Vol 2 No. 1 Februari 2025 P-ISSN : 3047-2113 E-ISSN : 3047-2121, Hal 46 - 51

JURNAL ILMIAH
MULTIDISIPLIN ILMU
Halaman Jurnal: https://journal.smartpublisher.id/index.php/jimi
Halaman UTAMA Jurnal : https://journal.smartpublisher.id/

DOI: https://doi.org/10.69714/s91z1k09

PENGUNAAN DATA MINIG UNTUK MENGIDENTIFIKASI PELANGGAN BERESIKO TINGGI DALAM PENJUALAN MENGUNAKAN ALGORITMA DECITION TREE C4.5

Zuhrian Nur Saputra [a*], Zaehol Fatah [b]
[a] Fakultas Sains & Teknologi, / Sistem Informasi zuhriannursaputra22 @gmail.com, Universitas Ibrahimy, Situbondo Jawa Timur
[b] Fakultas Sains & Teknologi, / Sistem Informasi zaeholfatah @gmail.com, Universitas Ibrahimy, Situbondo Jawa Timur
[*] Korespondensi

ABSTRACT

In the competitive world of business, identifying high-risk customers is critical to minimizing churn rates and increasing profitability. This research uses data mining techniques using the C4.5 decision tree algorithm to classify customers based on their churn risk. The research stages include data collection, cleaning, data processing, as well as dividing the data into training and testing sets. The implementation of this algorithm was carried out using RapidMiner software, which facilitates customer clustering and predicting behavior based on historical attributes. The evaluation results show the model has an accuracy of 74.59%, with precision and recall indicating the model's ability to identify high-risk customers. Thus, the Decision Tree C4.5 algorithm is proven to be effective in supporting decision making for customer churn risk mitigation strategies.

Keywords: *Data mining, Decision Tree C4.5 algorithm, customer churn, risk prediction, RapidMiner*.

Abstrak

Dalam dunia bisnis yang kompetitif, identifikasi pelanggan berisiko tinggi sangat penting untuk meminimalkan tingkat churn dan meningkatkan profitabilitas. Penelitian ini menggunakan teknik data mining dengan menggunakan algoritma pohon keputusan C4.5 untuk mengklasifikasikan pelanggan berdasarkan risiko churnnya. Tahapan penelitian mencakup pengumpulan data, pembersihan, pemrosesan data, Selain membagi data menjadi set pelatihan dan pengujian. Implementasi algoritma ini dilakukan dengan menggunakan perangkat lunak RapidMiner, yang memfasilitasi pengelompokan pelanggan dan memprediksi perilaku berdasarkan atribut historis. Hasil evaluasi menunjukkan model memiliki akurasi sebesar 74,59%, dengan precision dan recall yang menunjukkan kemampuan model dalam mengidentifikasi pelanggan berisiko tinggi. Dengan demikian, algoritma Decision Tree C4.5 terbukti efektif dalam mendukung pengambilan keputusan untuk strategi mitigasi risiko churn pada pelanggan.

Kata Kunci: *Data mining, algoritma Decision Tree C4.5, churn pelanggan, prediksi risiko, RapidMiner.*

1. PENDAHULUAN

Dalam dunia bisnis yang semakin kompetitif, memahami perilaku dan risiko nasabah merupakan salah satu faktor kunci dalam menjaga pertumbuhan dan kelangsungan bisnis. Pelanggan yang berisiko tinggi, terutama yang cenderung churn atau gagal memenuhi komitmen pembelian, dapat memberikan dampak yang signifikan terhadap keuntungan perusahaan. Oleh karena itu, sangat penting untuk dapat mengenali pelanggan sejak awal agar bisnis dapat melakukan langkah pencegahan yang sesuai.[1]

Penggunaan data mining menjadi pendekatan yang semakin populer untuk memecahkan masalah ini. Kegiatan penambangan data melibatkan analisis data dalam jumlah yang besar guna menemukan pola

Received Desember 19, 2024; Revised Desember 23, 2024; Accepted Januari 27, 2025; Published Februari 1, 2025

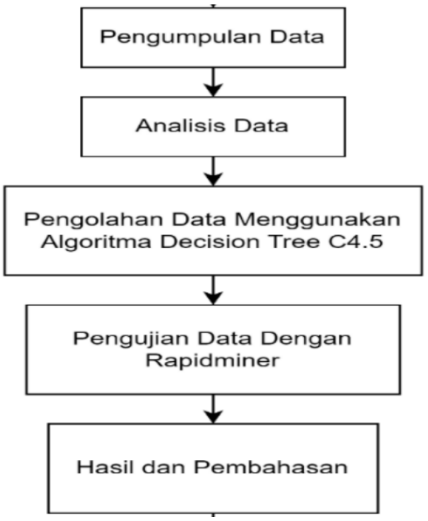
tersembunyi serta informasi yang bermanfaat. Sebagai bagian dari mengidentifikasi pelanggan berisiko tinggi, data mining dapat digunakan untuk menganalisis berbagai aspek interaksi pelanggan dengan perusahaan, seperti riwayat pembelian, preferensi produk, dan perilaku transaksi. Dengan menggunakan algoritma prediktif dan analisis pola, perusahaan dapat mengidentifikasi pelanggan yang berpotensi berisiko tinggi dan merancang strategi intervensi yang tepat, seperti penawaran khusus atau kampanye loyalitas pelanggan. [2] Data mining adalah metode yang diterapkan untuk mengolah dan menganalisis sejumlah besar data dengan tujuan menemukan pola dan informasi yang berguna. Teknik ini banyak digunakan dalam bidang medis untuk mendukung pengambilan keputusan yang lebih efektif. Salah satu algoritma yang cukup populer dalam data mining adalah pohon keputusan, terutama algoritma C4. 5 yang dikenal dengan tingkat akurasi yang tinggi dan model yang mudah dipahami.

Penambangan data juga memungkinkan bisnis membuat keputusan yang lebih efektif dan efisien berdasarkan data. Melalui analisis terperinci, perusahaan dapat mengkategorikan pelanggan berdasarkan tingkat risiko, memprediksi perilaku masa depan, dan memaksimalkan peluang untuk mempertahankan loyalitas pelanggan. Pendekatan ini tidak hanya mengurangi churn, namun juga meningkatkan profitabilitas jangka panjang.[3]

Artikel ini menjelaskan bagaimana penambangan data dapat digunakan untuk mengidentifikasi pelanggan berisiko tinggi dan teknik terkait seperti klasifikasi, pengelompokan, dan algoritma prediktif yang dapat digunakan untuk mendukung proses ini. Selain itu, studi kasus yang relevan juga dibahas untuk memberikan gambaran nyata tentang bagaimana teknologi ini diterapkan di industri.

2. METODE PENELITIAN

Dalam penelitian ini, data mining dilakukan oleh penulis menggunakan metode Decision Tree C4. 5 dalam proses Knowledge Discovery in Databases (KDD). Langkah-langkahnya meliputi pemahaman masalah dan tujuan penelitian, pengumpulan data dari sumber yang sesuai, pembersihan data, dan pemrosesan data termasuk pemilihan atribut penting. Selanjutnya, pohon keputusan dibangun menggunakan metode Decision Tree C4.5 dengan memperhitungkan nilai gain ratio yang optimal.[4] Data kemudian dikelompokkan menjadi set latihan dan uji untuk model pengujian. Evaluasi model dilakukan dengan menggunakan metrik seperti akurasi, precision, dan recall untuk mengukur kinerja klasifikasi. Hasil evaluasi ini digunakan untuk menemukan pola dan informasi baru dari data.



Gambar 1. Metode Penelitian

Diagram alir di atas menjelaskan proses sistematis dalam pengolahan data untuk menghasilkan hasil yang dapat dianalisis. Proses dimulai dengan pengumpulan data, di mana data dikumpulkan dari berbagai sumber untuk digunakan sebagai bahan analisis. Setelah data terkumpul, dilakukan analisis data awal guna memahami struktur data, membersihkan data dari kesalahan, dan menemukan pola-pola penting yang relevan untuk langkah berikutnya. Data yang telah dianalisis kemudian diolah menggunakan algoritma Decision Tree C4.5, sebuah metode machine learning yang berfungsi untuk membangun model berbasis pohon keputusan guna melakukan klasifikasi atau prediksi. Setelah model dibuat, dilakukan pengujian data menggunakan RapidMiner, sebuah perangkat lunak analitik yang memungkinkan evaluasi performa model,

seperti mengukur akurasi dan validitasnya.[5] Langkah terakhir adalah hasil dan pembahasan, di mana hasil pengolahan dan pengujian data dipaparkan serta dianalisis lebih lanjut. Pada tahap ini, dibuat interpretasi hasil, kesimpulan, dan rekomendasi berdasarkan temuan dari seluruh proses yang digunakan.

2.1. Metode Pengumpulan Data

Bagian ini menjelaskan metode pengumpulan data yang digunakan untuk mengenali pola dan kecenderungan churn pelanggan. Penelitian dilakukan dengan menggunakan pendekatan kuantitatif, di mana data sekunder dikumpulkan dari berbagai sumber termasuk database internal. Data ini dianalisis secara statistik untuk mengidentifikasi faktor-faktor yang berkontribusi terhadap tingkat churn (berhenti berlangganan).Misalnya, dilakukan menganalisis data demografi pelanggan, riwayat penggunaan layanan, dan tanggapan survei kepuasan pelanggan.[6]

2.1.1. Sumber Data

Salah satu referensi data yang digunakan adalah untuk analisis churn, termasuk informasi demografis, riwayat pembelian, interaksi dengan layanan, kepuasan pelanggan, dan faktor-faktor lain yang dapat mempengaruhi keputusan pelanggan untuk berhenti berlangganan.Tautan yang dapat di akses adalah sebagai berikut https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

2.1.2. Data Mining

Bagian ini menjelaskan metode pengumpulan data yang diterapkan untuk mengidentifikasi pola dan kecenderungan dalam perilaku churn pelanggan.[7] Dengan kata lain, data mining menjadi alat dan aplikasi yang mampu menganalisis data secara statistik untuk menemukan informasi yang belum diketahui sebelumnya. Secara sederhana, data mining merupakan proses ekstraksi informasi terbaru dengan mencari pola dan aturan spesifik di dalam sejumlah besar data. Prinsip kerja data mining adalah melalui analisis database besar untuk menemukan pola atau struktur baru yang mendukung proses pengambilan keputusan.[8]

2.1.3. Repid Miner

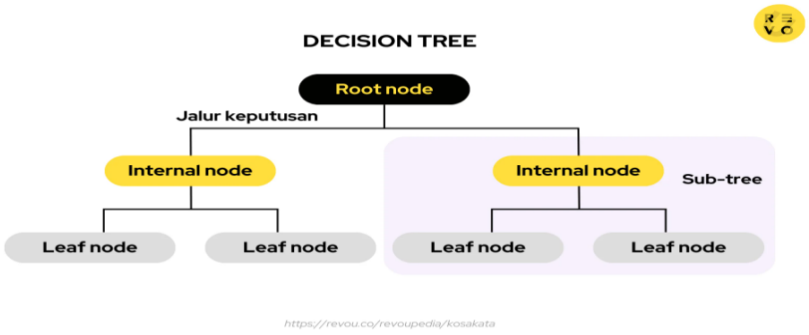
Rapid Miner merupakan alat analisis teks dan pengolahan data yang canggih, dirancang untuk mengekstrak pola dari data besar. Alat ini mengintegrasikan pendekatan statistik, kecerdasan buatan, dan basis data. Tujuan utama analisis teks a dalah untuk mendapatkan informasi yang relevan dan signifikan dari teks yang berukuran besar atau kompleks.[9]

2.1.4. Decision Tree

Pohon keputusan adalah algoritma untuk melakukan klasifikasi atau regresi.Pohon keputusan bekerja dengan cara membagi data menjadi beberapa cabang berdasarkan karakteristik tertentu di dalam data tersebut. Pembagian ini dilakukan secara iteratif (rekursif) hingga mencapai hasil akhir, suatu tangan yang memberikan keputusan atau prediksi berdasarkan nilai-nilai fitur tersebut.[10]

2.1.5. Algoritma C4.5

Algoritma C4.5 dikenal luas sebagai teknik yang umum digunakan dalam teknik data mining, terutama yang berkaitan dengan klasifikasi. Algoritma ini adalah hasil pengembangan dari algoritma ID3, yang dirancang untuk membangun pohon keputusan. Pohon keputusan ini dapat digunakan untuk memprediksi data terkini.[11]

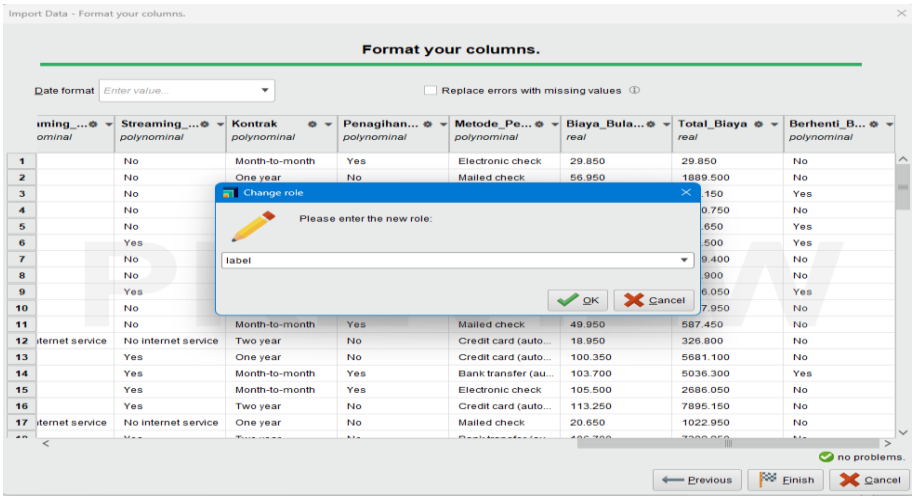


Gambar 2. Pohon Keputusan

3. HASIL DAN PEMBAHASAN

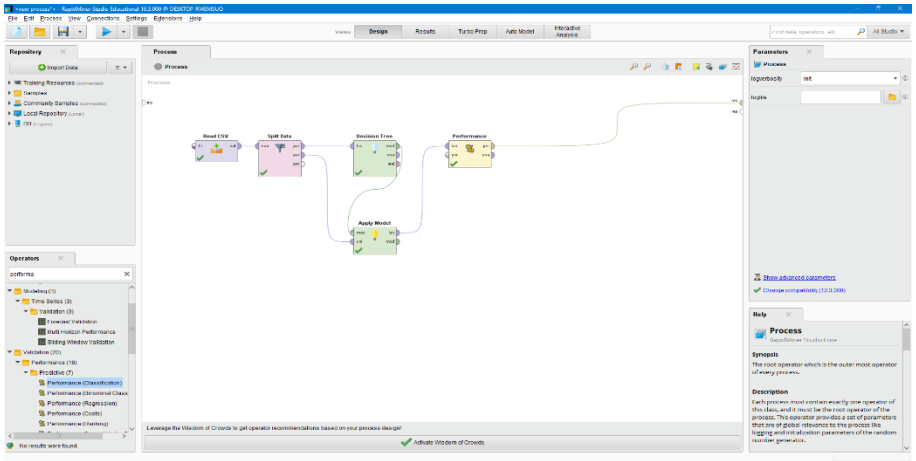
Data Mining adalah teknik yang sangat bermanfaat dalam mengungkap informasi yang tersembunyi di dalam dataset berukuran besar.[12] Dalam penjualan, penambangan data dapat digunakan untuk melacak pola perilaku pelanggan yang mungkin menunjukkan risiko tinggi, seperti potensi untuk beralih (berhenti berlangganan), kegagalan pembayaran, atau tindakan penipuan. Algoritma Decision Tree C4. 5 begitu terkenal di dunia Data Mining, terutama saat menangani permasalahan klasifikasi. Algoritma ini beroperasi dengan membuat pohon keputusan dari informasi yang terdapat dalam catatan masa lalu. Tiap cabang di pohon melambangkan sebuah atribut (contohnya, frekuensi pembelian, nilai transaksi, masa menjadi pelanggan), dan daun mana pun melambangkan kelasnya (seperti pelanggan berisiko tinggi atau rendah.

3.1 Model Data Mining



Gambar 3. Oprator Pemanggilan Data

Gambar di atas adalah antarmuka dari perangkat lunak RapidMiner dalam proses pengolahan data, khususnya pada tahap formatting atau penyesuaian kolom data. Pada bagian ini, [13]Jendela pop-up yang terlihat menunjukkan pengguna sedang mengubah peran (role) sebuah kolom menjadi label, yang biasanya digunakan sebagai variabel target atau output dalam analisis data atau model machine learning. Peran ini penting untuk memberi tahu sistem atribut mana yang akan diprediksi. Dataset yang ditampilkan memiliki berbagai atribut, termasuk jenis kontrak, metode pembayaran, biaya bulanan, total biaya, dan status berhenti berlangganan, yang diklasifikasikan sebagai nominal, polinomial, atau real berdasarkan jenis datanya. Langkah ini bertujuan untuk memastikan data sudah sesuai formatnya dan siap untuk diproses lebih lanjut, seperti pelatihan model atau analisis data lainnya.

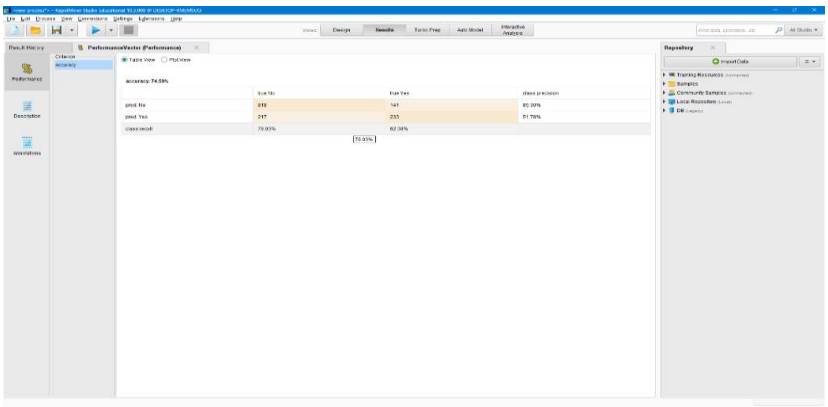


Gambar 4. Implementasi Algoritma C4.5

Pada gambar di atas, dapat melihat pengguna menggunakan RapidMiner Studio untuk membangun proses prediktif menggunakan algoritma pohon keputusan.Alur proses dimulai dengan Read CSV, yang mengimpor data dari file CSV.Kemudian data tersebut dipecah menjadi dua subset melalui proses Split

Data untuk memisahkan data latih dan data uji. Setelah itu, algoritma Decision Tree diaplikasikan pada data latih untuk membentuk model.[8] Model yang terbentuk kemudian diterapkan pada data uji menggunakan operator Apply Model. Hasil prediksi dievaluasi menggunakan operator performa yang menampilkan metrik performa model seperti akurasi, prcision, dan recall. Proses ini sering digunakan ketika menerapkan algoritma pohon keputusan untuk klasifikasi.Dalam hal ini, data dipecah menjadi data pelatihan dan pengujian untuk menghindari overfitting dan memungkinkan model melakukan generalisasi ke data baru.[9]

3.2 Hasil Akurasi



Gambar 5. Hasil Akurasi

Hasil akurasi model yang digunakan dalam penelitian ini adalah 74,59%. Ini berarti bahwa model pohon keputusan C4.5 yang diterapkan berhasil mengklasifikasikan data dengan benar sebesar 74,59% dari keseluruhan data yang diuji. Selain akurasi, model ini juga mengevaluasi metrik lain, yaitu: Precision - Untuk kelas “No” (tidak berisiko) sebesar 78,02% dan untuk kelas “Yes” (berisiko) sebesar 51,78%. Ini mengindikasikan proporsi prediksi “Yes” yang benar dibandingkan dengan keseluruhan prediksi “Yes”. Recall - Untuk kelas “No” adalah 79,03% dan untuk kelas “Yes” adalah 62,30%. Ini menunjukkan sejauh mana model berhasil mengidentifikasi pelanggan berisiko tinggi. Nilai akurasi ini menunjukkan bahwa model cukup baik dalam membedakan pelanggan berisiko dan tidak berisiko, tetapi precision dan recall untuk kategori “Yes” lebih rendah, yang berarti model masih memiliki keterbatasan dalam mendeteksi secara akurat semua pelanggan berisiko tinggi.

4. KESIMPULAN DAN SARAN

Berdasarkan penelitian ini dapat disimpulkan bahwa yang digunakan adalah algoritma Decision Tree C4.5 menunjukkan kinerja yang sangat memuaskan dalam mengidentifikasi pelanggan berisiko tinggi dalam penjualan, dengan tingkat akurasi sebesar 74,59%. Model ini dapat digunakan untuk mengklasifikasikan pelanggan menjadi berisiko tinggi atau berisiko rendah, membantu perusahaan menentukan strategi pencegahan yang tepat seperti penawaran khusus dan program loyalitas untuk pelanggan berisiko tinggi. Meskipun akurasi model ini sangat baik, namun memiliki kekurangan pada prescison dan recall untuk kategori risiko tinggi (kelas Ya). Hal ini menunjukkan bahwa model tersebut belum sepenuhnya dioptimalkan untuk mengidentifikasi semua pelanggan yang kemungkinan akan berhenti berlangganan atau berhenti berlangganan. Perbaikan model di masa depan diperlukan untuk membantu perusahaan mengidentifikasi pelanggan ini dengan lebih akurat dan mengambil tindakan yang lebih efektif.

DAFTAR PUSTAKA

[1] Y. Yudiana, A. Yulia Agustina, and dan Nur Khofifah, “Prediksi Customer Churn Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan,” *Indones. J. Islam. Econ. Bus.*, vol. 8, no. 1, pp. 01–20, 2023, [Online]. Available: http://e-journal.lp2m.uinjambi.ac.id/ojp/index.php/ijoieb

[2] I. Iddrus and D. W. Sari, “Penerapan Data Mining Menggunakan Algoritma Decision Tree C4.5 Untuk Memprediksi Mahasiswa Drop Out Di Universitas Wiraraja,” *J. Adv. Res. Inform.*, vol. 1, no. 02, pp. 1–7, 2023, doi: 10.24929/jars.v1i02.2684.

[3] Muhammad Rifqy Rifani and Andi Amri, “Pengaruh Kualitas Pelayanan terhadap Kepuasan

Pelanggan Miniso Big Mall Samarinda," *Lokawati J. Penelit. Manaj. dan Inov. Ris.*, vol. 2, no. 4, pp. 01–11, 2024, doi: 10.61132/lokawati.v2i4.934.

[4] A. Mahmood, H. Dhahri, M. Alhajla, and A. Almaslukh, "Enhanced Classification of Phonocardiograms using Modified Deep Learning," *IEEE Access*, vol. 12, no. November, pp. 178909–178916, 2024, doi: 10.1109/ACCESS.2024.3507920.

[5] W. H. Khoh, Y. H. Pang, S. Y. Ooi, L. Y. K. Wang, and Q. W. Poh, "Predictive Churn Modeling for Sustainable Business in the Telecommunication Industry: Optimized Weighted Ensemble Machine Learning," *Sustain.*, vol. 15, no. 11, 2023, doi: 10.3390/su15118631.

[6] T. Zhang, S. Moro, and R. F. Ramos, "A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation," *Futur. Internet*, vol. 14, no. 3, pp. 1–19, 2022, doi: 10.3390/fi14030094.

[7] M. J. Zaki, W. Meira, and W. Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press, 2020. [Online]. Available: https://books.google.co.id/books?id=oafDDwAAQBAJ

[8] P. Metode, C. Algoritma, and D. A. N. Naive, "Perbandingan metode algoritma c4.5 dan naive bayes untuk memprediksi penjualan kosmetik pada toko jelita 1,2," vol. 7, no. 2, pp. 220–225, 2024.

[9] A. Wasik *et al.*, "Implementasi data mining untuk memprediksi penjualan accessoris handphone dan handphone terlaris menggunakan metode k-nearest neighbor (k-nn) 1," vol. 1, no. 2, pp. 469–479, 2024.

[10] S. Syam *et al.*, *Data Mining : Teori dan Penerapannya dalam Berbagai Bidang*. PT. Sonpedia Publishing Indonesia, 2024. [Online]. Available: https://books.google.co.id/books?id=hTAxEQAAQBAJ

[11] Y. Ardilla *et al.*, *DATA MINING DAN APLIKASINYA*. Penerbit Widina, 2021. [Online]. Available: https://books.google.co.id/books?id=53FXEAAAQBAJ

[12] R. F. Putra *et al.*, *DATA MINING : Algoritma dan Penerapannya*. PT. Sonpedia Publishing Indonesia, 2023. [Online]. Available: https://books.google.co.id/books?id=zLHGEAAAQBAJ

[13] R. Bertolini, S. J. Finch, and R. H. Nehm, "Enhancing data pipelines for forecasting student performance: integrating feature selection with cross-validation," *Int. J. Educ. Technol. High. Educ.*, vol. 18, no. 1, 2021, doi: 10.1186/s41239-021-00279-6.

[14] E. Haerani *et al.*, "CLASSIFICATION ACADEMIC DATA USING MACHINE LEARNING FOR," vol. 4, no. 2, pp. 955–968, 2023.

[15] B. Sari, B. Sembiring, M. Pandia, H. Sembiring, and D. Margaretta, "Naïve Bayes Classifier and Decision Tree Algorithms for Classifying Payment Data," vol. 4, no. 1, pp. 592–600, 2023, doi: 10.30865/klik.v4i1.963.