p-ISSN: 2963-7856| e-ISSN: 2961-9890
Available online at jerkin.org/index.php/jerkin
Jurnal Pengabdian Masyarakat dan Riset Pendidikan
Volume 4 No. 3 Januari-Maret 2026, pp 16576-16579

Analisa Prediksi Kelulusan Mahasiswa Menggunakan Metode *Machine Learning*

Abdul Khalik Walya[1*], Hasbi Rizki Sulistyo[2], Ibnu Agustian Pratama[3], Sifatul Akmal[4], Imam Budiawan[5], Desmulyati[6]

[1-5]Sistem Informasi, [6]Informatika, Universitas Bina Sarana Informatika, Jl. Kramat Raya No.98, RT.2/RW.9, Kwitang, Kec. Senen, Kota Jakarta Pusat
E-mail: 17230601@bsi.ac.id
* Corresponding Author

https://doi.org/10.31004/jerkin.v4i3.4959

ARTICLE INFO

Article history
Received: 23 Nov 2025
Revised: 05 Dec 2025
Accepted: 30 Dec 2025

Kata Kunci:
Machine Learning, Prediksi Kelulusan, Klasifikasi, *Logistic Regression*, Random Forest, KNN

Keywords:
Machine Learning, Graduation Prediction, Classification, Logistic Regression, Random Forest, KNN

ABSTRACT

Prediksi kelulusan mahasiswa merupakan salah satu permasalahan penting dalam bidang pendidikan tinggi karena berkaitan dengan evaluasi keberhasilan proses akademik. Berbagai algoritma machine learning telah digunakan untuk membantu memprediksi kelulusan mahasiswa berdasarkan data akademik. Penelitian ini melakukan studi komparatif terhadap tiga algoritma klasifikasi, yaitu Logistic Regression, Random Forest, dan K-Nearest Neighbor, menggunakan data mahasiswa yang terdiri dari 200 data dengan atribut usia, program studi, IPK, dan tahun kelulusan. Tahapan penelitian meliputi pra-pemrosesan data, pembagian data latih dan data uji, pelatihan model, serta evaluasi performa menggunakan metrik klasifikasi. Hasil eksperimen menunjukkan bahwa Logistic Regression dan Random Forest menghasilkan performa terbaik dengan tingkat akurasi mencapai 100%, sedangkan K-Nearest Neighbor memperoleh akurasi sebesar 80%. Temuan ini menunjukkan bahwa karakteristik data dan pemilihan algoritma memiliki pengaruh signifikan terhadap hasil prediksi kelulusan mahasiswa.karakteristik data dan pemilihan algoritma memiliki pengaruh signifikan terhadap hasil prediksi kelulusan mahasiswa.

Student graduation prediction is an important issue in higher education as it is closely related to the evaluation of academic success. Various machine learning algorithms have been applied to predict student graduation based on academic data. This study conducts a comparative analysis of three classification algorithms, namely Logistic Regression, Random Forest, and K-Nearest Neighbor, using a simulated dataset consisting of 200 student records with attributes including age, department, GPA, and graduation year. The research stages include data preprocessing, data splitting, model training, and performance evaluation using classification metrics. Experimental results indicate that Logistic Regression and Random Forest achieve the best performance with an accuracy of 100%, while the K-Nearest Neighbor algorithm attains an accuracy of 80%. These findings highlight that data characteristics and algorithm selection significantly affect the accuracy of student graduation prediction.

How to Cite: Abdul Khalik Walya, et al (2025). Analisa Prediksi Kelulusan Mahasiswa Menggunakan Metode *Machine Learning*, 4(3) 16576-16579. https://doi.org/10.31004/jerkin.v4i3.4959

PENDAHULUAN

Perkembangan teknologi informasi mendorong pemanfaatan data dalam berbagai bidang, termasuk bidang pendidikan. Salah satu permasalahan yang sering dihadapi oleh institusi pendidikan adalah rendahnya tingkat kelulusan mahasiswa atau keterlambatan studi. Oleh karena itu, diperlukan suatu pendekatan yang mampu membantu pihak akademik dalam melakukan prediksi kelulusan mahasiswa secara lebih awal.

Machine learning merupakan salah satu pendekatan yang banyak digunakan untuk melakukan

prediksi dan klasifikasi berdasarkan pola data historis. Beberapa algoritma machine learning seperti Logistic Regression, Random Forest, dan K-Nearest Neighbor telah banyak diterapkan pada permasalahan klasifikasi, termasuk pada data akademik mahasiswa. Namun, setiap algoritma memiliki karakteristik dan performa yang berbeda tergantung pada jenis dan kompleksitas data yang digunakan.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membandingkan performa beberapa algoritma machine learning dalam memprediksi kelulusan mahasiswa menggunakan data mahasiswa, sebagaimana pendekatan komparatif yang banyak digunakan dalam educational data mining. Fokus penelitian ini bukan pada implementasi sistem nyata, melainkan pada analisis perbandingan performa algoritma dan pemahaman perilaku model terhadap karakteristik data akademik.

METODE

Penelitian ini menggunakan pendekatan eksperimen dengan tahapan utama meliputi pengumpulan dataset, pra-pemrosesan data, pemodelan machine learning, serta evaluasi performa model.

Pra-pemrosesan Data

Pada tahap pra-pemrosesan, dilakukan pemeriksaan missing value untuk memastikan tidak terdapat data kosong yang dapat memengaruhi performa model. Selanjutnya, data kategorikal pada atribut program studi dan label kelulusan dilakukan proses encoding agar dapat diproses oleh algoritma machine learning. Setelah itu, data dipisahkan menjadi data latih dan data uji dengan rasio 80:20.

Algoritma yang Digunakan

Algoritma yang digunakan dalam penelitian ini meliputi Logistic Regression sebagai model klasifikasi linear, Random Forest sebagai model ensemble berbasis pohon keputusan, dan K-Nearest Neighbor sebagai model berbasis jarak. Pemilihan ketiga algoritma ini bertujuan untuk membandingkan performa model dengan karakteristik yang berbeda.

Metode Evaluasi

Evaluasi performa model dilakukan dengan menyesuaikan standar jurnal terapan di bidang machine learning. Oleh karena itu, penelitian ini hanya menggunakan rumus evaluasi dasar yang umum digunakan, tanpa menurunkan rumus algoritma klasifikasi secara rinci.

Metrik evaluasi yang digunakan meliputi accuracy, precision, recall, dan F1-score. Accuracy digunakan untuk mengukur tingkat ketepatan prediksi secara keseluruhan, sedangkan precision dan recall digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan kelas tertentu. F1-score digunakan sebagai rata-rata harmonik antara precision dan recall.

HASIL DAN PEMBAHASAN

Deskripsi Atribut Data

Tabel 1. Deskripsi Atribut Dataset Mahasiswa

Atribut	Tipe Data	Keterangan
Age	Numerik	Usia mahasiswa
Department	Kategorikal	Program studi mahasiswa
GPA	Numerik	Indeks Prestasi Kumulatif
GraduationYear	Numerik	Tahun kelulusan
Lulus	Kategorikal	Status kelulusan

Tabel 1 menunjukkan atribut yang digunakan dalam penelitian ini. Atribut GPA merepresentasikan capaian akademik mahasiswa, sedangkan atribut lainnya berfungsi sebagai variabel pendukung. Kombinasi atribut numerik dan kategorikal memungkinkan penerapan berbagai algoritma klasifikasi untuk menganalisis pola kelulusan mahasiswa.

Perbandingan Performa Model

Tabel 2. Perbandingan Performa Algoritma Klasifikasi

Atribut	Akurasi	Precision	Recall	F1-Score	AUC
Logistic Regression	1.00	1.00	1.00	1.00	1.00

ISSN 2963-7856 (print), ISSN 2961-9890 (online)

Random Forest	1.00	1.00	1.00	1.00	1.00
KNN	0.80	0.75	0.80	0.75	0.91

Tabel 2. Logistic Regression dan Random Forest menunjukkan performa yang sangat tinggi dengan nilai akurasi dan AUC mencapai 1.00, sejalan dengan temuan penelitian sebelumnya pada data pendidikan. Sementara itu, algoritma KNN menghasilkan performa yang lebih rendah dengan akurasi sebesar 0.80. Perbedaan performa ini menunjukkan bahwa karakteristik data memiliki pengaruh signifikan terhadap efektivitas algoritma yang digunakan.

Analisis Confusion Matrix

Tabel 3. Ringkasan Confusion Matrix Setiap Model

Model	TP	FP	FN	TN
Logistic Regression	8	0	0	32
Random Forest	8	0	0	32
KNN	1	1	7	31

Tabel 3. Memperlihatkan bahwa Logistic Regression dan Random Forest tidak menghasilkan kesalahan klasifikasi pada data uji. Sebaliknya, KNN masih mengalami kesalahan prediksi, khususnya pada kelas minoritas, yang tercermin dari rendahnya nilai true positive.

Pembahasan

Performa tinggi yang dihasilkan oleh Logistic Regression dan Random Forest dipengaruhi oleh karakteristik dataset yang relatif sederhana. Analisis menunjukkan bahwa variabel GPA memiliki kontribusi yang sangat dominan terhadap penentuan status kelulusan. Dominasi satu fitur utama menyebabkan pemisahan kelas menjadi hampir sempurna, sehingga meningkatkan performa model secara signifikan.

Di sisi lain, algoritma KNN menunjukkan performa yang lebih rendah karena sensitivitasnya terhadap skala dan distribusi data. Pada penelitian ini, nilai parameter K ditetapkan sebesar 5. Pemilihan nilai K = 5 dilakukan karena merupakan nilai yang umum digunakan dalam penelitian klasifikasi untuk menjaga keseimbangan antara sensitivitas model terhadap noise dan kemampuan generalisasi. Nilai K yang terlalu kecil cenderung membuat model sensitif terhadap data pencilan, sedangkan nilai K yang terlalu besar dapat mengaburkan batas antar kelas. Tanpa proses standarisasi fitur dan tanpa optimasi nilai K, algoritma berbasis jarak cenderung mengalami penurunan performa, khususnya pada dataset dengan ketidakseimbangan kelas.

Meskipun hasil penelitian menunjukkan performa yang tinggi, penggunaan data dengan jumlah yang terbatas menjadi salah satu keterbatasan penelitian ini. Oleh karena itu, hasil yang diperoleh belum dapat digeneralisasikan secara langsung pada data akademik riil.

SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa Logistic Regression dan Random Forest menunjukkan performa yang sangat baik dalam memprediksi kelulusan mahasiswa pada data yang digunakan dalam penelitian ini. Sementara itu, KNN menghasilkan performa yang relatif lebih rendah dibandingkan kedua algoritma tersebut. Dominasi atribut GPA menjadi faktor utama yang memengaruhi tingginya performa model. Penelitian ini menunjukkan bahwa karakteristik data sangat berpengaruh terhadap hasil klasifikasi. Untuk penelitian selanjutnya, disarankan menggunakan dataset riil dengan jumlah data yang lebih besar serta menerapkan teknik pra-pemrosesan yang lebih lengkap.

UCAPAN TERIMA KASIH

Peneliti menyampaikan ucapan terima kasih kepada pihak yang sudah berkontribusi dalam pelaksanaan penelitian dan penyusunan artikel ini.

REFERENSI

C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.
L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
P. Cortez and A. Silva, “Using data mining to predict secondary school student performance,” Expert

Systems with Applications, vol. 35, no. 4, pp. 186–194, 2008.
T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Sebastopol, CA, USA: O'Reilly Media, 2019.
J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Waltham, MA, USA: Morgan Kaufmann, 2012.
D. W. Hosmer and S. Lemeshow, Applied Logistic Regression. Hoboken, NJ, USA: Wiley, 2013.
S. B. Kotsiantis, "Educational data mining: A review," International Journal of Artificial Intelligence in Education, vol. 21, no. 1, pp. 1–18, 2010.
A. Nugroho and B. Santosa, "Analisis klasifikasi data akademik mahasiswa menggunakan metode data mining," Jurnal Informatika, vol. 14, no. 2, pp. 85–92, 2018.
D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.
C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, vol. 40, no. 6, pp. 601–618, 2010.
Suyanto, "Machine learning untuk prediksi kelulusan mahasiswa," Jurnal Ilmu Komputer, vol. 11, no. 1, pp. 15–22, 2017.
Z. Arifin, "Prediksi kelulusan mahasiswa menggunakan algoritma Naive Bayes," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 8, no. 3, pp. 401–408, 2021.
I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques. Burlington, MA, USA: Morgan Kaufmann, 2016.

ISSN 2963-7856 (print), ISSN 2961-9890 (online)