Analisis Komparatif TabPFN dan XGBoost Untuk Prediksi Putus Kuliah Mahasiswa Multikelas pada Data Tidak Seimbang

Tanti Ayu Hardiningtyas*, Ulfa Siti Nuraini
Program Studi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Surabaya, Indonesia

Received: 4 November 2026 Accepted: 10 November 2026 Published: 1 December 2026

*tanti.23002@mhs.unesa.ac.id

Kata Kunci:
Ketidakseimbangan Kelas; *Educational Data Mining*; Prediksi Putus Kuliah Mahasiswa; TabPFN; XGBoost

DSI: Jurnal Data Science Indonesia is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

Abstrak : Identifikasi dini risiko putus kuliah memerlukan evaluasi yang memperhatikan ketidakseimbangan kelas dan waktu ketersediaan prediktor. Penelitian ini membandingkan TabPFN-3.5 dan XGBoost pada data 4.424 mahasiswa satu politeknik di Portugal (UCI 697) dengan 30 fitur yang tidak memuat data semester kedua. Evaluasi memakai sepuluh pembagian berstrata 80:20 dengan Macro-F1 sebagai metrik utama, uji t *resampled* terkoreksi, dan penyesuaian Holm. TabPFN tanpa penyetelan memperoleh ROC-AUC makro yang lebih tinggi daripada XGBoost tersetel, sedangkan selisih Macro-F1 tidak signifikan setelah koreksi. Penyesuaian prior meningkatkan deteksi kelas Enrolled dan Macro-F1 TabPFN, tetapi menurunkan *recall* Dropout, dan kesetaraannya dengan pelatihan berbobot tidak terbukti. Hasil tambahan menunjukkan sensitivitas terhadap cakupan dan waktu ketersediaan fitur serta perbedaan jenis biaya komputasi. Temuan terbatas pada satu institusi dan belum mendukung penggunaan operasional tanpa validasi eksternal, audit keadilan, dan peninjauan lisensi model.

PENDAHULUAN

Putus kuliah berdampak pada mahasiswa dan institusi, sehingga pemodelan risiko sebelum akhir masa studi memiliki nilai praktis[1][2]. Namun, ketidakseimbangan status akhir studi membuat akurasi agregat kurang memadai. Dalam konteks Indonesia, statistik pendidikan tinggi menjadi motivasi masalah, bukan bukti bahwa model berbasis data Portugal dapat langsung digunakan pada perguruan tinggi Indonesia.

Penelitian terdahulu berbeda dalam target, jadwal prediksi, dan prosedur validasi[3]–[5]. Studi Martins et al. menguji informasi pendaftaran[3], sedangkan Vo et al. menilai algoritme *ensemble* pada tiga kelas dengan himpunan dan desain evaluasi tersendiri[5]. Belum ada bukti dari telaah ini yang membenarkan klaim kebaruan absolut; kontribusi penelitian dibatasi pada perbandingan berpasangan TabPFN-3.5 dan XGBoost dalam protokol dan titik waktu yang didefinisikan.

Tujuan penelitian ialah (1) membandingkan kinerja diskriminasi dan klasifikasi multikelas melalui Macro-F1 sebagai metrik utama; (2) menilai manfaat dan konsekuensi pembobotan kelas, koreksi prior, serta cakupan prediktor; dan (3) menelaah biaya komputasi, interpretasi prediktor, dan keterbatasan implementasi. Hipotesis eksploratifnya adalah adanya perbedaan Macro-F1 antara keluarga model pada partisi identik; ROC-AUC, akurasi, dan *balanced accuracy* merupakan metrik pendukung. Karena pemilihan metrik utama dicatat setelah desain asli disusun, inferensi tetap ditafsirkan secara hati-hati.

TINJAUAN LITERATUR

Penelusuran literatur bersifat terarah, bukan tinjauan sistematis. Publikasi tahun 2019–November 2026 ditelusuri melalui indeks akademik Google Scholar, Crossref, dan arXiv dengan kombinasi kata kunci "student dropout", "academic success", "UCI 697", "multiclass", "TabPFN", "tabular foundation model",

"XGBoost", dan "class imbalance"; literatur metodologis yang lebih lama ditelusuri melalui rujukan silang. Studi diseleksi menurut definisi target (biner/multikelas), periode ketersediaan fitur, metode, dan validasi. Sebagai contoh, Kesuma dan Handayani (2026) menggunakan UCI 697 tetapi mengubah tiga kelas menjadi target biner[6]. Karena pencarian tidak dimaksudkan sebagai kajian sistematis, hasilnya tidak digunakan untuk menyatakan bahwa penelitian serupa sama sekali belum pernah dilakukan.

Data UCI 697 menggabungkan informasi saat pendaftaran dengan nilai akademik semester pertama dan kedua untuk tiga hasil, yaitu Dropout, Enrolled, dan Graduate[4]. Perbandingan dengan publikasi lain harus mempertimbangkan ketersediaan fitur; pelaporan hasil absolut tanpa protokol yang identik tidak membuktikan keunggulan metode. Kajian umum memperlihatkan daya saing metode pohon pada data tabular, sementara TabPFN menawarkan prediksi menggunakan konteks pelatihan tanpa optimasi bobot khusus tugas[7]–[9]. Ringkasan studi pembanding disajikan pada Tabel 1.

Penelitian pada data pendidikan dengan ketidakseimbangan kelas juga perlu dibandingkan dengan studi yang secara eksplisit menguji strategi mitigasi. Villar dan de Andrade (2024) membandingkan beberapa algoritme pembelajaran terawasi untuk memprediksi putus kuliah dan keberhasilan akademik; hasilnya tidak dapat langsung dibandingkan karena rancangan, set prediktor, dan prosedur evaluasinya berbeda[10]. Pada ranah tabular yang lebih luas[1], TabArena menunjukkan perlunya mempertimbangkan variasi dataset, prosedur validasi, dan anggaran komputasi ketika menilai model fondasi tabular terhadap metode pohon[11]. Oleh karena itu, perbandingan dalam penelitian ini dibatasi pada protokol yang identik, bukan peringkat universal antarmodel.

Tabel 1. Penelitian pembanding dan perbedaan desain

Sumber	Data dan fitur	Metode	Validasi/temuan
Martins et al. (2021) [3]	Pendaftaran awal; institusi Portugal	Beberapa pengklasifikasi	Prediksi sebelum rekam semester; hasil tidak langsung sebanding
Vo et al. (2026) [5]	UCI 697; cakupan fitur berbeda	RF, LightGBM, XGBoost, CatBoost	RF: akurasi 0,7797; LightGBM: Macro-F1 0,7082
Kesuma & Handayani (2026) [6]	UCI 697; tiga kelas diubah menjadi target biner	XGBoost dengan SHAP	Target biner sehingga tidak langsung sebanding dengan tiga kelas
Hollmann et al. (2025) [8]	Himpunan data tabular berukuran kecil lintas domain	TabPFN (model fondasi tabular) vs pohon	Unggul pada tolok ukur data kecil; bukan data pendidikan
Grinsztajn et al. (2022) [7]	45 himpunan tabular lintas domain	Pohon vs jaringan saraf	*Benchmark* lintas tugas; bukan hasil spesifik pendidikan
Studi ini	UCI 697; 30 fitur tanpa semester kedua	TabPFN-3.5, XGBoost + ablasi	10 *holdout* berpasangan; Macro-F1 utama

Pada kelas tidak seimbang, pendekatan umum mencakup pembobotan biaya, pengambilan sampel ulang, dan penyesuaian aturan keputusan [12]. Studi ini menguji pembobotan saat pemodelan dan koreksi prior pascaprediksi, bukan seluruh pendekatan mitigasi. Koreksi prior dapat meningkatkan sensitivitas kelas minoritas tetapi dapat mengurangi ketepatan probabilitas dan kinerja kelas lain[13].

METODE PENELITIAN

Data, Waktu Prediksi, dan Desain Evaluasi

Himpunan data UCI 697 berasal dari Polytechnic Institute of Portalegre, Portugal[4], berisi 4.424 catatan dan 36 fitur. Kelas terdiri atas Graduate 2.209 (49,93%), Dropout 1.421 (32,12%), dan Enrolled 794 (17,95%). Enam fitur semester kedua dikeluarkan sehingga terdapat 30 fitur: 11 latar belakang/demografi, 3 status finansial, 7 pendaftaran/penerimaan, 3 makroekonomi, dan 6 akademik semester pertama. Daftar lengkap beserta tipe tercantum pada Tabel 2. Dokumentasi publik menjelaskan asal informasi pendaftaran dan semester, tetapi tidak menetapkan tanggal pembaruan setiap status keuangan; oleh karena itu prediksi "akhir semester pertama" diperlakukan sebagai skenario, bukan kepastian bebas kebocoran informasi.

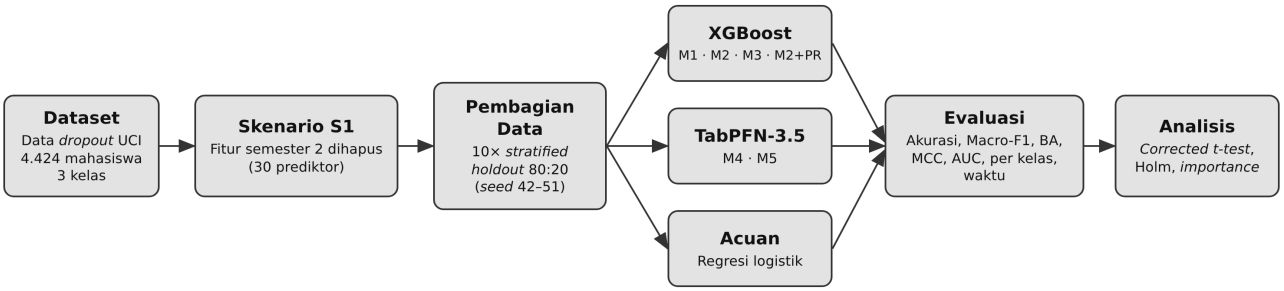
Tabel 2. Daftar 30 prediktor yang digunakan

No	Prediktor	Kelompok	Jenis pengukuran	Nilai unik	Rentang nilai

1	Marital status	Latar belakang/demografi	Kategorik	6	1–6
2	Application mode	Pendaftaran/penerimaan	Kategorik	18	1–57
3	Application order	Pendaftaran/penerimaan	Numerik	8	0–9
4	Course	Pendaftaran/penerimaan	Kategorik	17	33–9991
5	Daytime/evening attendance	Pendaftaran/penerimaan	Kategorik	2	0–1
6	Previous qualification	Pendaftaran/penerimaan	Kategorik	17	1–43
7	Previous qualification (grade)	Pendaftaran/penerimaan	Numerik	101	95–190
8	Nacionality	Latar belakang/demografi	Kategorik	21	1–109
9	Mother's qualification	Latar belakang/demografi	Kategorik	29	1–44
10	Father's qualification	Latar belakang/demografi	Kategorik	34	1–44
11	Mother's occupation	Latar belakang/demografi	Kategorik	32	0–194
12	Father's occupation	Latar belakang/demografi	Kategorik	46	0–195
13	Admission grade	Pendaftaran/penerimaan	Numerik	620	95–190
14	Displaced	Latar belakang/demografi	Kategorik	2	0–1
15	Educational special needs	Latar belakang/demografi	Kategorik	2	0–1
16	Debtor	Status finansial	Kategorik	2	0–1
17	Tuition fees up to date	Status finansial	Kategorik	2	0–1
18	Gender	Latar belakang/demografi	Kategorik	2	0–1
19	Scholarship holder	Status finansial	Kategorik	2	0–1
20	Age at enrollment	Latar belakang/demografi	Numerik	46	17–70
21	International	Latar belakang/demografi	Kategorik	2	0–1
22	Curricular units 1st sem (credited)	Akademik semester pertama	Numerik	21	0–20
23	Curricular units 1st sem (enrolled)	Akademik semester pertama	Numerik	23	0–26
24	Curricular units 1st sem (evaluations)	Akademik semester pertama	Numerik	35	0–45
25	Curricular units 1st sem (approved)	Akademik semester pertama	Numerik	23	0–26
26	Curricular units 1st sem (grade)	Akademik semester pertama	Numerik	805	0–18,88
27	Curricular units 1st sem (without evaluations)	Akademik semester pertama	Numerik	11	0–12
28	Unemployment rate	Makroekonomi	Numerik	10	7,6–16,2
29	Inflation rate	Makroekonomi	Numerik	9	−0,8 s.d. 3,7
30	GDP	Makroekonomi	Numerik	10	−4,06 s.d. 3,51

Keterangan: Nama prediktor mengikuti himpunan data UCI 697 (termasuk ejaan asli "Nacionality"). Kategorik=kategori yang dikodekan numerik.

Tidak ada nilai hilang pada seluruh prediktor Pengelompokan hanya untuk keperluan penyajian dan bukan definisi resmi codebook UCI. Sepuluh pembagian berstrata dengan benih 42-51 menggunakan 3.539 data latih dan 885 data uji per pembagian. Pemilihan hiperparameter, bobot kelas, dan prapemrosesan dilakukan di dalam data latih. Pengujian pada partisi tumpang tindih tidak boleh ditafsirkan sebagai sepuluh sampel independen. Skenario tambahan merupakan eksperimen pendukung, bukan pengganti angka utama, dan hasil dari protokol berbeda tidak digabungkan dengan angka utama artikel. Alur penelitian ditunjukkan pada Gambar 1. Ketergantungan variabel makro pada periode angkatan diperiksa melalui ablasi tanpa tiga variabel makro. Validasi GroupKFold berbasis angkatan tidak dilakukan karena identitas angkatan tidak tersedia; ablasi tersebut bukan pengganti pengujian generalisasi lintas angkatan.



Gambar 1. Alur penelitian. Seluruh langkah penyetelan dan pembobotan hanya menggunakan partisi latih pada setiap pembagian

Konfigurasi dan Ukuran Kinerja

Konfigurasi utama adalah M1 (XGBoost bawaan), M2 (XGBoost disetel), M3 (XGBoost disetel dan berbobot kelas), M4 (TabPFN-3.5 bawaan), M5 (M4 dengan penyesuaian prior), dan M6 (M2 dengan penyesuaian prior). Regresi logistik hanya menjadi acuan. XGBoost memakai metode histogram, kategori bawaan, dan empat utas CPU [14]. Penyetelan hiperparameter memakai Optuna TPE dengan 50 percobaan, validasi silang berstrata lima lipatan pada partisi latih untuk memaksimalkan Macro-F1, serta penghentian dini [15]. TabPFN memakai paket versi 9.2.0, *checkpoint* 3.5, delapan anggota *ensemble* bawaan, dan GPU [9]. Penyetelan jumlah anggota *ensemble* hanya diuji terpisah sebagai analisis sensitivitas, tanpa menyamakan anggaran optimasi. Sebagai analisis sensitivitas terpisah, TabPFN juga dievaluasi dengan empat anggota *ensemble* pada sepuluh pembagian yang sama; protokol tambahan ini tidak mengubah definisi M4 utama.

Fungsi objektif XGBoost menggunakan kerugian multinomial dengan regularisasi yang memuat penalti L2 dan L1, masing-masing dari hiperparameter reg_lambda dan reg_alpha (Persamaan (1) dan (2)).

$$\mathcal{L} = \sum_{i=1}^{n} \ell\,(y_i, \hat{y}_i) + \sum_{b=1}^{B} \Omega\,(f_b) \text{.....(1)}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda_2 \parallel \mathbf{w} \parallel_2^2 + \lambda_1 \parallel \mathbf{w} \parallel_1 \text{.....(2)}$$

dengan ℓ kerugian multinomial, f_b pohon ke-b, B jumlah pohon, T jumlah daun, $\mathbf{w}$ vektor bobot daun, γ penalti jumlah daun, serta λ_2 dan λ_1 koefisien penalti L2 dan L1. Bobot kelas dihitung dengan Persamaan (3), dengan n jumlah data latih, $K = 3$ jumlah kelas, dan n_k jumlah data latih kelas k.

$$w_k = \frac{n}{K\, n_k} \text{.....(3)}$$

Koreksi prior pascaprediksi mengubah probabilitas p_{ik} pengamatan ke-i untuk kelas k menjadi skor p'_{ik} menurut Persamaan (4), dengan α eksponen bobot ($\alpha = 1$ pada M5 dan M6).

$$p'_{ik} = \frac{w_k^{\alpha}\, p_{ik}}{\sum_{j=1}^{K} w_j^{\alpha}\, p_{ij}} \text{.....(4)}$$

Skor tersebut tidak diasumsikan terkalibrasi. Angka n_estimators dalam pencarian adalah batas yang disarankan untuk proses penghentian dini, bukan otomatis jumlah pohon akhir; median dan rentang antarkuartil seluruh sembilan hiperparameter disajikan pada Tabel 3; hasil utama dilaporkan untuk konfigurasi yang dipilih melalui prosedur penyetelan pada masing-masing pembagian.

Tabel 3. Hiperparameter XGBoost terpilih pada sepuluh pembagian (median dan rentang antarkuartil)

Hiperparameter	M2: median (Q1–Q3)	M3: median (Q1–Q3)
Macro-F1 CV terbaik	0,6738 (0,6712–0,6776)	0,6939 (0,6927–0,6983)
n_estimators	630,5 (360,8–752)	571,5 (515,2–666,2)
max_depth	4,5 (4–5,8)	6 (5,2–8)
learning_rate	0,0876 (0,0539–0,1230)	0,0376 (0,0180–0,0721)
subsample	0,8389 (0,7694–0,9147)	0,8313 (0,8024–0,8694)
colsample_bytree	0,6375 (0,5693–0,7710)	0,6463 (0,5858–0,8086)
min_child_weight	6 (3,2–7)	6 (4,2–10)
gamma	0,7603 (0,5602–1,0330)	0,6790 (0,2009–1,1863)
reg_lambda	2,9496 (0,4053–4,8183)	0,2544 (0,0682–1,9257)

reg_alpha	1,0771 (0,6162–1,2632)	1,1262 (0,3091–2,6358)

Keterangan: Nilai dihitung dari sepuluh pembagian (benih 42–51); kuartil memakai interpolasi linear. n_estimators adalah batas yang disarankan untuk penghentian dini, bukan otomatis jumlah pohon akhir.

Macro-F1 merupakan metrik utama, yaitu rata-rata F1 per kelas, sedangkan *balanced accuracy* adalah rata-rata *recall* per kelas (Persamaan (5) dan (6)). Hipotesis eksploratif dan penetapan metrik utama dinyatakan pada Pendahuluan; karena dicatat setelah desain awal, seluruh inferensi ditafsirkan secara hati-hati.

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K} \text{F1}_k\,, \qquad \text{F1}_k = \frac{2\,P_k R_k}{P_k + R_k} \quad \text{.....(5)}$$

$$\text{BA} = \frac{1}{K}\sum_{k=1}^{K} R_k \text{.....(6)}$$

dengan P_k dan R_k berturut-turut presisi dan *recall* kelas k. MCC multikelas dihitung menurut Persamaan (7) [16].

$$\text{MCC} = \frac{c\,s - \sum_{k=1}^{K} p_k t_k}{\sqrt{s^2 - \sum_{k=1}^{K} p_k^2}\,\sqrt{s^2 - \sum_{k=1}^{K} t_k^2}} \text{.....(7)}$$

dengan c jumlah prediksi benar, s jumlah pengamatan, serta p_k dan t_k berturut-turut jumlah prediksi dan jumlah aktual kelas k. Akurasi, macro ROC-AUC *one-vs-rest*, *log loss*, dan skor Brier merupakan metrik pelengkap.

Ringkasan kontras terdapat pada Tabel 4: A = M4−M1, B = M4−M2, C = M5−M3, C′ = M5−M6, dan D = M6 − M3 (eksploratif). Selisih berpasangan antarkonfigurasi pada pembagian ke-j, Δ_j, diuji dengan *corrected resampled t-test* [17] menurut Persamaan (8), dengan $\bar{\Delta}$ rerata selisih, s_d simpangan baku Δ_j, $J = 10$ jumlah pembagian, $n_{\text{test}} = 885$, dan $n_{\text{train}} = 3.539$.

$$t = \frac{\bar{\Delta}}{\text{SE}(\bar{\Delta})}, \qquad \text{SE}(\bar{\Delta}) = s_d\sqrt{\frac{1}{J} + \frac{n_{\text{test}}}{n_{\text{train}}}} \text{.....(8)}$$

Koreksi Holm diterapkan atas 16 uji utama [18]; interval keluarga dihitung konservatif untuk 16 perbandingan, sedangkan D dibentuk sebagai keluarga terpisah. Uji TOST untuk D memakai margin kesetaraan ±0,01 pada Macro-F1.

Importance permutasi berkelompok mengukur penurunan Macro-F1 ketika seluruh kolom dalam satu kelompok fitur g dipermutasi bersama pada data uji [19], sebagaimana Persamaan (9).

$$I_g = \text{Macro-F1}(f; \mathcal{D}) - \frac{1}{Q}\sum_{q=1}^{Q} \text{Macro-F1}\left(f; \widetilde{\mathcal{D}}_g^{(q)}\right) \text{.....(9)}$$

dengan f model terlatih, $\mathcal{D}$ data uji, $\widetilde{\mathcal{D}}_g^{(q)}$ data uji dengan kolom kelompok g dipermutasi pada pengulangan ke-q, dan Q jumlah pengulangan permutasi (Q = 3) yang sama untuk M3 dan M4.

Untuk setiap kelompok fitur, Q = 3 pengulangan permutasi digunakan pada masing-masing dari sepuluh pembagian data (S = 10). Kolom dalam satu kelompok dipermutasi bersama pada data uji, dengan indeks permutasi yang disamakan untuk M3 dan M4. Nilai yang dilaporkan merupakan rerata penurunan Macro-F1, bukan efek kausal.

Tabel 4. Rancangan perbandingan berpasangan antarkonfigurasi

Kontras	Konfigurasi	Pertanyaan Analitis
A	M4 vs M1	Dua keluarga model pada pengaturan bawaan
B	M4 vs M2	TabPFN bawaan vs XGBoost setelah penyetelan
C	M5 vs M3	Dua *pipeline* yang menangani ketidakseimbangan
C′	M5 vs M6	Kedua model dengan koreksi prior serupa
D	M6 vs M3	Koreksi prior vs pembobotan dalam XGBoost (eksploratif)

Analisis Sensitivitas dan Etika

Eksperimen tambahan mencakup CatBoost, LightGBM, Random Forest, fitur rasio lulus/diambil, penghapusan tiga indikator keuangan yang waktunya tidak pasti, pengeluaran mahasiswa dengan aktivitas akademik nol, penghapusan variabel makro, skenario pendaftaran dan 36 fitur, sensitivitas eksponen prior, kalibrasi, *importance* permutasi berkelompok dengan desain setara, kinerja risiko biner dan subkelompok, serta *benchmark* inferensi. Eksperimen menggunakan 10 benih kecuali pemeriksaan benih internal TabPFN

(3×3); batas validasi eksternal dan pembagian antarangatan tetap berlaku.

Data UCI berlisensi CC BY 4.0 [20]. Lisensi bobot TabPFN-3.5 versi 1.0 membatasi pemakaian nonkomersial dan nonproduksi; layanan operasional memerlukan izin yang sesuai [21]. Model memasukkan atribut demografis; analisis subkelompok eksploratif bukan sertifikasi keadilan. Kode dan hasil eksperimen memakai benih acak 42–51.

HASIL PENELITIAN

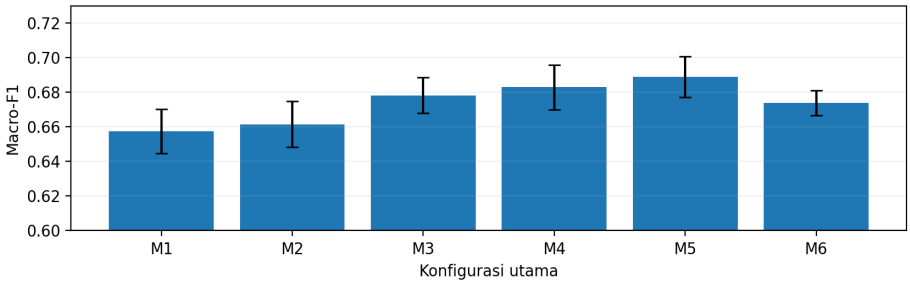
Perbandingan Utama

Tabel 5 menyajikan rerata ± simpangan baku sepuluh pembagian pada skenario 30 fitur. Macro-F1 tertinggi di antara konfigurasi utama adalah M5 (0,6889), sedangkan M4 mencapai akurasi dan diskriminasi ROC-AUC tertinggi. M3 dan M6 memiliki *balanced accuracy* (0,6810–0,6811) di atas M4 (0,6743); karena itu tidak benar bahwa TabPFN bawaan melampaui semua XGBoost pada semua ukuran. Sebaran Macro-F1 pada setiap pembagian ditunjukkan pada Gambar 2.

Tabel 5. Kinerja utama pada 30 fitur (rata-rata ± SB)

Konf.	Akurasi	Macro-F1*	*Balanced acc.*	ROC-AUC makro
M1	0,7382 ± 0,0120	0,6574 ± 0,0128	0,6504 ± 0,0124	0,8542 ± 0,0125
M2	0,7473 ± 0,0121	0,6615 ± 0,0133	0,6540 ± 0,0129	0,8663 ± 0,0091
M3	0,7261 ± 0,0127	0,6783 ± 0,0103	0,6810 ± 0,0103	0,8647 ± 0,0113
M4	0,7608 ± 0,0111	0,6830 ± 0,0129	0,6743 ± 0,0123	0,8828 ± 0,0098
M5	0,7262 ± 0,0120	0,6889 ± 0,0119	0,7001 ± 0,0115	0,8810 ± 0,0097
M6	0,7157 ± 0,0103	0,6739 ± 0,0071	0,6811 ± 0,0069	0,8629 ± 0,0087

Keterangan: *Macro-F1 merupakan metrik utama. M6 adalah XGBoost M2 dengan koreksi prior. Nilai ROC-AUC M5 menggunakan skor penyesuaian; perbandingan inferensial mengikuti Tabel 6.



Gambar 2. Macro-F1 sepuluh pembagian (rerata ± simpangan baku); perhatikan tumpang tindih antarpembagian.

Tabel 6. Uji berpasangan terpilih dan pengujian kesetaraan

Perbandingan	Metrik	Selisih	IK keluarga 95%	p Holm
A (M4−M1)	Macro-F1	+0,0256	[−0,0046; 0,0558]	0,0880
B (M4−M2)	Macro-F1	+0,0215	[−0,0130; 0,0560]	0,2428
B (M4−M2)	ROC-AUC	+0,0165	[0,0047; 0,0283]	0,0047
C (M5−M3)	Macro-F1	+0,0106	[−0,0172; 0,0384]	0,3247
C′ (M5−M6)	ROC-AUC	+0,0181	[0,0042; 0,0320]	0,0054
D (M6−M3)	Macro-F1	−0,0044	[−0,0160; 0,0072]	0,8026

Keterangan: IK keluarga mengendalikan multiplisitas pada keluarga uji yang ditentukan; p Holm D berasal dari empat uji eksploratif, sehingga tidak dibandingkan langsung dengan p keluarga utama. Pada D, TOST ±0,01 menghasilkan $p = 0{,}0842$ dan IK 90% [−0,01125; 0,00243], sehingga kesetaraan tidak terbukti. Perbedaan ROC-AUC lebih meyakinkan daripada perbedaan Macro-F1; selisih Macro-F1 pada perbandingan C (0,0106) berada dalam rentang simpangan baku antarpembagian (0,007-0,013), sehingga ukuran efeknya ditafsirkan hati-hati; sepuluh kemenangan bukan sepuluh bukti independen.

Eksperimen Tambahan dan Sensitivitas Waktu Fitur

Pada sepuluh pembagian, CatBoost bawaan memperoleh Macro-F1 0,6484, LightGBM 0,6603, Random Forest 0,6185; XGBoost M3 dengan rasio fitur dan konfigurasi dibekukan 0,6811, sedangkan versi rasio

disetel kembali 0,6845. Karena perlakuan penyetelan tidak identik, hasil ini menunjukkan ketergantungan pada rancangan *benchmark*, bukan dominasi universal. Skenario pendaftaran saja (Tabel 7) menghasilkan Macro-F1 0,5748 (M3) dan 0,5600 (M4); ketika semester kedua disertakan, nilainya menjadi 0,7049 dan 0,7230. Selisih terhadap 30 fitur memperlihatkan bahwa waktu tersedianya data memengaruhi batas keterbandingan dengan Vo et al. (2026)[5].

Analisis sensitivitas jumlah anggota *ensemble* menunjukkan bahwa TabPFN dengan empat anggota memperoleh Macro-F1 0,6858 ± 0,0149 dan balanced accuracy 0,6770 pada sepuluh pembagian; waktu prediksi rata-rata 17,76 detik untuk 885 mahasiswa. Sebagai pembanding deskriptif, M4 dengan delapan anggota mencapai Macro-F1 0,6830 ± 0,0129. Karena hasil waktu dan pengaturan berasal dari rangkaian pengukuran berbeda, perbedaan tersebut tidak dinyatakan signifikan ataupun sebagai bukti bahwa empat anggota lebih unggul.

Tabel 7. Sensitivitas cakupan fitur dan pembanding tambahan (Macro-F1)

Eksperimen	XGBoost M3 / varian	TabPFN M4 / varian	Keterangan
Pendaftaran saja	0,5748 ± 0,0137	0,5600 ± 0,0142	Data pendaftaran (24 fitur)
30 fitur, utama	0,6783 ± 0,0103	0,6830 ± 0,0129	Pembanding utama (30 fitur)
36 fitur penuh	0,7049 ± 0,0144	0,7230 ± 0,0147	Retrospektif (36 fitur)
Tanpa indikator keuangan[1]	0,6445 (27 fitur)	0,6455 (27 fitur)	Ablasi 3 indikator keuangan
Tanpa peserta 0 unit lulus	0,6397	0,6579	Eksklusi unit lulus nol
Tanpa 3 fitur makro	0,6732	0,6729	Ablasi indikator makro

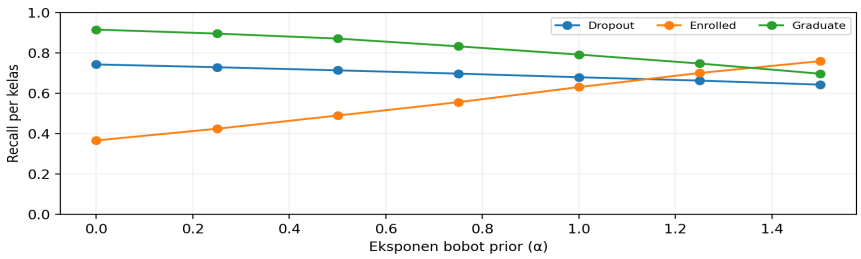
Tabel 8. Kinerja akademik semester pertama menurut kelas (rerata ± SB)

Kelas	n	Unit lulus	Nilai rata-rata
Dropout	1421	2,55 ± 2,86	7,26 ± 6,03
Enrolled	794	4,32 ± 2,29	11,13 ± 3,68
Graduate	2209	6,23 ± 2,58	12,64 ± 2,70

Tabel 8 memperlihatkan bahwa rata-rata unit lulus dan nilai semester pertama paling rendah pada kelas Dropout. Tiga indikator keuangan yang dikeluarkan ialah *Debtor*, *Tuition fees up to date*, dan *Scholarship holder*. Pada ablasi 27 fitur, Macro-F1 M3/M4 menjadi 0,6445/0,6455, lebih rendah daripada skenario 30 fitur; percobaan tersebut bukan bukti tanggal pencatatan atribut. Jumlah bernilai nol untuk unit semester pertama yang diambil, dievaluasi, dan lulus pada kelas Dropout masing-masing 77, 246, dan 570; pada Enrolled 28, 28, dan 71; serta pada Graduate 75, 75, dan 77. Analisis mengeluarkan mahasiswa dengan nol unit lulus menunjukkan penurunan Macro-F1, menyoroti risiko identifikasi status yang sudah tampak dalam rekam akademik.

Kalibrasi, Prior, Risiko Intervensi, dan Pentingnya Fitur

Pada M2, M4, dan M5, galat kalibrasi harapan (ECE; 15 *bin*) berturut-turut 0,0358, 0,0337, dan 0,0431; skor Brier multikelas 0,3524, 0,3297, dan 0,3656. Dengan demikian, M4 memiliki ECE dan Brier lebih rendah daripada M2, sedangkan penyesuaian prior (M5) memperburuk keduanya. Koreksi prior $\alpha = 0{,}5$ menghasilkan Macro-F1 M4 sebesar 0,6939 dibandingkan 0,6889 pada $\alpha = 1{,}0$ (Gambar 3), tetapi pencarian α secara retrospektif di data uji tidak dapat menjadi dasar klaim model terbaik. Penyetelan α pada bagian validasi di dalam data latih memberikan Macro-F1 0,6937 untuk TabPFN pada protokol tambahan, dan hasilnya tidak boleh disamakan langsung dengan M5 utama.



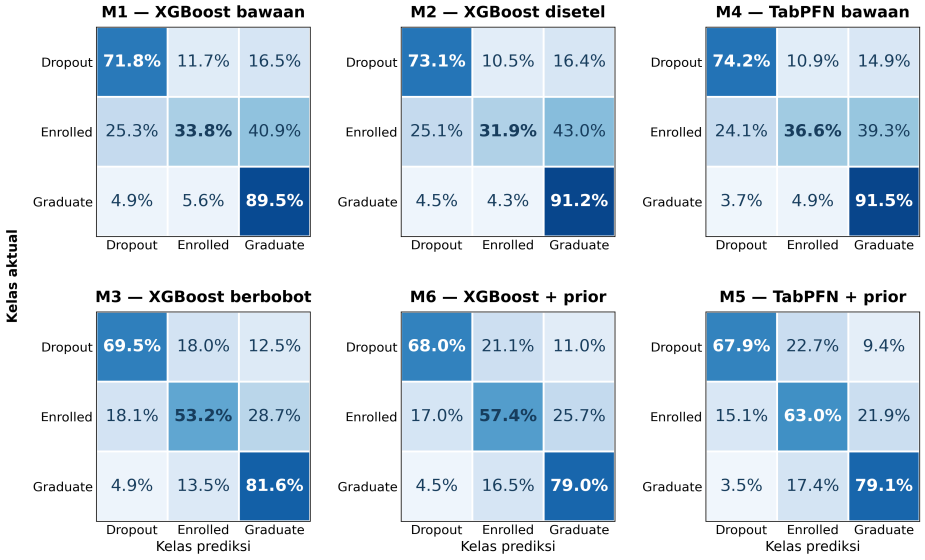
Gambar 3. Perubahan *recall* ketiga kelas terhadap eksponen bobot prior untuk M4; eksplorasi pada set uji, bukan prosedur penyetelan model akhir.

M4 ke M5 menaikkan *recall* Enrolled 0,366→0,630, tetapi menurunkan *recall* Dropout 0,742→0,679 serta Graduate 0,915→0,791. Jika risiko operasional didefinisikan sebagai prediksi Dropout atau Enrolled, M4 menandai 376,0 dari 885 mahasiswa (presisi 0,8999; *recall* 0,7637); M5 menandai 473,8 (presisi 0,8053; *recall* 0,8612). M5 menandai rata-rata 241,5 sebagai Enrolled dengan 141,3 positif palsu. Angka ini menggambarkan kebutuhan telaah manusia dan belum mengukur biaya nyata intervensi. Ukuran tersebut dirangkum pada Tabel 9.

Tabel 9. Perubahan *recall* per kelas dan risiko biner

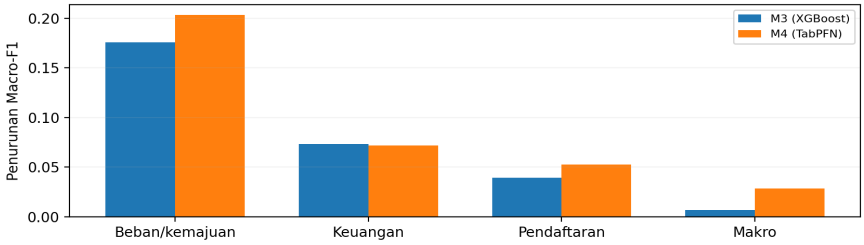
Ukuran	M4 TabPFN	M5 + prior
Recall Dropout	0,742	0,679
Recall Enrolled	0,366	0,630
Recall Graduate	0,915	0,791
Precision risiko bukan Graduate	0,900	0,805
Recall risiko bukan Graduate	0,764	0,861
Ditandai sebagai risiko (dari 885)	376,0	473,8

Gambar 4 memperlihatkan matriks kesalahan tiga kelas pada enam konfigurasi utama. Pada M4, 39,3% mahasiswa Enrolled diprediksi Graduate, turun menjadi 21,9% pada M5; kesalahan Dropout→Graduate turun dari 14,9% menjadi 9,4%. Sebaliknya, Graduate→Enrolled naik dari 4,9% menjadi 17,4%. Perubahan ini membantu mendeteksi mahasiswa yang belum selesai studi, tetapi meningkatkan jumlah peringatan palsu.



Gambar 4. Matriks kebingungan (*confusion matrix*) ternormalisasi menurut kelas aktual; angka pada sel menunjukkan persentase dari 10 pembagian. M6 adalah M2 dengan penyesuaian prior.

Jumlah pengamatan uji kumulatif untuk Dropout/Enrolled/Graduate adalah 2.840/1.590/4.420 (pengamatan antar pembagian bertumpang tindih).



Gambar 5. *Importance* permutasi berkelompok pada sepuluh pembagian yang sama; penurunan Macro-F1 tidak bersifat kausal.

Pada Gambar 5, *importance* permutasi berkelompok yang disamakan pada sepuluh pembagian

menunjukkan bahwa grup beban/kemajuan akademik memberi penurunan Macro-F1 rata-rata 0,1756 pada M3 dan 0,2034 pada M4; status keuangan 0,0733 dan 0,0717. Fitur berkorelasi dipermutasi bersama; angka ini menunjukkan ketergantungan prediktif, bukan sebab-akibat maupun bukti mekanisme representasi internal model[19].

Pada pemeriksaan benih internal TabPFN, tiga benih pembagian (42–44) disilangkan dengan tiga benih model (42–44), menghasilkan sembilan kombinasi yang tidak independen. Rerata Macro-F1 adalah 0,6921; simpangan baku deskriptif sembilan kombinasi 0,0035, rerata simpangan baku di dalam pembagian 0,0035, dan simpangan baku rerata antarpembagian 0,0008. Karena desain ini berbeda dengan sepuluh pembagian utama, angkanya tidak digunakan untuk mengklaim peningkatan kinerja M4.

Biaya Komputasi dan Subkelompok

Tabel 10. Waktu komputasi untuk 885 data uji (detik)

Konfigurasi / protokol	Penyetelan	*Fit*	Prediksi	Perangkat
M2; *holdout* utama	157,28	<2	≈0,01	CPU, 4 utas
M3; *holdout* utama	275,67	<2	≈0,01	CPU, 4 utas
M4; *holdout* utama	0	1,16	27,49	RTX 4050
M4; *benchmark* berulang	0	0,99	31,77	RTX 4050
M4; *cache* berulang	0	11,76	15,19	RTX 4050

Tabel 10 memisahkan biaya penyetelan, pelatihan (*fit*), dan prediksi. Waktu *holdout* utama dan *benchmark* berulang berasal dari protokol berbeda, sehingga tidak boleh diperlakukan sebagai ulangan berpasangan. *Benchmark first-use/warmed-up* tambahan juga memperlihatkan sensitivitas kondisi inferensi; data yang ada belum mengisolasi semua komponen pemuatan *checkpoint* dan sinkronisasi GPU. Pada laptop yang sama, XGBoost CPU lebih cepat daripada konfigurasi CUDA yang diuji, tetapi temuan tersebut tidak berlaku universal. Penyetelan merupakan biaya yang berpotensi diamortisasi, sedangkan prediksi berulang per angkatan merupakan biaya operasional. Sebagai bukti perbandingan CPU dan CUDA khusus pada laptop ini, M1 menghasilkan waktu *fit* 0,2756 vs 1,0722 detik dan prediksi 0,0084 vs 0,0243 detik; M2 0,3030 vs 1,4525 detik serta 0,0071 vs 0,0239 detik; M3 0,3441 vs 1,9826 detik serta 0,0065 vs 0,0281 detik (urutan CPU vs CUDA). Hasil ini bersifat konfigurasi-spesifik.

Analisis subkelompok pada jenis kelamin, usia, status perantau, dan beasiswa menunjukkan kesenjangan *recall* risiko; misalnya M4 memiliki rerata selisih *recall* 0,1246 antarkategori jenis kelamin dan 0,2171 antarkelompok usia. Penghapusan atribut demografis mengubah Macro-F1 M4 dari 0,6830 menjadi 0,6848 dalam eksperimen tambahan, tetapi tidak menghapus seluruh kesenjangan. Subkelompok kecil dan faktor perancu membatasi interpretasi keadilan (*fairness*).

PEMBAHASAN

Makna Perbandingan dan Implikasi Akademik

Temuan pada satu himpunan data menyokong keunggulan diskriminasi ROC-AUC TabPFN terhadap XGBoost tersetel pada protokol yang diuji, tetapi tidak mendukung klaim bahwa seluruh kinerja TabPFN lebih tinggi. Macro-F1 M5 relatif tinggi terutama karena peningkatan deteksi Enrolled, dengan kompromi pada Dropout dan beban peninjauan mahasiswa yang ditandai. Koreksi prior merupakan pilihan aturan keputusan yang memerlukan kebijakan biaya kesalahan serta validasi prospektif; nilai p nonsignifikan maupun kemiripan rata-rata bukan bukti kesetaraan.

Perbedaan rerata Macro-F1 M2 saat validasi silang penyetelan (0,6747) dan evaluasi uji (0,6615) mencapai 0,0132, sebanding dengan selisih sejumlah model, sehingga tidak tepat disebut "optimisme ringan". Analisis tambahan XGBoost dengan rekayasa fitur, model pembanding pohon, dan variasi jumlah *ensemble* TabPFN menunjukkan bahwa alokasi optimasi berpengaruh. Walaupun skenario 36 fitur mencapai kinerja lebih tinggi, data semester kedua tidak layak dipakai untuk keputusan di akhir semester pertama.

Perbandingan dengan Penelitian Terdahulu

Vo et al. (2026) melaporkan RF berakurasi 0,7797 dan LightGBM dengan Macro-F1 0,7082 pada cakupan prediktor dan validasi berbeda [5]. Uji dalam penelitian ini dengan 36 fitur menghasilkan Macro-F1 lebih tinggi daripada skenario 30 fitur pada kedua keluarga, tetapi tidak menjadikan desain Vo et al. Identik [5].

Studi *benchmark* lintas domain tidak dapat digugurkan maupun dikukuhkan oleh satu himpunan pendidikan[7].

Keterbatasan dan Penerapan

Data berasal dari satu institusi Portugal; arti Enrolled terkait status pada akhir durasi nominal program, sedangkan struktur studi, beasiswa, dan variabel makro di Indonesia dapat berbeda. Data tidak memuat tahun masuk yang diperlukan untuk validasi temporal yang dapat diaudit. Tanggal pencatatan indikator keuangan tidak terkonfirmasi; mahasiswa dengan unit semester pertama nol mungkin telah menunjukkan kondisi putus kuliah sebelum titik prediksi. Model juga diuji pada pembagian yang saling bergantung; uji terkoreksi dan IK keluarga tidak menghilangkan semua ketidakpastian akibat pemilihan model dan penyetelan.

Analisis subkelompok masih deskriptif, dan profil biaya intervensi institusional belum tersedia. Lisensi TabPFN-3.5 membatasi pemakaian produksi. Dengan demikian keluaran belum layak menjadi dasar otomatis klasifikasi mahasiswa; penggunaan di institusi memerlukan pengujian eksternal, pencatatan tanggal fitur, izin penggunaan model, audit manfaat-risiko, serta persetujuan tata kelola data.

KESIMPULAN

Pertama, pada 30 fitur dan sepuluh pembagian berstrata, TabPFN-3.5 menunjukkan ROC-AUC makro lebih baik daripada XGBoost tersetel ($\Delta = 0{,}0165$; $p_{\mathrm{Holm}} = 0{,}0047$), tetapi perbedaan Macro-F1 tidak signifikan setelah koreksi. Kedua, penyesuaian prior meningkatkan deteksi Enrolled dan Macro-F1 pada TabPFN, namun mengurangi *recall* Dropout dan meningkatkan positif palsu; uji TOST tidak membuktikan kesetaraan dengan pelatihan berbobot XGBoost. Ketiga, TabPFN menghindari biaya penyetelan khusus tugas tetapi lebih mahal saat inferensi, dan kelompok kemajuan akademik menjadi prediktor penting secara asosiatif. Hasil tambahan menegaskan sensitivitas terhadap informasi semester, status keuangan, dan subkelompok. Sebelum penerapan, dibutuhkan validasi eksternal/temporal pada konteks pendidikan Indonesia, audit keadilan, penaksiran biaya intervensi, serta kesesuaian lisensi.

DAFTAR PUSTAKA

[1] D. Andrade-Girón *et al.*, "Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review," *ICST Trans. Scalable Inf. Syst.*, Jul. 2023.

[2] M. Quimiz-Moreira, R. Delgadillo, J. Parraga-Alava, N. Maculan, and D. Mauricio, "Factors, Prediction, Explainability, and Simulating University Dropout Through Machine Learning: A Systematic Review, 2012–2024," *Computation*, vol. 13, no. 8, p. 198, Aug. 2025.

[3] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, "Early Prediction of student's Performance in Higher Education: A Case Study," in *Trends and Applications in Information Systems and Technologies*, 2021, pp. 166–175.

[4] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," *Data*, vol. 7, no. 11, p. 146, Oct. 2022.

[5] N. T. Vo, Q. N. Le, and H. Le, "Artificial Intelligence-Driven Early Prediction of Student Dropout and Academic Outcomes in Higher Education: A Comparative Study of Advanced Machine Learning Approaches," *EAI Endorsed Trans. Ind. Networks Intell. Syst.*, vol. 13, no. 1, Mar. 2026.

[6] C. Kesuma and V. R. Handayani, "Student Dropout Prediction using XGBoost and Explainable AI (SHAP): A Case Study of Portuguese Student Dataset," *Int. J. Informatics, Econ. Manag. Sci.*, vol. 5, no. 2, 2026.

[7] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?," *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022.

[8] N. Hollmann *et al.*, "Accurate Predictions on Small Data with A Tabular Foundation Model," *Nature*, vol. 637, no. 8045, pp. 319–326, Jan. 2025.

[9] B. Jäger *et al.*, "TabPFN-3.5: Technical Report," *arXiv*, 2026.

[10] A. Villar and C. R. V. de Andrade, "Supervised Machine Learning Algorithms for Predicting Student Dropout and Academic Success: A Comparative Study," *Discov. Artif. Intell.*, vol. 4, no. 1, p. 2, Jan.

2024.
[11] N. Erickson *et al.*, "TabArena: A Living Benchmark for Machine Learning on Tabular Data," *Adv. Neural Inf. Process. Syst. (NeurIPS 2025). Datasets Benchmarks Track*, 2025.
[12] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[13] M. Saerens, P. Latinne, and C. Decaestecker, "Adjusting the Outputs of A Classifier to New A Priori Probabilities: A Simple Procedure," *Neural Comput.*, vol. 14, no. 1, pp. 21–41, Jan. 2002.
[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.
[16] J. Gorodkin, "Comparing Two K-category Assignments by A K-Category Correlation Coefficient," *Comput. Biol. Chem.*, vol. 28, no. 5–6, pp. 367–374, Dec. 2004.
[17] C. Nadeau and Y. Bengio, "Inference for the Generalization Error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003.
[18] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.
[19] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning A Variable's Importance by Studying An Entire Class of Prediction Models Simultaneously," *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2017.
[20] V. Realinho, M. Vieira Martins, J. Machado, and L. Baptista, *Predict Students' Dropout and Academic Success*. UCI Machine Learning Repository, 2021.
[21] Prior Labs, "TABPFN-3.5 License v1.0," 2026. [Online]. Available: https://huggingface.co/Prior-Labs/tabpfn_3_5/blob/main/LICENSE.