppert, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," *J. Am. Stat. Assoc.*, 2004, doi: 10.1198/jasa.2004.s339.

[17] R. D. Kurniawan, H. Rintis, and Setiono, "MENGISI DATA HUJAN YANG HILANG DENGAN METODE AUTOREGRESSIVE DAN METODE RECIPROCAL DENGAN PENGUJIAN DEBIT KALA ULANG (STUDI KASUS DI DAS BAKALAN)," *e-Jurnal MATRIKS Tek. SIPI*, 2017.

[18] F. ; D. Q. N. David, "School of Mathematics," *Anal. Crit. Think.*, vol. 15, no. 4, pp. 12–14, 2009.

[19] E. Harrison *et al.*, "Machine learning model demonstrates stunting at birth and systemic inflammatory biomarkers as predictors of subsequent infant growth – a four-year prospective study," *BMC Pediatr.*, vol. 20, no. 1, pp. 1–10, 2020, doi: 10.1186/s12887-020-02392-3.

[20] J. W. Puspita, S. Gunadharma, S. W. Indratno, and E. Soewono, "Bayesian approach to identify spike and sharp waves in EEG data of epilepsy patients," *Biomed. Signal Process. Control*, vol. 35, 2017, doi: 10.1016/j.bspc.2017.02.016.