



Klasifikasi Kelulusan Mahasiswa dengan KNN Berbasis Data Akademik

Faiz Bimo, Erza Sofian

Program Studi Sistem Informasi, Universitas Ary Ginanjar

f.bimo@students.esqbs.ac.id, erza.sofian@esqbs.ac.id

Abstrak

Kelulusan tepat waktu merupakan salah satu indikator efektivitas proses akademik dan dapat didukung melalui identifikasi dini mahasiswa yang berisiko mengalami keterlambatan studi. Penelitian ini mengevaluasi algoritma K-Nearest Neighbor (KNN) untuk klasifikasi biner status kelulusan mahasiswa berdasarkan rekam akademik. Naskah sumber melaporkan 540 data lulusan valid dari program studi Manajemen, Sistem Informasi, dan Ilmu Komputer, dengan variabel akademik berupa indeks prestasi kumulatif (IPK), tahun masuk, tahun lulus, dan jumlah SKS lulus. Status kelulusan ditetapkan tepat waktu apabila masa studi tidak melebihi empat tahun. Implementasi yang terdokumentasi meliputi pembersihan data, transformasi StandardScaler, Synthetic Minority Over-sampling Technique (SMOTE), pembagian data 80:20, serta GridSearchCV dengan validasi silang lima lipatan untuk memilih nilai K pada rentang 1-30. Eksperimen yang konsisten pada bagian hasil memilih K=1 dengan rerata akurasi validasi silang 90,97%. Pada 161 data uji, model mengklasifikasikan 148 observasi secara benar dan menghasilkan akurasi 91,93%, macro-precision 92,10%, macro-recall 91,91%, dan macro-F1 91,91%. Hasil tersebut menunjukkan KNN mampu memisahkan dua kelas kelulusan dengan baik pada dataset yang dilaporkan. Namun, karena scaling dan SMOTE pada alur yang terdokumentasi dilakukan sebelum pemisahan data latih dan uji, performa tersebut perlu ditafsirkan secara hati-hati karena berpotensi mengalami kebocoran informasi. Validasi lanjutan perlu menempatkan scaling dan SMOTE hanya pada data latih atau di dalam setiap fold validasi silang serta mempertahankan data uji yang benar-benar tidak tersentuh sebelum model digunakan sebagai sistem pendukung keputusan akademik.

Kata kunci: K-Nearest Neighbor, Kelulusan Mahasiswa, Klasifikasi, Data Akademik, SMOTE

Abstract

Timely graduation is an important indicator of academic process effectiveness and can be supported by data-driven early identification of students at risk of delayed completion. This study evaluates the K-Nearest Neighbor (KNN) algorithm for binary classification of student graduation status using academic records. The source manuscript reports 540 valid graduate records drawn from management, information systems, and computer science programs, with academic variables including grade point average (GPA), admission year, graduation year, and completed credits. Graduation status is defined as on time when the study period is no more than four years. The reported implementation applies data cleaning, StandardScaler transformation, Synthetic Minority Over-sampling Technique (SMOTE), an 80:20 train-test split, and GridSearchCV with five-fold cross-validation to select K from 1 to 30. The documented experiment selects K=1 with a mean cross-validation accuracy of 90.97%. On 161 test observations, the model correctly classifies 148 cases and produces an accuracy of 91.93%, macro-precision of 92.10%, macro-recall of 91.91%, and macro-F1 of 91.91%. The findings indicate that KNN can separate the two graduation classes effectively in the reported dataset. Nevertheless, because scaling and SMOTE were executed before the train-test split in the documented workflow, the reported performance should be interpreted cautiously due to possible information leakage. Future validation should place scaling and SMOTE exclusively inside the training folds and evaluate the model on an untouched test set before operational deployment.

Keywords: K-Nearest Neighbor, student Graduation, Classification, Academic Data, SMOTE

1. Pendahuluan

Ketepatan waktu kelulusan merupakan indikator yang relevan bagi evaluasi proses pendidikan tinggi karena mencerminkan kemampuan mahasiswa menyelesaikan beban akademik sesuai kurikulum dan kemampuan institusi mengelola layanan akademik. Pemanfaatan rekam akademik untuk memprediksi kelulusan memungkinkan perguruan tinggi bergerak dari evaluasi retrospektif menuju identifikasi risiko yang lebih dini. Dalam konteks regulasi Indonesia, penjaminan mutu pendidikan tinggi juga menempatkan capaian pembelajaran, beban studi, dan penyelesaian program sebagai bagian dari pengelolaan mutu pendidikan tinggi.

Pendekatan machine learning telah digunakan untuk memetakan pola kelulusan dari data historis mahasiswa. KNN relevan karena melakukan klasifikasi berdasarkan kedekatan observasi pada ruang fitur dan relatif mudah diinterpretasikan. Studi Sari et al., menerapkan KNN pada prediksi kelulusan mahasiswa [1], sedangkan Atmaja et al., menggunakan normalisasi dan jarak Euclidean dalam konteks serupa [2]. Penelitian yang lebih baru juga melaporkan penggunaan KNN untuk prediksi ketepatan waktu kelulusan dengan variasi atribut dan nilai K yang berbeda [3], [4], [5]. Hasil lintas studi tersebut menunjukkan bahwa performa KNN sangat bergantung pada representasi fitur, skala variabel, distribusi kelas, pemilihan K, dan rancangan validasi. Penelitian lain juga memperlihatkan penggunaan KNN untuk prediksi kelulusan mahasiswa dan peramalan status kelulusan dengan karakteristik data yang berbeda [6], [7], [8]. Pada konteks pendidikan menengah, Sari et al., juga menerapkan KNN untuk prediksi kelulusan, yang memperlihatkan portabilitas metode berbasis kedekatan pada data pendidikan dengan unit analisis yang berbeda [9].

Masalah penting pada data kelulusan adalah ketidakseimbangan kelas. Ketika satu kategori jauh lebih dominan, akurasi agregat dapat terlihat tinggi meskipun kemampuan mengenali kelas minoritas lemah. SMOTE merupakan salah satu pendekatan yang sering digunakan untuk membentuk sampel sintesis pada kelas minoritas. Hermanto et al., menunjukkan bahwa penanganan ketidakseimbangan perlu dipertimbangkan secara eksplisit dalam prediksi kelulusan [10]. Namun, oversampling harus ditempatkan hanya pada data latih atau di dalam fold validasi agar observasi sintesis yang diturunkan dari keseluruhan dataset tidak mencemari data evaluasi. Selain KNN, studi komparatif menunjukkan bahwa kinerja klasifikasi kelulusan dapat berbeda ketika dibandingkan dengan Naive Bayes dan SVM, sehingga pemilihan algoritma perlu mempertimbangkan karakteristik dataset dan distribusi kelas [11], [12].

Berdasarkan naskah penelitian awal, dataset kelulusan yang digunakan memiliki distribusi kelas tidak seimbang dan implementasi menggabungkan StandardScaler, SMOTE, KNN, serta GridSearchCV. Celah yang hendak dijawab adalah bagaimana performa KNN pada data akademik tersebut ketika nilai K dipilih secara sistematis, sekaligus bagaimana hasilnya dibaca secara kritis menggunakan precision, recall, F1-score, dan confusion matrix, bukan hanya akurasi. Kebaruan praktis penelitian terletak pada integrasi optimasi K dan evaluasi dua kelas pada data lulusan institusi, disertai penilaian metodologis terhadap alur preprocessing yang menentukan validitas estimasi performa.

Penelitian ini bertujuan mengklasifikasikan status kelulusan tepat waktu dan tidak tepat waktu menggunakan KNN berdasarkan data akademik, menentukan nilai K terbaik melalui GridSearchCV, serta mengevaluasi performa model menggunakan confusion matrix, accuracy, precision, recall, dan F1-score. Hasil penelitian diharapkan menjadi dasar pengembangan sistem peringatan dini akademik, tetapi penggunaan operasional harus didahului validasi ulang dengan prosedur yang bebas kebocoran data.

2. Metode Penelitian

2.1. Desain dan Sumber Data

Penelitian menggunakan pendekatan kuantitatif berbasis klasifikasi supervised learning. Berdasarkan bagian pengolahan data pada naskah sumber, data berasal dari arsip lulusan Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ/Universitas Ary Ginanjar dan memuat 540 baris valid pada sheet 'Daftar Lulusan'. Rentang tahun masuk yang dilaporkan adalah 2013-2023 dan tahun lulus 2017-2025. Dataset mencakup mahasiswa dari program studi Manajemen, Sistem Informasi, dan Ilmu Komputer. Perbedaan penamaan institusi pada naskah sumber dipertahankan sebagai keterbatasan dokumentasi dan perlu diseragamkan oleh penulis sebelum submit final.

Variabel yang secara eksplisit digunakan pada kode pemodelan adalah IPK, tahun masuk, tahun lulus, dan jumlah SKS lulus. Label target 'Tepat Waktu' dibentuk dari masa studi, dengan kelas 1 untuk masa studi tidak lebih dari empat tahun dan kelas 0 untuk masa studi lebih dari empat tahun. Tahun masuk dan tahun lulus merupakan variabel yang secara langsung membentuk masa studi; oleh sebab itu, penggunaannya sebagai prediktor perlu dipahami sebagai klasifikasi retrospektif lulusan, bukan prediksi dini mahasiswa aktif.

2.2. Prapemrosesan dan Pemodelan

Tahap prapemrosesan yang terdokumentasi meliputi penghapusan baris tanpa NIM, konversi atribut numerik, penanganan nilai hilang, pemilihan fitur, standarisasi menggunakan StandardScaler, dan penyeimbangan kelas

menggunakan SMOTE dengan `random_state=42`. Setelah resampling, data dibagi menjadi 80% data latih dan 20% data uji menggunakan stratifikasi. Jumlah hasil resampling yang ditampilkan adalah 804 observasi, masing-masing 402 observasi pada dua kelas; pembagian tersebut menghasilkan 643 data latih dan 161 data uji.

KNN menentukan kelas suatu observasi berdasarkan mayoritas kelas dari K tetangga dengan jarak terdekat. Penelitian menggunakan jarak Euclidean. Untuk dua observasi x dan y dengan p fitur, jarak dihitung sebagai $d(x,y) = \sqrt{\sum_{i=1}^p ((x_i - y_i)^2)}$, untuk $i=1, \dots, p$. Karena KNN berbasis jarak, standarisasi diperlukan agar fitur dengan skala numerik besar tidak mendominasi perhitungan kedekatan. Secara teoritis, karakter KNN sebagai metode berbasis kedekatan membuat pemilihan metrik jarak dan skala fitur menjadi penting; perbandingan Euclidean dan Manhattan juga telah dikaji pada klasifikasi kelulusan mahasiswa [13], [14]

Nilai K dioptimasi menggunakan GridSearchCV dengan validasi silang lima lipatan dan kandidat `n_neighbors=1` sampai 30. Kriteria seleksi yang terdokumentasi adalah accuracy. Model terbaik selanjutnya dievaluasi pada data uji menggunakan confusion matrix, accuracy, precision, recall, dan F1-score. Untuk menjaga reproduktibilitas pada replikasi penelitian, pipeline yang direkomendasikan adalah split data terlebih dahulu, kemudian fitting scaler dan SMOTE hanya pada data latih di dalam setiap fold validasi silang.

2.3. Validasi dan Reproduksiabilitas

Rancangan validasi menentukan apakah skor model merepresentasikan kemampuan generalisasi atau hanya kedekatan terhadap data yang pernah memengaruhi proses pelatihan. Dalam KNN, dua operasi yang sangat sensitif terhadap batas data latih dan data uji adalah standarisasi dan oversampling. StandardScaler menghitung rerata serta simpangan baku dari data yang diberikan pada tahap fitting. Jika statistik tersebut dihitung menggunakan keseluruhan dataset, informasi distribusi data uji ikut memengaruhi transformasi data latih. Demikian pula, SMOTE membentuk observasi sintetis berdasarkan lingkungan tetangga kelas minoritas. Karena itu, SMOTE yang dijalankan sebelum split dapat membuat struktur data uji ikut memengaruhi sampel yang kemudian digunakan untuk melatih model.

Untuk replikasi yang lebih ketat, dataset asli perlu dibagi lebih dahulu secara stratified menjadi training set dan hold-out test set. Selanjutnya, pada setiap fold cross-validation di training set, scaler di-fit hanya pada fold pelatihan, kemudian SMOTE diterapkan pada fold pelatihan yang telah ditransformasi. Fold validasi hanya ditransformasi menggunakan parameter scaler dari fold pelatihan dan tidak di-SMOTE. Setelah nilai K dipilih, pipeline akhir di-fit pada seluruh training set dan hanya sekali dievaluasi pada hold-out test set. Prosedur ini mencegah test set berperan dalam pemilihan parameter maupun pembentukan observasi sintetis.

Reproduksiabilitas juga memerlukan dokumentasi versi perangkat lunak, random seed, definisi setiap atribut, aturan penanganan missing value, dan identitas unit analisis. Naskah sumber telah mendokumentasikan `random_state=42` untuk SMOTE dan split, rentang $K=1-30$, serta lima fold GridSearchCV, tetapi belum melaporkan versi Python/scikit-learn/imbalanced-learn. Informasi tersebut sebaiknya ditambahkan pada replikasi berikutnya karena perubahan implementasi library dapat menghasilkan perbedaan kecil pada proses validasi dan resampling.

2.4. Evaluasi Model

Confusion matrix membedakan prediksi benar dan salah untuk dua kelas. Accuracy dihitung sebagai proporsi seluruh prediksi yang benar. Precision mengukur proporsi prediksi suatu kelas yang benar, recall mengukur proporsi anggota aktual suatu kelas yang berhasil dikenali, sedangkan F1-score merupakan rata-rata harmonik precision dan recall. Karena distribusi kelas awal tidak seimbang, interpretasi penelitian menekankan metrik per kelas dan macro-average selain accuracy.

3. Hasil dan Pembahasan

3.1. Karakteristik Data

Naskah sumber melaporkan 540 data lulusan valid dengan komposisi 348 mahasiswa Manajemen (64,44%), 139 Sistem Informasi (25,74%), dan 53 Ilmu Komputer (9,81%). Statistik deskriptif yang dilaporkan menunjukkan rata-rata masa studi 4,02 tahun (minimum 0,5; maksimum 7; simpangan baku 0,67), rata-rata IPK 3,37 (2,46-3,92; simpangan baku 0,31), serta rata-rata SKS lulus 148,23 (143-160; simpangan baku 2,73).

Tabel 1. Distribusi Program Studi

Program Studi	Jumlah	Persentase
Manajemen	348	64,44%
Sistem Informasi	139	25,74%
Ilmu Komputer	53	9,81%

Dokumen juga menampilkan distribusi target 402 observasi tepat waktu dan 703 tidak tepat waktu (total 1.105), yang tidak konsisten dengan keterangan 540 data lulusan valid. Di bagian preprocessing, setelah SMOTE justru ditampilkan 402 observasi pada masing-masing kelas. Karena sumber tidak menyediakan dataset mentah untuk rekonsiliasi, angka distribusi 402:703 tidak digunakan untuk menghitung ulang performa; inkonsistensi ini harus diverifikasi terhadap file data asli sebelum publikasi.

3.2. Integritas Dataset dan Definisi Target

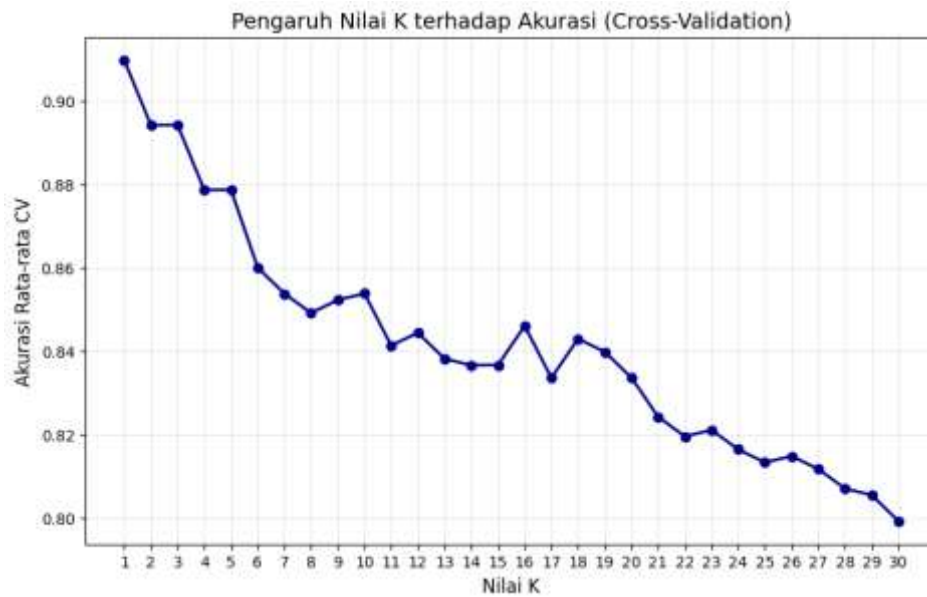
Pemeriksaan konsistensi memperlihatkan beberapa angka yang tidak dapat direkonsiliasi hanya dari manuskrip. Uraian awal hasil menyebut 540 lulusan, dan distribusi program studi 348+139+53 juga tepat berjumlah 540. Sebaliknya, tabel target mencantumkan 402 tepat waktu dan 703 tidak tepat waktu sehingga totalnya 1.105. Screenshot pembersihan data kembali menunjukkan 540 baris dan distribusi target 402 berbanding 138, sedangkan screenshot setelah SMOTE menunjukkan kedua kelas menjadi 402:402. Rangkaian screenshot tersebut lebih konsisten dengan 540 observasi awal yang terdiri dari 402 kelas mayoritas dan 138 kelas minoritas, kemudian kelas minoritas di-oversample menjadi 402. Oleh karena itu, tabel 402:703 pada naskah awal sangat mungkin merupakan artefak dari versi data atau analisis yang berbeda, tetapi tanpa dataset asli hal tersebut tidak dapat dipastikan.

Definisi target juga berimplikasi pada pemilihan fitur. Target tepat waktu dibentuk dari masa studi yang merupakan fungsi tahun masuk dan tahun lulus. Pada kode pemodelan, tahun masuk dan tahun lulus dimasukkan bersama IPK dan jumlah SKS lulus sebagai fitur. Untuk klasifikasi setelah mahasiswa lulus, fitur tersebut memang tersedia, tetapi nilai prediktifnya sebagian berasal dari informasi yang secara konseptual sangat dekat dengan definisi label. Untuk tujuan intervensi dini, fitur yang lebih tepat adalah informasi yang tersedia pada titik prediksi, misalnya IPS per semester, SKS kumulatif sampai semester tertentu, jumlah mata kuliah tidak lulus atau diulang, presensi, dan indikator administratif yang telah terjadi sebelum outcome kelulusan diketahui.

Konsekuensinya, interpretasi penelitian ini dibatasi pada kemampuan algoritma membedakan catatan lulusan berdasarkan fitur yang didokumentasikan. Klaim bahwa model dapat langsung mengidentifikasi mahasiswa aktif yang akan terlambat belum didukung oleh rancangan fitur saat ini. Pemisahan antara klasifikasi retrospektif dan prediksi prospektif penting agar sistem pendukung keputusan tidak menggunakan variabel yang baru tersedia setelah outcome terjadi.

3.3. Optimasi Nilai K

GridSearchCV pada rentang $K=1-30$ menghasilkan $K=1$ sebagai parameter terbaik dengan rerata accuracy validasi silang 0,909750 atau 90,97%. Grafik pada eksperimen menunjukkan kecenderungan akurasi menurun ketika K membesar. Hal ini konsisten dengan karakter KNN: K kecil mempertahankan struktur lokal data, sedangkan K besar menghaluskan batas keputusan. Akan tetapi, $K=1$ memiliki variansi tinggi dan sensitif terhadap noise sehingga pemilihannya sebaiknya dikonfirmasi pada pipeline validasi yang mencegah kebocoran data.



Gambar 1. Pengaruh nilai K terhadap akurasi validasi silang

3.4. Kinerja Model pada Data Uji

Pada 161 data uji, confusion matrix yang konsisten dengan classification report adalah 77 observasi kelas tidak tepat waktu diprediksi benar, 4 observasi kelas tidak tepat waktu diprediksi sebagai tepat waktu, 71 observasi kelas tepat waktu diprediksi benar, dan 9 observasi kelas tepat waktu diprediksi sebagai tidak tepat waktu. Dengan demikian, terdapat 148 prediksi benar dan 13 kesalahan klasifikasi.

Tabel 2. Confusion Matrix Model KNN

Kelas Aktual	Prediksi Tidak Tepat	Prediksi Tepat
Tidak Tepat Waktu	77	4
Tepat Waktu	9	71

Tabel 3. Metrik Evaluasi pada Data Uji

Kelas/Metrik	Precision	Recall	F1-score	Support
Tidak Tepat Waktu	0,8953	0,9506	0,9222	81
Tepat Waktu	0,9467	0,8875	0,9161	80
Macro average	0,9210	0,9191	0,9191	161
Accuracy	-	-	0,9193	161

Akurasi pengujian sebesar 91,93% menunjukkan proporsi prediksi benar yang tinggi. Kelas tidak tepat waktu memiliki recall 95,06%, sehingga sebagian besar mahasiswa pada kelas tersebut berhasil dikenali, tetapi precision-nya 89,53%. Sebaliknya, kelas tepat waktu memiliki precision 94,67% dan recall 88,75%. Macro-F1 sebesar 91,91% menunjukkan performa relatif seimbang pada kedua kelas setelah resampling. Nilai ini juga mengoreksi narasi awal yang menyebut macro-F1 90,49%; angka yang konsisten dengan classification report adalah sekitar 91,91%.

CLASSIFICATION REPORT				
	precision	recall	f1-score	support
Tidak Tepat Waktu	0.8953	0.9506	0.9222	81
Tepat Waktu	0.9467	0.8875	0.9161	80
accuracy			0.9193	161
macro avg	0.9210	0.9191	0.9191	161
weighted avg	0.9208	0.9193	0.9192	161

Gambar 2. Classification report model KNN

3.5. Pembahasan Kritis

Hasil 91,93% berada pada kisaran performa yang kompetitif dibandingkan beberapa penelitian KNN pada prediksi kelulusan, meskipun perbandingan langsung harus hati-hati karena dataset, definisi target, fitur, dan skema validasi berbeda. Amandha et al., melaporkan akurasi 93,33% dengan K=5 [15], sementara studi Masse et al., Riadi et al., dan Sofiyanto et al., kembali menegaskan bahwa pemilihan fitur dan parameter K menentukan performa [3], [4], [5]. Literatur juga menunjukkan perlunya memperhatikan distribusi kelas dan teknik resampling ketika target tidak seimbang [10]. Literatur yang lebih luas menunjukkan bahwa prediksi kelulusan berbasis machine learning memanfaatkan beragam indikator akademik dan memerlukan validasi yang konsisten agar hasil antarpemeriksaan dapat dibandingkan secara wajar [16]. Pendekatan selain KNN, seperti neuro-fuzzy dan deep learning, juga telah digunakan untuk memodelkan performa atau ketepatan waktu kelulusan mahasiswa [17], [18]

Terdapat dua isu validitas yang penting. Pertama, kode pada naskah sumber melakukan `StandardScaler.fit_transform` dan `SMOTE.fit_resample` sebelum `train_test_split`. Prosedur tersebut membuat parameter scaling memanfaatkan informasi seluruh dataset dan memungkinkan sampel sintetis dibentuk sebelum pemisahan data uji. Kondisi ini berpotensi menghasilkan estimasi performa yang optimistis. Desain yang lebih kuat adalah memisahkan hold-out test set terlebih dahulu dan menjalankan scaler serta SMOTE hanya di dalam data latih/fold cross-validation. Kedua, fitur tahun lulus bersama tahun masuk secara langsung mengandung informasi tentang lama studi yang digunakan membentuk target. Jika tujuan sistem adalah peringatan dini sebelum mahasiswa lulus, tahun lulus tidak tersedia pada saat prediksi dan harus dikeluarkan. Karena itu, hasil saat ini lebih tepat diposisikan sebagai pembuktian klasifikasi retrospektif daripada model early-warning yang siap implementasi.

Selain itu, naskah awal menyatakan K=3 dan akurasi 98,15% pada bagian kesimpulan serta membandingkan KNN dengan Decision Tree, Random Forest, Naive Bayes, dan Logistic Regression. Tidak terdapat eksperimen pendukung yang konsisten untuk klaim tersebut pada bagian hasil yang tersedia. Revisi ini tidak mempertahankan klaim tersebut agar kesimpulan hanya berasal dari bukti eksperimen yang terdokumentasi, yaitu K=1 dan akurasi pengujian 91,93%.

3.6. Implikasi Akademik dan Batas Generalisasi

Dalam konteks pengelolaan akademik, keluaran model klasifikasi seharusnya diperlakukan sebagai indikator risiko, bukan keputusan final mengenai mahasiswa. Recall yang tinggi pada kelas tidak tepat waktu berguna ketika institusi memprioritaskan sensitivitas untuk menjangkau mahasiswa yang memerlukan perhatian. Namun, false positive tetap menimbulkan biaya intervensi karena mahasiswa yang sebenarnya tidak bermasalah dapat ikut ditandai. Sebaliknya, false negative lebih kritis apabila mahasiswa berisiko justru tidak terdeteksi. Pemilihan threshold atau strategi klasifikasi pada pengembangan berikutnya karena itu perlu dikaitkan dengan biaya relatif dari kedua jenis kesalahan tersebut.

Sistem pendukung keputusan yang baik juga perlu menyediakan penjelasan yang dapat ditindaklanjuti oleh dosen pembimbing akademik. KNN pada dasarnya dapat menjelaskan prediksi melalui kasus-kasus tetangga terdekat, tetapi penggunaan fitur tahun lulus tidak menghasilkan rekomendasi intervensi yang bermakna. Model prospektif sebaiknya menggunakan indikator yang dapat diubah atau ditindaklanjuti, seperti kemajuan SKS, tren IPS, pengulangan mata kuliah, atau kehadiran. Dengan demikian, prediksi dapat dihubungkan dengan bentuk pendampingan yang spesifik, misalnya perencanaan beban studi, remediasi mata kuliah, atau konseling akademik.

Generalisasi hasil juga terbatas pada karakteristik dataset institusi dan periode yang digunakan. Perubahan kurikulum, kebijakan kelulusan, sistem pembelajaran, maupun komposisi mahasiswa dapat menyebabkan concept drift sehingga pola kedekatan yang dipelajari KNN tidak selalu stabil pada kohort baru. Validasi temporal dengan melatih model pada angkatan lama dan mengujinya pada angkatan yang lebih baru akan memberikan estimasi yang lebih relevan untuk penggunaan nyata. Validasi lintas program studi juga diperlukan karena distribusi IPK, struktur kurikulum, dan pola pengambilan SKS dapat berbeda.

Aspek tata kelola data perlu diperhatikan ketika model dikembangkan menjadi sistem institusional. Identitas mahasiswa seperti NIM dan nama tidak diperlukan sebagai fitur prediksi dan sebaiknya dipisahkan dari dataset analitik. Evaluasi fairness antarkelompok juga perlu dilakukan apabila atribut demografis atau jalur penerimaan kelak digunakan. Tujuannya bukan sekadar memperoleh akurasi tinggi, tetapi memastikan bahwa intervensi akademik berbasis model dapat diaudit, proporsional, dan tidak menghasilkan perlakuan yang tidak dapat dipertanggungjawabkan.

4. Kesimpulan

Berdasarkan eksperimen yang terdokumentasi, KNN mampu mengklasifikasikan status kelulusan pada data akademik dengan $K=1$ sebagai parameter terbaik menurut GridSearchCV lima lipatan. Pada 161 data uji, model menghasilkan akurasi 91,93%, macro-precision 92,10%, macro-recall 91,91%, dan macro-F1 91,91%, dengan 148 prediksi benar dari 161 observasi. Hasil tersebut menunjukkan potensi KNN sebagai komponen analitik untuk klasifikasi kelulusan. Namun, hasil belum dapat diperlakukan sebagai estimasi final untuk implementasi sistem peringatan dini karena preprocessing yang terdokumentasi melakukan scaling dan SMOTE sebelum pembagian data, serta memasukkan tahun lulus sebagai fitur. Penelitian lanjutan perlu melakukan validasi ulang dengan hold-out test set yang tidak tersentuh, pipeline scaling-SMOTE hanya pada data latih, fitur yang benar-benar tersedia sebelum kelulusan, serta pengujian lintas angkatan atau lintas program studi. Dengan perbaikan tersebut, model dapat dievaluasi secara lebih valid sebagai dasar intervensi akademik bagi mahasiswa berisiko terlambat lulus.

Referensi

- [1] D. R. Sari, V. Julianto, and H. Rhomadana, "Prediction Of Student Graduation Using The K-Nearest Neighbor Method Case Study in Politeknik Negeri Tanah Laut," *J. Ilm. Inform.*, vol. 8, no. 1, pp. 74–88, 2023, doi: <https://doi.org/10.35316/v8i1.3105>.
- [2] I. P. Y. P. Atmaja, G. I. Setiawan, I. W. Dika, and I. A. P. F. Imawati, "Data Mining Memprediksi Kelulusan Mahasiswa Menggunakan Metode K-Nearest Neighbors (KNN) Studi Kasus Universitas PGRI Mahadewa Indonesia," *J. Manaj. dan Teknol. Inf.*, vol. 13, no. 2, pp. 10–23, 2023, doi: <https://doi.org/10.59819/jmti.v13i2.3082>.
- [3] B. A. Masse, R. Rasyid, I. Irsan, F. A. Masse, D. Grace, and D. Kusumawati, "Prediksi Ketepatan Waktu Kelulusan Mahasiswa Di Universitas Islam Negeri Datokarama Palu Menggunakan Algoritma K-Nearest Neighbor," *J. Sains Komput. Dan Teknol. Inf.*, vol. 7, no. 1, pp. 12–18, 2024, doi: <https://doi.org/10.33084/jsakti.v7i1.8433>.
- [4] I. Riadi, R. Umar, and R. Anggara, "Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode K-Nearest Neighbor," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 2, pp. 249–256, 2024, doi: <https://doi.org/10.25126/jtiik.20241127330>.
- [5] I. N. Sofiyanto, A. B. Saputra, R. F. Rakhmadianto, H. N. E. Febrianto, A. M. Endiwan, and A. A. Sari, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbor (KNN) dengan Klasifikasi Biner," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, 2025, pp. 40–45. doi: <https://doi.org/10.47701/gpqsm502>.
- [6] N. Ahmad, S. Hafizh, and R. Sulthanah, "Prediksi Kelulusan Mata Kuliah Mahasiswa Teknologi Informasi Menggunakan Algoritma K-Nearest Neighbor," *J. Manaj. Inform.*, vol. 14, no. 2, pp. 135–149, 2024, doi: <https://doi.org/10.34010/jamika.v14i2.12454>.
- [7] F. A. Putra and M. H. Pratama, "Algoritma K-Nearest Neighbor Untuk Peramalan Kelulusan Mahasiswa," *J. Sist. Informasi, Tek. Inform. dan Teknol. Pendidik.*, vol. 4, no. 1, pp. 15–22, 2024, doi: <https://doi.org/10.55338/justikpen.v4i1.128>.
- [8] O. S. D. Putri and A. Y. P. Putri, "Prediksi Kelulusan Mahasiswa Dengan Metode K-Nearest Neighbor Pada Jurusan Sistem Informasi Institut XYZ," in *Prosiding Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika*, 2023, pp. 199–206. doi: <https://doi.org/10.31284/p.snestik.2023.4388>.
- [9] D. P. Sari, S. S. Hilabi, and A. Hananto, "Penerapan Data Mining Metode K-Nearest Neighbor Untuk Memprediksi Kelulusan Siswa Sekolah Menengah Pertama," *SMARTICS*, vol. 9, no. 1, pp. 14–19, 2023, doi: <https://doi.org/10.21067/smartics.v9i1.8088>.
- [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," *Sci. J. Informatics*, vol. 11, no. 3, pp. 559–568, 2024, doi: <https://doi.org/10.15294/sji.v11i3.5528>.
- [11] M. F. A. Shiddiqi and H. D. Bhakti, "Perbandingan Algoritme K-Nearest Neighbor Dan Naïve Bayes Dalam Klasifikasi Kelulusan Mahasiswa," *J. Mhs. Tek. Inform.*, vol. 8, no. 5, pp. 10970–10979, 2024, doi: <https://doi.org/10.36040/jati.v8i5.10955>.
- [12] A. Putri et al., "Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 20–26, 2023, doi: <https://doi.org/10.57152/malcom.v3i1.610>.
- [13] N. Hidayati and A. Hermawan, "K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation," *J. Eng. Appl. Technol.*, vol. 2, no. 2, pp. 86–91, 2021, doi: <https://doi.org/10.21831/jeatech.v2i2.42777>.
- [14] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014. doi: <https://doi.org/10.1017/CBO9781107298019>.
- [15] S. Amandha, H. Rohayani, and K. Kurniawansyah, "Implementation of Data Mining for Predicting Student Graduation Using the K-Nearest Neighbor Algorithm at Jambi Muhammadiyah University," *Indones. J. Artif. Intell. Data Min.*, vol. 7, no. 1, pp. 134–140,

DOI: <https://doi.org/10.69693/ijmst.v4i3.13830>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

- 2024, doi: <https://doi.org/10.24014/ijaidm.v7i1.26150>.
- [16] L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 23451–23465, 2024, doi: [10.1109/ACCESS.2024.3361479](https://doi.org/10.1109/ACCESS.2024.3361479).
- [17] R. Mehdi and M. Nachouki, "A neuro-fuzzy model for predicting and analyzing student graduation performance in computing programs," *Educ. Inf. Technol.*, vol. 28, pp. 2455–2484, 2023, doi: <https://doi.org/10.1007/s10639-022-11205-2>.
- [18] N. V. Darenoh, F. A. Bachtiar, and R. S. Perdana, "Prediction of on-time student graduation with deep learning method," *J. ICT Res. Appl.*, vol. 18, no. 1, pp. 1–20, 2024, doi: <https://doi.org/10.5614/itbj.ict.res.appl.2023.18.1.1>.