https://jurnal.seaninstitute.or.id/index.php/jukomi
Jurnal Komputer Indonesia (JU-KOMI)
Volume 3 No 02, 2025
JU-KOMI
E-ISSN: 2963-0460

Implementation of C4.5 Algorithm for Diarrhea Prediction

[1]Sipra Barutu, [2]Siska Simamora
[1]Program Studi Teknologi Informasi, Universitas Panca Budi, [2]Program Studi Teknologi Informasi, Universitas Putra Abadi Langkat

ARTICLE INFO

Keywords:
C4.5 Algorithm, Prediction, Diarrhea, Data Mining, Classification

Email :
barutusipra@gmail.com

ABSTRACT

Diarrheal disease remains one of the major health problems among toddlers in Indonesia. Environmental factors such as drinking water quality, sanitation, mothers' hand hygiene, and immunization status play an important role in influencing the occurrence of diarrhea. This study aims to analyze the application of the C4.5 algorithm in developing a predictive model for diarrhea among toddlers using secondary data from a Public Health Center (Puskesmas), consisting of 200 records divided into 150 training data and 50 testing data. The analysis process was carried out through entropy calculation, information gain assessment, and decision tree construction to obtain classification patterns. The results showed that the C4.5 model achieved an accuracy of 92%, precision of 87.5%, recall of 87.5%, F1-score of 87.5%, and specificity of 94.12%. These values indicate that the C4.5 algorithm is capable of making predictions with a good level of accuracy and balance in detecting both positive and negative cases. This study contributes to the utilization of data mining, particularly the C4.5 algorithm, as a decision-support tool in the health sector for the prevention of diarrheal disease among toddlers.

INTRODUCTION

Diarrhea is one of the leading causes of morbidity and mortality among toddlers in many developing countries, including Indonesia. Risk factors such as poor drinking water quality, inadequate environmental sanitation, lack of hand hygiene, and incomplete immunization remain major challenges in efforts to prevent this disease. Therefore, a predictive method is needed to help healthcare workers identify factors that contribute to the occurrence of diarrhea cases in toddlers.

The detection rate of diarrhea cases among toddlers in Indonesia remains low. In 2021, the detection rate reached only 22.18% of the established target—around 818,687 toddlers out of the target of 3,690,984. This indicates a significant challenge in achieving effective public health program goals, due to various factors such as frequent changes in program management, low accuracy and completeness of reports, and lack of data integration across programs (Ministry of Health of the Republic of Indonesia, 2022).

Based on the information presented, it is necessary to address diarrhea among toddlers through the utilization of technology, one of which is predictive analysis using machine learning algorithms. The C4.5 algorithm is a widely used technique in disease classification. Related research conducted by Sepharni (2022) applied the C4.5 algorithm to heart disease classification and achieved an accuracy rate of 75.92%.

The development of data mining technology enables more systematic analysis of health data through the application of various classification algorithms. One of the most widely used is the C4.5 algorithm, which can generate decision trees by considering the most informative attributes based on information gain values. Through this method, the relationship patterns between environmental and health factors can be mapped to facilitate early detection and disease prevention. The purpose of this study is to present data used for diarrhea prediction, apply the C4.5 algorithm to determine predictive patterns of diarrheal disease, and implement precision, recall, and F1-score techniques as part of algorithm comparison and evaluation.

METHOD

The research method used in this study is a quantitative approach with an experimental design. The data utilized are secondary data obtained from medical records and health reports available at the Public Health Center (Puskesmas).

Subsequently, the data were processed and analyzed using the C4.5 algorithm to develop a predictive model that provides information on the factors influencing the occurrence of diarrhea among toddlers. The dataset consists of 200 records, divided into training and testing data—150 used as training data and 50 as testing data.

This section focuses on detailing the variables used in the study systematically to ensure clarity in the data analysis process. Each variable, including water quality, environmental sanitation, hand hygiene, immunization, and diarrhea status, is described through its operational definition and role within the analytical framework.

The study includes several key stages, starting with data collection, data preprocessing, and continuing to the application and evaluation of the predictive model. The evaluation results are then used to compare the performance of the algorithms based on various performance metrics such as accuracy, precision, recall, and F1-score. The findings of this research are expected to contribute to determining the most effective algorithm for predicting diarrheal disease among toddlers in Public Health Centers.

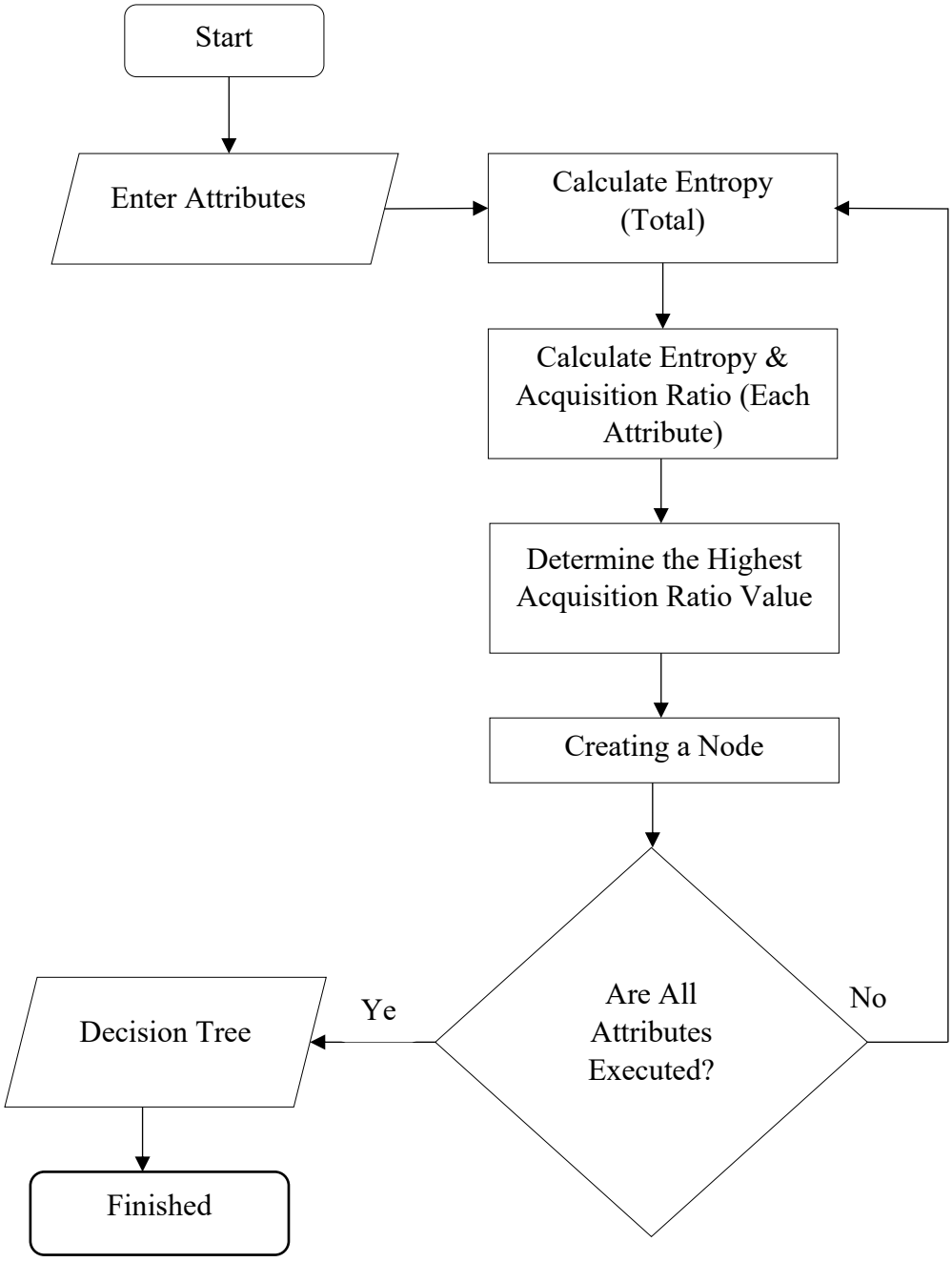

Figure 1 Flowchart of the C4.5 Algorithm

$$\boldsymbol{Enthropy(S) - (x + a)^n = \sum_{j=1}^{k} -\mathrm{P\ j\ Log2\ P\ j}}$$

Explanation:
S = set of cases
n = number of partitions of S
P_i = proportion of S_i to S

$$\mathrm{Gain}(S, A) = Enthropy(S) - (x + a)^n + \sum_{i=1}^{n} -\frac{\mathrm{Si}}{\mathrm{S}} * Enthropy(Si)$$

Explanation:

S : set of cases
A : attribute
n : number of partitions of attribute A
Si\ : Proportion of S_i to S
|S^r| : number of cases in S

RESULTS AND DISCUSSION

Based on the results of the calculations carried out, both through the evaluation of total entropy and the computation of the information gain of each attribute, the attribute with the highest gain value was identified and selected as the root of the decision tree. The subsequent tree construction process was performed recursively on each branch until all data could be optimally classified. The result of this process is visualized in the form of a decision tree structure, as shown in the following figure. Each node in the tree represents an attribute selected based on its contribution to data separation, while each branch indicates the corresponding attribute value. The leaf nodes represent the final output or classification result based on the combination of attributes traversed from the root to the terminal node.

The resulting tree structure provides an illustration of how combinations of attributes such as Environmental Sanitation, Drinking Water Quality, and Mother's Hand Hygiene collectively influence the classification outcome. This demonstrates the effectiveness of the C4.5 algorithm in identifying patterns within the data through the selection of the most informative attributes at each level of branching.

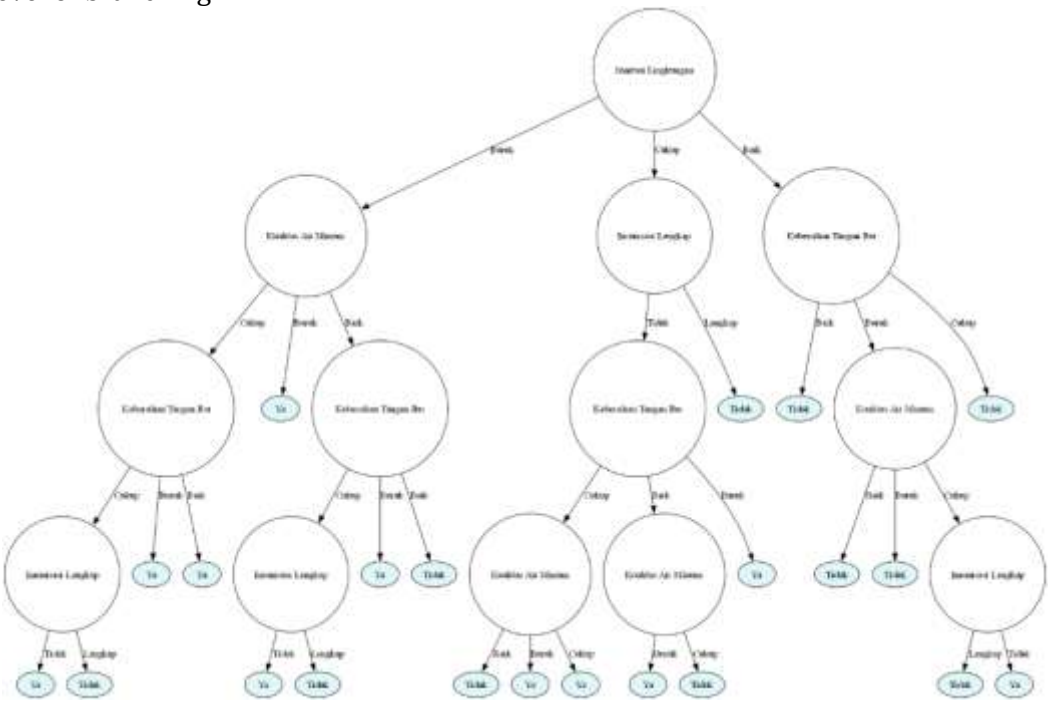

Figure 2. C4.5 Decision Tree

Based on the structure of the constructed decision tree, classification patterns can be formulated to represent the decision paths from the root to the leaf nodes. These patterns are summarized in the following table, which presents the combination of attributes in each path along with the corresponding classification results, as shown in Table 1 below:

Table 1 Prediction Rules Using the C4.5 Method

Rule	Description
Rule 1	IF Environmental Sanitation = Good AND Mother's Hand Hygiene = Good THEN Diarrhea = No
Rule 2	IF Environmental Sanitation = Good AND Mother's Hand Hygiene = Fair THEN Diarrhea = No

Rule 3	IF Environmental Sanitation = Good AND Mother's Hand Hygiene = Poor AND Drinking Water Quality = Good THEN Diarrhea = No
Rule 4	IF Environmental Sanitation = Good AND Mother's Hand Hygiene = Poor AND Drinking Water Quality = Poor THEN Diarrhea = No
Rule 5	IF Environmental Sanitation = Good AND Mother's Hand Hygiene = Poor AND Drinking Water Quality = Fair AND Immunization = Complete THEN Diarrhea = No
Rule 6	IF Environmental Sanitation = Good AND Mother's Hand Hygiene = Poor AND Drinking Water Quality = Fair AND Immunization = Incomplete THEN Diarrhea = Yes
Rule 7	IF Environmental Sanitation = Fair AND Immunization = Complete THEN Diarrhea = No
Rule 8	IF Environmental Sanitation = Fair AND Immunization = Incomplete AND Mother's Hand Hygiene = Poor THEN Diarrhea = Yes
Rule 9	IF Environmental Sanitation = Fair AND Immunization = Incomplete AND Mother's Hand Hygiene = Good AND Drinking Water Quality = Fair THEN Diarrhea = No
Rule 10	IF Environmental Sanitation = Fair AND Immunization = Incomplete AND Mother's Hand Hygiene = Good AND Drinking Water Quality = Poor THEN Diarrhea = Yes
Rule 11	IF Environmental Sanitation = Fair AND Immunization = Incomplete AND Mother's Hand Hygiene = Fair AND Drinking Water Quality = Good THEN Diarrhea = No
Rule 12	IF Environmental Sanitation = Fair AND Immunization = Incomplete AND Mother's Hand Hygiene = Fair AND Drinking Water Quality = Fair THEN Diarrhea = Yes
Rule 13	IF Environmental Sanitation = Fair AND Immunization = Incomplete AND Mother's Hand Hygiene = Fair AND Drinking Water Quality = Poor THEN Diarrhea = Yes
Rule 14	IF Environmental Sanitation = Poor AND Drinking Water Quality = Poor THEN Diarrhea = Yes
Rule 15	IF Environmental Sanitation = Poor AND Drinking Water Quality = Good AND Mother's Hand Hygiene = Good THEN Diarrhea = No
Rule 16	IF Environmental Sanitation = Poor AND Drinking Water Quality = Good AND Mother's Hand Hygiene = Poor THEN Diarrhea = Yes
Rule 17	IF Environmental Sanitation = Poor AND Drinking Water Quality = Good AND Mother's Hand Hygiene = Fair AND Immunization = Complete THEN Diarrhea = No
Rule 18	IF Environmental Sanitation = Poor AND Drinking Water Quality = Good AND Mother's Hand Hygiene = Fair AND Immunization = Incomplete THEN Diarrhea = Yes
Rule 19	IF Environmental Sanitation = Poor AND Drinking Water Quality = Fair AND Mother's Hand Hygiene = Good THEN Diarrhea = Yes
Rule 20	IF Environmental Sanitation = Poor AND Drinking Water Quality = Fair AND Mother's Hand Hygiene = Poor THEN Diarrhea = Yes
Rule 21	IF Environmental Sanitation = Poor AND Drinking Water Quality = Fair AND Mother's Hand Hygiene = Fair AND Immunization = Complete THEN Diarrhea = No
Rule 22	IF Environmental Sanitation = Poor AND Drinking Water Quality = Fair AND Mother's Hand Hygiene = Fair AND Immunization = Incomplete THEN Diarrhea = Yes

C4.5 Model Testing

The testing of the C4.5 model aims to evaluate the algorithm's ability to classify data based on relevant attributes through the construction of a decision tree. The model's prediction results are then compared with the actual data to calculate evaluation metrics, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), which are subsequently used to assess the model's accuracy, precision, recall, and F1-score. This evaluation is essential to understand how well the model can accurately distinguish between target classes and identify potential classification errors. Based on these prediction results, a summary of the classification outcomes is presented in the following table.

Table 2 C4.5 Model Testing

No	Drinking Water Quality	Environmental Sanitation	Mother's Hand Hygiene	Immunization	Actual	Predicted	Confusion Matrix
1	Poor	Good	Fair	Complete	No	No	TN
2	Fair	Fair	Poor	Complete	No	No	TN
3	Good	Good	Poor	Complete	No	No	TN
4	Good	Fair	Fair	Complete	No	No	TN
5	Poor	Good	Good	Complete	No	No	TN
6	Fair	Poor	Poor	Incomplete	Yes	Yes	TP
7	Good	Fair	Good	Incomplete	No	No	TN
8	Poor	Good	Good	Incomplete	No	No	TN
9	Good	Good	Fair	Complete	No	No	TN
10	Fair	Good	Fair	Complete	No	No	TN
11	Fair	Fair	Good	Complete	No	No	TN
12	Poor	Good	Fair	Complete	No	No	TN
13	Good	Fair	Fair	Complete	No	No	TN
14	Good	Good	Good	Complete	No	No	TN
15	Fair	Good	Fair	Complete	No	No	TN
16	Fair	Good	Poor	Complete	No	No	TN
17	Poor	Poor	Poor	Complete	Yes	Yes	TP
18	Good	Poor	Poor	Complete	No	Yes	FP
19	Poor	Good	Fair	Complete	No	No	TN
20	Poor	Good	Fair	Complete	No	No	TN
21	Fair	Fair	Fair	Complete	No	No	TN
22	Fair	Good	Fair	Incomplete	No	No	TN
23	Poor	Fair	Poor	Complete	Yes	No	FN
24	Fair	Fair	Poor	Incomplete	Yes	Yes	TP
25	Good	Fair	Good	Incomplete	No	No	TN
26	Fair	Fair	Fair	Complete	No	No	TN
27	Fair	Good	Fair	Incomplete	No	No	TN
28	Fair	Good	Fair	Incomplete	No	No	TN
29	Poor	Poor	Good	Incomplete	Yes	Yes	TP
30	Fair	Fair	Good	Complete	No	No	TN
31	Good	Poor	Good	Incomplete	No	No	TN
32	Good	Fair	Good	Complete	No	No	TN
33	Good	Fair	Good	Incomplete	No	No	TN
34	Poor	Poor	Fair	Incomplete	Yes	Yes	TP
35	Fair	Good	Poor	Complete	No	No	TN
36	Fair	Poor	Good	Incomplete	Yes	Yes	TP
37	Good	Poor	Good	Incomplete	No	No	TN
38	Poor	Fair	Fair	Complete	No	No	TN
39	Poor	Poor	Fair	Complete	Yes	Yes	TP
40	Poor	Poor	Poor	Incomplete	Yes	Yes	TP
41	Poor	Poor	Poor	Incomplete	Yes	Yes	TP
42	Poor	Good	Fair	Complete	No	No	TN
43	Fair	Poor	Poor	Incomplete	Yes	Yes	TP
44	Poor	Fair	Poor	Complete	Yes	No	FN
45	Poor	Poor	Good	Complete	No	Yes	FP
46	Poor	Poor	Fair	Incomplete	Yes	Yes	TP

47	Fair	Poor	Poor	Incomplete	Yes	Yes	TP
48	Good	Poor	Fair	Incomplete	Yes	Yes	TP
49	Poor	Poor	Poor	Incomplete	Yes	Yes	TP
50	Good	Poor	Good	Complete	No	No	TN

To gain a more comprehensive understanding of the performance of the classification model used, an analysis of the prediction results was conducted using a confusion matrix. A confusion matrix is a tabular representation that describes the comparison between the model's prediction results and the actual conditions in the test data. With this approach, the quality of predictions can be evaluated systematically through four main components: True Negative (TN) with a total of 32, True Positive (TP) with a total of 14, False Positive (FP) with a total of 2, and False Negative (FN) with a total of 2. This information provides an initial overview that the model is able to classify most of the data accurately and only produces a few prediction errors. All of these summaries are outlined in the following graph.

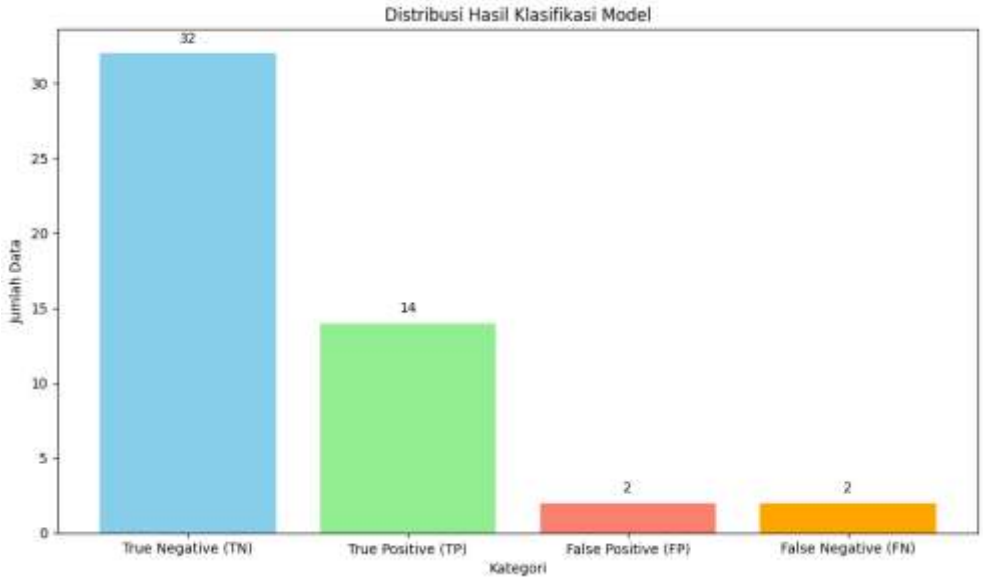

Figure 3 Confusion matrix for the C4.5 method

The evaluation of the classification model's performance is not only carried out through the confusion matrix but also through the calculation of several quantitative evaluation metrics. These metrics provide a more detailed picture of the model's predictive quality from various perspectives, such as overall accuracy, precision in predicting the positive class, sensitivity, and the ability to avoid classification errors.

Table 3 C4.5 Evaluation Results

Metric	**Value**
Accuracy	0.92
Precision	0.875
Recall (Sensitivity / TPR)	0.875
F1-Score	0.875
Specificity (TNR)	0.9412
False Positive Rate (FPR)	0.0588
True Positive Rate (TPR)	0.875

Based on the calculation results, the obtained Accuracy value is 0.9200, indicating that 92% of the model's total predictions fall into the correct class. The Precision value of 0.8750 suggests that among all instances predicted as positive (diarrhea), 87.5% are indeed true positives. The Recall or True Positive Rate (TPR), also at 0.8750, demonstrates that the model successfully detected 87.5% of all actual diarrhea cases. Furthermore, the F1-Score value of 0.8750 represents

the balance between precision and recall, which is crucial in situations where the data distribution between classes is imbalanced. The Specificity (TNR) value of 0.9412 indicates that the model performs very well in identifying negative cases (no diarrhea). Meanwhile, the False Positive Rate (FPR) of only 0.0588 reflects the model's low error rate in misclassifying negative data as positive.

CONCLUSION

Based on the results of the C4.5 algorithm using various measurement metrics, it can be concluded that the C4.5 algorithm shows excellent classification performance with an accuracy value of 0.92. The precision, recall, and F1-score values, which are all 0.875, indicate a balance between the ability to detect positive classes and the accuracy of the predictions provided. The high specificity, which is 0.9412, shows that this algorithm is able to recognize negative data well, which is supported by a low false positive rate of 0.0588. Further research is recommended to test the algorithm on datasets with different characteristics, such as unbalanced data distribution or higher dimensions, in order to evaluate the stability and generalization of the model more comprehensively.

REFERENCES

Kementerian Kesehatan Republik Indonesia. (2018). Riset Kesehatan Dasar 2018. Jakarta: Kemenkes RI.

Kementerian Kesehatan Republik Indonesia. (2020). Profil Kesehatan Indonesia 2020. Jakarta: Kemenkes RI.

Kementerian Kesehatan Republik Indonesia. (2022). Laporan Kesehatan Nasional 2022. Jakarta: Kemenkes RI.

Sepharni. (2022). Klasifikasi Penyakit Jantung Menggunakan Algoritma C4.5. Jurnal Informatika, 10(2), 75-92.

Depari, Widiastiwi, & Santoni. (2022). Perbandingan Algoritma Machine Learning dalam Klasifikasi Penyakit Jantung. Jurnal Kesehatan Digital, 15(3), 70-85.

Munggaran, & Hidayatulloh. (2015). Penerapan Algoritma C4.5 untuk Diagnosa Penyakit Diare Pada Anak Balita Berbasis Mobile. Jurnal Sistem Informasi, 8(1), 55-67.

Ente, et al. (2020). Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5. Jurnal Ilmu Komputer, 12(2), 98-112.

Afifuddin, & Hakim. (2023). Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5. Jurnal Teknologi Informasi, 19(1), 33-45.

Prabowo, et al. (2023). Komparasi Tingkat Akurasi Random Forest dan Decision Tree C4.5 Pada Klasifikasi Data Penyakit Infertilitas. Jurnal Kesehatan Digital, 17(4), 88-102.

Kalimah. (2022). Klasifikasi Penyakit Diabetes Menggunakan Metode Decision Tree dan Random Forest. Jurnal Informatika Medis, 14(3), 67-79.

Masriadi. (2017). Epidemiologi Penyakit Diare. Makassar: Universitas Hasanuddin Press.

Purnama. (2016). Penyakit Diare dan Faktor Risikonya. Jakarta: Pustaka Kesehatan.

Simatupang. (2004). Rotavirus dan Perannya dalam Diare pada Anak. Jurnal Kedokteran Indonesia, 10(2), 44-55.

Nikma Kumala Sari, & Almansyah Lukito. (2017). Faktor Penyebab dan Pencegahan Diare pada Balita. Jurnal Kesehatan Masyarakat, 12(1), 78-91.

Hassan, & Alatas. (1985). Patogenesis dan Pencegahan Diare pada Anak. Jakarta: Balai Pustaka.

Kliegman, Marcdante, & Jenson. (2006). Nelson Textbook of Pediatrics. Philadelphia: Elsevier.

Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.