Vol. 10, No.1: 12 — 15
May 2021
DOI: 10.22487/25411969.2021.v10.i1.15466

Natural Science
Journal of Science and Technology

Original article

Implementation of boosting on the C5.0 algorithm in the health development index data

Mohammad Fajri[1*], Ashari Ramadhan[1], Rahmat Hidayatullah[2]

[1]Department of Statistic, Faculty of Mathematics and Natural Sciences, Tadulako University, Jalan Raya Soekarno–Hatta, Tondo, Palu, 94117, Central Sulawesi, Indonesia.
[2]Department of Natural Education Science, Faculty of Teacher Training and Education Science, Tadulako University, Jalan Raya Soekarno-Hatta, Tondo, Palu 94117, Central Sulawesi, Indonesia

Keywords: PHDI, C5.0 algorithm, boosting, accuracy

Article history:
Received 16 March 2021
Accepted 29 May 2021
Published 31 May 2021

* Coresponding Author :
m.fajri@untad.ac.id

Abstract

The public health development index is a collection of health indicators that can be easily and directly measured to describe health problems down to the district/city level. The last calculation of the public health development index in Indonesia was carried out in 2018 with the result that not all districts/cities are in a high condition. This research was conducted to classify districts/cities that have a high and low public health development index using the boosted C5.0 algorithm. The results showed that the accuracy of the model increased with the increase in the number of iterations and was constant at the 60th iteration. Error training was also smaller and constant at the 10th iteration. The final accuracy, sensitivity and specificity were obtained respectively 97.09, 96.72, and 97.62.

INTRODUCTION

Public health is one of the main capitals in order to advance the welfare of the nation. The State of Indonesia in Law Number 17 of 2007 concerning the Long-Term National Development Plan (LTNDP) for 2005-2025 states briefly the direction of national development, including health development (Kementerian Kesehatan RI, 2015).

The Public Health Development Index (PHDI) is a collection of health indicators that can be easily and directly measured to describe health problems. This set of health indicators can directly or indirectly contribute to increasing long and healthy life expectancy. The selected indicators in the PHDI show more of the impact of health development in the previous year and become a reference for planning health development programs for the following year (Kementerian Kesehatan RI., 2014).

The Public Health Development Index results in 2018 showed a national average of 0.6087. The results were used as a comparison to see the high or low PHDI from all districts/cities in Indonesia. So that the research will be conducted by classifying districts/cities in Indonesia including high or low PHDI using several indicators.

The methods used to carry out the classification process are quite diverse (Cahyani & Muslim, 2020). One of them is the decision tree method. The decision tree has several algorithms, one of which is the C.50 boosting algorithm. The C.50 boosting algorithm produces classification rules in the form of a decision tree which is then used for data classification. The boosting process in the C5.0 algorithm will change the weak learner to become the strong learner (Suprianto et al., 2013; Tanyu et al., 2021).

MATERIALS AND METHODS

The Public Health Development Index (PHDI) is a collection of health indicators that can be easily and directly measured to describe health problems. The public health development index is a key indicator to see health development down to the district/city level

(Kementerian Kesehatan RI., 2014). Some of the health development indicators are the under-five health index, reproductive health index, health service index, health behavior index, non-communicable disease index, infectious disease index, and environmental health index.

The data used in this study are sourced from the official website of the Ministry of Health of the Republic of Indonesia, in the 2018 PHDI book. The data consists of 514 observations of districts/cities about Health Development Index. The data is divided into 2 variables, namely the target variable and the predictor variable. The target variable is PHDI, while the predictor variables consist of the under-five health index, reproductive health index, health service index, health behavior index, non-communicable disease index, infectious disease index, environmental health index. The target variable is then divided into 2 classes, namely low if the PHDI value is below the average (0.6087) and high if the PHDI value is above the average (0.6087).

The C5.0 algorithm is an algorithm for classification with the output of classification rules that produces a decision tree. The C5.0 algorithm is developed from the ID3 and C.45 algorithms (Wijaya et al., 2018). The accuracy of the C5.0 algorithm can be optimized using boosting. In making classification rules the C5.0 algorithm using entropy and information gain. Here is the formula for calculating total entropy (Han et al., 2012) :

$$info(D) = -\sum_{i=0}^{n} p_i \, log \qquad (1)$$

info(D) = Total entropy
p_i = Class opportunities
i = Number of classes

After obtaining the total entropy, then calculate the entropy of the predictor variable using the following formula (Han et al., 2012) :

$$info_A(D) = -\sum_{j=1}^{v} \frac{D_j}{D} . info(D_j) \qquad (2)$$

Info A (D) = entropy of variable A
Info (D j) = entropy factor j in variable A
D j = number of target classes in variable A in factor j
D = total number of target classes
v = many classes in variable A

Furthermore to determine the variable that becomes the root node, used information gain. The variable with the highest information gain will be the root node and then followed by the variable with the lower information gain. Here's the formula for calculating information gain (Han et al., 2012) :

Gain (A) = Info(D) – Info A (D) (3)

Gain (A) = the information gain value of variable A

Based on research conducted by (Id & Astried, 2018), the classification accuracy of the C5.0 algorithm can be improved by boosting. Boosting is the process of turning a weak learner into a strong learner. These changes are made by training several models (multiple models) with the same algorithm so that the best model can be found (Id & Astried, 2018).

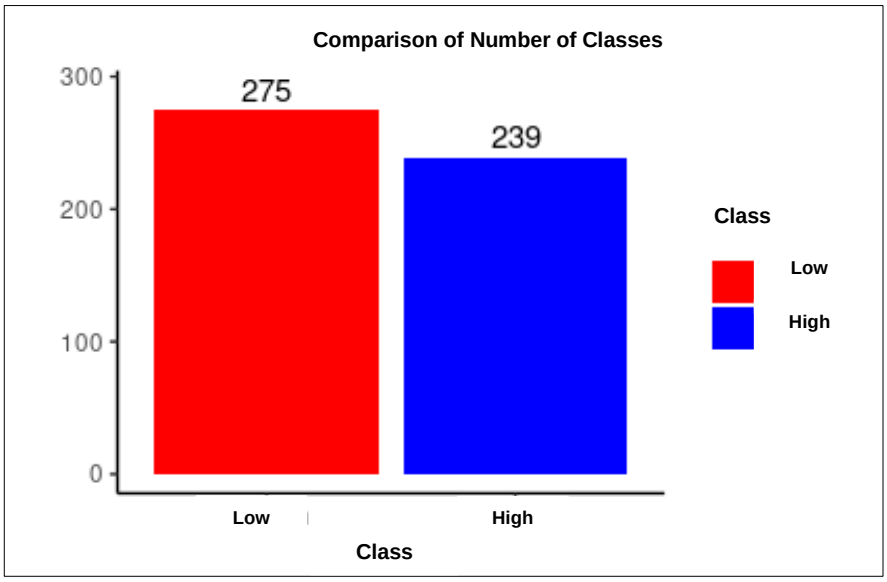

Fig 1. Queuing System for Ahmad Yani International Airport Passenger

RESULT AND DISCUSSION

Before creating a classification model using the C5.0 algorithm, it is necessary to check the class balance of the predictor variables, which consists of Toodler Health Index, Public Health Index, Health Service Index, Health Behavior Index, Non Contagious Disease Index, Contagious Disease Index and Environment Health Index. Classification in the imbalance class will tend to ignore classes that have a small number of samples so that it can adversely affect the performance of the classification algorithm (Triyanto & Kusumaningrum, 2017; Sundaramurthy & Jayavel, 2020).

A dataset where the most common class is twice the least class is only slightly unbalanced, while the dataset with an imbalance ratio of 10:1 will be unbalanced and the dataset with an imbalance ratio of 1000:1 is very unbalanced (Kurniawati, 2019). In the picture below, it can be seen that the classes in the PHDI variable are relatively balanced.

Furthermore, the data will be divided into training data and testing data with a ratio of 80:20. The training data will be used to build the model, while the testing data is used to test the model. The measures used to test the model consist of accuracy, sensitivity, and specificity.

Many model produced use C5.0 boosting algorithms (Rajeswari & Suthendran, 2019; Ayinla, I. B. & Akinola, S. O., 2020), so the result focus on the accuracy of the method. C5.0 boosting algorithms are performed up to 100 iterations and accuracy tends to be stable from 60 iterations. The model is tested with 103 testing data with classification accuracy as in Table 1.

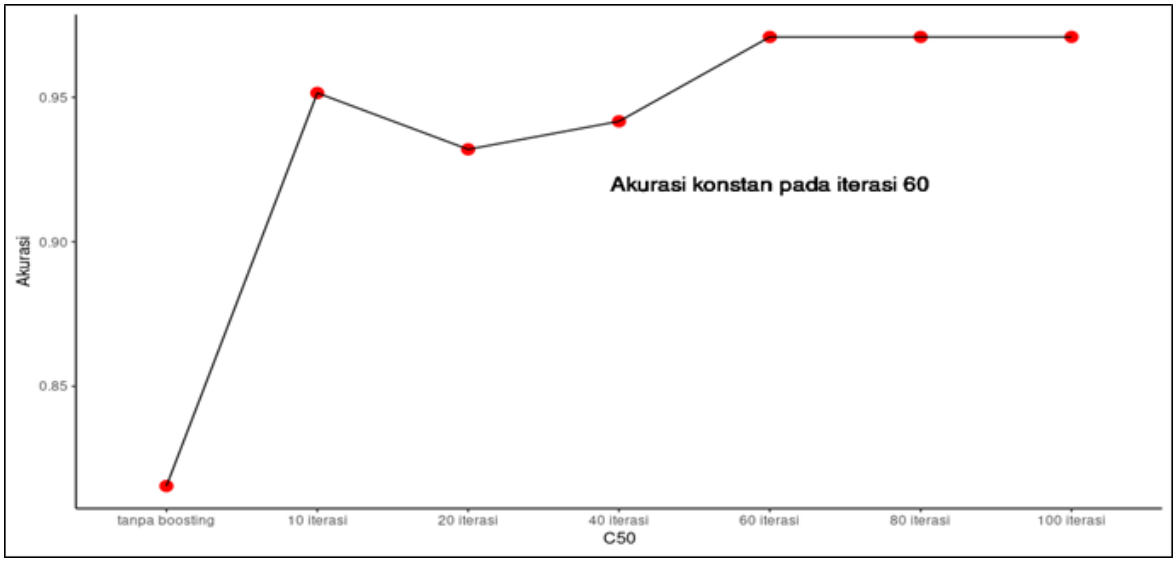

Fig 2. The Accuracy of The Model

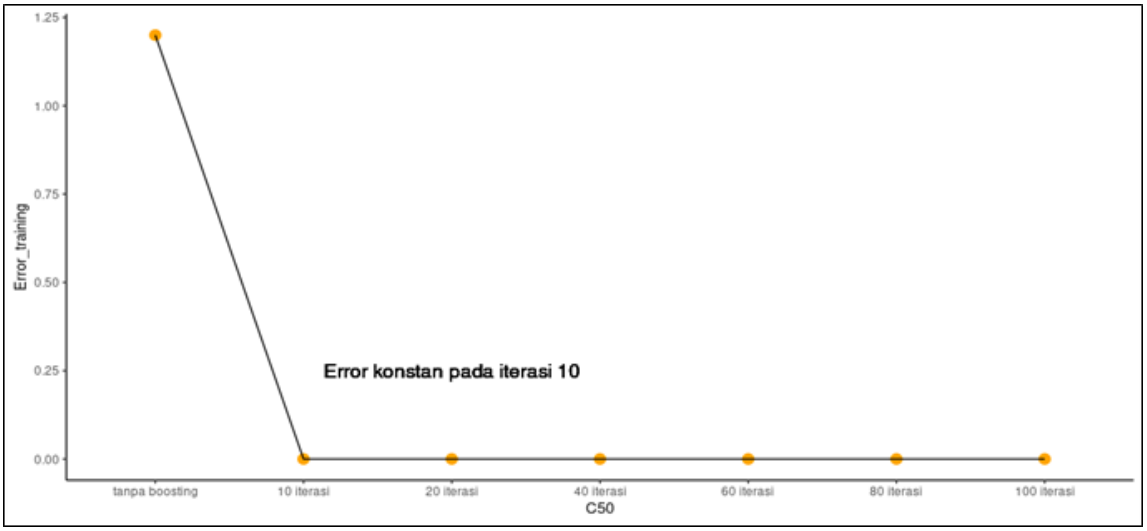

Fig 3. The Constant Training Error

Model testing obtained 97.09% accuracy where only 3 data were misclassified out of a total of 103 testing data. Here's the confusion matrix table:

Table 1. Classification Accuracy

Prediction	Actual	
	Low	High
Low	59	1
High	2	41

Table 2. Confusion Matrix

No	Value
Accuracy	97.09
Sensitivity	96.72%
Specificity	97.62%

Research testing shows the accuracy of the model continues to increase along with the number of iterations. Accuracy was constant at the 60th iteration. The following is a plot of the accuracy of the model. The error rate during model training is also reduced when boosting. The following image shows the constant training error starting from the 10th iteration.

CONCLUSION

The conclusion obtained in this study is that the accuracy of the model increases with the increase in the number of iterations and is constant at the time of the 60th iteration. Error training is also smaller and constant at the 10th iteration. In the other word, boosting can reduce the misclassification and increase the accuracy. So boosting can be used to produce the better model. The final accuracy, sensitivity and specificity were obtained 97.09, 96.72, and 97.62.

ACKNOWLEDGEMENTS

The authors would like to thank profusely to Balitbangkes Indonesian Health Ministry for allowing the authors to retrieve the public health development index data for completion of this study.

REFERENCES

Ayinla, I. B., & Akinola, S. O. (2020). An Improved Collaborative Pruning Using Ant Colony Optimization and Pessimistic Technique of C5.0 Decision Tree Algorithm. https://doi.org/10.5281/ZENODO.4427699

Cahyani, N., & Muslim, M. A. (2020). Increasing Accuracy of C4.5 Algorithm by Applying Discretization and Correlation-based Feature Selection for Chronic Kidney Disease Diagnosis. 12(1), 8.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufman.

Id, I. D. & Astried. (2018). Perbandingan Kinerja Algoritma C5.0 Dan Boosted C5.0 Untuk Klasifikasi Kanker Payudara. Conference: Konverensi Nasional Matematika XIX.

Kementerian Kesehatan RI. (2014). Indeks Pembangunan Kesehatan Masyarakat (IPKM) (1st ed.). Badan Penelitian dan Pengembangan Kesehatan.

Kementerian Kesehatan RI. (2015). Rencana Strategis Kementerian Kesehatan Tahun 2015-2019.

Kurniawati, Y. E. (2019). Class Imbalanced Learning Menggunakan Algoritma Synthetic Minority Over-sampling Technique – Nominal (SMOTE-N) pada Dataset Tuberculosis Anak. Jurnal Buana Informatika, 10(2), 134. https://doi.org/10.24002/jbi.v10i2.2441

Rajeswari, S., & Suthendran, K. (2019). C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. Computers and Electronics in Agriculture, 156, 530–539. https://doi.org/10.1016/j.compag.2018.12.013

Sundaramurthy, S., & Jayavel, P. (2020). A hybrid Grey Wolf Optimization and Particle Swarm Optimization with C4.5 approach for prediction of Rheumatoid Arthritis. Applied Soft Computing, 94, 106500. https://doi.org/10.1016/j.asoc.2020.106500

Suprianto, D., Hasanah, R. N., & Budi S, P. (2013). Sistem Pengenalan Wajah Secara Real-Time dengan Adaboost, Eigenface PCA & MySQL. Jurnal EECCIS, 7(2), 179–184.

Tanyu, B. F., Abbaspour, A., Alimohammadlou, Y., & Tecuci, G. (2021). Landslide susceptibility analyses using Random Forest, C4.5, and C5.0 with balanced and unbalanced datasets. CATENA, 203, 105355. https://doi.org/10.1016/j.catena.2021.105355

Triyanto, A. Y., & Kusumaningrum, R. (2017). Implementasi Teknik Sampling untuk Mengatasi Imbalanced Data pada Penentuan Status Gizi Balita dengan Menggunakan Learning Vector Quantization. Jurnal IPTEK-KOM, 19(17), 39–50.

Wijaya, A. C., Hasibuan, N. A., & Ramadhani, P. (2018). Implementasi Algoritma C5.0 dalam Klasifikasi Pendapatan Masyarakat (Studi Kasus: Kelurahan Masjid Kecamatan Medan Kota). Majalah INTI, 13(2), 192–198.