BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian mengenai prediksi kinerja akademik mahasiswa telah mengalami perkembangan pesat seiring dengan meluasnya implementasi teknik educational data mining (EDM) dan learning analytics di institusi pendidikan tinggi [2]. Di sektor pengembangan sistem deteksi dini, Arnold dan Pistilli mempelopori sistem Course Signals di Purdue University untuk mengelompokkan tingkat risiko mahasiswa berdasarkan data demografi, riwayat akademik, dan log aktivitas interaksi pada sistem manajemen pembelajaran [1]. Langkah pengkondisian intervensi institusional ini diperkuat oleh proyek Open Academic Analytics Initiative (OAAI) oleh Jayaprakash et al., yang menunjukkan bahwa model prediktif risiko kegagalan akademik dapat ditransfer antar-institusi secara andal untuk menyokong program intervensi tepat waktu [5].

Dari sudut pandang metodologi pengujian, Shahiri et al. melakukan tinjauan sistematis terhadap teknik penggalian data pendidikan dan menegaskan bahwa penentuan atribut prediktor yang paling berpengaruh merupakan salah satu objektif utama dalam riset performa akademik [13]. Relevansi manajemen data praproses sebelum model dievaluasi secara eksternal ditekankan oleh Costa et al., yang membuktikan bahwa keberhasilan dan stabilitas generalisasi prediksi sangat sensitif terhadap kualitas manipulasi fitur awal [14]. Keselarasan alur eksperimen dari seleksi fitur hingga tahap evaluasi akhir kemudian dibakukan dalam tinjauan komprehensif praktik terbaik oleh Alyahyan dan Düştegör [2].

Pada perkembangan terkini, riset mulai bergeser ke arah pemodelan algoritmik yang lebih kompleks. Meshari dan Nasir mengeksplorasi arsitektur *deep learning* yang dioptimasi secara sistematis untuk

mendongkrak akurasi mentah prediksi keberhasilan mahasiswa [18]. Namun, tingginya kompleksitas model algoritmik memicu tantangan baru terkait ketahanan performa. Brodersen et al. mengidentifikasi bahwa evaluasi model yang hanya bergantung pada akurasi konvensional (*point estimate*) berisiko tinggi menghasilkan kesimpulan yang menyesatkan (*optimistic bias*) apabila dataset yang diuji mengalami ketidakseimbangan kelas (*class imbalance*) [19]. Oleh karena itu, diperlukan metrik evaluasi alternatif seperti *Matthews Correlation Coefficient* (MCC) untuk menjamin perbandingan performa yang adil (*fair comparison*) [13], [14].

Lebih lanjut, pengujian keandalan visualisasi interpretasi dan ketahanan penjelasan model prediktif yang kompleks menuntut adanya pengujian metrik formal yang mencakup ketahanan (*robustness*) serta stabilitas [6]. Validasi model juga tidak boleh terbatas pada pengujian internal acak, melainkan harus lolos uji ketahanan berbasis waktu (*temporal generalizability*) melalui skema *temporal validation* untuk mendeteksi penurunan performa model (*performance drift*) akibat pergeseran kovariat dari tahun ke tahun [11], [12].

Ringkasan penelitian terdahulu serta posisinya terhadap penelitian ini disajikan pada Tabel 2.1.

Tabel 2.1 Tabel Data

Peneliti (Tahun)	**Fokus dan Metode**	**Temuan Utama**	**Relevansi terhadap Penelitian Ini**
Brodersen et al. (2010) [19]	Klasifikasi biner dan estimasi distribusi posterior *balanced accuracy*.	Menemukan bias optimis pada akurasi konvensional saat menangani data tidak seimbang (*imbalanced*).	Mendasari urgensi implementasi metrik MCC dan *Balanced Accuracy* untuk menjamin evaluasi berimbang.

Meshari & Nasir (2025) [18]	Prediksi keberhasilan akademik mahasiswa menggunakan *Deep Learning* teroptimasi.	Pembuktian bahwa proses penyetelan hiperparameter yang ketat meningkatkan akurasi komputasi model secara signifikan.	Menjustifikasi implementasi modul pencarian parameter (*hyperparameter search*) pada arsitektur FNN dan LSTM.
Arnold & Pistilli (2012) [1]	Purwarupa sistem *Course Signals* berbasis interaksi LMS, nilai, dan demografi.	Kategorisasi tingkat risiko dini secara proaktif terbukti menurunkan tingkat kegagalan studi mahasiswa.	Mendasari perancangan Sistem Peringatan Dini (*Early Warning System*) berbasis klasifikasi risiko biner.
Jayaprakash et al. (2014) [5]	Model prediksi risiko akademik lintas institusi dalam proyek OAAI.	Membuktikan model prediktif dapat dipindahkan (*transferable*) antar-institusi secara valid untuk program intervensi.	Menjadi dasar pengujian validitas eksternal model saat ditransfer lintas lingkungan *cohort*.
Shahiri et al. (2015) [13]	Tinjauan sistematis (*systematic review*) pada penerapan EDM untuk performa siswa.	Menyimpulkan bahwa penemuan variabel penentu utama adalah esensi krusial dari riset data pendidikan.	Mendasari analisis kontribusi fitur prediktor melalui audit fungsional penggerak eksternal.
Costa et al. (2017) [14]	Evaluasi efektivitas EDM untuk identifikasi dini potensi kegagalan mata kuliah dasar.	Membuktikan bahwa kualitas pra-pemrosesan data menentukan batas atas performa generalisasi model.	Menguatkan penekanan pada standardisasi *clean pipeline* praproses (imputasi, threshold, scaler).
Alyahyan & Düştegör (2020) [2]	Tinjauan literatur sistematis dan kompilasi praktik terbaik prediksi performa tinggi.	Merumuskan alur standar EDM dari seleksi atribut, konfigurasi arsitektur, hingga	Menjadi acuan baku struktural penulisan alur metodologi eksperimen

		metrik evaluasi akhir.	dual-pipeline komparatif.
Saarela dan Geogieva (2022) [6]	Kuantifikasi kualitas penjelasan model lewat metrik *robustness, stability,* dan *fidelity*.	Model kompleks (*black-box*) menuntut evaluasi metrik formal untuk menjamin konsistensi interpretasinya.	Mendasari pentingnya pengujian stabilitas interpretasi fitur (*Permutation Feature Importance*).
Al-azazi & Ghurab (2023) [14]	Model hibrida ANN-LSTM untuk prediksi performa harian (*day-wise*) pada lingkungan MOOC.	Integrasi komponen memori LSTM berhasil menangkap dependensi laten sekuensial interaksi siswa.	Mendasari implementasi arsitektur LSTM untuk mengekstrak data tabular melalui representasi sekuensial.
Arévalo-Cordovilla & Peña (2024) [12]	Analisis komparatif ML dalam memprediksi kesuksesan belajar berbasis data LMS dan faktor luar.	Model non-parametrik terbukti lebih unggul dalam menyerap pola interaksi variabel non-linear yang kompleks.	Menjustifikasi pemilihan algoritma *Ensemble* dan *Deep Learning* sebagai objek komparasi.
Alghamdi et al. (2022) [20]	Studi *benchmark* klasifikasi teks berita menggunakan algoritma ML klasik dan Transformer.	Tidak ada teknik tunggal yang unggul di seluruh dataset; efektivitas model sangat sensitif pada karakteristik data.	Menjustifikasi urgensi studi komparatif multi-arsitektur untuk mendeteksi batas ketahanan algoritma.
Suaza-Medina et al. (2024) [3]	Prediksi performa akademik wilayah tertinggal menggunakan 9 model ML dan analisis SHAP.	Faktor sosial-ekonomi, wilayah geografis, dan usia bertindak sebagai prediktor dengan pengaruh paling dominan.	Mendasari isolasi dan pemetaan kekuatan variabel sosiokultural non-kognitif sebagai fitur prediktor.
Adefemi & Mutanga (2025) [5]	Model hibrida CNN-LSTM untuk mereduksi *noise* pada	Kombinasi abstraksi fitur spasial (CNN) dan rangkaian	Menjustifikasi perombakan arsitektur dari FNN tabular

	data perilaku pembelajaran siswa.	temporal (LSTM) menembus batas akurasi model tunggal.	spasial menuju rantai sekuensial LSTM.
Islam et al. (2022) [21]	Eksperimen komparatif 11 algoritma ML untuk prediksi klinis dengan teknik SMOTE.	Algoritma ANN mencatatkan performa terbaik (F1-score 0,957) saat dikombinasikan dengan penanganan *imbalance*.	Menginspirasi penanganan data asimetris, namun diganti dengan *Pruned Binary* guna menjaga keaslian data.
Jiang et al. (2023) [22]	Pengembangan metode SSMCCA untuk analisis integrasi data lintas *cohort* independen.	Perbaikan ortogonalitas variabel berhasil meningkatkan kemampuan transfer (*transferability*) antar-kohort.	Mengonfirmasi perlunya validasi ketat lintas *cohort* angkatan untuk menguji kekuatan generalisasi model.
Li et al. (2024) [10]	Kerangka kerja *Type-Driven Lazy Drift Adaptor* untuk mengatasi pergeseran konsep (*concept drift*).	Identifikasi tipe pergeseran (*sudden, gradual*) secara daring (*online*) menaikkan efisiensi relearning model.	Mendasari analisis visualisasi *performance drift* akibat pergeseran data mahasiswa antar-generasi.
Hartmann et al. (2024) [11]	Pengembangan dan validasi temporal model prediksi risiko klinis pada kohort jangka panjang.	Model yang sangat akurat pada data internal mengalami degradasi kalibrasi sebesar 20-35% pada data masa depan.	Menjadi acuan kritis untuk mengukur tingkat degradasi akurasi global seiring berjalannya waktu.
Deliosa et al. (2022) [17]	Pengujian sistematis tingkat generalisabilitas temuan dari data arsip ke konteks baru.	Tingkat reproduksibilitas murni hanya mencapai 45%, menegaskan pentingnya transparansi spesifikasi analisis.	Menjustifikasi penerapan kerangka kerja *reproducible* melalui enkapsulasi arsitektur *Clean Pipeline*.

Singh et al. (2022) [7]	Evaluasi validitas eksternal dan metrik keadilan (*fairness*) model prediksi multi-center.	Model mengalami kejatuhan akurasi dan bias keadilan saat ditransfer antar-lokasi tanpa penyesuaian domain.	Mendasari integrasi metrik keadilan evaluasi melalui penyeimbangan kuadran sensitivitas (*recall*).
Melek & Melek (2025) [13]	Pendekatan geometris metrik *Golden Distance* untuk evaluasi klasifikasi tidak seimbang.	Metrik ringkas mampu merangkum multi-dimensi performa secara adil tanpa memanipulasi kelas data minoritas.	Memperkuat landasan teoretis untuk menolak metrik akurasi global tunggal pada distribusi data asimetris.
Bijalwan et al. (2021) [23]	Pengenalan pola aktivitas rehabilitasi pasca-cedera menggunakan model hibrida CNN-LSTM.	extended Kalman Filter efektif mereduksi oklusi dan derau (*noise*) pada aliran data sensor yang bising.	Menginspirasi perlunya eliminasi fitur konstan menggunakan filter *Variance Threshold* pada data kuesioner.
Yang et al. (2022) [8]	Studi generalisabilitas temporal model ML EHR pada empat wilayah klaster kesehatan independen.	Teknik *transfer learning* dan kustomisasi spesifik lokasi terbukti mengungguli model siap pakai (*ready-made*).	Memberikan justifikasi teoritis mengenai perlunya kalibrasi ulang keputusan ambang batas (*threshold*).
Arya et al. (2023) [24]	Peningkatan stabilitas model prognosis melalui autoencoder variasional multi-modal.	Penggabungan multi-modalitas data secara konsisten mampu mereduksi masalah dimensi tinggi (*curse of dimensionality*).	Mendasari teknik seleksi fitur *Mutual Information* untuk membatasi ruang input menjadi top-40 fitur.
Li et al. (2022) [25]	Penemuan proteomic signature (POSREG) melalui optimasi simultan stabilitas dan akurasi.	Pembobotan konsistensi konsisten menghasilkan tanda fitur yang	Mendasari visualisasi *Rank Migration* untuk melacak konsistensi urutan

		stabil dan *reproducible* lintas pengujian.	peringkat fitur terpenting.
Kalita et al. (2025) [26]	Kerangka kerja Bi-LSTM untuk prediksi performa akademik dengan interpretabilitas nilai SHAP.	Model mampu menangkap dependensi dua arah secara saksama dan memecahkan batasan kotak hitam (*black-box*).	Menjustifikasi penggunaan lapisan *Bidirectional* pada LSTM dan integrasi *Permutation Importance*.
Airlangga (2024) [27]	Studi komparatif arsitektur MLP, CNN, BiLSTM, dan LSTM dengan mekanisme Attention.	Model spasial mencatatkan akurasi tinggi pada data statis, sedangkan varian LSTM menunjukkan fluktuasi metrik.	Mendasari pengujian kestabilan hiperparameter lintas 6 variasi FNN dan 4 variasi sekuensial LSTM.
Adewale et al. (2018) [28]	Pemodelan hubungan non-linear variabel psikologis terhadap prestasi menggunakan ANN.	Pendekatan jaringan saraf tiruan terbukti melampaui limitasi asumsi linier dari statistik parametrik klasik.	Mendasari justifikasi teoretis transisi dari regresi linier menuju ruang diskrit klasifikasi biner.
Cai et al. (2024) [29]	Model peringatan dini ketidakstabilan lereng berbasis algoritma CatBoost teroptimasi.	Penggunaan *Grid Search Cross-Validation* (GSCV) mengunci stabilitas model dari ancaman *overfitting*.	Mendasari kerangka validasi silang berulang berlapis (*Repeated Stratified K-Fold 5x3*) sebagai *engine* utama.
Ariyanto & Wulandhari (2024) [4]	Seleksi subtes seleksi masuk paling berpengaruh terhadap prestasi menggunakan Random Forest.	Penggunaan metode seleksi fitur berbasis *impurity* sukses mengefisiensikan beban instrumen pengujian.	Menjustifikasi penempatan algoritma *Random Forest* (Mahasiswa 1) sebagai basis *baseline comparator*.
Boujmiraz et al. (2026) [30]	Tinjauan komprehensif implementasi Explainable AI (XAI)	Transparansi logika berpikir mesin merupakan	Mendasari penyusunan visualisasi

	pada sistem prediksi pendidikan tinggi.	prerequisite absolut sebelum model diintegrasikan ke institusi.	analisis risiko prediktif melalui *Feature Governance Quadrant*.

Berdasarkan tinjauan komprehensif terhadap 30 literatur ilmiah tersebut, teridentifikasi sebuah kesenjangan metodologis yang nyata: mayoritas penelitian terdahulu masih menggabungkan seluruh faktor internal-eksternal secara acak dan menghentikan pengujiannya pada batas evaluasi internal snapshot (satu waktu) [2], [13]. Penelitian ini menempatkan diri secara tegas pada celah metodologis tersebut dengan mengusung tiga pilar kontribusi ilmiah utama:

1. **Pemisahan dan Isolasi Faktor Eksternal:** Mengevaluasi ketahanan model murni berdasarkan variabel lingkungan non-kognitif di luar pencapaian akademik formal (seperti log LMS atau riwayat nilai kuis) untuk menakar kekuatan prediktif riil dari faktor luar kampus [3], [19].
2. **Standardisasi Evaluasi Berbasis Keadilan (*Fair Comparison*):** Mengimplementasikan metrik korelasi multi-dimensi Matthews Correlation Coefficient (MCC) dan Balanced Accuracy untuk melengkapi sekaligus mengaudit evaluasi akurasi konvensional dari Mahasiswa 1 dan Mahasiswa 2 yang rentan terdistorsi oleh ketidakseimbangan kelas (class imbalance) natural [13], [19].
3. **Pengujian Stabilitas Dinamis Lintas Batas:** Menyematkan lapisan validasi eksternal berupa validasi temporal sekuensial lintas cohort (antar-angkatan) serta memetakan karakteristik degradasi performa model (performance drift) guna mengukur masa pakai efektif model (model lifespan) sebelum mengalami kegagalan prediksi di dunia nyata [6], [7], [10].

2.2 Teori yang berkaitan

2.2.1 Kinerja Akademik Mahasiswa dan Faktor Eksternal

Kinerja akademik mahasiswa dalam penelitian ini direpresentasikan secara operasional melalui indeks prestasi kumulatif kategorikal (*Clean-Margin GPA Category*) [19]. Pemahaman konseptual bahwa prestasi akademik tidak semata-mata ditentukan oleh variabel kognitif internal individu, melainkan dipengaruhi secara masif oleh dinamika lingkungan luar kampus, berakar kuat pada Teori Sistem

Ekologis dari Bronfenbrenner [1]. Perspektif ini memandang perkembangan manusia sebagai produk dari interaksi berkelanjutan antara faktor personal dan berlapis-lapis dimensi lingkungan, yang melingkupi *microsystem* (keluarga dan teman sebaya), *mesosystem* (relasi antar-lingkungan terdekat), *exosystem* (kondisi yang tidak bersinggungan langsung namun berdampak, seperti pendapatan orang tua), *macrosystem* (budaya dan status sosial-ekonomi), hingga *chronosystem* (dimensi perubahan waktu) [1]. Kerangka ini disempurnakan menjadi model bioekologis yang menempatkan peran proses, person, konteks, dan waktu (*Process-Person-Context-Time*/PPCT) sebagai fondasi utama untuk menilai seberapa kuat faktor eksternal dapat bertindak sebagai prediktor yang andal seiring berjalannya waktu [1].

Secara empiris, dukungan terhadap pengaruh faktor sosio-demografis luar kampus ini diperkuat oleh konsep modal sosial dari Coleman, yang menjelaskan bagaimana ekosistem keluarga menyokong pembentukan modal manusia (*human capital*) dan capaian pendidikan [1], [2]. Sirin melalui studi meta-analisisnya membuktikan adanya hubungan positif yang kuat dan konsisten antara status sosial-ekonomi dengan capaian akademik mahasiswa di berbagai klaster institusi [3]. Pengaruh pendapatan dan latar belakang pendidikan orang tua juga dibuktikan oleh Davis-Kean beroperasi secara tidak langsung melalui transmisi keyakinan serta pola perilaku di dalam rumah [3], [4]. Lebih lanjut, adanya "krisis generalisasi" (*generalizability crisis*) dan batasan dari pembuatan stimulus eksperimen faktor eksternal non-kognitif ini menuntut pengujian ketahanan yang ketat lintas *cohort* [7], [9]. Hal ini diperlukan agar model tidak sekadar menghafal pola data statis secara bias (*optimistic bias*) [11], mengingat motivasi, gaya hidup, dan lingkungan belajar mahasiswa bersifat adaptif serta dinamis dari tahun ke tahun [6], [10].

2.2.2 *Educational Data Mining* dan *Learning Analytics*

Educational Data Mining (EDM) merupakan domain interdisipliner yang fokus pada pengembangan metode komputasi untuk mengeksplorasi data berskala besar dari lingkungan pendidikan [1], [29]. Baker dan Inventado memosisikan EDM sebagai rumpun riset yang beririsan erat dengan *learning analytics*, namun dengan penekanan teknis yang lebih spesifik pada penemuan pola tersembunyi (*pattern discovery*) dan pemodelan prediktif tingkat tinggi [29]. Romero dan Ventura dalam survei komprehensifnya

menegaskan bahwa evolusi metode EDM dirancang untuk mendukung pengambilan keputusan akademis yang berbasis pada bukti data nyata (*data-driven decision making*) di institusi pendidikan tinggi [28], [29].

Di sisi lain, *learning analytics* lebih menitikberatkan pada pengukuran, pengumpulan, analisis, dan pelaporan karakteristik pembelajar beserta konteks situasionalnya [1]. Long dan Siemens mengingatkan bahwa potensi analitika pembelajaran akan mencapai efisiensi tertinggi ketika mampu menjangkau variabel yang lebih luas di luar batasan sistem manajemen pembelajaran internal kampus (*Learning Management System* / LMS) [1]. Meskipun demikian, Viberg et al. menemukan kesenjangan di mana mayoritas penerapan analitika di perguruan tinggi masih tertahan pada level deskriptif (menjelaskan apa yang telah terjadi) dan belum sepenuhnya membuktikan dampak prediktif yang kokoh dalam meningkatkan hasil belajar secara konsisten [1]. Penelitian ini menempatkan diri pada koridor prediktif tersebut melalui evaluasi komparatif yang ketat untuk mengukur daya generalisasi dan stabilitas algoritma ketika menghadapi pergeseran waktu (*temporal drift*) lintas angkatan [8], [9], [14].

2.2.3 Kualitas Data dan Penanganan Data Hilang

Kualitas data merupakan determinan fundamental yang menentukan validitas dan reliabilitas akhir dari suatu model prediksi [14]. Wang dan Strong merumuskan kualitas data sebagai konstruksi multidimensi yang mencakup aspek akurasi, kelengkapan (*completeness*), relevansi, dan ketepatan waktu (*timeliness*) dari kacamata konsumen data [12], [23]. Konsep pengukuran kualitas data ini kemudian diperluas oleh Pipino et al. melalui integrasi penilaian objektif dan subjektif yang fungsional pada data tabular [12]. Teori kualitas data ini melandasi pengkondisian evaluasi dalam riset ini, khususnya dalam memantau tingkat kesenyapan (*sparsity ratio*) fitur eksternal mahasiswa yang bersumber dari kuesioner mandiri (*self-reported data*) dan biodata [3], [13].

Dalam mengelola ketidaksempurnaan data, Little dan Rubin mengklasifikasikan mekanisme data hilang (*missing data*) ke dalam tiga kategori utama: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR), dan *Missing Not at Random* (MNAR) [14], [15]. Mereka menegaskan bahwa bias dan stabilitas inferensi model prediktif sangat dipengaruhi oleh ketepatan pemilihan teknik imputasi data hilang [14], [15]. Kesalahan dalam mengidentifikasi

mekanisme ini dapat mengakibatkan penurunan drastis pada kemampuan generalisasi model (*generalizability crisis*) saat diuji secara eksternal pada populasi baru atau *cohort* angkatan yang berbeda [14], [16].

2.2.4 Kalibrasi Probabilitas dan Reliabilitas Prediksi

Evaluasi terhadap performa arsitektur *machine learning* tidak boleh berhenti pada akurasi klasifikasi biner semata, melainkan wajib menguji keandalan nilai probabilitas yang dihasilkan [14]. Kalibrasi merujuk pada tingkat kesesuaian sistematis antara estimasi probabilitas yang diprediksi model dengan frekuensi kejadian aktual di dunia nyata [14], [15]. Brier memperkenalkan metrik verifikasi akurasi probabilistik berbasis skor galat kuadrat (*Brier Score*) yang menjadi standar evaluasi presisi nilai probabilitas [14], [15]. Niculescu-Mizil dan Caruana membuktikan bahwa algoritma yang berbeda menghasilkan karakteristik probabilitas dengan distorsi kalibrasi yang unik, sehingga memerlukan intervensi teknik kalibrasi tambahan (seperti *Platt Scaling* atau *Isotonic Regression*) untuk menyelaraskan nilai prediksi [14], [15].

Dalam aplikasinya pada model risiko, Van Calster et al. mendefinisikan hierarki kalibrasi dari yang paling longgar (*calibration-in-the-large*) hingga tingkat kalibrasi individu yang sangat ketat [14]. Pengabaian terhadap aspek kalibrasi dalam analitika prediktif berisiko tinggi memicu keputusan intervensi yang menyesatkan akibat kesalahan estimasi tingkat risiko objek, yang pada akhirnya dapat merugikan alokasi sumber daya institusi [14], [15]. Teori ini menjadi dasar bagi tugas Mahasiswa 3 untuk menguji apakah model *ensemble* (*Optimized Random Forest*) dan *deep learning* teroptimasi (*Feedforward Neural Network* dan *Bidirectional* LSTM) yang dihasilkan telah memiliki nilai probabilitas yang terkalibrasi secara andal seiring berjalannya waktu sebelum direkomendasikan kepada pihak universitas [8], [9], [14].

2.2.5 Metrik Evaluasi Klasifikasi yang Adil (*Fair Comparison*)

Ketika mengevaluasi kinerja model klasifikasi, pemilihan metrik harus disesuaikan secara cermat dengan karakteristik distribusi data. Evaluasi berbasis akurasi konvensional (*point estimate*) rentan memicu bias optimis (*optimistic bias*) yang menyesatkan apabila diimplementasikan pada dataset yang mengalami ketidakseimbangan kelas (*class imbalance*) [11]. Untuk menjamin perbandingan performa yang adil (*fair comparison*), Chicco dan Jurman

membuktikan bahwa *Matthews Correlation Coefficient* (MCC) merupakan metrik yang paling informatif, jujur, dan berimbang untuk klasifikasi biner karena mengintegrasikan seluruh proporsi dari keempat kuadran matriks konfusi (*True Positive, True Negative, False Positive, False Negative*) secara simetris [5], [6].

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Di samping itu, Brodersen et al. menawarkan metrik *Balanced Accuracy* sebagai alternatif evaluasi yang adil pada kondisi data timpang dengan menghitung nilai rata-rata dari sensitivitas dan spesifisitas beserta sebaran distribusi posteriornya [11]. Guna menghasilkan analisis komparatif yang komprehensif dan mendalam, penggunaan metrik korelasi berimbang ini disandingkan secara ketat dengan parameter pelengkap komputasi seperti *precision, recall, F1-score macro*, serta analisis kurva *Area Under the ROC Curve (ROC-AUC)* dan *Area Under the Precision-Recall Curve (PR-AUC)* [5], [21].

2.2.6 Karakteristik *Performance Drift* dan Validasi Temporal

Validasi internal menggunakan metode validasi silang acak standar (*random cross-validation*) terbukti tidak mencukupi untuk menjamin keandalan model di dunia nyata karena cenderung mengabaikan dimensi kronologis waktu [9], [14]. Perubahan karakteristik populasi mahasiswa antar-generasi memicu terjadinya pergeseran kovariat (*covariate shift atau concept drift*) yang berujung pada penurunan akurasi model secara bertahap seiring berjalannya waktu (*performance drift*) [7], [14]. Karakteristik kelemahan ketahanan model ini dikonseptualisasikan oleh Yarkoni sebagai bagian dari krisis generalisasi (*generalizability crisis*) ilmiah yang masif, di mana model yang tampak superior di dalam laboratorium mengalami degradasi performa yang parah saat diuji secara eksternal pada lingkungan baru [16].

Untuk mendeteksi dan mengukur stabilitas temporal tersebut, metodologi pengujian harus mengadopsi skema *temporal validation* (validasi sekuensial berbasis waktu) lintas *cohort* (antar-angkatan) [8], [14]. Eksperimen ini mensimulasikan penerapan model secara prospektif dengan membekukan model pada data masa lalu (*frozen*

model) dan mengujinya pada data angkatan baru [8], [31]. Analisis ketahanan ini diperkuat oleh implementasi *Statistical Process Control* (SPC) menggunakan grafik kendali untuk memantau pergerakan stabilitas nilai AUC dan *Brier Score* dari waktu ke waktu secara statistik, sehingga batas usia efektif model (*model lifespan*) sebelum memerlukan kalibrasi ulang (model refitting) dapat diidentifikasi secara presisi [9], [31].

2.2.7 Etika dan Tata Kelola Kecerdasan Buatan

Penerapan sistem kecerdasan buatan prediktif dalam ranah pendidikan membawa konsekuensi etis dan risiko bias algoritmik yang tinggi jika luaran model klasifikasinya bersifat tidak adil [14], [20]. Bias algoritmik dalam komputasi pendidikan sering kali berakar dari ketidakseimbangan representasi data historis, yang berpotensi memicu diskriminasi sistematis dan memperlebar ketimpangan deteksi (*recall disparity*) terhadap kelompok minoritas atau mahasiswa kategori berisiko tinggi [20]. Keadilan algoritmik (*algorithmic fairness*) harus dinilai secara komprehensif untuk memastikan bahwa probabilitas risiko yang dihasilkan model tidak mengeksklusi mahasiswa yang murni membutuhkan intervensi akademik [20], [31].

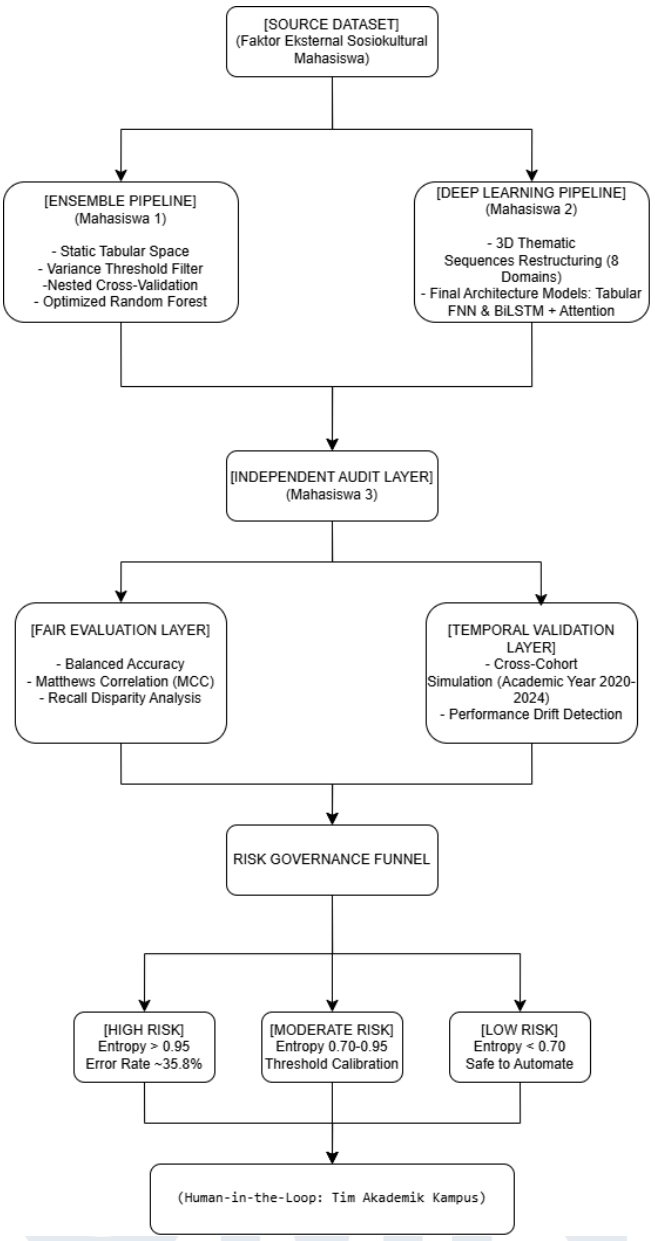
Pada tatanan tata kelola, pedoman etika untuk kecerdasan buatan yang terpercaya (*Trustworthy AI*) bertumpu secara mutlak pada tiga pilar utama, yaitu prinsip transparansi, akuntabilitas, dan keadilan evaluasi [14], [31]. Pendekatan pengelolaan risiko ini dipertegas oleh kerangka kerja manajemen risiko internasional (*Risk Management Framework/RMF*) untuk mengidentifikasi, memantau, dan memitigasi dampak negatif kegagalan prediksi AI [14]. Dari aspek manajemen operasional model, panduan risiko menegaskan bahwa aktivitas validasi eksternal secara independen, pemantauan stabilitas probabilitas berkelanjutan (*Reliability Error*), serta visualisasi pemetaan ketidakpastian (*Risk Governance Funnel*) merupakan pilar utama untuk meminimalkan risiko kegagalan sistem prediktif [14], [28]. Prinsip tata kelola risiko independen inilah yang diwujudkan melalui pengujian batas-batas generalisasi performa lintas pipeline dalam penelitian ini [14], [16].

2.3 Framework/Algoritma yang digunakan

2.3.1 Kerangka Kerja Algoritma Klasifikasi (*Ensemble vs Deep Learning*)

Dalam melakukan evaluasi prediktif yang adil (*fair comparison*), penelitian ini mengomparasikan dua rumpun besar pemodelan algoritmik melalui kerangka kerja terpadu [14]. Rumpun pertama adalah metode *ensemble learning* berbasis pohon keputusan (*decision trees*), yaitu *Random Forest*, CatBoost, dan XGBoost [3], [4], [27]. *Random Forest*, yang dikembangkan oleh Breiman, menggabungkan banyak versi pohon prediktor yang dibangun secara independen melalui teknik *bootstrap aggregating* (*baging*) untuk menekan varians prediksi tanpa mendistorsi bias [30].

Guna memperluas dimensi komparasi, algoritma berbasis *gradient boosting* seperti XGBoost dan CatBoost dilibatkan karena kemampuannya dalam meminimalkan fungsi kerugian melalui optimasi gradien sekuensial yang efisien [27], [29]. Pemilihan XGBoost sangat relevan karena kemampuannya dalam menangani interaksi nonlinear yang rumit antar-fitur eksternal secara tangguh [3], [27]. Integrasi dinamis dari keseluruhan alur pemrosesan ini, mulai dari pemisahan jalur data spasial-tabular hingga restrukturisasi rangkaian temporal semu, divisualisasikan secara komprehensif pada Gambar 2.1.



Gambar 2.1 Skema Alur Kerja Eksperimen Komparatif dan Tata Kelola Risiko

Rumpun kedua yang dijadikan objek komparasi adalah pendekatan *deep learning* yang dikembangkan oleh Mahasiswa 2, mencakup *Feedforward Neural Network* (FNN) dan *Bidirectional Long Short-Term Memory* (BiLSTM) [4], [26]. Lapisan memori BiLSTM yang diperkuat oleh *Attention Mechanism* digunakan secara khusus untuk menilai apakah terdapat pola sekuensial kronologis atau ketergantungan urutan (*thematic sequences*) pada 8 domain kehidupan variabel faktor eksternal mahasiswa dari satu fase ke fase berikutnya [4], [24].

Menurut Dietterich, pengujian komparatif antara arsitektur *ensemble* dan *deep learning* ini krusial untuk mengidentifikasi *trade-off* terbaik dari sisi statistik, efisiensi komputasi, stabilitas fitur (*Rank Migration*), serta keterwakilan representasi data (*representational power*) dalam domain pendidikan tinggi [1], [13].

2.3.2 Strategi Optimasi Hyperparameter dan Desain Penyekatan Data

Konfigurasi hiperparameter internal (seperti jumlah estimator, kedalaman maksimum pohon, tingkat regulasi L_2, unit *dense*, rasio *dropout*, dan *learning rate*) sangat menentukan batas atas performa akhir dari masing-masing model [2], [25]. Penyetelan parameter secara algoritmik terbukti mampu mendongkrak ketahanan model secara substansial dibandingkan penggunaan konfigurasi bawaan standar (*default settings*) [18], [27].

Untuk mencapai efisiensi tersebut, proses pencarian parameter mengombinasikan metode acak (*Random Search*) serta optimasi berbasis Bayesian (*Bayesian Optimization*) menggunakan Optuna [2, 18]. Bergstra dan Bengio membuktikan bahwa pengalokasian ruang pencarian secara acak dan probabilistik jauh lebih efektif dibandingkan pencarian menyeluruh (*Grid Search*), karena mampu menghindari pemborosan daya komputasi pada hiperparameter yang kurang sensitif terhadap fungsi loss [18], [27].

Guna menjamin bahwa metrik akurasi akhir yang dilaporkan bebas dari kontaminasi kebocoran data (*data leakage*) akibat proses optimasi, eksperimen ini menerapkan skema penyekatan data yang ketat melalui *Nested Cross-Validation* [19], [27]. Kuhn dan Johnson menjelaskan bahwa pemisahan berlapis antara proses pemilihan parameter terbaik (pada *inner loop* menggunakan 3-Fold CV) dan proses estimasi ketahanan performa model (pada *outer loop* menggunakan *Repeated Stratified K-Fold* 5x3) sangat efektif untuk mengeliminasi risiko penilaian performa yang terlalu optimistis (*optimistic bias*) [11], [27]. Desain pengujian berlapis inilah yang mempersiapkan model sebelum diserahkan kepada Mahasiswa 3 untuk menjalani prosedur validasi temporal eksternal lintas *cohort* angkatan yang jauh lebih ketat [8], [9], [14].

2.4 Tools/software yang digunakan

Seluruh rangkaian eksperimen komparatif, standardisasi metrik evaluasi, validasi sekuensial temporal, hingga visualisasi degradasi model diimplementasikan menggunakan bahasa pemrograman Python pada lingkungan pengembangan terintegrasi *Jupyter Notebook*. Pemilihan ekosistem ini menjamin transparansi kode agar seluruh alur pengujian bersifat dapat direproduksi secara mandiri (*reproducible*) oleh peneliti lain [13], [19].

Pustaka-pustaka utama yang digunakan disajikan pada Tabel 2.2.

Tabel 2.2 Perangkat Lunak dan Pustaka Pendukung Validasi Performa

Perangkat / Pustaka	**Fungsi Spesifik dalam Penelitian**
Draw.io	Pembuatan diagram alur penelitian, skema arsitektur komparatif, dan visualisasi pemodelan [13].
Python	Bahasa pemrograman utama untuk eksekusi praproses, pelatihan model, dan komputasi metrik evaluasi [13], [19].
Jupyter Notebook	Lingkungan komputasi interaktif untuk dokumentasi eksperimen, audit kualitas data, dan pengujian model [13].
pandas	Manipulasi data tabular eksternal, integrasi lintas file csv sekuensial, dan transformasi fitur [13].
NumPy	Operasi array multidimensi, perhitungan matematis matriks konfusi, dan manipulasi aljabar linier [13].
SciPy	Operasi komputasi ilmiah dan eksekusi fungsi statistik penunjang kalibrasi probabilitas [13].
scikit-learn	Pemodelan *Random Forest*, pembagian lipatan *Nested CV*, kalibrasi probabilitas, dan ekstraksi metrik utama [11], [13].

XGBoost	Pustaka khusus untuk kompilasi dan optimasi model *ensemble gradient boosting* [4], [27].
Keras / TensorFlow	Penyusunan lapisan arsitektur *deep learning* (FNN dan LSTM) beserta regularisasinya [13], [24].
Optuna	Kerangka kerja otomatisasi pencarian *hyperparameter* berbasis *Bayesian Optimization* [13], [18].
Alibi-Explain	Pustaka untuk mendukung audit visualisasi interpretasi dan keandalan peta fitur kepentingan model [15], [31].
Matplotlib	Visualisasi kurva performa dasar, diagram batang perbandingan, dan histogram distribusi [13].
Seaborn	Pembuatan grafik statistik, matriks korelasi, kualifikasi *Brier Score*, dan plot kalibrasi [13].
adjustText	Penyesuaian otomatis tata letak teks label pada visualisasi grafik performa seiring waktu [13].

Implementasi ekosistem terbuka (*open-source*) ini didasarkan pada tingkat kematangan pustaka, kelengkapan algoritma evaluasi, serta dukungannya yang luas terhadap alur kerja metodologi validasi model di domain sains data medis maupun sosial [14], [15].