Prediksi dan Analisis Faktor Putus Studi Mahasiswa dengan Machine Learning pada Perguruan Tinggi Swasta

Prediction and Analysis of Student Dropout Factors Using Machine Learning at a Private Higher Education Institution

Luvia Friska Narulita*[1]
Universitas 17 Agustus 1945 Surabaya, Jl. Semolowaru No.45 Surabaya, 031-5931800
*E-mail : luvia@untag-sby.ac.id**[1]
**Corresponding author*
Received 7 May 2026; Revised 22 May 2026; Accepted 24 May 2026

Abstract - Putus studi mahasiswa masih menjadi salah satu permasalahan yang dihadapi oleh perguruan tinggi swasta karena dapat mempengaruhi kualitas akademik dan keberlangsungan institusi. Risiko putus studi mahasiswa perlu dideteksi sejak awal agar intervensi akademik dapat dilakukan secara lebih cepat dan tepat. Penelitian ini menggunakan pendekatan *Educational Data Mining* untuk memprediksi putus studi mahasiswa serta mengidentifikasi faktor-faktor yang berkontribusi terhadap risiko putus studi. Data penelitian diperoleh dari sistem informasi akademik pada salah satu perguruan tinggi swasta di Jawa Timur dengan total sebanyak 4.730 data mahasiswa. Variabel yang digunakan meliputi Indeks Prestasi Semester 1 (IPS1), Indeks Prestasi Semester 2 (IPS2), jumlah SKS yang diambil, frekuensi cuti, dan tren IPS. Penelitian menggunakan dua algoritma klasifikasi, yaitu Decision Tree dan Logistic Regression. Hasil penelitian menunjukkan bahwa model Decision Tree menghasilkan performa yang lebih baik dibandingkan Logistic Regression dengan tingkat *accuracy* sebesar 76,56% dan *recall* sebesar 82% dalam mendeteksi mahasiswa putus studi, sedangkan Logistic Regression menghasilkan *accuracy* sebesar 64,52% dengan *recall* sebesar 73%. Hasil analisis *feature importance*menunjukkan bahwa jumlah SKS dan IPS semester pertama merupakan faktor yang paling dominan dalam klasifikasi putus studi, sedangkan tren IPS memberikan kontribusi tambahan yang relatif lebih kecil. Temuan penelitian menunjukkan bahwa data akademik awal mahasiswa dapat dimanfaatkan untuk mendukung sistem *early warning* dalam mengidentifikasi mahasiswa berisiko putus studi secara lebih dini.

Kata kunci: putus studi, educational data mining, decision tree, logistic regression, prediksi mahasiswa

Abstract *— Student dropout remains one of the major challenges faced by private higher education institutions, as it can affect academic quality and institutional sustainability. The risk of student dropout needs to be detected early so that academic interventions can be carried out more quickly and accurately. This study applies an Educational Data Mining approach to predict student dropout and identify factors contributing to dropout risk. The research data were obtained from the academic information system of a private higher education institution in East Java, consisting of 4,730 student records. The variables used include Semester Grade Point Average 1 (GPA1), Semester Grade Point Average 2 (GPA2), the number of enrolled credits, leave frequency, and GPA trends. The study employed two classification algorithms, namely Decision Tree and Logistic Regression. The results show that the Decision Tree model achieved better performance than Logistic Regression, with an accuracy of 76.56% and a recall of 82% in detecting student dropout, while Logistic Regression achieved an accuracy of 64.52% with a recall of 73%. Feature importance analysis indicates that the number of enrolled credits and first-semester GPA are the most dominant factors in dropout classification, while GPA trends provide a relatively smaller additional contribution. The findings suggest that students' early academic*

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151

data can be utilized to support an early warning system for identifying students at risk of dropout at an earlier stage.

Keywords: *student dropout, educational data mining, decision tree, logistic regression, student prediction*

1. PENDAHULUAN

Putus studi mahasiswa masih dianggap sebagai hal yang kompleks dengan penyebab yang beragam, hal tersebut tertulis pada penelitian yang dilakukan pada tahun 2018 di Korea Selatan [1]. Berdasarkan data yang dikeluarkan oleh dinas komunikasi dan informatika provinsi Jawa Barat, angka kasus putus studi mahasiswa pada tahun 2020 adalah sebesar 602.208 mahasiswa dari total mahasiswa terdaftar sebanyak 8.483.213. Kasus putus studi paling rentan dialami oleh Perguruan Tinggi Swasta dengan total 478.826 orang atau 79,5% pada tahun 2020. Jumlah tersebut jauh lebih tinggi dibandingkan dengan total mahasiswa putus studi di Perguruan Tinggi Negeri yang mengalami mahasiswa putus studi sebesar 101.758 orang. Tingginya tingkat putus studi di Perguruan Tinggi Swasta tentunya mempengaruhi stabilitas institusi, karena Perguruan Tinggi Swasta masih mengandalkan pembayaran uang kuliah dari mahasiswa sebagai pendapatan utama.

Faktor – faktor yang menyebabkan mahasiswa mengalami putus studi antara lain tidak mampu memenuhi persyaratan akademik, tidak lulus mata kuliah wajib, keterbatasan jumlah SKS serta kendala dalam penyelesaian studi. Faktor administratif dan kondisi personal juga mempengaruhi keberlangsungan studi mahasiswa. Upaya sistematis diperlukan untuk mengidentifikasi mahasiswa yang berpotensi mengalami putus studi sejak dini. [2]

Educational Data Mining merupakan salah satu pendekatan yang dapat digunakan untuk menganalisis data pendidikan dan mengekstraksi pola yang relevan[3]. Teknik *Machine Learning* dapat digunakan untuk membangun model prediktif berdasarkan data riwayat akademis mahasiswa. Identifikasi mahasiswa yang berpotensi putus studi biasanya dilakukan berdasarkan pengamatan manual oleh bagian akademik yang selanjutnya diteruskan kepada fakultas, kaprodi dan dosen wali. Pendekatan yang lebih terstruktur dan otomatis dalam deteksi risiko putus studi sejak dini di universitas dibutuhkan karena pendekatan secara manual masih memiliki kelemahan dalam hal subjektifitas dan tidak bisa dilakukan pada skala besar.

Machine learning dapat digunakan untuk mengolah data akademik historis menjadi model prediktif untuk mengestimasi probabilitas putus studi mahasiswa di perguruan tinggi.[4] Penelitian tentang penggunaan machine learning untuk prediksi putus studi mahasiswa di perguruan tinggi telah banyak dilakukan oleh peneliti di luar negeri[5]. L.Kemper et al dalam penelitiannya [6] telah membuat model machine learning untuk memprediksi putus studi di Institut Teknologi Karlsruhe menggunakan Decision Tree dan Logistic Regression. Data yang digunakan berasal dari transkrip nilai yang menghasilkan tingkat akurasi sampai dengan 95% untuk mahasiswa semester 3 ke atas dan akurasi 83% untuk mahasiswa yang baru mencapai satu semester. Dalam penelitian tersebut, digunakan kombinasi angka rata – rata dan jumlah ujian yang berhasil dan gagal dilalui. Penelitian lain menggunakan model Gradient Boosting, Random Forest, Support Vector Machine dan Ensembel untuk memprediksi tingkat putus studi di perguruan tinggi.[7] [8][9][10]

Penelitian yang telah dilakukan berfokus pada putus studi, belum memperhatikan tahapan-tahapan yang dilalui oleh mahasiswa, padahal pada kondisi sebenarnya, putus studi dapat terjadi pada tahapan yang berbeda selama masa studi mahasiswa. Terjadinya putus studi mahasiswa dapat terjadi pada semester awal perkuliahan maupun pada tahap akhir perkuliahan.

Penelitian sebelumnya banyak menggunakan indikator kinerja akademik dan belum semuanya menggunakan variabel perilaku, yaitu hal - hal yang berhubungan dengan proses akademik, seperti pola cuti akademik, perilaku pendaftaran, kendala keuangan, dan intervensi

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151

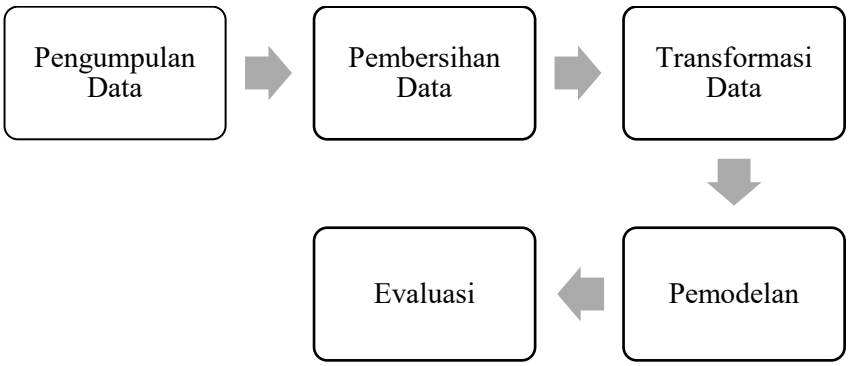
akademik. Faktor-faktor ini mungkin memainkan peran penting dalam membentuk ketekunan mahasiswa tetapi belum dieksplorasi secara memadai dalam model yang ada.

Untuk mengatasi kesenjangan tersebut, penelitian ini bertujuan untuk mengembangkan pendekatan berbasis pembelajaran mesin yang tidak hanya memprediksi putus studi mahasiswa tetapi juga mengidentifikasi faktor-faktor kunci yang berkontribusi. Penggabungan variabel yang lebih komprehensif antara kinerja akademik dan pola perilaku membuat penelitian ini berupaya memberikan pemahaman yang lebih holistik tentang putus studi mahasiswa. Temuan diharapkan dapat mendukung pengembangan sistem peringatan dini dan strategi intervensi berbasis data untuk meningkatkan retensi mahasiswa di perguruan tinggi swasta.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dalam kerangka Educational Data Mining untuk menganalisis putus studi mahasiswa pada perguruan tinggi swasta. Tujuan dari penelitian ini adalah untuk memprediksi status putus studi mahasiswa serta mengidentifikasi faktor dominan yang berkontribusi pada putus studi. Pendekatan yang digunakan adalah dengan mengkategorikan mahasiswa menjadi dua kelas, yaitu putus studi dan tidak putus studi.

Proses penelitian yang ditunjukkan pada gambar 1 dimulai dari pengumpulan data akademik mahasiswa, kemudian dilanjutkan dengan tahap pembersihan data. Setelah itu, tahapan berikutnya adalah proses transformasi data dan pembentukan variabel. Data yang sudah siap diteliti kemudian dilakukan pemodelan dengan algoritma Decision Tree dan Logistic Regression. Model yang dihasilkan dari kedua algoritma tersebut selanjutnya dievaluasi menggunakan metrik kinerja untuk menentukan performa terbaik.



Gambar 1. Proses Penelitian

Data yang digunakan dalam penelitian ini diperoleh dari sistem informasi akademik pada salah satu perguruan tinggi swasta di Jawa Timur dengan total data sebanyak 4730 data. Dari keseluruhan data tersebut, selanjutnya dibagi ke dalam data latih dan data uji, dengan pembagian sebesar 80% untuk data latih dan 20% untuk data uji. Pembagian data menggunakan metode stratified split untuk mempertahankan proporsi distribusi kelas target pada data latih dan data uji. Pendekatan tersebut penting terutama pada kasus klasifikasi dengan distribusi kelas yang tidak seimbang, karena pembagian data secara acak berpotensi menghasilkan evaluasi model yang bias terhadap kelas mayoritas[11]. Sedangkan untuk meningkatkan validitas evaluasi model dan mengurangi risiko overfitting, digunakan evaluasi tambahan menggunakan metode 5-fold cross-validation [12], [13]. Dari 4730 data tersebut, sejumlah 3196 data terdistribusi ke dalam kelas Tidak DO dan sejumlah 1534 terdistribusi ke dalam kelas DO. Distribusi tersebut menunjukkan bahwa dataset penelitian memiliki distribusi kelas yang tidak seimbang, untuk itu, pada model Logistic Regression digunakan mekanisme pembobotan kelas *(class weighting)* untuk mengurangi bias model terhadap kelas mayoritas.

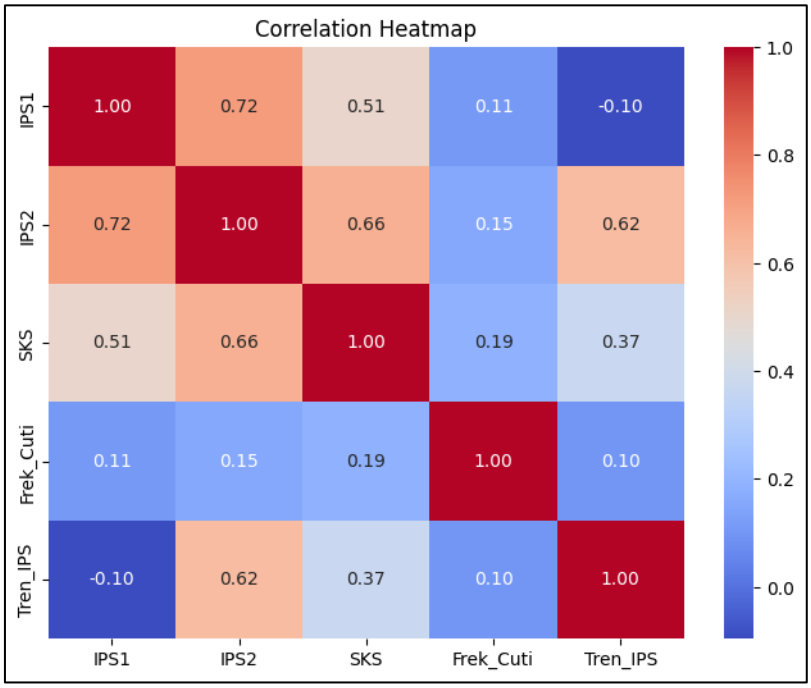
Setiap data merepresentasikan mahasiswa dengan atribut akademik dan perilaku yang diperoleh dari variabel - variabel penelitian yang digunakan, yaitu Indeks Prestasi semester 1,

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151

Indeks Prestasi semester 2, Jumlah SKS yang diambil dan frekuensi cuti. Variabel tersebut digunakan untuk merepresentasikan kondisi akademik awal dan perilaku mahasiswa yang mungkin berpengaruh terhadap putus studi. Seluruh identitas mahasiswa dianonimkan pada data yang akan digunakan untuk menjaga kerahasiaan data. Data hanya akan digunakan untuk kepentingan penelitian.

Tahapan *pre-processing* diperlukan untuk memastikan kualitas data dengan tahapan yang meliputi pembersihan data, seleksi variabel yang relevan, pengelompokan variabel target dan pembagian dataset. Proses pembersihan data diperlukan untuk menemukan fitur – fitur yang telah disaring untuk proses pembentukan model [14].

Pada penelitian ini, semua variabel independen yang digunakan memiliki tipe data numerik, sehingga tidak memerlukan proses encoding dari data kategorikal menjadi numerik. Sedangkan untuk variabel target, hanya menggunakan Binary Encoding sederhana untuk menentukan putus studi atau tidak. Keterkaitan variabel yang digunakan dalam penelitian ini ditunjukkan pada gambar 2 berupa heatmap. Heatmap pada gambar 2 menunjukkan keterkaitan yang kuat antara variabel IPS 1 dan IPS 2, serta antara IPS 2 dan SKS.



Gambar 2. Heatmap Variabel

Proses klasifikasi dari dataset yang sudah ada dilakukan dengan menggunakan metode Decision Tree dan Logistic Regression yang dapat menghasilkan model yang mudah diinterpretasikan dalam bentuk aturan keputusan.

Decision Tree merupakan metode klasifikasi yang direpresentasikan dalam bentuk struktur pohon, di mana setiap node menggambarkan pemilihan atribut atau pengambilan keputusan, sedangkan setiap cabang merepresentasikan hasil dari keputusan tersebut, dan node akhir menunjukkan label kelas dari suatu data. Proses klasifikasi dilakukan dengan menelusuri data dari node akar hingga mencapai leaf berdasarkan nilai atribut yang diuji pada setiap node. Setiap node internal berfungsi sebagai pengujian terhadap suatu atribut, dan setiap cabang yang keluar dari node tersebut menunjukkan kemungkinan hasil dari pengujian tersebut. Decision Tree banyak digunakan dalam pengambilan keputusan non-linear karena mampu menangkap hubungan kompleks antar variabel, meskipun menggunakan aturan keputusan yang sederhana. Selain itu, model Decision Tree juga dapat dengan mudah diubah menjadi bentuk aturan klasifikasi yang lebih mudah dipahami[15][16].

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151

Logistic Regression digunakan sebagai model pembanding karena kemampuannya dalam memodelkan hubungan antara variabel independen dan probabilitas terjadinya putus studi secara matematis. Model ini mengasumsikan hubungan linear antar variabel, sehingga dapat digunakan untuk mengidentifikasi arah pengaruh masing – masing variabel terhadap putus studi.

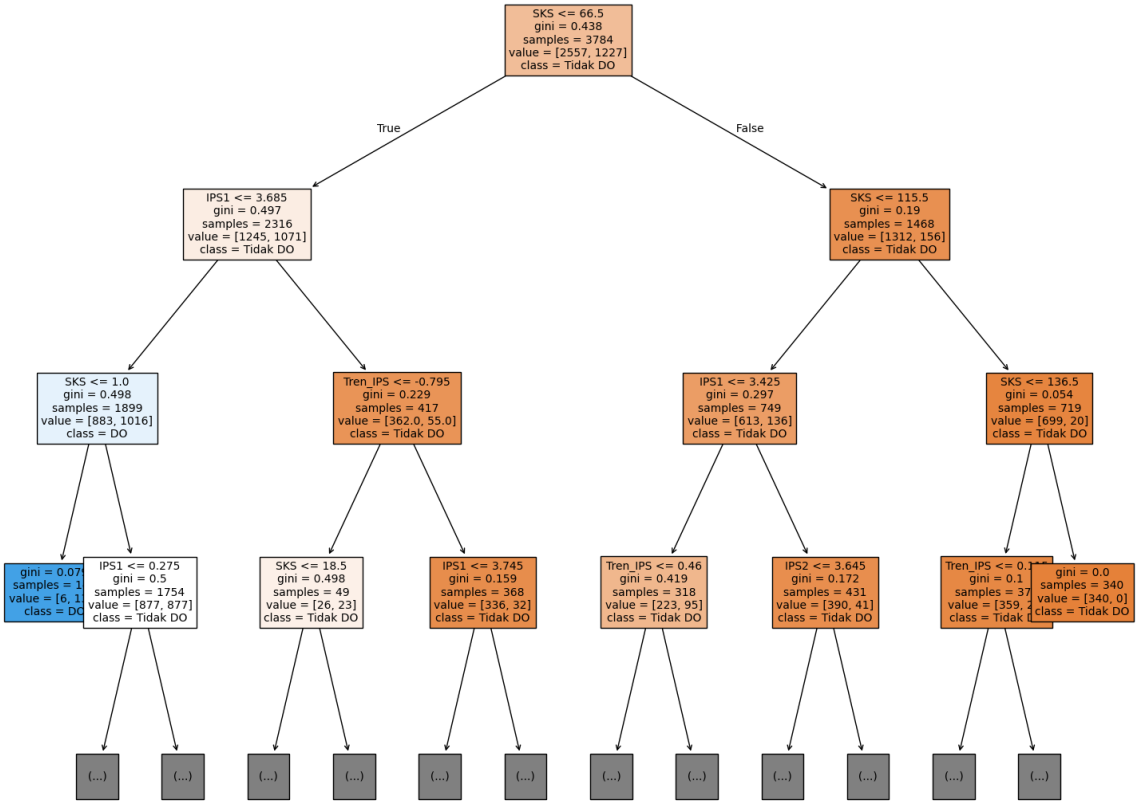
Model yang dibangun menggunakan data latih berfungsi untuk mempelajari pola hubungan antara variabel input dan status mahasiswa. Data uji digunakan untuk menguji kemampuan generalisasi model.

Kinerja dari model yang dibangun selanjutnya dievaluasi dengan menggunakan metrik *accuracy, precision, recall dan F1 score* [17] untuk mengukur kemampuan model dalam melakukan klasifikasi mahasiswa ke dalam kategori putus studi dan tidak putus studi secara tepat. Model Decision Tree yang dibuat juga digunakan untuk mengidentifikasi faktor – faktor dominan yang berkontribusi terhadap putus studi. Variabel yang sering muncul pada node awal serta aturan keputusan utama dianggap sebagai faktor yang memiliki pengaruh besar terhadap putus studi mahasiswa.

3. HASIL DAN DISKUSI

3.1 Hasil

Berdasarkan struktur Decision Tree yang ditunjukkan pada gambar 3, didapatkan bahwa ketika mahasiswa dengan jumlah SKS sangat rendah dengan performa awal akademik yang rendah memiliki kecenderungan lebih tinggi mengalami putus studi. Hal tersebut menunjukkan bahwa progres akademik pada semester awal menjadi indikator penting dalam klasifikasi risiko putus studi.



Gambar 3. Hasil Decision Tree

Pada tabel 1 ditunjukkan hasil *feature importance* dari model Decision Tree yang dibangun. Berdasarkan hasil tersebut, variabel SKS merupakan variabel dengan kontribusi terbesar dalam proses klasifikasi putus studi dengan nilai importance sebesar 0.551, diikuti dengan IPS pada semester 1 dengan nilai importance sebeasr 0.413. Selanjutnya tren perubahan

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151

IPS memberikan kontribusi yang relatif kecil, yaitu sebesar 0.033. Dari *feature importance* tersebut didapatkan bahwa progres akademik menjadi indikator utama dalam mengidentifikasi risiko putus studi.

Tabel 1. Hasil Feature Importance Model Decision Tree

Variabel	Importance
SKS	0.551
IPS1	0.413
Tren_IPS	0.033
IPS2	0.003
Frek_Cuti	0.000

Hasil pengujian menunjukkan bahwa model Decision Tree menghasilkan tingkat akurasi sebesar 76,56%, sedangkan Logistic Regression menghasilkan akurasi sebesar 64,52%. Dalam penelitian ini menunjukkan bahwa Decision Tree memiliki performa yang lebih baik dalam mengklasifikasikan status mahasiswa.

Pada tabel 2 yang berisi nilai odds ratio pada Logistic Regression menunjukkan bahwa nilai odds ratio pada variabel SKS yang mengindikasikan progres akademik bernilai sebesar 0.329. Hal tersebut menunjukkan bahwa variabel SKS berkaitan dengan penurunan probabilitas putus studi. Selain itu, hasil analisis korelasi menunjukkan adanya hubungan positif antara jumlah SKS dan performa akademik mahasiswa, khususnya terhadap IPS semester kedua (r = 0,66).

Tabel 2. Odds Ratio Logistic Regression

Variabel	Koefisien	Odds Ratio	Interpretasi
IPS1	0.431	1.539	asosiasi positif
IPS2	-0.474	0.623	menurunkan risiko
SKS	-1.112	0.329	menurunkan risiko
Frek_Cuti	0.024	1.025	pengaruh rendah
Tren_IPS	0.106	1.112	asosiasi kecil

Jika diperhatikan dari confusion matrix, model Decision Tree mampu mengklasifikasikan mahasiswa tidak putus studi dengan baik, dengan nilai recall sebesar 0,74, serta mampu mendeteksi mahasiswa putus studi dengan kemampuan yang baik. Nilai recall dalam mendeteksi mahasiswa putus studi sebesar 0,82 yang menunjukkan bahwa model Decision Tree cukup efektif dalam mengidentifikasi mahasiswa berisiko putus studi.

Nilai recall yang dihasilkan oleh Logistic Regression adalah 0,73 untuk kelas putus studi, tetapi memiliki nilai precision yang lebih rendah, yaitu 0,46. Model logistic regression cenderung menghasilkan lebih banyak prediksi positif terhadap putus studi, namun memiliki tingkat kesalahan yang lebih tinggi. Performa dari kedua model ditunjukkan pada tabel 3.

Tabel 3. Perbandingan performa model

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.77	0.60	0.82	0.69
Logistic Regression	0.65	0.47	0.73	0.57

Dalam penelitian ini menunjukkan bahwa Decision Tree memberikan performa yang lebih seimbang dibandingkan dengan Logistic Regression dalam memprediksi putus studi mahasiswa.

3.2 Diskusi

Perbedaan kinerja antara Decision Tree dan Logistic Regression menunjukkan adanya karakteristik data yang tidak sepenuhnya linear. Decision Tree memiliki kemampuan untuk

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151

menangkap hubungan non-linear antar variabel, sehingga mampu menghasilkan performa yang lebih baik dalam klasifikasi dibandingkan Logistic Regression yang berbasis hubungan linear.

Selain itu, nilai recall yang tinggi pada kedua model, khususnya Decision Tree, menunjukkan bahwa variabel yang digunakan dalam penelitian ini cukup representatif dalam mengidentifikasi mahasiswa yang berisiko putus studi. Hal ini menunjukkan bahwa faktor akademik awal dan perilaku mahasiswa memiliki peran penting dalam menentukan keberlangsungan studi.

Berdasarkan hasil analisis model, variabel jumlah SKS menunjukkan pengaruh yang paling signifikan terhadap putus studi mahasiswa. Rendahnya jumlah SKS dapat mengindikasikan keterlambatan progress akademik atau rendahnya keterlibatan mahasiswa. Mahasiswa dengan jumlah SKS yang lebih tinggi cenderung memiliki kemungkinan yang lebih kecil untuk mengalami putus studi. Selain itu, variabel IPS semester kedua juga menunjukkan pengaruh negatif terhadap putus studi, yang berarti semakin tinggi IPS semester kedua, semakin kecil kemungkinan mahasiswa mengalami putus studi.

Dalam model Logistic Regression, IPS semester pertama menunjukkan asosiasi positif terhadap probabilitas putus studi. Namun, interpretasi terhadap variabel ini perlu dilakukan secara hati-hati karena adanya korelasi yang cukup kuat antar variabel akademik, khususnya antara IPS semester pertama dan IPS semester kedua.

Variabel frekuensi cuti menunjukkan kontribusi yang relatif rendah terhadap model klasifikasi. Temuan ini konsisten dengan hasil korelasi dan nilai odds ratio yang menunjukkan hubungan yang lemah terhadap putus studi

Menariknya, IPS semester pertama menunjukkan kontribusi yang jauh lebih besar dibandingkan IPS semester kedua dalam proses klasifikasi. Temuan ini mengindikasikan bahwa semester pertama merupakan fase kritis dalam mendeteksi mahasiswa berisiko putus studi, sehingga intervensi akademik sejak awal perkuliahan menjadi penting dalam implementasi sistem *early warning*.

Hasil ini sejalan dengan konsep dalam Educational Data Mining yang menyatakan bahwa data akademik awal dapat digunakan untuk mengidentifikasi pola perilaku mahasiswa dan memprediksi keberhasilan studi[17], [18]. Penggunaan pendekatan berbasis data memberikan keunggulan dibandingkan metode manual karena mampu mengolah data dalam jumlah besar secara sistematis dan objektif.

Namun demikian, penelitian ini memiliki keterbatasan pada variabel yang digunakan, yang masih terbatas pada aspek akademik dan administratif. Faktor lain seperti kondisi sosial ekonomi, motivasi belajar, serta interaksi akademik tidak dapat dimasukkan dalam model karena keterbatasan data. Hal ini kemungkinan mempengaruhi kemampuan model dalam mendeteksi mahasiswa putus studi secara lebih akurat.

Akurasi model pada penelitian ini lebih rendah dibandingkan beberapa penelitian sebelumnya yang melaporkan akurasi di atas 90%. Perbedaan ini kemungkinan disebabkan oleh keterbatasan variabel yang digunakan, di mana penelitian ini hanya menggunakan variabel akademik dan administratif awal tanpa memasukkan faktor demografis, sosial ekonomi, maupun aktivitas pembelajaran digital.

4. KESIMPULAN

Putus studi mahasiswa lebih banyak dipengaruhi oleh progres akademik awal, khususnya jumlah SKS dan performa semester pertama, dibandingkan faktor administratif seperti frekuensi cuti, sehingga intervensi akademik sejak semester awal menjadi strategi penting untuk mencegah putus studi..

Penelitian ini memiliki keterbatasan variabel penelitian yang digunakan, meliputi motivasi mahasiswa, kondisi keuangan dan faktor psikologis mahasiswa selama menjalani perkuliahan. Penelitian lanjutan tentang penyebab putus studi mahasiswa di perguruan tinggi

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151

swasta dapat memasukkan faktor – faktor tersebut untuk mengetahui keterkaitan antara faktor non akademik terhadap putus studi mahasiswa.

Hasil penelitian ini menunjukkan bahwa data akademik awal mahasiswa dapat dimanfaatkan untuk menghasilkan pengetahuan yang mendukung pengambilan keputusan akademik berbasis data.

REFERENSI

[1] D. Kim and S. Kim, "Sustainable education: Analyzing the determinants of university student dropout by nonlinear panel data models," *Sustainability (Switzerland)*, vol. 10, no. 4, Mar. 2018, doi: 10.3390/su10040954.

[2] Jabar Digital Service, "Alasan Mahasiswa Drop Out atau Putus Kuliah, Apakah Gara-gara Skripsi Susah."

[3] E. Kalita *et al.*, "Educational data mining: a 10-year review," *Discover Computing*, vol. 28, no. 1, p. 81, May 2025, doi: 10.1007/s10791-025-09589-z.

[4] M. Soni and N. Jain, "MACHINE LEARNING-BASED DROPOUT PREDICTION FOR UNDERGRADUATES," *ShodhKosh: Journal of Visual and Performing Arts*, vol. 5, no. 5, May 2024, doi: 10.29121/shodhkosh.v5.i5.2024.4551.

[5] S. D. A. Bujang *et al.*, "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," *IEEE Access*, vol. 9, pp. 95608–95621, 2021, doi: 10.1109/ACCESS.2021.3093563.

[6] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, Jan. 2020, doi: 10.1080/21568235.2020.1718520.

[7] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, Jan. 2022, doi: 10.1016/j.caeai.2022.100066.

[8] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, "Early Prediction of student's Performance in Higher Education: A Case Study," 2021, pp. 166–175. doi: 10.1007/978-3-030-72657-7_16.

[9] C. Beaulac and J. S. Rosenthal, "Predicting University Students' Academic Success and Major using Random Forests," Jan. 2019, doi: 10.1007/s11162-019-09546-y.

[10] S. Anwar, "Predicting School Dropout Risk Using Machine Learning Models: A Comparative Study of Random Forest, Gradient Boosting, and Neural Network," *Journal Of Economich, Technology and Business (JETBIS)*, vol. 4, no. 6, 2025, [Online]. Available: https://jetbis.al-makkipublisher.com/index.php/al/index

[11] T. Huo, D. H. Glueck, E. A. Shenkman, and K. E. Muller, "Stratified split sampling of electronic health records," *BMC Med. Res. Methodol.*, vol. 23, no. 1, p. 128, May 2023, doi: 10.1186/s12874-023-01938-0.

[12] Z. Lyu *et al.*, "Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam," *Materials*, vol. 15, no. 4, p. 1477, Feb. 2022, doi: 10.3390/ma15041477.

[13] M. K. Mayangsari, I. Syarif, and A. Barakbah, "Evaluation of Stratified K-Fold Cross Validation for Predicting Bug Severity in Game Review Classification," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Jul. 2023, doi: 10.22219/kinetik.v8i3.1740.

[14] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: a case study with high school students," *Expert Syst.*, vol. 33, no. 1, pp. 107–124, Feb. 2016, doi: 10.1111/exsy.12135.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151

[15] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," *Procedia Technology*, vol. 25, pp. 326–332, 2016, doi: 10.1016/j.protcy.2016.08.114.

[16] C. Starbuck, "Logistic Regression," in *The Fundamentals of People Analytics*, Cham: Springer International Publishing, 2023, pp. 223–238. doi: 10.1007/978-3-031-28674-2_12.

[17] D. M. W. Powers and Ailab, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION."

[18] M. P. Colpo, T. Thompsen Primo, M. S. de Aguiar, and C. Cechinel, "Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges," *Revista Brasileira de Informática na Educação*, vol. 32, pp. 220–256, May 2024, doi: 10.5753/rbie.2024.3559.

e-ISSN : 2356-2579 | p-ISSN : 1412-2693, DOI : 10.62411/tc.v25i2.16151