Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 5 No. 2 (2026) pp: 8127-8135
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Prediksi Kelulusan Tepat Waktu Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (K-NN)

Tasya Parmi, Fitri Yunita
Program Studi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Islam Indragiri
tasyaparmii@gmail.com, fitriyun@gmail.com

Abstrak

Memprediksi ketepatan waktu kelulusan mahasiswa merupakan kebutuhan yang sangat penting namun belum banyak dieksplorasi dalam tata kelola pendidikan tinggi, terutama di perguruan tinggi yang mana keterlambatan penyelesaian studi menimbulkan dampak merugikan terhadap status akreditasi dan efektivitas institusi. Penelitian ini berupaya mengoperasionalkan algoritma K-Nearest Neighbor (K-NN) sebagai pendekatan klasifikasi selektif untuk memprediksi apakah mahasiswa yang terdaftar di program Sistem Informasi Universitas Islam Indragiri akan lulus dalam batas waktu yang ditentukan. Dengan memanfaatkan data 153 catatan mahasiswa yang mencakup nilai IPK, semester dan SKS dataset tersebut melalui prapemrosesan yang terstruktur yang mencakup pembersihan data dan normalisasi Min-Max Scaling, dilanjutkan dengan pemisahan bertingkat menjadi subset pelatihan dan pengujian dengan rasio 90:10 melalui 10-fold Cross Validation. Eksperimen membuktikan bahwa k = 7 menghasilkan keseimbangan klasifikasi yang paling optimal. Penilaian evaluatif melalui perincian Confussion Matrix dan analisis Kurva ROC membuktikan keakuratan selektif model, yang menghasilkan akurasi keseluruhan sebesar 99,33%, recall 100% untuk kelompok lulusan tepat waktu, presisi 98,33%, dan skor F1 sebesar 99,09% , Area Under the Curve (AUC) mencapai nilai 0,990, yang mengonfirmasi potensi klasifikasi yang kuat dan ambang batas "sangat baik". Temuan ini menegaskan bahwa K-NN merupakan mekanisme yang efisien secara komputasi dan dapat diandalkan secara epistemologis untuk memprediksi risiko kegagalan kelulusan sejak dini, sehingga memberikan landasan informasi yang dapat ditindaklanjuti kepada pengelola akademik guna melakukan intervensi yang tepat waktu dan meningkatkan indikator akreditasi institusi.

Kata Kunci: *Data Mining, K- Nearest Neighbor, Kelulusan Mahasiswa, Prediksi, RapidMiner*

1. Latar Belakang

Intuisi pendidikan tinggi mempunyai prean yang cukup strategis dalam membentuk kualitas suber daya manusia sutau bangsa. Jenis-jenis intuisi pndidikan tinggi di Indoensia mencakup universitas, institutt, sekolah tinggi, akademi, dan politeknik. Keberlangsungan dan kualitas sebuah institusi pendidikan sangat ditentukan olh mehasiswanya, yang merupakan modal utama dalam proses akademik [1]. Di era digitalisasi sekarnag, ledakan volume data akademik telah menjadi tantangan yang cukup serius terlebih dihadapi oleh institusi pendidikan, terutama dalam mengolah dan menginterprestasi data tersebut menjadi pengetahuan yang bermanfaat bagi sivitas akademika dan pengelola institusi [2]. Prediksi keberhasilan akademik semakin menarik perhatian di bidang pendidikan. Prestasi akademik yang baik dapat meningkatkan peringkat universitas [3]. Status kelulusan mahasiswa telah menjadi salah satu prioritas pengelolaan data dan investasi yang paling menantang dan penting bagi para pendidik. Banyak penelitian menunjukkan bahwa dampak atau efektivitas sistem pendidikan dapat ditentukan dengan menganalisis tingkat kelulusan siswa [4]. Kelulusan selalu diukur dengan IPK (indeks Prestasi Kumulatif), SKS, semester yang ditempuh dan lainnya sesuai dengan syarat tertentu dari suatu institusi yang sudah disahkan. Nilai yang lebih rendah kerap menjadi tolak ukur yang mengakibatkan kelulusan yang terlambat.

Dalam sistem penjaminan mutu pendidikan tinggi Indonesia, salah satu hal penilaian akrediktasi yng paling penting adaalah proporsi mahasiswa yang berhasil menyelesaikan studi sesuai dengan batas waktu yang telah ditetapkan [5]. Badan Akreditasi Nasional menetapkan bahwasannya ma masa stusi ideal mahasiswa asalah empat tahun, sehingga kemampuan untuk mempiksi status kelulusan sejak awal menjadi suatu kebutuhan penting. Tingginya angka mahasiswa yang melampaui batas waktu studi tidak hanya berdampak pada penurunan nilai akreditasi program studi, tetapi jug amencerminkan inefsiensi daalam pengelolaan akademik. Meninjau besarnya volume data yang harus dikelola berbagai algoritma klasifikasi telah dikembangakan untuk memopang pemrosesan data yang besar di ruang lingkup pendidikan [6]. Oleh karenanya, penerapan teknik predikstif berbasis Data Mining menjadi solusi yang baik dan relevan menekan risiko keterlambatan kelulusan, sekaligus menyediakan informasi strategis yang sekiranya akan sangat bermanfaat oleh pengelola akdemik dan mahasiswa itu sesndiri.

EDM (Educational Data Mining) ataucabang ilmu yang berfokus pada transformasi data mentah yang ekosistem pensisikan menjadi pengetahuan bermakna yang dapat dimanfaatkan untuk berbagai pihak, mulai dari pengambil kebijakan, tenaga pendidik, sampai mahasiswa itu sendiri, dalam rangka pengambilan keputusan yang lebih baik sasaran terkait pengelola sumber daya akademik. Sebagai inti dari proses ini, Data Mining fungsinya adalah sebagai mekanisme untuk mengetahui pola-pola tersembunyi yang terdapat dalam kumpulan data berskala besar [7]. Dalam konteks pendidiakan, Data Mining mencakup berbagai pendekatan analisis seperti prdiksi hasil belajar, identifikasi hubungan antar variabel akademik, klasifikasi kelompok mahasiswa, penggalian pola tersembunyi dalam model, serta reduksi kompleksitas data guna mendukung interprestasi yang lebih mudah dipahami [8]. Penggunaan metode ini bertujuan untuk menghasilkan prediksi yang andal demi mendukung intervensi dini bagi mahasiswa yang berisiko terlambat lulus. Sejumlah studi terdahulu telah mengeksplorasi berbagai pendekatan, di antaranya penelitian Dinda Safitri yang mengimplementasikan algoritma K-NN (K-Nearest Neighbor) melalui platform Orange Data Mining dan berhasil membuktikan tingkat akurasi yang memuaskan [9]. Kinerja algoritma K-NN dipengaruhi secara signifikan oleh penetapan nilai K serta karakteristik permasalahan yang dihadapi. Studi lain yang relevan berupaya membangun sistem prediksi kelulusan dengan memanfaatkan perhitungan jarak Euclidean sebagaii dasar klasifikasi [10]. K-NN mengklasifikasikan data baru berdasarkan kemiripannya dengan rekam data historis, menggunakan representasi numeric (angka) yang paling mendekati satu sama lain. Prediksi atau peramalan merupakan suatu perhitungan untuk meramalkan keadaan di masa yang akan datang dengan melakukan pengujian terhadap keadaan di masa lalu. Salah satu fungsi prediksi adalah untuk membantu dalam pengambilan keputusan sehingga resiko bisa ditekan seminimal mungkin. algoritma K-NN bisa mengenali kelulusan mahasiswa pada permasalahan terkini dengan metode mengadopsi pemecahan dari permasalahan yang mempunyai kedekatan dengan permasalahan terkini [11] dan model ini dapat mengukur kualitas informasi yang diperoleh secara lebih objektif [12]. Pada penelitian yang dulu berjudul "Penerapan Algoritma K-Nearest Neighbor (K-NN) untk memprediksi kelulusan mahasiswa menggunakan RapidMiner" hasil dari algoritma K-NN yang memiliki akurasi tinggi dan memberikan range pada klasifikasi yang digunakan untuk perhitungan [13]. Meskipun demikian, kajian tersebut menggunakan dataset mahasiswa yang bersifat umum tanpa pembatasan program studi tertentu, sehingga generalisabilitasnya pada konteks institusional yang spesifik masih terbatas.

Penelitian ini mengatasi kekurangan tersebut dengan membatasi cakupan analisis pada satu program studi di Universitas Islam Indragiri yaitu prodi Sistem Informasi sehingga memungkinkan pengembangan model yang lebih terperinci dan dapat diterapkan secara lebih baik di lingkungan intuisi. Pada penelitian kali ini metode prediksi K-NN (K-Nearest Neighbor) akan mengambil sample mahasiswa program studi sistem informasi dengan harapan dapat memberi peringatan dini, mendukung keputusan akademik, meningkatkan akreditasi institusi dan perguruan tinggi serta yang paling penting membantu prediksi kelulusan mahasiswa guna mengambil langkah untuk meningkatkan lulusan yang tepat waktu.

2. Metode Penelitian

Studi ini dirancang menggunakan paradigma kuantitatif melalui pendekatan Data Mining dengan algoritma K-Nearest Neighbor (KNN) yang dijalankan pada Machine Learning RapidMiner. Disamping itu, diperlukan kerangka kerja penelitian yang menggambarkan proses yang sistematis, sehingga tujuan dari penelitian dapat dicapai dan berjalan dengan baik. Kerangka kerja dalam penelitian ini ditunjukkan pada Gambar 1.

Gambar 1. Kerangka Penelitian

DOI: https://doi.org/10.31004/riggs.v5i2.10092

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini bersumber dari administrasi Program Studi Sistem Informasi, dengan seluruh identitas mahasiswa telah dianonimkan untuk menjaga privasi. Rekam data tersebut mencerminkan hasil evaluasi akademik yang dilakukan oleh dosen pengampu sepanjang semester berjalan. Setelah data tersedia secara lengkap, tahap berikutnya adalah penentuan variabel-variabel yang akan dijadikan fitur dalam model prediksi.

2.2 Prapemrosesan Data

Setelah pengumpulan, data yang diperoleh melewati tahap prapemrosesan yang bertujuan untuk menyiapkan dataset agar bebas dari anomali dan inkonsistensi. Proses ini mencakup seleksi variabel yang relevan serta eliminasi data duplikat agar dataset yang digunakan benar-benar bersih dan representatif [14]. Dengan demikian, dataset akhir yang digunakan untuk analisis telah melalui proses pembersihan yang memadai.

2.2.1 Data Cleaning

Tahap pembersihan data merupakan langkah fondasi yang tidak dapat dilewati sebelum proses analisis dilakukan, mengingat keakurtan suatu model danagat bergantung pada tinggi kualitas data masukan yang digunakan. Proses ini bersifat interatif karna strategi pembersihan yang diterapkan akan menyesuaikan diri dengan karakteristik spesifik dataset dan kebutuhan algoritmik dari metode *Machine Learning* yang dipilih.

2.2.2 Normalisasi Data

Tahap normalisasi dilakukan untuk mengubah nilai atribut dalam dataset ke dalam skala yang seragam tanpa mengubah informasi inti dari data tersebut karna K-NN adlah algoritma yang keputusannya ditentukan oleh rentang data. Di penelitian ini normalisasi menggunakan metode Min-Max Scaling. proses ini terutama bergantung pada nilai maksimum dan minimum dalam atribut tersebut. Pada metode normalisasi ini, nilai komponen dari sampel data asli akan ditransformasikan ke dalam rentang 0-1. Rumus normalisasi data akan ditampilkan pada rumus 1.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

X adalah sample data asli, X min merupakan nilai terkecil pada atribut dan X max adalah nilai tertinggi atribut pada data yang akan dinormalisasi. Tujuan utama normalisasi ini adalah untuk memastikan bahwa semua atribut data memiliki skala yang seragam, sehingga menghindari dominasi atribut dengan skala besar terhadap atribut lainnya. Dengan menskalakan ulang rentang nilai data, interpretasi hasil analisis menjadi lebih konsisten, dan kinerja model klasifikasi dapat ditingkatkan [15].

2.3 Pembagian Data

Dataset yang telah melalui tahap prapemrosesan selanjutnya dibagi menggunakan teknik Cross Validation dengan skema 10-fold. Dalam konfigurasi ini, 90% data difungsikan sebagai data latih yang membentuk basis referensi historis, sementara 10% sisanya dijadikan data uji untuk mengevaluasi generalisasi model. Setiap fold diproses secara bergiliran sehingga seluruh data berkesempatan menjadi data uji, dan hasil akhir diperoleh melalui agregasi performa klasifikasi algoritma K-NN pada semua fold tersebut.

2.4 K-Nearest Neigbor

Dalam penelitian ini, algoritma K-NN diimplementasikan sebagai inti dari sistem klasifikasi berbasis Data Mining. Cara kerjanya adalah mengidentifikasi K titik data terdekat di dalam ruang fitur, kemudian menggunakan label mayoritas dari titik-titik tersebut untuk menghasilkan prediksi terhadap data baru yang belum diketahui kelasnya. Salah satu algoritma klasifikasi terawasi yang paling banyak diteliti dan paling dikenal di bidang Machine Learning adalah aturan *k*-Nearest Neighbor (*k*NN).. Algoritma K-NN termasuk dalam paradigma Lazy Learning, yang berarti bahwa algoritma ini tidak memerlukan proses pelatihan untuk membangun model yang sesuai dengan kumpulan data, melainkan langsung menggunakan sampel dari kumpulan data untuk melakukan perbandingan dengan elemen kueri yang akan diklasifikasikan [16]. Dengan menggunakan perangkat lunak RapidMiner untuk implementasi yang dalam proses ini, terdapat 3 operator: Pembaca data eksternal dari Microsoft Excel Read, , Normalisasi, dan K-NN untuk implementasi klasifikasi, dan Validasi. Setelah membuat model Klasifikasi K-NN di operator Validasi [17].

2.4.1 Menentukan Nilai K

Dalam K-Nearest Neighbors (KNN), ketentuan nilai parameter k berdasar di pertimbangan yang berhubungan dengn kebutuhan khusus, yang mana diarahkan untuk memperoleh dan mendapatkan hasil peramalan yang optimal

DOI: https://doi.org/10.31004/riggs.v5i2.10092

dan relevan. Pada konteks prediksi status kelulusan mahasiswa, pendekatan ini dilakukan dengan memilih mayoritas label tetangga terdekat. Pada penentuan nilai K tentu memiliki batas maksimal K yang diambil berikut akan di tampilkan pada rumus 2.

$$K = \sqrt{n} \qquad (2)$$

K adalah jumlah tetangga terdekat dan n sebagai jumlahkeseluruahan data yang digunakan dalam perhitungan.

2.4.2 Menghitung Jarak Euclidean

K dtentukan berdasarkan nilai yang paling optimal dan tinggi persenannya. Nilai K yang optimal dihasilkan melalui percobaan yang melibatkan perhitungan rentang jarak antara setiap objek dan data-data yang disediakan.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (3)$$

Berbasis rumus di atas, di mana matriks ED itu jarak skala antara dua besaran vektor x (data pelatihan) dan y (data uji) dari sebuah matriks dengan dimensi tertentu. Kelompokkan hasil objek-objek ini ke dalam kelompok-kelompok dengan jarak Euclidean terkecil. Kelompokkan kategori-kategori y (klasifikasi Tetangga Terdekat berdasarkan nilai k) dengan menggunakan kategori tetangga yang paling umum, sehingga kategori objek tersebut dapat diprediksi.

2.5 Evaluasi

Di penelitian ini, penilaiannya akan menggunakan matriks kebingungan (confusion matrix), yang digunakan untuk menjelaskan hasil dari model klasifikasi dengan cara mengadu prediksi model dengan nilai aktual di data uji. matriks kebingungan juga dipakai buat mengukur kinerja metode klasifikasi. Confusion matrix bisa akan beri info detail soal kinerja model KNN dan hasil klasifikasi yang akurat. Tabel confusion matrix akan ditampilkan pada Tabel 1. Confusion marix dan destail perhitungan pada rumus 4,5, 6 dan 7 sebagai berikut.

Tabel 1. Confussion Matrix

Clasification	Classification	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\backslash\% \qquad (4)$$

$$Precision = \frac{TP}{TP+FP} \qquad (5)$$

$$Recall = \frac{TP}{TP+FN} \qquad (6)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \qquad (7)$$

Positif Benar (TP): data yang secara benar diklasifikasikan menjadi positif; Negatif Benar (TN): data yang secara benar diklasifikasikan menjadi negatif; Positif Palsu (FP): data yang secara keliru diklasifikasikan menjadi positif; dan Negatif Palsu (FN): data yang secara keliru diklasifikasikan menjadi negatif. Pada confussion matrix terdapat perhitungan nilai akurasi yang menunjukkan kemampuan prediktif yang kuat., recall yang menunjukkan bahwa data positif yang benar-benar positif berhasil diidentifikasi oleh model, presisi dan f-measure menunjukkan kinerja klasifikasi yang seimbang berikut akan ditampilkan rumusnya.

2.5.1 ROC

ROC curve (Receiver Operating Characteristic) merupakan salah satu metode untuk menganalisis dan mengevaluasi berbagai aspek dari sebuah metode klasifikasi, seperti akurasi, kecepatan, keandalan, skalabilitas, dan interpretabilitas. Analisis ini kemudian divisualisasikan dalam bentuk grafik untuk memudahkan pemahaman dan interpretasi hasil evaluasi tersebut. Kurva ROC banyak digunakan untuk menggambarkan seberapa baik model kerentanan terhadap kesalahan (misalnya, model probabilitas) mengklasifikasikan modul perangkat lunak sebagai bermasalah atau tidak bermasalah. AUC, yaitu Area di Bawah Kurva ROC, biasanya digunakan untuk mengukur daya pembeda keseluruhan dari suatu model kerentanan terhadap kesalahan. Sebaliknya, metrik kinerja seperti Precision, F-score, Accuracy, dan sebagainya digunakan untuk mengevaluasi suatu klasifikasi tertentu. Metrik-metrik ini didefinisikan sebagai fungsi dari Confusion Matrix, yang berisi jumlah modul yang rusak dan tidak

DOI: https://doi.org/10.31004/riggs.v5i2.10092

rusak yang diperkirakan dengan benar dan salah. Dengan demikian, setiap titik pada kurva ROC juga sesuai dengan Confusion Matrix [18].

3. Hasil dan Diskusi

Dalam prediksi kelulusan mahasiswa Sistem Informasi Universitas Islam Indragiri dengan metode K-Nearest Neigbor peneliti menggunakan 153 data mahasiswa yang didapatkan dari admin data Sistem Informasi yang terdiri dari beberapa kategori yaitu IPK, jumlah SKS yang diambil, semester yang ditempuh, dan status kelulusan sebagai variabel prediksi dengan 2 kelas yaitu lulus tepat waktu dan lulus tidak tepat waktu.

3.1 Pengumpulan Data

Setelah dilakukan pengumpulan data dengan ketentuan dan syarat pada pihak terkait administrasi bidang kampus didapatkan total data mahasiswa sebanyak 153 angkatan 2020-2021 sebagai data historis yang nantinya akan dijadikan data latih sebagai acuan prediksi penelitian. Berikut dituliskan beberapa data yang didapatkan pada Tabel 2 berikut.

Tabel 2. Data Mahasiswa

NIM	SKS	IPK	Semester	Status
403201010001	150	3,73	8	Lulus Tepat Waktu
403201010002	142	3,32	12	Lulus TidakTepat Waktu
403201010003	150	3,22	12	Lulus Tidak Tepat Waktu
403201010004	141	3,50	8	Lulus Tepat Waktu
403201010005	150	3,61	8	Lulus Tepat Waktu
403201010006	138	3,46	8	Lulus Tepat Waktu

3.2 Preprocessing Data

Pada tahap ini berfokus pada keseimbangan data untuk mengurangi noise, data kekosongan dan juga kesalahan agar tidak terjadi kesalahan dalam perhitungan prediksi.

3.1 Data Cleaning (pembersihan data)

Pada tahap ini program akan membersihkan data dan menghapus secara otomatis data yang tidak bisa terbaca atau data kosong untuk kelancaran dalam menghitung hasil pada tahap K-Nearest Neighbor pada data mahasiswa. Data yang digunakan memiliki 1 data dengan keterangan K (Keluar) dan 2 data kosong yang otomatis akan dihapus sehingga dalam penelitian ini menggunakan 150 data mahasiswa.

3.2 Normalisasi

Tahap ini bertujuan untuk menyeimbangkan data dan mengubah data SKS, IPK, dan semester menjadi angka yang berkisar dari 0-1 agar tidak terjadi ketimpangan ataupun kesalahan dalam perhitungan. Berikut contoh perhitungannya.

$$x' = \frac{X - X_{min}}{X_{max} - X_{min}} = \frac{150-0}{159-0} = \frac{150}{159} = 0.94339$$

Hasil akan ditampilkan sebagai contoh perhitungan dari normalisasi yang akan ditunjukkan pada Tabel 3 berikut.

Tabel 3. Contoh Normalisasi Min-Max Scalling

SKS	IPK	Semester	SKS	Normalisasi SKS	Normalisasi Semester	Normalisasi IPK
150	3,73	8	150	0,943396226	0,00	0,9325
142	3,32	12	142	0,893081761	1,00	0,83
150	3,22	12	150	0,943396226	1,00	0,805
141	3,50	8	141	0,886792453	0,00	0,875
150	3,61	8	150	0,943396226	0,00	0,9025
138	3,46	8	138	0,867924528	0,00	0,865

3.3 Pembagian Data

Pada 150 data mahasiswa akan dilakukan splitting/pembagian data secara random untuk memisahkan data training sebagai data historis acuan untuk perhitungan yang sudah diketahui statusnya dan data testing/latih sebagai data yang akan dihitung dan tidak diketahui statusnya dengan perbandinngan 90:10 yaitu metode untuk mengevaluasi kinerja dari suatu model atau algoritma dengan membagi data menjadi beberapa bagian, kemudian melatih dan

DOI: https://doi.org/10.31004/riggs.v5i2.10092

menguji model tersebut secara berulang-ulang. Data diambil secara random untuk memisahkan data training sebagai data historis acuan untuk perhitungan yang sudah diketahui statusnya dan data testing/latih sebagai data yang akan dihitung dan diketahui statusnya dengan perbandinngan 90:10 yang mana ini akan diuji dengan 10 kali fold validasi dengan menguji semua data tmetode ini dikenal juga dengan cross validation.

Tabel 4. Contoh Pembagian data

NIM	SKS	IPK	Normalisasi Semester	Status	Keterangan
403201010001	0,943396226	0,9325	0,00	Lulus Tepat Waktu	training
403201010002	0,893081761	0,83	1,00	Lulus TidakTepat Waktu	training
403201010003	0,943396226	0,805	1,00	Lulus Tidak Tepat Waktu	training
403201010004	0,886792453	0,875	0,00	Lulus Tepat Waktu	training
403201010005	0,943396226	0,9025	0,00	Lulus Tepat Waktu	training
403201010006	0,867924528	0,865	0,00	Lulus Tepat Waktu	training
Data Uji					
403211010087	0,962264151	8625	1,00	?	testing

Pada Tabel 4 dijelaskan bahwa data training adalah 90% kumpulan data yang statusnya sudah diketahui dan merupakan data historis yang sudah melewati beberapa pengujian. Dan data testing/latih adalah 10% dari dari semua data yang statusnya tidak diketahui dan akan dijadikan bahan penelitian prediksi.

3.4 K-Nearest Neighbor

Prosedur K-Nearest Neighbor (K-NN) dalam sistem inilah langkah awal yang harus dilakukan di metode ini, nilai K (dari jumlah titik data/tetangga terdekat) diputuskan oleh peneliti. Kesalahan dalam memilih nilai K dapat mengubah keakuratan prediksi.

3.4.1 Menentukan Nilai K

Dalam metode K-Nearest Neighbors (KNN), penentuan nilai parameter k didasarkan pada pertimbangan yang berkaitan dengan kebutuhan khusus, yang diarahkan untuk memperoleh hasil peramalan yang optimal. Pada konteks peramalan status kelulusan, pendekatan ini dilaksanakan dengan memilih mayoritas label tetangga terdekat.

$$K = \sqrt{n} = \sqrt{150} \approx 12.247$$

Batas maksimal dari K pada 150 data 12,247= 13, setelah itu akan dicari dari 1-13 nilai K mana yang memiliki akurasi tertinggi yang akan ditampilkan pada tabel 5 berikut.

Tabel 5. Perhitungan Nilai K

Jumlah K	Persentasi *Accuracy*
K1-K6	98,67%
K7-13	99,33%

Pada Tabel 5. Dapat diketahui bahwa K-NN bergantung pada nilai Accuracy nilai K sehingga persentase tertinggi merupakan nilai K terbaik. Dan nilai K terbaik adalah K=7 dengan Accuracy 99,33%. Nilai akurasi yang serupa yang diperoleh untuk K = 1 sampai K = 6 menunjukkan pola klasifikasi yang stabil dalam dataset. Akurasi yang lebih tinggi yang dicapai untuk K = 7 hingga K = 20 menunjukkan bahwa ukuran lingkungan yang moderat meningkatkan ketahanan model terhadap noise.

3.4.2 Menghitung Jarak Euclidean

Menghitung jarak ini menentukan nilai untuk ketetanggan dengan menghitung variabel yang digunakan yang nantinya akan di diambilbeberapa berdasarkan jumlah K yang diputuskan untuk prediksi. Berikut contoh perhitungannya.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$d = \sqrt{(0{,}943396226 - 0{,}893081761)^2 + (0{,}9325 - 0{,}83)^2 + (0.00 - 1{,}00)^2}$$

$$d = 1{,}006497787$$

Perhitungan selanjutnya akan mengetahui jarak ketetanggaan atau *Eulidean Distance* pada data mahasiswa.

DOI: https://doi.org/10.31004/riggs.v5i2.10092

Sebagai contoh akan ditampilkan beberapa hasil perhitungan Eulidean Distance pada Tabel 6.

Tabel 6. Contoh Hasil *Eulidean Distance*

Normalisasi SKS	Normalisasi IPK	Normalisasi semester	*Eulidean Distance*
0,943396226	0,9325	0,00	1,006497787
0,893081761	0,83	1,00	0,056183142
0,943396226	0,805	1,00	1,004043817
0,886792453	0,875	0,00	0,062930415
0,943396226	0,9025	0,00	0,084274713
0,867924528	0,865	0,00	0,075513093

Pada Tabel 5. Setelah dilakukan perhitungan jarak maka selanjutnya mengurutkan hasil berdasarkan Eulidean Distance dari angka yang terkecil ke angka terbesar. Hal ini dilakukan untuk melihat jarak mana yang paling dekat untuk menentukan hasil ketetanggaan seperti yang sudah dilakukan pada tabel di Tabel. 3 di atas Penyusunan kategori Y (klasifikasi tetangga terdekat) Setelah mendapat peringkat sesuai jarak setiap titik data, langkah berikut adalah menyusun kategori Y (klasifikasi tetangga terdekat). Di langkah 1, parameternya (k) ditentukan 7. Didasarkan pada perhitungan KNN, ini hasil klasifikasinya pada Tabel 7 sebgai berikut.

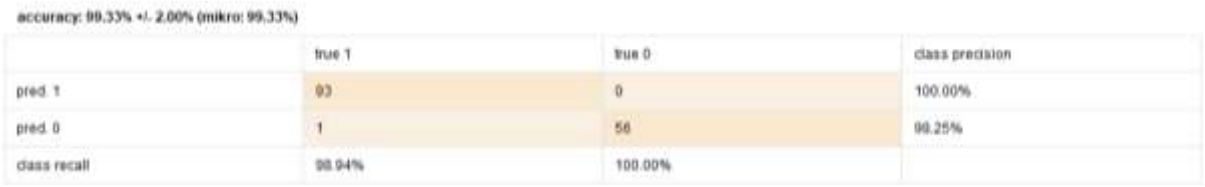
Tabel 7. Hasil Klasifikasi

NIM	SKS	IPK	Normalisasi Semester	Status	ED	No
403211010002	0,886792	0,875	1,00	lulus tepat waktu	0,056183142	1
403211010004	0,943396	0,805	0,00	lulus tidak tepat waktu	0,062930415	2
403211010006	0,867925	0,865	0,00	lulus tepat waktu	0,075513093	3
403211010005	0,893082	0,83	0,00	lulus tidak tepat waktu	0,084274713	4
403201010007	0,943396	0,8625	0,00	lulus tepat waktu	1,000377928	5
403211010003	0,943396	0,9025	1,00	lulus tepat waktu	1,004043817	6
403211010001	0,943396	0,9325	0,00	lulus tepat waktu	1,006497787	7

Seperti yang ditunjukkan pada Tabel, sesudah mengelompokkan kategori Y ke dalam K=7, hasil uji terhadap 150 titik data training menunjukkan bagian terbanyak data diklasifikasikan menjadi "lulus tepat waktu," dan hanya dua titik data yang diklasifikasikan sebagai "lulus terlambat." Tabel 7 menyajikan contoh sebagian besar hasil klasifikasi tersebut.

3.5 Pengujian Model

Pada pengujian ini akan menggunakan confusion matrix sebagai pengujian model dengan menggunakan kurva ROC



Gambar 2. Confusion Matrix

Dari hasil perhitungan confusion matrix pada Gambar.2 Confusion Matrix didapatkan Accuracy dengan nilai tinggi yaitu 99,33% dengan hasil setiap fold berada pada rata-rata 2,00% dan TP dengan jumlah 93 FP 0, TN 56 dan FN 1. Berikut dijelaskan perhitungan Accuracy dengan rumus dan perhitungan yang telah ditentukan

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{93+56}{93+56+0+1} = \frac{149}{150} = 0{,}9933 \times 100\% = 99{,}33\%$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{93}{93+0} = \frac{93}{93} = 1 \times 100\% = 100\%$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{93}{93+1} = 0{,}9894 \times 100\% = 98{,}94\%$$

DOI: https://doi.org/10.31004/riggs.v5i2.10092

$$\text{F1/F-Measure} = 2\times \frac{\text{Presicion} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{100 \times 98{,}94}{100 + 98{,}94} = 2\times \frac{9.894}{198{,}94} = 49{,}73 = 99{,}46\ \%$$

3.6 Kurva ROC

Analisis Receiver Operating Characteristic (ROC) menunjukkan model klasifikasi K-Nearest Neighbor (K-NN) mencapai nilai Area di Bawah Kurva (AUC) sebesar 0,990 ± 0,030. Hasil ini menunjukkan kemampuan diskriminatif yang tinggi dalam membedakan antara mahasiswa yang lulus tepat waktu dan yang tidak. Berikut akan ditampilkan hasil dari RapidMiner dalam memvisualisasikan kurva ROC.

Gambar 3. ROC Curve

Pada Gambar 3. ROC Curve/ Kurva ROC menunjukkan bahwa kurva ROC berada jauh di atas garis dasar klasifikasi acak, yang menunjukkan bahwa model ini memiliki kinerja prediktif yang kuat serta keseimbangan yang baik antara sensitivitas dan spesifisitas. Menurut standar klasifikasi yang diterima secara luas, nilai AUC sebesar 0,090 termasuk dalam kategori “baik” dan berada pada ambang batas kinerja klasifikasi “sangat baik”. Oleh karena itu, model K-NN yang diusulkan dapat dianggap andal dan efektif untuk memprediksi hasil kelulusan mahasiswa.

4. Kesimpulan

Hasil temuan penelitian ini mengonfirmasi bahwa algoritma K-Nearest Neighbor (K-NN) terbukti efektif sebagai alat prediksi ketepatan waktu kelulusan mahasiswa, diuji pada 153 data dari Program Studi Sistem Informasi Universitas Islam Indragiri. Akurasi keseluruhan yang diraih sebesar 99,33% (±2,00%) mencerminkan kinerja klasifikasi yang sangat baik dan konsisten di seluruh fold validasi. Disamping itu, analisis confusion matrix menunjukkan bahwa model klasifikasi prediksi ini berhasil mengidentifikasi semua mahasiswa yang lulus tepat waktu, sehingga menghasilkan nilai recall sebesar 98,94%, dengan tetap mempertahankan angka presisi tinggi sebesar 100 % dan skor F1 sekitar 99,46%. Evaluasi menggunakan kurva Receiver Operating Characteristic (ROC) menghasilkan nilai AUC sebesar 0,990 (±0,030), yang menempatkan model dalam kategori kinerja "sangat baik" berdasarkan standar klasifikasi yang berlaku umum. Jarak yang signifikan antara kurva ROC dan garis referensi acak secara visual menegaskan kapasitas diskriminatif model dalam membedakan mahasiswa yang berpotensi lulus tepat waktu dari yang tidak. Dengan demikian, secara keseluruhan dapat disimpulkan bahwa K-NN merupakan metode yang tepat dan efisien untuk memodelkan prediksi kelulusan dalam konteks ini. Dari perspektif praktis, model ini berpotensi difungsikan sebagai sistem pendukung keputusan akademik yang membantu institusi dalam memantau trajektori studi mahasiswa dan melakukan intervensi sejak dini terhadap mereka yang teridentifikasi berisiko terlambat lulus. Terlepas dari hasil yang menjanjikan, penelitian ini memiliki beberapa keterbatasan yang perlu diakui. Penggunaan atribut yang terbatas pada IPK, SKS, dan semester mengakibatkan model belum mampu menangkap kompleksitas faktor-faktor yang memengaruhi ketepatan waktu kelulusan secara menyeluruh. Penelitian lanjutan disarankan untuk memperluas cakupan fitur dengan memasukkan variabel-variabel tambahan yang lebih komprehensif—seperti keaktifan organisasi, kondisi sosial-ekonomi, maupun pola kehadiran—sehingga kapabilitas prediktif model dapat ditingkatkan secara signifikan..

DOI: https://doi.org/10.31004/riggs.v5i2.10092

Referensi

1. E. F. Wati, E. S. Perangin-angin, and A. P. Sari, "Prediction of Student Graduation using the K-Nearest Neighbors Method," vol. 7, no. 158, pp. 211–216, 2023, doi: https://doi.org/10.30645/ijistech.v7i3.318.
2. C. Romero and S. Ventura, "Educational Data mining and Learning Analytics : An updated survey," vol. 10, no. 3, 2020, doi: p. e1355, 2020, doi: 10.1002/widm.1355.
3. E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," *Appl. Comput. Intell. Soft Comput.*, vol. 2024, 2024, doi: 10.1155/2024/4067721.
4. K. Okoye, J. T. Nganji, J. Escamilla, and S. Hosseini, "Computers and Education : Artificial Intelligence Machine learning model (RG-DMML) and ensemble algorithm for prediction of students ' retention and graduation in education," *Comput. Educ. Artif. Intell.*, vol. 6, no. January, p. 100205, 2024, doi: 10.1016/j.caeai.2024.100205.
5. D. A. Enggar Novianto, Arief Hermawan, "Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, Decision Tree Untuk Prediksi Status Kelulusan Mahasiswa S1," vol. 8, no. 2, pp. 146–154, 2023, doi: https://doi.org/10.36341/rabit.v8i2.3434.
6. A. Putri, C. S. Hardiana, E. Novfuja, F. Try, and P. Siregar, "Comparison of K-NN , Naive Bayes and SVM Algorithms for Final-Year Student Graduation Prediction Komparasi Algoritma K-NN , Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," vol. 3, no. April, pp. 20–26, 2023, doi: https://doi.org/10.57152/malcom.v3i1.610.
7. E. S. Rachardian Seprima, "Prediksi Kelulusan Tepat Waktu Mahasiswa Untuk Pemantauan Program Studi Menggunakan Metode Data Mining," vol. 21, no. 2, pp. 168–182, 2024, doi: https://doi.org/10.24246/aiti.v21i2.168-182.
8. S. Sarker, M. K. Paul, S. Tasnimul, H. Thasin, and A. M. Hasan, "Analyzing students academic performance using educational data mining," *Comput. Educ. Artif. Intell.*, vol. 7, no. December 2023, p. 100263, 2024, doi: 10.1016/j.caeai.2024.100263.
9. D. Safitri, S. S. Hilabi, and F. Nurapriani, "Analisis penggunaan algoritma klasifikasi dalam prediksi kelulusan menggunakan orange data mining," vol. 8, no. 1, pp. 75–81, 2023, doi: 10.36341/rabit.v8i1.3009.
10. N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," vol. 2, no. 2, pp. 86–91, 2021, doi: 10.21831/jeatech.v2i2.42777.
11. J. Zeniarja *et al.*, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," vol. 18, no. 2, pp. 102–108, 2022, doi: 10.17529/jre.v18i2.24047.
12. D. Y. Sari and J. Wang, "K-Nearest Neighbors (K-NN) Algorithm Model in Predicting the Graduation Rate of Teacher Professional Education Students in Indonesia," vol. 4, no. August, pp. 291–310, 2024, doi: 10.47134/ijsl.v4i3.277.
13. H. S. M. Zaehol Fatah, "Penerapan Algoritma K-Nearest Neighbor (K-NN) untuk Memprediksi Kelulusan Mahasiswa Menggunakan RapidMiner," *Jamastika*, vol. 4, 2025, doi: 10.35473/jamastika.v4i2.4482.
14. S. C. Matz, C. S. Bukow, H. Peters, C. Deacons, A. Dinu, and C. Stachl, "Using machine learning to predict student retention from socio - demographic characteristics and app - based engagement metrics," *Sci. Rep.*, pp. 1–16, 2023, doi: 10.1038/s41598-023-32484-w.
15. [M. R. Firmansyah and Y. P. Astuti, "Stroke Classification Comparison with KNN through Standardization and Normalization Techniques," vol. 6, no. 1, pp. 1–8, 2024, doi: 10.26877/asset.v6i1.17685.
16. J. J. Valero-mas, C. Penarrubia, F. J. Castellanos, J. Gallego, and J. Calvo-zaragoza, "Insights into imbalance-aware Multilabel Prototype Generation mechanisms for *k* -Nearest Neighbor classification in noisy scenarios," *Pattern Recognit.*, vol. 169, no. June 2025, p. 111884, 2026, doi: 10.1016/j.patcog.2025.111884.
17. S. Amandha, H. Rohayani, K. Kurniawansyah, S. Teknologi, and Jambi, "Implementation of Data Mining for Predicting Student Graduation Using the K-Nearest Neighbor Algorithm at Jambi Muhammadiyah University," vol. 7, no. 1, pp. 134–140, 2024, doi: 10.24014/ijaidm.v26150.
18. L. Lavazza, S. Morasca, and G. Rotoloni, "Software Defect Prediction evaluation : New metrics based on the ROC curve," *Inf. Softw. Technol.*, vol. 187, no. April 2024, p. 107865, 2025, doi: 10.1016/j.infsof.2025.107865.

DOI: https://doi.org/10.31004/riggs.v5i2.10092