Jurnal Inovasi Artificial Intelligence & Komputasional Nusantara (JIKOMNUS)
Vol. 5, No. 1, Juli 2026 – Halaman 46-51 p-ISSN: 3063-XXXX
SK.ISSN: 30634830/II.7.4/SK.ISSN/09/2024 e-ISSN: 3063-4830

ANALISIS KINERJA LOGISTIC REGRESSION, K-NEAREST NEIGHBOR, DECISION TREE, DAN RANDOM FOREST UNTUK PREDIKSI PUTUS STUDI MAHASISWA BERBASIS DATA AKADEMIK TABULAR

Angga Dwi Cahyono[*1], Khansa Alyssa Fauziyah[2], Arif Dwi Putra[3], Nafiendra Praba Hendyka[4], Aryo Bagus Satrio Wicaksono[5], Ani Dijah Rahajoe[6]

[1, 2, 3, 4, 5, 6)]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Email: 24081010127@student.upnjatim.ac.id[*1], 24081010110@student.upnjatim.ac.id[2], 24081010021@student.upnjatim.ac.id[3], 24081010285@student.upnjatim.ac.id[4], 24081010328@student.upnjatim.ac.id[5], anidijah.if@upnjatim.ac.id[6]

Abstrak

Putusnya studi mahasiswa menjadi masalah bagi institusi perguruan tinggi karena berdampak pada kualitas lulusan, reputasi institusi, dan efisiensi pengelolaan pendidikan. Perguruan tinggi seringkali menggunakan pendekatan secara konvensional dalam melihat perkembangan mahasiswa. Namun, pendekatan konvensional tidaklah efektif karena bersifat subjektif dan memiliki bias kognitif. Pendekatan dengan menggunakan data mining dapat menjadi solusi bagi perguruan tinggi untuk dapat dijadikan sebagai sistem peringatan dini dalam memprediksi mahasiswa yang berpotensi mengalami dropout. Penelitian ini bertujuan untuk membandingkan empat algoritma klasifikasi data mining dalam mendeteksi potensi putusnya studi mahasiswa yaitu *Logistic Regression, KNN, Decision Tree,* dan *Random Forest* dengan menggunakan dataset sekunder yang didapatkan dari website Kaggle yang terdiri dari 4.424 data mahasiswa dengan variabel yang berbeda-beda. Evaluasi performa model dilakukan dengan menggunakan metrik *accuracy, precision, recall,* dan *F1-Score.* Hasil penelitian menunjukkan bahwa model Random Forest adalah model dengan performa terbaik dengan nilai *precision* dan *F1-Score* tertinggi dibandingkan algoritma lainnya. Model ini dapat digunakan sebagai dasar sistem peringatan dini untuk mendeteksi mahasiswa yang berpotensi mengalami dropout.

Kata kunci: *data mining, klasifikasi, prediksi putus studi mahasiswa, random forest, data akademik tabular*

PERFORMANCE ANALYSIS OF LOGISTIC REGRESSION, K-NEAREST NEIGHBOR, DECISION TREE, AND RANDOM FOREST FOR PREDICTING STUDENT DROPOUT USING TABULAR ACADEMIC DATA

Abstract

Student dropout has become a major issue for higher education institutions as it affects graduate quality, institutional reputation, and the efficiency of educational management. Universities often rely on conventional approaches to monitor student progress. However, these conventional approaches are not effective because they are subjective and prone to cognitive bias. A data mining based approach can serve as a solution for higher education institutions by functioning as an early warning system to predict students who are at risk of dropping out. This study aims to compare four data mining classification algorithms in detecting potential student dropout, namely Logistic Regression, K Nearest Neighbor, Decision Tree, and Random Forest, using a secondary dataset obtained from the Kaggle platform consisting of 4.424 student records with diverse variables. Model performance is evaluated using accuracy, precision, recall, and F1-score metrics. The results show that the Random Forest model achieves the best performance with the highest precision and F1-score compared to the other algorithms. This model can be used as a basis for developing an early warning system to identify students who are at risk of dropping out.

Keywords: *data mining, classification, student dropout prediction, random forest, tabular academic data*

Jurnal Inovasi Artificial Intelligence & Komputasional Nusantara (JIKOMNUS)
Vol. 5, No. 1, Juli 2026 – Halaman 46-51 p-ISSN: 3063-XXXX
SK.ISSN: 30634830/II.7.4/SK.ISSN/09/2024 e-ISSN: 3063-4830

1. PENDAHULUAN

Perguruan tinggi adalah institusi yang memiliki peran penting dalam memfasilitasi pendidikan dan mencetak Sumber Daya Manusia (SDM) yang ahli dan unggul. Akan tetapi, salah satu permasalahan yang dialami oleh perguruan tinggi adalah tingginya angka mahasiswa tidak menyelesaikan studinya secara tepat waktu, bahkan hingga mengalami *drop-out.* Angka *dropout* mahasiswa di perguruan tinggi menunjukkan hasil yang signifikan. Data dari Pangkalan Data Pendidikan Tinggi (PDDIKTI) menunjukkan bahwa dari total 9,6 juta mahasiswa, ribuan mahasiswa mengalami *dropout.* Salah satu daerah dengan tingkat *dropout* yang tinggi di Indonesia adalah Jawa Timur dengan angka yang mencapai 14,84% (Regina Pramnesti et al., 2025).

Putusnya studi mahasiswa tidak hanya berdampak bagi individu yang bersangkutan, tetapi memiliki dampak yang lebih luas bagi perguruan tinggi, keluarga, dan sistem pendidikan. Penelitian (Fauziyah Laili et al., 2024) menunjukkan bahwa putusnya studi mahasiswa dapat mempengaruhi reputasi perguruan tinggi. Hal ini dapat terjadi karena fenomena tersebut menjadi penanda bahwa perguruan tinggi gagal dalam melakukan pembinaan dan pendampingan akademik. Tingginya angka putus mahasiswa juga dapat berdampak pada efektivitas pengelolaan mahasiswa, pencapaian target akademik institusi, dan kualitas layanan pendidikan (Rahmadani and Mukti, 2020).

Fenomena putusnya studi mahasiswa dapat diprediksi dengan dua pendekatan, yakni dengan cara konvensional dan cara modern berbasis data. Penelitian (Eegdeman et al., 2022) menunjukkan bahwa pendekatan konvesional memiliki keterbatasan dalam memprediksi potensi *dropout* mahasiswa karena terdapat unsur subjektivitas dan sensitif terhadap bias kognitif, sedangkan pendekatan modern berbasis data dapat mengatasi keterbatasan yang ada dengan pendekatan berbasis machine learning yang mampu memanfaatkan data akademik secara sistematis dan berkelanjutan. Untuk mengatasi kekurangan yang ada pada pendekatan konvensional, maka di sinilah peran data mining dan dibutuhkan. Data mining mampu mengolah data secara sistematis serta bisa menangkap potensi putus studi mahasiswa yang sulit diidentifikasi secara manual (Tumbilung and Efendi, 2025).

Penelitian ini bertujuan untuk dapat menganalisis potensi putus studi mahasiswa dengan berbasis data akademik tabular berdasarkan 35 variabel pada data sekunder yang didapatkan dari website Kaggle dengan label target awal yakni *Graduated, Enrolled,* dan *Dropout* yang bersifat *open source.* Variabel-variabel yang ada pada dataset meliputi data demografis, data administratif dan pendidikan, data sosial ekonomi individu, data akademik, serta data ekonomi makro. Penelitian ini akan membandingkan performa beberapa klasifikasi data mining dengan menggunakan empat algoritma yaitu KNN, Logistic Regression, Decision Tree, dan Random Forest untuk mendeteksi mahasiswa yang berpotensi mengalami putus studi dengan memanfaatkan sumber akademik dan non akademik. Hasil penelitian diharapkan dapat memberikan gambaran dalam mengidentifikasi potensi putus studi mahasiswa dengan menggunakan data mining sebagai sistem peringatan dini.

2. TINJAUAN PUSTAKA

Data mining adalah proses yang menggunakan teknik matematika, statistika, *artificial intelligence,* dan *machine learning*. Dalam pengolahan data, terdapat tahapan-tahapan yang harus dilakukan ketika mengolah sebuah data yaitu KDD *(Knowledge Discovery in Database).* Data mining merupakan salah satu tahapan pada KDD yang bertujuan untuk mendapatkan sebuah pengetahuan baru dari data yang diolah. Proses dari KDD meliputi pembersihan data *(data cleansing),* integrasi data *(data integration),* pemilihan data *(data selection)*, transformasi data *(data transformation)*, penggalian data *(data mining)*, evaluasi pola *(pattern evaluation)*, dan penyajian pengetahuan *(knowledge presentation)* (Kharis and Zili, 2022a).

Data mining dapat menjadi sebuah sistem penyelesaian masalah dalam pendidikan khususnya pada tingkatan perguruan tinggi. Dalam konteks perguruan tinggi, data mining dapat digunakan untuk mengolah data mahasiswa yang berpotensi putus studi. Salah satu cabang khusus pada data mining adalah *Educational Data Mining (EDM).* EDM berperan penting sebagai sistem untuk mendukung pengambilan keputusan sehingga relevan untuk digunakan dalam analisis dan prediksi potensi putus studi mahasiswa di perguruan tinggi.

3. METODE PENELITIAN

Penelitian ini menggunakan metode kuantitatif dengan pendekatan eksperimen. Pendekatan eksperimen dilakukan untuk menguji performa empat algoritma klasifikasi yang digunakan dalam pengujian data yaitu algoritma KNN, *Logistic Regression*, *Decision Tree*, dan *Random Forest*. Secara umum, objek yang diteliti adalah prediksi putus studi mahasiswa di perguruan tinggi. Pendekatan yang bersifat kuantitatif diterapkan karena studi ini berfokus pada pengolahan data berupa angka serta penilaian kinerja model dengan memakai metrik kuantitatif seperti *accuracy, precision, recall,* dan *F1-score.* Sedangkan evaluasi kinerja model klasifikasi pada penelitian ini dilakukan menggunakan confusion matrix yang menghasilkan empat kemungkinan prediksi, yaitu

True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN).

Accuracy dihitung menggunakan persamaan (1), digunakan untuk mengukur tingkat ketepatan model dalam melakukan klasifikasi secara keseluruhan. Precision dihitung menggunakan persamaan (2), digunakan untuk mengetahui tingkat ketepatan prediksi model terhadap kelas tertentu. Recall dihitung menggunakan persamaan (3), digunakan untuk mengukur kemampuan model dalam mendeteksi data pada kelas tertentu secara benar. Sementara itu, *F1-score* merupakan nilai rata-rata harmonik antara precision dan *recall* yang digunakan untuk menyeimbangkan kedua metrik tersebut dan dihitung menggunakan persamaan (4). Secara matematis, perhitungan metrik evaluasi tersebut dirumuskan sebagai berikut.

a. *Accuracy*
Mengukur persentase prediksi yang benar terhadap seluruh data.

$$\frac{TP+T}{TP+TN+FP+FN} \quad (1)$$

b. *Precision*
Mengukur ketepatan prediksi kelas positif.

$$\frac{TP}{TP+FP} \quad (2)$$

c. *Recall*
Mengukur kemampuan model dalam mendeteksi seluruh data positif.

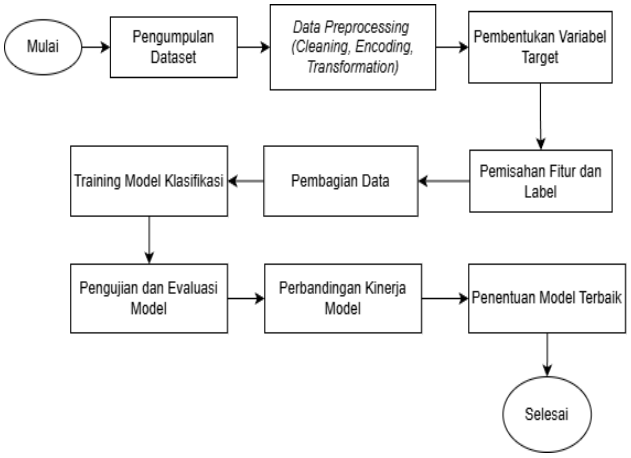
$$\frac{TP}{TP+FN} \quad (3)$$

d. *F1-Score*
Merupakan rata-rata harmonik antara Precision dan Recall.

$$2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Data yang digunakan adalah data sekunder yang didapatkan dari platform open source Kaggle berupa data akademik tabular (Realinho et al., 2022). Dataset yang digunakan berisi 35 variabel yang berbeda yakni variabel demografis, variabel akademik, variabel sosial-ekonomi, variabel administratif dan pendidikan awal, dan variabel ekonomi makro. Pada dataset yang digunakan, terdapat satu variabel target yang akan digunakan sebagai representasi hasil dari perhitungan. Terdapat tiga label pada target yaitu *Graduate* (lulus), *Enrolled* (masih menjalani studi), dan *Dropout* (putus studi). Dataset tersebut dipilih karena cocok untuk merepresentasikan data mahasiswa dengan variabel-variabel yang relevan dengan studi kasus prediksi *dropout* mahasiswa, dan disajikan dalam bentuk data tabular sehingga sesuai untuk klasifikasi data mining.

Penelitian ini dilakukan dalam beberapa tahapan dimulai dari persiapan data hingga evaluasi model. Tujuan dilakukannya tahapan tersebut agar analisis dapat dilakukan secara sistematik dan terstruktur. Agar dapat memahami lebih jelas terkait tahapan-tahapan yang dilakukan, akan kami gambarkan dalam diagram alir di bawah ini.



Gambar 1. Diagram Alir Tahapan Penelitian

4. HASIL DAN PEMBAHASAN

Pada bagian ini dilakukan analisis kinerja empat algoritma klasifikasi data mining, yaitu K-Nearest Neighbor (KNN), Logistic Regression, Decision Tree, dan Random Forest, dalam memprediksi status putus studi mahasiswa berdasarkan data akademik tabular. Penilaian performa model dilakukan menggunakan metrik *accuracy, precision, recall,* dan *F1-score,* dengan tujuan untuk memperoleh gambaran menyeluruh mengenai kemampuan masing-masing algoritma dalam melakukan klasifikasi, khususnya dalam mendeteksi mahasiswa yang berpotensi mengalami dropout.

4.1. Hasil Pengujian Model

4.1.1. Perbandingan Kinerja Model

Tabel 1. Perbandingan Kinerja Model

Model	Accuracy	Precision	Recall	F1-Score
LR	0.715	0.699	0.715	0.704
DT	0.673	0.674	0.673	0.673
RF	**0.764**	**0.758**	**0.764**	**0.730**
KNN	0.610	0.592	0.610	0.594

Berdasarkan Tabel 1, dapat dilihat bahwa algoritma Random Forest menghasilkan performa terbaik dengan nilai *accuracy*, *precision*, *recall*, dan *F1-score* tertinggi dibandingkan model lainnya.

Logistic Regression menunjukkan performa yang cukup stabil, namun masih berada di bawah *Random Forest*. *Decision Tree* memiliki performa menengah, sedangkan KNN menunjukkan performa terendah, yang mengindikasikan bahwa metode berbasis jarak kurang optimal pada dataset dengan jumlah variabel yang cukup banyak.

4.1.2. Analisis Hasil Pengujian Model

Hasil pengujian menunjukkan bahwa setiap algoritma memiliki karakteristik dan tingkat performa yang berbeda. *Logistic Regression* menghasilkan performa yang cukup stabil serta memiliki keunggulan dalam hal interpretabilitas model. Namun, algoritma ini memiliki keterbatasan dalam menangani hubungan non-linear antar variabel, sehingga nilai akurasi dan kemampuan deteksi kelas tertentu masih berada di bawah beberapa algoritma lainnya.

Algoritma *K-Nearest Neighbor* (KNN) bekerja dengan mengandalkan jarak kedekatan antar data, sehingga kinerjanya sangat dipengaruhi oleh pemilihan parameter *k* dan proses normalisasi data (Manurung et al., 2025). Pada dataset jumlah variabel yang relatif banyak, performa KNN cenderung menurun akibat meningkatnya kompleksitas dimensi data, yang menyebabkan proses klasifikasi menjadi kurang optimal (Halder et al., 2024).

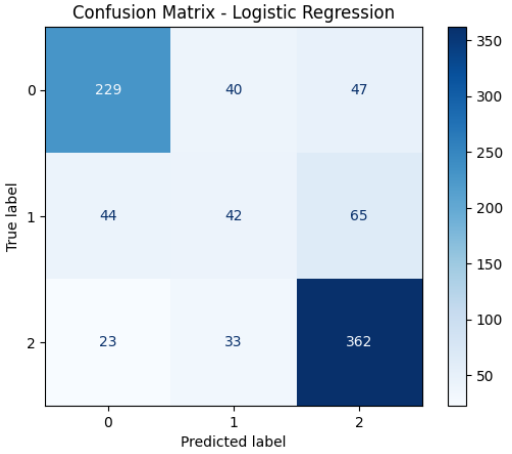
Berbeda dengan dua algoritma sebelumnya, *Decision Tree* mampu menangkap hubungan non-linear antar fitur dengan lebih baik dan melakukan pemilihan atribut secara otomatis. Hal ini membuat performa *Decision Tree* lebih baik dibandingkan Logistic Regression dan KNN. Meskipun demikian, model ini memiliki kelemahan berupa kecenderungan mengalami *overfitting,* terutama ketika struktur pohon terlalu dalam dan kompleks.

Algoritma *Random Forest* menunjukkan kinerja paling unggul dibandingkan algoritma lainnya. Model ini secara konsisten menghasilkan nilai *accuracy*, *precision*, *recall* dan *F1-score* tertinggi. Keunggulan tersebut diperoleh karena *Random Forest* merupakan metode *ensemble* yang mengombinasikan banyak pohon keputusan, sehingga mampu mengurangi risiko *overfitting* serta meningkatkan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya (Sulehu et al., 2025).

4.2. Analisis Confusion Matrix

Confusion Matrix digunakan untuk mengevaluasi performa model klasifikasi secara lebih rinci dengan membandingkan label aktual dan hasil prediksi model. Melalui Confusion Matrix, dapat diketahui jumlah prediksi yang benar maupun salah pada masing-masing kelas, sehingga memberikan gambaran yang lebih jelas mengenai kemampuan setiap model dalam melakukan klasifikasi data multikelas.

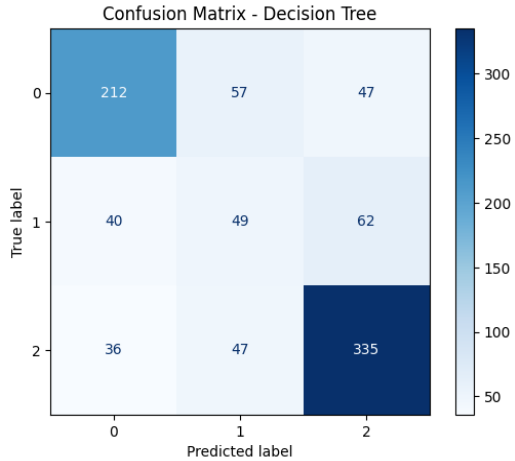
4.2.1. Confusion Matrix Logistic Regression



Gambar 2. *Confusion Matrix Logistic Regression*

Berdasarkan *Confusion Matrix* pada Gambar 2, dapat dilihat bahwa prediksi yang berada pada diagonal utama menunjukkan jumlah klasifikasi yang benar. Model *Logistic Regression* menunjukkan performa yang baik pada kelas *Dropout*, dengan sebanyak 362 data berhasil diprediksi secara benar. Hal ini menandakan bahwa model mampu mengenali mahasiswa yang berisiko putus studi dengan cukup akurat. Pada kelas *Graduate*, model berhasil memprediksi 229 data secara benar, namun masih terdapat kesalahan prediksi ke kelas *Enrolled* dan *Dropout*. Hal ini menunjukkan bahwa meskipun model cukup baik dalam mengenali mahasiswa yang lulus, masih terdapat beberapa data yang memiliki karakteristik yang mirip dengan kelas lain. Sementara itu, performa terendah terdapat pada kelas *Enrolled*, di mana hanya 42 data yang berhasil diprediksi dengan benar. Banyak data pada kelas ini yang salah diklasifikasikan sebagai *Graduate* maupun *Dropout*.

4.2.2. Confusion Matrix Decision Tree

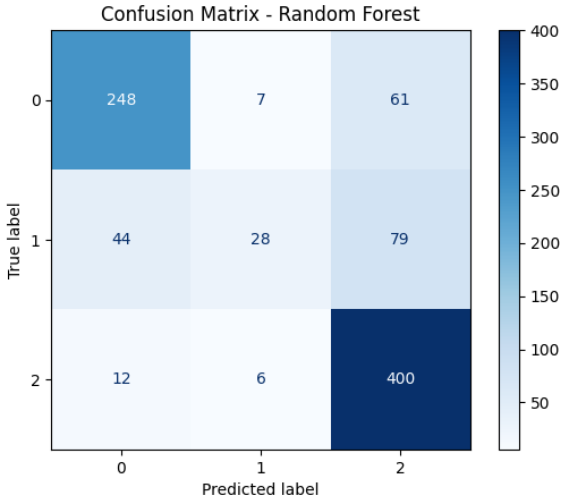


Gambar 3. *Confusion Matrix Decision Tree*

Berdasarkan *Confusion Matrix* pada Gambar 3, model *Decision Tree* mampu mengklasifikasikan kelas *Dropout* dengan cukup baik, ditunjukkan oleh

335 data yang berhasil diprediksi secara benar. Pada kelas *Graduate*, sebanyak 212 data berhasil diklasifikasikan dengan benar, meskipun masih terdapat kesalahan prediksi ke kelas *Enrolled* dan dropout. Sementara itu, pada kelas *Enrolled*, model hanya mampu memprediksi 49 data secara benar. Hal ini menunjukkan bahwa kelas *Enrolled* masih sulit dibedakan karena memiliki karakteristik yang cenderung tumpang tindih dengan dua kelas lainnya. Secara keseluruhan, *Decision Tree* memiliki performa yang lebih baik dibandingkan *Logistic Regression* dalam menangkap hubungan non-linear antar variabel, namun masih menunjukkan keterbatasan dalam mengklasifikasikan kelas *enrolled* secara optimal.

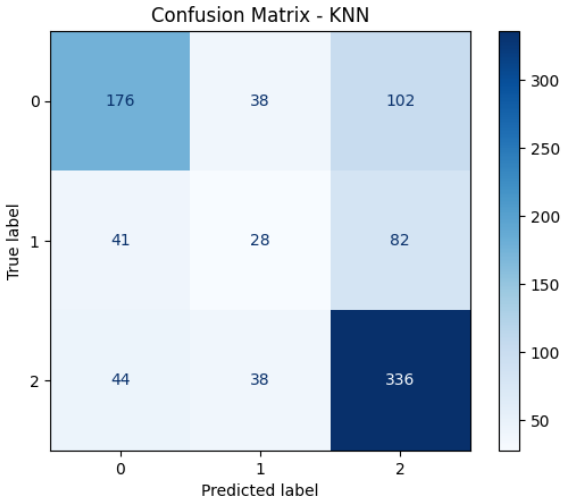
4.2.3. Confusion Matrix Random Forest



Gambar 4. *Confusion Matrix Random Forest*

Berdasarkan *Confusion Matrix* pada Gambar 4, model *Random Forest* menunjukkan performa yang sangat baik dalam mengklasifikasikan kelas *Dropout*, dengan 400 data berhasil diprediksi secara benar. Hal ini menunjukkan bahwa model memiliki kemampuan yang sangat tinggi dalam mengidentifikasi mahasiswa yang berisiko putus studi. Pada kelas *Graduate*, sebanyak 248 data berhasil diklasifikasikan dengan benar, dengan jumlah kesalahan prediksi yang relatif kecil dibandingkan model lainnya. Sementara itu, pada kelas *Enrolled*, model berhasil memprediksi 28 data secara benar, meskipun masih terdapat kesalahan klasifikasi ke kelas *Graduate* dan *Dropout*. Secara keseluruhan, *Random Forest* memberikan hasil klasifikasi paling optimal dibandingkan algoritma lain yang diuji. Keunggulan ini diperoleh karena pendekatan *ensemble learning* yang mampu menggabungkan banyak pohon keputusan, sehingga meningkatkan akurasi prediksi dan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

4.2.4. Confusion Matrix KNN



Gambar 5. *Confusion Matrix KNN*

Berdasarkan *Confusion Matrix* pada Gambar 5, model KNN mampu mengklasifikasikan kelas *Dropout* dengan cukup baik, ditunjukkan oleh 336 data yang berhasil diprediksi secara benar. Namun, pada kelas *Graduate*, hanya 176 data yang diklasifikasikan dengan benar, dengan cukup banyak kesalahan prediksi ke kelas *Dropout*. Sementara itu, performa terendah terdapat pada kelas *Enrolled*, di mana hanya 28 data yang berhasil diprediksi secara benar. Hal ini menunjukkan bahwa model KNN mengalami kesulitan dalam membedakan kelas *Enrolled*, terutama pada dataset dengan jumlah fitur yang relatif banyak.

4.3. Pembahasan Prediksi Putus Studi Mahasiswa

Berdasarkan hasil penelitian, metode klasifikasi berbasis *ensemble learning,* khususnya *Random Forest*, terbukti sangat efektif dalam memprediksi potensi putus studi mahasiswa. Kemampuan algoritma ini dalam mengelola data dengan jumlah variabel yang besar serta mendeteksi pola yang kompleks menjadikannya lebih unggul dalam konteks prediksi *dropout.*

Jika dibandingkan dengan pendekatan konvensional yang umumnya bergantung ada penilaian subjektif, pendekatan berbasis data mining menawarkan metode yang lebih objektif dan terstruktur. Model prediksi yang dihasilkan berpotensi untuk diimplementasikan sebagai sistem peringatan dini (*early warning system*), sehingga pihak perguruan tinggi dapat melakukan intervensi akademik secara lebih cepat dan tepat sasaran.

Secara keseluruhan, temuan dalam penelitian ini sejalan dengan hasil penelitian sebelumnya yang menyatakan bahwa *Random Forest* memiliki performa yang lebih baik dalam kasus klasifikasi data pendidikan, khususnya pada prediksi putus studi mahasiswa (Kharis and Zili, 2022b). Oleh karena itu, penerapan metode data mining pada data akademik tabular dapat menjadi solusi yang efektif dalam membantu perguruan tinggi menekan angka *dropout*

serta meningkatkan kualitas pengelolaan dan layanan pendidikan.

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa data mining dapat dilakukan untuk memprediksi potensi putus studi mahasiswa berbasis data akademik tabular. Dari empat algoritma klasifikasi data mining yang telah diuji, didapatkan hasil bahwa algoritma *Random Forest* menjadi model terbaik dengan nilai *Accuracy* 87,01%, *Precision* 87,89%, *Recall* 69,01%, dan *F1-Score* 77,32%. *Logistic Regression* menempati peringkat kedua dengan nilai *Accuracy* 85,20%, *Precision* 79,77%, *Recall* 72,18%, dan *F1-Score* 75,79%. *Decision Tree* menempati peringkat ketiga dengan nilai *Accuracy* 81,92%, *Precision* 72,14%, *Recall* 71,13%, dan *F1-Score* 71,63%. KNN menempati posisi terakhir dengan nilai *Accuracy* 65,20%, *Precision* 47,06%, *Recall* 67,61%, dan *F1-Score* 55,49%. Berdasarkan distribusi label pada dataset, kelas *Graduate* berjumlah 49,93%, *Dropout* 32,11%, dan *Enrolled* 17,95%. Setelah diubah menjadi klasifikasi biner, data terdiri dari *Non-dropout* 67,89% dan *Dropout* 32,11%.

Hasil penelitian ini dapat dimanfaatkan sebagai dasar dari sistem peringatan dini untuk membantu perguruan tinggi dalam memprediksi mahasiswa yang berpotensi mengalami *dropout* sehingga dapat membantu perguruan tinggi dalam mengelola sumber daya pendidikan dengan baik. Keterbatasan dalam penelitian ini terletak pada penggunaan dataset sekunder dan ruang lingkup data yang terbatas. Penelitian selanjutnya diharapkan dapat mengembangkan dataset secara luas, dengan menggunakan algoritma lain.

6. DAFTAR PUSTAKA

EEGDEMAN, I., CORNELISZ, I., VAN KLAVEREN, C. and MEETER, M., 2022. Computer or teacher: Who predicts dropout best? *Frontiers in Education*, 7. https://doi.org/10.3389/feduc.2022.976922.

FAUZIYAH LAILI, U., TANZEH, A., EFENDI, N. and GUFRON, M., 2024. K-Means Clustering Method On Academic Advising Management And Early Detection Of Student Dropout (Sequential Explanatory Mixed Method Study At UIN Sunan Ampel Surabaya And IAIN Kediri). *International Journal of Educational Research & Social Sciences*. [online] Available at: <https://ijersc.org/>.

HALDER, R.K., UDDIN, M.N., UDDIN, M.A., ARYAL, S. and KHRAISAT, A., 2024. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11(1). https://doi.org/10.1186/s40537-024-00973-y.

KHARIS, S.A.A. and ZILI, A.H.A., 2022a. Learning Analytics dan Educational Data Mining pada Data Pendidikan. *JURNAL RISET PEMBELAJARAN MATEMATIKA SEKOLAH*, [online] 6(1), pp.12–20. https://doi.org/10.21009/jrpms.061.02.

KHARIS, S.A.A. and ZILI, A.H.A., 2022b. Learning Analytics dan Educational Data Mining pada Data Pendidikan. *JURNAL RISET PEMBELAJARAN MATEMATIKA SEKOLAH*, [online] 6(1), pp.12–20. https://doi.org/10.21009/jrpms.061.02.

MANURUNG, J., SARAGIH, H., PRABUKUSUMO, M.A. and FIRDAUS, E.A., 2025. Optimizing the performance of the K-Nearest Neighbors algorithm using grid search and feature scaling to improve data classification accuracy. *Jurnal Mandiri IT*, [online] 14(2), pp.260–268. Available at: <www.ejournal.isha.or.id/index.php/Mandiri>.

RAHMADANI, A. and MUKTI, Y.R., 2020. Adaptasi akademik, sosial, personal, dan institusional : studi college adjustment terhadap mahasiswa tingkat pertama. *Jurnal Konseling dan Pendidikan*, 8(3), p.159. https://doi.org/10.29210/145700.

REALINHO, V., MACHADO, J., BAPTISTA, L. and MARTINS, V.M., 2022. *Predict students' dropout and academic success*. [online] https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention. Available at: <https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention> [Accessed 22 December 2025].

REGINA PRAMNESTI, A., DIJAH RAHAJOE, A., MUMPUNI, R., 2025. CICES (Cyberpreneurship Innovative and Creative Exact and Social Science) Optimasi Model Prediksi Kelulusan Mahasiswa Berbasis Principal Component Analysis dan Modified K-Nearest Neighbor. *Agustus*, 11(2). https://doi.org/10.33050/cices.v11i2.3914.

SULEHU, M., WISDA, W., WANITA, F. and MARKANI, M., 2025. Optimasi Prediksi Kelulusan Mahasiswa Menggunakan Random Forest untuk Meningkatkan Tingkat Retensi. *Jurnal Minfo Polgan*, 13(2), pp.2364–2374. https://doi.org/10.33395/jmp.v13i2.14472.

TUMBILUNG, C.F. and EFENDI, R., 2025. PREDIKSI RISIKO DROP OUT MAHASISWA MENGGUNAKAN MODEL MACHINE LEARNING BERBASIS DATA AKADEMIK (Studi Kasus : Universitas XYZ). *JTIK (Jurnal Teknologi Informasi dan Komunikasi)*,

[online] 16(2). Available at: <http://ejurnal.provisi.ac.id/index.php/JTIKP>.