Application Of The Metaheuristic Algorithm To Optimize The K Value In K-NN In Grouping Factors Causing Stunting

Antoni[1)], Mbera Mehuli[2)], Satria Yudha Prayogi[3)],
[1,3)]Dosen Program Studi Teknik Informatika, Fakultas Teknik, UISU-Medan
[2)]Dosen FKIP UISU, Medan
antonigtg@ft.uisu.ac.id; mberahuli@fkip.uisu.ac.id; satria.yp@ft.uisu.ac.id

Abstrak

Stunting is a health problem that has a long-term impact on the quality of life of individuals and the development of a nation. Accurately identifying the factors that cause stunting is an important step in developing effective mitigation strategies. K-Nearest Neighbor (K-NN) is a machine learning algorithm that is widely used in data classification and grouping, but its performance is greatly influenced by the selection of optimal K value parameters. This research proposes the application of metaheuristic algorithms, such as genetic algorithms (GA) and Particle Swarm Optimization (PSO), to optimize the K value in K-NN in grouping factors that cause stunting. This method integrates the power of exploration and exploitation of metaheuristic algorithms to find K parameters that produce optimal accuracy. Based on the results of applying the metaheuristic algorithm, it was found that without optimization, K-NN only produces an accuracy of 63%, which shows the importance of choosing the right K value. The use of GA in K-NN optimization provides a substantial increase in accuracy, reaching 73%, thanks to its ability to explore the solution space effectively. Meanwhile, PSO also increases accuracy by up to 74%. It is hoped that the findings of this research will be a significant contribution in the development of a more accurate grouping model for analyzing factors causing stunting, so that it can support data-based decision making in an effort to address the stunting problem holistically.

Keywords : *K-Nearest Neighbor Genetics, PSO, Grouping*

I. INTRODUCTION

Stunting is a public health problem that has become a global concern, especially in developing countries like Indonesia. Stunting is defined as a chronic growth disorder in children caused by malnutrition that lasts for a long period of time, recurrent infections, and lack of psychosocial stimulation during growth and development[1][2]. Children who experience stunting have a height that is far below the average for children their age. In addition, stunting not only has an impact on physical growth, but also on cognitive development, economic productivity, and a person's overall quality of life[3][4]. The stunting problem requires a comprehensive approach that involves systematically identifying the causal factors. Various previous studies have shown that stunting is not only influenced by a single factor, but is the result of complex interactions between various variables[5][6]. Data grouping is an analytical technique that aims to divide data into groups based on similar attributes[7]. This technique can provide deep insight into patterns in the data, enabling researchers and policymakers to identify community groups most vulnerable to stunting. One of the most widely used clustering algorithms is K-Means Clustering.

K-Means is an unsupervised machine learning algorithm that is used to divide data into a certain number of clusters, which is determined by the K parameter[8][9]. This parameter determines the number of clusters to be created, which affects the accuracy of the grouping results. However, one of the main challenges in implementing the K-Means algorithm is determining the optimal K value. Determining an incorrect K value can produce unrepresentative or biased groupings, thereby affecting the validity of the analysis results[10].[11]

The problem in this research is related to the limitations of determining the K value in K-Means. Determining the K value in K-Means is usually done using methods such as the Elbow Method, which is a technique that involves observing changes in inertia values (in clusters) based on the number of clusters and Silhouette Score: Measuring how close a data with data in its cluster compared with other clusters. Although these two methods are quite popular, they have limitations when applied to large or complex datasets. Both approaches also require manual evaluation, which is time consuming and tends to be impractical[12][13]. Therefore, a more efficient and accurate automatic optimization method is needed to determine the optimal K value.

Metaheuristics are algorithmic approaches designed to search for the best solution in a large and complex search space. This algorithm has the ability to find optimal solutions even in conditions where mathematical solutions are difficult to find.

Some popular metaheuristic algorithms include Particle Swarm Optimization (PSO), genetic algorithm and Simulated Annealing[14][15]. The integration of metaheuristic algorithms with K-Means provides an innovative approach for optimizing K values. By using metaheuristics, K values can be determined automatically based on cluster quality evaluations, such as inertia or Davies-Bouldin index[16]. This approach not only improves the accuracy of clustering results, but also reduces manual intervention in the analysis process. Based on this, in the context of stunting, the K-Means algorithm optimized with metaheuristics has great potential to group factors that cause stunting effectively. With more accurate grouping, the most vulnerable groups of society can be identified based on a combination of health, socio-economic and environmental factors. This information is critical for designing more specific and efficient intervention programs. For example, if a particular group shows poor sanitation patterns and low education levels as dominant factors, interventions can be focused on health education and improving sanitation in that area. The aim of this research is to integrate the metaheuristic algorithm with K-Means to optimize the K value, to accurately group factors causing stunting based on multidimensional data and to evaluate the effectiveness of the metaheuristic algorithm in improving grouping results compared to conventional methods.

II. LITERATURE REVIEW

2.1 Stunting data

The dataset used in this research comes from the stunting disease classification competition which will be held in 2022. The data consists of 1000 rows which have each characteristic of stunting disease such as:

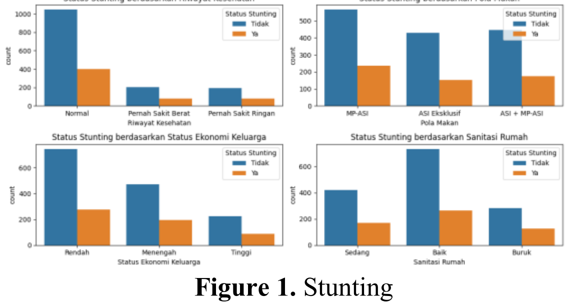

Figure 1. Stunting

2.2 Preprocessing

The preprocessing stage in data processing is a very important first step in the data analysis or machine learning process. The goal is to prepare raw data so that it can be used more effectively by the model or algorithm used. This process involves several techniques to improve data quality, remove noise, and convert it into a suitable format for further analysis.

2.3 K-Nearest Neighbors (KNN)

In this research, the K-Nearest Neighbors (KNN) algorithm was applied to group the factors that cause stunting, a nutritional problem that affects children's growth. KNN is a proximity-based algorithm that works by classifying data based on its nearest neighbors. To get optimal results, one of the challenges in using KNN is determining the K value, namely the number of neighbors used in classification. Inappropriate selection of K values can affect the quality of grouping. Therefore, this research utilizes metaheuristic algorithms, such as Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), to optimize the K value in KNN. With this approach, it is hoped that the K value obtained can increase the accuracy of the model in classifying factors that cause stunting, such as poor nutrition, sanitation and access to health services. Optimizing the KKK value using a metaheuristic algorithm allows finding better values, resulting in a more valid grouping and can provide deeper insight into stunting prevention efforts. In this research, the mathematical formulation in equation 1 is used

$$d\left(x_{test,}x_i\right) = \sqrt{\sum_{j=1}^{m}(x_{test,j-}\ x_{i,j})^2} \quad (1)$$

Basically, the KNN algorithm is used to group data based on its proximity to neighboring data. For an Xtest test data, KNN calculates the distance between the test data and the training data x1, optimize clustering accuracy. The optimal value can be found by minimizing the error rate

$$Error\ rate(K) = \frac{1}{N_{test}}\sum_{i=1}^{N_{test}}[](y_{test,i} \neq y_{true,i}) \quad (2)$$

2.4 Evaluation Model

Evaluation of the clustering performance will be carried out using a confusion matrix and metrics such as F1 Score, precision and recall which are calculated to assess the effectiveness of clustering factors causing stunting. The formula used for performance assessment is presented in Equation 1-4. In this equation, TP corresponds to True Positive, TN represents True Negative, FP represents False Positive, and FN represents False Negative.

$$accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (3)$$

$$Precission = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\ score = \frac{2\ X\ precission\ X\ recall}{precission+recall} \quad (6)$$

III. METHODOLOGY

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and This research uses a quantitative approach by applying a metaheuristic algorithm to optimize the K value in K-Means in grouping factors that cause stunting. This research method was designed systematically and divided into five main stages such as data collection, data pre-processing, data cleaning, grouping with KNN and performance evaluation which are explained in Figure 2 below:

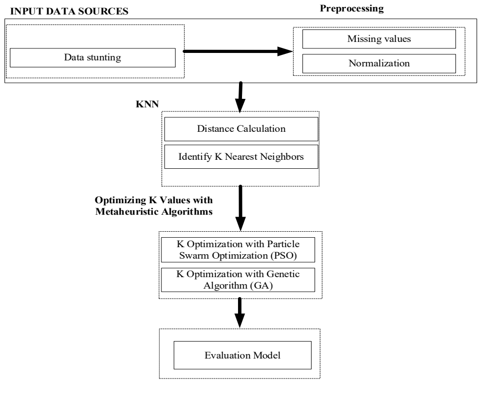

Figure 2. Proposed Method

IV. RESULT AND DISCUSSION

In this section we will discuss the results of the metaheuristic algorithm analysis to optimize the K value of K-Nearest Neighbors (K-NN) in grouping factors that cause stunting. Based on the process carried out in optimizing the K value using the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) algorithms, it results in an increase in the accuracy of the K-NN model compared to the baseline method without optimization. This research carried out various experiments aimed at seeing a comparison between the KNN algorithm without optimizing the K value with the PSO and genetic algorithms. This research will use stunting data with a total of 1000 data. This research focuses on grouping stunting cases. The following is a discussion regarding the implementation of the algorithm carried out.

4.1 Application of KNN algorithm results without optimization

The results of applying the K-Nearest Neighbors (K-NN) algorithm without optimization show less than optimal performance in classifying factors that cause stunting. In testing with various K values manually, the highest accuracy achieved was only 63%, which is relatively lower than the results after optimization. Selection of K values by trial-and-error indicates that K-NN without optimization often experiences overfitting at small K values or underfitting at too large K values. In addition, models without optimization tend to be less adaptive in capturing variations in data patterns, especially in datasets that have an unbalanced distribution or a lot of noise. This shows that although K-NN is simple and easy to implement, without an optimization approach, this model has limitations in optimally utilizing data information to produce accurate grouping. The following are the accuracy results in Figure 3

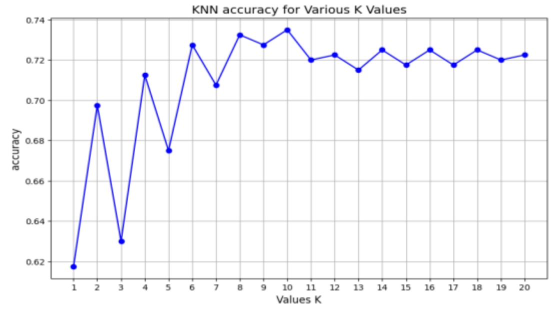

Figure 3. Accuracy in the application of genetic algorithms

4.2 Application of KNN results with genetic algorithm optimization

The application of the K-Nearest Neighbors (K-NN) algorithm with optimization using the Genetic Algorithm (GA) algorithm shows a significant increase in the accuracy of classification of factors causing stunting. In this application, GA is used to find the optimal K value, which is an important parameter in the K-NN algorithm. By utilizing selection, crossover and mutation mechanisms, GA can explore various combinations of K values to find better solutions compared to manually or randomly selecting K values. As a result, the K-NN model optimized with GA achieved an accuracy of 73%, much higher compared to K-NN without optimization which only produced an accuracy of 63%. The advantage of GA lies in its ability to explore a wider solution space and avoid the pitfalls of local solutions that may occur in traditional search methods.

With crossover and mutation mechanisms, GA is able to find more precise K values for complex data, thereby increasing accuracy in grouping factors that cause stunting, such as nutritional consumption patterns, sanitation status, and socio-economic factors. In addition, GA provides flexibility in adjusting the K value to the characteristics of the dataset used, which vary in each condition. The following are the accuracy results for the genetic algorithm in Figure 4

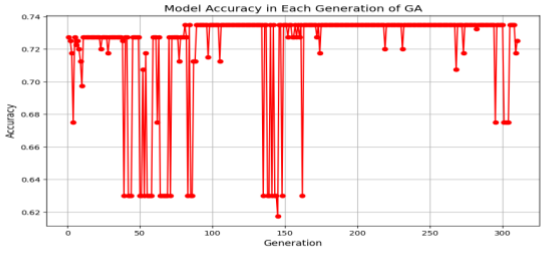

Figure 4. Accuracy in the application of genetic algorithms

4.3 Application of KNN results with PSO algorithm optimization

The application of the K-Nearest Neighbors (K-NN) algorithm with optimization using the Particle Swarm Optimization (PSO) algorithm also shows a significant increase in the accuracy of grouping factors that cause stunting. In this approach, PSO is used to optimize the K value in K-NN, which plays an important role in determining the number of nearest neighbors used for classification. PSO, with its swarm intelligence mechanism, works by moving "particles" (potential solutions) in the search space to find the optimal K value, relying on the evaluation of fitness values to iteratively improve the position of the particles. The results obtained from applying PSO to K-NN show an average accuracy of 74%, slightly higher than the optimization results using the Genetic Algorithm (GA) algorithm, but still higher than K-NN without optimization. The following are the accuracy results in Figure 5

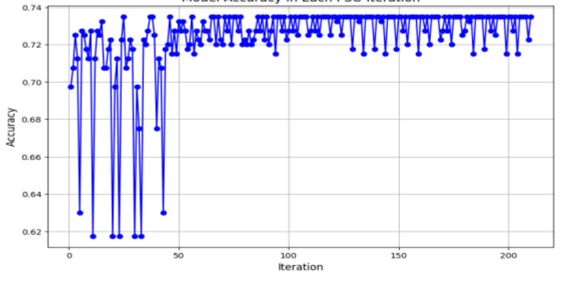

Figure 5. Accuracy in the application of genetic algorithms

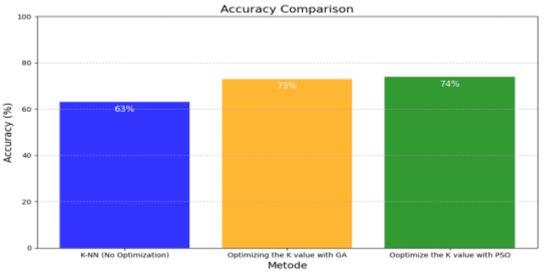

Figure 6. Accuracy Comparison

V. DISCUSSION

In this research, the application of the K-Nearest Neighbors (K-NN) algorithm in grouping factors causing stunting with optimization using metaheuristic algorithms, namely Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), shows significant results in increasing model accuracy compared to K-NN without optimization. K-NN is a very popular algorithm in classification, but its performance is very dependent on choosing the optimal K value. Selecting an inappropriate K value can cause the model to experience overfitting or underfitting, which in turn will reduce prediction accuracy. Therefore, optimizing the K value is very important to obtain optimal results in grouping factors that cause stunting. The application of GA in K-NN optimization provides a significant increase in accuracy, reaching 73%, which is higher compared to the K-NN model without optimization which only reaches 63%. This shows that GA is able to explore the solution space more effectively to find the K value that best suits the characteristics of the data. The advantage of GA lies in its selection, crossover and mutation mechanisms which make it possible to explore various combinations of solutions in a broad and adaptive manner. Meanwhile, optimization with PSO also provides a significant improvement over K-NN, with accuracy reaching 74%. Although slightly inferior compared to GA, PSO has an advantage in terms of time efficiency. PSO bekerja dengan cara yang lebih sederhana, menggunakan swarm intelligence untuk mencari optimal solution, thus requiring fewer iterations and faster convergence compared to GA. This advantage makes PSO more efficient in applications with limited computing time, especially on large datasets. However, PSO also has disadvantages related to sensitivity to initial parameter selection and the risk of getting stuck in local solutions. In other words, although PSO can provide good results in a shorter time, the results can be influenced by less than optimal parameter settings. To overcome this, further experiments are needed regarding the selection of appropriate parameters for PSO, in order to maximize its performance.

Comparison between K-NN without optimization and K-NN optimized with GA and PSO shows that metaheuristic optimization provides a clear improvement in model performance. K-NN without optimization is unable to accommodate data complexity effectively, while the application of GA and PSO brings great benefits in classification accuracy. Although GA provides the best results in accuracy, PSO offers higher efficiency in computing time, making it a better choice in situations with tight time constraints. This research shows that the integration of metaheuristic algorithms with K-NN is very useful in classifying factors causing stunting, and provides important insights in improving the quality of public health analysis. For future research, it can be considered to combine GA and PSO or integrate other optimization techniques to obtain an optimal balance between accuracy and time efficiency, as well as to increase model stability on larger and more complex datasets.

VI. CONCLUSION

In this research, the application of the K-Nearest Neighbors (K-NN) algorithm to group factors causing stunting by optimizing using metaheuristic algorithms, namely Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), showed significant results. Without optimization, K-NN only produces an accuracy of 63%, which shows the importance of choosing the right K value. The use of GA in K-NN optimization provides a substantial increase in accuracy, reaching 73%, thanks to its ability to explore the solution space effectively. Meanwhile, PSO also increases accuracy by up to 74%, with the advantage of better time efficiency compared to GA. Although GA provides the highest accuracy, longer computing time is an obstacle, whereas PSO offers a faster solution albeit with a slight decrease in accuracy. Overall, optimization using GA and PSO proved effective in improving K-NN performance, both in terms of accuracy and time efficiency. This research shows that the combination of K-NN with metaheuristic optimization algorithms can be a very useful tool in classifying factors causing stunting, providing important insights in public health analysis. In the future, further research can combine the two algorithms to obtain an optimal balance between accuracy and computing time.

REFERENCES

[1] S. M. E. Sharifnia *et al.*, "Decision support framework for home health caregiver allocation using optimally tuned spectral clustering and genetic algorithm," *Healthc. Anal.*, vol. 5, p. 100339, 2024.

[2] S.Aryza et al (2024) A ROBUST OPTIMIZATION TO DYNAMIC SUPPLIER DECISIONS AND SUPPLY ALLOCATION PROBLEMS IN THE MULTI-RETAIL INDUSTRY. *Eastern-European Journal of Enterprise Technologies*, *129*(3).

[3] Aryza, S., Efendi, S., & Sihombing, P. (2024). Enhanced optimization model decision efficient multi product retail. *International Journal of Electronics and Telecommunications*, 661-666.

[4] N. Hasdyna, R. K. Dinata, Rahmi, and T. I. Fajri, "Hybrid Machine Learning for Stunting Prevalence: A Novel Comprehensive Approach to Its Classification, Prediction, and Clustering Optimization in Aceh, Indonesia," in *Informatics*, MDPI, 2024, p. 89.

[5] O. N. Chilyabanyama *et al.*, "Performance of machine learning classifiers in classifying stunting among under-five children in Zambia," *Children*, vol. 9, no. 7, p. 1082, 2022.

[6] 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, p. 8621, 2020.

[7] H. Shen, H. Zhao, and Y. Jiang, "Machine learning algorithms for predicting stunting among under-five children in Papua New Guinea," *Children*, vol. 10, no. 10, p. 1638, 2023.

[8] C. Fannany, P. H. Gunawan, and N. Aquarini, "Machine Learning Classification Analysis for Proactive Prevention of Child Stunting in Bojongsoang: A Comparative Study," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, 2024, pp. 1–5.

[9] A. Al Jawaldeh *et al.*, "Tackling childhood stunting in the Eastern Mediterranean Region in the context of COVID-19," *Children*, vol. 7, no. 11, p. 239, 2020.

[10] R. Hidayat, A. Bejo, S. Sumaryono, and A. Winursito, "Denoising speech for MFCC feature extraction using wavelet transformation in speech recognition system," *Proc. 2018 10th Int. Conf. Inf. Technol. Electr. Eng. Smart Technol. Better Soc. ICITEE 2018*, pp. 280–284, 2018, doi: 10.1109/ICITEED.2018.8534807.

[9] Pratibha, A. Gahalot, Uprant, S. Dhiman, and L. Chouhan, "Crime prediction and analysis," *2nd Int. Conf. Data, Eng. Appl. IDEA 2020*, 2020, doi: 10.1109/IDEA49133.2020.9170731.

[10] M. Arfah, F. Fachrizal, and O. Nugroho, "Developing A Model Of Association Rules With Machine Learning In Determining User Habits On Social Media.," *Eastern-European J. Enterp. Technol.*, no. 2, 2024.

[11] A. Rahmatika, O. Nugroho, and T. A. AnuR, "Using Relational Learning In Exploring The Effectiveness Of Using Hashtags In Future Topics And User Relations In X.," *Eastern-European J. Enterp. Technol.*, no. 2, 2024.

[12] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Eng.*, vol. 69, no. November 2013, pp. 1356–1364, 2014, doi: 10.1016/j.proeng.2014.03.129.

[13] A. R. Lubis, S. Prayudani, Al-Khowarizmi, Y. Y. Lase, and Y. Fatmi, "Similarity Normalized Euclidean Distance on KNN Method to Classify Image of Skin Cancer," in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2021, pp. 68–73. doi: 10.1109/ISRITI54043.2021.9702826.

[14] H. Janawisuta and P. H. Gunawan, "Early Detection of Stunting in Indonesian Toddlers: A Machine Learning Approach," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, 2024, pp. 12–16.

[15] L. Atamou, D. C. Rahmadiyah, H. Hassan, and A. Setiawan, "Analysis of the determinants of stunting among children aged below five years in stunting locus villages in Indonesia," in *Healthcare*, MDPI, 2023, p. 810.

[16] D. I. Yani, L. Rahayuwati, C. W. M. Sari, M. Komariah, and S. R. Fauziah, "Family household characteristics and stunting: an update scoping review," *Nutrients*, vol. 15, no. 1, p. 233, 2023.