OPTIMIZATION OF CART ALGORITHM BASED ON ANT BE COLONY FEATURE SELECTION FOR STUNTING DIAGNOSIS

Pungkas Subarkah[1*], Ali Nur Ikhsan[1], Rizki Wahyudi[1], Dayana Rofiqoh[1]
[1]Informatics, Universitas Amikom Purwokerto
email: *subarkah@amikompurwokerto.ac.id

Abstract: One of the main health problems in children is stunting which is one of the concerns in the Sustainable Development Goals (SDGs). Specifically in Indonesia, the prevalence of stunting in 2024 is 21.6%. This figure is still relatively high, because the target prevalence of stunting is 14%. This study aims to implement machine learning knowledge through the Classification And Regression Trees (CART) algorithm based on Ant Be Colony (ABC) feature selection which aims to determine the increase in accuracy in analyzing stunting datasets. The data used comes from Kaggle which consists of 16500 datasets. The dataset consists of gender, age, birth length, birth weight, body length, body weight, breastfeeding and stunting status. The research methods used are data collection, data preprocessing, classification, and evaluation using K-fold cross validation. The results obtained in this research are the implementation of the CART algorithm obtained a value of 89.86% and the results of CART with Ant Be Colony (ABC) feature selection, which obtained an accuracy value of 93.65%. This shows that there is an increase in the accuracy value in the use of CART algorithm optimization and Ant Be Colony (ABC) feature selection by 3.76%. With the research results that have been obtained, it can be categorized as excellent accuracy value excellent. It is hoped that further research can be carried out by adding other classification algorithms or adding feature selection.

Keywords: classification; feature selection; optimazation; stunting

Abstrak: Salah satu masalah kesehatan utama pada anak adalah stunting yang menjadi salah satu perhatian dalam Sustainable Development Goals (SDGs). Khusus di Indonesia angka Pravelensi stunting pada tahun 2024 di angka 21.6%. Angka ini masih tergolong tinggi, karena target angka pravelensi stunting ialah 14%. Penelitian ini bertujuan untuk mengimplementasikan pengetahuan machine learning melalui algoritma Classification And Regression Trees (CART) berbasis seleksi fitur Ant Be Colony (ABC) yang bertujuan untuk mengetahui peningkatan akurasi dalam menganalisis dataset stunting. Data yang digunakan bersumber dari Kaggle yang terdiri dari 16500 dataset. Dataset terdiri dari jenis kelamin, usia, panjang lahir, berat lahir, panjangg badan, berat badan, menyusui dan status stunting. Metode penelitian yang digunakan adalah pengumpulan data, preprocessing data, klasifikasi, dan evaluasi menggunakan K-fold cross validation. Hasil yang diperoleh pada penelitian ini adalah Implementasi algoritma CART memperoleh nilai sebesar 89,86% dan hasil seleksi fitur CART dengan Ant Be Colony (ABC) memperoleh nilai akurasi sebesar 93,65%. Hal ini menunjukkan adanya peningkatan nilai akurasi pada penggunaan optimasi algoritma CART dan pemilihan fitur Ant Be Colony (ABC) sebesar 3,76%. Dengan hasil penelitian yang telah diperoleh dapat dikategorikan nilai akurasi yang diperoleh sangat baik. Diharapkan dapat dilakukan penelitian selanjutnya dengan menambahkan algoritma klasifikasi lain atau menambahkan seleksi fitur.

Kata kunci: klasifikasi; optimalisasi; seleksi fitur; stunting

INTRODUCTION

Stunting is a condition of chronic malnutrition where toddlers fail to grow properly[1]. The impact of stunting on children under five years old is diverse, potentially negative long-term and short-term impacts short-term, stunted children may experience a weakened immune system, increased susceptibility to disease, and poor cognitive development. In the long term, stunting can lead to reduced economic productivity and perpetuate the cycle of poverty[2].

The prevalence of stunting in Indonesia is 21.6% in 2023, which is still high as the target in 2024 is 14%[3]. Due to its severe and diverse impacts, it is important to detect stunting early[4]. There is a need for a new breakthrough through an innovative approach to early detection of stunting, allowing the use of data on various attributes of toddlers to predict stunting conditions[5]. This requires a recent development that is suitable for handling this data to generate a health plan that may give priority to disease management strategies as an early prediction of stunting so that it does not become widespread in Indonesia[6].

Data Mining is a process used to manipulate data by extracting previously unknown information from large datasets[7]. Currently, the role of using data mining is often used in the world of health, one of which is in diagnosing diseases. The use of data mining techniques in predicting stunting diagnosis aims to analyze patient data and develop patient data statistically based on the patient's track record history. In solving prediction problems, the use of methods or techniques aims to simplify the prediction process. In addition, the use of feature selection is used as a process in determining relevant features. This technique or method helps in determining several attributes that will be used in the classification or prediction process[8] .

One of the machine learning algorithms used in classification is the Classfication And Regression Trees (CART) algorithm, this algorithm is a derivative of the trees algorithm. The application of this algorithm has been carried out by previous researchers in various aspects, including the health sector [9], the education sector[10], to other fields. The accuracy results obtained in using this algorithm are quite good[11][12], therefore researchers used the algorithm in this study coupled with Ant Be Colony (ABC) feature selection. The existence of this feature selection is expected to improve the accuracy results in this study, ABC feature selection has also been used by previous researchers, both in the health sector or in other fields[13].

Some previous studies that use the Classification and Regression Trees (CART) algorithm and Ant Be Colony (ABC) feature selection, as follows: First, regarding research conducted by reference [14], the use of the Classification And Regression Trees (CART) algorithm in stroke disease analysis, in his research the researcher used 5110 data, then tested using this algorithm with a division of 70% as training data and 30% as testing data. The results obtained are obtaining an accuracy value of 86.82% with a good category.

While research conducted by Nurhaeka and Putri using the CART algorithm for the classification of BPJS

patient diagnosis results. The purpose of Using the CART algorithm, this study attempts to categorize the outcomes of BPJS patient diagnoses according to the correlation between illness kinds and symptoms. The information utilized is secondary data that was gathered from 200 medical records of BPJS patients who had examination findings at Puskesmas Umbulharjo 1. The analysis's findings indicate that the CART categorization indicates that headache, stomach pain, nausea, and vomiting symptoms are associated with the categories of TB, penunomia, gastritis, and hypertension disorders. Considering the outcomes of the classification accuracy utilizing the CART technique of 69%[15]. Furthermore, research was also conducted by Agiel Fadillah Hermawan et al, the purpose of the research conducted was the initial prediction using the CART algorithm regarding stroke disease. The result obtained is to get an accuracy of 74.73%[16].

The use of Ant Be Colony (ABC) in feature selection, from previous research conducted by reference [17], in analyzing heart disease using ABC, the results obtained increased accuracy to 94.23%. The next research conducted by the reference, that the use of feature selection is very diverse, one of which is in article classification. This is used to determine the suitability of the type of article. The results obtained are an increase in the accuracy obtained, which is 96.4%[18].

. After the above description of the literature review, our research has the novelty of using the Classification And Regression Trees (CART) algorithm based on Ant Be Colony (ABC) feature selection applied to stunting detection cases. This research presents the use of the CART algorithm by combining Ant Be Colony (ABC) feature selection, this

makes the method approach that has never been explored by other researchers.

METHOD

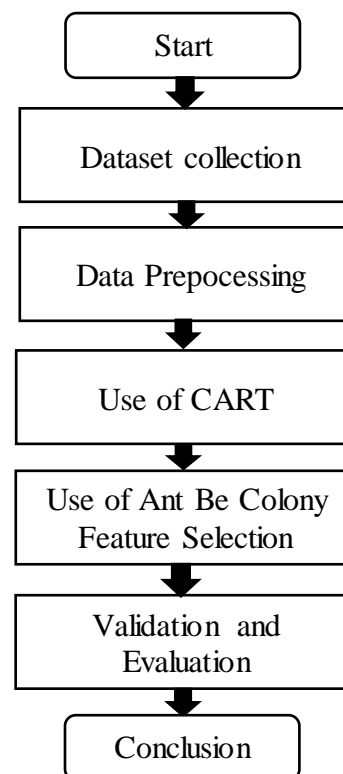Below are the stages in the research, can be seen in image 1.

Image 1. Research Steps

Data Collection

The secondary data utilized in this study comes from the Stunting Dataset on Kaggle, which has eight attributes total—seven feature attributes and one target class attribute. 16500 records with 8 characteristics (seven attributes and one target attribute) make up the data retrieved from the Kaggle Dataset website [19]. The target attribute consists of two outputs, namely patients affected by stunting and patients in undiagnosed by medical.

Data Prepocessing

This stage is carried out a data selection process that aims to get clean data and is ready to be used as research material. The data pre-processing stage includes attribute identification and selection, handling missing values and discretizing values.

CART

Because of its inherent feature selection capabilities, simplicity, and adaptability to more complex tree algorithms like RF and augmented trees, CART was selected. According to earlier research, CART is a potent feature selection algorithm that automatically determines the most crucial characteristics during the tree induction process[20].

Ant Be Colony Feature Selection

Three types of Ant Be Colony (ABC) algorithms may be distinguished: scout, observer, and employed bees. The food supply is the answer to the optimization issue in the ABC algorithm. A appropriateness function associated with the food source's location can be used to gauge its quality. The best solution in the solution space may be found by using the method by which bees look for food sources[21].

Validation and Evaluation

During this phase, validation and assessment are conducted, and the correctness of the outcomes produced by the model is measured using methods included in the Weka program, namely 10-fold cross validation and confusion matrix. In addition, to support the validation and evaluation process, conclusions can be drawn by concluding the use of the CART algorithm based on Ant Be Colony (ABC) optimization based on the precision, recall and F-Measure values.

RESULT AND DISCUSSION

Dataset

Stunting Dataset from Kaggle, consisting of 8 attributes which consist of 7 Feature attributes and 1 goal class attribute. The data obtained from the Kaggle Dataset site is 16500 records which have 8 attributes (7 attributes and 1 target attribute). Here are the details of the stunting dataset in this study:

Table 1. Stunting dataset details

No.	Attribute Name	Description
1.	Gender	Describes the condition by gender classification, male or female
2.	Age	Describes the age of the dataset
3.	Birth Weight	Describes the weight of a person in the dataset
4.	Birth Length	Describes the birth weight of a person in the dataset
5.	Body Weight	Describes the body weight of a person in the dataset
6.	Body Length	Describes the body length of a person in the dataset
7.	Breastfreding	Describes the condition of a person whether breastfeeding or not, indicated by Yes or No.
8.	Stunting	Describes that a person's condition whether diagnosed with stunting or not, marked with Yes or No

Data Preprocessing

After the dataset is carried out the preprocessing stage, including to find out

the emptiness of data or redundant data, it can be generated with the dataset as follows in table 2 below:

Table 2. Data that has been preprocessed

No	Original Data	Preprocessed Data	Decription
1.	F	F	Describes the condition by gender classification, male or female
2.	19	19	Describes the age of the dataset
3.	3	3	Describes the weight of a person in the dataset
4.	52	52	Describes the birth weight of a person in the dataset
5.	7.7	7.7	Describes the body weight of a person in the dataset
6.	57	57	Describes the body length of a person in the dataset
7.	Yes	Yes	Describes the condition of a person whether

No	Original Data	Preprocessed Data	Decription
			breastfeeding or not, indicated by Yes or No.
8.	Yes	Yes	Describes that a person's condition whether diagnosed with stunting or not, marked with Yes or No

.

Furthermore, after the dataset has gone through the preprocessing stage, a clean dataset and the number of classes are also obtained, as follows:

Table 3. Number of datasets by class

No.	Classification	Number of Datasets
1.	Yes Stunting Indicated	11627
2.	Indicated No Stunting	5233
	Total	16500

Use CART and Feature Selection ABC

Then, after the data has been cleaned, the next step is to use the Classification And Regression Trees (CART) classification method and the use of Ant Be Colony (ABC) feature selection, then the following results are obtained:

Table 4. Results of Using CART Algorithm and ABC feature selection

No.	Algoritm	Accuracy result	Precision	Recall	F-Measure
1.	CART	89.86%	0.89	0.76	0.82
2.	CART + ABC	93.65%	0.93	0.84	0.92

From table 4. it can be seen that the use of the CART algorithm and the use of ABC feature selection can increase the accuracy value, from 89.86 to 93.65% there is an increase of 3.79%. The results of using the Ant Be Colony feature selection can be concluded to be excellent classification results. The following is a graph of the increase in accuracy on the stunting dataset using the CART algorithm with Ant Be Colony (ABC) feature selection.

CONCLUSION

Based on the results of the research that has been done, it is obtained that in the optimization of the CART algorithm based on Ant Be Colony feature selection regarding stunting detection, using 16500 datasets of stunting data, the results of the implementation of the CART algorithm obtained a value of 89.86% and the results of CART with Ant Be Colony (ABC) feature selection, which obtained an accuracy value of 93.65%. This shows that there is an increase in the accuracy value in the use of CART algorithm optimization and Ant Be Colony (ABC) feature selection by 3.76%. With the research results that have been obtained, it can be categorized as excellent accuracy value obtained. It is hoped that further research can be carried out by adding other classification algorithms or adding feature selection.

ACKNOWLEGMENTS

BIBLIOGRAPHY

[1] I. Irdawati *et al.*, "Efforts of Increase Cadre Capacity about Stunting Prevention," *MethodsX*, vol. 13, no. February, pp. 1–6, 2024, doi: 10.1016/j.mex.2024.102720.

[2] S. Lonang and D. Normawati, "Klasifikasi Status Stunting Pada Balita Menggunakan K-Nearest Neighbor Dengan Feature Selection Backward Elimination," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 49, 2022, doi: 10.30865/mib.v6i1.3312.

[3] Eko, "149 Juta Anak di Dunia Alami Stunting Sebanyak 6,3 Juta di Indonesia, Wapres Minta Keluarga Prioritaskan Kebutuhan Gizi," PAUDPEDIA. Accessed: Feb. 20, 2024. [Online]. Available: https://paudpedia.kemdikbud.go.id /berita/149-juta-anak-di-dunia-ala mi-stu nting-sebanyak-63-juta-di-indonesia-wapres-minta-keluarga-prioritaskan-kebutuhan- gizi

[4] O. Martony, "Stunting Di Indonesia: Tantangan Dan Solusi Di Era Modern," *Int. J. Technol.*, vol. 47, no. 1, p. 100950, 2023, doi: https://doi.org/10.31539/j oting.v5i2.6930.

[5] A. Yuda, Z. Septina, A. Maharani, and Y. Nurdiantami, "Tinjauan Literatur : Perkembangan Program Penanggulangan Stunting di Indonesia," *J. Epidemiol. Kesehat. Indones.*, vol. 6, no. 2, pp. 53–58, 2023, doi: 10.7454/epidkes.v6i2. 6049.

[6] E. Y. Reza and T. W. Widyaningsih, "Analisis untuk Memprediksi Kualitas Tumbuh Kembang Balita dengan Menerapkan Metode kNN dan Naïve Bayes," *Sist. J. Sist. Inf.*, vol. 13, no. 5, pp. 1865–1875, 2024, doi: https://doi.org/10.32520/stmsi.v13i5.4121.

[7] X. Yin, Y. Zuo, and G. Fu, "Design of intelligent detection method for electricity transmission line equipment defect based on data mining algorithm," *Int. J. Thermofluids*, vol. 24, no. August, p. 100814, 2024, doi: 10.1016/j.ijft.2024.100814.

[8] A. C. Kurniawan and A. Salam, "Seleksi Fitur Information Gain untuk Optimasi Klasifikasi Penyakit Tuberkulosis," vol. 8, pp. 70–82, 2024, doi: 10.30865/mib.v8i1.7122.

[9] G. Bermejo-Martín *et al.*, "Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy," *IEEE Access*, vol. 9, no. 1, p. 121, 2021, doi: 10.1007/s12652-020-02883-2.

[10] P. Subarkah, A. N. Ikhsan, and A. Setyanto, "The effect of the number of attributes on the selection of study program using classification and regression trees algorithms," in *Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*, 2018. doi: 10.1109/ICITISEE.2018.8721030.

[11] P. Subarkah, E. P. Pambudi, S. Oktaviani, and N. Hidayah, "Perbandingan Metode Klasifikasi Data Mining untuk Nasabah Bank Telemarketing," *Matrik J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 1, pp. 139–148, 2020, doi: 10.30812/matrik.v20i1.826.

[12] A. Gayen and H. R. Pourghasemi, *Spatial Modeling of Gully Erosion*. Elsevier Inc., 2019. doi: 10.1016/b978-0-12-815226-3.00030-2.

[13] Q. Gao, H. Xu, and A. Li, "The analysis of commodity demand predication in supply chain network based on particle swarm optimization algorithm," *J. Comput. Appl. Math.*, vol. 400, p. 113760, 2022, doi: 10.1016/j.cam.2021.113760.

[14] M. Firdaus Banjar, Irawati, F. Umar, and L. N. Hayati, "Analysis of stroke classification using Random Forest method," vol. 14, no. 3, pp. 186–193, 2022.

[15] N. Tou and P. M. Endraswari, "Implementasi Data Mining Dalam Klasifikasi Hasil Diagnosa Pasien BPJS Menggunakan Algoritma CART," *JIKA (Jurnal Inform.*, vol. 6, no. 2, p. 170, 2022, doi: 10.31000/jika.v6i2.6164.

[16] A. Fadillah Hermawan, F. Rakhmat Umbara, and F. Kasyidi, "Prediksi Awal Penyakit Stroke Berdasarkan Rekam Medis menggunakan Metode Algoritma CART (Classification and Regression Tree)," *J. MIND J. / ISSN*, vol. 7, no. 2, pp. 151–164, 2022, [Online]. Available: https://doi.org/10.26760/mindjournal.v7i2.151-164

[17] P. D. Putra, S. Sukemi, and D. P. Rini, "Peningkatan Akurasi Klasifikasi Backpropagation Mengguna kan Artificial Bee Colony dan K-

NN Pada Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 208, 2021, doi: 10.30865/mib.v5i1.2634.

[18]   A. K. Nugroho and I. Permadi, "Ant Colony Optimization Untuk Menyeleksi Fitur Dan Klasifikasi Artikel," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 10, no. 1, pp. 223–232, 2019, doi: 10.24176/simet.v10i1.2944.

[19]   MUHTAROM, "Dataset Stunting," Kaggle. Accessed: Jan. 16, 2024. [Online]. Available: https://www.kaggle.com/datasets/ muhtarom/stunting-dataset

[20]   T. H. Wu, P. Y. Chen, C. C. Chen, M. J. Chung, Z. K. Ye, and M. H. Li, "Classification and Regression Tree (CART)-based estimation of soil water content based on meteorological inputs and explorations of hydrodynamics behind," *Agric. Water Manag.*, vol. 299, no. October 2023, 2024, doi: 10.1016/j.agwat.2024.108869.

[21]   L. Wang, L. Xie, and C. Xi, "Improved artificial bee colony algorithm for pressure source parameter inversion of Sakurajima volcano from InSAR data," *Geod. Geodyn.*, vol. 15, no. 6, pp. 635–641, 2024, doi: 10.1016/j.geog.2024.05.004.