Volume 20 Nomor 02 (2025) 63-73
E – Link P-ISSN 1858-2109 E-ISSN 2656-5676
Jurnal Teknik Elektro dan Informatika

PEMODELAN PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA D3 TEKNOLOGI INFORMASI MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS (KNN)

Salnan Ratih Asriningtias[1)] , Dwi Utari Surya[2)] , Eka Ratri Noor Wulandari[3)] , Mochamad Dimas Putra Hermawan[4)] , Rifqi Alfiansyah Kamil[5)] , David Kurniawan[6)]
[1,2,3)] Program Studi D-III Teknologi Informasi, Fakultas Vokasi, Universitas Brawijaya
JL. Veteran No 12 – 14, Malang, Jawa Timur
E-mail : [1)]salnan@ub.ac.id, [2)]d.utarisurya@ub.ac.id, [3)]ekaratri@ub.ac.id, [4)] dimasputra21@student.ub.ac.id, [5)] rifqialfiansyah@student.ub.ac.id, [6)] david.kurniawan@student.ub.ac.id

ABSTRAK

Peningkatan jumlah mahasiswa pada Program Studi D3 Teknologi Informasi menyebabkan pengelolaan kelulusan tepat waktu menjadi semakin menantang. Dalam konteks ini, metode machine learning digunakan untuk menganalisis pola secara komprehensif dan memprediksi tingkat kelulusan tepat waktu. Dataset yang digunakan berjumlah 608 mahasiswa yang terdiri dari angkatan 2018 hingga 2020, dengan atribut seperti IPK Semester 1–6, jenis kelamin, serta jalur masuk, yang keseluruhannya berkontribusi pada pemahaman faktor-faktor yang memengaruhi kelulusan tepat waktu. Penelitian ini menerapkan dan mengevaluasi algoritma machine learning, khususnya K-Nearest Neighbors (KNN). Proses pemodelan dilakukan melalui pembagian data menggunakan metode train-test split dengan rasio 70:30 dan divalidasi lebih lanjut menggunakan 10-fold cross-validation untuk memastikan kemampuan generalisasi model. Evaluasi kinerja model dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa model KNN menghasilkan performa terbaik dengan nilai accuracy sebesar 84%, precision 79%, recall 98%, dan F1-score 87%, melampaui performa Decision Tree dan Random Forest. Tingginya nilai recall menunjukkan kemampuan KNN yang sangat baik dalam mendeteksi mahasiswa yang berpotensi lulus tepat waktu. Kebaruan penelitian ini terletak pada penerapan dan evaluasi komprehensif model KNN dengan validasi berlapis (train-test split dan cross-validation) pada konteks Program Studi D3 Teknologi Informasi di Indonesia, yang masih jarang dikaji dalam penelitian sebelumnya. Temuan ini diharapkan dapat menjadi dasar pengambilan keputusan akademik dalam perencanaan intervensi dini guna meningkatkan capaian kelulusan tepat waktu mahasiswa.

Kata kunci : K-Nearest Neighbors, kelulusan tepat waktu, D3 Teknologi Informasi, prediksi kelulusan

ABSTRACT

The increasing number of students in the Diploma 3 (D3) Information Technology Study Program has made the management of on-time graduation increasingly challenging. In this context, machine learning methods are employed to comprehensively analyze patterns and predict on-time graduation rates. The dataset used in this study consists of 608 students from the 2018–2020 cohorts, with attributes including Grade Point Average (GPA) from semesters 1 to 6, gender, and admission

pathways, all of which contribute to understanding the factors influencing on-time graduation. This study applies and evaluates machine learning algorithms, with a particular focus on the K-Nearest Neighbors (KNN) algorithm. The modeling process is conducted using a train–test split with a 70:30 ratio and further validated through 10-fold cross-validation to ensure the model's generalization capability. Model performance is evaluated using accuracy, precision, recall, and F1-score metrics. The experimental results indicate that the KNN model achieves the best performance, with an accuracy of 84%, precision of 79%, recall of 98%, and an F1-score of 87%, outperforming Decision Tree and Random Forest models. The high recall value demonstrates KNN's strong capability in identifying students with a high potential to graduate on time. The novelty of this study lies in the comprehensive application and evaluation of the KNN model with layered validation (train–test split and cross-validation) in the context of the Diploma in Information Technology Study Program in Indonesia, which remains relatively underexplored in previous studies. These findings are expected to serve as a basis for academic decision-making in planning early interventions to improve students' on-time graduation outcomes.

***Keywords**: K-Nearest Neighbors, on-time graduation, Diploma in Information Technology, graduation prediction*

1. PENDAHULUAN

Kelulusan tepat waktu merupakan salah satu indikator utama keberhasilan pengelolaan akademik dan menjadi komponen penting dalam penilaian akreditasi program studi di perguruan tinggi Indonesia[1]. Tingginya tingkat kelulusan tepat waktu berkontribusi langsung terhadap reputasi institusi serta kesiapan lulusan dalam memasuki dunia kerja atau melanjutkan pendidikan ke jenjang yang lebih tinggi[2]. Oleh karena itu, perguruan tinggi dituntut untuk memiliki sistem yang mampu memantau dan meningkatkan capaian akademik mahasiswa secara berkelanjutan.

Namun demikian, banyak institusi pendidikan masih menghadapi kendala dalam mengidentifikasi mahasiswa yang berpotensi tidak lulus tepat waktu sejak dini [3]. Pendekatan konvensional yang bersifat reaktif menyebabkan intervensi akademik sering dilakukan setelah masalah akademik mahasiswa berkembang lebih jauh, sehingga kurang efektif [4]. Kondisi ini menunjukkan perlunya pendekatan berbasis data yang mampu memberikan prediksi kelulusan secara lebih akurat dan proaktif.

Berbagai penelitian sebelumnya telah memanfaatkan teknik machine learning untuk memprediksi kelulusan mahasiswa, seperti penggunaan K-Nearest Neighbors (KNN) [5], Random Forest [6], Regresi Logistik [7], dan Support Vector Machine [8]. Hasil penelitian-penelitian tersebut menunjukkan bahwa performa akademik awal dan pola nilai semester menjadi prediktor penting dalam menentukan kelulusan. Namun, sebagian besar penelitian tersebut dilakukan pada jenjang pendidikan sarjana (S1) atau menggunakan dataset lintas program studi dengan karakteristik yang heterogen.

Research gap dari penelitian ini terletak pada masih terbatasnya studi yang secara khusus membahas prediksi kelulusan tepat waktu pada jenjang Diploma III (D3), khususnya Program Studi D3 Teknologi Informasi di Indonesia. Karakteristik pendidikan vokasi yang lebih menekankan pada praktik dan durasi studi yang lebih singkat menyebabkan pola kelulusan mahasiswa D3 berpotensi berbeda dengan mahasiswa S1 [9], sehingga hasil penelitian terdahulu belum tentu dapat digeneralisasikan secara langsung.

Selain itu, penelitian-penelitian sebelumnya umumnya hanya melaporkan nilai akurasi sebagai indikator utama, tanpa memberikan analisis komprehensif terhadap metrik evaluasi lain seperti recall, F1-score, dan kurva ROC yang penting dalam konteks identifikasi mahasiswa berisiko [10][11].

Penelitian ini menegaskan perbedaannya dengan mengevaluasi performa model secara lebih menyeluruh serta membandingkan tiga algoritma machine learning populer, yaitu KNN, Decision Tree, dan Random Forest, pada dataset mahasiswa D3 Teknologi Informasi yang terstruktur.

Kontribusi utama penelitian ini diantaranya sebagai berikut :

1. Menyediakan studi empiris prediksi kelulusan tepat waktu khusus pada mahasiswa D3 Teknologi Informasi berbasis data akademik multi-semester.
2. Menunjukkan bahwa algoritma KNN memiliki performa terbaik dibandingkan Decision Tree dan Random Forest, terutama dalam aspek recall yang sangat penting untuk deteksi dini mahasiswa berisiko.
3. Menyajikan evaluasi model yang komprehensif menggunakan accuracy, precision, recall, F1-score, ROC-AUC, dan learning curve untuk memastikan kemampuan generalisasi model.

Berdasarkan uraian tersebut, tujuan penelitian ini adalah mengembangkan dan mengevaluasi model prediksi kelulusan tepat waktu mahasiswa D3 Teknologi Informasi menggunakan algoritma machine learning, serta mengidentifikasi model dengan kinerja terbaik yang dapat dijadikan dasar dalam pengambilan keputusan dan intervensi akademik oleh pengelola program studi.

2. METODE PENELITIAN

Penelitian terkini menggunakan teknik pembelajaran mesin tradisional telah membuktikan efektivitasnya dalam memprediksi kelulusan mahasiswa. Sebagai contoh, Salim dkk. [12] memanfaatkan algoritma KNN untuk memprediksi kelulusan dengan membagi data akademik selama beberapa tahun. Menggunakan K-Fold Cross-Validation, penelitian tersebut menunjukkan bahwa akurasi prediksi meningkat setiap tahun. Altabrawee dkk. [13] menggunakan algoritma Random Forest untuk memprediksi kelulusan mahasiswa di sebuah universitas besar dan menemukan bahwa variabel seperti kehadiran dan nilai ujian merupakan prediktor yang signifikan. Selain itu, Desfiandi and Soewito [14] menerapkan model Regresi Logistik untuk menganalisis data akademik mahasiswa dan berhasil memprediksi kelulusan dengan tingkat akurasi tinggi berdasarkan performa awal semester. Junaidi dkk. [15] menggunakan model Support Vector Machine (SVM) untuk membedakan mahasiswa yang akan lulus tepat waktu dan yang tidak, menekankan pentingnya variabel non-akademik seperti partisipasi kegiatan organisasi.

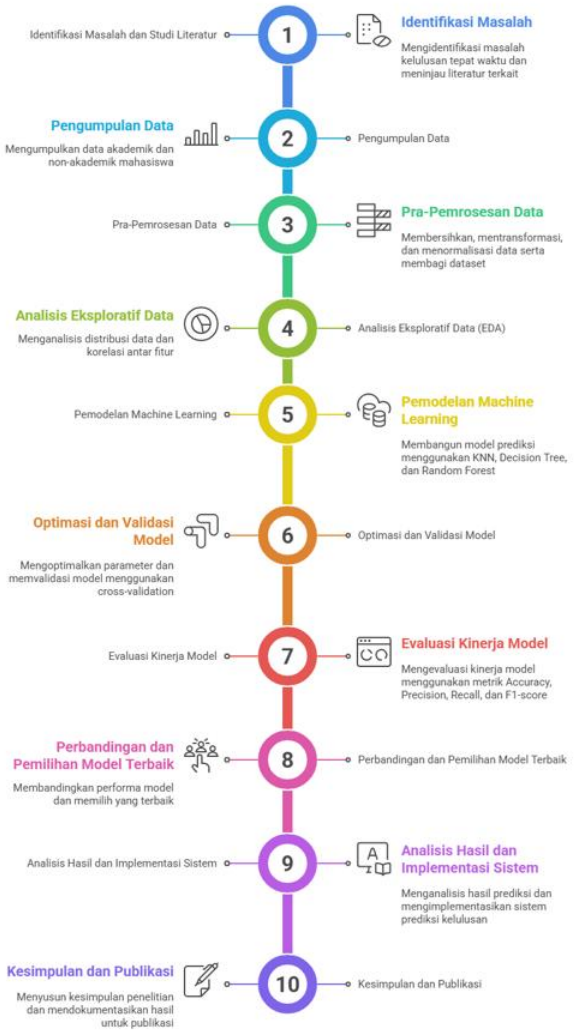
Di Indonesia, khususnya pada Program Studi D3 Teknologi Informasi, pemanfaatan machine learning dalam memprediksi kelulusan tepat waktu belum banyak dieksplorasi. Penelitian ini mengisi kesenjangan tersebut dengan menerapkan teknik machine learning pada data mahasiswa D3 Teknologi Informasi, sehingga dapat mendukung peningkatan akreditasi dan mempersiapkan mahasiswa menghadapi dunia kerja atau pendidikan lanjut [16].

Penelitian ini mengeksplorasi penerapan tiga model machine learning: K-Nearest Neighbors (KNN), Random Forest, dan Decision Tree untuk memprediksi kelulusan tepat waktu mahasiswa D3 Teknologi Informasi. Dengan mengintegrasikan berbagai atribut mahasiswa, penelitian ini bertujuan mengembangkan model prediktif komprehensif yang dapat memberikan pemahaman mendalam tentang faktor-faktor yang memengaruhi kelulusan tepat waktu, sekaligus mendukung strategi akademik yang lebih efektif bagi institusi pendidikan tinggi di Indonesia [17].

Tahapan penelitian ini terdiri dari proses pengumpulan data, penentuan variabel penelitian, tahapan pra-pemrosesan data, serta penerapan model machine learning, sebagaimana ditunjukkan pada Gambar 1. Setiap tahapan disusun secara sistematis untuk memastikan data yang digunakan memiliki kualitas yang baik dan siap dianalisis secara komputasional.

Tujuan utama dari tahapan penelitian ini adalah memberikan pemahaman yang komprehensif mengenai proses persiapan dan analisis data dalam memprediksi kelulusan tepat waktu mahasiswa Program Studi D3 Teknologi

Informasi. Melalui pendekatan ini, diharapkan model machine learning yang dibangun mampu menghasilkan prediksi yang akurat dan dapat dijadikan dasar dalam pengambilan keputusan akademik.



Gambar 1. Tahapan Penelitian.

2.1 Data

Data penelitian berasal dari rekam akademik mahasiswa Program Studi D3 Teknologi Informasi, meliputi angkatan 2018 hingga 2020. Seluruh data telah dikumpulkan dan dianonimkan sesuai standar perlindungan data pribadi.

2.2 Fitur Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 608 data mahasiswa dengan 12 variabel yang mencakup informasi akademik seperti yang ditunjukan pada Tabel 1. Fitur akademik berupa Indeks Prestasi Semester (IPS) dari semester 1 hingga semester 6 digunakan sebagai representasi perkembangan performa akademik mahasiswa selama masa studi. Nilai IPS dipilih sebagai fitur utama karena menjadi indikator paling langsung dan kuantitatif terhadap pencapaian akademik mahasiswa setiap semester.

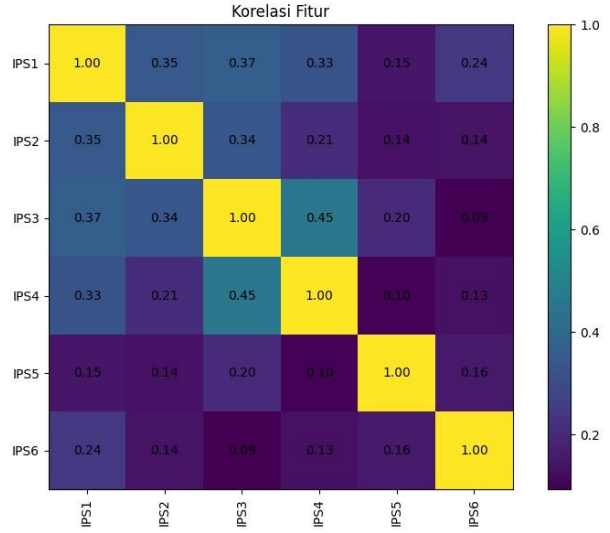
Variabel target dari penelitian ini adalah "Lulus Tepat Waktu," yang dikonversi menjadi variabel biner untuk memudahkan proses klasifikasi oleh model machine learning. Selain itu, dilakukan analisis proporsi kelas pada variabel target Kelulusan Tepat Waktu untuk mengidentifikasi potensi ketidakseimbangan kelas (class imbalance) dalam dataset. Hasil analisis menunjukkan bahwa distribusi kelas "Ya" (lulus tepat waktu) dan "Tidak" (tidak lulus tepat waktu) berada pada proporsi 72% : 28%.

Meskipun terdapat perbedaan proporsi antar kelas, ketimpangan tersebut masih berada dalam batas yang dapat ditangani oleh model klasifikasi tanpa penerapan teknik penyeimbangan data tambahan, seperti oversampling atau undersampling. Oleh karena itu, dataset digunakan dalam kondisi asli untuk menjaga karakteristik distribusi data yang merepresentasikan kondisi riil mahasiswa.

Tabel 1. Definisi dan Deskripsi Fitur Dataset.

Variabel	Definisi
IPS Semester 1	Indeks Prestasi Semester pada semester pertama.
IPS Semester 2	Indeks Prestasi Semester pada semester kedua
IPS Semester 3	Indeks Prestasi Semester pada semester ketiga
IPS Semester 4	Indeks Prestasi Semester pada semester keempat
IPS Semester 5	Indeks Prestasi Semester pada semester kelima
IPS Semester 6	Indeks Prestasi Semester pada semester keenam
Lulus Tepat Waktu	Indikator kelulusan tepat waktu mahasiswa (Ya/Tidak)

Analisis korelasi pada gambar 2 menunjukkan bahwa seluruh nilai IPS memiliki korelasi positif satu sama lain, terutama pada IPS3 dan IPS4, yang mengindikasikan adanya konsistensi performa akademik dari semester ke semester.



Gambar 2. Peta Korelasi Antar Fitur Dataset : Hasil pengolahan data penelitian

2.3 Label Encoding

Dataset penelitian ini memiliki banyak atribut yang menggambarkan kualitas akademik mahasiswa. IPK semester 1 sampai 6 menggunakan nilai numerik. Variabel target, 'Kelulusan Tepat Waktu', merupakan indikator biner yang menunjukkan apakah seorang mahasiswa lulus tepat waktu, dengan 0 mewakili 'Tidak' dan 1 mewakili 'Ya'. Proses pengkodean yang ditunjukkan pada tabel 2 memungkinkan data kategorikal diproses secara efektif oleh algoritma pembelajaran mesin yang digunakan dalam penelitian ini.

Tabel 2. Definisi Variabel dan Tipe Nilai pada Dataset Penelitian.

Variabel	Definisi	Nilai
IPK Semester 1-6	Indeks Prestasi Kumulatif (IPK) pada semester tertentu	Nilai numerik
Kelulusan Tepat Waktu	Indikator kelulusan tepat waktu mahasiswa	0 (Tidak), 1 (Ya)

Sebelum proses pemodelan, dilakukan tahap normalisasi fitur numerik untuk memastikan setiap atribut berada pada skala yang sebanding. Hal ini sangat penting terutama untuk algoritma K-Nearest Neighbors (KNN), yang menghitung kedekatan antar data berdasarkan jarak *Euclidean*.

Pada penelitian ini, fitur numerik berupa IPS Semester 1 hingga Semester 6 dinormalisasi menggunakan metode *MinMaxScaler*, yang mengubah data sehingga memiliki nilai rata-rata nol dan standar deviasi satu. Proses normalisasi ini bertujuan untuk mencegah fitur dengan rentang nilai yang lebih besar mendominasi perhitungan jarak dan memengaruhi hasil klasifikasi.

2.4 Pemodelan

Dalam penelitian ini, proses pemodelan dilakukan setelah tahap prapemrosesan data. Algoritma machine learning dilatih untuk memprediksi kelulusan tepat waktu mahasiswa program D3 Teknologi Informasi berdasarkan karakteristik akademik seperti IPK Semester 1-6, Jalur Penerimaan, dan Jenis Kelamin. Algoritma yang digunakan meliputi K-Nearest Neighbors (KNN), Decision Tree (DT), dan Random Forest (RF).

KNN mengklasifikasikan data baru berdasarkan kedekatannya dengan sejumlah tetangga terdekat. Pendekatan ini efisien untuk dataset berukuran kecil hingga sedang dan mudah diimplementasikan [18]. Persamaan (1) merupakan rumus Jarak *Euclidean* untuk dua titik $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + .. + (p_n - q_n{}^2)} \quad .....(1)$$

dimana $d(p,q)$ adalah jarak Euclidean antara titik p dan titik q. Nilai p_i dalah nilai atribut ke-i dari titik p. Nilai q_i adalah nilai atribut ke-i dari titik q . Nilai n menyatakan jumlah atribut atau dimensi yang digunakan dalam perhitungan jarak.

Dalam KNN, algoritma menemukan k titik terdekat berdasarkan jarak ini, lalu menetapkan kelas berdasarkan suara mayoritas tetangga terdekat.

Decision Tree adalah struktur berbasis pohon yang membagi data menjadi subset berdasarkan aturan keputusan sederhana yang dinyatakan dalam persamaan (2). Struktur ini dikenal karena interpretabilitasnya dan kesesuaiannya untuk data dengan banyak fitur [19].

$$Gini = 1 - \sum_{i=1}^{n} p_i{}^2 \quad \text{.....(2)}$$

dimana p_i adalah proporsi sampel kelas III pada suatu simpul tertentu. Nilai Gini menunjukkan seberapa beragam kelas pada suatu simpul tertentu (semakin rendah semakin baik).

Random Forest membangun beberapa pohon keputusan selama pelatihan, dan keluarannya ditentukan melalui pemungutan suara di antara pohon-pohon tersebut yang ditunjukkan dalam persamaan (3). Algoritma ini dikenal karena stabilitas dan ketahanannya terhadap overfitting [20].

$$RF\,(x) = \frac{1}{T}\sum_{t=1}^{T} Tree_t(x) \quad \text{.....(3)}$$

dimana T adalah jumlah total pohon, $Tree_t(x)$ adalah prediksi yang dibuat oleh pohon keputusan ke-t untuk input x. Prediksi akhir untuk klasifikasi ditentukan oleh suara mayoritas di semua pohon.

Model-model tersebut dilatih menggunakan 70% dataset sebagai set pelatihan, dengan 30% sisanya digunakan untuk pengujian. Metode uji terpisah ini bertujuan untuk mengevaluasi performa dan efektivitas model dalam memprediksi kelulusan tepat waktu [21]. Hasil dari model-model ini akan dibandingkan untuk mengidentifikasi model yang memberikan prediksi paling akurat.

2.5 Indikator Kinerja

Dalam studi ini, kinerja berbagai model pembelajaran mesin dievaluasi menggunakan indikator kinerja yang diakui secara luas [22]. Salah satu indikator utama yang digunakan adalah Confusion Matrix, yang memvisualisasikan hasil klasifikasi [23]. Matriks ini membandingkan hasil prediksi dengan nilai aktual, menghasilkan empat kategori: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), and *False Negative* (FN). Dengan menggunakan matriks ini, dapat dihitung metrik lainnya yaitu *Precision*, *Recall*, dan *F1-score*. Rumus untuk metrik ini adalah sebagai berikut [24]:

Accuracy adalah rasio prediksi yang benar terhadap total data sesuai dengan persamaan (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(4)}$$

dimana TP adalah *True Positives* (prediksi positif dengan benar), FP merepresentasikan *False Positives* (prediksi positif yang salah), FN merepresentasikan *False Negatives* (kejadian positif yang terlewatkan) dan TN adalah *True Negatives* (Prediksi negatif yang benar).

Precision mengukur kemampuan model untuk mengidentifikasi hanya contoh yang relevan dengan tepat. Nilai ini dihitung menggunakan persamaan (5) sebagai berikut :

$$Precision = \frac{TP}{TP+FP} \quad \text{.....(5)}$$

Recall (Sensitivity) mengukur kemampuan model untuk mendeteksi semua kejadian yang relevan. Hal ini dihitung menggunakan persamaan (6).

$$Recall = \frac{TP}{TP+FN} \quad \text{.....(6)}$$

F1-score adalah rata-rata harmonis antara presisi dan perolehan, yang menyeimbangkan trade-off di antara keduanya. Nilai ini dihitung menggunakan persamaan (7).

$$F1\text{-}Score = 2\,x\frac{Precision\ x\ Recall}{Precision+Recall} \quad \text{.....(7)}$$

Metrik ini memberikan pemahaman yang lebih detail tentang kinerja model dibandingkan dengan akurasi semata. Dengan menghitung indikator ini, berbagai model dapat dibandingkan untuk mengidentifikasi model yang paling efektif dalam memprediksi kelulusan tepat waktu mahasiswa.

2.6 *Training* dan *Validation*

Studi ini membagi dataset menjadi dua subset: 70% untuk pelatihan dan 30% untuk validasi [25]. Pembagian ini mengikuti praktik standar machine learning untuk memastikan

model dapat digeneralisasi dengan baik ke data yang belum pernah dilihat sebelumnya. Evaluasi dilakukan untuk memprediksi kelulusan tepat waktu mahasiswa program D3 Teknologi Informasi.

Model yang digunakan Adalah K-Nearest Neighbors, Decision Tree, dan Random Forest. Seluruh model dikembangkan dan dilatih menggunakan pustaka *scikit-learn* pada bahasa pemrograman Python. Parameter model dioptimalkan menggunakan GridSearchCV untuk menemukan kombinasi terbaik guna meningkatkan akurasi dan kinerja [26]. Parameter yang digunakan dalam penelitian ditunjukkan dalam Tabel 3.

Tabel 3. Parameter Model yang Digunakan dalam Penelitian.

Model	Definisi
K-Nearest Neighbors (KNN)	n_neighbors (1-50), metric (minkowski), weights (uniform)
Decision Tree	Criterion (gini), max_depth (None), random_state (42)
Random Forest	n_estimators (100), criterion (gini), random_state (42)
GridSearchCV	param_grid (C = [0.1, 1, 10, 100]; gamma = [scale, auto, 0.001, 0.01, 0.1, 1])

Kinerja model dievaluasi menggunakan metrik Accuracy, Precision, Recall, dan F1-score. Pengukuran ini memberikan gambaran tentang sejauh mana model mampu mengklasifikasikan data dengan tepat. Model dengan hasil terbaik ditentukan berdasarkan performanya pada data validasi.

Penelitian ini menggunakan 10-fold cross-validation pada tahap pelatihan, dengan membagi dataset pelatihan menjadi 10 subset [27]. Model dilatih dan divalidasi sebanyak 10 kali, kemudian hasil akhirnya dihitung sebagai rata-rata dari performa pada 10 proses pelatihan tersebut. Pendekatan ini membantu mengurangi risiko terjadinya overfitting.

3. HASIL DAN DISKUSI

Setiap model dilatih menggunakan metode *train-test split* dengan rasio 70:30, dan kinerjanya dievaluasi menggunakan metrik *Accuracy, Precision, Recall, dan F1-Score* seperti yang ditunjukkan dalam tabel 4.

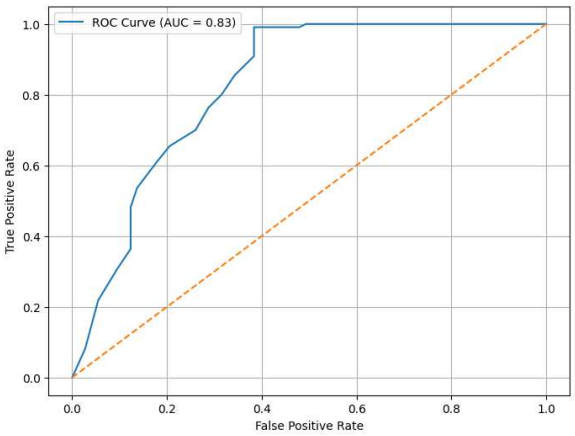
Tabel 4. Perbandingan Kinerja Model Machine Learning.

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
KNN	84%	79%	98%	87%
Decision Tree	70%	78%	72%	75%
Random Forest	73%	79%	74%	76%

Hasil penelitian menunjukkan bahwa metode K-Nearest Neighbors (KNN) memberikan kinerja terbaik, dengan akurasi sebesar 84%, precision 79%, dan recall yang sangat tinggi yaitu 98%. Hal ini menunjukkan bahwa KNN mampu mendeteksi mahasiswa yang akan lulus tepat waktu dengan sangat baik, meskipun precision sedikit lebih rendah, kemungkinan akibat adanya beberapa prediksi positif palsu. Dengan F1-Score sebesar 87%, model ini menunjukkan keseimbangan yang baik antara kemampuan mendeteksi kasus dan meminimalkan kesalahan.

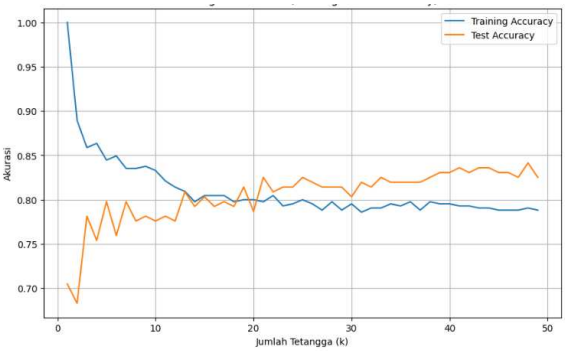
Model Decision Tree menunjukkan akurasi terendah 70% dan recall 0.72%, yang mengindikasikan bahwa model ini kurang efektif dalam mendeteksi kelulusan mahasiswa. Meskipun precision sebesar 78% masih cukup baik, nilai F1-Score sebesar 75% menunjukkan adanya ketidakseimbangan dalam mendeteksi kasus dan mengurangi kesalahan prediksi.

Model Random Forest memiliki kinerja yang lebih baik dibandingkan Decision Tree, dengan akurasi 73% dan F1-Score sebesar 76%. Precision sebesar 79% dan recall 74% menunjukkan bahwa model ini lebih stabil dan akurat dalam mengklasifikasikan kelulusan mahasiswa, meskipun masih berada di bawah performa KNN.



Gambar 3. Kurva ROC Model K-Nearest Neighbors (KNN) : Hasil pengolahan data penelitian

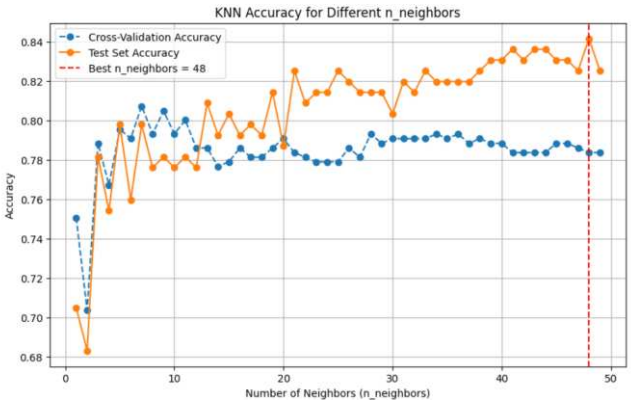
Selain menggunakan metrik akurasi, precision, recall, dan F1-Score, performa model KNN juga dianalisis menggunakan kurva *Receiver Operating Characteristic* (ROC) seperti ditunjukkan pada Gambar 3. Kurva ROC memperlihatkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai ambang keputusan. Hasil menunjukkan bahwa model KNN memiliki *Area Under the Curve* (AUC) sebesar 0.83, yang mengindikasikan bahwa model mampu membedakan dengan baik antara mahasiswa yang lulus tepat waktu dan yang tidak. Nilai AUC di atas 0.80 termasuk dalam kategori sangat baik, sehingga memperkuat hasil penelitian bahwa KNN merupakan model paling akurat untuk memprediksi kelulusan tepat waktu.



Gambar 4. Kurva Pembelajaran Model KNN terhadap Variasi Jumlah Tetangga : Hasil pengolahan data penelitian

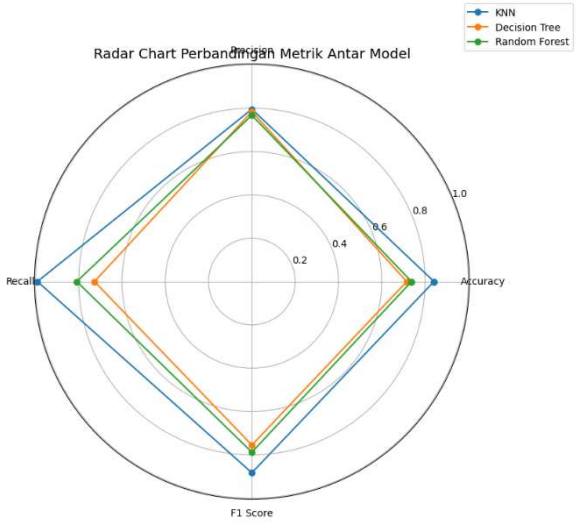
Kurva ROC yang berada jauh di atas garis diagonal (*baseline classifier*) menunjukkan bahwa model KNN tidak hanya unggul dalam mendeteksi mahasiswa yang lulus tepat waktu (sesuai nilai recall tinggi), tetapi juga memiliki kemampuan generalisasi yang baik terhadap data uji.

Learning Curve pada Gambar 4 menunjukkan perubahan akurasi model KNN pada data pelatihan dan data pengujian terhadap variasi jumlah tetangga (k). Seiring bertambahnya nilai k, akurasi pelatihan cenderung menurun dan menjadi lebih stabil, sementara akurasi pengujian meningkat dan mendekati akurasi pelatihan. Pola ini menunjukkan bahwa model mulai mencapai keseimbangan yang lebih baik antara bias dan varians. Pada rentang k sekitar 20 hingga 45, akurasi pengujian berada pada titik yang paling stabil (antara 0.81 hingga 0.84), yang mengindikasikan bahwa model memiliki generalisasi terbaik pada rentang tersebut.



Gambar 5. Akurasi Model KNN pada Berbagai Nilai Jumlah Tetangga (k) : Hasil pengolahan data penelitian

Fenomena ini selaras dengan hasil *tuning hyperparameter* yang mendeteksi bahwa nilai k optimal berada pada angka 48, di mana akurasi pengujian mencapai nilai tertinggi dibandingkan nilai k lainnya. Dengan demikian, *learning curve* mengonfirmasi bahwa KNN tidak hanya memiliki performa yang baik berdasarkan metrik evaluasi (accuracy, precision, recall, F1-score), tetapi juga menunjukkan karakteristik generalization yang kuat ketika jumlah tetangga dipilih secara optimal.



Gambar 6. Radar Chart Perbandingan Metrik Evaluasi antar Model

Radar chart pada Gambar 6 membandingkan empat metrik evaluasi utama—Accuracy, Precision, Recall, dan F1-Score—antara tiga model yang diuji, yaitu KNN, Decision Tree, dan Random Forest. Visualisasi ini menunjukkan bahwa KNN memiliki performa paling unggul secara konsisten, ditandai dengan nilai Accuracy tertinggi dan Recall yang sangat dominan, sehingga mampu mendeteksi mahasiswa yang lulus tepat waktu dengan lebih efektif. Random Forest menampilkan performa menengah dengan nilai Precision dan F1-Score yang stabil, namun masih berada di bawah KNN, sedangkan Decision Tree menunjukkan kinerja terendah, terutama pada metrik Recall, yang mencerminkan kurangnya kemampuan dalam mengidentifikasi kasus positif. Secara keseluruhan, bentuk poligon yang paling luas pada model KNN menunjukkan bahwa model ini memiliki kinerja paling seimbang dan unggul di antara ketiga model dalam memprediksi kelulusan tepat waktu.

4. KESIMPULAN DAN SARAN

Penelitian ini menerapkan berbagai model machine learning untuk memprediksi kelulusan tepat waktu mahasiswa. Di antara seluruh model yang diuji, K-Nearest Neighbors (KNN) menunjukkan performa terbaik dalam mendeteksi kelulusan mahasiswa, dengan akurasi tinggi sebesar 84% dan nilai recall yang sangat baik sebesar 98%. Model Random Forest memberikan keseimbangan yang lebih baik dibandingkan Decision Tree, dengan akurasi yang lebih tinggi dan stabilitas prediksi yang lebih baik. Secara keseluruhan, KNN merupakan pilihan terbaik untuk mendeteksi kelulusan tepat waktu, sedangkan Random Forest lebih andal dalam menangani data yang lebih kompleks.

Penelitian ini memberikan kontribusi praktis dan akademik. Secara praktis, hasil penelitian dapat dimanfaatkan oleh Program Studi D3 Teknologi Informasi, sebagai dasar pengembangan sistem peringatan dini untuk mendeteksi mahasiswa yang berisiko tidak lulus tepat waktu. Secara akademik, penelitian ini memperkaya kajian educational data mining pada konteks pendidikan vokasi di Indonesia serta menunjukkan keunggulan algoritma K-Nearest Neighbors dalam memprediksi kelulusan tepat waktu berbasis data akademik multi-semester.

Penelitian ini memiliki beberapa keterbatasan, antara lain penggunaan dataset yang terbatas pada mahasiswa Program Studi D3 Teknologi Informasi Universitas Brawijaya dengan rentang angkatan 2018–2020, sehingga generalisasi hasil masih terbatas. Selain itu, variabel yang digunakan didominasi oleh faktor akademik, sementara faktor non-akademik belum diakomodasi dalam model.

Untuk penelitian mendatang, disarankan untuk menambahkan proses seleksi fitur serta mengeksplorasi teknik machine learning lainnya guna meningkatkan akurasi prediksi. Integrasi machine learning dalam bidang pendidikan terbukti efektif dalam memprediksi capaian belajar mahasiswa. Dengan model prediksi yang kuat, institusi pendidikan dapat memberikan dukungan yang lebih terarah kepada mahasiswa, sehingga membantu mereka meraih keberhasilan akademik dan lulus tepat waktu.

4. DAFTAR PUSTAKA

[1] G. Bandiera, J. Frank, F. Scheele, J. Karpinski, and I. Philibert, “Effective accreditation in postgraduate medical education: from process to outcomes and

back," *BMC Med Educ*, vol. 20, no. Suppl 1, p. 307, 2020.

[2] S. L. DesJardins, D. A. Ahlburg, and B. P. McCall, "A temporal investigation of factors related to timely degree completion," *J Higher Educ*, vol. 73, no. 5, pp. 555–581, 2002.

[3] A. Sarra, L. Fontanella, and S. Di Zio, "Identifying students at risk of academic failure within the educational data mining framework," *Soc Indic Res*, vol. 146, no. 1, pp. 41–60, 2019.

[4] D. Bañeres, M. E. Rodríguez, A. E. Guerrero-Roldán, and A. Karadeniz, "An early warning system to detect at-risk students in online higher education," *Applied Sciences*, vol. 10, no. 13, p. 4427, 2020.

[5] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," *Journal of Engineering and Applied Technology*, vol. 2, no. 2, pp. 86–91, 2021.

[6] A. Hartono, L. A. Dewi, E. Yuniarti, S. T. H. Putri, and T. S. Harahap, "Machine learning classification for detecting heart disease with K-NN algorithm, decision tree and random forest," *Eksakta: Berkala Ilmiah Bidang MIPA*, vol. 24, no. 4, pp. 513–522, 2023.

[7] M. N. Yakubu and A. M. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," *Kybernetes*, vol. 51, no. 2, pp. 916–934, 2022.

[8] L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting university student graduation using academic performance and machine learning: a systematic literature review," *IEEE Access*, vol. 12, pp. 23451–23465, 2024.

[9] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, 2020, p. 32019.

[10] A. Merghadi and others, "Machine learning methods for landslide susceptibility studies," *Earth Sci Rev*, vol. 207, p. 103225, 2020.

[11] D. Valero-Carreras and others, "Comparing two SVM models through different metrics," *Comput Oper Res*, vol. 152, p. 106131, 2023.

[12] A. P. Salim, K. A. Laksitowening, and I. Asror, "Time series prediction on college graduation using kNN algorithm," in *Proceedings of the 8th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2020, pp. 1–4.

[13] H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi, "Predicting students' performance using machine learning techniques," *Journal of University of Babylon for Pure and Applied Sciences*, vol. 27, no. 1, pp. 194–205, 2019.

[14] Desfiandi and Soewito, "Student graduation time prediction using logistic regression, decision tree, support vector machine, and AdaBoost ensemble learning," 2023.

[15] S. Junaidi, R. V Anggela, and D. Kariman, "Classification data mining methods for on-time graduation prediction," *Journal of Applied Computer Science and Technology*, vol. 5, no. 1, pp. 109–119, 2024.

[16] M. I. Hasbullah and V. Yasin, "Prediksi kelulusan mahasiswa menggunakan machine learning," *Jurnal Teknologi Informasi dan Komunikasi*, vol. 16, no. 2, pp. 1–10, 2025.

[17] I. Saputra and others, "Integration of Artificial Intelligence in Education: Opportunities, Challenges, Threats and Obstacles," *Indonesian Journal of Computer Science*, vol. 12, no. 4, 2023.

[18] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, 2019, pp. 1255–1260.

[19] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: a review," *Procedia Manuf*, vol. 35, pp. 698–703, 2019.

[20] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst Appl*, vol. 134, pp. 93–101, 2019.

[21] Q. H. Nguyen and others, "Influence of data splitting on performance of machine learning models," *Math Probl Eng*, vol. 2021, p. 4832864, 2021.

[22] D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, and R. A. Abougalala, "Prediction of student exam performance using data mining classification algorithms," *Educ Inf Technol (Dordr)*, vol. 29, pp. 21621–21645, 2024, doi: 10.1007/s10639-024-12619-w.

[23] M. Yağcı, "Educational data mining: prediction of students' academic performance," *Smart Learning Environments*, vol. 9, p. 11, 2022, doi: 10.1186/s40561-022-00192-z.

[24] R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and f1 score," in *Workshop on Evaluation and Comparison of NLP Systems*, 2020, pp. 79–91.

[25] Brilliance and others, "Data splitting in machine learning models with stratified sampling," *Brilliance: Jurnal Ilmu Komputer dan Rekayasa*, vol. 5, no. 2, 2025.

[26] D. P. Mishra, H. K. Gupta, G. Saajith, and R. Bag, "Optimizing heart disease prediction model with GridSearchCV," in *International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, IEEE, 2024, pp. 1–6.

[27] S. M. Malakouti, "Improving the prediction of wind speed and power production using ensemble method," *Case Studies in Chemical and Environmental Engineering*, vol. 8, p. 100351, 2023.