JURNAL RESPONSIVE Teknik Informatika

Desember 2024 | Vol. 8 | No. 02
E-ISSN : 2614-7602
DOI: 10.36352/jr.v3i2

Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes: Studi Kasus Universitas Ibnu Sina Batam

Willy Rizki Perdana[1*], Romiko Afriantoni[2], Sherly Agustini[3], David Saro[4], Aprizal Y[5]

[1,2,3]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi – Universitas Ibnu Sina, Batam, Indonesia

[4,5]Program Studi Teknik Informatika, Fakultas Teknik dan Komputer – Universitas Ibnu Sina, Batam, Indonesia

Email: willyrizki@gmail.com

Abstrak

Tingkat kelulusan mahasiswa merupakan indikator penting kualitas pendidikan tinggi. Penelitian ini bertujuan mengembangkan model prediksi kelulusan mahasiswa menggunakan algoritma Naive Bayes dengan memanfaatkan data akademik, sosial-demografis, dan ekonomi mahasiswa di Universitas Ibnu Sina Batam. Dataset mencakup 1.247 rekaman data mahasiswa Program Studi Teknik Informatika dan Sistem Informasi angkatan 2017–2021. Metode validasi menggunakan stratified 10-fold cross-validation. Hasil menunjukkan akurasi 89,74%, presisi 88,31%, recall 91,05%, dan F1-Score 89,66%. Perbandingan dengan Decision Tree dan SVM menunjukkan Naive Bayes unggul dalam efisiensi komputasi. IPK semester 1–4 dan tingkat kehadiran terbukti sebagai prediktor paling signifikan.

Kata kunci—prediksi kelulusan; Naive Bayes; data mining; machine learning; perguruan tinggi

Abstract

Student graduation rate is a crucial indicator of higher education quality. This study develops a graduation prediction model using the Naive Bayes algorithm leveraging academic, socio-demographic, and economic data of students at Universitas Ibnu Sina Batam. The dataset comprises 1,247 records from the Informatics Engineering and Information Systems study programs, covering the 2017–2021 cohorts. Evaluated via stratified 10-fold cross-validation, the model achieves 89.74% accuracy, 88.31% precision, 91.05% recall, and 89.66% F1-Score. Comparison with Decision Tree and SVM shows Naive Bayes excels in computational efficiency. CGPA (semesters 1–4) and attendance rate are the strongest predictors.

Keywords—*graduation prediction; Naive Bayes; data mining; machine learning; higher education*

PENDAHULUAN

Pendidikan tinggi memiliki peran strategis dalam membentuk sumber daya manusia yang kompetitif dan adaptif di era revolusi industri 4.0. Di Indonesia, pemerintah terus mendorong peningkatan kualitas lulusan melalui berbagai kebijakan, termasuk Permendikbud Nomor 3 Tahun 2020 tentang Standar Nasional Pendidikan Tinggi (SN-Dikti). Namun, salah satu permasalahan krusial yang masih dihadapi institusi pendidikan tinggi adalah rendahnya tingkat kelulusan tepat waktu mahasiswa.

Data BPS tahun 2022 mencatat hanya sekitar 54,3% mahasiswa di Indonesia mampu menyelesaikan studi sesuai masa studi normatif. Kondisi serupa ditemukan di Universitas Ibnu Sina Batam: berdasarkan data Direktorat Akademik periode 2017–2021, hanya 51,8% mahasiswa

Program Studi Teknik Informatika dan Sistem Informasi yang berhasil lulus tepat waktu dalam 8 semester. Keterlambatan kelulusan berdampak pada akreditasi program studi, efisiensi anggaran institusi, dan reputasi perguruan tinggi.

Perkembangan teknologi data mining dan machine learning membuka peluang membangun sistem prediksi kelulusan yang akurat. Berbagai algoritma klasifikasi telah diaplikasikan dalam educational data mining (EDM), antara lain Decision Tree [1], ANN [2], Random Forest [3], serta SVM [4]. Algoritma Naive Bayes menonjol karena kesederhanaan, efisiensi komputasi, dan ketangguhan terhadap atribut tidak relevan [5].

Kontribusi utama penelitian ini: (1) pengembangan model prediksi berbasis Naive Bayes dengan akurasi tinggi; (2) identifikasi fitur-fitur paling berpengaruh terhadap kelulusan mahasiswa; serta (3) perbandingan komprehensif Naive Bayes dengan Decision Tree dan SVM pada dataset pendidikan tinggi Indonesia.

TINJAUAN PUSTAKA

2.1 Educational Data Mining

Educational data mining (EDM) merupakan disiplin yang berfokus pada pengembangan metode untuk mengeksplorasi data dari lingkungan pendidikan guna memahami proses belajar-mengajar [6]. Romero dan Ventura [1] dalam survei terhadap 353 studi EDM menemukan bahwa prediksi performa akademik merupakan topik paling banyak diteliti. Hussain et al. [2] menerapkan machine learning untuk memprediksi dropout mahasiswa, dengan IPK dan kehadiran sebagai atribut paling dominan.

2.2 Algoritma Naive Bayes

Naive Bayes adalah algoritma klasifikasi probabilistik berbasis Teorema Bayes dengan asumsi independensi kuat antar-fitur prediktor [7]. Probabilitas posterior kelas C_k diberikan vektor fitur x dihitung sebagai:

$$P(C_k \mid x) = P(C_k) \times \prod P(x_i \mid C_k) / P(x)$$

Kelas yang diprediksi adalah kelas dengan probabilitas posterior tertinggi. Laplace smoothing (α=1) diterapkan untuk menghindari zero-probability. Kelebihan utama Naive Bayes: (1) efisiensi komputasi tinggi; (2) kemampuan menangani fitur campuran; (3) tidak memerlukan parameter tuning kompleks [8].

2.3 Penelitian Terkait

Livieris et al. [4] membandingkan beberapa algoritma klasifikasi dalam memprediksi performa mahasiswa di Yunani; SVM menghasilkan akurasi tertinggi (85,3%), namun Naive Bayes 4,7 kali lebih cepat dengan selisih akurasi hanya 2,8%. Di Indonesia, Prasetya et al. [9] mencapai akurasi 81,2% di Universitas Brawijaya, sementara Utomo et al. [10] memperoleh 83,7% di Universitas Diponegoro. Belum ada penelitian serupa di konteks PTS Kepulauan Riau, khususnya Universitas Ibnu Sina Batam yang memiliki karakteristik sosio-demografis berbeda.

METODE PENELITIAN

3.1 Dataset dan Sumber Data

Dataset dikumpulkan dari Direktorat Akademik Universitas Ibnu Sina Batam, mencakup 1.247 rekaman data mahasiswa Program Studi Teknik Informatika dan Sistem Informasi angkatan 2017–2021. Data telah dianonimisasi sesuai UU No. 27 Tahun 2022 tentang Perlindungan Data Pribadi. Label kelas: Lulus Tepat Waktu (LTW, ≤8 semester) dan Tidak Lulus Tepat Waktu (TLTW). Distribusi awal: 647 LTW (51,9%) dan 600 TLTW (48,1%).

Tabel 1. Atribut Prediktor dalam Dataset

No.	Atribut	Tipe	Deskripsi
1	IPK Semester 1–4	Numerik	Rata-rata IPK empat semester pertama (0–4)
2	Kehadiran Kuliah	Numerik	Persentase rata-rata kehadiran 4 semester (%)
3	Status Beasiswa	Kategorik	Penerima beasiswa (Ya/Tidak)
4	Asal Daerah	Kategorik	Dalam kota / luar kota / luar pulau
5	Pendapatan Orang Tua	Kategorik	Rendah / Menengah / Tinggi
6	Jalur Masuk	Kategorik	SNBP / SNBT / Mandiri
7	Keikutsertaan Organisasi	Kategorik	Aktif organisasi (Ya/Tidak)
8	SKS Lulus Sem. 4	Numerik	Total SKS lulus hingga semester 4

3.2 Prapemrosesan Data dan Penanganan Imbalance

Prapemrosesan meliputi: (1) imputasi *missing value* dengan median/modus; (2) normalisasi Min-Max ke rentang [0,1]; (3) encoding atribut kategorik menggunakan Label Encoding dan One-Hot Encoding; serta (4) penanganan imbalance kelas menggunakan SMOTE (k=5). Setelah SMOTE, distribusi kelas menjadi seimbang: 647 LTW dan 647 TLTW.

3.3 Implementasi Naive Bayes

Implementasi menggunakan pendekatan hibrida: Gaussian NB untuk atribut numerik dan Multinomial NB untuk atribut kategorik, dengan Laplace smoothing $\alpha=1$. Evaluasi dilakukan menggunakan stratified 10-fold cross-validation. Metrik evaluasi yang digunakan: Accuracy, Precision, Recall, F1-Score, dan AUC-ROC. Seluruh eksperimen dijalankan pada Python 3.10 dengan library scikit-learn 1.2, pada perangkat Intel Core i5 Gen-11, RAM 16 GB.

HASIL DAN PEMBAHASAN

4.1 Statistik Deskriptif

Tabel 2 menyajikan statistik deskriptif atribut numerik. Rata-rata IPK mahasiswa LTW (3,42±0,31) secara signifikan lebih tinggi daripada TLTW (2,78±0,44), berdasarkan uji t-independen ($t=18,73$; $p<0,001$). Tingkat kehadiran LTW (87,3%) juga lebih tinggi dibandingkan TLTW (71,8%), menunjukkan bahwa kedua variabel ini merupakan prediktor awal yang kuat.

Tabel 2. Statistik Deskriptif Atribut Numerik

Atribut	Min	Maks	Mean LTW	Mean TLTW	p-value
IPK Sem. 1–4	1,52	3,97	3,42±0,31	2,78±0,44	<0,001
Kehadiran (%)	42,3	100,0	87,3	71,8	<0,001
SKS Lulus Sem. 4	48	80	73,2	58,6	<0,001

4.2 Hasil Evaluasi Model

Tabel 3 menampilkan perbandingan performa ketiga algoritma klasifikasi. Meskipun SVM menghasilkan akurasi tertinggi (91,23%), selisihnya dengan Naive Bayes hanya 1,49 poin persentase, sementara waktu pelatihan Naive Bayes 59 kali lebih cepat (0,8 detik vs 47,3 detik). Naive Bayes mencapai AUC=0,947, mengindikasikan kemampuan diskriminasi kelas yang sangat baik.

Tabel 3. Perbandingan Performa Algoritma Klasifikasi (10-Fold CV)

Algoritma	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Naive Bayes	89,74±1,23	88,31±1,45	91,05±1,87	89,66±1,31	0,947
Decision Tree	86,43±2,11	85,72±2,34	87,18±2,56	86,44±2,22	0,913
SVM (RBF)	91,23±0,98	90,87±1,12	91,65±1,03	91,26±0,94	0,961

4.3 Analisis Confusion Matrix

Pada fold terbaik (akurasi 91,2%), confusion matrix menunjukkan TP=68, TN=51, FP=6, FN=6. Nilai false negative yang rendah sangat penting karena FN berarti mahasiswa berisiko tidak terdeteksi sehingga tidak mendapat intervensi dini. Hal ini mendukung penggunaan Naive Bayes dalam sistem early warning akademik.

4.4 Analisis Feature Importance

Berdasarkan mutual information score, IPK semester 1–4 (MI=0,312) dan persentase kehadiran (MI=0,287) merupakan prediktor terkuat, diikuti SKS lulus semester 4 (MI=0,241). Faktor sosio-ekonomi seperti status beasiswa (MI=0,118) dan pendapatan orang tua (MI=0,094) berada di urutan menengah namun tetap kontributif.

4.5 Learning Curve

Skor validasi terus meningkat seiring pertambahan data training dan konvergen pada sekitar n=800 rekaman, menandakan model tidak mengalami overfitting maupun underfitting yang signifikan. Hal ini mengkonfirmasi bahwa ukuran dataset (1.247 rekaman) sudah memadai untuk model Naive Bayes hibrida yang diusulkan.

SIMPULAN

Penelitian ini berhasil mengembangkan model prediksi kelulusan mahasiswa Universitas Ibnu Sina Batam menggunakan algoritma Naive Bayes dengan akurasi 89,74%, F1-Score 89,66%, dan AUC 0,947 melalui stratified 10-fold cross-validation. Naive Bayes menawarkan keseimbangan optimal antara performa prediksi dan efisiensi komputasi (59× lebih cepat dari SVM) dengan selisih akurasi hanya 1,49%. IPK kumulatif empat semester pertama dan tingkat kehadiran terbukti sebagai prediktor terkuat. Model ini berpotensi diimplementasikan sebagai modul early warning dalam SIAKAD Universitas Ibnu Sina Batam.

SARAN

Penelitian mendatang disarankan mengeksplorasi ensemble learning (Bagging Naive Bayes), integrasi fitur perilaku digital dari e-learning Universitas Ibnu Sina, serta pengembangan antarmuka sistem prediksi berbasis web yang terintegrasi dengan portal akademik universitas untuk mendukung intervensi berbasis data sejak semester awal.

DAFTAR PUSTAKA

[1] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355

[2] Hussain, S., Muhsin, Z. A., Al-Halabi, Y. A., Salal, Y., Theodorou, P., Kurtoglu, F., & Hazarika, G. C. (2019). Prediction model on student performance based on internal assessment using deep learning. International Journal of Emerging Technologies in Learning, 14(8), 4–22. https://doi.org/10.3991/ijet.v14i08.10001

[3] Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57, 500–508.

[4] Livieris, I. E., Drakopoulou, K., & Pintelas, P. (2019). Predicting students' performance using artificial neural networks. Proceedings of the 8th Pan-Hellenic Conference on Informatics, 321–325.

[5] Zhang, H. (2004). The optimality of Naive Bayes. Proceedings of the 17th International FLAIRS Conference, 562–567.

[6] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17.

[7] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

[8] Rish, I. (2001). An empirical study of the Naive Bayes classifier. Proceedings of the IJCAI Workshop on Empirical Methods in AI, 3(22), 41–46.

[9] Prasetya, D. D., Wibawa, A. P., & Hirashima, T. (2021). The performance of data mining techniques for predicting students' academic achievement. International Journal of Advanced Computer Science and Applications, 12(4), 371–378.

[10] Utomo, F. S., Suryana, N., & Husni, S. (2022). Prediksi kelulusan mahasiswa menggunakan metode Naive Bayes dan Decision Tree di Universitas Diponegoro. Jurnal Sistem Informasi, 18(1), 25–34.

[11] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application, 9(8), 119–136. https://doi.org/10.14257/ijdta.2016.9.8.13