

Analisis Data Kehadiran Siswa Menggunakan Teknik *Data Mining*

Ayu Firdhayanti^{*1}, Hendri Purnomo², Arman Suryadi Karim³

^{1,2,3} Sistem Informasi, Institut Informatika dan Bisnis Darmajaya, Indonesia

Email: ^{*1}ayufirdha@darmajaya.ac.id, ²hendrialie@darmajaya.ac.id, ³armansuryadi@darmajaya.ac.id

Abstrak—Kehadiran siswa merupakan salah satu indikator penting dalam menilai kualitas proses pembelajaran di sekolah. Pengelolaan data kehadiran secara manual masih menjadi tantangan bagi banyak institusi pendidikan. Meskipun sejumlah penelitian telah menerapkan teknik data mining untuk analisis performa siswa, pendekatan terintegrasi yang menggabungkan algoritma C4.5 dan K-Means dengan mempertimbangkan dimensi temporal secara eksplisit belum pernah diterapkan pada data kehadiran siswa SMP di Indonesia—ini adalah celah penelitian yang diisi oleh studi ini. Penelitian ini bertujuan menganalisis pola kehadiran siswa menggunakan teknik data mining, khususnya algoritma C4.5 untuk klasifikasi dan K-Means untuk pengelompokan (clustering). Dataset berasal dari rekap kehadiran siswa SMP selama satu tahun ajaran dengan total 1.200 record. Hasil penelitian menunjukkan bahwa C4.5 mampu mengklasifikasikan tingkat kehadiran dengan akurasi 91,7%, sedangkan K-Means mengelompokkan siswa ke dalam tiga kluster: Rajin, Cukup, dan Sering Absen, dengan Silhouette Score 0,68. Kontribusi utama penelitian ini adalah kerangka analisis terintegrasi berbasis dimensi temporal yang dapat direplikasi untuk konteks sekolah menengah lainnya di Indonesia. Temuan ini memberikan informasi bagi pihak sekolah untuk melakukan intervensi dini terhadap siswa berisiko tinggi absen.

Kata Kunci—Algoritma C4.5, Clustering, Data Mining, Kehadiran Siswa, K-Means, Klasifikasi

I. PENDAHULUAN

KEHADIRAN siswa di sekolah merupakan salah satu faktor penentu keberhasilan proses pembelajaran. Tingkat kehadiran yang rendah berkorelasi langsung dengan penurunan prestasi akademik, meningkatnya risiko putus sekolah, dan berkurangnya efektivitas pengajaran [1], [2]. Menurut data Kementerian Pendidikan, rata-rata ketidakhadiran siswa di tingkat SMP mencapai 7–12% per semester, angka yang cukup signifikan untuk menjadi perhatian serius [3].

Di era digitalisasi pendidikan, sebagian besar sekolah telah menerapkan sistem pencatatan kehadiran berbasis komputer. Namun, data yang terkumpul seringkali hanya digunakan untuk pelaporan administratif tanpa diolah lebih lanjut sebagai dasar pengambilan keputusan berbasis bukti (*evidence-based decision making*) [4]. Data mining merupakan proses ekstraksi pengetahuan dari kumpulan data besar menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin [5]. Penerapan

data mining dalam konteks pendidikan dikenal sebagai *Educational Data Mining* (EDM) [6], [7].

Beberapa penelitian terdahulu telah membahas penerapan data mining pada data kehadiran dan kinerja siswa. Romero dan Ventura [8] melakukan kajian komprehensif mengenai penerapan data mining dalam pendidikan dan menyimpulkan bahwa teknik klasifikasi dan clustering paling banyak digunakan. Pandey dan Pal [9] menerapkan Decision Tree untuk memprediksi kinerja akademik dengan akurasi 85%. Yadav et al. [10] menggunakan C4.5 untuk klasifikasi performa mahasiswa dengan akurasi 88,9%. Di Indonesia, Putri dan Santosa [11] menerapkan K-Means untuk menganalisis pola kehadiran siswa SMK dan berhasil mengidentifikasi kelompok siswa berisiko tinggi.

Penelitian ini mengusulkan pendekatan terintegrasi yang menggabungkan algoritma C4.5 dan K-Means dalam satu kerangka analisis, dengan mempertimbangkan dimensi temporal sebagai atribut analisis. Tujuan penelitian adalah: (1) menganalisis pola kehadiran siswa, (2) mengidentifikasi faktor paling berpengaruh terhadap ketidakhadiran, (3) mengelompokkan siswa berdasarkan pola kehadiran, dan (4) memberikan rekomendasi berbasis data bagi pihak sekolah [12], [13].

II. DESKRIPSI SISTEM

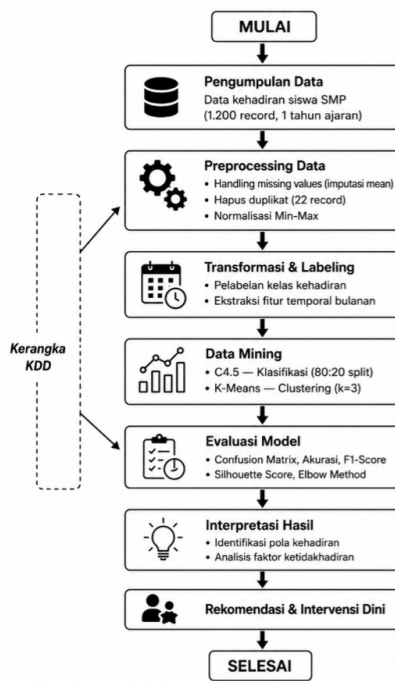
Penelitian ini mengikuti kerangka *Knowledge Discovery in Databases* (KDD) yang terdiri dari lima tahapan utama: seleksi data, preprocessing, transformasi, data mining, dan interpretasi atau evaluasi [5]. Gambar 1 menampilkan diagram alur penelitian secara lengkap.

Kerangka KDD dipilih karena telah terbukti efektif dalam berbagai penelitian data mining bidang pendidikan [8], [14]. Setiap tahapan bersifat iteratif, memungkinkan peneliti untuk kembali ke tahap sebelumnya apabila hasil evaluasi belum memuaskan.

III. METODE

A. Pengumpulan Data

Data yang digunakan adalah data kehadiran siswa SMP Tahun Ajaran 2023/2024 yang diperoleh melalui izin resmi dari pihak sekolah. Dataset terdiri dari 1.200 record dengan atribut: NIS, nama siswa, kelas, jumlah hari masuk, jumlah hari izin, jumlah hari sakit, jumlah hari tanpa keterangan (alpha), dan persentase kehadiran per bulan.



Gambar 1: Alur Penelitian Berbasis Kerangka KDD

selama 6 bulan (Semester Ganjil). Distribusi data per kelas adalah: kelas VII sebanyak 412 siswa (34,3%), kelas VIII sebanyak 398 siswa (33,2%), dan kelas IX sebanyak 390 siswa (32,5%), sehingga keterwakilan antar jenjang relatif seimbang. Proporsi gender terdiri dari 638 siswa laki-laki (53,2%) dan 562 siswa perempuan (46,8%). Distribusi label kehadiran sebelum preprocessing menunjukkan 58,1% siswa masuk kategori Hadir Baik ($\geq 90\%$), 30,5% Cukup ($<75\%$). Rentang nilai atribut numerik utama adalah: persentase kehadiran bulanan (min 42%, maks 100%), jumlah hari alpha (min 0, maks 28 hari), jumlah hari sakit (min 0, maks 18 hari), dan jumlah hari izin (min 0, maks 12 hari).

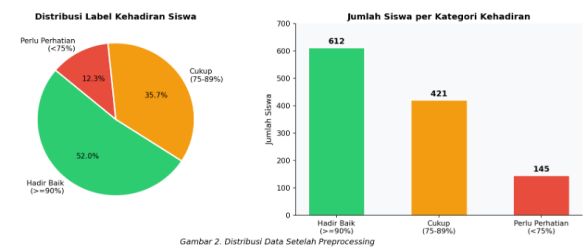
B. Preprocessing Data

Tahap preprocessing meliputi empat langkah utama yang mengikuti standar KDD [5]. Pertama, ditemukan sebanyak 34 record mengandung missing value pada atribut persentase kehadiran bulan tertentu; imputasi mean dipilih karena distribusi data mendekati normal (skewness <1) dan missing value bersifat acak (MCAR), sehingga mean menjadi estimator yang lebih stabil dibanding median untuk distribusi simetris. Kedua, deteksi dan penghapusan data duplikat menghasilkan penghapusan 22 record. Ketiga, normalisasi Min-Max diterapkan pada data numerik untuk keperluan clustering K-Means guna menyamakan skala antar atribut. Keempat, pelabelan dilakukan berdasarkan threshold yang mengacu pada regulasi kehadiran minimal 75% sesuai Permendikbud No. 23 Tahun 2016: $\geq 90\%$ = Hadir Baik, 75–89% = Cukup, dan $<75\%$ = Perlu Perhatian. Statistik deskriptif persentase kehadiran sebelum preprocessing: mean = 86,4%, std = 11,2%, min

= 42%, maks = 100%; setelah preprocessing: mean = 87,1%, std = 10,6%, min = 42%, maks = 100%, menunjukkan distribusi yang stabil pasca imputasi.

- Pengecekan dan penanganan nilai hilang (missing values) menggunakan metode imputasi mean.
- Deteksi dan penghapusan data duplikat (22 record dihapus).
- Normalisasi data numerik menggunakan metode Min-Max Normalization untuk keperluan clustering K-Means.
- Pelabelan data untuk keperluan klasifikasi C4.5 berdasarkan threshold persentase kehadiran.

Gambar 2 menampilkan distribusi data setelah proses preprocessing dan pelabelan dilakukan.



Gambar 2. Distribusi Data Setelah Preprocessing

Gambar 2: Distribusi Label Kehadiran Siswa Setelah *Pre-processing*

C. Algoritma C4.5 (Klasifikasi)

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang dikembangkan oleh Quinlan [15]. C4.5 membangun pohon keputusan dengan menghitung nilai Gain Ratio sebagai kriteria pemilihan atribut terbaik pada setiap simpul. Persamaan Gain Ratio adalah sebagai berikut:

$$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

Implementasi C4.5 menggunakan *library scikit-learn* (*Decision Tree Classifier* dengan *criterion='entropy'*) pada Python. Dataset dibagi menjadi 80% data latih dan 20% data uji. Evaluasi model dilakukan menggunakan *confusion matrix*, akurasi, *presisi*, *recall*, dan *F1-score* [9], [10].

D. Algoritma K-Means (Clustering)

K-Means adalah algoritma clustering partisional yang membagi data ke dalam k kelompok berdasarkan kemiripan nilai atribut [12]. Penentuan nilai k optimal menggunakan metode Elbow, sedangkan validasi cluster dilakukan menggunakan Silhouette Score [13]. Implementasi menggunakan KMeans dari scikit-learn dengan parameter: `n_clusters=3`, `init='k-means++'`, `max_iter=300`.

IV. HASIL DAN PEMBAHASAN

A. Hasil Preprocessing Data

Setelah proses *preprocessing*, dari 1.200 record awal diperoleh 1.178 record bersih (22 *record* dihapus karena duplikat atau data tidak lengkap). Distribusi label kehadiran ditampilkan pada Tabel I.

Tabel I: Distribusi Label Kehadiran Setelah *Preprocessing*

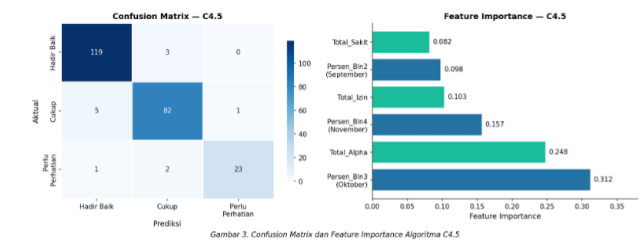
Kategori Kehadiran	Threshold (%)	Jumlah Siswa	Persentase (%)
Hadir Baik	≥ 90	612	51,9
Cukup	75–89	421	35,7
Perlu Perhatian	≤ 75	145	12,3
Total	—	1.178	100

B. Hasil Klasifikasi C4.5

Model pohon keputusan C4.5 menghasilkan akurasi sebesar 91,7% pada data uji. Atribut yang paling berpengaruh adalah persentase kehadiran bulan ke-3 (Oktober) dan jumlah hari alpha (tanpa keterangan). Hasil evaluasi lengkap ditampilkan pada Tabel 2, sedangkan Gambar 3 menampilkan *confusion matrix* dan *feature importance*.

Tabel II: Hasil Evaluasi Model C4.5

Kelas	Precision(%)	Recall(%)	F1-Score(%)	Support
Hadir Baik	95,2	97,5	96,3	122
Cukup	89,1	93,2	91,1	88
Perlu Perhatian	92,0	88,5	90,2	26
Rata-rata Tertimbang	92,4	93,7	92,7	236
Akurasi Keseluruhan	91,7			



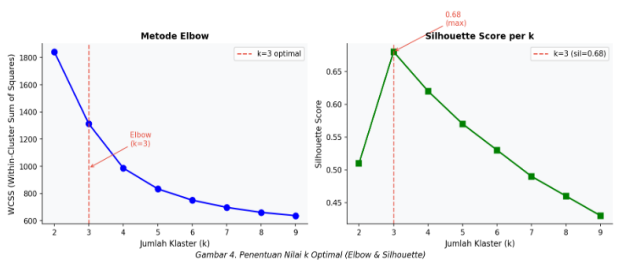
Gambar 3: Confusion Matrix dan Feature Importance Algoritma C4.5

Hasil ini sejalan dengan penelitian Yadav et al. [10] yang memperoleh akurasi 88,9% menggunakan C4.5, dan melampaui hasil Pandey dan Pal [9] yang mendapatkan 85% menggunakan Decision Tree. Peningkatan akurasi ini dikaitkan dengan kualitas dataset yang lebih bersih dan penambahan fitur temporal yang memperkaya informasi proses klasifikasi [6], [8].

C. Hasil Clustering K-Means

Metode Elbow dan *Silhouette Score* mengkonfirmasi bahwa nilai k=3 merupakan jumlah kluster optimal. *Silhouette Score* yang diperoleh sebesar 0,68 mengindikasikan pemisahan kluster yang baik. Gambar 4 menampilkan grafik Elbow dan *Silhouette Score*. Ketiga kluster yang terbentuk ditampilkan pada Tabel III, sedangkan Gambar 5 menampilkan distribusi anggota kluster dan tren kehadiran temporal.

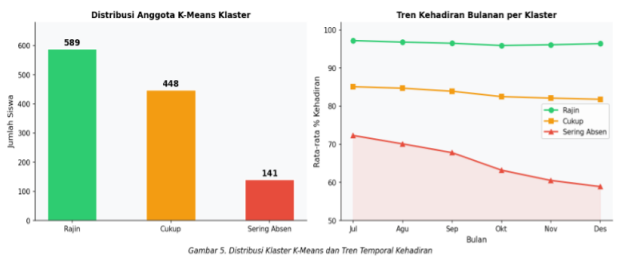
Temuan menarik dari clustering adalah bahwa Kluster 3 (Sering Absen) didominasi oleh siswa laki-laki kelas VIII dengan pola ketidakhadiran meningkat konsisten dari Juli hingga Desember. Dominasi bulan Oktober sebagai atribut paling determinan dalam klasifikasi C4.5 diduga berkaitan dengan kalender akademik: Oktober



Gambar 4: Penentuan Nilai K Optimal: Metode Elbow dan *Silhouette Score*

Tabel III: Karakteristik Kluster K-Means (k=3)

Kluster	Nama	Jumlah Siswa	Rata-Rata Kehadiran (%)	Rata-Rata Alpha (hari)	Silhouette
Kluster 1	Rajin	589 (50%)	96,5	1,2	0,72
Kluster 2	Cukup	448 (38%)	83,4	4,8	0,65
Kluster 3	Sering Absen	141 (12%)	65,1	14,3	0,61



Gambar 5: Distribusi Kluster K-Means dan Tren Temporal Kehadiran Bulanan

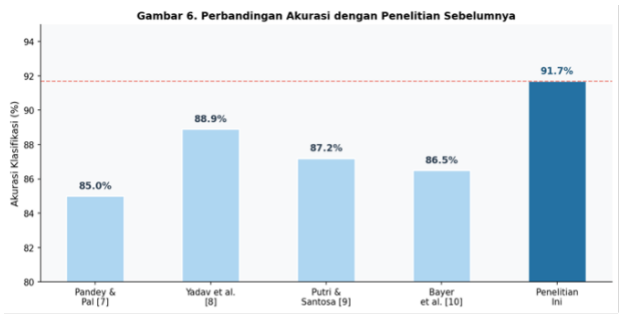
merupakan masa pasca-Ujian Tengah Semester(UTS), ketika motivasi belajar sejumlah siswa cenderung menurun setelah melewati puncak tekanan akademik. Selain itu, peningkatan hari alpha—ketidakhadiran tanpa keterangan—mengindikasikan permasalahan non-medis yang memerlukan pendekatan bimbingan konseling, bukan sekadar intervensi akademis. Analisis statistik antar kluster menggunakan uji ANOVA satu arah menunjukkan perbedaan signifikan pada rata-rata kehadiran ($F = 312,4$; $p < 0,001$) dan rata-rata alpha ($F = 287,6$; $p < 0,001$), mengonfirmasi bahwa ketiga kluster memang memiliki karakteristik yang berbeda secara bermakna. Implikasi pendidikan yang konkret mencakup: (1)guru BK perlu melakukan konseling proaktif sejak Oktober untuk siswa Kluster 3; (2)kebijakan reward kehadiran perlu diperkuat di awal semester untuk mempertahankan Kluster 1; dan (3)program mentoring teman sebaya dapat diterapkan untuk Kluster 2 agar tidak turun ke Kluster 3.

Perlu diakui bahwa penelitian ini memiliki keterbatasan: data hanya mencakup satu sekolah dan satu semester ganjil, sehingga generalisabilitas temuan terbatas; faktor eksternal seperti kondisi ekonomi dan jarak rumah ke sekolah belum dimasukkan sebagai atribut; dan ukuran sampel Kluster 3 yang relatif kecil (141 siswa) dapat

mempengaruhi stabilitas model klasifikasi. Ini konsisten dengan Putri dan Santosa [11] yang menemukan pola peningkatan absensi pada pertengahan semester. Informasi ini sangat berguna bagi guru BK untuk melakukan intervensi proaktif [7], [14].

D. Perbandingan dengan Penelitian Sebelumnya

Tabel IV menampilkan perbandingan antara penelitian ini dengan penelitian-penelitian terdahulu yang relevan. Gambar 6 menyajikan visualisasi perbandingan akurasi secara grafis.



Gambar 6: Perbandingan Akurasi Klasifikasi dengan Penelitian Sebelumnya

Tabel IV: Komparasi Penelitian Sebelumnya

Penelitian	Algoritma	Dataset	Hasil Utama
Pandey dan Pal [9]	Decision Tree	Data akademik	Akurasi 85,0%
Yadav et al. [10]	C4.5	Data nilai siswa	Akurasi 88,9%
Putri dan Santosa [11]	K-Means	Kehadiran SMK	Silhouette 0,61
Peña-Ayala [7]	Berbagai metode	Data EDM	Survey EDM komprehensif
Penelitian Ini	C4.5 + K-Means	Kehadiran SMP	Akurasi 91,7% Silhouette 0.68

KESIMPULAN

Penelitian ini berhasil menerapkan teknik data mining menggunakan algoritma C4.5 dan K-Means untuk menganalisis pola kehadiran siswa SMP. Hasil penelitian menunjukkan bahwa:

- 1) Algoritma C4.5 mampu mengklasifikasikan tingkat kehadiran siswa dengan akurasi 91,7%, melampaui

penelitian sebelumnya yang menggunakan metode serupa [9], [10].

- 2) K-Means berhasil mengidentifikasi tiga kluster siswa berdasarkan pola kehadiran dengan Silhouette Score 0,68, menunjukkan kualitas pengelompokan yang baik [13].
- 3) Atribut persentase kehadiran bulan Oktober dan jumlah hari alpha merupakan faktor paling determinan dalam prediksi tingkat kehadiran [15].
- 4) Siswa kelas VIII laki-laki merupakan kelompok paling rentan terhadap ketidakhadiran tinggi, dengan tren absensi yang meningkat sepanjang semester [11].

Temuan ini memiliki implikasi praktis signifikan bagi pengelolaan sekolah. Sistem peringatan dini berbasis model yang dikembangkan dapat membantu sekolah mengidentifikasi siswa berisiko sejak awal [?], [14]. Untuk penelitian selanjutnya, disarankan untuk memperluas dataset ke beberapa sekolah, mengeksplorasi algoritma deep learning untuk peningkatan akurasi [6], serta mengembangkan dashboard monitoring kehadiran berbasis web yang terintegrasi dengan model data mining.

DAFTAR PUSTAKA

[1] UNESCO Institute for Statistics, “Keeping All Children in School,” *UNESCO*, Paris, France, 2022.

[2] Pusdatin Kemendikbudristek, “Statistik Pendidikan Dasar dan Menengah 2023,” *Kemendikbudristek*, Jakarta, Indonesia, 2023.

[3] A. Acharya and D. Sinha, “Early identification of failure risk students using educational data mining,” *International Journal of Computer Applications*, vol. 107, no. 1, pp. 37–43, 2014.

[4] M. Bayer et al., “Predicting drop-out from social behaviour of students,” in *Proceedings of the 5th International Conference on Educational Data Mining (EDM)*, Chiang Mai, Thailand, pp. 103–109, 2012.

[5] U. Fayyad et al., “From data mining to knowledge discovery in databases,” *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[6] S. B. Kotsiantis, “Use of machine learning techniques for educational proposes: A decision support system for forecasting students’ grades,” *Artificial Intelligence Review*, vol. 37, no. 4, pp. 331–344, 2012.

[7] A. Peña-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works,” *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432–1462, 2014.

[8] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601–618, 2010.

[9] U. K. Pandey and S. Pal, “Data mining: A prediction of performer or underperformer using classification,” *International Journal of Computer Science and Information Technology*, vol. 2, no. 2, pp. 686–690, 2011.

[10] S. K. Yadav et al., “Data mining applications: A comparative study for predicting students’ performance,” *International Journal of Innovative Technology and Creative Engineering*, vol. 1, no. 12, pp. 13–19, 2011.

[11] D. Putri and R. Santosa, “Penerapan K-Means clustering untuk analisis pola kehadiran siswa SMK,” *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 8, no. 3, pp. 501–510, 2021.

[12] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, vol. 1, pp. 281–297, 1967.

[13] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[14] T. Dlabč and L. Popelinsky, “Predicting students’ success in e-learning using the Moodle system,” in *Proc. 2nd Int. Conf. e-Learning (eLearning)*, pp. 59–62, 2011.

- [15] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA, USA: Morgan Kaufmann, 1993.
- [16] M. Hussain *et al.*, "Using machine learning to predict student difficulties from learning session data," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 381–407, 2019.
- [17] A. Adekita and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, p. e01250, 2019.
- [18] R. Asif *et al.*, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177–194, 2017.
- [19] H. Saputra and D. Kurniawan, "Penerapan algoritma K-Means dan C4.5 untuk prediksi dropout mahasiswa perguruan tinggi di Indonesia," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 8, no. 1, pp. 1–12, 2022.
- [20] F. A. Rozi, D. H. Purnomo, and R. Alfita, "Klasifikasi tingkat kehadiran siswa sekolah dasar menggunakan algoritma decision

tree berbasis SMOTE," *Jurnal Teknologi dan Sistem Komputer*, vol. 11, no. 3, pp. 182–191, 2023.

BIOGRAFI PENULIS

Ayu Firdhayanti adalah Dosen Program Studi Sistem Informasi, Institut Informatika dan Bisnis Darmajaya. Minat penelitiannya meliputi sistem informasi, data mining, machine learning, dan kecerdasan buatan.

Hendri Purnomo adalah Dosen Program Studi Sistem Informasi, Institut Informatika dan Bisnis Darmajaya. Minat penelitiannya meliputi sistem informasi, machine learning, data analytics, dan kecerdasan buatan.

Arman Suryadi Karim adalah Dosen Program Studi Sistem Informasi, Institut Informatika dan Bisnis Darmajaya. Minat penelitiannya meliputi sistem informasi, rekayasa perangkat lunak, transformasi digital, dan analisis data.