**JUTEK – JURNAL TEKNOLOGI**
https://e-journalbattuta.ac.id/index.php/jutek

# Komparasi Algoritma K-Nearest Neighbor, Naïve Bayes, dan Random Forest dengan SMOTE untuk Prediksi Dropout Mahasiswa

Randi Sumitro[1], Juni Ismail[2]
[1]Program Studi Sistem Informasi, STIKOM Tunas Bangsa, Kota Pematangsiantar, Sumatera Utara, Indonesia
[2]Program Studi Teknik Komputer, Politeknik Bisnis Indonesia, Simalungun, Indonesia
Corespondensi : randisumitro2@gmail.com

**ABSTRAK**

Angka putus kuliah (*dropout*) merupakan persoalan serius bagi perguruan tinggi karena berdampak langsung pada akreditasi institusi, efisiensi pembiayaan pendidikan, dan masa depan mahasiswa itu sendiri. Deteksi dini terhadap mahasiswa berisiko menjadi kunci agar intervensi akademik dapat dilakukan tepat waktu. Penelitian ini membandingkan kinerja tiga algoritma klasifikasi, yaitu K-Nearest Neighbor (KNN), Naïve Bayes, dan Random Forest, untuk memprediksi status mahasiswa pada dataset publik Predict Students' Dropout and Academic Success yang memuat 4.424 instance dengan 34 fitur dan tiga kelas target (Dropout, Enrolled, Graduate). Ketidakseimbangan kelas pada data latih ditangani menggunakan Synthetic Minority Over-sampling Technique (SMOTE), sedangkan data uji dipertahankan pada distribusi aslinya dengan pembagian 80:20 secara stratified. Hasil pengujian menunjukkan Random Forest memberikan kinerja terbaik dengan akurasi 76,72% dan F1-score makro 0,7134, mengungguli Naïve Bayes (akurasi 65,99%) dan KNN (akurasi 61,24%). Validasi silang lima lipatan memperkuat temuan tersebut dengan F1 makro 0,6887 ± 0,0099 sekaligus simpangan baku terkecil di antara ketiga model. Analisis *feature importance* mengungkap bahwa kinerja akademik pada dua semester pertama, khususnya jumlah mata kuliah yang lulus dan rata-rata nilai, merupakan prediktor paling dominan. Temuan ini dapat dimanfaatkan perguruan tinggi sebagai dasar pembangunan sistem peringatan dini (*early warning system*) untuk menekan angka putus kuliah.

Kata kunci: *Prediksi Dropout, Random Forest, K-Nearest Neighbor, Naïve Bayes, SMOTE*

## PENDAHULUAN

Putus kuliah masih menjadi salah satu persoalan paling krusial dalam penyelenggaraan pendidikan tinggi. Mahasiswa yang berhenti sebelum menyelesaikan studinya tidak hanya kehilangan kesempatan memperoleh kualifikasi akademik, tetapi juga menimbulkan kerugian bagi institusi berupa menurunnya rasio kelulusan, terganggunya penilaian akreditasi, serta tidak efisiennya sumber daya pembelajaran yang telah dialokasikan (Mariano et al., 2022). Pada konteks perguruan tinggi di Indonesia, tingginya angka *dropout* juga berkaitan erat dengan reputasi program studi di mata calon mahasiswa dan pemangku kepentingan (Samasil et al., 2022). Oleh sebab itu, kemampuan mengenali mahasiswa berisiko sedini mungkin menjadi kebutuhan nyata bagi pengelola institusi pendidikan.

Perkembangan *machine learning* membuka peluang besar untuk menjawab kebutuhan tersebut. Rekam jejak akademik, latar belakang demografis, hingga kondisi sosial ekonomi mahasiswa yang tersimpan dalam sistem informasi akademik dapat diolah menjadi model prediktif status studi. Realinho et al. (2022) bahkan secara khusus menyediakan dataset benchmark publik yang menggabungkan ketiga dimensi tersebut agar penelitian prediksi *dropout* dapat dilakukan secara terbuka dan dapat direplikasi. Ketersediaan data semacam ini

memungkinkan perbandingan antaralgoritma dilakukan pada pijakan yang sama, sesuatu yang sulit dicapai apabila setiap penelitian hanya menggunakan data internal institusinya masing-masing.

Di antara berbagai algoritma klasifikasi, K-Nearest Neighbor (KNN), Naïve Bayes, dan Random Forest termasuk yang paling sering digunakan pada data tabular pendidikan. KNN mengklasifikasikan sebuah instance berdasarkan mayoritas label tetangga terdekatnya di ruang fitur (Cover & Hart, 1967), Naïve Bayes memanfaatkan teorema Bayes dengan asumsi independensi antarfitur sehingga sangat ringan secara komputasi, sedangkan Random Forest membangun gabungan banyak pohon keputusan melalui mekanisme *bagging* dan pemilihan fitur acak sehingga tangguh terhadap *overfitting* (Breiman, 2001). Ketiganya mewakili tiga paradigma berbeda, yakni berbasis jarak, berbasis probabilitas, dan berbasis *ensemble*, sehingga menarik untuk diadu pada kasus yang sama.

Sejumlah penelitian terdahulu telah membandingkan algoritma-algoritma tersebut pada ranah pendidikan tinggi. Sejati et al. (2022) membandingkan Naïve Bayes, KNN, dan Random Forest pada data penerimaan mahasiswa baru dan menemukan Random Forest paling unggul dengan akurasi 73,61%. Rofi et al. (2024) memperbandingkan KNN dan Random Forest untuk klasifikasi mahasiswa berpotensi *dropout* pada data internal sebuah perguruan tinggi dan kembali menempatkan Random Forest di posisi teratas. Putri et al. (2023) mengkaji KNN, Naïve Bayes, dan SVM untuk prediksi kelulusan mahasiswa tingkat akhir, sementara Samasil et al. (2022) menerapkan Naïve Bayes dan Decision Tree untuk mengklasifikasikan mahasiswa berpotensi putus studi.

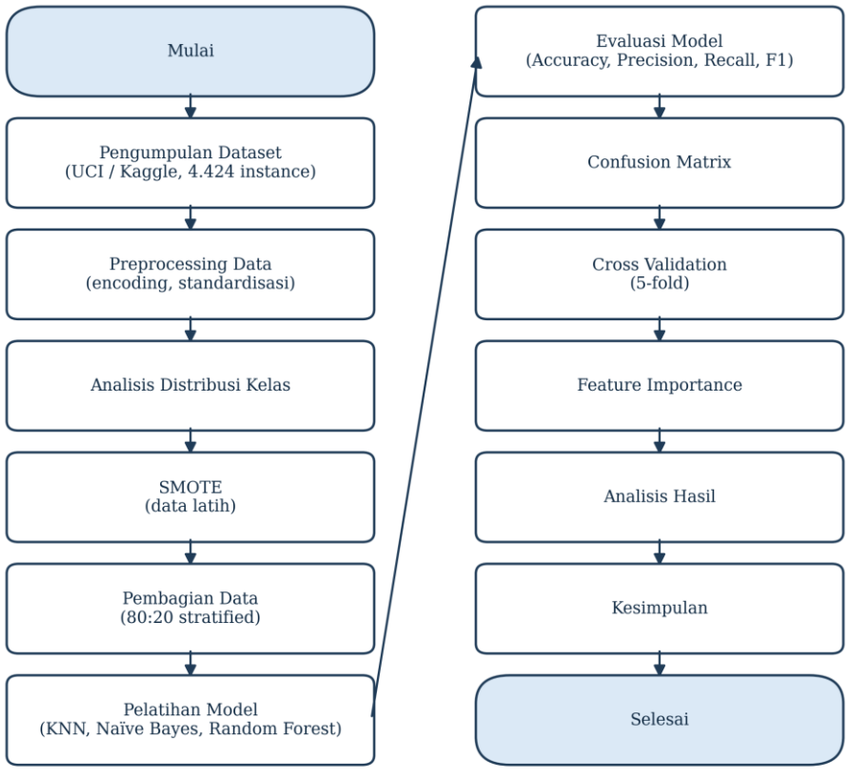
Meskipun temuan-temuan tersebut konsisten menunjukkan keunggulan metode *ensemble*, terdapat beberapa keterbatasan yang berulang. Pertama, mayoritas penelitian menggunakan data internal institusi yang tidak tersedia untuk publik, sehingga hasilnya sulit direplikasi dan dibandingkan secara adil. Kedua, persoalan ketidakseimbangan kelas yang hampir selalu muncul pada data status studi jarang ditangani secara eksplisit, padahal kelas minoritas justru sering menjadi kelas yang paling penting untuk dikenali. Ketiga, sebagian besar studi menyederhanakan persoalan menjadi klasifikasi biner, padahal status mahasiswa pada praktiknya tidak hanya lulus atau berhenti, tetapi juga masih aktif menempuh studi.

Berangkat dari celah tersebut, penelitian ini mengomparasikan KNN, Naïve Bayes, dan Random Forest pada dataset benchmark publik tiga kelas dengan penanganan ketidakseimbangan kelas menggunakan Synthetic Minority Over-sampling Technique atau SMOTE (Chawla et al., 2002) yang diterapkan hanya pada data latih agar protokol evaluasi tetap sahih. Pembeda penelitian ini terhadap studi sebelumnya terletak pada kombinasi penggunaan dataset publik tiga kelas, penerapan SMOTE, validasi silang lima lipatan sebagai uji kestabilan, serta analisis *feature importance* yang diarahkan menjadi rekomendasi kebijakan institusi.

Tujuan penelitian ini adalah memperoleh model klasifikasi terbaik untuk memprediksi status studi mahasiswa sekaligus mengidentifikasi faktor-faktor yang paling menentukan risiko putus kuliah. Kontribusi yang diberikan mencakup bukti empiris perbandingan tiga paradigma klasifikasi pada protokol evaluasi yang ketat dan dapat direplikasi, serta rumusan implikasi praktis berupa rancangan dasar sistem peringatan dini akademik bagi perguruan tinggi.

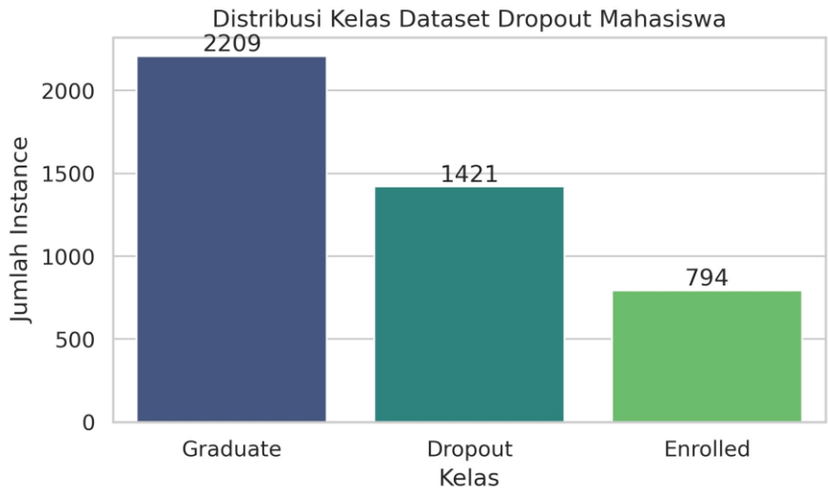
## METODE PENELITIAN

Penelitian ini merupakan penelitian kuantitatif eksperimental dengan pendekatan komparatif terhadap tiga algoritma klasifikasi. Seluruh tahapan eksperimen dijalankan pada lingkungan Google Colaboratory menggunakan pustaka *scikit-learn* (Pedregosa et al., 2011) dan *imbalanced-learn* dengan *random seed* 42 agar hasil dapat direproduksi. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1, dimulai dari pengumpulan dataset hingga penarikan kesimpulan.




**Gambar 1. Diagram Alir Penelitian**

Dataset yang digunakan adalah Predict Students' Dropout and Academic Success yang dipublikasikan Realinho et al. (2022) dan tersedia secara terbuka pada repositori UCI Machine Learning maupun Kaggle; penelitian ini menggunakan salinan publik pada Kaggle. Dataset ini berisi 4.424 catatan mahasiswa sebuah institusi pendidikan tinggi dengan 34 fitur yang mencakup dimensi akademik (jumlah mata kuliah yang diambil, dievaluasi, dan diluluskan beserta nilainya pada semester pertama dan kedua), demografis (usia saat mendaftar, jenis kelamin, status pernikahan, pekerjaan dan pendidikan orang tua), serta makroekonomi (tingkat pengangguran, inflasi, dan PDB). Variabel target terdiri atas tiga kelas, yaitu Graduate sebanyak 2.209 instance (49,9%), Dropout sebanyak 1.421 instance (32,1%), dan Enrolled sebanyak 794 instance (17,9%). Hasil eksplorasi menunjukkan tidak terdapat nilai hilang maupun data duplikat, sehingga tidak diperlukan imputasi. Distribusi kelas yang tidak seimbang tersebut divisualisasikan pada Gambar 2 dan menjadi dasar penerapan teknik *oversampling*.




**Gambar 2. Distribusi Kelas Dataset**

Tahap prapemrosesan diawali dengan *label encoding* pada variabel target sehingga ketiga kelas terpetakan ke nilai numerik. Data kemudian dibagi menjadi data latih dan data uji dengan proporsi 80:20 secara *stratified* agar komposisi kelas pada kedua himpunan tetap proporsional, menghasilkan 3.539 instance data latih (1.767 Graduate, 1.137 Dropout, 635 Enrolled) dan 885 instance data uji (442 Graduate, 284 Dropout, 159 Enrolled). Standardisasi fitur dilakukan menggunakan *StandardScaler* yang hanya di-*fit* pada data latih untuk mencegah

kebocoran data (*data leakage*), mengingat KNN sangat sensitif terhadap perbedaan skala fitur. Ketidakseimbangan kelas selanjutnya ditangani dengan SMOTE (Chawla et al., 2002), yaitu teknik *oversampling* yang membangkitkan instance sintetis kelas minoritas melalui interpolasi antartetangga terdekat. SMOTE diterapkan secara eksklusif pada data latih sehingga setiap kelas berjumlah sama dengan kelas mayoritas, yakni 1.767 instance per kelas atau total 5.301 instance, sedangkan data uji tetap berada pada distribusi aslinya agar pengukuran kinerja mencerminkan kondisi nyata.

Tiga algoritma kemudian dilatih pada data hasil SMOTE. Pada KNN, nilai k optimal ditentukan melalui validasi silang lima lipatan pada data latih terhadap kandidat k = 3, 5, 7, 9, dan 11; nilai terpilih adalah k = 3. Naïve Bayes diimplementasikan dalam varian Gaussian dengan parameter bawaan karena seluruh fitur telah berbentuk numerik. Random Forest dibangun dengan 100 pohon keputusan (*n_estimators* = 100) mengikuti mekanisme *bootstrap aggregating* dan pemilihan subhimpunan fitur secara acak pada setiap simpul (Breiman, 2001). Pemilihan ketiganya didasarkan pada keterwakilan paradigma klasifikasi yang berbeda sebagaimana diuraikan pada bagian pendahuluan.

Kinerja model diukur pada data uji menggunakan empat metrik yang diturunkan dari *confusion matrix*, yaitu akurasi, presisi, *recall*, dan F1-score sebagaimana Persamaan (1) sampai (4). Mengingat distribusi kelas tidak seimbang, presisi, *recall*, dan F1-score dilaporkan dalam rata-rata makro agar setiap kelas memperoleh bobot yang sama. Kestabilan model diuji lebih lanjut melalui validasi silang *stratified* lima lipatan dengan metrik F1 makro yang dilaporkan sebagai rerata beserta simpangan bakunya. Sebagai pelengkap analisis, tingkat kepentingan fitur (*feature importance*) diekstraksi dari model Random Forest berdasarkan penurunan impuritas Gini untuk mengidentifikasi faktor-faktor yang paling menentukan prediksi.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

## HASIL

Ketiga model yang telah dilatih pada data hasil SMOTE diuji menggunakan 885 instance data uji yang mempertahankan distribusi kelas asli. Rekapitulasi kinerja keempat metrik beserta waktu pelatihan disajikan pada Tabel 1. Random Forest mencatatkan kinerja tertinggi pada seluruh metrik dengan akurasi 76,72% dan F1 makro 0,7134, terpaut lebih dari sepuluh poin persentase di atas Naïve Bayes (65,99%) dan lima belas poin di atas KNN (61,24%). Konsekuensi dari kompleksitas *ensemble* terlihat pada waktu pelatihan Random Forest yang mencapai 2,19 detik, jauh lebih lama dibandingkan KNN (0,0015 detik) maupun Naïve Bayes (0,0123 detik), meskipun durasi tersebut masih sangat ringan untuk kebutuhan praktis.

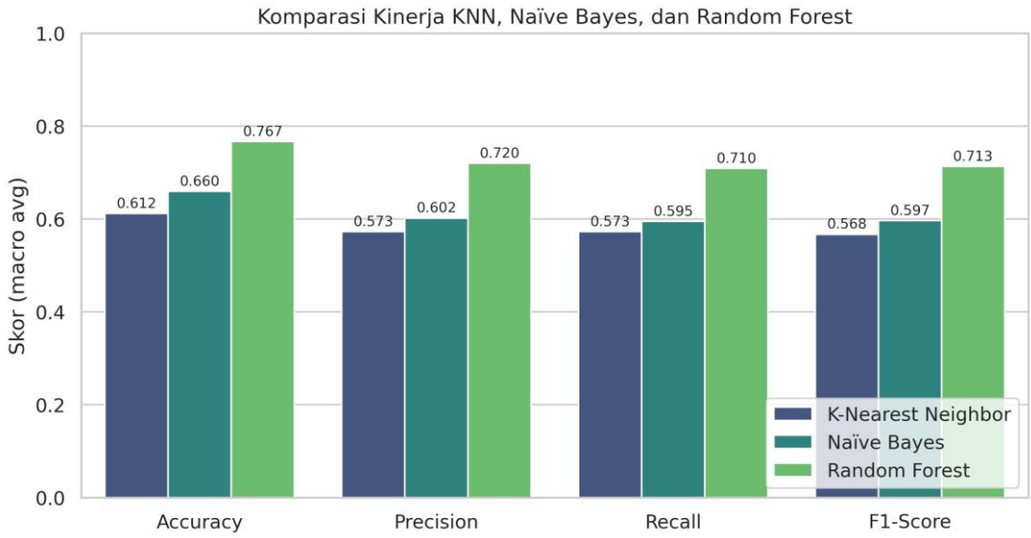
**Tabel 1. Komparasi Kinerja Model pada Data Uji (rata-rata makro)**

| Model | Accuracy | Precision | Recall | F1-Score | Waktu Latih (s) |
|---|---|---|---|---|---|
| K-Nearest Neighbor | 0,6124 | 0,5730 | 0,5728 | 0,5675 | 0,0015 |
| Naïve Bayes | 0,6599 | 0,6024 | 0,5950 | 0,5969 | 0,0123 |
| Random Forest | 0,7672 | 0,7202 | 0,7095 | 0,7134 | 2,1898 |

Rincian kinerja per kelas pada Tabel 2 memperlihatkan pola yang konsisten pada ketiga model: kelas Graduate dan Dropout relatif mudah dikenali, sedangkan kelas Enrolled menjadi titik terlemah. Pada Random Forest, F1-score kelas Dropout mencapai 0,7692 dengan presisi 0,8233, dan kelas Graduate mencapai 0,8537, sementara kelas Enrolled hanya 0,5171. Pola serupa dengan tingkat yang lebih rendah terjadi pada Naïve Bayes (F1 Enrolled 0,3248) dan KNN (F1 Enrolled 0,3351). Perbandingan visual keempat metrik makro antarmodel disajikan pada Gambar 3.

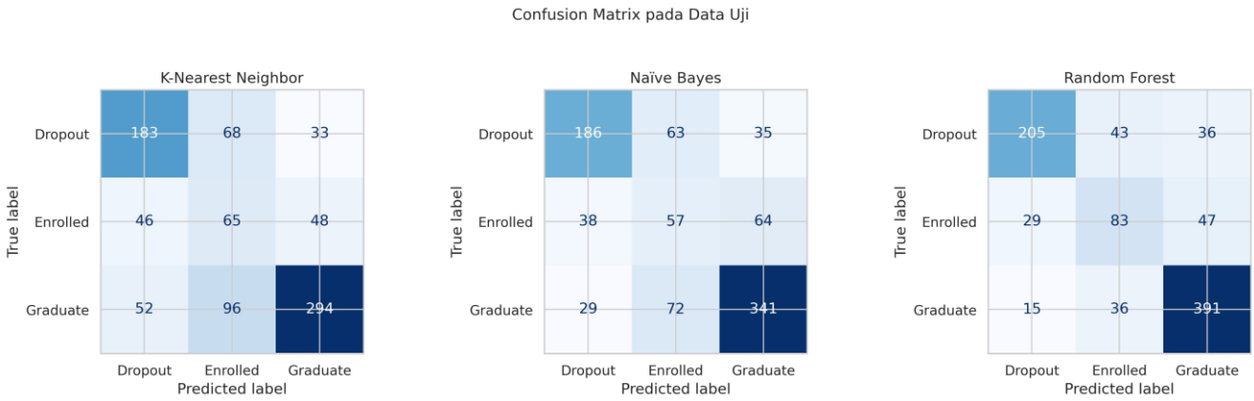
**Tabel 2. F1-Score per Kelas pada Data Uji**

| Kelas | KNN | Naïve Bayes | Random Forest |
|---|---|---|---|
| Dropout (n=284) | 0,6478 | 0,6927 | 0,7692 |
| Enrolled (n=159) | 0,3351 | 0,3248 | 0,5171 |
| Graduate (n=442) | 0,7197 | 0,7732 | 0,8537 |




**Gambar 3. Perbandingan Metrik Kinerja Antarmodel**

Pola kesalahan klasifikasi setiap model dapat ditelusuri melalui *confusion matrix* pada Gambar 4. Pada Random Forest, dari 284 mahasiswa Dropout pada data uji, sebanyak 205 terprediksi benar, 43 keliru sebagai Enrolled, dan hanya 36 yang keliru sebagai Graduate. Dari 159 mahasiswa Enrolled, kesalahan terbesar mengarah ke Graduate (47 instance) dan Dropout (29 instance). Sebaliknya pada KNN, sebanyak 96 dari 442 mahasiswa Graduate justru terprediksi sebagai Enrolled, yang menjelaskan rendahnya akurasi model tersebut.




**Gambar 4. Confusion Matrix Ketiga Model pada Data Uji**

Hasil validasi silang lima lipatan pada Tabel 3 mengonfirmasi konsistensi peringkat ketiga model. Random Forest kembali menempati posisi teratas dengan F1 makro 0,6887 sekaligus simpangan baku terkecil (0,0099), sedangkan KNN (0,5968 ± 0,0145) dan Naïve Bayes (0,5978 ± 0,0105) berada pada level yang hampir sama. Simpangan baku yang kecil pada seluruh model menunjukkan bahwa keunggulan Random Forest bukan kebetulan satu kali pembagian data, melainkan stabil pada berbagai partisi.

**Tabel 3. Hasil Validasi Silang 5-Fold (F1 Makro)**

| Model | F1-Makro (rerata) | Simpangan Baku |
|---|---|---|
| K-Nearest Neighbor | 0,5968 | 0,0145 |
| Naïve Bayes | 0,5978 | 0,0105 |
| Random Forest | 0,6887 | 0,0099 |

## PEMBAHASAN

Keunggulan Random Forest dapat dijelaskan dari karakteristik datanya. Dataset ini memuat 34 fitur heterogen

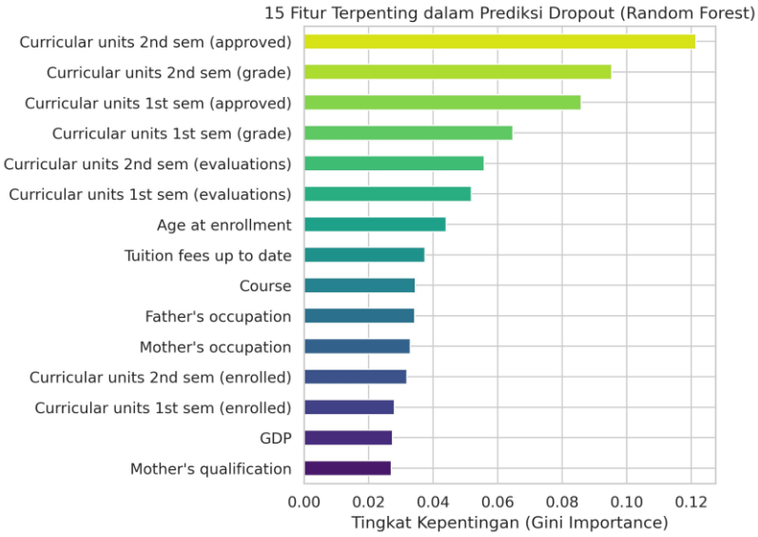
yang saling berinteraksi secara nonlinier; mekanisme *ensemble* yang menggabungkan ratusan pohon keputusan dengan subhimpunan fitur acak mampu menangkap interaksi semacam itu sekaligus meredam varians model tunggal (Breiman, 2001). Sebaliknya, KNN bekerja berbasis jarak pada ruang berdimensi 34 sehingga terdampak fenomena *curse of dimensionality* yang membuat jarak antarinstance menjadi kurang diskriminatif. Adapun Naïve Bayes dirugikan oleh asumsi independensi antarfitur, padahal fitur-fitur akademik semester pertama dan kedua pada dataset ini jelas berkorelasi kuat.

Penerapan SMOTE memberikan kontribusi nyata terutama pada kemampuan model mengenali kelas minoritas. Tanpa penyeimbangan, kelas Enrolled yang hanya 17,9% dari populasi cenderung terabaikan oleh model yang mengejar akurasi global. Setelah SMOTE, *recall* kelas Enrolled mencapai 0,4088 pada KNN, 0,3585 pada Naïve Bayes, dan 0,5220 pada Random Forest, suatu tingkat pengenalan yang sulit dicapai pada data sangat timpang. Meskipun demikian, presisi kelas Enrolled tetap moderat, yang menandakan *oversampling* sintetis tidak sepenuhnya menghapus tumpang-tindih antar kelas pada ruang fitur.

Rendahnya kinerja seluruh model pada kelas Enrolled justru memberikan wawasan substantif. Berbeda dengan Dropout dan Graduate yang merupakan status akhir, Enrolled adalah status transisional: mahasiswa pada kelas ini secara profil akademik berada di antara mahasiswa yang akan lulus dan yang akan berhenti, sehingga batas kelasnya secara alami kabur. *Confusion matrix* Random Forest memperlihatkan kesalahan kelas Enrolled tersebar hampir berimbang ke arah Graduate dan Dropout, bukan terkonsentrasi ke satu kelas, yang memperkuat dugaan tumpang-tindih profil tersebut. Temuan serupa mengenai sulitnya kelas antara juga tersirat pada penelitian tiga kelas lain yang memakai dataset sejenis (Realinho et al., 2022).

Dari sudut pandang praktis, pola kesalahan Random Forest tergolong menguntungkan. Presisi kelas Dropout sebesar 0,8233 berarti ketika model menandai seorang mahasiswa berisiko berhenti, sekitar delapan dari sepuluh penandaan tersebut tepat sasaran, sehingga sumber daya pendampingan akademik tidak banyak terbuang. Sementara itu, hanya 12,7% mahasiswa Dropout yang keliru diprediksi sebagai Graduate, jenis kesalahan paling berbahaya karena membuat mahasiswa berisiko luput dari perhatian. Karakteristik inilah yang menjadikan Random Forest layak diposisikan sebagai mesin inferensi pada sistem peringatan dini (*early warning system*) akademik yang terintegrasi dengan sistem informasi kampus.

Analisis *feature importance* pada Gambar 5 menunjukkan dominasi fitur akademik dua semester pertama. Jumlah mata kuliah yang diluluskan pada semester kedua menempati peringkat teratas (0,121), disusul rata-rata nilai semester kedua (0,095), jumlah mata kuliah yang diluluskan semester pertama (0,086), rata-rata nilai semester pertama (0,065), serta jumlah evaluasi pada kedua semester. Fitur nonakademik baru muncul pada peringkat berikutnya, yaitu usia saat mendaftar dan status ketepatan pembayaran uang kuliah, yang menandakan dimensi ekonomi dan demografis tetap berkontribusi meskipun tidak sekuat capaian akademik. Implikasinya jelas: jendela intervensi paling efektif berada pada tahun pertama perkuliahan, ketika sinyal risiko sudah terbaca dari kelulusan mata kuliah dan nilai, namun status akhir mahasiswa belum terbentuk.




**Gambar 5. Lima Belas Fitur Terpenting Berdasarkan Random Forest**

Apabila disandingkan dengan penelitian terdahulu, temuan ini sejalan sekaligus melengkapi. Akurasi Random

Forest sebesar 76,72% pada penelitian ini sedikit melampaui capaian Sejati et al. (2022) sebesar 73,61% pada tugas prediksi yang lebih sederhana (dua kelas) dengan data jauh lebih besar. Adapun akurasi 99,05% yang dilaporkan Rofi et al. (2024) dicapai pada data internal berukuran kecil dengan karakteristik berbeda, sehingga tidak dapat diperbandingkan secara langsung; justru di sinilah nilai penggunaan dataset benchmark publik pada penelitian ini, yaitu memungkinkan replikasi dan perbandingan yang adil oleh peneliti lain. Kelebihan penelitian ini terletak pada protokol evaluasi yang ketat, meliputi standardisasi dan SMOTE yang hanya diterapkan pada data latih, pelaporan metrik makro, serta validasi silang sebagai uji kestabilan.

Penelitian ini tidak lepas dari keterbatasan. Pertama, dataset bersumber dari institusi pendidikan tinggi di Portugal sehingga generalisasi temuan ke konteks perguruan tinggi Indonesia memerlukan validasi pada data lokal. Kedua, perbandingan dibatasi pada tiga algoritma klasik tanpa penalaan hiperparameter ekstensif maupun uji signifikansi statistik antarmodel. Ketiga, kinerja kelas Enrolled yang masih moderat menunjukkan ruang perbaikan, misalnya melalui pendekatan *cost-sensitive learning*, teknik penyeimbangan hibrida, atau perumusan ulang masalah menjadi klasifikasi dua tahap.

## KESIMPULAN

Penelitian ini berhasil membandingkan KNN, Naïve Bayes, dan Random Forest untuk prediksi status studi mahasiswa pada dataset publik tiga kelas dengan penanganan ketidakseimbangan menggunakan SMOTE. Random Forest terbukti menjadi model terbaik dengan akurasi 76,72%, presisi makro 0,7202, *recall* makro 0,7095, dan F1 makro 0,7134, serta menunjukkan kestabilan tertinggi pada validasi silang lima lipatan (0,6887 ± 0,0099). Penerapan SMOTE pada data latih meningkatkan kemampuan model mengenali kelas minoritas tanpa mencederai keabsahan evaluasi. Analisis *feature importance* menegaskan bahwa capaian akademik dua semester pertama, terutama jumlah mata kuliah yang lulus dan rata-rata nilai, merupakan penentu utama risiko putus kuliah, disusul faktor usia saat mendaftar dan ketepatan pembayaran uang kuliah. Bagi institusi pendidikan, temuan ini menegaskan pentingnya pemantauan intensif mahasiswa tahun pertama serta kelayakan Random Forest sebagai mesin prediksi pada sistem peringatan dini akademik. Penelitian lanjutan disarankan menguji model pada data perguruan tinggi Indonesia, menambahkan penalaan hiperparameter dan uji signifikansi statistik seperti uji McNemar, mengeksplorasi algoritma *gradient boosting* serta teknik penyeimbangan alternatif, dan mengembangkan purwarupa sistem peringatan dini yang terintegrasi dengan sistem informasi akademik agar manfaat praktisnya dapat diukur secara langsung.

## DAFTAR PUSTAKA


Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Mariano, A. M., Ferreira, A. B. de M. L., Santos, M. R., Castilho, M. L., & Bastos, A. C. F. L. C. (2022). Decision trees for predicting dropout in engineering course students in Brazil. Procedia Computer Science, 214, 1113–1120. https://doi.org/10.1016/j.procs.2022.11.285

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Putri, A., Hardiana, C. S., Novfuja, E., Siregar, F. T. P., Rahmaddeni, Fatma, Y., & Wahyuni, R. (2023). Komparasi algoritma K-NN, Naive Bayes dan SVM untuk prediksi kelulusan mahasiswa tingkat akhir. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 3(1), 20–26. https://doi.org/10.57152/malcom.v3i1.610

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. Data, 7(11), 146. https://doi.org/10.3390/data7110146

Rofi, M. M., Setiawan, F. A., & Riana, F. (2024). Perbandingan metode K-NN dan Random Forest pada klasifikasi mahasiswa berpotensi dropout. INFOTECH Journal, 10(1). https://doi.org/10.31949/infotech.v10i1.8856

Samasil, S., Yuyun, & Hazriani. (2022). Klasifikasi mahasiswa berpotensi drop out menggunakan algoritma Naive Bayes dan Decision Tree. Jurnal Ilmiah Ilmu Komputer, 8(2), 108–114. https://doi.org/10.35329/jiik.v8i2.242

Sejati, P., Munawar, Pilliang, M., & Akbar, H. (2022). Studi komparasi Naive Bayes, K-Nearest Neighbor, dan Random Forest untuk prediksi calon mahasiswa yang diterima atau mundur. Jurnal Teknologi Informasi dan Ilmu Komputer, 9(7). https://doi.org/10.25126/jtiik.2022976737