https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Ketahanan Model Machine Learning terhadap Ketidaklengkapan Data Survei untuk Prediksi Kinerja Akademik

Intra Swadaya Hidayat[1,*], Yarza Aprizal[2], Rendy Almaheri Adhi Pratama[3], Muhammad Jhonsen Syafriandi[4]
[1] Digital Business Study Program, Institut Teknologi dan Bisnis Palcomtech, Palembang, Indonesia
[2,3,4] Informatics Study Program, Institut Teknologi dan Bisnis Palcomtech, Palembang, Indonesia

E-mail: [1,*]intra.swadaya@palcomtech.ac.id, [2]yarza_afrizal@palcomtech.ac.id, [3]rendy_almaheri@palcomtech.ac.id, [4]m_jhonsen@palcomtech.ac.id
Corresponding Author E-mail: intra.swadaya@palcomtech.ac.id

ABSTRAK

Model prediksi berbasis survei biasanya dievaluasi pada respons lengkap, padahal ketidaklengkapan dapat muncul secara acak maupun terstruktur. Penelitian ini mengevaluasi kapan informasi survei masih cukup untuk memprediksi persepsi perubahan kinerja akademik pada dataset publik berisi 361 respons, 28 prediktor, dan tiga kelas yang tidak seimbang. Tiga konfigurasi fitur—seluruh prediktor (F1), tanpa tiga item berisiko proxy (F2), dan hanya fitur kontekstual (F3)—dievaluasi menggunakan CatBoost dengan repeated stratified 5-fold cross-validation sebanyak lima pengulangan. Macro F1 baseline masing-masing adalah 0,414±0,047, 0,384±0,042, dan 0,351±0,046. Recall kelas menurun hanya 0,023 pada F1 dan 0,010 pada F2 serta F3, sehingga performa absolut belum memadai untuk identifikasi mahasiswa berisiko. Pada F2, kehilangan 50% item secara acak masih mempertahankan 94,8% Macro F1 baseline ketika hanya data uji tidak lengkap dan 97,7% ketika data latih serta uji sama-sama tidak lengkap. Sebaliknya, monotone dropout dengan hanya 50% atau 25% item awal teramati menurunkan retensi menjadi 79,3% dan 78,8% (Holm-adjusted p<0,001). Hasil menunjukkan bahwa kecukupan relatif bergantung pada pola kehilangan, tetapi retensi tinggi tidak menjamin kelayakan prediksi ketika baseline lemah dan kelas minoritas hampir tidak terdeteksi.

Kata kunci: ***Educational Data Mining; ketidaklengkapan data; ketahanan model; target proxy; selective prediction***

ABSTRACT

Survey-based prediction models are commonly evaluated on complete responses, although missingness can occur randomly or structurally. This study examines when survey information remains sufficient to predict self-reported changes in academic performance using a public dataset of 361 responses, 28 predictors, and three imbalanced classes. Three feature configurations—all predictors (F1), predictors excluding three high-risk proxy items (F2), and contextual predictors only (F3)—were evaluated with CatBoost using five-times repeated stratified five-fold cross-validation. Baseline Macro F1 was 0.414±0.047, 0.384±0.042, and 0.351±0.046 for F1, F2, and F3, respectively. Recall for the decreased class was only 0.023 for F1 and 0.010 for both F2 and F3, indicating inadequate absolute performance for identifying at-risk students. For F2, 50% random item loss retained 94.8% of baseline Macro F1 under test-time missingness and 97.7% when both training and test data were incomplete. In contrast, monotone dropout with only the first 50% or 25% of items observed reduced retention to 79.3% and 78.8% (Holm-adjusted p<0.001). Thus, relative information sufficiency depends on the missingness pattern, but high retention does not imply deployability when baseline performance is weak and the minority class is nearly undetected.

Keywords: ***Educational Data Mining; missing data; model robustness; target proxy; selective prediction***

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

1. PENDAHULUAN

Educational Data Mining (EDM) menggunakan teknik statistik, pembelajaran mesin, dan penambangan data untuk menemukan pola yang dapat mendukung pemahaman proses belajar serta pengambilan keputusan pendidikan. Prediksi kinerja mahasiswa merupakan salah satu tema dominan karena dapat membantu institusi mengenali kebutuhan dukungan secara lebih dini. Tinjauan sistematis menunjukkan bahwa penelitian bidang ini banyak berfokus pada pemilihan algoritma, akurasi klasifikasi, dan identifikasi prediktor dominan (Xiao et al., 2022; Yağcı, 2022). Walaupun bermanfaat, fokus tersebut belum otomatis menjawab apakah model tetap layak digunakan ketika informasi masukan tidak lengkap.

Data survei memiliki karakter berbeda dari log sistem pembelajaran maupun nilai resmi. Jawaban bergantung pada persepsi responden, desain instrumen, urutan pertanyaan, serta kemauan untuk menyelesaikan seluruh bagian. Ketidaklengkapan dapat muncul sebagai satu atau beberapa item terlewat, satu halaman tidak tersimpan, pola skip, atau penghentian pengisian. Mitra et al. (2023) menegaskan bahwa missingness yang terstruktur perlu diperlakukan sebagai bagian dari masalah pemodelan, bukan sekadar gangguan yang selalu dapat diselesaikan melalui imputasi. Dampak kehilangan informasi juga bergantung pada lokasi dan makna item; kehilangan 20% sel secara acak tidak ekuivalen dengan hilangnya satu kelompok konstruk yang sangat prediktif.

Dataset Students' Academic Performance in Video-Conference-Assisted Online Learning During the COVID-19 Pandemic dikumpulkan melalui Google Form dari mahasiswa yang mengikuti pembelajaran daring berbantuan konferensi video pada beberapa institusi di Jakarta. Dokumentasi resmi menyatakan bahwa target memiliki label decreased, stable, dan good yang dikodekan sebagai 0, 1, dan 2 serta dataset dilisensikan CC BY 4.0 (Miranda et al., 2024a). Artikel asal melaporkan 361 respons, 28 atribut, dan lima perspektif prediktor: video conference application (VC), learning material (LM), internet connection (IC), students' ability to learn (SL), dan student knowledge (SK) (Miranda et al., 2024b).

Penelitian asal mengembangkan Random Forest, Support Vector Machine, dan Gaussian Naive Bayes, menangani ketidakseimbangan menggunakan SMOTE dan augmentasi, serta menggunakan SHAP untuk interpretasi. Random Forest dengan penyeimbangan data dilaporkan paling unggul, sedangkan Performance, Frequency Constraint, dan Increase Value termasuk fitur dominan (Miranda et al., 2024b). Temuan tersebut penting, tetapi menimbulkan dua persoalan metodologis lanjutan. Pertama, survei asal mewajibkan seluruh pertanyaan sehingga tidak menyediakan missing value alami untuk menguji ketahanan deployment. Kedua, fitur seperti Performance dan Increase Value secara semantik berpotensi sangat dekat dengan target, sehingga performa tinggi dapat sebagian merefleksikan ketergantungan pada target proxy.

Istilah target perlu dibatasi secara hati-hati. Karena data bersumber dari survei mahasiswa dan dokumentasi yang tersedia belum menunjukkan pencocokan dengan transkrip atau nilai resmi, artikel ini menggunakan istilah persepsi perubahan kinerja akademik atau kinerja akademik yang dilaporkan mahasiswa. Istilah prestasi akademik aktual tidak digunakan. Audit instrumen tetap diperlukan untuk memastikan bunyi pertanyaan target, skala respons, serta relasinya dengan item Performance, IncreaseValue, Ability, Knowledge, Competence, dan CompletingProject.

Research gap penelitian ini adalah belum adanya evaluasi sistematis pada dataset tersebut mengenai degradasi prediksi ketika informasi berkurang, perbedaan dampak antar pola missingness, sensitivitas setiap perspektif, serta batas kelengkapan operasional yang masih mempertahankan performa. Klaim ini dibatasi pada hasil penelusuran awal dan tidak dinyatakan sebagai penelitian pertama. Fokus kajian bukan menentukan algoritma terbaik, melainkan menjawab seberapa banyak dan jenis informasi apa yang masih diperlukan agar prediksi tetap andal.

Penelitian bertujuan: (1) mengukur ketahanan prediksi terhadap penurunan kelengkapan survei; (2) membandingkan item-level random missingness, section-level missingness, monotone dropout, dan target-proximal nonresponse; (3) mengidentifikasi perspektif paling kritis; (4) mengaudit ketergantungan pada item dekat target; (5) menguji model contextual-only; dan (6)

menentukan kondisi penolakan prediksi. Kontribusi yang dituju ialah kerangka evaluasi kecukupan informasi, performance retention, ambang kelengkapan berbasis data, serta rekomendasi penggunaan model yang mempertimbangkan coverage dan reliability.

Tabel 1. Posisi penelitian dan kontribusi yang diuji

Aspek	Studi asal	Penelitian ini
Tujuan	Prediksi dan interpretasi	Kecukupan informasi
Missingness	Data lengkap	Empat pola simulasi
Fitur	Seluruh prediktor	F1, F2, dan F3
Keputusan	Selalu prediksi	Prediksi/abstain

2. LANDASAN TEORI

2.1 Educational Data Mining dan prediksi kinerja

Model prediksi dalam EDM dapat menggunakan data demografis, aktivitas pada learning management system, asesmen formatif, maupun survei pengalaman belajar. Validitas kesimpulan bergantung pada kesesuaian konstruk target dan prediktor, bukan hanya nilai metrik. Prediksi nilai resmi dari riwayat asesmen berbeda secara substantif dari prediksi evaluasi diri mahasiswa melalui item survei. Karena itu, artikel ini memosisikan output sebagai klasifikasi persepsi perubahan kinerja akademik dan membatasi implikasinya untuk analitik kualitas data, bukan keputusan kelulusan.

Kelas target yang tidak seimbang menuntut metrik yang tidak didominasi kelas mayoritas. Macro F1 menghitung F1 setiap kelas dengan bobot sama, sedangkan balanced accuracy merata-ratakan recall antarkelas. Matthews correlation coefficient memberi ringkasan korelasi prediksi dan label yang lebih informatif daripada accuracy pada distribusi tidak seimbang (Chicco & Jurman, 2020). Recall per kelas tetap disajikan agar kegagalan mengenali kelas decreased tidak tertutupi oleh performa kelas good.

2.2 Missingness, kelengkapan, dan ketahanan

Teori klasik membedakan missing completely at random, missing at random, dan missing not at random berdasarkan hubungan mekanisme ketidaklengkapan dengan data teramati dan tidak teramati (Rubin, 1976). Penelitian ini tidak mengklaim bahwa simulasi merekonstruksi mekanisme nyata. Simulasi dirancang sebagai stress test prediktif yang terkontrol: random missingness mewakili item terlewat tanpa struktur; section loss mewakili blok pertanyaan yang tidak tersedia; monotone dropout mengikuti urutan instrumen; dan target-proximal nonresponse mewakili ketidaksediaan jawaban evaluatif.

Kelengkapan respons didefinisikan sebagai proporsi fitur yang teramati pada satu record. Ketahanan diukur terhadap baseline data lengkap dengan performance retention. Pendekatan relatif ini membantu membandingkan konfigurasi fitur yang memiliki tingkat baseline berbeda. Nilai retention lebih dari satu tidak langsung dianggap peningkatan substantif; nilai tersebut perlu diperiksa terhadap variasi fold, seed simulasi, dan interval kepercayaan.

Retention(m) = Macro F1(m) / Macro F1(lengkap)

2.3 Target proxy dan validitas konstruk

Target proxy adalah prediktor yang menyandi secara langsung atau hampir langsung informasi yang hendak diprediksi. Proxy tidak selalu merupakan kebocoran teknis; sebuah item dapat tersedia secara sah saat inferensi, tetapi membuat klaim prediktif menjadi trivial atau melingkar. Oleh sebab itu, audit dilakukan pada bunyi pertanyaan, bukan nama kolom saja. Setiap item diklasifikasikan sebagai direct target proxy, proximal predictor, contextual predictor, atau behavioural predictor.

Tiga konfigurasi fitur digunakan untuk memisahkan performa maksimal dan validitas konstruk. F1 mereplikasi penggunaan seluruh atribut. F2 menghapus direct proxy dan prediktor yang sangat proksimal berdasarkan audit semantik definisional. F3 hanya mempertahankan kondisi pengalaman belajar yang secara logis tersedia sebelum mahasiswa menilai perubahan kinerjanya. Selisih performa F1–F2 dan F2–F3 ditafsirkan sebagai ketergantungan informasi, bukan efek kausal.

2.4 Selective prediction

Selective prediction memasangkan klasifikator dengan fungsi seleksi yang menerima atau menolak sebuah prediksi. Tujuannya bukan memaksimalkan jumlah

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

prediksi, melainkan memperoleh trade-off yang transparan antara coverage dan risiko pada kasus yang diterima (Pugnana & Ruggieri, 2023). Dalam penelitian ini, seleksi didasarkan pada dua sinyal yang mudah diaudit, yaitu kelengkapan respons dan probabilitas kelas maksimum.

Confidence model tidak otomatis identik dengan probabilitas kebenaran. Karena itu, analisis reliability diagram, log loss, dan kesalahan ber-confidence tinggi diperlukan. Aturan abstention yang baik harus meningkatkan kualitas prediksi yang diterima tanpa menolak secara tidak proporsional kelas minoritas.

3. METODOLOGI

3.1 Desain penelitian dan sumber data

Penelitian menggunakan desain eksperimen komputasional dengan data sekunder publik. Dataset versi 1 diterbitkan pada Mendeley Data dengan DOI 10.17632/x8mc6y45x2.1 dan lisensi CC BY 4.0 (Miranda et al., 2024a). Audit file memverifikasi 361 baris, 28 prediktor, satu target, tanpa missing value alami, serta satu baris duplikat tambahan. Seluruh prediktor pada file analitik berada pada rentang 0–2, meskipun dokumentasi transformasi menjelaskan beberapa skala asal yang lebih lebar. Karena itu, eksperimen memperlakukan file sebagai data yang telah ditransformasikan dan tidak merekonstruksi skala respons mentah.

Unit analisis adalah respons mahasiswa dan target tidak pernah dibuat missing. Lima perspektif prediktor—VC, LM, IC, SL, dan SK—dipetakan menggunakan tabel pernyataan kuesioner pada dokumentasi resmi. Urutan pada tabel tersebut digunakan sebagai dasar skenario monotone dropout. Keputusan ini didokumentasikan karena urutan kolom pada file analitik tidak sepenuhnya sama dengan urutan pernyataan pada data dictionary.

Tabel 2. Fakta dataset yang telah diverifikasi dari sumber primer

Elemen	Informasi
Observasi	361 respons
Prediktor	28 atribut
Perspektif	VC, LM, IC, SL, SK
Target	0 menurun; 1 stabil; 2 baik
Kelas	17; 82; 262
Lisensi	CC BY 4.0

3.2 Audit kualitas dan konstruk

Audit kualitas menemukan nol missing value alami, tidak ada fitur konstan, satu baris duplikat identik, dan rentang 0–2 pada seluruh prediktor. Duplikat dipertahankan karena audit tidak menghapus observasi secara otomatis dan jumlahnya hanya satu baris tambahan. Ketidaksesuaian antara rentang file dan deskripsi transformasi dicatat sebagai keterbatasan dokumentasi. Hasil audit dan pemetaan fitur disimpan dalam dataset_audit.csv serta data_dictionary.csv.

Audit semantik terhadap bunyi item menandai Increase_Value, Performance, dan Positive_effect sebagai item berisiko proxy tinggi karena secara langsung menanyakan nilai atau performa. Ketiganya dikeluarkan pada F2. F3 mempertahankan 19 fitur kontekstual dari perspektif VC, IC, serta tiga item LM yang tidak secara langsung menanyakan hasil akademik. Klasifikasi fitur ditetapkan menggunakan kedekatan definisional antara bunyi prediktor dan konstruk target.

Tabel 3. Konfigurasi fitur penelitian

Konf.	Cakupan	Tujuan
F1	Semua prediktor	Baseline maksimal
F2	Tanpa proxy/proksimal	Audit ketergantungan
F3	Kontekstual saja	Konstruk defensif

3.3 Model dan validasi

CatBoostClassifier dipilih sebagai model utama karena sesuai untuk data tabular kecil, mendukung multikelas, fitur kategorikal, nilai hilang, bobot kelas, serta probabilitas prediksi. Ordered boosting pada CatBoost dirancang untuk mengurangi bias target statistics pada fitur kategorikal (Prokhorenkova et al., 2018). DummyClassifier digunakan sebagai sanity check. Jika struktur data tidak kompatibel, analisis alternatif menggunakan pipeline SimpleImputer, OneHotEncoder, dan

RandomForestClassifier; pergantian ini harus diputuskan sebelum melihat hasil skenario.

Evaluasi menggunakan repeated stratified 5-fold cross-validation dengan lima pengulangan sehingga menghasilkan 25 fold yang identik pada seluruh perbandingan. CatBoost dikonfigurasi dengan 250 iterasi, depth 5, learning rate 0,05, regularisasi L2 sebesar 5, dan auto_class_weights=Balanced. Konfigurasi ditetapkan tanpa melihat hasil skenario. Transformasi missingness dilakukan terpisah pada fold pelatihan dan pengujian untuk mencegah kebocoran informasi.

Ketidakseimbangan ditangani dengan auto_class_weights=Balanced. SMOTE tidak digunakan pada eksperimen utama agar pengaruh missingness tidak bercampur dengan sampel sintetis. Keputusan ini juga membedakan tujuan studi dari penelitian asal yang berfokus pada penyeimbangan dan perbandingan algoritma.

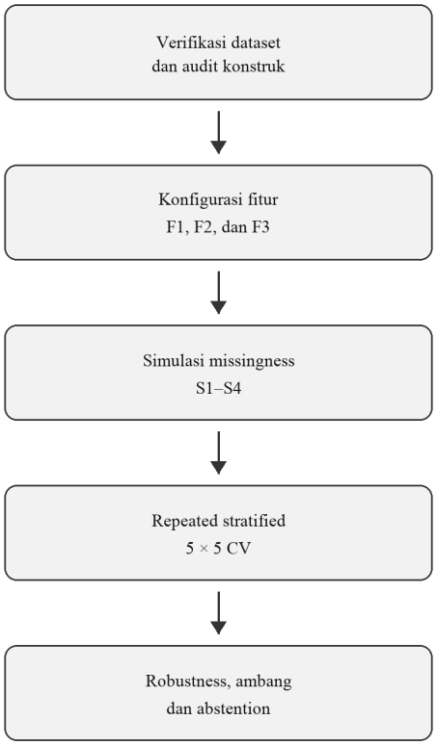
3.4 Simulasi ketidaklengkapan data

Tabel 4. Skenario missingness yang diuji

Sken.	Pola	Tingkat/unit
S1	Acak per item	10–50%; 3 seed
S2	Hilang satu bagian	Satu perspektif
S3	Monotone dropout	25, 50, 75, 100% urutan
S4	Nonrespons proksimal	Item hasil audit

S1 menerapkan masker Bernoulli pada sel prediktor dengan tingkat kehilangan 10%, 20%, 30%, 40%, dan 50%. Setiap tingkat diulang menggunakan tiga seed simulasi. Kondisi pertama hanya memodifikasi validation fold untuk merepresentasikan test-time missingness; kondisi kedua memodifikasi training dan validation fold secara terpisah. Target tidak pernah dimasking dan kelengkapan aktual setiap fold dicatat.

S2 menghilangkan seluruh fitur dari satu perspektif secara bergantian. S3 memotong respons berdasarkan urutan pertanyaan asli dan karena itu hanya valid jika urutan instrumen terverifikasi. S4 menghilangkan direct target proxy dan prediktor sangat proksimal. Perbandingan antar pola dilakukan pada tingkat informasi yang sedapat mungkin ekuivalen, namun kehilangan satu perspektif juga dilaporkan sebagai eksperimen kelompok tanpa memaksakan persentase yang sama.



Gambar 1. Alur penelitian kecukupan informasi

3.5 Metrik, ambang, dan analisis statistik

Metrik primer adalah Macro F1. Metrik sekunder terdiri atas balanced accuracy, MCC, precision, recall dan F1 setiap kelas, confusion matrix, log loss, confidence rata-rata, waktu training, serta waktu inferensi. Accuracy hanya disajikan sebagai informasi tambahan. Performance retention dihitung pada setiap fold terhadap baseline fold yang sama.

Batas minimum kelengkapan didefinisikan sebagai kelengkapan terendah yang masih mempertahankan sekurangnya 90% Macro F1 baseline. Analisis sensitivitas menggunakan batas 95% dan 97%. Karena tingkat yang diuji bersifat diskrit, estimasi ambang dilaporkan sebagai interval antartingkat, bukan nilai kontinu yang terlalu presisi. Ambang merupakan batas operasional pada dataset dan pipeline ini, bukan standar universal.

Aturan abstention menggabungkan kelengkapan c dan confidence pmax. Grid ambang kelengkapan dan confidence dievaluasi pada prediksi out-of-fold. Setiap kombinasi menghasilkan coverage, rejection

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

rate, Macro F1 kasus diterima, error rate, dan recall per kelas. Pemilihan aturan dilakukan berdasarkan trade-off yang ditetapkan sebelum pembacaan hasil akhir dan diperiksa terhadap risiko menolak kelas minoritas.

Distribusi hasil pada fold yang sama dianalisis secara berpasangan. Interval kepercayaan 95% dilaporkan bersama rata-rata dan simpangan baku. Wilcoxon signed-rank digunakan untuk perbandingan skenario terhadap baseline, sedangkan koreksi Holm mengendalikan family-wise error pada pengujian berganda (Demšar, 2006; Holm, 1979).

3.6 Interpretabilitas, reproduksibilitas, dan etika

Interpretasi utama menggunakan audit semantik, perbandingan konfigurasi F1–F3, serta ablation pada lima perspektif. Pendekatan ini dipilih karena langsung mengukur ketergantungan prediktif ketika suatu kelompok informasi tidak tersedia. SHAP tidak digunakan sebagai bukti utama dan hubungan prediktif tidak ditafsirkan sebagai hubungan kausal.

Seluruh konfigurasi, seed, fold, dan keluaran per-fold disimpan agar analisis dapat direplikasi. Berkas utama meliputi dataset_audit.csv, data_dictionary.csv, all_fold_results.csv, all_predictions.csv, aggregate_results.csv, baseline_confusion_matrices.csv, statistical_tests.csv, completeness_thresholds.csv, abstention_results.csv, dan experiment_summary.json.

Model tidak digunakan untuk menentukan kelulusan, memberi sanksi, atau menggantikan penilaian dosen. Output hanya digunakan untuk menilai kecukupan informasi dan ketahanan analitik. Dataset yang dipublikasikan tetap diperiksa terhadap kemungkinan atribut sensitif dan risiko identifikasi ulang.

4. HASIL DAN PEMBAHASAN

4.1 Hasil audit dataset dan target

File analitik memiliki 361 observasi dan 29 kolom yang terdiri atas 28 prediktor serta satu target. Tidak ditemukan missing value alami atau fitur konstan. Satu observasi merupakan duplikat identik, tanpa konflik label. Semua prediktor bernilai 0–2, sedangkan dokumentasi menyebut transformasi dari beberapa skala asal yang lebih lebar. Distribusi target sangat tidak seimbang: 17 menurun, 82 stabil, dan 262 baik.

Dokumentasi mendefinisikan target sebagai perubahan kinerja akademik yang dilaporkan mahasiswa, bukan nilai dari transkrip. Bunyi target tidak disajikan selengkap bunyi 28 prediktor, sehingga validitas konstruk belum dapat dipastikan sepenuhnya. Audit semantik menandai Increase_Value, Performance, dan Positive_effect sebagai proxy berisiko tinggi; keputusan ini digunakan untuk analisis sensitivitas, bukan sebagai bukti bahwa terjadi kebocoran teknis.

4.2 Performa baseline dan konfigurasi fitur

Dummy prior memperoleh accuracy 0,726 tetapi Macro F1 hanya 0,280 karena selalu memihak kelas baik. F1 menghasilkan Macro F1 0,414±0,047, balanced accuracy 0,424±0,045, dan MCC 0,234±0,095. Setelah tiga item berisiko proxy dihapus, F2 mencapai Macro F1 0,384±0,042 dan MCC 0,168±0,097. F3 memperoleh Macro F1 0,351±0,046 dan MCC 0,094±0,111. Pada seluruh konfigurasi, recall kelas menurun berada pada 0,010–0,023.

Tabel 5. Rata-rata performa baseline pada 25 fold

Model	Macro F1	MCC	Recall menurun
Dummy	0,280	0,000	0,000
F1	0,414	0,234	0,023
F2	0,384	0,168	0,010
F3	0,351	0,094	0,010

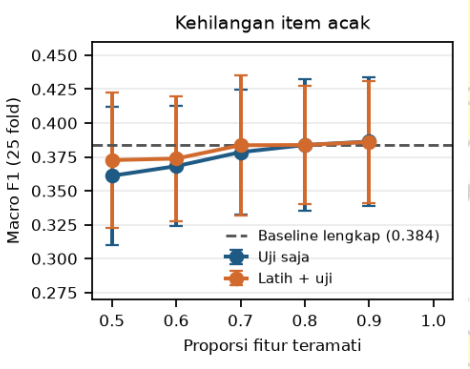
Selisih Macro F1 F1–F2 sebesar 0,030 menunjukkan ketergantungan moderat pada tiga item berisiko proxy, sedangkan pengurangan F2–F3 sebesar 0,033 menunjukkan kontribusi tambahan fitur

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

nonkontekstual. Namun, ketiga konfigurasi memiliki performa absolut rendah. Accuracy tidak digunakan untuk menyatakan keberhasilan karena Dummy justru mencapai accuracy tertinggi dengan mengabaikan dua kelas minoritas.

4.3 Degradasi performa dan batas kelengkapan

Pada F2, Macro F1 test-time missingness berubah dari 0,384 pada data lengkap menjadi 0,386, 0,384, 0,379, 0,368, dan 0,361 ketika 10%–50% item hilang secara acak. Retensi pada kehilangan 50% adalah 0,948. Ketika missingness diterapkan pada data latih dan uji, Macro F1 pada kehilangan 50% adalah 0,373 dengan retensi 0,977. Setelah koreksi Holm, perbedaan random missingness terhadap baseline tidak signifikan pada tingkat yang diuji.

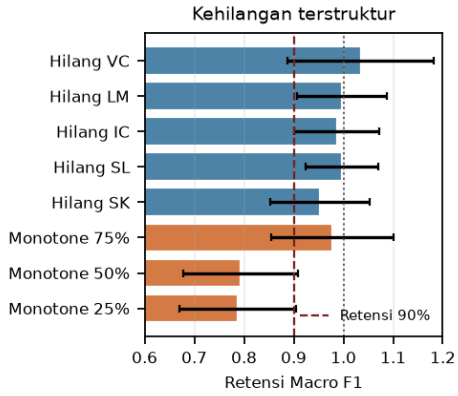


Gambar 2. Macro F1 terhadap kelengkapan acak pada F2

Berdasarkan kriteria relatif, kelengkapan 50% masih memenuhi retensi 90% pada kedua kondisi random missingness; retensi 95% membutuhkan kelengkapan 60% pada test-time missingness. Akan tetapi, ambang ini tidak dapat dipakai sebagai izin deployment karena baseline F2 hanya 0,384 dan hampir tidak pernah mengenali kelas menurun. Dengan demikian, data dapat disebut cukup untuk mempertahankan performa agregat yang sudah ada, tetapi belum cukup untuk prediksi individual yang andal.

4.4 Sensitivitas perspektif dan item proksimal

Kehilangan seluruh perspektif SK memberi penurunan terbesar pada F2, dengan retensi 0,953, diikuti IC 0,987, SL 0,997, dan LM 0,997; penghilangan VC tidak menurunkan rata-rata Macro F1 (retensi 1,035), tetapi variasi antarfold lebar. Setelah koreksi Holm, tidak ada section loss yang berbeda signifikan dari baseline. Sebaliknya, monotone dropout dengan hanya 50% dan 25% item awal teramati menurunkan retensi menjadi 0,793 dan 0,788 ($p_{Holm}<0,001$).



Gambar 3. Retensi pada kehilangan terstruktur

SK merupakan kelompok paling kritis secara prediktif dalam F2, tetapi efeknya kecil dan tidak signifikan setelah koreksi multipel. Pola monotone yang lebih merusak menunjukkan bahwa kombinasi serta urutan informasi lebih penting daripada jumlah kolom dari satu perspektif saja. Hasil ini tidak ditafsirkan secara kausal, terutama karena dua fitur SK yang tersisa masih berdekatan secara konseptual dengan target.

4.5 Selective prediction dan implikasi operasional

Grid selective prediction tidak menghasilkan aturan yang layak. Pada ambang kelengkapan 0,60 dan confidence 0,50, coverage turun menjadi 43,1% dan Macro F1 naik menjadi 0,416, tetapi recall kelas menurun tetap 0. Aturan yang mempertahankan coverage di atas 50% hanya mencapai Macro F1 sekitar 0,386 dan recall kelas menurun sekitar 0,016.

Tidak ada ambang abstention yang direkomendasikan untuk penggunaan individual. Peningkatan Macro F1 diperoleh terutama dengan menolak banyak kasus tanpa memperbaiki deteksi kelas menurun. Dalam kondisi ini, abstention harus dipandang

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

sebagai bukti keterbatasan model, bukan mekanisme yang membuat model aman untuk keputusan akademik.

4.6 Pembahasan umum dan keterbatasan

Hasil memperlihatkan perbedaan antara robustness relatif dan usefulness absolut. Missingness acak hingga 50% hanya sedikit mengubah Macro F1, sedangkan monotone dropout pada 50% atau 25% urutan awal menyebabkan penurunan yang jelas. Stabilitas terhadap kehilangan acak kemungkinan dipengaruhi redundansi informasi dan rendahnya baseline, sehingga tidak boleh ditafsirkan sebagai bukti bahwa separuh pertanyaan selalu dapat dihapus.

Dibandingkan studi asal, penelitian ini tidak berupaya mengungguli algoritma tertentu, melainkan menguji validitas informasi di bawah kehilangan terkontrol. Hasil yang lebih rendah juga tidak dapat dibandingkan langsung karena tujuan, konfigurasi fitur, penyeimbangan kelas, dan protokol validasi berbeda. Kontribusi utama terletak pada temuan bahwa retensi tinggi dapat menyesatkan ketika kelas minoritas hampir tidak pernah diprediksi.

Keterbatasan utama mencakup target self-report, hanya 17 observasi pada kelas menurun, missingness sintetis, satu dataset kecil, tidak adanya external test set, serta dokumentasi skala yang tidak sepenuhnya konsisten dengan file. Penetapan tiga proxy didasarkan pada audit semantik dan belum didukung validasi konstruk eksternal. Simulasi monotone bergantung pada urutan tabel kuesioner dan tidak membuktikan pola nonrespons populasi nyata. Hubungan yang ditemukan bersifat prediktif, bukan kausal.

5. KESIMPULAN

Data survei dapat disebut cukup secara relatif ketika tujuan terbatas pada mempertahankan Macro F1 baseline di bawah kehilangan item acak: pada konfigurasi tanpa tiga proxy, kelengkapan 50% masih mempertahankan 94,8%–97,7% performa baseline. Namun, data belum cukup untuk prediksi individual yang andal karena Macro F1 baseline hanya 0,384 dan recall kelas menurun 0,010. Kehilangan terstruktur lebih berbahaya; monotone dropout dengan 50% atau 25% item awal teramati menurunkan retensi menjadi sekitar 79%. SK menjadi perspektif paling sensitif, tetapi efek section loss tidak signifikan setelah koreksi Holm. Penghapusan proxy menurunkan Macro F1 sebesar 0,030, F3 lebih rendah lagi, dan tidak ada aturan abstention yang memperbaiki deteksi kelas menurun secara memadai.

Secara metodologis, penelitian ini menunjukkan bahwa ambang kelengkapan harus selalu dilaporkan bersama performa absolut dan recall per kelas. Retensi 90% bukan bukti kelayakan ketika baseline lemah. Penelitian lanjutan memerlukan lebih banyak observasi kelas menurun, target akademik yang tervalidasi, pola missingness empiris, external validation, dan validasi konstruk independen.

6. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada para penyusun dataset Students' Academic Performance in Video-Conference-Assisted Online Learning During the COVID-19 Pandemic atas penyediaan data penelitian secara terbuka melalui Mendeley Data.

DAFTAR PUSTAKA

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6. https://doi.org/10.1186/s12864-019-6413-7

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Advances in Neural Information Processing Systems, 30.

Miranda, E., Aryuni, M., Rahmawati, M. I., Hiererra, S. E., & Sano, A. V. D. (2024a). Students' academic performance in video-conference-assisted online learning during the COVID-19 pandemic (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/x8mc6y45x2.1

Miranda, E., Aryuni, M., Rahmawati, M. I., Hiererra, S. E., & Sano, A. V. D. (2024b). Machine learning's model-agnostic interpretability on the prediction of students' academic performance in video-conference-assisted online learning during the COVID-19 pandemic. Computers and Education: Artificial Intelligence, 7, 100312. https://doi.org/10.1016/j.caeai.2024.100312

Mitra, R., McGough, S. F., Chakraborti, T., Holmes, C., Copping, R., Hagenbuch, N., et al. (2023). Learning from data with structured missingness. arXiv. https://doi.org/10.48550/arXiv.2304.01429

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6639–6649.

Pugnana, A., & Ruggieri, S. (2023). AUC-based selective classification. Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, 206, 2494–2514.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. Engineering Reports, 4(5), e12482. https://doi.org/10.1002/eng2.12482

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9, 11. https://doi.org/10.1186/s40561-022-00192-z