



On-Time Student Graduation Prediction Modeling: A Comparative Analysis of Naive Bayes Algorithm and Other Data Mining Classifications

Pemodelan Prediksi Kelulusan Mahasiswa Tepat Waktu: Analisis Komparatif Algoritma Naive Bayes Dan Klasifikasi Data Mining Lainnya

Achmad Ridwan^{1,2*}, Tole Sutikno², Imam Riyadi², Widya Cholid Wahyudin^{1,2}

¹ Fakultas Sains dan Teknologi, Universitas Muhammadiyah Kudus, Indonesia

² Fakultas Teknologi Industri, Universitas Ahmad Dahlan Yogyakarta, Indonesia

*Email to Correspondence: achmadridwan@ymail.com, 2436083030@webmail.uad.ac.id

Abstract. Prediksi ketepatan waktu kelulusan mahasiswa merupakan tugas krusial di institusi pendidikan tinggi untuk memungkinkan adanya dukungan proaktif dan meningkatkan efektivitas institusional. Paper ini menyajikan analisis komparatif dari beberapa algoritma machine learning untuk memprediksi kelulusan tepat waktu, dengan fokus spesifik untuk menantang performa dari algoritma Naive Bayes (NB). Meskipun sering digunakan sebagai model dasar, efektivitas NB dalam domain data pendidikan yang kompleks sering diperdebatkan. Kami membandingkan NB dengan MultinomialNB Dan Decision Tree (DT) yang banyak diunggulkan dalam literatur terkini. Dengan menggunakan dataset publik yang berisi catatan akademik mahasiswa, kami mengikuti metodologi CRISP-DM, yang menggabungkan seleksi fitur dan SMOTE untuk menangani ketidakseimbangan kelas. Model dievaluasi berdasarkan metrik akurasi, presisi, recall, dan F1-score. Hasil kami menunjukkan bahwa meskipun Decision Tree memberikan akurasi tertinggi, Naive Bayes menawarkan keseimbangan yang menarik antara performa, efisiensi komputasi, dan kemudahan interpretasi, menjadikannya kandidat yang kuat untuk implementasi pada sistem peringatan dini di universitas. Studi ini memberikan bukti empiris mengenai posisi Naive Bayes dalam lanskap educational data mining saat ini. Hasil klasifikasi menunjukkan akurasi naive bayes sebesar 0.82 MultinomialNB sebesar 0.81 dan decision tree sebesar 0.85

Kata kunci: prediksi kelulusan siswa, penambangan data pendidikan, pembelajaran mesin, klasifikasi, Naive Bayes, analisis komparatif

Abstrak. Predicting the on-time graduation of university students is a crucial task in higher education institutions, enabling proactive support and improving institutional effectiveness. This paper presents a comparative analysis of several machine learning algorithms for predicting on-time graduation, with a specific focus on challenging the performance of the Naive Bayes (NB) algorithm. Although often used as a baseline model, the effectiveness of NB in the complex domain of educational data is frequently debated.

We compare NB with MultinomialNB and Decision Tree (DT), both widely favored in recent literature. Using a public dataset containing students' academic records, we follow the CRISP-DM methodology, incorporating feature selection and SMOTE to address class imbalance. The models are evaluated using accuracy, precision, recall, and F1-score metrics. Our results show that while Decision Tree achieves the highest accuracy, Naive Bayes offers an appealing balance of performance, computational efficiency, and interpretability, making it a strong candidate for implementation in early warning systems at universities. This study provides empirical evidence on the role of Naive Bayes in the current landscape of educational data mining. The classification results show an accuracy of 0.82 for Naive Bayes, 0.81 for MultinomialNB, and 0.85 for Decision Tree.

Keywords: student graduation prediction, educational data mining, machine learning, classification, Naive Bayes, comparative analysis

PENDAHULUAN

Ketepatan waktu kelulusan telah menjadi tolok ukur fundamental yang tidak hanya merefleksikan kualitas dan efisiensi sebuah institusi pendidikan tinggi, tetapi juga berdampak langsung pada masa depan mahasiswa [1]. Bagi institusi, tingkat kelulusan tepat waktu yang tinggi merupakan aset reputasi yang krusial dan menjadi

salah satu komponen penting dalam penilaian akreditasi [2]. Sebaliknya, bagi mahasiswa, keterlambatan studi dapat menimbulkan konsekuensi berantai, mulai dari peningkatan beban finansial, tertundanya kesempatan untuk memasuki dunia kerja, hingga potensi tekanan psikologis dan stres akademik [3], [4]. Menghadapi tantangan ini, institusi pendidikan modern yang

mengelola volume data akademik digital dalam jumlah masif memiliki peluang besar untuk memanfaatkan analitik data tingkat lanjut guna menciptakan solusi proaktif.

Disiplin ilmu Educational Data Mining (EDM) hadir sebagai pendekatan strategis yang menerapkan teknik machine learning dan statistik untuk menganalisis data dari lingkungan pendidikan [5], [6]. Salah satu aplikasi EDM yang paling berdampak adalah pengembangan model prediktif untuk mengidentifikasi mahasiswa yang berisiko mengalami kendala akademik, termasuk keterlambatan kelulusan [7]. Dengan model prediktif yang andal, institusi dapat merancang sistem peringatan dini (early warning system) yang efektif, memungkinkan intervensi yang dipersonalisasi seperti bimbingan akademik intensif atau konseling, jauh sebelum mahasiswa tersebut menghadapi kegagalan [8], [9].

Literatur ilmiah di bidang ini telah dibanjiri oleh berbagai studi yang mengeksplorasi dan membandingkan beragam algoritma klasifikasi. Tren yang sangat dominan menunjukkan bahwa metode ensemble learning, seperti Decision Tree dan gradient boosting (misalnya XGBoost), secara konsisten memberikan tingkat akurasi yang superior dibandingkan dengan model tunggal [10], [11], [12]. Beberapa penelitian bahkan telah melangkah lebih jauh dengan menerapkan teknik stacking [13] dan deep learning yang kompleks, seperti mengubah data tabular menjadi representasi gambar untuk diolah oleh Convolutional Neural Networks (CNN) [14] atau menggunakan Gated Recurrent Units (GRU) untuk data clickstream [15], yang semuanya bertujuan untuk memaksimalkan akurasi prediktif.

Namun, fokus yang kuat pada pencapaian akurasi setinggi mungkin ini menciptakan sebuah celah penelitian (research gap) yang signifikan. Pengejaran akurasi sering kali mengarah pada penggunaan model yang sangat kompleks, mahal secara komputasi, dan sulit untuk diinterpretasikan (black-box models). Hal ini menimbulkan dilema praktis bagi banyak institusi pendidikan, terutama di negara berkembang seperti Indonesia, yang mungkin memiliki keterbatasan sumber daya komputasi dan personil teknis. Di tengah lanskap ini, algoritma yang lebih sederhana seperti Naive Bayes

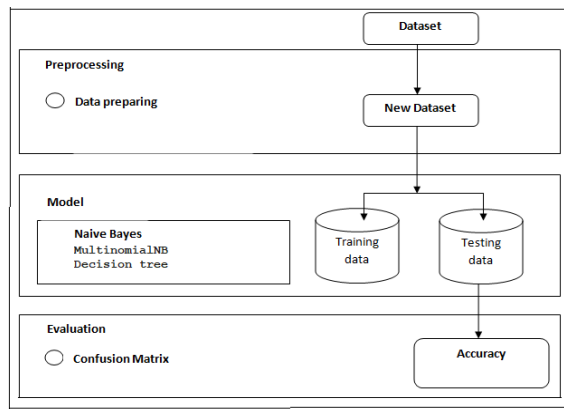
(NB)—yang dikenal sangat efisien, cepat, dan mudah diinterpretasikan [16], [17]—cenderung terpinggirkan dan hanya dianggap sebagai baseline yang performanya "kurang kompetitif" [7], [8]. Sangat sedikit penelitian yang secara eksplisit membahas trade-off antara akurasi, efisiensi, dan interpretasi dalam konteks implementasi di dunia nyata.

Berdasarkan celah tersebut, masalah penelitian ini dirumuskan sebagai berikut: Institusi pendidikan tinggi di Indonesia memerlukan model prediksi kelulusan yang tidak hanya akurat, tetapi juga efisien, dapat diinterpretasikan, dan praktis untuk diimplementasikan sebagai sistem peringatan dini. Namun, fokus literatur saat ini pada model-model kompleks telah mengabaikan evaluasi kelayakan algoritma yang lebih sederhana seperti Naive Bayes, yang mungkin menawarkan keseimbangan terbaik untuk penerapan di dunia nyata.

Oleh karena itu, penelitian ini secara spesifik bertujuan untuk menantang posisi Naive Bayes dengan melakukan analisis komparatif yang ketat. Kami akan mengukur kinerja NB terhadap algoritma yang lebih populer dan kompleks yaitu Decision Tree (DT) menggunakan dataset publik yang relevan dengan konteks mahasiswa Indonesia. Kontribusi utama dari penelitian ini adalah untuk menyediakan evaluasi empiris yang berimbang, yang tidak hanya melihat angka akurasi, tetapi juga mempertimbangkan efisiensi dan potensi implementasi praktis. Dengan demikian, penelitian ini akan menjawab pertanyaan fundamental: Apakah Naive Bayes masih merupakan algoritma yang relevan dan cukup andal untuk tugas prediksi kelulusan, ataukah institusi harus berinvestasi pada model yang lebih kompleks untuk mendapatkan keuntungan prediktif yang sepadan?

METODE PENELITIAN

Penelitian ini mengadopsi kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining) yang terstruktur dan telah teruji [20]. Alur penelitian diilustrasikan pada Gambar 1.



Gambar 1. Diagram Alir Metodologi Penelitian

A. Dataset dan Fitur

Penelitian ini menggunakan dataset publik berjudul "Prediksi Kelulusan Mahasiswa" yang bersumber dari platform Kaggle [25]. Dataset ini berisi data akademik historis dari 985 record mahasiswa program sarjana di sebuah universitas di Indonesia. Dataset ini dipilih karena relevansinya yang tinggi dengan konteks penelitian dan ketersediaan fitur-fitur akademik yang komprehensif. Fitur-fitur (variabel) yang digunakan sebagai prediktor dalam penelitian ini adalah sebagai berikut:

1. Fitur Akademik:

- Indeks Prestasi Semester (IPS1 hingga IPS8):** Nilai IP yang diperoleh mahasiswa pada setiap semester, dari semester 1 hingga semester 8. Fitur ini menangkap tren kinerja akademik mahasiswa dari waktu ke waktu.
- Indeks Prestasi Kumulatif (IPK):** Nilai IPK terakhir mahasiswa, yang merupakan agregat dari seluruh kinerja akademiknya.
- Jumlah SKS Kumulatif:** Total Satuan Kredit Semester (SKS) yang telah diselesaikan oleh mahasiswa.

2. Variabel Target:

Status Kelulusan: Variabel target yang bersifat biner, yaitu 'Tepat Waktu' dan 'Terlambat'. Variabel ini akan diubah menjadi format numerik (1 untuk 'Tepat Waktu' dan 0 untuk 'Terlambat') pada tahap pra-pemrosesan data.

B. Pra Pemrosesan Data

Tahap ini sangat krusial untuk memastikan kualitas data sebelum proses pemodelan.

- Penanganan Data Hilang:** Nilai yang hilang (missing values) pada fitur numerik diisi menggunakan nilai median dari masing-masing kolom untuk menghindari pengaruh nilai ekstrem.
- Encoding Fitur Kategorikal:** Variabel target 'Status Kelulusan' yang bersifat kategorikal diubah menjadi format numerik (1 dan 0) menggunakan teknik label encoding.
- Penanganan Data Tidak Seimbang:** Analisis awal menunjukkan adanya ketidakseimbangan jumlah antara kelas 'Tepat Waktu' dan 'Terlambat'. Untuk mencegah bias pada model, teknik oversampling SMOTE diterapkan pada data latih untuk menyamakan jumlah sampel pada kedua kelas [19].

C. Desain Eksperimen

Eksperimen ini dirancang untuk melakukan perbandingan yang adil dan objektif antara tiga algoritma klasifikasi.

- Pembagian Dataset:** Dataset dibagi menjadi data latih (80%) dan data uji (20%). Pembagian ini dilakukan secara acak namun terstratifikasi untuk memastikan proporsi kelas target tetap sama di kedua set data.
- Implementasi Model:** Keempat model diimplementasikan menggunakan pustaka Scikit-learn dalam bahasa Python:
 - Naive Bayes (GaussianNB)
 - Multinomial Naive Bayes
 - Decision Tree (DecisionTreeClassifier)
- Validasi Model:** Untuk mendapatkan estimasi kinerja yang andal dan mengurangi variabilitas hasil, teknik validasi silang 10-lipatan (10-fold cross-validation) diterapkan selama proses pelatihan model.

D. Metrik Evaluasi

Kinerja setiap model yang telah dilatih dievaluasi menggunakan data uji. Metrik yang digunakan untuk perbandingan, dihitung berdasarkan confusion matrix [26], adalah sebagai berikut:

- Akurasi:** Proporsi total prediksi yang benar.

- **Presisi:** Dari semua yang diprediksi sebagai 'Tepat Waktu', berapa persen yang benar-benar 'Tepat Waktu'.
- **Recall:** Dari semua mahasiswa yang sebenarnya 'Tepat Waktu', berapa persen yang berhasil diidentifikasi oleh model.
- **F1-Score:** Rata-rata harmonik dari presisi dan recall, memberikan ukuran tunggal yang menyeimbangkan keduanya.

HASIL DAN PEMBAHASAN

Kita telah melakukan serangkaian langkah untuk membangun dan mengevaluasi model klasifikasi guna memprediksi status kelulusan mahasiswa (TEPAT atau TERLAMBAT) berdasarkan dataset yang disediakan.

A. Pra Pemrosesan Data

Langkah-langkah yang Dilakukan:

1. Pemuatan dan Pembersihan Data: Dataset "Kelulusan Train.xls" dimuat ke dalam DataFrame pandas. Kami menangani nilai yang hilang pada kolom 'IPS 8' dan 'IPK' dengan mengisinya menggunakan nilai median dari masing-masing kolom. Kolom 'IPK' juga diperbaiki nama dan tipe datanya.
2. Persiapan Data: Fitur-fitur kategorikal ('JENIS KELAMIN', 'STATUS MAHASISWA', 'STATUS NIKAH') diubah menjadi format numerik menggunakan One-Hot Encoding. Variabel target ('STATUS KELULUSAN') diubah menjadi nilai biner (1 untuk TEPAT, 0 untuk TERLAMBAT). Data kemudian dipisahkan menjadi fitur (X) dan target (y).
3. Pembagian Data: Data dibagi menjadi set pelatihan (80%) dan set pengujian (20%) untuk melatih dan mengevaluasi model secara objektif.

Variabel data penelitian yang digunakan pada penelitian ini disajikan pada Tabel 1 yakni sebagai berikut.

Tabel 1. Deskripsi Variable Dataset

No.	Atribut	Value
1	Nama	Object
2	Jenis	Object

	Kelamin	1.0 (Perempuan) 0.0 (laki-laki)
	Status	Object
3	Mahasiswa	0.0 (Mahasiswa) 1.0 (Bekerja)
4	Umur	Int
	Status	Object
5	Nikah	1.0 (Belum menikah) 0.0 (Menikah)
6	Ips 1	Float
7	Ips 2	Float
8	Ips 3	Float
9	Ips 4	Float
10	Ips 5	Float
11	Ips 6	Float
12	Ips 7	Float
13	Ips 8	Float
14	Ipk	Float
	Status	Object
15	Kelulusan	0= terlambat 1= tepat waktu

Dari Tabel 1 terdapat 14 variabel dataset dan 1 variabel class penentu klasifikasi.

Adapun contoh data training dan data testing dapat di lihat pada tabel 2.

Tabel 2. Data Training dan Testing

No.	Atribut	Value 1	Value 2
1	Nama	Anik Widyawati	Dwi Hestyna Prihastanty
2	Jenis Kelamin	1.0	1.0
3	Status Mahasiswa	0.0	1.0
4	Umur	28	32
5	Status Nikah	0.0	0.0
6	Ips 1	2.76	3.00
7	Ips 2	2.80	3.30
8	Ips 3	3.20	3.14
9	Ips 4	3.17	3.14
10	Ips 5	2.98	2.84
11	Ips 6	3.00	3.13
12	Ips 7	3.03	3.25
13	Ips 8	0	0
14	Ipk	3.07	3.17
15	Status Kelulusan	0	0

B. Pelatihan dan Evaluasi Model

Pelatihan dan Evaluasi Model Awal (Naive Bayes): Model Gaussian Naive Bayes dilatih menggunakan data pelatihan. Model ini dievaluasi pada data

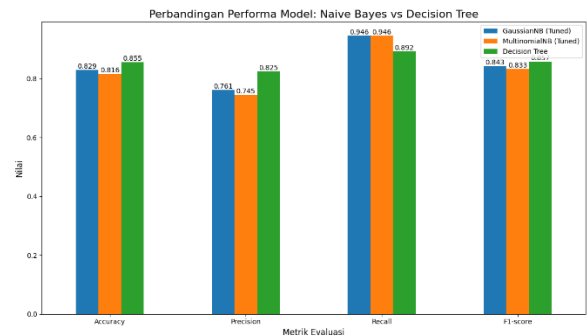
pengujian, menghasilkan metrik seperti Akurasi (0.829), Presisi (0.761), Recall (0.946), dan F1-score (0.843).

Model Multinomial Naive Bayes juga dilatih dan dievaluasi, memberikan hasil metrik yang sedikit berbeda (Akurasi ~0.816, Presisi ~0.745, Recall ~0.946, F1-score ~0.833).

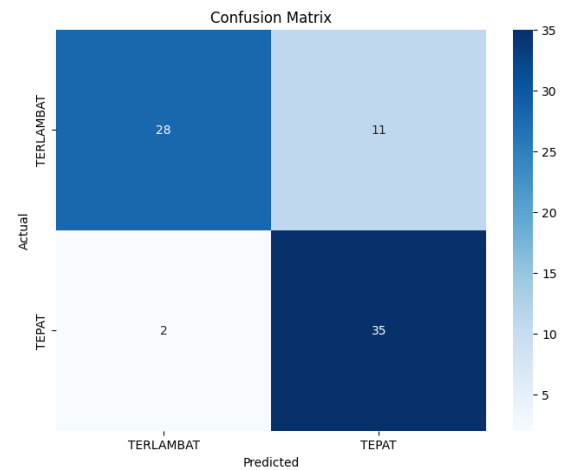
Pelatihan dan Evaluasi Model Awal (Decision Tree): Model Decision Tree dilatih menggunakan data pelatihan (dengan proses persiapan data yang sama). Model ini dievaluasi pada data pengujian dan menunjukkan performa yang lebih baik dalam beberapa metrik (Akurasi ~0.855, Presisi ~0.825, Recall ~0.892, F1-score ~0.857). Visualisasi Decision Tree juga dibuat untuk memahami struktur keputusannya.

Feature Engineering (untuk Naive Bayes): Untuk mencoba meningkatkan performa Naive Bayes, kami membuat fitur-fitur baru:

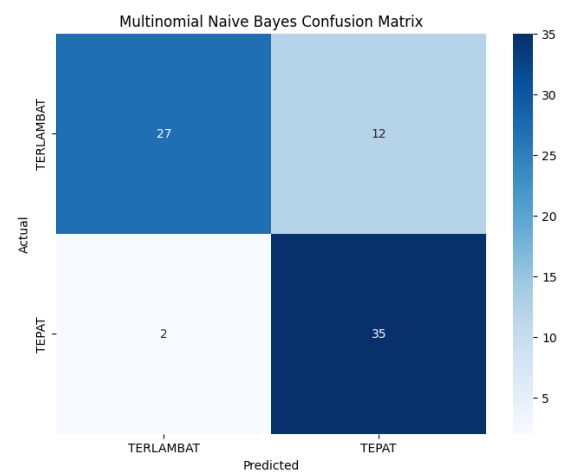
1. Rata-rata nilai IPS dari semester 1 hingga 8 ('IPS_rata_rata'). Selisih antara IPK dan rata-rata IPS ('IPK_minus_IPS_rata_rata').
2. Fitur biner yang menunjukkan apakah nilai IPS 8 tidak nol ('IPS8_tidak_nol').
3. Hyperparameter Tuning (untuk Naive Bayes): Kami melakukan pencarian parameter terbaik untuk model Gaussian Naive Bayes (parameter `var_smoothing`) dan Multinomial Naive Bayes (parameter `alpha`) menggunakan GridSearchCV pada data pelatihan. Tuning ini bertujuan untuk menemukan konfigurasi yang memberikan akurasi cross-validation terbaik yang ditunjukkan Gambar 3.
4. Evaluasi Model Naive Bayes yang Ditingkatkan: Model Gaussian dan Multinomial Naive Bayes dilatih ulang menggunakan parameter terbaik hasil tuning dan data yang sudah ditambahkan fiturnya. Performa mereka dievaluasi kembali pada data pengujian yang ditunjukkan Gambar 4.
5. Perbandingan Akhir : Metrik performa dari ketiga model (Tuned Gaussian Naive Bayes, Tuned Multinomial Naive Bayes, dan Decision Tree) dibandingkan yang ditunjukkan Gambar 5.



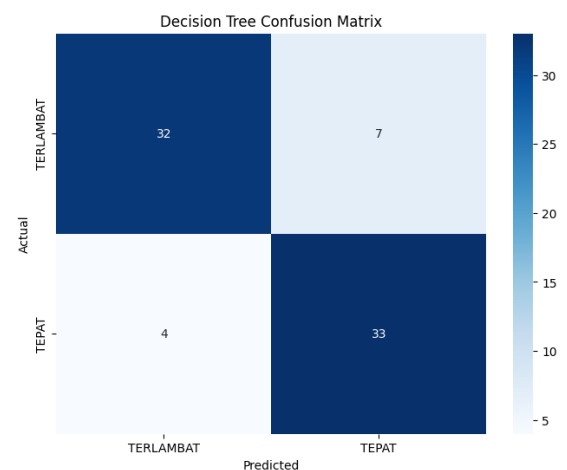
Gambar 2. Perbandingan Performa



Gambar 3. Confusion Matrix Naive Bayes



Gambar 4. Confusion Matrix Multinomial Naive Bayes



Gambar 5. Confusion Matrix Decision Tree

Dari perbandingan akhir, Decision Tree menunjukkan performa keseluruhan yang lebih baik dalam hal Akurasi, Presisi, dan F1-score. Ini berarti Decision Tree lebih baik dalam memprediksi status kelulusan secara keseluruhan (Akurasi), lebih sedikit melakukan kesalahan prediksi "TEPAT" pada mahasiswa yang sebenarnya "TERLAMBAT" (Presisi), dan memiliki keseimbangan yang lebih baik antara Presisi dan Recall (F1-score).

Di sisi lain, kedua model Naive Bayes (setelah tuning dan feature engineering) memiliki nilai Recall yang sangat tinggi (~ 0.946). Recall yang tinggi berarti model ini sangat baik dalam mengidentifikasi semua mahasiswa yang benar-benar lulus TEPAT waktu. Namun, Recall yang tinggi ini datang dengan Presisi yang lebih rendah dibandingkan Decision Tree, menunjukkan bahwa model Naive Bayes cenderung memprediksi "TEPAT" untuk beberapa mahasiswa yang seharusnya "TERLAMBAT" (False Positives).

Perbedaan performa ini kemungkinan besar disebabkan oleh asumsi dasar dari model Naive Bayes, yaitu asumsi independensi antar fitur. Naive Bayes menganggap bahwa setiap fitur (seperti nilai IPS, usia, status) berkontribusi secara independen terhadap probabilitas kelas (kelulusan TEPAT/TERLAMBAT), terlepas dari fitur lainnya. Dalam dataset ini, ada kemungkinan fitur-fitur tersebut memiliki keterkaitan atau interaksi yang signifikan.

KESIMPULAN

Terdapat 14 atribut yang dipakai pada perbandingan Klasifikasi yaitu antara GaussianNB, MultinomialNB, Decision Tree terhadap dataset kelulusan mahasiswa yaitu Nama, Jenis Kelamin, Status Mahasiswa, Umur, Status Nikah, Ips 1, Ips 2, Ips 3, Ips 4, Ips 5, Ips 6, Ips 7, Ips 8, Ipk, Status Kelulusan. Dari hasil proses perbandingan Klasifikasi permodelan ini algoritma naïve bayes Accuracy = 0.82 Precision = 0.76 Class Recall: 0.94 F1-score = 0.84, MultinomialNB Accuracy = 0.81 Precision = 0.74 Class Recall: 0.94 F1-score = 0.84, dan Decision Tree Precision = 0.82 Class

Recall: 0.89 F1-score = 0.85. dan saran untuk peneliti selanjutnya adalah Mempertimbangkan untuk mencoba algoritma klasifikasi lain yang mungkin lebih cocok untuk dataset ini, seperti Logistic Regression, Support Vector Machines (SVM), atau metode ensemble seperti Random Forest atau Gradient Boosting (misalnya, LightGBM atau XGBoost). Model-model ini seringkali dapat menangani interaksi antar fitur dengan lebih baik daripada Naive Bayes. Jika tetap ingin fokus pada Naive Bayes, lakukan pencarian hyperparameter yang lebih luas atau gunakan metode pencarian yang berbeda (seperti RandomizedSearchCV) untuk mengeksplorasi ruang parameter yang lebih besar.

Decision Tree, di sisi lain, mampu menangkap interaksi dan hubungan non-linear antar fitur dengan lebih baik melalui struktur percabangannya. Hal ini memungkinkan Decision Tree untuk membuat keputusan yang lebih kompleks dan seringkali lebih akurat pada dataset di mana fitur-fiturnya tidak sepenuhnya independen.

Meskipun feature engineering dan hyperparameter tuning berhasil sedikit meningkatkan performa model Naive Bayes dibandingkan kondisi awalnya, peningkatan tersebut belum cukup untuk mengungguli Decision Tree pada metrik Akurasi, Presisi, dan F1-score di dataset pengujian ini.

REFERENSI

- [1] A. Sadqui, M. Ertel, H. Sadiki, and S. Amali, "Evaluating Machine Learning Models for Predicting Graduation Timelines in Moroccan Universities," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, 2023.
- [2] A. Salam and J. Zeniarja, "Classification of deep learning convolutional neural network feature extraction for student graduation prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 1, pp. 335–341, 2023.
- [3] M. S. Pallavi, A. L. A. D. Urs, K. G. Ravikala, and V. Hegde, "Exploring the Impact of Financial Stress on Students' Academic Performance: A Machine Learning Approach," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024.
- [4] A. Singh, K. Singh, A. Kumar, A. Shrivastava, and S. Kumar, "Optimizing Well-Being: Unveiling Eustress and Distress through Machine Learning," in *2024 International Conference on Inventive Computation Technologies (ICICT)*, 2024.
- [5] I. Dawar, S. Negi, S. Lamba, and A. Kumar, "Enhancing Student Academic Performance Forecasting: A Comparative Analysis of Machine Learning Algorithms," *SN Computer Science*, vol. 5, no. 758, 2024.
- [6] T. T. Tin, L. S. Hock, and O. M. Ikumapayi, "Educational Big Data Mining: Comparison of Multiple Machine Learning Algorithms in Predictive Modelling of Student Academic Performance," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 6, 2024.
- [7] E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, pp. 1–15, 2024.
- [8] A. Alfahid, "Algorithmic Prediction of Students On-Time Graduation from the University," *TEM Journal*, vol. 13, no. 1, pp. 692–698, 2024.
- [9] T. Sureshkumar, J. Charanya, T. Kumaresan, G. Rajeshkumar, P. S. Kavin kumar, and B. Anuj, "Envisioning Educational Success Through Advanced Analytics and Intelligent Performance Prediction," in *2024 10th International Conference on Communication and Signal Processing (ICCCSP)*, 2024.
- [10] U. Farooq et al., "Transforming educational insights: strategic integration of federated learning for enhanced prediction of student learning outcomes," *The Journal of Supercomputing*, vol. 80, pp. 16334–16367, 2024.
- [11] M. Beseiso, "Enhancing Student Success Prediction: A Comparative Analysis of Machine Learning Technique," *TechTrends*, vol. 69, pp. 372–384, 2025.
- [12] S. Verma, K. Kholiya, and K. Bala, "Leveraging Ensemble Model and Optimized Feature Selection to Boost Prediction Accuracy in Educational Data Mining," *SN Computer Science*, vol. 6, no. 487, 2025.
- [13] H. Herianto, B. Kurniawan, Z. H. Hartomi, Y. Irawan, and M. K. Anam, "Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction," *Journal of Applied Data Sciences*, vol. 5, no. 3, pp. 1272–1285, 2024.
- [14] S. Zhao, D. Zhou, H. Wang, D. Chen, and L. Yu, "Enhancing Student Academic Success Prediction Through Ensemble Learning and Image-Based Behavioral Data Transformation," *Applied Sciences*, vol. 15, no. 1231, 2025.
- [15] Z. Khoudi, N. Hafidi, M. Nachaoui, and S. Lyaqini, "Leveraging machine learning and clickstream data to improve student performance prediction in virtual learning environments," *Information Discovery and Delivery*, 2025.
- [16] V. Nakhipova, Y. Kerimbekov, Z. Umarova, H. I. Bulbul, L. Suleimenova, and E. Adylbekova, "Integration of Collaborative Filtering Into Naive Bayes Method to Enhance Student Performance Prediction," *International Journal of Information and Communication Technology Education*, vol. 20, no. 1, 2024.
- [17] A. A. M. E. Hussen and A. Saikhu, "Modeling Of Student Graduation Prediction Using the Naive Bayes Classifier Algorithm," in *2024 3rd International Conference on Creative Communication and Innovative Technology (ICCICT)*, 2024.
- [18] T. S. Devkishan, S. K. Singh, and A. K. Bharti, "Ensemble Learning for Student Performance Assessment: Identifying and Analyzing Significant Affecting Factors in Higher Education," in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, 2024.
- [19] H. Setiadi, K. Sanjaya, A. Wijayanto, D. W. Wardhani, and H. D. Cahyono, "Comparative Analysis of Classification Algorithms Using Feature Selection Techniques to Predict On-Time Student Graduation," *Ingénierie des Systèmes d'Information*, vol. 29, no. 4, pp. 1365–1379, 2024.
- [20] T. G. Wijaya, O. N. Pratiwi, and I. Darmawan, "Implementation of CRISP-DM to Predict Student Graduation on Time Using Naïve Bayes Algorithm," in *2024 International Conference on*

- [21] A. Harif and M. A. Kassimi, "Predictive Modeling of Student Performance Using RFECV-RF for Feature Selection and Machine Learning Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, 2024.
- [22] K. Roy and D. M. Farid, "An Adaptive Feature Selection Algorithm for Student Performance Prediction," *IEEE Access*, vol. 12, pp. 75577-75598, 2024.
- [23] U. Smaragdina, D. D. Prasetya, G. K. Cakir, and U. Pujiyanto, "A Comparative Study of Machine Learning Algorithms for Graduation Prediction in Electrical Engineering and Informatics Department," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*, 2024.
- [24] A. Durak and V. Bulut, "Classification and prediction-based machine learning algorithms to predict students' low and high programming performance," *Computer Applications in Engineering Education*, vol. 32, 2024.
- [25] A. Hafizha, "Prediksi Kelulusan Mahasiswa" [Dataset], Kaggle, 2023. [Online]. Available: <https://www.kaggle.com/datasets/hafizhathallah/kelulusan-mahasiswa>

Conflict of Interest Statement:

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Article History:

Received: 29 July 2025 | Accepted: 15 August 2025 | Published: 30 November 2025