

ANALISIS RISIKO *DROP OUT* MAHASISWA MENGGUNAKAN METODE *HYBRID K-MEANS* DAN *CART*

Ridwan Afriansah^{*1}, Ahmad Syarif Sukri¹, Mansur¹, Muhammad Nadzirin Anshari Nur¹, Hasmina Tari Mokui¹, Tambi¹

¹Program Studi Magister Manajemen Rekayasa, Fakultas Teknik, Universitas Halu Oleo, Kendari, Indonesia

Penulis korespondensi: Ridwan Afriansah (ridwan.afriansah@gmail.com)

Drop out mahasiswa merupakan persoalan penting dalam pendidikan tinggi karena berdampak pada keberlanjutan studi, efisiensi sumber daya, dan capaian institusi. Penelitian ini bertujuan membangun model analisis risiko drop out mahasiswa menggunakan pendekatan hybrid K-Means dan Classification and Regression Tree (CART) dalam kerangka two-stage modeling. Data bersumber dari Sistem Informasi Akademik STMIK X pada tahun akademik 2015/2016-2018/2019. Setelah pembersihan dan seleksi diperoleh 631 rekaman mahasiswa, terdiri atas 119 mahasiswa drop out dan 512 mahasiswa aktif/lulus. Tahap pertama menggunakan K-Means untuk menemukan struktur segmentasi mahasiswa, sedangkan tahap kedua menggunakan CART dengan label klaster sebagai fitur tambahan. Hasil K-Means menunjukkan K = 10 sebagai jumlah klaster statistik terbaik dengan silhouette score 0,6330; sepuluh klaster tersebut kemudian diinterpretasikan menjadi tiga kelompok risiko, yaitu rendah, sedang, dan tinggi. Pada data uji, model hybrid menghasilkan accuracy 92,25%, precision 60,87%, recall 87,50%, dan F1-score 71,79%, lebih baik daripada CART tunggal dengan accuracy 88,03%, precision 44,44%, recall 25,00%, dan F1-score 32,00%. Validasi menggunakan Orange Data Mining juga menunjukkan arah peningkatan yang konsisten dengan accuracy model hybrid sebesar 92,60%. Variabel kehadiran menjadi prediktor paling dominan dengan feature importance 0,71. Temuan ini menunjukkan bahwa segmentasi K-Means dapat memperkaya representasi data sebelum klasifikasi CART dan meningkatkan kemampuan deteksi dini mahasiswa berisiko drop out.

Kata Kunci - CART, Drop Out Mahasiswa, Educational Data Mining, K-Means Clustering, Two-Stage Modeling.

I. PENDAHULUAN

Drop out mahasiswa merupakan salah satu indikator kritis dalam keberhasilan penyelenggaraan pendidikan tinggi. Mahasiswa yang menghentikan studi sebelum menyelesaikan program akademik kehilangan investasi waktu dan biaya, sedangkan institusi menghadapi penurunan efisiensi pemanfaatan sumber daya, retensi, dan capaian kelulusan. Literatur internasional

menunjukkan bahwa fenomena ini tidak dapat dijelaskan hanya oleh satu faktor, karena performa akademik, kondisi sosial-ekonomi, keterlibatan mahasiswa, dan dukungan institusional saling berinteraksi dalam membentuk keputusan untuk bertahan atau meninggalkan studi [1]–[4].

Perkembangan sistem informasi akademik memungkinkan perguruan tinggi menyimpan data mahasiswa dalam jumlah besar, seperti indeks prestasi, kehadiran, kemajuan studi, status pembayaran, dan karakteristik demografis. Data tersebut dapat dimanfaatkan melalui *Educational Data Mining* (EDM) untuk menemukan pola yang tidak mudah teridentifikasi melalui pemantauan administratif konvensional. Kajian EDM menunjukkan bahwa klasifikasi dan klusterisasi merupakan dua teknik yang banyak digunakan untuk prediksi performa dan *dropout* karena mampu mengubah data historis menjadi informasi yang mendukung keputusan akademik [5].

Pada tugas prediksi *dropout*, algoritma klasifikasi seperti *decision tree*, *random forest*, *support vector machine*, dan *boosting* telah digunakan dengan tingkat kinerja yang beragam. Oqaidi et al. menekankan pentingnya pemilihan fitur, algoritma, dan metrik evaluasi dalam membangun model prediksi *dropout* [6], sedangkan Villar dan de Andrade menunjukkan bahwa algoritma *supervised learning* dapat memberikan kemampuan diskriminasi yang baik pada data keberhasilan akademik dan *dropout* [7]. Namun, penggunaan model klasifikasi tunggal berpotensi kurang optimal ketika populasi mahasiswa memiliki struktur yang sangat heterogen.

Salah satu strategi untuk menangkap heterogenitas tersebut adalah menggunakan pemodelan bertahap. Pada pendekatan ini, proses *unsupervised learning* digunakan terlebih dahulu untuk menemukan segmen mahasiswa, kemudian informasi segmentasi dimanfaatkan pada tahap *supervised learning*. Pendekatan berlapis dan berbasis klaster telah dilaporkan dapat meningkatkan skalabilitas maupun kemampuan prediksi pada konteks pendidikan [8],[9]. Penelitian *K-Means* untuk pengelompokan risiko *dropout* juga menunjukkan bahwa klusterisasi mampu membentuk kelompok risiko yang mudah diinterpretasikan, tetapi masih memerlukan perluasan dengan algoritma prediktif agar hasil segmentasi dapat

ditransformasikan menjadi keputusan pada tingkat individu [10].

Model *hybrid* menjadi relevan karena menggabungkan keunggulan eksploratif dari klasterisasi dan kemampuan prediktif dari klasifikasi. Studi mengenai metode *hybrid* dan model fusion menunjukkan bahwa penggabungan beberapa pendekatan dapat meningkatkan *recall*, *precision*, dan akurasi dibandingkan model tunggal pada kasus *dropout* [11]. Rabelo dan Zarate juga menunjukkan bahwa kombinasi beberapa model dapat meningkatkan kemampuan identifikasi mahasiswa yang berpotensi keluar dari pendidikan tinggi [12]. Meskipun demikian, masih terdapat ruang untuk mengembangkan model yang tetap mudah dijelaskan oleh pengelola akademik. Dalam konteks tersebut, CART memiliki kelebihan karena menghasilkan pohon keputusan dan aturan if-then yang transparan, sehingga lebih mudah diterjemahkan menjadi tindakan akademik daripada model yang bersifat *black-box* [13].

Penelitian ini dilaksanakan pada data mahasiswa STMIK X di Kota Kendari. Data akademik, demografis, dan sosial-ekonomi telah tersedia dalam SIAKAD, tetapi belum dimanfaatkan sebagai pipeline analitik terpadu untuk deteksi risiko *dropout*. Karakteristik mahasiswa yang berasal dari wilayah perkotaan maupun luar kota, perbedaan capaian akademik, variasi kehadiran, dan kondisi pembayaran kuliah menimbulkan heterogenitas yang perlu ditangkap sebelum proses klasifikasi dilakukan. Faktor akademik dan sosial-ekonomi tersebut juga konsisten dengan hasil telaah mengenai determinan *dropout* pada pendidikan tinggi [2], [3], [14].

Kebaruan penelitian ini terletak pada penerapan *two-stage modeling* yang mengintegrasikan *K-Means* dan *CART* dalam satu alur analisis. *K-Means* tidak digunakan hanya untuk menghasilkan tiga kelompok risiko secara langsung, melainkan terlebih dahulu menentukan jumlah klaster yang paling baik secara statistik. Klaster yang terbentuk kemudian diinterpretasikan menjadi kelompok risiko rendah, sedang, dan tinggi. Label klaster tersebut digunakan sebagai fitur tambahan pada *CART* agar klasifikasi individual memperoleh informasi kontekstual mengenai segmen mahasiswa. Pendekatan serupa pada ensemble berbasis klaster menunjukkan bahwa transformasi representasi data dapat membantu klasifikasi ketika data memiliki tipe dan pola yang beragam [15].

II. LANDASAN TEORI

A. Drop Out Mahasiswa dan Educational Data Mining

Drop out pada pendidikan tinggi merujuk pada kondisi mahasiswa meninggalkan program studi sebelum menyelesaikan pendidikan. Penelitian ini menggunakan definisi operasional berbasis catatan akademik: mahasiswa dikategorikan *drop out* apabila tidak melakukan registrasi ulang selama dua semester berturut-turut tanpa keterangan resmi, tidak sedang cuti akademik, dan tidak berada dalam proses mutasi atau pindah program studi. Pemisahan status tersebut penting agar label target benar-benar merepresentasikan penghentian studi, bukan ketidakaktifan administratif sementara.

Faktor penyebab *dropout* bersifat multidimensional. Review sosio-ekonomi menempatkan latar belakang

ekonomi, kondisi keluarga, dan performa akademik sebagai faktor yang berulang dalam berbagai konteks [2]. Kajian lain menunjukkan kontribusi kehadiran, motivasi, integrasi sosial, serta kepuasan mahasiswa terhadap keberlanjutan studi [3], [4], [14]. Karena variabel-variabel tersebut tidak selalu tersedia dalam data administratif, EDM menyediakan pendekatan pragmatis untuk menggunakan informasi yang tersedia di SIAKAD sebagai indikator risiko yang dapat dipantau secara periodik [5].

B. K-Means Clustering

K-Means merupakan algoritma unsupervised learning yang membagi objek ke dalam K klaster berdasarkan kemiripan jarak terhadap centroid. Tujuan optimasinya adalah meminimalkan variasi di dalam klaster dan memaksimalkan perbedaan antarklaster. Dalam data mahasiswa, K-Means dapat digunakan untuk menangkap kombinasi profil akademik, demografis, dan sosial-ekonomi tanpa terlebih dahulu menggunakan label *dropout*. Penelitian terdahulu menunjukkan bahwa K-Means dapat membentuk segmentasi risiko mahasiswa yang interpretable dan menjadi dasar intervensi [9], [10].

C. CART dan Two-Stage Modeling

Classification and Regression Tree (CART) membangun pohon keputusan melalui pemisahan rekursif pada variabel yang memberikan penurunan *impurity* terbesar. Untuk klasifikasi biner, Gini *impurity* dapat digunakan sebagai kriteria pemisahan. Struktur pohon terdiri atas *root node*, internal node, cabang, dan leaf node yang merepresentasikan kelas prediksi. Keunggulan CART adalah kemampuannya menghasilkan aturan keputusan yang dapat dibaca secara langsung oleh pengguna non-teknis serta *feature importance* berdasarkan kontribusi suatu variabel terhadap penurunan *impurity* [16], [17].

Keterbatasan CART adalah sensitivitas terhadap perubahan data dan risiko overfitting apabila pohon tumbuh terlalu dalam. Karena itu, *pruning* dan *hyperparameter tuning* diperlukan untuk memperoleh pohon yang cukup sederhana namun tetap memiliki daya prediksi. Pada penelitian pendidikan, decision tree telah digunakan untuk prediksi *dropout* dan dapat memberikan keseimbangan antara akurasi dan interpretabilitas [6], [13].

Dalam *two-stage modeling*, output tahap pertama menjadi input atau fitur tambahan pada tahap kedua. Pada penelitian ini, label klaster K-Means digunakan untuk memperkaya representasi setiap mahasiswa sebelum CART dibangun. Dengan cara tersebut, CART tidak hanya membaca nilai fitur individual, tetapi juga memperoleh informasi tentang posisi mahasiswa di dalam struktur populasi. Logika ini sejalan dengan penelitian cluster-based prediction dan ensemble berbasis klaster yang menunjukkan bahwa segmentasi dapat membantu klasifikasi pada data yang heterogen [9], [15].

III. METODE PENELITIAN

A. Data Penelitian dan Definisi Status

Penelitian ini merupakan penelitian kuantitatif terapan dengan desain eksperimen komparatif. Kerangka proses mengikuti *Knowledge Discovery in Databases* (KDD), yaitu seleksi data, *preprocessing*, transformasi, data mining, evaluasi, dan interpretasi. Model baseline menggunakan CART tunggal, sedangkan model utama menggunakan hybrid K-Means-CART. Unit analisis adalah mahasiswa program sarjana yang tercatat pada SIAKAD STMIK X di Kendari.

Periode data mencakup tahun akademik 2015/2016 sampai 2018/2019. Setelah proses seleksi dan pembersihan, dataset akhir berisi 631 rekaman. Mahasiswa dengan cuti resmi dan mahasiswa yang melakukan mutasi/pindah dikeluarkan dari dataset karena tidak merepresentasikan dropout. Status target dikodekan biner, yaitu 1 untuk dropout dan 0 untuk aktif/lulus. Distribusi kelas disajikan pada Tabel 1.

Tabel 1. Distribusi Kelas Dataset Penelitian

Status	Jumlah	Persentase
Drop out (kelas 1)	119	19%
Aktif/Lulus (kelas 0)	512	81%
Total	631	100%

Sumber: Hasil pengolahan data SIAKAD STMIK X (2026)

Komposisi 119 mahasiswa dropout dan 512 mahasiswa aktif/lulus menunjukkan rasio kelas sekitar 1:4,30. Distribusi ini menjadi alasan utama untuk memberikan perhatian khusus pada recall dan F1-score, karena accuracy yang tinggi dapat tetap terjadi apabila model terlalu dominan memprediksi kelas mayoritas.

B. Pra-Pemrosesan dan Pembentukan Fitur

Pra-pemrosesan mencakup pemeriksaan duplikasi, konsistensi format, penanganan nilai hilang, normalisasi, dan encoding. Variabel rasio dipetakan ke rentang 0-1 menggunakan Min-Max Scaling agar perhitungan jarak Euclidean pada K-Means tidak didominasi oleh variabel

dengan skala lebih besar. Variabel kategorikal dikonversi ke representasi numerik. Setelah preprocessing dan feature engineering, dataset siap pemodelan terdiri atas 631 observasi dengan 11 fitur yang mewakili aspek akademik, demografis, dan sosial-ekonomi.

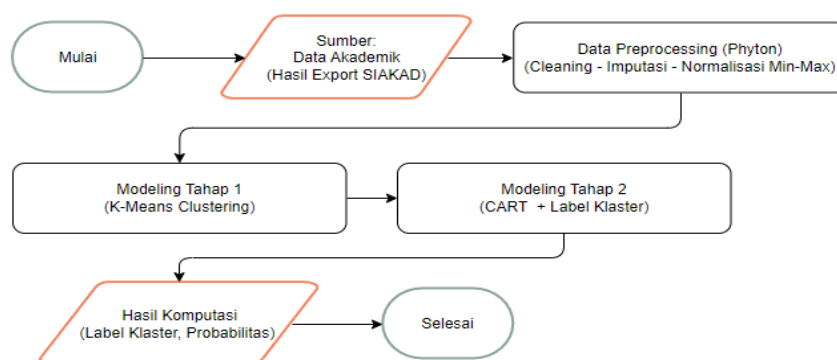
C. Tahap 1: K-Means Clustering

K-Means dijalankan menggunakan inisialisasi K-Means++ pada rentang $K = 2$ sampai $K = 14$. Setiap solusi dinilai menggunakan WCSS dan *Silhouette Score*. Jumlah kluster yang dipilih adalah nilai K yang memberikan kombinasi titik elbow yang jelas dan kualitas silhouette terbaik. Setelah kluster statistik diperoleh, profil rata-rata setiap kluster dianalisis berdasarkan performa akademik, kehadiran, progres SKS, jumlah mata kuliah tidak lulus, dan status UKT. Kluster-kluster yang memiliki pola serupa kemudian diagregasikan ke dalam tiga label risiko operasional: rendah, sedang, dan tinggi.

D. Tahap 2: CART dan Validasi

Pada model *baseline*, CART dibangun menggunakan fitur asli tanpa label kluster. Pada model hybrid, label kluster hasil K-Means ditambahkan sebagai fitur sebelum pembangunan CART. *Hyperparameter* dioptimalkan melalui *grid search* dan *cross-validation* pada data training, serta *pruning* digunakan untuk mengurangi *overfitting*. Model memanfaatkan Gini Index sebagai kriteria *split*. Skema validasi utama menggunakan temporal *split*: data tahun akademik 2015/2016-2017/2018 sebagai data training dan 2018/2019 sebagai data testing, sehingga model diuji pada periode yang secara waktu belum pernah digunakan untuk pelatihan.

Kinerja dievaluasi menggunakan accuracy, *precision*, *recall*, dan *F1-score*. Seluruh eksperimen utama dijalankan menggunakan Python dan scikit-learn. Untuk validasi tambahan, skema pemodelan direplikasi menggunakan Orange Data Mining dengan dataset dan pembagian waktu yang sama. Perbandingan lintas platform digunakan untuk memastikan bahwa arah peningkatan kinerja tidak hanya muncul pada satu lingkungan implementasi.



Gambar 1. Arsitektur Pipeline Komputasi Model Hybrid K-Means-CART

IV. HASIL PENELITIAN

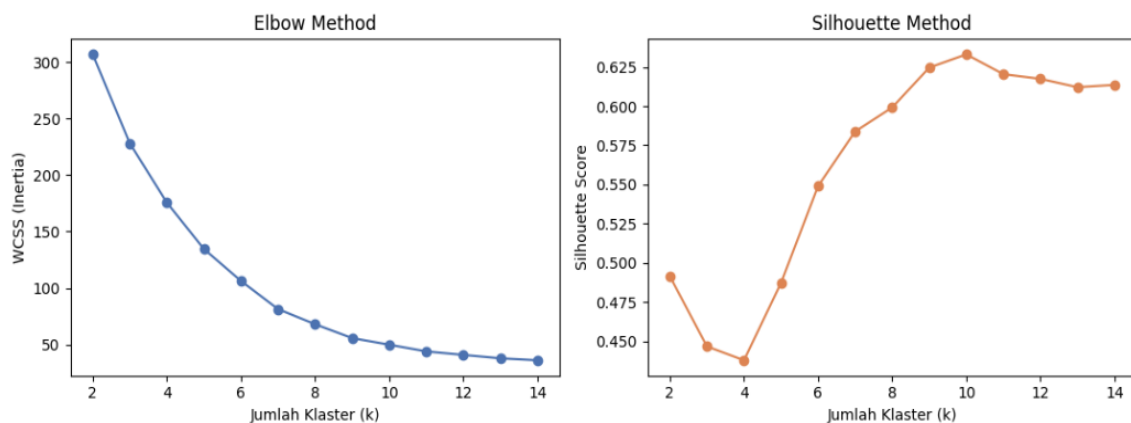
A. Deskripsi Dataset

Rata-rata IPK mahasiswa pada dataset sebesar 2,13 dengan median 3,02, sedangkan rata-rata IPS sebesar 1,98 dengan median 2,75. Jumlah mata kuliah tidak lulus rata-rata sebanyak 3 mata kuliah dengan nilai maksimum 31.

Tingkat kehadiran rata-rata sebesar 49,46% dengan median 68,75%. Perbedaan antara mean dan median pada IPK, IPS, dan kehadiran menunjukkan adanya kelompok mahasiswa dengan capaian sangat rendah yang menekan nilai rata-rata keseluruhan.

Komposisi jenis kelamin didominasi mahasiswa laki-laki sebanyak 410 orang (65%), sedangkan mahasiswa

perempuan sebanyak 221 orang (35%). Asal daerah relatif seimbang, yaitu 312 mahasiswa urban (49%) dan 319 mahasiswa rural (51%). Pada status pembayaran UKT, 410 mahasiswa (64%) tercatat lunas atau tepat waktu dan 221 mahasiswa (36%) menunggak. Variasi ini menunjukkan bahwa dataset memiliki heterogenitas yang cukup untuk eksplorasi klaster pada dimensi akademik dan sosial-ekonomi.



Gambar 2. *Elbow Method* dan *Silhouette Score* untuk Penentuan Jumlah Klaster

Pemilihan $K = 10$ tidak berarti terdapat sepuluh tingkat risiko. Nilai K tersebut digunakan untuk mempertahankan struktur sub-populasi yang secara statistik lebih homogen. Jika K dipaksakan menjadi 3, *silhouette score* hanya

B. Hasil Segmentasi K-Means

Hasil Elbow Method menunjukkan penurunan WCSS mulai melambat secara nyata pada sekitar $K = 10$. *Silhouette Score* meningkat dari 0,4914 pada $K = 2$ menjadi 0,6330 pada $K = 10$, kemudian menurun menjadi 0,6204 pada $K = 11$ dan 0,6174 pada $K = 12$. Nilai 0,6330 merupakan skor tertinggi pada rentang yang diuji, sehingga $K = 10$ dipilih sebagai jumlah klaster statistik terbaik.

0,4466, jauh lebih rendah daripada 0,6330 pada $K = 10$. Karena itu, sepuluh klaster terlebih dahulu dipertahankan untuk analisis profil, kemudian diagregasikan ke tiga kategori risiko berdasarkan kesamaan karakteristik.

Tabel 2. Interpretasi Klaster K-Means menjadi Tiga Kelompok Risiko

Kelompok Risiko	Klaster	Karakteristik Utama (skala normalisasi)	Implikasi
Rendah	1, 3, 6, 7	IPK sekitar 0,85; IPS dan kehadiran 0,82-0,89; MK tidak lulus 0,02-0,03; status UKT 0,00.	Monitoring rutin dan pemeliharaan engagement.
Sedang	9, 10	IPK 0,54-0,55; IPS dan kehadiran 0,23-0,29; MK tidak lulus 0,31-0,43.	Pendampingan akademik dan pemantauan progres.
Tinggi	2, 4, 5, 8	IPK 0,02-0,24; IPS dan kehadiran 0,00; status UKT 1,00; MK tidak lulus relatif lebih tinggi.	Prioritas intervensi dini akademik dan finansial.

Kelompok risiko rendah terdiri atas klaster 1, 3, 6, dan 7. Kelompok ini memiliki performa akademik relatif tinggi, tingkat kehadiran yang baik, jumlah mata kuliah tidak lulus yang sangat rendah, dan tidak menunjukkan indikasi tunggakan UKT. Sebaliknya, klaster 2, 4, 5, dan 8 membentuk kelompok risiko tinggi karena kombinasi capaian akademik yang sangat rendah, kehadiran yang mendekati nol pada skala normalisasi, dan indikasi tunggakan UKT. Klaster 9 dan 10 berada di antara kedua kelompok tersebut dan dikategorikan risiko sedang.

C. Hasil Klasifikasi CART

Model CART tunggal menggunakan seluruh variabel tanpa label klaster. Setelah tuning, pohon baseline menggunakan *max_depth* = 6, *min_samples_split* = 20, *min_samples_leaf* = 1, dan *Gini Index*. Pada model tersebut, status UKT menjadi *root node*, menunjukkan bahwa kondisi pembayaran memiliki daya diskriminasi tinggi sebelum informasi segmentasi dimasukkan. Beberapa aturan baseline mengkombinasikan status UKT, IPK, jumlah mata kuliah tidak lulus, asal daerah, semester aktif, SKS kumulatif, dan jenis kelamin.

Pada model hybrid, label klaster K-Means ditambahkan sebelum CART dilatih. Struktur pohon berubah secara

substantif: *kehadiran_normalized* menjadi *root node* dengan Gini impurity 0,368 pada 489 sampel training. Kondisi kehadiran lebih dari 0,02 dan semester aktif tidak lebih dari 0,7083 menghasilkan kelompok dominan yang diprediksi tidak dropout. Sebaliknya, kombinasi kehadiran sangat rendah, jumlah mata kuliah tidak lulus yang lebih tinggi, serta SKS kumulatif rendah membentuk beberapa jalur keputusan menuju kelas dropout.

D. Confusion Matrix dan Kinerja Model

Tabel 3. Confusion Matrix pada Data Uji

Model	TN	FP	FN	TP
CART Tunggal	121	5	12	4
Hybrid K-Means-CART	117	9	2	14

Confusion matrix menunjukkan perbedaan utama pada kemampuan mendeteksi kelas dropout. CART tunggal hanya mendeteksi 4 dari 16 mahasiswa dropout pada data uji, sehingga menghasilkan 12 *false negative*. Model *hybrid* mendeteksi 14 dari 16 mahasiswa dropout dan hanya menghasilkan 2 *false negative*. Meskipun *false positive* meningkat dari 5 menjadi 9, penurunan *false negative* memiliki nilai operasional yang lebih besar

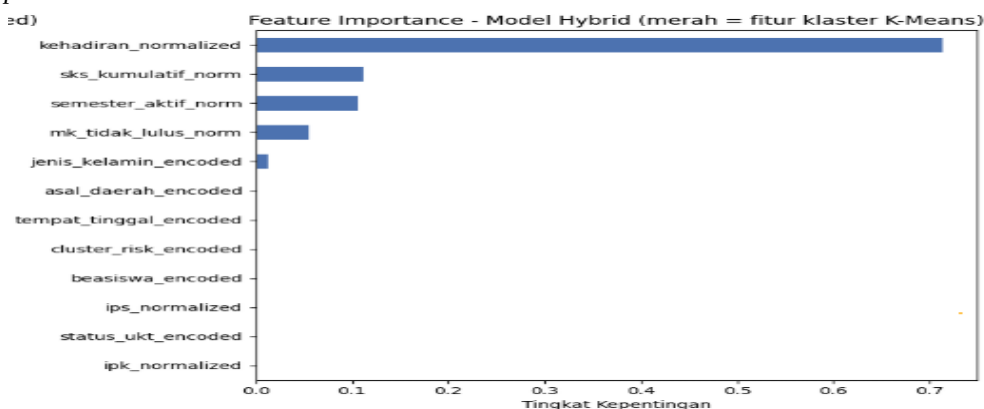
dalam sistem peringatan dini karena mahasiswa berisiko yang terlewat jauh berkurang.

Tabel 4. Perbandingan Kinerja CART Tunggal dan Hybrid K-Means-CART

Metrik	CART Tunggal	Hybrid	Peningkatan
Accuracy	88,03%	92,25%	4,22 poin
Precision	44,44%	60,87%	16,42 poin
Recall	25,00%	87,50%	62,50 poin
F1-Score	32,00%	71,79%	39,79 poin

Model *hybrid* meningkatkan seluruh metrik utama. Accuracy naik dari 88,03% menjadi 92,25%. Precision meningkat dari 44,44% menjadi 60,87%, sedangkan recall meningkat paling besar dari 25,00% menjadi 87,50%. F1-score meningkat dari 32,00% menjadi 71,79%. Peningkatan recall sebesar 62,50 poin persentase menunjukkan bahwa penambahan informasi segmentasi K-Means terutama memperkuat kemampuan model dalam menemukan mahasiswa yang benar-benar dropout, bukan

F. Feature Importance



Gambar 3. Feature Importance Model Hybrid K-Means-CART

Tabel 6. Peringkat Feature Importance Utama

Rank	Variabel	Hybrid	CART Tunggal
1	Kehadiran	0,71	0,00
2	SKS Kumulatif	0,11	0,06
3	Semester Aktif	0,10	0,13
4	MK Tidak Lulus	0,05	0,05

Kehadiran menjadi variabel paling dominan pada model hybrid dengan *feature importance* 0,71, jauh di atas SKS kumulatif 0,11, semester aktif 0,10, dan jumlah mata kuliah tidak lulus 0,05. Pada CART tunggal, importance kehadiran tercatat 0,00, sedangkan model lebih banyak menggunakan status UKT sebagai pemisah utama. Pergeseran ini menunjukkan bahwa pengelompokan awal membantu memunculkan kontribusi kehadiran yang sebelumnya tertutupi oleh struktur global data.

G. Pembahasan

Hasil penelitian menunjukkan bahwa manfaat utama hybrid K-Means-CART bukan sekadar kenaikan accuracy, tetapi perubahan kualitas deteksi terhadap kelas dropout. Pada baseline, model sangat baik mengenali

hanya meningkatkan proporsi prediksi benar secara keseluruhan.

E. Validasi Antar-Platform

Tabel 5. Perbandingan Akurasi pada Dua Platform

Platform	CART Tunggal	Hybrid	Peningkatan
Python/scikit-learn	88,03%	92,25%	4,22 poin
Orange Data Mining	89,40%	92,60%	3,20 poin

Replikasi menggunakan Orange Data Mining menghasilkan accuracy 89,40% untuk CART tunggal dan 92,60% untuk model hybrid. Meskipun nilai absolutnya sedikit berbeda dari Python, arah hasil tetap sama: model hybrid lebih baik daripada baseline. Konsistensi ini memperkuat interpretasi bahwa peningkatan bukan sekadar artefak dari satu implementasi pustaka. Pendekatan validasi lintas platform ini sejalan dengan studi komparatif yang menggunakan Orange untuk menguji algoritma klasifikasi pada data mahasiswa [18].

mahasiswa non-dropout, tetapi gagal menangkap banyak mahasiswa yang benar-benar dropout. Setelah label kluster ditambahkan, *false negative* turun dari 12 menjadi 2. Dalam konteks *early warning system*, perubahan ini lebih penting daripada sekadar beberapa poin peningkatan accuracy karena tujuan intervensi adalah menemukan sebanyak mungkin mahasiswa berisiko sebelum terlambat.

Peningkatan tersebut mendukung argumen bahwa populasi mahasiswa bersifat heterogen. Model klasifikasi yang langsung bekerja pada seluruh populasi dapat menggunakan satu fitur dominan yang secara global paling kuat, seperti status UKT, tetapi melewatkan pola lokal pada sub-populasi tertentu. K-Means memecah populasi menjadi segmen dengan karakteristik lebih homogen. Informasi segmentasi ini kemudian memperkaya representasi pada CART sehingga model dapat memanfaatkan indikator seperti kehadiran dan progres SKS secara lebih efektif. Temuan ini konsisten dengan penelitian cluster-based prediction, ensemble berbasis kluster, serta studi hybrid yang menunjukkan peningkatan kinerja dibandingkan model tunggal.

Pemilihan $K = 10$ juga memberikan implikasi metodologis. Jika tiga kategori risiko dipaksakan sejak awal, kualitas kluster berdasarkan *silhouette score*

menjadi lebih rendah. Dengan memisahkan proses pembentukan klaster statistik dari proses interpretasi risiko, penelitian ini mempertahankan struktur data yang lebih kuat sekaligus tetap menghasilkan tiga kategori yang mudah digunakan oleh institusi. Strategi tersebut membedakan jumlah klaster sebagai hasil optimasi komputasional dari jumlah label risiko sebagai keputusan interpretatif dan operasional.

Feature importance memperlihatkan bahwa kehadiran merupakan indikator paling dominan pada model hybrid. Temuan tersebut logis dari sudut pandang pengelolaan akademik karena kehadiran merupakan indikator perilaku yang dapat diamati lebih awal dan berulang sepanjang semester. Apabila dikombinasikan dengan SKS kumulatif, semester aktif, serta jumlah mata kuliah tidak lulus, pengelola dapat membentuk daftar prioritas intervensi. Literatur juga menunjukkan bahwa faktor akademik dan engagement memiliki hubungan yang konsisten dengan risiko *dropout*.

Status UKT tetap penting meskipun tidak lagi menjadi *root node* pada model hybrid. Klaster risiko tinggi justru memperlihatkan kombinasi performa akademik rendah dan tunggakan UKT. Dengan demikian, hasil model tidak seharusnya diterjemahkan menjadi intervensi akademik tunggal.

V. KESIMPULAN DAN SARAN

Penelitian ini berhasil membangun model analisis risiko *drop out* mahasiswa dengan pendekatan *two-stage hybrid K-Means-CART*. *K-Means* menghasilkan sepuluh klaster statistik dengan *silhouette score* tertinggi 0,6330 pada $K = 10$. Sepuluh klaster tersebut kemudian diinterpretasikan menjadi tiga kelompok risiko, yaitu rendah, sedang, dan tinggi. Strategi ini memungkinkan segmentasi mempertahankan kualitas struktur data sekaligus menghasilkan kategori risiko yang operasional bagi pengelola perguruan tinggi.

Model *hybrid* menunjukkan kinerja lebih baik daripada CART tunggal. Accuracy meningkat dari 88,03% menjadi 92,25%, precision dari 44,44% menjadi 60,87%, recall dari 25,00% menjadi 87,50%, dan F1-score dari 32,00% menjadi 71,79%. *Confusion matrix* menunjukkan *false negative* turun dari 12 menjadi 2. Validasi dengan Orange Data Mining juga menghasilkan arah yang sama, yaitu accuracy 89,40% pada CART tunggal dan 92,60% pada model *hybrid*. Kehadiran menjadi variabel paling dominan pada model hybrid dengan *feature importance* 0,71, diikuti SKS kumulatif, semester aktif, dan jumlah mata kuliah tidak lulus.

Berdasarkan temuan tersebut, institusi disarankan mengintegrasikan keluaran klaster dan aturan CART ke dalam mekanisme pemantauan akademik secara bertahap. Model sebaiknya digunakan sebagai alat bantu early warning, bukan sebagai dasar keputusan otomatis tentang status mahasiswa. Penelitian berikutnya perlu melibatkan lebih dari satu perguruan tinggi, menambahkan variabel psikologis dan engagement mahasiswa, mengevaluasi fairness model pada kelompok demografis, serta melakukan retraining berkala agar model tetap relevan terhadap perubahan karakteristik mahasiswa dan kebijakan institusi.

DAFTAR PUSTAKA

- [1] OECD, "Education at a Glance 2019: OECD Indicators," OECD Publishing, Paris, 2019. doi: 10.1787/f8d7880d-en.
- [2] C. Aina, E. Baici, G. Casalone, and F. Pastore, "The determinants of university dropout: A review of the socio-economic literature," *Socioecon. Plann. Sci.*, vol. 79, p. 101102, 2022, doi: 10.1016/j.seps.2021.101102.
- [3] Nurmalitasari, Z. Awang Long, and M. F. Mohd Noor, "Factors Influencing Dropout Students in Higher Education," *Educ. Res. Int.*, vol. 2023, pp. 1–13, 2023, doi: 10.1155/2023/7704142.
- [4] M. Esteban, A. B. Bernardo, E. Blanco, and P. Oserin, "An Analysis of Student Satisfaction and its Relationship with Academic and Social Factors in University Dropout," *Int. J. Educ. Psychol.*, vol. 13, no. 3, pp. 219–239, 2024, doi: 10.17583/ijep.14268.
- [5] M. P. Colpo, T. Thompsen Primo, M. S. de Aguiar, and C. Cechinel, "Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges," *Rev. Bras. Inform. na Educ.*, vol. 32, pp. 220–256, 2024, doi: 10.5753/rbie.2024.3559.
- [6] K. Oqaidi, S. Aouhassi, and K. Mansouri, "Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 18, pp. 103–117, 2022, doi: 10.3991/ijet.v17i18.25567.
- [7] A. Villar and C. R. V de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discov. Artif. Intell.*, vol. 4, no. 1, p. 2, 2024, doi: 10.1007/s44163-023-00079-z.
- [8] J. Niyogisubizo, L. Liao, E. Nziyumba, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Comput. Educ. Artif. Intell.*, vol. 3, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.
- [9] L.-V. Poellhuber, B. Poellhuber, M. Desmarais, C. Leger, N. Roy, and M.-C. Vu, "Cluster-Based Performance of Student Dropout Prediction as a Solution for Large Scale Models in a Moodle LMS," *LAK23: 13th International Learning Analytics and Knowledge Conference*. pp. 592–598, 2023, doi: 10.1145/3576050.3576146.
- [10] M. R. Al Fatah, A. A. Murtopo, and E. U. S. Utami, "Implementasi Algoritma K-Means untuk Pengelompokan Risiko Drop out Mahasiswa," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 3, pp. 2719–2726, 2025, doi: 10.31004/riggs.v4i3.2374.
- [11] D. D. Ndunguru, F. Xing, and C. Z. Oroni, "Hybrid methods for student dropout prediction: feature selection and model fusion," *Human-Intelligent Syst. Integr.*, vol. 7, no. 1, pp. 263–278, 2025, doi: 10.1007/s42454-025-00067-x.
- [12] A. M. Rabelo and L. E. Zarate, "A model for predicting dropout of higher education students," *Data Sci. Manag.*, vol. 8, no. 1, pp. 72–85, 2025, doi: 10.1016/j.dsm.2024.07.001.
- [13] C. H. Cho, Y. W. Yu, and H. G. Kim, "A Study on Dropout Prediction for University Students Using Machine Learning," *Appl. Sci.*, vol. 13, no. 21, p. 12004, 2023, doi: 10.3390/app132112004.
- [14] O. Lorenzo-Quiles, S. Galdon-Lopez, and A. Lendinez-Turon, "Factors contributing to university dropout: a review," *Front. Educ.*, vol. 8, 2023, doi: 10.3389/educ.2023.1159864.
- [15] N. Iam-On and T. Boongoen, "Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings," *Int. J. Mach. Learn. Cybern.*, vol. 8, no. 2, pp. 497–510, 2017, doi: 10.1007/s13042-015-0341-x.
- [16] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning*. Springer International Publishing, 2023.
- [17] J. Han, J. Pei, and H. Tong, *Data Mining*. Elsevier, 2024.
- [18] I. Attyatullatifah and M. Kamayani, "Comparison of Classification Algorithms for Predicting Graduation of Informatics Engineering Students with Orange Data Mining," *Indones. J. Comput. Sci.*, vol. 13, no. 2, 2024, doi: 10.33022/ijcs.v13i2.3796.