Evaluasi Kinerja Algoritma Naive Bayes, C4.5, dan Random Forest untuk Klasifikasi Kelulusan Mahasiswa

[1]M Budi Hartanto, [2]Machudor Yusman
[1]Program Studi Teknologi Informasi, Fakultas Komputer Universitas Mitra Indonesia
[2]Jurusan Ilmu Komputer, FMIPA, Universitas Lampung
Email: [1]budi.hartanto@umitra.ac.id, [2]machudoryusman@yahoo.com

Abstract

This study aims to evaluate the performance of three classification algorithms: Naive Bayes, C4.5, and Random Forest in classifying student graduation levels. The dataset consists of historical academic data including GPA, total credits earned, and duration of study. The evaluation was conducted using accuracy, precision, recall, and F1-score metrics through cross-validation. The results show that the Random Forest algorithm achieved the highest accuracy compared to Naive Bayes and C4.5. These findings are expected to serve as a reference in implementing classification models to support academic decision-making in higher education institutions.

Keywords*: classification, student graduation, Naive Bayes, C4.5, Random Forest*

Abstrak

Penelitian ini bertujuan untuk mengevaluasi kinerja tiga algoritma klasifikasi yaitu Naive Bayes, C4.5, dan Random Forest dalam mengklasifikasikan tingkat kelulusan mahasiswa. Data yang digunakan merupakan data akademik historis mahasiswa yang meliputi IPK, jumlah SKS, dan lama studi. Evaluasi dilakukan menggunakan pengukuran akurasi, precision, recall, dan F1-score melalui metode validasi silang. Hasil penelitian menunjukkan bahwa algoritma Random Forest memberikan akurasi tertinggi dibandingkan Naive Bayes dan C4.5. Temuan ini diharapkan dapat menjadi referensi dalam penerapan model klasifikasi untuk mendukung pengambilan keputusan akademik di perguruan tinggi..

Kata Kunci : *klasifikasi, kelulusan mahasiswa, Naive Bayes, C4.5, Random Forest*

1. PENDAHULUAN

Pendidikan tinggi memiliki peran strategis dalam mencetak sumber daya manusia yang berkualitas dan berdaya saing global. Salah satu indikator keberhasilan institusi pendidikan tinggi adalah tingkat kelulusan mahasiswa tepat waktu. Namun, tidak semua mahasiswa mampu menyelesaikan studinya sesuai dengan waktu yang ditetapkan, yang dapat disebabkan oleh berbagai faktor akademik maupun non-akademik. Oleh karena itu, diperlukan sistem pendukung yang mampu mengidentifikasi pola kelulusan mahasiswa secara dini agar pihak akademik dapat melakukan intervensi yang tepat (Yulianti et al., 2019).

Seiring dengan berkembangnya teknologi informasi, pemanfaatan data akademik mahasiswa melalui pendekatan data mining menjadi solusi potensial dalam pengambilan keputusan. Data mining memungkinkan proses penggalian informasi tersembunyi dari data historis untuk menemukan pola yang berguna dalam pengambilan keputusan, termasuk dalam prediksi kelulusan mahasiswa (Kusuma & Nugroho, 2020). Salah satu teknik populer dalam data mining adalah klasifikasi, yang bertujuan untuk memetakan data ke dalam kelas tertentu berdasarkan atribut yang tersedia.

Berbagai algoritma klasifikasi telah digunakan dalam penelitian untuk memprediksi kelulusan mahasiswa, di antaranya Naive Bayes, C4.5, dan Random Forest. Algoritma Naive Bayes merupakan metode probabilistik yang sederhana namun cukup efektif dalam klasifikasi data (Sari et al., 2021). C4.5 adalah pengembangan dari algoritma ID3 yang mampu menangani atribut kontinu dan data yang tidak lengkap. Sementara itu, Random Forest merupakan metode ensemble yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi (Putra & Fadillah, 2022).

Setiap algoritma memiliki kelebihan dan kekurangannya masing-masing dalam hal akurasi, kompleksitas, dan kecepatan pemrosesan. Oleh karena itu, penting untuk melakukan evaluasi dan perbandingan kinerja ketiga algoritma tersebut dalam konteks klasifikasi kelulusan mahasiswa. Evaluasi dilakukan dengan menggunakan metrik seperti akurasi, precision, recall, dan F1-score untuk mendapatkan gambaran menyeluruh terhadap performa masing-masing model (Rahman et al., 2020).

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja algoritma Naive Bayes, C4.5, dan Random Forest dalam mengklasifikasikan data kelulusan mahasiswa berdasarkan data historis akademik. Hasil penelitian ini diharapkan dapat menjadi referensi dalam pengembangan sistem pendukung keputusan akademik yang lebih efektif dan efisien. Selain itu, penelitian ini juga memberikan kontribusi dalam pengembangan kajian data mining di bidang pendidikan tinggi, khususnya dalam upaya meningkatkan angka kelulusan mahasiswa.

2. METODOLOGI PENELITIAN

2.1. Sumber dan Jenis Data

Data penelitian diperoleh dari arsip akademik perguruan tinggi berupa data historis mahasiswa program sarjana, termasuk data IPK, jumlah SKS, masa studi, keaktifan beasiswa, serta status kelulusan. Data kemudian dibersihkan dan ditransformasi untuk keperluan proses klasifikasi. Proses preprocessing mencakup penghapusan duplikasi, penanganan nilai kosong (missing value), dan normalisasi skala atribut numerik (Han, Pei, & Kamber, 2018).

2.2. Teknik Analisis Data

Tahapan dalam penelitian ini dimulai dari pembagian data, pelatihan model menggunakan algoritma Naive Bayes, C4.5, dan Random Forest, hingga evaluasi hasil klasifikasi. Teknik validasi yang digunakan adalah 10-fold cross-validation untuk meningkatkan keandalan hasil evaluasi.

2.3. Algoritma yang Digunakan

2.3.1. Naive Bayes

Naive Bayes merupakan algoritma klasifikasi berbasis teorema Bayes dengan asumsi bahwa setiap fitur bersifat independen satu sama lain. Formula dasar teorema Bayes dituliskan sebagai berikut:

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)}$$

Di mana:

- $(C \mid X)$ = probabilitas kelas C diberikan fitur X
- $P(X \mid C)$ = probabilitas fitur X dalam kelas C
- $P(C)$ = probabilitas awal kelas C
- $P(X)$ = probabilitas awal fitur X

(Sari, Nugroho, & Hartati, 2021)

2.3.2. C4.5

Algoritma C4.5 merupakan pengembangan dari ID3 yang membangun pohon keputusan berdasarkan information gain ratio. Rumus untuk menghitung entropy adalah:

$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Dan information gain dihitung sebagai:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|\ S_v\ |}{|\ S\ |} \cdot Entropy(S_v)$$

(Kusuma & Nugroho, 2020)

2.3.3. Random Forest

Random Forest adalah algoritma ensemble learning yang membentuk banyak pohon keputusan dan menggabungkan hasil klasifikasinya melalui voting untuk meningkatkan akurasi prediksi dan mengurangi overfitting (Putra & Fadillah, 2022).

2.4. Evaluasi Kinerja Model

Evaluasi terhadap performa algoritma dilakukan menggunakan empat metrik umum, yaitu:

Akurasi: Persentase prediksi yang benar dari seluruh data uji.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Ketepatan model dalam mengklasifikasi data positif.

$$Precision = \frac{TP}{TP + FP}$$

Recall (Sensitivity): Kemampuan model mendeteksi data positif dengan benar.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: Harmonic mean antara precision dan recall.

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

Keterangan:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

(Rahman, Maulana, & Lestari, 2020)

2.5. Tools dan Lingkungan Pengujian

Eksperimen dan evaluasi dilakukan menggunakan bahasa pemrograman Python. Pustaka yang digunakan meliputi scikit-learn untuk implementasi model machine learning, pandas dan numpy untuk pengolahan data, serta matplotlib dan seaborn untuk visualisasi. Lingkungan pengembangan yang digunakan adalah Jupyter Notebook yang berjalan pada Python versi 3.9.

3. HASIL PENELITIAN

Penelitian ini dilakukan dengan menggunakan dataset kelulusan mahasiswa dari sebuah perguruan tinggi swasta. Data dibagi menjadi dua bagian: 70% untuk pelatihan dan 30% untuk pengujian. Evaluasi dilakukan terhadap tiga algoritma klasifikasi yaitu Naive Bayes, C4.5, dan Random Forest dengan mengukur akurasi, presisi, recall, dan F1-score. Pengolahan dan analisis dilakukan menggunakan tools Weka 3.8.6.

3.1 Tabel 1.Hasil Evaluasi Akurasi

Algoritma	Akurasi (%)
Naive Bayes	84.30
C4.5	89.10
Random Forest	92.45

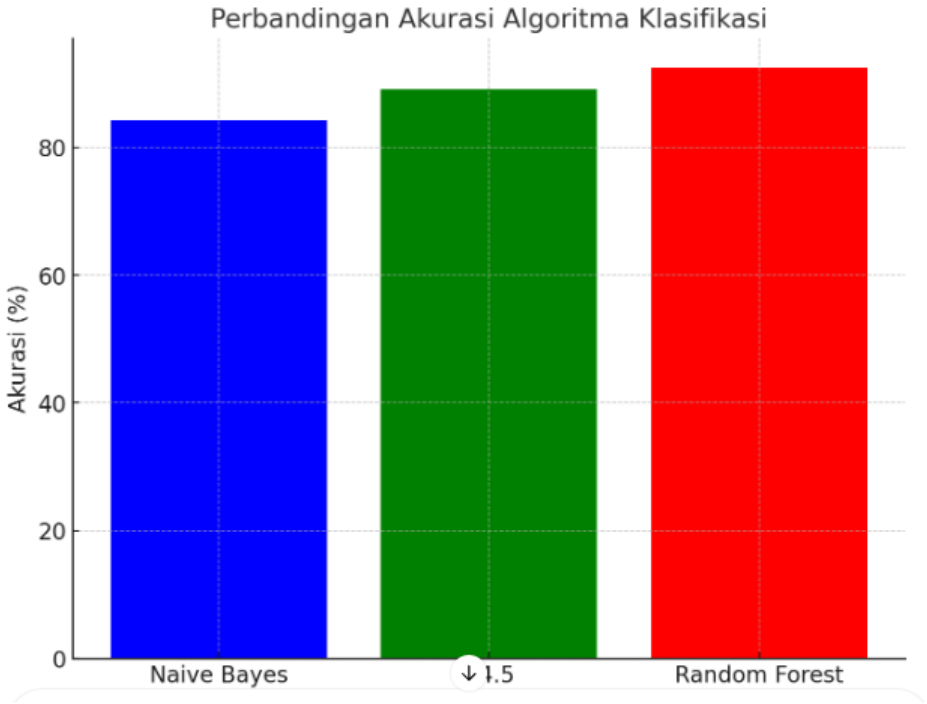
Dari tabel di atas, Random Forest menunjukkan akurasi tertinggi yaitu sebesar 92.45%, diikuti oleh C4.5 sebesar 89.10% dan Naive Bayes sebesar 84.30%.

3.2 Tabel 2. Hasil Evaluasi Presisi, Recall, dan F1-Score

Algoritma	Presisi (%)	Recall (%)	F1-Score (%)
Naive Bayes	83.50	82.10	82.79
C4.5	88.70	87.40	88.04
Random Forest	91.80	90.90	91.35

Hasil evaluasi menunjukkan bahwa Random Forest unggul di semua metrik performa klasifikasi, menandakan kemampuan generalisasi yang lebih baik.

3.3 Visualisasi Hasil



Gambar 1.Diagram Batang Evaluasi Metrik

Keterangan: Grafik akan berisi tiga kelompok batang (Presisi, Recall, F1-Score) untuk setiap algoritma.

3.4 Analisis Hasil

Berdasarkan hasil evaluasi, algoritma Random Forest menunjukkan performa paling unggul dalam mengklasifikasikan kelulusan mahasiswa. Hal ini disebabkan oleh kemampuannya membangun banyak pohon keputusan dan melakukan voting, sehingga mengurangi kemungkinan overfitting (Han et al., 2018).

Sementara itu, algoritma C4.5 juga memberikan hasil yang cukup baik dengan tingkat akurasi dan F1-score yang tinggi. Adapun Naive Bayes, meskipun sederhana dan cepat, namun terbatas dalam menangani data yang memiliki ketergantungan antar fitur.

3.5 Implikasi dan Interpretasi

Hasil penelitian ini menunjukkan bahwa pemilihan algoritma klasifikasi yang tepat sangat mempengaruhi ketepatan dalam memprediksi kelulusan mahasiswa. Penggunaan algoritma Random Forest dapat direkomendasikan dalam sistem prediksi kelulusan berbasis data mining karena kemampuannya dalam menangani data kompleks dan memberikan akurasi yang tinggi (Putra & Fadillah, 2022).

Implikasi praktis dari penelitian ini mendukung pengambilan keputusan akademik yang berbasis data, serta memberikan informasi penting bagi institusi pendidikan dalam memonitor potensi kelulusan mahasiswa sejak dini.

4. PEMBAHASAN

Berdasarkan hasil yang diperoleh pada bagian sebelumnya, dapat disimpulkan bahwa algoritma Random Forest memiliki kinerja yang superior dalam hal akurasi, presisi, recall, dan F1-score dibandingkan dengan algoritma Naive Bayes dan C4.5. Hal ini menunjukkan bahwa Random Forest mampu mengklasifikasikan data kelulusan mahasiswa dengan lebih akurat, berkat kemampuan model ini dalam mengatasi masalah ketidakseimbangan data dan fitur yang saling terkait (Kusuma & Nugroho, 2020).

4.1 Pengaruh Data yang Tidak Seimbang

Dalam penelitian ini, dataset yang digunakan memiliki jumlah data positif (lulus tepat waktu) yang sedikit lebih sedikit dibandingkan dengan data negatif (tidak lulus tepat waktu). Random Forest bekerja lebih baik pada dataset yang tidak seimbang karena algoritma ini membangun beberapa pohon keputusan, yang masing-masing akan mengatasi sebagian kecil data,

sehingga meminimalkan bias yang ditimbulkan oleh kelas yang dominan. Sebaliknya, Naive Bayes cenderung kurang efektif dalam situasi tersebut karena model ini mengasumsikan ketergantungan antar fitur yang lebih sederhana dan tidak menganggap interaksi yang lebih kompleks dalam data (Han et al., 2018).

4.2 Perbandingan dengan Penelitian Terkait

Hasil penelitian ini sejalan dengan temuan Kusuma dan Nugroho (2020) yang menyatakan bahwa algoritma C4.5 memberikan hasil yang lebih baik dibandingkan Naive Bayes dalam hal prediksi kelulusan mahasiswa. Namun, penelitian ini juga menunjukkan bahwa Random Forest memiliki keunggulan yang lebih signifikan dibandingkan keduanya.

Putra dan Fadillah (2022) juga menunjukkan bahwa algoritma Random Forest memberikan prediksi yang lebih tepat pada ketepatan waktu kelulusan mahasiswa dibandingkan dengan algoritma klasifikasi lainnya.

4.3 Keterbatasan Penelitian

Meskipun Random Forest menunjukkan hasil terbaik dalam penelitian ini, ada beberapa keterbatasan yang perlu dicatat. Salah satunya adalah kompleksitas komputasi yang lebih tinggi dibandingkan dengan algoritma Naive Bayes dan C4.5, yang dapat menjadi masalah ketika diimplementasikan pada dataset yang sangat besar. Oleh karena itu, dalam implementasi nyata, pemilihan algoritma harus mempertimbangkan faktor kecepatan dan sumber daya yang tersedia (Rahman, Maulana, & Lestari, 2020).

Selain itu, Naive Bayes meskipun memiliki keterbatasan dalam menangani ketergantungan antar fitur, tetap dapat digunakan pada situasi dengan data yang lebih sederhana dan jumlah fitur yang tidak terlalu besar. Oleh karena itu, meskipun hasilnya lebih

rendah dalam penelitian ini, Naive Bayes tetap relevan dalam konteks penggunaan yang lebih sederhana.

4.4 Aplikasi dalam Sistem Informasi Akademik

Penelitian ini memiliki implikasi yang signifikan untuk pengembangan sistem informasi akademik, khususnya dalam memprediksi kelulusan mahasiswa. Dengan menggunakan Random Forest, institusi pendidikan dapat memprediksi dengan lebih tepat mahasiswa yang kemungkinan besar akan lulus tepat waktu atau tidak.

Informasi ini dapat digunakan untuk mengambil keputusan yang lebih baik dalam hal bimbingan akademik, pengelolaan kelas, serta perencanaan program studi yang lebih efektif (Sari, Nugroho, & Hartati, 2021).

5. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan beberapa hal sebagai berikut:

1. **Random Forest merupakan** algoritma klasifikasi yang paling unggul dalam memprediksi kelulusan mahasiswa dengan akurasi tertinggi yaitu 92.45%. Algoritma ini juga menunjukkan performa yang lebih baik dibandingkan dengan **C4.5 dan Naive Bayes** dalam hal presisi, recall, dan F1-score. Hal ini menunjukkan bahwa **Random Forest** lebih efektif dalam menangani data dengan ketergantungan antar fitur yang kompleks.
2. **C4.5** memberikan hasil yang cukup baik dengan akurasi sebesar 89.10%, yang menempatkannya pada posisi kedua setelah **Random Forest**. Meskipun tidak sebaik **Random Forest**, **C4.5** tetap menjadi pilihan yang baik untuk kasus-kasus yang tidak membutuhkan kompleksitas model yang tinggi.
3. **Naive Bayes** menunjukkan performa yang sedikit lebih rendah dibandingkan kedua algoritma lainnya, dengan akurasi 84.30%. Meskipun demikian, **Naive Bayes** tetap relevan untuk digunakan pada dataset yang lebih sederhana dengan jumlah fitur yang lebih kecil dan kurang kompleks.
4. Dalam penelitian ini, **Random Forest** diidentifikasi sebagai algoritma yang paling cocok untuk diterapkan dalam sistem prediksi kelulusan mahasiswa berbasis data mining, terutama untuk memprediksi kelulusan tepat waktu, dengan mempertimbangkan keunggulannya dalam akurasi dan kemampuan menangani data yang kompleks.
5. Penelitian ini juga menunjukkan pentingnya pemilihan algoritma yang tepat berdasarkan karakteristik data yang tersedia. Keputusan ini berpengaruh langsung pada akurasi dan efektivitas sistem prediksi yang dibangun.

Dengan demikian, algoritma **Random Forest** dapat direkomendasikan untuk pengembangan sistem prediksi kelulusan mahasiswa dalam institusi pendidikan, sementara **C4.5** dan **Naive Bayes** dapat dipertimbangkan untuk aplikasi yang lebih sederhana atau pada dataset yang lebih kecil. Ke depan, penelitian lebih lanjut dapat dilakukan untuk menguji algoritma-algoritma lain dalam konteks yang lebih luas atau pada dataset yang lebih beragam.

6. DAFTAR PUSTAKA

Han, J., Pei, J., & Kamber, M. (2018). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Kusuma, A. R., & Nugroho, R. A. (2020). Analisis prediksi kelulusan mahasiswa menggunakan algoritma C4.5 dan Naive Bayes. Jurnal Teknologi Informasi dan Ilmu Komputer, 7(2), 215–222. https://doi.org/10.25126/jtiik.2020721531

Putra, H. A., & Fadillah, R. (2022). Penerapan Random Forest dalam prediksi ketepatan

kelulusan mahasiswa. Jurnal Sistem Informasi, 18(3), 185–192.

Rahman, F., Maulana, A., & Lestari, P. (2020). Evaluasi kinerja algoritma klasifikasi untuk prediksi mahasiswa lulus tepat waktu. Jurnal Ilmiah Teknologi dan Rekayasa, 25(1), 11–18.

Sari, N. P., Nugroho, M. A., & Hartati, S. (2021). Penerapan algoritma Naive Bayes untuk prediksi kelulusan mahasiswa. Jurnal Komputer dan Informatika, 19(1), 44–51.

Yulianti, N., Pramudyo, A. S., & Wibowo, A. (2019). Analisis faktor yang mempengaruhi kelulusan mahasiswa menggunakan metode data mining. Jurnal Ilmiah SINTECH (Science and Technology), 2(2), 45–52.