

KLASIFIKASI TINGKAT KELULUSAN MAHASISWA TEPAT WAKTU BERDASARKAN DATA AKADEMIK MENGUNAKAN NAÏVE BAYES

Nasril¹, Adelia Berliana Syariyanti², Nesha Audi Maulidia³

^{1,2,3}Program Studi Manajemen Informatika, Jurusan Komputer, Politeknik LP3I Jakarta

Email: ¹nasril477@gmail.com, ²adeliabrlns01@gmail.com, ³audinesha8@gmail.com

Abstrak

Kelulusan mahasiswa tepat waktu merupakan salah satu indikator penting dalam menilai kualitas penyelenggaraan pendidikan tinggi, namun masih terdapat mahasiswa yang mengalami keterlambatan menyelesaikan studi. Penelitian ini bertujuan menerapkan metode Naïve Bayes untuk mengklasifikasikan tingkat kelulusan mahasiswa tepat waktu berdasarkan data akademik serta mengevaluasi kinerja model yang dihasilkan. Dataset yang digunakan terdiri dari 1.687 data mahasiswa dengan atribut ip1, ip2, ip3, ip4, dan tepat sebagai label klasifikasi. Tahapan penelitian meliputi pengumpulan data, preprocessing (data cleaning, data selection, data transformation, dan pelabelan data), pembagian data training 80% dan testing 20%, klasifikasi menggunakan RapidMiner, serta evaluasi model menggunakan confusion matrix, accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan model memperoleh accuracy sebesar 88,72%, precision sebesar 93,59%, recall sebesar 94,19%, dan F1-score sebesar 93,89%. Confusion matrix menghasilkan 292 True Positive, 7 True Negative, 20 False Negative, dan 18 False Positive. Performa yang tinggi pada kelas “Ya” menunjukkan model sangat baik dalam mengidentifikasi mahasiswa yang lulus tepat waktu, meskipun performa pada kelas “Tidak” masih relatif rendah akibat ketidakseimbangan distribusi data. Disimpulkan bahwa algoritma Naïve Bayes memiliki performa yang baik dan layak digunakan sebagai metode pendukung keputusan akademik dalam memprediksi kelulusan mahasiswa tepat waktu.

Kata Kunci: Data Mining, Naïve Bayes, Klasifikasi, Kelulusan Mahasiswa, RapidMiner.

Abstract

Timely student graduation is a key indicator of higher education quality; however, many students still experience delays in completing their studies. This research aims to apply the Naïve Bayes algorithm to classify students' timely graduation status based on academic data and to evaluate the resulting model's performance. The dataset consists of 1,687 student records with five attributes: ip1, ip2, ip3, ip4, and tepat as the class label. The research stages include data collection, preprocessing (data cleaning, selection, transformation, and labeling), an 80:20 training-testing split, classification using RapidMiner, and evaluation using a confusion matrix, accuracy, precision, recall, and F1-score. The experimental results show an accuracy of 88.72%, precision of 93.59%, recall of 94.19%, and F1-score of 93.89%, with a confusion matrix of 292 True Positives, 7 True Negatives, 20 False Negatives, and 18 False Positives. The high performance on the “Yes” class indicates that the model is very effective at identifying students who graduate on time, although performance on the “No” class remains relatively low due to class imbalance. It is concluded that the Naïve Bayes algorithm performs well and is suitable as a decision-support method for predicting timely student graduation.

Keywords: Data Mining, Naïve Bayes, Classification, Student Graduation, RapidMiner.

1. PENDAHULUAN

Pendidikan tinggi memiliki peran penting dalam menghasilkan sumber daya manusia yang berkualitas. Salah satu indikator keberhasilan perguruan tinggi adalah tingkat kelulusan mahasiswa tepat waktu, yang mencerminkan proses pembelajaran dan pengelolaan akademik yang baik. Namun, pada kenyataannya masih terdapat mahasiswa yang mengalami keterlambatan dalam menyelesaikan studinya, yang dipengaruhi oleh berbagai faktor akademik seperti Indeks Prestasi, nilai mata kuliah, dan jumlah SKS yang ditempuh [1].

Perkembangan teknologi informasi mendorong perguruan tinggi mengelola data akademik secara digital, sehingga data tersebut berpotensi diolah menjadi informasi bermanfaat melalui data mining, yaitu proses menemukan pola atau pengetahuan baru dari sekumpulan data menggunakan teknik statistik, basis data, dan machine learning [1]. Dalam bidang pendidikan, penerapan ini dikenal sebagai Educational Data Mining (EDM), yang bertujuan menganalisis perilaku belajar mahasiswa dan mendukung pengambilan keputusan berbasis data [2].

Salah satu teknik data mining yang banyak digunakan adalah klasifikasi. Penelitian ini menerapkan algoritma Naïve Bayes, yaitu metode berbasis probabilitas menggunakan Teorema Bayes dengan asumsi setiap atribut bersifat independen. Naïve Bayes memiliki proses komputasi yang relatif sederhana, cepat, mudah diterapkan, serta mampu memberikan hasil klasifikasi yang baik pada berbagai jenis data [3]. Berbagai penelitian terdahulu menunjukkan bahwa Naïve Bayes memiliki performa yang baik dalam memprediksi kelulusan mahasiswa, meskipun pemanfaatannya pada data akademik lokal masih perlu terus dikembangkan [4]–[6].

Berdasarkan uraian tersebut, penelitian ini bertujuan menerapkan Naïve Bayes untuk mengklasifikasikan tingkat kelulusan mahasiswa tepat waktu berdasarkan data akademik berupa Indeks Prestasi Semester 1 sampai Semester 4, serta mengevaluasi kinerja model yang dihasilkan menggunakan confusion matrix, accuracy, precision, recall, dan F1-score.

2. METODE PENELITIAN

Penelitian ini menggunakan metode kuantitatif dengan pendekatan data mining untuk klasifikasi. Tahapan penelitian meliputi pengumpulan data, preprocessing data, pembagian data training dan testing, klasifikasi menggunakan Naïve Bayes, dan evaluasi model, sebagaimana diuraikan pada sub-bagian berikut.

A. Pengumpulan Data

Data yang digunakan merupakan dataset `data_lulus_tepat_waktu.csv` yang berisi 1.687 data mahasiswa dengan lima atribut, yaitu `ip1`, `ip2`, `ip3`, `ip4` (Indeks Prestasi semester 1–4), dan `tepat` sebagai label klasifikasi dengan dua kategori: Ya (lulus tepat waktu, 1.552 data) dan Tidak (tidak lulus tepat waktu, 135 data). Distribusi kelas yang tidak seimbang ini menjadi pertimbangan penting dalam evaluasi model.

B. Preprocessing Data

Data cleaning dilakukan menggunakan Microsoft Excel untuk memeriksa data kosong dan duplikat; hasil pemeriksaan tidak menemukan keduanya sehingga seluruh 1.687 data digunakan. Data selection dilakukan dengan operator Select Attributes pada RapidMiner untuk memilih atribut `ip1`–`ip4` dan `tepat`. Data transformation menetapkan `ip1`–`ip4` sebagai atribut numerik dan `tepat` sebagai atribut nominal berkategori Ya/Tidak. Pelabelan data dilakukan menggunakan operator Set Role dengan menjadikan atribut `tepat` sebagai label.

C. Pembagian Data

Dataset dibagi menjadi data training (80%) dan data testing (20%) menggunakan operator Split Data dengan metode shuffled agar data terdistribusi secara acak. Data training digunakan untuk membangun model, sedangkan data testing digunakan untuk mengukur performa model pada data yang belum pernah dipelajari.

D. Klasifikasi Menggunakan Naïve Bayes

Naïve Bayes menghitung probabilitas suatu data termasuk ke dalam kelas tertentu berdasarkan Teorema Bayes, dirumuskan sebagai berikut.

$$P(C|X) = [P(X|C) \times P(C)] / P(X) \quad (1)$$

dengan $P(C|X)$ adalah probabilitas posterior kelas C terhadap data X, $P(X|C)$ adalah likelihood, $P(C)$ adalah prior probability, dan $P(X)$ adalah evidence. Alur proses pada RapidMiner terdiri dari operator Retrieve, Set Role, Split Data, Naïve Bayes, Apply Model, dan Performance Classification, di mana model dilatih pada data training dan diterapkan pada data testing untuk menghasilkan prediksi status kelulusan.

E. Evaluasi Model

Evaluasi dilakukan menggunakan confusion matrix yang terdiri atas True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN), serta empat metrik berikut.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \times 100\% \quad (2)$$

$$\text{Precision} = TP / (TP + FP) \times 100\% \quad (3)$$

$$\text{Recall} = TP / (TP + FN) \times 100\% \quad (4)$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

3. HASIL DAN PEMBAHASAN

Pengujian model dilakukan menggunakan operator Performance Classification pada RapidMiner terhadap data testing (20% dari 1.687 data, yaitu 337 data). Hasil confusion matrix disajikan pada Tabel 1.

Tabel 1. Confusion Matrix Hasil Pengujian Naïve Bayes

	Prediksi Tidak	Prediksi Ya
Aktual Tidak	7 (TN)	18 (FP)
Aktual Ya	20 (FN)	292 (TP)

Berdasarkan Tabel 1, sebanyak 292 data mahasiswa yang lulus tepat waktu berhasil diprediksi benar sebagai kelas Ya, dan 7 data mahasiswa yang tidak lulus tepat waktu berhasil diprediksi benar sebagai kelas Tidak. Namun, terdapat 20 data yang sebenarnya tidak lulus tepat waktu tetapi diprediksi lulus tepat waktu (False Negative), serta 18 data yang sebenarnya lulus tepat waktu tetapi diprediksi tidak lulus tepat waktu (False Positive).

Berdasarkan nilai TP, TN, FP, dan FN tersebut, diperoleh accuracy sebesar $(292+7)/(292+7+20+18) \times 100\% = 88,72\%$. Hasil evaluasi lengkap disajikan pada Tabel 2.

Tabel 2. Hasil Evaluasi Model Naïve Bayes

Metrik	Nilai
Accuracy	88,72%
Precision	93,59%
Recall	94,19%
F1-score	93,89%

Nilai precision, recall, dan F1-score per kelas disajikan pada Tabel 3 untuk memperlihatkan pengaruh ketidakseimbangan data terhadap performa masing-masing kelas.

Tabel 3. Precision, Recall, dan F1-score per Kelas

Metrik	Kelas Ya	Kelas Tidak
Precision	93,59%	28,00%
Recall	94,19%	25,93%
F1-score	93,89%	26,93%

Berdasarkan Tabel 3, nilai precision, recall, dan F1-score pada kelas Ya masing-masing sebesar 93,59%, 94,19%, dan 93,89%, menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengidentifikasi mahasiswa yang lulus tepat waktu. Sebaliknya, nilai precision, recall, dan F1-score pada kelas Tidak masing-masing hanya sebesar 28,00%, 25,93%, dan 26,93%. Rendahnya performa pada kelas Tidak disebabkan oleh jumlah data kelas Tidak (135 data) yang jauh lebih sedikit dibandingkan kelas Ya (1.552 data), sehingga model Naïve Bayes lebih banyak mempelajari pola dari kelas mayoritas dan kurang optimal dalam mengenali kelas minoritas.

Secara keseluruhan, hasil pengujian dengan accuracy 88,72% menunjukkan bahwa algoritma Naïve Bayes mampu mengklasifikasikan kelulusan mahasiswa tepat waktu dengan tingkat ketepatan yang tinggi berdasarkan atribut Indeks Prestasi Semester 1 sampai Semester 4. Meskipun demikian, ketidakseimbangan distribusi kelas perlu menjadi perhatian pada penelitian selanjutnya, misalnya dengan menerapkan teknik resampling seperti SMOTE untuk meningkatkan representasi kelas minoritas.

4. KESIMPULAN

Metode Naïve Bayes berhasil diterapkan untuk mengklasifikasikan kelulusan mahasiswa tepat waktu berdasarkan data akademik berupa Indeks Prestasi Semester 1 sampai Semester 4, melalui tahapan preprocessing data, pembagian data training dan testing, pembentukan model menggunakan RapidMiner, serta evaluasi menggunakan confusion matrix. Model yang dihasilkan memperoleh accuracy sebesar 88,72%, precision sebesar 93,59%, recall sebesar 94,19%, dan F1-score sebesar 93,89% pada kelas Ya, yang menunjukkan bahwa algoritma Naïve Bayes memiliki kemampuan yang baik dalam mengidentifikasi mahasiswa yang berpotensi lulus tepat waktu, meskipun performa pada kelas Tidak masih perlu ditingkatkan akibat ketidakseimbangan data. Hasil klasifikasi ini dapat dimanfaatkan sebagai informasi pendukung bagi program studi dalam mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan sejak dini. Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar dan beragam, menambahkan atribut lain seperti IPK dan jumlah SKS, menerapkan teknik penyeimbangan data

seperti SMOTE, serta membandingkan Naïve Bayes dengan algoritma klasifikasi lain seperti Decision Tree, Random Forest, atau Support Vector Machine.

DAFTAR PUSTAKA

- [1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.
- [2] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, e1355, 2020.
- [3] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Waltham, MA, USA: Morgan Kaufmann, 2017.
- [4] R. H. Tanjung, Y. Yunus, and G. W. Nurcahyo, "Comparison of C4.5 and Naive Bayes algorithms in predicting student graduation," *Jurnal CoSciTech*, vol. 4, no. 1, pp. 193–199, 2023, doi: 10.37859/coscitech.v4i1.4755.
- [5] L. Lianah, S. Z. Harahap, and I. Irmayati, "Implementation of the C4.5 and Naive Bayes algorithms to predict student graduation," *Sinkron*, vol. 8, no. 3, pp. 1741–1757, 2024, doi: 10.33395/sinkron.v8i3.13860.
- [6] R. Sukarna and Y. Ansori, "Implementasi data mining menggunakan metode Naïve Bayes dengan feature selection untuk prediksi kelulusan mahasiswa tepat waktu," *Jurnal Ilmiah Sains dan Teknologi*, vol. 6, pp. 50–61, 2022.
- [7] G. Gunawan, Hanes, and Catherine, "C4.5, K-Nearest Neighbor, Naïve Bayes, and Random Forest Algorithms Comparison to Predict Students' On Time Graduation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 4, no. 2, 2021, doi: 10.24014/ijaidm.v4i2.10833.
- [8] D. Dikriani and A. T. I. Karim, "Comparison of C4.5 and Naive Bayes Algorithm Methods in Prediction of Student Graduation on Time (Case Study: Information Systems Study Program)," *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, vol. 3, no. 1, 2023, doi: 10.20895/dinda.v3i1.782.
- [9] M. Y. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, art. 11, 2022, doi: 10.1186/s40561-022-00192-z.
- [10] D. Hermanto, D. I. Ricoida, D. Pibriana, Rusbandi, and M. R. Pribadi, "Analysis of Student Graduation Prediction Using Machine Learning Techniques on an Imbalanced Dataset: An Approach to Address Class Imbalance," *Scientific Journal of Informatics*, vol. 11, no. 3, pp. 559–568, 2024, doi: 10.15294/sji.v11i3.5528.
- [11] A. N. Anwar, D. Dani, and A. Napila, "Naïve Bayes Algorithm Analysis For Student Graduation Timeliness Prediction," *bit-Tech*, vol. 8, no. 1, 2025, doi: 10.32877/bt.v8i1.2825.

- [12] J. Ibrahimy and Z. Fatah, "Data Mining Classification to Predict Student Graduation Using the Naive Bayes Method," *JITTER: Jurnal Ilmiah Teknologi dan Komputer*, vol. 5, no. 3, 2024, doi: 10.24843/JTRTI.2024.v05.i03.p07.
- [13] M. I. Syaifudin, B. Mulyawan, and N. J. Perdana, "Design of Student Graduation Prediction System Using Naive Bayes and Website-Based Decision Tree," *International Journal of Application on Sciences, Technology and Engineering*, vol. 1, no. 4, pp. 1599–1609, 2023, doi: 10.24912/ijaste.v1.i4.1599-1609.
- [14] A. U. Khasanah and Harwati, "Educational Data Mining Techniques Approach to Predict Student's Performance," *International Journal of Information and Education Technology*, vol. 9, no. 2, pp. 115–118, 2019, doi: 10.18178/ijiet.2019.9.2.1184.
- [15] D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, et al., "Prediction of student exam performance using data mining classification algorithms," *Education and Information Technologies*, vol. 29, pp. 21621–21645, 2024, doi: 10.1007/s10639-024-12619-w.