JTIS : Journal of Technopreneurship and Information Systems
Vol. 08, No. 1, Februari 2025, hal. 1~7
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 5
DOI: 10.36085

Analisis Algoritma C4.5 untuk Prediksi Minat Baca

[1]Muhajir Yunus, [2]Azminuddin I. S. Aziz, [3]Fitriah
[1]Universitas Muhammadiyah Gorontalo, Indonesia
[2]Institut Teknologi Bacharuddin Jusuf Habibie, Indonesia
[3]Universitas Muhammadiyah Bengkulu, Indonesia
[1]muhajiryunus@gmail.com; [2]azminuddinaziz@ith.ac.id ; [3]fitriah@umb.ac.id;

Article Info

Article history:

Received, 2025-01-15
Revised, 2025-01-25
Accepted, 2025-02-15

Kata Kunci:

Decision tree C4.5
Minat baca
Machine learning
Supervised Learning

ABSTRAK

Minat baca merupakan indikator penting dari tingkat literasi dan berkorelasi langsung dengan kemampuan berpikir analitis dan kualitas pendidikan secara keseluruhan. Prediksi minat baca dapat diselesaikan dengan pendekatan machine leraning menggunakan algoritma C4.5 yang handal dalam mengolah data. Berdasarkan hasil analisis yang telah dilakukan, diperoleh pohon keputusan C4.5 untuk prediksi minat baca, di mana variabel lingkungan membaca tidak mempengaruhi prediksi minat baca, sedangkan variabel umur yang paling berpengaruh terhadap prediksi minat baca. Sedangkan hasil evaluasi model menggunakan *confusion matrix* menghasilkan akurasi sebesar 71.14%, dimana menurut tafsiran *guilford empirical rules* akurasi tersebut termasuk tinggi/handal. Hasil interval kepercayaan didapatkan batas atas = 0.743437, dan batas bawah = 0.6771. Dengan demikian diperoleh model C4.5 untuk prediksi minat baca yang akurasinya tinggi/handal.

Keywords:

Decision tree C4.5
Interest reading
Machine learning
Supervised Learning

ABSTRACT

Reading interest is an important indicator of literacy level and is directly correlated with analytical thinking skills and overall quality of education. Reading interest prediction can be solved with a machine learning approach using a C4.5 algorithm that is reliable in processing data. Based on the results of the analysis that has been carried out, a C4.5 decision tree was obtained for the prediction of reading interest, where the reading environment variable does not affect the prediction of reading interest, while the age variable has the most influence on the prediction of reading interest. Meanwhile, the results of the model evaluation using the confusion matrix produced an accuracy of 71.14%, which according to the interpretation of the Guilford Empirical Rules included high/reliable. The result of the confidence interval is obtained that the upper limit = 0.743437, and the lower limit = 0.6771. Thus, Model C4.5 was obtained for the prediction of reading interest with high accuracy/reliability.

Penulis Korespondensi:

Muhajir Yunus,
Program Studi Bisnis Digital,
Universitas Muhammadiyah Gorontalo
Email: muhajiryunus@umgo.ac.id

1. PENDAHULUAN

Hingga saat ini minat baca masyarakat Indonesia masih tergolong rendah orang-orang lebih cenderung menonton ketimbang membaca [1]. UNESCO menyebut Indeks minat baca masyarakat Indonesia hanya diangka 0,001% atau dari 1000 orang Indonesia, cuma 1 orang yang rajin membaca [2]. Berdasarkan studi "Most Littered Nation In the World" yang dilakukan oleh Central Connecticut State University, Indonesia dinyatakan menduduki peringkat ke-60 dari 61 negara soal minat membaca. Minat baca merupakan salah satu indikator penting dalam mencerminkan tingkat literasi dan kualitas sumber daya manusia suatu bangsa [3]. Tingkat minat baca yang tinggi umumnya berkorelasi erat dengan kemampuan analisis yang baik, wawasan yang luas, serta peningkatan kualitas pendidikan dan kehidupan intelektual masyarakat. Indonesia masih

JTIS : Journal of Technopreneurship and Information Systems
Vol. 08, No. 1, Februari 2025, hal. 1~7
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 5
DOI: 10.36085

menghadapi tantangan serius terkait rendahnya minat baca masyarakatnya, terutama jika dibandingkan dengan negara-negara lain di kawasan maupun global [4]. Oleh karena itu, prediksi minat baca perlu dilakukan dengan akurat agar dapat mendukung pengambilan keputusan terkait kepustakaan.

Kemajauan teknologi informasi menyebabkan kemudahan bagi setiap orang memperoleh data dengan mudah bahkan cenderung berlebihan. Data yang sedemikian besar tentunya memiliki informasi yang tersembunyi di dalamnya, namun kemampuan manusia terbatas dalam menganalisis data atau menggali pengetahuan dari data tesebut [5]. Pengetahuan tersebut tentunya sangat berguna untuk mendukung pengambilan kebijakan atau keputusan. Selain itu, kemampuan komputasi yang semakin canggih dan terjangkau, serta persaingan bisnis yang semakin kompetitif merupakan faktor-faktor lainnya mengapa *machine learning* semakin berperan dalam mendukung pengambilan keputusan [6].

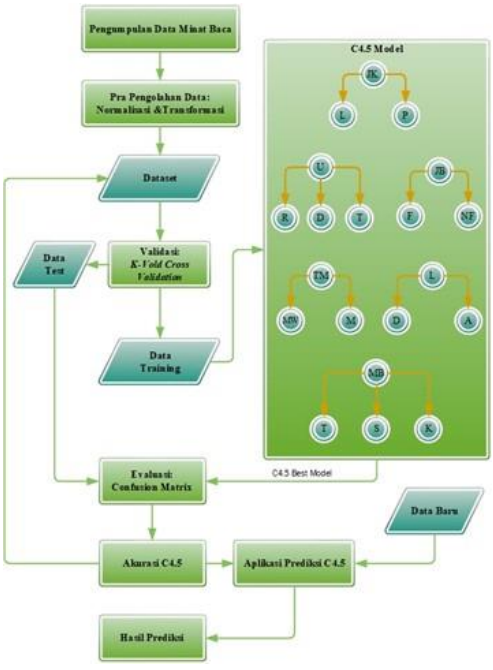
Beberapa metode machine learning yang dapat digunakan untuk prediksi minat baca, antara lain: (1) Artificial Neural Network (ANN) [7]; (2) Support Vector Machine (SVM) [8][9]; (3) K-Nearest Neighbor (K-NN) [10][11]; (4) C4.5 [12]. Di antara metode-metode tersebut, SVM berada sekelas dengan ANN yang merupakan metode yang sering digunakan untuk prediksi, keduanya merupakan metode supervised learning [13]. ANN memiliki kelebihan dalam memprediksi data yang berdimensi besar dan melakukan pelatihan terhadap seluruh data [14]. Sementara SVM tidak melakukan pelatihan terhadap seluruh data sehingga dapat mempercepat waktu komputasi dan memiliki kelebihan pada masalah yang hanya 2 class, namun lemah pada dimensi data yang besar, dan tidak cocok untuk multi classification [15].

Decision tree C4.5 adalah salah satu metode belajar yang sangat popular dan banyak digunakan secara praktis [16][17]. Metode ini merupakan metode yang berusaha menemukan fungsi-fungsi pendekatan yang bernilai diskrit dan tahan terhadap data-data yang memiliki kesalahan (*noise*) [18]. Kelebihan algoritma C4.5 dapat menghasilkan pohon keputusan yang mudah diinterprestasikan, memiliki tingkat akurasi yang dapat diterima, efisisen dalam menangani atribut bertipe diskret dan dapat menangani atribut bertipe diskret dan numerik [19][20]. Penelitian ini menggunakan metode C4.5 untuk prediksi minat baca karena sesuai karakter data minat baca yang seluruh variabelnya bertipe diskret, yaitu: Jenis Kelamin; Jenis Bacaan; Tujuan Membaca; Lingkungan Baca, kecuali variabel Umur saja yang beripe numerik. Sedangkan variabel outputnya, yaitu: Minat Baca yang terdiri dari label: Tinggi; Sedang; dan Rendah.

Pemodelan C4.5 untuk prediksi minat baca dilakukan menggunakan alat bantu Rapid Miner. Dalam mengukur kinerja dari model tersebut, metode Confussion Matrix dapat digunakan untuk mengetahui akurasinya. Selanjutnya, dilakukan pengembangan sistem, software dari model tersebut yang direkayasa menggunakan tools Ms. Visual Studio dengan Bahasa Pemrograman Visual Basic .NET dan database SQL Server.

2. METODE PENELITIAN

Tahapan penelitian yang digunakan dalam penelitian ini dapat dilihat pada gambar 1.



Gambar 1. Alur tahapan penelitian

JTIS : Journal of Technopreneurship and Information Systems
Vol. 08, No. 1, Februari 2025, hal. 1~7
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 5
DOI: 10.36085

Gambar 1 menunjukkan alur proses penelitian ini dibagi menjadi beberapa tahap. Pertama, peneliti mengumpulkan data yang dibutuhkan untuk penelitian berupa data minat baca. Kedua, pra-pemrosesan data termasuk mentransformasi atribut dan menormalisasi data. Ketiga, data dibagi menjadi dua subset, yaitu data pelatihan dan data pengujian. Keempat, pembentukan model algoritma Decision Tree C4.5 menggunakan tools rapid miner untuk melatih model prediksi minat baca. Kelima, validasi model menggunakan metode k-fold cross validation untuk memeriksa apakah model yang dilatih dengan data pelatihan juga berkinerja baik pada data yang belum pernah terlihat sebelumnya. Setelah model dilatih dilakukan evaluasi untuk mengukur sejauh mana kinerja algoritma. Keenam mengimplementasikan dalam bentuk aplikasi sistem prediksi dan menginterpretasikan hasil prediksi untuk membantu pengambilan keputusan.

3. HASIL DAN ANALISIS

Berdasarkan evaluasi model yang telah dilakukan mengguakan Confusion Matrix diperoleh akurasi sebesar 71.14% dan Kappa sebesar 0.108 seperti pada gambar berikut.

accuracy: 71.14% +/- 2.98% (mikro: 71.14%)

	true Sedang	true Tinggi	true Rendah	class precision
pred. Sedang	138	43	11	71.88%
pred. Tinggi	2	1	1	25.00%
pred. Rendah	1	0	4	80.00%
class recall	97.87%	2.27%	25.00%	

Gambar 2. Hasil Confusion Matrix

Hasil percobaan validasi yang telah dilakukan mengguakan *k-fold cross validation* ditunjukkan pada tabel berikut.

Tabel 1. Hasil *k-fold cross validation*

Kriteria	Depth	Pruning	Prepruning	Confidence	Min Gain	Leaf Size	Split Size	Prepruning Number	Akurasi
Gain Ratio	20	True	True	0.25	0.1	2	4	3	71.14%
Gain Ratio	20	True	False	0.25	0.1	2	4	3	63.18%
Gain Ratio	20	True	True	0.50	0.1	2	4	3	70.65%
Gain Ratio	20	True	True	0.25	0.99	2	4	3	70.15%
Gain Ratio	20	True	True	0.25	0.1	1	4	3	70.65%
Gain Ratio	20	True	True	0.25	0.1	2	4	0	70.65%

Berdasarkan uji coba yang telah dilakukan pada Tabel 1. Percobaan pertama mendapatkan hasil yang terbaik dengan akurasi sebesar 71,14%. Untuk variabel yang bertipe numerik (umur), maka terlebih dahulu harus dipecahkan dengan cara melakukan diskretisasi. Teknik pemecahan biner digunakan dengan teknik diskretisasi berbasis entropy. Pada node akar, diperoleh pemecahan umur > 27.500 dan <= 27.500.

Tabel 2. Node Akar

		Jumlah	Tinggi	Sedang	Rendah	Entropy	Gain
		201	**44**	**141**	**16**	**1.1292**	
Umur	> 27.500	18	4	7	7	1.5420	**0.0602**
	<= 27.500	183	40	134	9	1.0225	
Jenis Kelamin	Pria	48	7	35	6	1.1123	**0.0116**
	Wanita	153	37	106	10	1.1193	
Jenis Bacaan	Non Fiksi	87	19	62	6	1.0937	**0.0009**
	Fiksi	114	25	79	10	1.1547	

JTIS : Journal of Technopreneurship and Information Systems
Vol. 08, No. 1, Februari 2025, hal. 1~7
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 5
DOI: 10.36085

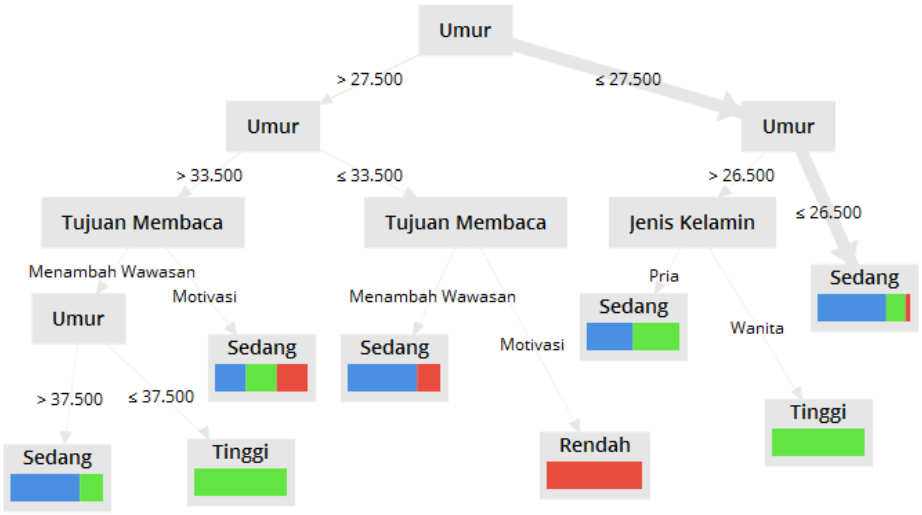
Tujuan Membaca	Wawasan	146	33	107	6	1.0028	**0.0340**
	Motivasi	55	11	34	10	1.3405	
Lingkungan Baca	Ruangan	86	22	58	6	1.1544	**0.0046**
	Di luar	115	22	83	10	1.1024	

Untuk hasil prediksi menggunakan algoirtma C4.5 ditunjukkan pada tabel berikut.

Tabel 3. Hasil Prediksi Algoritma C4.5

No	Class	Prediksi	Confidence Tinggi	Confidence Sedang	Confidence Rendah	Keterangan
1	Sedang	Sedang	0.211180124	0.739130435	0.049689441	Benar
2	Sedang	Sedang	0.211180124	0.739130435	0.049689441	Benar
3	Tinggi	Sedang	0.211180124	0.739130435	0.049689441	Salah
4	Tinggi	Sedang	0	0.5	0.5	Salah
5	Sedang	Sedang	0.211180124	0.739130435	0.049689441	Benar
6	Tinggi	Sedang	0.211180124	0.739130435	0.049689441	Salah
7	Sedang	Sedang	0.211180124	0.739130435	0.049689441	Benar
…	…	…	…	…	…	…
201	Sedang	Sedang	0.2	0.75	0.05	Benar

Berdasarkan hasil penelitan yang dilakukan diperoleh graph pohon keputusan Algoritma C4.5 untuk prediksi minat baca seperti yang ditunjukkan pada gambar berikut.



Gambar 2. Graph Pohon Keputusan C4.5

Berdasarkan Gambar 2. graph pohon keputusan yang dihasilkan bahwa variabel lingkungan membaca tidak mempengaruhi prediksi minat baca, sedangkan variabel umur yang paling berpengaruh terhadap prediksi minat baca. Sedangkan hasil evaluasi model menggunakan Confusion Matrix menghasilkan akurasi sebesar 71.14% dan Kappa sebesar 0.108. Untuk memberikan tafsiran pada nilai akurasi tersebut, dapat digunakan referensi *Guilford Empirical Rules* berikut ini.

Tabel 4. *Guilford Empirical Rules*

Besar Akurasi (%)	**Penafsiran**
0 - < 20	Akurasi sangat buruk.
>= 20 - < 40	Akurasi sangat rendah.
>= 40 - < 70	Akurasi sangat sedang atau cukup tinggi.
>= 70 - < 90	Akurasi sangat tinggi.
>= 90 - < 100	Akurasi sangat sangat tinggi.

JTIS : Journal of Technopreneurship and Information Systems
Vol. 08, No. 1, Februari 2025, hal. 1~7
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 5
DOI: 10.36085

Dengan berdasarkan referensi *Guilford Empirical Rules* kinerja Model C4.5 untuk prediksi minat baca termasuk tinggi/handal, sehingga boleh dilanjutkan ke tahap pengembangan sistemnya. Namun untuk memperkirakan interval kepercayaan perlu digunakan distribusi probabilitas yang mengatur ukuran akurasi. Dalam penelitian ini digunakan distribusi normal.

$$P\left(-Z_{\alpha/2} \leq \frac{acc - p}{\sqrt{\frac{p(1-p)}{n}}} \leq Z_{1-\alpha/2}\right) = 1 - \alpha$$

Dimana:
n : jumlah data uji;
p : akurasi sebenarnya;
acc : akurasi dari model tersebut;
$Z_{\alpha/2}$: batas atas kepercayaan akurasi;
$Z_{1-\alpha/2}$: batas bawah kepercayaan akurasi.

Tabel 5. Interval Kepercayaan $Z_{\alpha/2}$ dalam Distribusi Normal

1-α	0.5	0.7	0.8	0.85	0.9	0.95	0.98	0.99	0.998	0.999
$Z_{\alpha/2}$	0.67	1.04	1.282	1.440	1.645	1.967	2.326	2.576	3.090	3.291

Dengan menyederhanakan pertidaksamaan di atas, dihasilkan interval kepercayaan sebagai berikut untuk *p*.

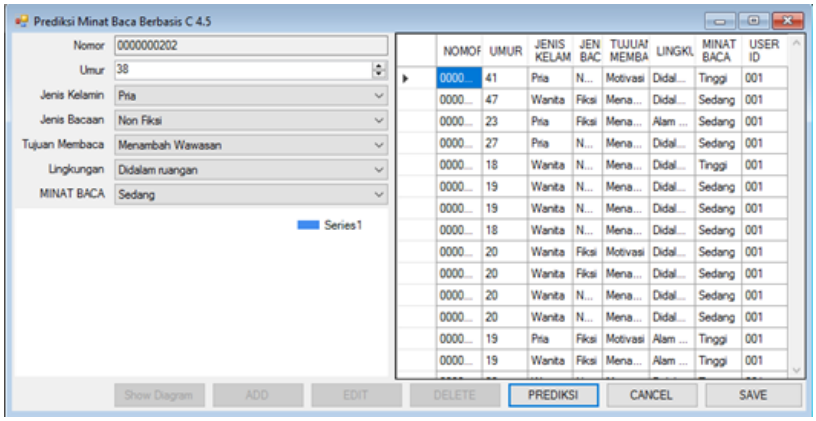
$$\frac{2 * n * acc + Z_{\alpha/2}^2 \pm Z_{\alpha/2}\sqrt{Z_{\alpha/2}^2 + 4 * n * acc - 4 * n * acc^2}}{2(n + Z_{\alpha/2}^2)}$$

Akurasi dari Model C4.5 = 71.14% dengan data uji n = 201, berapa interval kepercayaan akurasi tersebut pada tingkat kepercayaan 70%.

$$\frac{2 * 201 * 0.7114 + 1.04^2 \pm 1.04\sqrt{1.04^2 + 4 * 201 * 0.7114 - 4 * 201 * 0.7114^2}}{2(201 + 1.04^2)}$$

Didapatkan batas atas = 0.743437, dan batas bawah = 0.6771. Jadi untuk tingkat kepercayaan 70%, di dapatkan interval kepercayaan akurasi dari 67.71% s/d 74.34%. Interval kepercayaan yang masih masuk dalam akurasi 70% dan selisih intervalnya tidak terlalu signifikan.

Untuk Sistem Prediksi Minat Baca berbasis C4.5 diuji kinerjanya dengan meggunakan metode White Box Testing menghasilkan V(G) = CC = 8, di mana setiap jalurnya dapat dieksekusi, sehingga dinyatakan bahwa sistem ini telah memenuhi syarat logika pemrograman dan tidak kompleks. Sedangkan dengan pengujian Black Box menyatakan bahwa sistem telah bebas dari berbagai kesalahan komponen-komponennya. Dengan demikian, diperoleh Sistem Prediksi Minat Baca berbasis C4.5 yang handal sehingga dapat diimplementasikan seperti pada gambar berikut.



Gambar 3. Sistem Prediksi Minat Baca

4. KESIMPULAN

Berdasarkan hasil percobaan yang telah dilakukan, diperoleh pohon keputusan C4.5 untuk prediksi minat baca, di mana variabel lingkungan membaca tidak mempengaruhi prediksi minat baca, sedangkan variabel umur yang paling berpengaruh terhadap prediksi minat baca. Sedangkan hasil evaluasi model menggunakan Confusion Matrix menghasilkan akurasi sebesar 71.14% dan Kappa sebesar 0.108, di mana menurut tafsiran Guilford Empirical Rules akurasi tersebut termasuk tinggi/handal. Namun untuk memperkirakan interval kepercayaan perlu digunakan distribusi probabilitas yang mengatur ukuran akurasi. Dalam penelitian ini menggunakan distribusi normal, didapatkan batas atas = 0.743437, dan batas bawah = 0.6771. Jadi untuk tingkat kepercayaan 70%, di dapatkan interval kepercayaan akurasi dari 67.71% s/d 74.34%. Interval kepercayaan yang masih masuk dalam akurasi 70% dan selisih intervalnya tidak terlalu signifikan. Dengan demikian, diperoleh Model C 4.5 untuk prediksi minat baca yang akurasinya tinggi/handal.

REFERENSI

[1] P. N. RI, "Statistik Perpustakaan dan Minat Baca Masyarakat," 2023. https://perpusnas.go.id/

[2] U. I. for Statistics, "Literacy rates and UIS literacy survey," 2022. https://uis.unesco.org/en/literacy-rates

[3] A. Habók, A. Magyar, M. B. Németh, dan B. Csapó, "Motivation and self-related beliefs as predictors of academic achievement in reading and mathematics: Structural equation models of longitudinal data," *Int. J. Educ. Res.*, vol. 103, hal. 101634, 2020, doi: 10.1016/j.ijer.2020.101634.

[4] A. A. Nguyen, T. T. M., & Putra, "Comparative analysis of reading habits in ASEAN countries: Implications for educational policy," *Asian Educ. Dev. Stud.*, vol. 9, no. 3, hal. 321–335, 2020, doi: https://doi.org/10.1108/AEDS-03-2019-0056.

[5] R. S. Wahono, "Literature Review: Pengantar Dan Metode." https://romisatriawahono.net/2016/05/07/literature-review-pengantar-dan-metode/

[6] R. T. Huang, Z., & Rust, "Artificial intelligence in service," *J. Serv. Res.*, vol. 24, no. 2, hal. 155–176, 2021, doi: https://doi.org/10.1177/1094670520986712.

[7] A. Tuan Hoang *dkk.*, "A review on application of artificial neural network (ANN) for performance and emission characteristics of diesel engine fueled with biodiesel-based fuels," *Sustain. Energy Technol. Assessments*, vol. 47, hal. 101416, Okt 2021, doi: 10.1016/j.seta.2021.101416.

[8] D. Valero-Carreras, J. Alcaraz, dan M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, vol. 152, hal. 106131, Apr 2023, doi: 10.1016/j.cor.2022.106131.

[9] W. C. Leong, A. Bahadori, J. Zhang, dan Z. Ahmad, "Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM)," *Int. J. River Basin Manag.*, vol. 19, no. 2, hal. 149–156, Apr 2021, doi: 10.1080/15715124.2019.1628030.

[10] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, dan A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *J. Big Data*, vol. 11, no. 1, hal. 113, Agu 2024, doi: 10.1186/s40537-024-00973-y.

[11] A. Laios, A. Gryparis, D. DeJong, R. Hutson, G. Theophilou, dan C. Leach, "Predicting complete cytoreduction for advanced ovarian cancer patients using nearest-neighbor models," *J. Ovarian Res.*, vol. 13, no. 1, hal. 117, Des 2020, doi: 10.1186/s13048-020-00700-0.

[12] M. Yunus, M. K. Biddinika, dan A. Fadlil, "Classification of Stunting in Children Using the C4.5 Algorithm," *J. Online Inform.*, vol. 8, no. 1, hal. 99–106, Jun 2023, doi: 10.15575/join.v8i1.1062.

[13] A. Kurani, P. Doshi, A. Vakharia, dan M. Shah, "A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting," *Ann. Data Sci.*, vol. 10, no. 1, hal. 183–208, Feb 2023, doi: 10.1007/s40745-021-00344-x.

[14] S. Thudumu, P. Branch, J. Jin, dan J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *J. Big Data*, vol. 7, no. 1, hal. 42, Des 2020, doi: 10.1186/s40537-020-00320-x.

[15] J. He, C. Song, Q. Luo, L. Lan, C. Yang, dan W. Gui, "Noise-robust self-adaptive support vector machine for residual oxygen concentration measurement," *IEEE Trans. Instrum. Meas.*, hal. 1–1, 2020, doi: 10.1109/TIM.2020.2987049.

[16] R. T. Wiyono dan N. D. W. Cahyani, "Performance Analysis of Decision Tree C4.5 as a Classification Technique to Conduct Network Forensics for Botnet Activities in Internet of Things," in *2020 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, Agu 2020, hal. 1–5. doi: 10.1109/ICoDSA50139.2020.9212932.

[17] A. A. Dehghani, N. Movahedi, K. Ghorbani, dan S. Eslamian, "Decision tree algorithms," in *Handbook of Hydroinformatics*, Elsevier, 2023, hal. 171–187. doi: 10.1016/B978-0-12-821285-1.00004-X.

[18] V.-H. Nhu, T. T. Bui, L. N. My, H. Vuong, dan H. N. Duc, "A new approach based on integration of random subspace and C4.5 decision tree learning method for spatial prediction of shallow landslides," *Vietnam J. Earth Sci.*, Feb 2022, doi: 10.15625/2615-9783/16929.

[19] P. Chen, "The Application of an Improved C4.5 Decision Tree," in *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, IEEE, Jul 2021, hal. 392–396. doi: 10.1109/ICNISC54316.2021.00078.

[20] M. Ahmad, N. A. Al-Shayea, X.-W. Tang, A. Jamal, H. M. Al-Ahmadi, dan F. Ahmad, "Predicting the Pillar Stability of Underground Mines with Random Trees and C4.5 Decision Trees," *Appl. Sci.*, vol. 10, no. 18, hal. 6486, Sep 2020, doi: 10.3390/app10186486.