Jurnal Informatika Terpadu

https://journal.nurulfikri.ac.id/index.php/JIT
ISSN ONLINE : 2460-8998

DETEKSI POLA KETIDAKAKTIFAN AKADEMIK MAHASISWA PADA MATA KULIAH PEMBELAJARAN MESIN MENGGUNAKAN DBSCAN

Mohamad Ardi[1], Endyk Noviyantono[2], Eviana Tjatur Putri[3]

[1,2]Teknik Informatika, STMIK PPKIA Tarakanita Rahmawati
[3]Sistem Informasi, STMIK PPKIA Tarakanita Rahmawati
Tarakan, Kalimantan Utara, Indonesia 77113
mohamadardi@ppkia.ac.id, endyk@ppkia.ac.id, eviana@ppkia.ac.id

Abstract

Lecturers often recognize students at risk of receiving a grade E only after the final examination. This study explores whether Density-Based Spatial Clustering of Applications with Noise (DBSCAN) can identify academic inactivity patterns associated with grade E before the final examination. The data comprised 142 records of a machine learning course across two odd semesters of the 2024/2025 and 2025/2026 academic years, described by five pre-final features: participation, project, assignment, quiz, and midterm scores. In this dataset, every grade E coincided with a zero final-examination score, so "at risk" denotes non-completion of the final examination. DBSCAN parameters were selected using internal validity indices without using final grades. DBSCAN formed two clusters and 17 noise records. In a retrospective evaluation, the 25 records outside the main cluster included 21 of 28 grade-E records (precision 0.84, recall 0.75, F1-score 0.79, MCC 0.75). Isolation Forest performed identically, and cross-semester validation showed limited generalization (F1-score 0.39–0.40). DBSCAN therefore offers an interpretable exploratory approach for identifying inactivity patterns, not a validated early-warning system. Future work should use prospective validation and attendance or learning-log data.

Keywords: *Anomaly Detection, At-Risk Students, DBSCAN, Educational Data Mining, Unsupervised Learning*

Abstrak

Dosen sering baru mengenali mahasiswa yang berisiko memperoleh nilai E setelah ujian akhir. Penelitian ini mengeksplorasi apakah *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) dapat mengidentifikasi pola ketidakaktifan akademik yang berasosiasi dengan nilai E sebelum ujian akhir. Data terdiri atas 142 rekaman nilai mata kuliah Pembelajaran Mesin dari dua semester ganjil tahun akademik 2024/2025 dan 2025/2026, yang dianonimkan dan dideskripsikan oleh lima fitur pra-UAS: partisipasi, proyek, tugas, kuis, dan UTS. Pada data ini seluruh nilai E bertepatan dengan UAS bernilai nol, sehingga "berisiko" berarti tidak menyelesaikan ujian akhir. Parameter DBSCAN dipilih dengan indeks validitas internal tanpa memakai nilai akhir. DBSCAN membentuk dua klaster dan 17 rekaman *noise*. Sebanyak 25 rekaman di luar klaster utama mencakup 21 dari 28 rekaman bernilai E (*precision* 0,84, *recall* 0,75, *F1-score* 0,79, MCC 0,75) pada evaluasi retrospektif. *Isolation Forest* menghasilkan kinerja identik, dan validasi lintas semester menunjukkan generalisasi yang terbatas (*F1-score* 0,39–0,40). DBSCAN karenanya merupakan pendekatan eksploratif yang mudah diinterpretasi untuk mengidentifikasi pola ketidakaktifan, bukan sistem peringatan dini yang telah tervalidasi. Penelitian lanjutan perlu menggunakan validasi prospektif serta data kehadiran atau log pembelajaran.

Kata kunci: DBSCAN, Deteksi Anomali, *Educational Data Mining*, Mahasiswa Berisiko, *Unsupervised Learning*

1. PENDAHULUAN

Perguruan tinggi bidang informatika membutuhkan cara yang sistematis untuk memantau capaian belajar mahasiswa. *Educational data mining* dan *learning analytics* telah berkembang pesat untuk tujuan itu [1], [2], dan pemanfaatannya di pendidikan tinggi terus meluas [3]. Sebagian besar penelitian berfokus pada prediksi capaian atau keberhasilan studi dengan pendekatan terawasi (*supervised*) yang memerlukan label hasil belajar [4], [5], [6].

Pada praktiknya, label tersebut baru tersedia setelah ujian akhir selesai, sehingga dosen kerap mengenali mahasiswa berisiko ketika ruang untuk intervensi sudah tertutup. Pada tingkat satu mata kuliah, jumlah data juga kecil sehingga model terawasi mudah tidak stabil. Pendekatan tak terawasi

Diterima 14 Oktober 2026, Direvisi 16 Oktober 2026, Diterima untuk publikasi 3 November 2026

(*unsupervised*) menawarkan alternatif karena tidak membutuhkan label. Pengelompokan nilai mahasiswa selama ini didominasi K-Means [7], [8], [9], padahal K-Means mensyaratkan jumlah klaster di awal dan peka terhadap *outlier* [10], [11], [12].

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [13], [14] tidak memerlukan jumlah klaster, dapat menemukan klaster berbentuk sembarang, dan secara eksplisit melabeli titik yang tidak padat sebagai *noise*. Dalam konteks pendidikan, DBSCAN telah dipakai pada pembelajaran campuran [15], jejak digital LMS [16], dan pemetaan kesiapan karier [17]. Masalah peringatan dini mahasiswa yang kesulitan juga telah dirumuskan sebagai masalah deteksi anomali [18].

Objek penelitian ini adalah mata kuliah Pembelajaran Mesin dengan kode TMA05 pada STMIK PPKIA Tarakanita Rahmawati selama dua semester ganjil tahun akademik 2024/2025 dan 2025/2026. Dari 142 rekaman nilai, 28 (19,7%) berakhir dengan nilai E, dan proporsinya naik dari 14,8% pada 2024/2025 menjadi 22,7% pada 2025/2026. Pada seluruh rekaman tersebut, nilai E bertepatan dengan nilai UAS sama dengan nol. Dengan demikian, komponen penilaian sebelum UAS dapat diuji untuk mengidentifikasi pola ketidakaktifan yang berasosiasi dengan nilai E.

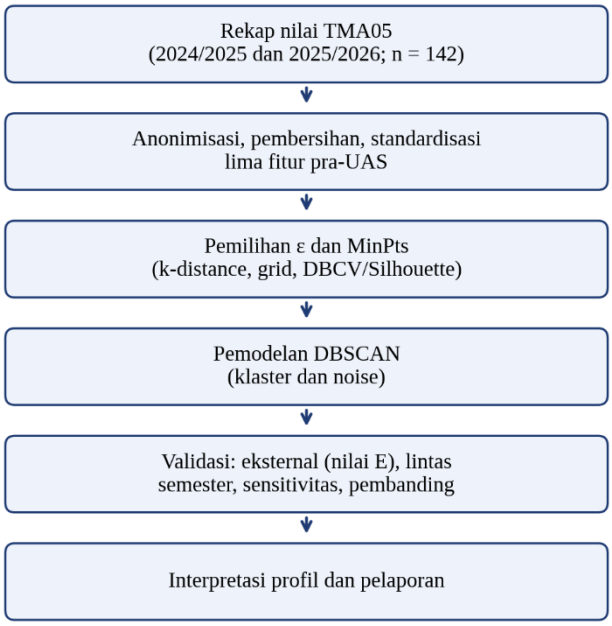
Studi terdahulu dapat dibedakan menurut peran DBSCAN dan cara validasinya. Pada kelompok pertama, DBSCAN berfungsi sebagai penyaring *outlier* sebelum K-Means [15], [17], sehingga mutu deteksinya sendiri tidak dievaluasi terhadap luaran belajar. Pada kelompok kedua, klasterisasi diterapkan pada log LMS berskala besar untuk prediksi *dropout* [16], dan metode terbaik ternyata bergantung pada data sehingga tidak ada metode yang selalu unggul. Pada kelompok ketiga, peringatan dini dirumuskan sebagai deteksi anomali [18], tetapi pada konteks dan data yang berbeda dari skor penilaian satu mata kuliah. Dari ketiganya tampak tiga keterbatasan: DBSCAN jarang diuji langsung sebagai pendeteksi mahasiswa berisiko pada skor mata kuliah; pemilihan parameter jarang dijelaskan dan diuji sensitivitasnya [17]; serta hasil deteksi jarang divalidasi lintas semester dan dibandingkan dengan detektor anomali lain pada anggaran penanda yang sama [19], [20], [21].

Penelitian ini berbeda dari studi terdahulu pada empat hal. Pertama, DBSCAN diperlakukan langsung sebagai pendeteksi, dengan definisi operasional penanda yang eksplisit. Kedua, parameter dipilih dengan aturan yang dapat direproduksi, dan hasilnya terbukti identik di seluruh wilayah plateau parameter. Ketiga, hasil divalidasi terhadap nilai E dengan metrik yang memperhitungkan kelas minoritas dan dibandingkan dengan detektor lain, termasuk aturan sederhana, pada anggaran penanda yang seragam. Keempat, generalisasi diuji lintas semester dan hasil yang kurang baik dilaporkan secara terbuka.

Berdasarkan hal tersebut, rumusan masalah penelitian ini adalah: (1) pola capaian belajar apa yang dibentuk DBSCAN pada lima komponen penilaian pra-UAS; (2) seberapa baik rekaman yang ditandai DBSCAN berasosiasi dengan mahasiswa bernilai E; (3) seberapa sensitif dan stabil hasil tersebut terhadap parameter, penskalaan, dan rekaman berulang, serta bagaimana kinerjanya dibandingkan dengan detektor lain pada anggaran penanda yang seragam; dan (4) apakah hasil tersebut bertahan lintas semester. Tujuan penelitian adalah menjawab keempat pertanyaan itu. Penelitian ini bersifat eksploratif dan retrospektif, dan tidak dimaksudkan membuktikan kemampuan prediksi. Batasan penelitian: satu mata kuliah, dua semester ganjil, lima fitur nilai, tanpa data kehadiran dan demografi, serta nilai E dipakai hanya sebagai acuan validasi dan bukan sebagai label latih.

2. METODE PENELITIAN

Penelitian ini bersifat kuantitatif eksploratif dengan pendekatan *unsupervised learning*. Tahapan mengikuti alur *knowledge discovery in databases* seperti pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1 Data, Instrumen, dan Metode Pengujian

Data berupa rekap nilai mata kuliah Pembelajaran Mesin semester ganjil 2024/2025 (54 rekaman) dan 2025/2026 (88 rekaman), total 142 rekaman. Lima mahasiswa tercatat pada kedua tahun ajaran (10 rekaman), sehingga setiap kemunculan diperlakukan sebagai rekaman terpisah. Kelimanya seluruhnya bernilai E pada 2024/2025 dan tercatat kembali pada 2025/2026 (tiga kembali bernilai E dan dua bernilai AB), sehingga rekaman berulang diperkaya nilai E dan observasi tidak sepenuhnya independen. Dampaknya diuji melalui analisis sensitivitas tanpa kemunculan kedua (n = 137) dan tanpa seluruh rekaman berulang (n = 132). Identitas mahasiswa diganti ID acak, data dianalisis dalam bentuk anonim, dan hasil dilaporkan

secara agregat. Tabel 1 menunjukkan sebagian *dataset* yang digunakan pada penelitian ini.

Tabel 1. *Dataset* Penelitian (Part. = Partisipatif)

ID	Tahun	Part.	Proyek	Tugas	Kuis	UTS	UAS	*Grade*
M001	2024/2025	90	70	70	90	60	70	B
M002	2024/2025	65	70	70	65	60	0	E
M003	2024/2025	65	70	70	65	75	69	B
M004	2024/2025	90	100	100	90	65	70	B
M005	2024/2025	90	100	100	90	80	75	AB
...	...	...	...	...	...	...	...	...
M053	2024/2025	90	100	100	90	87	89	A
M054	2024/2025	90	100	100	90	70	85	AB
M055	2025/2026	85	100	0	100	85	0	E
M056	2025/2026	85	85	100	100	77	70	AB
M057	2025/2026	85	85	70	100	77	70	AB
...	...	...	...	...	...	...	...	...
M138	2025/2026	0	0	0	0	0	0	E
M139	2025/2026	85	100	70	100	75	70	AB
M140	2025/2026	0	0	0	0	0	0	E
M141	2025/2026	85	0	0	100	70	0	E
M142	2025/2026	85	100	70	100	70	80	AB

Fitur yang dipakai adalah lima komponen yang tersedia sebelum UAS: aktivitas partisipatif, hasil proyek, tugas, kuis, dan UTS (skala 0–100). Kolom UAS, Total, dan *Grade* tidak dimasukkan ke pemodelan. Alasannya, nilai E pada data ini identik dengan UAS bernilai nol, sehingga memasukkan UAS akan membuat deteksi menjadi sepele dan membocorkan label (*data leakage*). Pada 2024/2025, rekap mencatat nilai yang identik antara partisipatif dan kuis serta antara proyek dan tugas pada seluruh rekaman. Hal ini dicatat sebagai keterbatasan data.

Dalam penelitian ini, *mahasiswa berisiko* didefinisikan secara operasional sebagai mahasiswa yang rekamannya ditandai DBSCAN sebagai *noise* atau berada di luar klaster utama berdasarkan lima komponen pra-UAS, dengan nilai akhir E sebagai acuan validasi. Pada *dataset* ini seluruh nilai E bertepatan dengan UAS bernilai nol, sehingga E merepresentasikan tidak diselesaikannya ujian akhir, yang dapat berkaitan dengan ketidakhadiran, tidak mengerjakan ujian, atau kebijakan penilaian, dan tidak dapat dibedakan dari penguasaan materi yang rendah. Istilah "berisiko" dalam artikel ini karenanya berarti berisiko memperoleh nilai E, bukan kegagalan akademik dalam pengertian umum. Tabel 2 menunjukkan ringkasan *dataset* yang digunakan pada riset ini.

Tabel 2. Ringkasan *Dataset*

Tahun Ajaran	Rekaman	Nilai E	% E	Rerata Total
2024/2025	54	8	14,8	72,6
2025/2026	88	20	22,7	70,4
Gabungan	142	28	19,7	71,3

2.2 Tahapan Penelitian

Tahap pertama adalah praproses: pembuangan baris kosong, anonimisasi, dan standardisasi *z-score* pada lima fitur. Standardisasi dilakukan pada *dataset* gabungan (142 rekaman) dengan rerata dan simpangan baku populasi dari seluruh rekaman, sehingga ε bermakna pada skala yang sama untuk kedua tahun ajaran. Mengingat perbedaan distribusi nilai dan bobot penilaian antartahun, standardisasi per tahun ajaran dan penskalaan Min-Max diuji sebagai analisis sensitivitas. Nilai nol diperlakukan sebagai nilai sebenarnya (komponen tidak dikerjakan), bukan data hilang.

Tahap kedua adalah pemilihan parameter DBSCAN tanpa memakai nilai akhir. Rentang ε diperkirakan dari *k-distance plot*, lalu diuji pada *grid* MinPts 3–12 dan ε 0,3–2,0 dengan langkah 0,1 (180 kombinasi). Pemilihan mengikuti tiga langkah: (1) hanya kombinasi valid yang dipertimbangkan, yaitu yang menghasilkan minimal dua klaster dengan proporsi *noise* lebih dari nol dan tidak melebihi 30%; (2) kombinasi valid diurutkan menurut DBCV [22]; dan (3) bila selisih DBCV tidak lebih dari 0,001, dipilih *Silhouette Coefficient* [23] tertinggi yang dihitung tanpa titik *noise*. *Davies-Bouldin Index* dilaporkan sebagai pembanding. Bila beberapa kombinasi menghasilkan partisi identik, salah satunya dipilih sebagai konfigurasi akhir.

Tahap ketiga adalah pemodelan dengan DBSCAN (jarak Euclidean). Dua definisi penanda risiko dipakai: (a) titik *noise* saja, dan (b) seluruh rekaman di luar klaster terbesar (*noise* ditambah klaster kecil). Definisi (b) dipakai karena DBSCAN dapat membentuk klaster padat yang tidak mewakili perilaku mayoritas. Kedua definisi dilaporkan berdampingan beserta dekomposisinya agar kontribusi masing-masing terhadap *recall* dapat dinilai. Tahap keempat adalah validasi eksternal terhadap nilai E dengan *precision*, *recall*, *F1-score*, *Matthews correlation coefficient* (MCC) [24], dan uji eksak Fisher.

Tahap kelima adalah validasi lintas semester dua arah (2024/2025 sebagai data latih dan 2025/2026 sebagai data uji, dan sebaliknya). Penskala dan DBSCAN dipasang hanya pada semester latih. Karena DBSCAN tidak memiliki fungsi prediksi bawaan, rekaman semester uji ditugaskan ke klaster titik inti terdekat bila jaraknya tidak melebihi ε, dan selain itu dianggap *noise*. Rekaman yang tidak ditugaskan ke klaster terbesar semester latih ditandai berisiko. Dua skema parameter dibandingkan: parameter tetap (ε = 1,5;

MinPts = 6) dan parameter yang dipilih ulang pada semester latih dengan aturan yang sama.

Tahap keenam adalah perbandingan dengan detektor lain, yaitu *Local Outlier Factor* (LOF) [25], *Isolation Forest* [26], dan HDBSCAN [27]. Karena DBSCAN menentukan sendiri jumlah penandanya, LOF, *Isolation Forest*, dan skor *outlier* HDBSCAN (GLOSH) dibandingkan pada tiga anggaran penanda yang seragam (17, 25, dan 36), yaitu jumlah *noise* DBSCAN, jumlah rekaman di luar klaster utama, dan jumlah *noise* HDBSCAN dengan konfigurasi bawaan [19], [20]. Dua aturan sederhana ditambahkan sebagai pembanding: rerata lima komponen di bawah 50, dan adanya komponen bernilai nol. Ambang 50 dipilih sebagai angka bulat dan bukan hasil penyetelan terhadap nilai E.

Tahap ketujuh adalah uji ketahanan yang terdiri atas peta sensitivitas ε dan MinPts, analisis per tahun ajaran, ablasi tanpa fitur UTS, serta dua prosedur kestabilan. Prosedur pertama adalah *repeated subsampling*: 200 subsampel 80% tanpa pengembalian, dengan standardisasi dan DBSCAN dipasang ulang pada setiap ulangan. *Adjusted Rand Index* (ARI) dihitung antara label model penuh yang dibatasi pada anggota subsampel dan label model subsampel; indeks Jaccard dihitung pada irisan anggota yang sama antara himpunan rekaman ditandai oleh model penuh dan oleh model subsampel; interval *F1-score* adalah persentil 2,5 dan 97,5 dari *F1-score* terhadap nilai E pada anggota subsampel. Prosedur kedua adalah *bootstrap* standar dengan pengambilan sampel berpengembalian (200 ulangan), dengan evaluasi pada rekaman *out-of-bag* yang ditugaskan melalui titik inti terdekat. Implementasi memakai Python 3, scikit-learn 1.8 [28], dan pustaka hdbscan untuk DBCV, dengan *random seed* 42.

3. HASIL DAN PEMBAHASAN

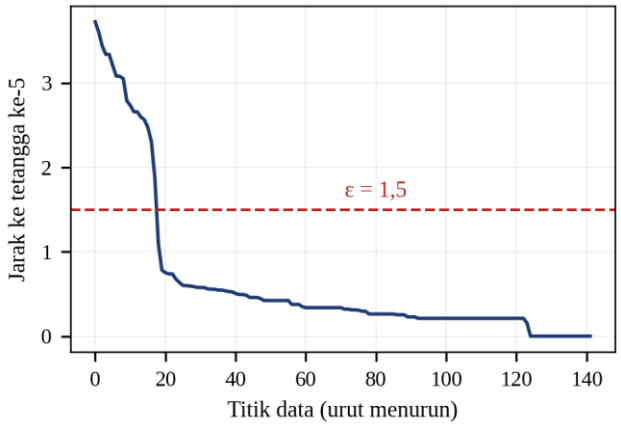
3.1 Karakteristik Data

Tabel 3 menunjukkan bahwa komposisi nilai berbeda antartahun. Nilai tugas turun tajam dari 87,8 menjadi 46,7, sedangkan simpangan baku komponen lain membesar pada 2025/2026. Sebaran nilai huruf juga bergeser: tahun 2024/2025 memuat 6 A, 15 AB, 24 B, 1 BC, dan 8 E, sedangkan 2025/2026 memuat 30 A, 38 AB, 20 E, serta tidak ada B dan BC. Regresi linear nilai total terhadap enam komponen memberi bobot yang berbeda antartahun (UTS 30% dan UAS 40% pada 2024/2025, sedangkan partisipatif dan proyek masing-masing 25% serta UTS dan UAS masing-masing 20% pada 2025/2026). Perbedaan skema ini memperkuat alasan analisis sensitivitas dan analisis per tahun. Terdapat 6 rekaman dengan seluruh komponen pra-UAS bernilai nol (1 pada 2024/2025 dan 5 pada 2025/2026), dan seluruhnya bernilai E.

Tabel 3. Statistik Deskriptif Komponen (Rerata ± Simpangan Baku)

Komponen	2024/2025	2025/2026
Partisipatif	80,6 ± 16,2	71,5 ± 32,8
Proyek	87,8 ± 18,5	73,8 ± 38,2
Tugas	87,8 ± 18,5	46,7 ± 28,8
Kuis	80,6 ± 16,2	83,0 ± 37,8
UTS	69,2 ± 17,3	76,0 ± 26,8
UAS (tidak dipakai)	66,4 ± 29,0	61,7 ± 34,7

3.2 Penentuan Parameter



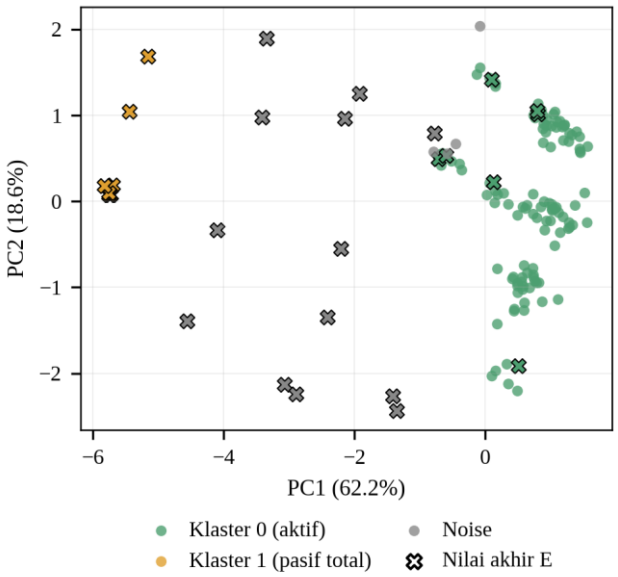
Gambar 2. *K-distance Plot* (MinPts = 6)

K-distance plot pada Gambar 2 memperlihatkan celah lebar antara titik padat (jarak kecil) dan titik jarang (jarak besar), sehingga ε dapat dipilih pada rentang sekitar 1,3–2,0. Dari 180 kombinasi *grid*, 105 kombinasi memenuhi kriteria validitas. DBCV tertinggi (0,650), *Silhouette* tertinggi (0,775), dan *Davies-Bouldin* terendah (0,258) terdapat pada solusi yang sama, yaitu dua klaster dengan 12,0% *noise* dan *Calinski-Harabasz* 231,0. Korelasi peringkat Spearman antara DBCV dan *Silhouette* pada seluruh kombinasi valid adalah 0,968, sehingga tidak terjadi konflik antarindeks. Sebanyak 27 kombinasi (MinPts 5–8; ε 1,3–2,0) menghasilkan partisi yang identik (ARI = 1,000). Konfigurasi akhir ε = 1,5 dan MinPts = 6 dipilih di dalam wilayah tersebut, sehingga pilihan di dalamnya tidak memengaruhi hasil. Tabel 4 memperlihatkan bahwa konfigurasi yang terlalu ketat memecah data menjadi tiga hingga empat klaster dengan mutu rendah, sedangkan MinPts yang terlalu besar menyatukan seluruh data menjadi satu klaster.

Tabel 4. Contoh Konfigurasi DBSCAN Pada *Grid*

MinPts	ε	Klaster	*Noise* (%)	*Silhouette*	DBCV
3	1,0	4	10,6	0,322	0,279
4	1,5	4	6,3	0,580	0,529
6	1,0	3	13,4	0,305	0,256
6	1,5	2	12,0	0,775	0,650
6	2,0	2	12,0	0,775	0,650
8	1,5	2	12,0	0,775	0,650
10	1,5	1	17,6	–	–

3.3 Klaster yang Terbentuk



Gambar 3. Proyeksi PCA 2D Klaster DBSCAN

DBSCAN membentuk dua klaster dan 17 titik *noise* yang ditunjukkan pada Gambar 3. Klaster 0 berisi 117 rekaman (82,4%) yang aktif secara konsisten dengan rerata partisipatif 84,6, proyek 91,2, kuis 92,2, dan UTS 79,6 yang ditunjukkan pada Tabel 5. Klaster 1 berisi 8 rekaman (5,6%) yang pasif total, yaitu hampir seluruh komponen bernilai nol, dan seluruhnya bernilai E. Kelompok *noise* heterogen: sebagian hanya aktif pada satu atau dua komponen (misalnya partisipatif dan proyek nol tetapi kuis 100), sebagian lagi kuat di semua komponen kecuali kuis yang bernilai nol. Menurut tahun ajaran, *noise* berjumlah 1 rekaman pada 2024/2025 dan 16 rekaman pada 2025/2026, sedangkan Klaster 1 berjumlah 1 dan 7 rekaman. Pola ketidakaktifan parsial jauh lebih banyak muncul pada 2025/2026.

Tabel 5. Profil Klaster (Rerata ± Simpangan Baku)

Fitur	Klaster 0 (n=117)	Klaster 1 (n=8)	*Noise* (n=17)
Partisipatif	84,6 ± 8,6	0,0 ± 0,0	43,8 ± 42,8
Proyek	91,2 ± 10,9	0,0 ± 0,0	32,6 ± 46,1
Tugas	68,8 ± 28,2	11,9 ± 23,0	41,2 ± 33,0
Kuis	92,2 ± 11,7	0,0 ± 0,0	50,9 ± 50,1
UTS	79,6 ± 11,4	0,0 ± 0,0	65,4 ± 31,7
Nilai E	7	8	13

3.4 Validasi Eksternal Terhadap Nilai E

Penanda berupa *noise* saja menangkap 13 dari 28 mahasiswa bernilai E (*precision* 0,765; *recall* 0,464; *F1-score* 0,578; MCC 0,526; rasio odds 23,8; $p < 0,001$). Penanda berupa seluruh rekaman di luar klaster utama (25 rekaman) menangkap 21 dari 28 mahasiswa bernilai E dengan 4 positif palsu (*precision* 0,840; *recall* 0,750; *F1-score* 0,792; MCC 0,747; rasio odds 82,5; $p < 0,001$). Dari 117 rekaman Klaster 0, hanya 7 (6,0%) yang bernilai E, sedangkan pada rekaman yang ditandai proporsinya 84,0%. Kesesuaian klaster dengan nilai huruf lengkap rendah (ARI 0,124; NMI 0,274). Dengan demikian klaster memisahkan pola ketidakaktifan, bukan membedakan A dari B.

Dekomposisi menunjukkan bahwa 17 titik *noise* terdiri atas 13 rekaman bernilai E dan 4 rekaman bukan E, sedangkan Klaster 1 (8 rekaman) seluruhnya bernilai E. Kenaikan *recall* dari 0,464 pada penanda *noise* menjadi 0,750 pada penanda di luar klaster utama seluruhnya berasal dari Klaster 1. Enam dari delapan rekaman Klaster 1 memiliki seluruh komponen pra-UAS bernilai nol sehingga dapat dikenali tanpa DBSCAN. Kemampuan DBSCAN yang tidak trivial tampak pada titik *noise*: dari 20 rekaman bernilai E di luar Klaster 1, 13 (65%) ditandai sebagai *noise* dengan *precision* 0,765. Evaluasi ini bersifat retrospektif pada sampel yang sama dan bukan bukti generalisasi.

Empat positif palsu seluruhnya mahasiswa bernilai AB yang kuisnya nol tetapi lulus UAS. Tujuh negatif palsu seluruhnya berada di Klaster 0 dengan komponen awal yang memadai (partisipatif 65–90 dan UTS 60–85), tetapi tidak mengikuti UAS. Kegagalan jenis ini mungkin berkaitan dengan faktor di luar skor dan tidak dapat dideteksi dari nilai komponen. Tabel 6 menunjukkan validasi eksternal DBSCAN terhadap nilai E.

Tabel 6. Validasi Eksternal DBSCAN Terhadap Nilai E (TP/FP/FN: Positif Benar/Positif Palsu/Negatif Palsu)

Penanda	Ditandai	TP/FP/FN	*Prec.*	*Rec.*	F1	MCC
Noise Saja	17	13/4/15	0,765	0,464	0,578	0,526
Di Luar Klaster Utama	25	21/4/7	0,840	0,750	0,792	0,747

3.5 Perbandingan Dengan Detektor Lain

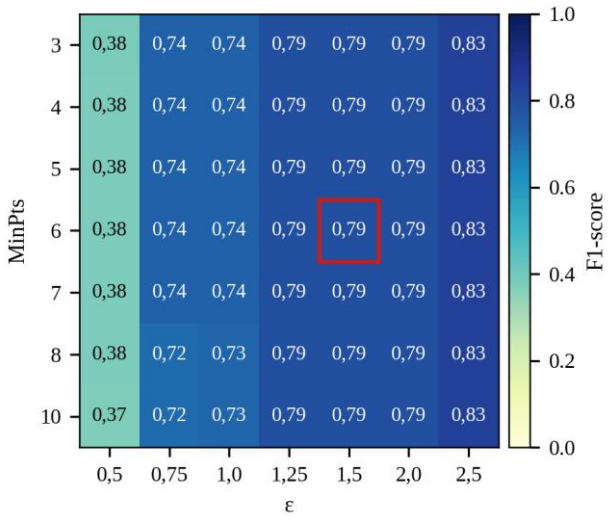
Tabel 7 membandingkan DBSCAN dengan tiga detektor pada anggaran penanda yang seragam. *Isolation Forest* menghasilkan kinerja identik dengan DBSCAN pada 25 penanda (21 TP, 4 FP) dan sedikit lebih baik pada 17 penanda (F1 0,622 terhadap 0,578). Pada 36 penanda, LOF memberi *F1-score* tertinggi (0,719), sedangkan skor HDBSCAN paling rendah (0,469); HDBSCAN dengan konfigurasi bawaan menghasilkan tepat 36 titik *noise* dan tujuh klaster sehingga *precision*-nya rendah (0,417) dan tidak sepenuhnya sebanding dengan metode lain. Nilai AUC skor kontinu LOF, *Isolation Forest*, dan skor HDBSCAN masing-masing 0,906, 0,926, dan 0,670. DBSCAN tidak dapat dipaksa ke anggaran tertentu karena jumlah penandanya ditentukan oleh kepadatan data. Dua aturan sederhana mencapai *F1-score* 0,727 dan 0,746, hanya sedikit di bawah DBSCAN (0,792). Dengan demikian DBSCAN setara, bukan lebih unggul, daripada detektor yang lebih sederhana pada data kecil ini. Nilai tambahnya terletak pada interpretabilitas profil klaster dan penentuan jumlah penanda secara otomatis.

Tabel 7. Perbandingan Detektor pada Anggaran Penanda yang Seragam

Metode	Ditandai	*Prec.*	*Rec.*	F1	MCC
DBSCAN (*Noise*)	17	0,765	0,464	0,578	0,526
DBSCAN (Luar Klaster Utama)	25	0,840	0,750	0,792	0,747
Isolation Forest	17	0,824	0,500	0,622	0,581
Isolation Forest	25	0,840	0,750	0,792	0,747
Isolation Forest	36	0,611	0,786	0,688	0,606
LOF	17	0,706	0,429	0,533	0,472
LOF	25	0,600	0,536	0,566	0,468
LOF	36	0,639	0,821	0,719	0,647
HDBSCAN (GLOSH)	17	0,765	0,464	0,578	0,526
HDBSCAN (GLOSH)	25	0,600	0,536	0,566	0,468
HDBSCAN (GLOSH)	36	0,417	0,536	0,469	0,321
Aturan: Rerata < 50	16	1,000	0,571	0,727	0,719
Aturan: Ada Komponen = 0	31	0,710	0,786	0,746	0,681

3.6 Sensitivitas, Kestabilan, dan Ablasi

Gambar 4 menunjukkan bahwa *F1-score* penanda di luar klaster utama stabil pada 0,79 untuk ε 1,25–2,0 pada seluruh MinPts yang diuji (0,83 pada ε 2,5), dan turun menjadi sekitar 0,72–0,74 pada ε 0,75–1,0 serta 0,37–0,38 pada ε 0,5 saat data terfragmentasi. Pada *repeated subsampling* (200 ulangan), rerata ARI adalah 0,993 (simpangan baku 0,011; minimum 0,967), indeks Jaccard pada irisan anggota sebesar 1,000, dan *F1-score* rerata 0,790 (rentang 95% 0,732–0,851). Kestabilan setinggi ini sebagian dipengaruhi oleh duplikasi data: hanya terdapat 90 vektor fitur unik dari 142 rekaman. Pada *bootstrap* standar, evaluasi pada rekaman *out-of-bag* (rerata 52 rekaman) menghasilkan *F1-score* rerata 0,784 (interval 95% 0,615–0,923) dan MCC 0,739 (0,565–0,915), dengan jumlah klaster yang bervariasi: satu klaster pada 27 ulangan, dua pada 107, tiga pada 59, dan empat pada 7 ulangan. Ketidakpastian ini lebih besar daripada yang tampak pada *subsampling*. Penskalaan Min-Max dengan ε 0,4–0,5 menghasilkan solusi yang sama (dua klaster, 25 penanda, F1 0,792). Ablasi tanpa UTS menghasilkan tiga klaster, 9 *noise*, dan 24 penanda dengan *precision* 0,833, *recall* 0,714, dan *F1-score* 0,769, sehingga sebagian besar sinyal risiko sudah terlihat sebelum UTS. Tabel 8 merangkum sensitivitas terhadap rekaman berulang dan penskalaan.



Gambar 4. Peta Sensitivitas F1-Score Terhadap ε dan MinPts (Kotak Merah: Konfigurasi Terpilih)

Hasil tanpa rekaman berulang serupa dengan hasil utama, sedangkan penskalaan per tahun mengubah keluaran secara nyata. Standardisasi gabungan dipertahankan karena penskalaan per tahun menghilangkan perbedaan tingkat keaktifan antarkohort dan mengubah makna jarak.

Tabel 8. Sensitivitas terhadap Rekaman Berulang dan Penskalaan (Kl. = Jumlah Klaster; Tanda = Jumlah Rekaman Ditandai)

Skenario	n	Kl.	Tanda	*Prec.*	*Rec.*	F1	MCC
Utama (Gabungan)	142	2	25	0,840	0,750	0,792	0,747
Tanpa Kemunculan Kedua	137	2	22	0,864	0,760	0,809	0,771
Tanpa Seluruh Rekaman Berulang	132	1	20	0,850	0,850	0,850	0,823
z per Tahun (ε 1,5; MinPts 6)	142	3	32	0,656	0,750	0,700	0,622
z per Tahun (otomatis: ε 0,4; MinPts 5)	142	7	90	0,289	0,929	0,441	0,303

Hasil per tahun ajaran berbeda tajam. Pada 2025/2026 saja, penanda di luar klaster utama menangkap 19 dari 20 mahasiswa bernilai E (*precision* 0,826; *recall* 0,950; *F1-score* 0,884). Pada 2024/2025 saja, hanya dua rekaman menjadi *noise* (keduanya E) dan *F1-score* penanda di luar klaster utama hanya 0,370. Profil mahasiswa bernilai E menjelaskan perbedaan ini. Pada 2025/2026, rerata komponen pra-UAS mahasiswa bernilai E adalah 29,0 dibanding 82,3 pada mahasiswa lain, dan hanya 2 dari 20 yang memiliki rerata di atas atau sama dengan 60. Pada 2024/2025, perbedaannya jauh lebih kecil (64,7 dibanding 84,1), dan 6 dari 8 mahasiswa bernilai E memiliki rerata pra-UAS di atas atau sama dengan 60 serta nilai UTS lebih dari nol. Dengan demikian nilai E pada 2025/2026 sebagian besar didahului ketidakaktifan yang terlihat pada komponen pra-UAS, sedangkan pada 2024/2025 umumnya terjadi pada mahasiswa yang cukup aktif tetapi tidak mengikuti UAS, suatu pola yang secara prinsip tidak dapat dideteksi dari fitur pra-UAS. Perbedaan bobot penilaian antartahun (UTS 30% dan UAS 40% pada 2024/2025 dibanding partisipatif dan proyek masing-masing 25% serta UTS dan UAS masing-masing 20% pada 2025/2026) menjadi faktor tambahan yang tidak dapat dipisahkan dari perbedaan kohort pada data ini.

3.7 Validasi Lintas Semester

Tabel 9 menunjukkan bahwa generalisasi lintas semester terbatas. Tabel tersebut memiliki beberapa penjelasan yakni (Tetap: ε 1,5 dan MinPts 6; Ulang: Parameter Dipilih Ulang pada Data Latih, ε 1,1 dan MinPts 4 untuk 2024/2025 serta ε 1,0 dan MinPts 5 untuk 2025/2026; Tanda = Jumlah Rekaman Ditandai/Jumlah Rekaman Uji). Model yang dilatih pada 2024/2025 menandai 83–84 dari 88 rekaman 2025/2026 sehingga *precision* hanya sekitar 0,24 meskipun *recall* mencapai 1,000. Hal ini terjadi karena klaster utama 2024/2025 sempit dan bernilai tinggi, sedangkan mahasiswa 2025/2026 lebih beragam sehingga sebagian besar berada di luar klaster tersebut. Pada arah sebaliknya, parameter tetap menandai hanya 2 dari 54 rekaman (*recall* 0,250), sedangkan parameter yang dipilih ulang menandai 17 rekaman dengan *precision* 0,294. *F1-score* pada seluruh skema berkisar 0,385–0,400, jauh di bawah hasil retrospektif pada data gabungan (0,792).

Tabel 9. Validasi Lintas Semester

Latih → uji	Param.	Tanda	TP/FP/FN	*Prec.*	*Rec.*	F1	MCC
2024/25 → 2025/26	Tetap	83/88	20/63/0	0,241	1,000	0,388	0,133
2024/25 → 2025/26	Ulang	84/88	20/64/0	0,238	1,000	0,385	0,118
2025/26 → 2024/25	Tetap	2/54	2/0/6	1,000	0,250	0,400	0,470
2025/26 → 2024/25	Ulang	17/54	5/12/3	0,294	0,625	0,400	0,279

3.8 Pembahasan

Seluruh nilai E pada data ini bertepatan dengan UAS bernilai nol. Kondisi tersebut dapat berkaitan dengan ketidakhadiran, tidak mengerjakan ujian, atau kebijakan penilaian, dan tidak dapat dibedakan dari rendahnya penguasaan materi. Karena itu temuan penelitian ini harus dibaca sebagai asosiasi antara pola ketidakaktifan pada komponen pra-UAS dan nilai E, bukan sebagai bukti bahwa nilai rendah sebelum UAS menyebabkan atau memprediksi kegagalan akademik. Tujuh mahasiswa bernilai E yang berada di klaster utama, dengan komponen pra-UAS yang memadai, menunjukkan bahwa sebagian ketidakikutsertaan UAS tidak meninggalkan jejak pada skor sebelumnya.

Temuan ini sejalan dengan penelitian terdahulu pada beberapa aspek dan berbeda pada aspek lain. Pecuchova dan Drlik [16] membandingkan BIRCH, DBSCAN, dan GMM pada log LMS dan melaporkan BIRCH sebagai yang terbaik, sehingga tidak ada metode yang selalu menang; hal ini konsisten dengan temuan bahwa DBSCAN hanya setara dengan *Isolation Forest* di sini. Lauría [18] menunjukkan bahwa peringatan dini dapat dibingkai sebagai deteksi anomali ketika mahasiswa kesulitan jumlahnya sedikit, dan hasil kami memperkuat bingkai tersebut pada data skor mata kuliah. Eriyani dkk. [17] memakai DBSCAN untuk menyaring *outlier* sebelum K-Means tanpa analisis sensitivitas formal; penelitian ini menambahkan analisis sensitivitas, kestabilan, validasi eksternal, dan pembanding. Pengaruh *outlier* pada mutu klaster yang dilaporkan oleh Nowak-Brzezińska dan Gaibei [12] serta Sánchez Vinces dkk. [21] juga mendukung perlunya memperlakukan titik *noise* secara eksplisit.

DBSCAN setara dengan *Isolation Forest* dan hanya sedikit di atas aturan sederhana seperti rerata komponen di bawah 50. Hal ini menunjukkan bahwa pada data berukuran kecil dengan pola ketidakaktifan yang jelas, sebagian besar sinyal dapat ditangkap tanpa klasterisasi berbasis kepadatan. Keunggulan DBSCAN yang dapat dipertahankan adalah penentuan jumlah penanda secara otomatis dan keluaran berupa profil klaster yang dapat dijelaskan kepada dosen pengampu.

Generalisasi lintas semester yang terbatas (*F1-score* 0,385–0,400) menunjukkan bahwa klaster acuan dan parameter dari satu semester tidak dapat dipakai langsung pada semester lain. Pada data ini, penyebabnya mencakup perbedaan sebaran nilai tugas, perbedaan bobot penilaian, dan perbedaan perilaku mahasiswa antartahun. Konsekuensi praktisnya adalah bahwa parameter dan klaster acuan harus dikalibrasi ulang pada setiap semester, dan bahwa hasil penelitian ini tidak boleh dipakai sebagai model tetap untuk semester berikutnya.

Keterbatasan penelitian ini meliputi: (1) satu mata kuliah dan dua semester ganjil dengan jumlah data kecil, dengan hanya 90 vektor fitur unik dari 142 rekaman; (2) nilai E identik dengan tidak mengikuti UAS, sehingga "berisiko"

berarti berisiko memperoleh nilai E, bukan penguasaan materi yang rendah; (3) evaluasi bersifat retrospektif pada sampel yang sama, dan validasi lintas semester menunjukkan generalisasi yang terbatas; (4) skema bobot dan sebaran nilai antartahun berbeda, serta pada 2024/2025 dua pasang komponen tercatat identik; (5) lima mahasiswa muncul pada kedua tahun sehingga rekaman tidak sepenuhnya independen, meskipun analisis sensitivitas menunjukkan hasil serupa; (6) tidak ada data kehadiran, aktivitas LMS, atau demografi; (7) hasil bersifat asosiatif dan tidak boleh ditafsirkan sebagai hubungan sebab-akibat; dan (8) hasil peka terhadap penskalaan sehingga memerlukan kalibrasi ulang bila fitur atau kohort berubah.

3.9 Implikasi dan Pertimbangan Etis

Secara teoretis, penelitian ini menunjukkan bahwa pada data skor mata kuliah berskala kecil, struktur berbasis kepadatan dapat memisahkan pola ketidakaktifan total dan ketidakaktifan parsial yang tidak tampak pada nilai akhir, dan bahwa kinerja deteksi ditentukan oleh sifat kohort dan skema penilaian, bukan semata oleh pilihan algoritma. Temuan ini menambah bukti bahwa evaluasi *unsupervised learning* dalam *educational data mining* perlu menyertakan validasi lintas kohort dan pembanding sederhana, bukan hanya indeks validitas internal.

Secara praktis, daftar penanda dapat ditinjau oleh dosen pengampu dan koordinator program studi setelah UTS sebagai dasar untuk menghubungi dan mendampingi mahasiswa, bukan sebagai dasar keputusan akademik. Pada evaluasi retrospektif, dari 142 rekaman, 25 ditandai dan 84% di antaranya berakhir E. Dua pola yang terungkap, yaitu ketidakaktifan total dan ketidakaktifan parsial pada satu komponen, menyarankan bentuk tindak lanjut yang berbeda. Efektivitas pendampingan semacam ini belum diuji dalam penelitian ini dan menjadi agenda penelitian lanjutan, misalnya melalui rancangan kuasi-eksperimen dan evaluasi prospektif pada semester berikutnya.

Dari sisi etika, penanda risiko membawa konsekuensi bagi mahasiswa. Dari 25 rekaman yang ditandai, 4 (16%) tidak bernilai E, dan dari 28 mahasiswa bernilai E, 7 (25%) tidak ditandai, sehingga penanda dapat keliru ke dua arah. Penanda tidak boleh dicatat permanen pada rekam akademik atau dipakai sebagai dasar sanksi, aksesnya perlu dibatasi pada pihak yang berkepentingan, dan data harus dianonimkan dalam analisis dan pelaporan. Komunikasi kepada mahasiswa perlu bersifat mendukung dan tidak melabeli mahasiswa sebagai "berisiko gagal".

4. KESIMPULAN

Penelitian ini menjawab empat rumusan masalah. Pertama, DBSCAN pada lima komponen pra-UAS membentuk dua klaster (117 rekaman aktif konsisten dan 8 rekaman pasif total) serta 17 titik *noise* dengan pola ketidakaktifan parsial. Kedua, pada evaluasi retrospektif, rekaman di luar klaster utama menangkap 21 dari 28 mahasiswa bernilai E dengan *precision* 0,84, *recall* 0,75, *F1-score* 0,79, dan MCC 0,75, dengan catatan bahwa kenaikan *recall* dari penanda *noise* (0,464) berasal dari klaster pasif total yang mudah dikenali. Ketiga, hasil stabil terhadap *subsampling* dan berada pada wilayah parameter yang luas, tetapi kinerjanya setara dengan *Isolation Forest* dan hanya sedikit di atas aturan sederhana, serta peka terhadap penskalaan dan sifat kohort. Keempat, generalisasi lintas semester terbatas (*F1-score* 0,385–0,400).

DBSCAN karenanya merupakan pendekatan eksploratif yang mudah diinterpretasi untuk mengidentifikasi pola ketidakaktifan akademik yang berasosiasi dengan nilai E, dengan potensi penerapan sebagai alat bantu *triase* setelah UTS yang dikalibrasi ulang tiap semester. Penelitian ini tidak membuktikan sistem peringatan dini yang dapat digeneralisasikan lintas semester maupun lintas mata kuliah. Penelitian selanjutnya disarankan melakukan validasi prospektif, menambahkan data kehadiran dan log aktivitas pembelajaran, memperluas ke beberapa mata kuliah dan institusi, serta mengevaluasi efektivitas intervensi yang mengikuti penanda.

Ucapan Terima Kasih

Penulis menyampaikan apresiasi dan terima kasih yang mendalam kepada segenap jajaran sivitas akademika STMIK PPKIA Tarakanita Rahmawati atas dedikasi, bimbingan, serta kemudahan administratif yang senantiasa diberikan selama proses penyelesaian jurnal ini. Ungkapan terima kasih juga penulis haturkan kepada rekan-rekan seperjuangan atas dukungan moral, motivasi, serta saran yang membangun dalam penyelesaian jurnal ini.

DAFTAR PUSTAKA

[1] C. Romero and S. Ventura, “Educational data mining and learning analytics: An updated survey,” *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 3, May 2020, doi: 10.1002/widm.1355.

[2] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, “Educational data mining and learning analytics for 21st century higher education: A review and synthesis,” *Telemat. Informatics*, vol. 37, pp. 13–49, Apr. 2019, doi: 10.1016/j.tele.2019.01.007.

[3] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, “The current landscape of learning analytics in higher education,” *Comput. Human Behav.*, vol. 89, pp. 98–110, Dec. 2018, doi: 10.1016/j.chb.2018.07.027.

[4] E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: literature review and best practices,” *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, p. 3, Dec. 2020, doi: 10.1186/s41239-020-0177-7.

[5] A. Namoun and A. Alshanqiti, “Predicting Student Performance Using Data Mining and Learning

Analytics Techniques: A Systematic Literature Review," *Appl. Sci.*, vol. 11, no. 1, p. 237, Dec. 2020, doi: 10.3390/app11010237.

[6] M. Riestra-González, M. del P. Paule-Ruíz, and F. Ortin, "Massive LMS log data analysis for the early prediction of course-agnostic student performance," *Comput. Educ.*, vol. 163, p. 104108, Apr. 2021, doi: 10.1016/j.compedu.2020.104108.

[7] R. Vankayalapati, K. B. Ghutugade, R. Vannapuram, and B. P. S. Prasanna, "K-Means Algorithm for Clustering of Learners Performance Levels Using Machine Learning Techniques," *Rev. d'Intelligence Artif.*, vol. 35, no. 1, pp. 99–104, Feb. 2021, doi: 10.18280/ria.350112.

[8] I. Akbar, I. S. Samad, R. Rahmat, and S. Rosmiana, "Data Mining Analysis of K-means Algorithm and Decision Tree for Early Detection of Students at Risk of Dropping Out," *J. Informatics Inf. Syst. Softw. Eng. Appl.*, vol. 7, no. 2, pp. 148–162, May 2025, doi: 10.20895/inista.v7i2.1630.

[9] S. Križanić, "Educational data mining using cluster analysis and decision tree technique: A case study," *Int. J. Eng. Bus. Manag.*, vol. 12, p. 184797902090867, Jan. 2020, doi: 10.1177/1847979020908675.

[10] N. H. M. M. Shrifan, M. F. Akbar, and N. A. M. Isa, "An adaptive outlier removal aided k-means clustering algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6365–6376, Sep. 2022, doi: 10.1016/j.jksuci.2021.07.003.

[11] I. K. Khan *et al.*, "Addressing limitations of the K-means clustering algorithm: outliers, non-spherical data, and optimal cluster selection," *AIMS Math.*, vol. 9, no. 9, pp. 25070–25097, 2024, doi: 10.3934/math.20241222.

[12] A. Nowak-Brzezińska and I. Gaibei, "How the Outliers Influence the Quality of Clustering?," *Entropy*, vol. 24, no. 7, p. 917, Jun. 2022, doi: 10.3390/e24070917.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Elsevier, 1996, pp. 226–231. doi: 10.5555/3001460.3001507.

[14] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, Sep. 2017, doi: 10.1145/3068335.

[15] L. Shu and G. Li, "Application of improved clustering algorithm in mixed teaching of modern educational technology," *Discov. Artif. Intell.*, vol. 5, no. 1, p. 195, Aug. 2025, doi: 10.1007/s44163-025-00393-8.

[16] J. Pecuchova and M. Drlik, "Enhancing the Early Student Dropout Prediction Model Through Clustering Analysis of Students' Digital Traces," *IEEE Access*, vol. 12, pp. 159336–159367, 2024, doi: 10.1109/ACCESS.2024.3486762.

[17] N. S. Eriyani, D. F. Surianto, Nurhidayat, F. N. D. Salam, and Muliadi, "Outlier-Aware Clustering for Mapping Computer Science Students' Career Readiness: A Hybrid DBSCAN–K-Means Approach," *Int. J. Electron. Commun. Syst.*, vol. 6, no. 1, pp. 1–17, Jun. 2026, doi: 10.24042/ijecs.v6i1.28865.

[18] E. Lauría, "Framing Early Alert of Struggling Students as an Anomaly Detection Problem: An Exploration," in *Proceedings of the 13th International Conference on Computer Supported Education*, SCITEPRESS - Science and Technology Publications, 2021, pp. 26–35. doi: 10.5220/0010471900260035.

[19] G. O. Campos *et al.*, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Min. Knowl. Discov.*, vol. 30, no. 4, pp. 891–927, Jul. 2016, doi: 10.1007/s10618-015-0444-8.

[20] M. Goldstein and S. Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data," *PLoS One*, vol. 11, no. 4, p. e0152173, Apr. 2016, doi: 10.1371/journal.pone.0152173.

[21] B. V. Sánchez Vinces, E. Schubert, A. Zimek, and R. L. F. Cordeiro, "A comparative evaluation of clustering-based outlier detection," *Data Min. Knowl. Discov.*, vol. 39, no. 2, p. 13, Mar. 2025, doi: 10.1007/s10618-024-01086-z.

[22] D. Moulavi, P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander, "Density-Based Clustering Validation," in *Proceedings of the 2014 SIAM International Conference on Data Mining*, Philadelphia, PA: Society for Industrial and Applied Mathematics, Apr. 2014, pp. 839–847. doi: 10.1137/1.9781611973440.96.

[23] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[24] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1

score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020, doi: 10.1186/s12864-019-6413-7.

[25] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, New York, NY, USA: ACM, May 2000, pp. 93–104. doi: 10.1145/342009.335388.

[26] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, Dec. 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

[27] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *J. Open Source Softw.*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.

[28] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html