BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Prediksi kinerja akademik mahasiswa menjadi topik yang semakin penting seiring meningkatnya kebutuhan institusi pendidikan untuk melakukan intervensi dini terhadap potensi penurunan prestasi belajar. Berdasarkan berbagai penelitian terdahulu, penerapan machine learning dalam memprediksi performa akademik mahasiswa menunjukkan peningkatan signifikan, dengan algoritma yang umum digunakan meliputi Random Forest, model boosting seperti LightGBM, serta jaringan saraf Multi-Layer Perceptron (MLP), yang mampu menangkap pola non-linear pada data akademik dan menghasilkan tingkat akurasi yang tinggi, umumnya di atas 80% dalam berbagai studi dan konteks [12]. Random Forest dikenal memiliki stabilitas dan kemampuan generalisasi yang baik, LightGBM unggul dalam efisiensi komputasi dan akurasi pada data berdimensi tinggi, sementara MLP efektif dalam memodelkan hubungan kompleks antar fitur akademik. Namun demikian, sebagian besar penelitian tersebut masih berfokus pada performa prediksi dan cenderung bersifat *black-box*, sehingga belum memberikan pemahaman yang memadai mengenai faktor-faktor yang memengaruhi hasil prediksi. Untuk mengatasi keterbatasan tersebut, pendekatan Explainable Artificial Intelligence (XAI), khususnya SHapley Additive exPlanations (SHAP), mulai banyak diterapkan guna meningkatkan transparansi dan kepercayaan terhadap model prediksi [13]. Berikut disajikan Tabel 2.1 yang merangkum beberapa penelitian terdahulu terkait pengembangan model prediksi kinerja akademik mahasiswa serta penerapan *Explainable Machine Learning* sebagai dasar pemilihan metode pada penelitian ini.

Tabel 2. 1 Penelitian terdahulu berdasarkan metode yang digunakan

No	Tahun	Masalah	Model	Hasil	*Main Findings*
1	2024	Prediksi kelulusan tepat	Random Forest	Akurasi 88%, precision 81%,	Random Forest efektif memprediksi kelulusan

		waktu mahasiswa untuk mendukung efektivitas pembelajaran dan akreditasi institusi		recall 97%, specificity 80%	dan mengidentifikasi faktor signifikan berdasarkan *variable importance*; model diimplementasikan dalam aplikasi web berbasis Streamlit. [14]
2	2022	Optimasi pemilihan fitur untuk meningkatkan akurasi prediksi performa akademik mahasiswa berbasis LMS	Random Forest + Correlation -Based Feature Selection (CFS)	Akurasi meningkat dari 91.66% menjadi 97.22% setelah optimasi fitur	Fitur paling berpengaruh adalah waktu belajar, penyelesaian tugas, dan partisipasi kuis; optimasi CFS terbukti meningkatkan performa model. [15]
3	2024	Prediksi ketepatan waktu kelulusan mahasiswa dengan data tidak seimbang	Random Forest + Random Oversampl ing (ROS)	Akurasi 90.04%, precision 87.05%, recall 90.04%	Teknik Random Oversampling efektif mengatasi ketidakseimbangan data dan meningkatkan performa model Random Forest dalam memprediksi kelulusan tepat waktu. [16]
4	2022	Prediksi kinerja akademik matematika berbasis profil siswa	Boosting Algorithms (LightGB M, AdaBoost, Gradient Boosting) dengan feature selection (Informati on Gain + Recursive Feature Eliminatio n)	Accuracy : 96–97% (LightGBM terbaik)	LightGBM mampu memberikan akurasi prediksi tertinggi ketika dikombinasikan dengan metode seleksi fitur Information Gain dan RFE, serta efektif dalam memodelkan kinerja akademik mahasiswa berbasis profil siswa. [17]
5	2024	Prediksi performa akademik ujian standar di wilayah tertinggal + interpretabilita s	LightGBM , XGBoost, Gradient Boosting, dan algoritma klasifikasi lainnya dengan	Accuracy : 88–92% (LightGBM & XGBoost tertinggi)	Pendekatan SHAP berhasil mengidentifikasi variabel paling berpengaruh seperti tingkat sosial ekonomi, gender, wilayah, usia, dan lokasi sekolah, sehingga meningkatkan transparansi model dan

			pendekatan SHAP		mendukung perumusan kebijakan pendidikan. [18]
6	2022	Prediksi kinerja akademik berbasis perilaku pembelajaran daring	CART, RF, XGBoost, LightGBM , Bagging, Stacking	CART: 68%, RF: 71%, XGBoost: 73% , LightGBM: 76%, Bagging: 83%, Stacking: 84%	Model Individual LightBGM memiliki akurasi terbaik, namun Model fusion (bagging & stacking) meningkatkan akurasi secara signifikan dibandingkan single model; stacking fusion RF–CART–XGBoost–LightGBM memberikan performa terbaik. [19]
7	2023	Prediksi tingkat kinerja akademik mahasiswa untuk mendukung pemantauan akademik dan perencanaan studi	Fuzzy C-Means Clustering, Multi-Layer Perceptron, Logistic Regression , Random Forest, Fuzzy C-Means–Multi-Layer Perceptron, Fuzzy C-Means–Logistic Regression , Fuzzy C-Means–Random Forest	Fuzzy C-Means: ≈78–80% · Multi-Layer Perceptron: ≈82–85% · Logistic Regression: ≈79–82% · Random Forest: ≈83–86% · Fuzzy C-Means–Multi-Layer Perceptron: ≈88–92% (tertinggi)	Integrasi Fuzzy C-Means dengan Multi-Layer Perceptron menghasilkan akurasi prediksi tertinggi dan menurunkan false alarm rate dibandingkan model tunggal. [20]
8	2021	Prediksi kinerja akademik siswa tingkat menengah untuk deteksi dini risiko kegagalan belajar	Multilayer Perceptron, J48 Decision Tree, PART Rule-Based Classifier, Bagging, MultiBoost , Voting, serta kombinasi	MultiBoost–Multilayer Perceptron: 98.7% (Precision: 98.6% · Recall: 98.6% · F1-score: 98.6%)	Fusion model berbasis MultiBoost dengan Multilayer Perceptron memberikan performa tertinggi dan secara signifikan mengungguli single classifier maupun ensemble tunggal [7]

			fusion single dan ensemble classifiers		
9	2025	Prediksi kinerja akademik mahasiswa untuk meningkatkan pengambilan keputusan institusi pendidikan dengan pendekatan explainable machine learning	K-Nearest Neighbors Regressor, Linear Regression, CatBoost Regressor, Extreme Gradient Boosting Regressor, AdaBoost Regressor, serta Ensemble Voting Regression	Dataset 1 (10.000 sampel): RMSE = 0.1050 · MAE = 0.0837 · R^2 = 0.9890 (Voting Regression terbaik); Dataset 2 (6.607 sampel): R^2 = 0.7716 (Voting Regression terbaik)	Ensemble Voting Regression menunjukkan performa paling robust dan konsisten pada dua dataset berbeda; integrasi SHAP dan LIME berhasil mengidentifikasi faktor akademik dan perilaku belajar sebagai penentu utama kinerja [21]
10	2024	Sulitnya memahami hubungan antar faktor yang memengaruhi keberhasilan akademik mahasiswa menggunakan model ML tradisional.	XGB-SHAP (gabungan XGBoost dan SHAP)	MAE ≈ 6 dan R^2 ≈ 0.82, lebih baik dari tiga model lain yang diuji.	Model memberikan interpretasi faktor penting yang memengaruhi prestasi, seperti mode pembelajaran dan kemampuan belajar mandiri (*self-directed learning skills*), serta menekankan pentingnya *customized feature selection* sesuai konteks pembelajaran. [22]

Berdasarkan Tabel 2.1, penelitian terkait prediksi kinerja akademik mahasiswa menunjukkan bahwa algoritma berbasis *ensemble tree* dan *neural network* merupakan pendekatan yang paling dominan digunakan, khususnya Random Forest, LightGBM, XGBoost, serta Multi-Layer Perceptron (MLP). Random Forest banyak diterapkan karena kemampuannya dalam menangani data berdimensi tinggi, data tidak seimbang, serta menghasilkan interpretasi melalui analisis *feature importance*. Hal ini terlihat pada beberapa studi yang menunjukkan bahwa Random Forest mampu memprediksi kelulusan tepat waktu mahasiswa dengan akurasi 88%, precision 81%, dan recall 97% [14]. Selain itu, optimasi fitur menggunakan

Correlation-Based Feature Selection (CFS) terbukti meningkatkan akurasi Random Forest secara signifikan dari 91.66% menjadi 97.22%, menegaskan pentingnya seleksi fitur dalam meningkatkan performa prediksi [15]. Pendekatan penanganan data tidak seimbang seperti Random Oversampling (ROS) juga dilaporkan mampu meningkatkan stabilitas dan kinerja model Random Forest dengan akurasi mencapai 90.04% [16].

Model boosting seperti LightGBM dan XGBoost menunjukkan performa unggul dalam memodelkan pola non-linear yang kompleks pada data pendidikan. Penelitian berbasis profil siswa melaporkan bahwa LightGBM mencapai akurasi tertinggi pada rentang 96–97% setelah dikombinasikan dengan Information Gain dan Recursive Feature Elimination [17]. Temuan serupa juga diperoleh pada studi prediksi performa akademik di wilayah tertinggal, di mana LightGBM dan XGBoost mencapai akurasi tertinggi pada kisaran 88–92%, sekaligus meningkatkan interpretabilitas model melalui pendekatan SHAP untuk mengidentifikasi faktor dominan seperti kondisi sosial ekonomi, gender, usia, dan lokasi sekolah [18]. Selain itu, pendekatan *ensemble fusion* seperti bagging dan stacking secara konsisten meningkatkan performa dibandingkan model individual, dengan akurasi meningkat dari 76% pada LightGBM tunggal menjadi 84% pada model stacking RF–CART–XGBoost–LightGBM [19].

Di sisi lain, pendekatan berbasis *hybrid learning* dan *neural network* juga menunjukkan hasil yang kompetitif. Integrasi Fuzzy C-Means dengan Multi-Layer Perceptron menghasilkan akurasi tertinggi pada kisaran 88–92% dibandingkan model tunggal lainnya, serta mampu menurunkan *false alarm rate* [20]. Pada tingkat pendidikan menengah, pendekatan fusion berbasis MultiBoost dan Multilayer Perceptron bahkan mencapai akurasi sangat tinggi sebesar 98.7%, mengungguli single classifier maupun ensemble tunggal [7]. Penelitian terkini juga menekankan pentingnya *explainable machine learning*, di mana model Ensemble Voting Regression yang dikombinasikan dengan SHAP dan LIME menunjukkan performa yang sangat robust pada dua dataset berbeda, dengan nilai R^2 mencapai

0.9890 pada dataset sederhana dan 0.7716 pada dataset yang lebih kompleks [21]. Lebih lanjut, pendekatan XGB-SHAP mampu menghasilkan prediksi yang tidak hanya akurat ($R^2 \approx 0.82$), tetapi juga memberikan interpretasi mendalam terhadap faktor-faktor penting seperti mode pembelajaran dan kemampuan belajar mandiri mahasiswa [22]. Secara keseluruhan, literatur menunjukkan bahwa model *ensemble* dan *hybrid* mampu meningkatkan akurasi prediksi kinerja akademik secara signifikan dibandingkan model tunggal. Meskipun demikian, model individual seperti LightGBM, Random Forest, dan Multi-Layer Perceptron (MLP) tetap banyak digunakan dalam penelitian terdahulu karena memiliki karakteristik yang saling melengkapi. LightGBM dikenal efisien dalam menangani data berukuran besar dan berdimensi tinggi, Random Forest unggul dalam stabilitas serta kemampuannya mengurangi *overfitting*, sementara MLP efektif dalam memodelkan hubungan non-linear yang kompleks. Namun, sebagian besar model tersebut masih memiliki keterbatasan dari sisi interpretabilitas, sehingga integrasi pendekatan *explainable machine learning* seperti SHAP menjadi aspek penting untuk meningkatkan transparansi dan pemahaman terhadap faktor-faktor yang memengaruhi hasil prediksi dalam konteks pendidikan.

2.2 Teori Terkait Penelitian

2.2.1 Kinerja Akademik Mahasiswa

Kinerja akademik mahasiswa merupakan gambaran kemampuan individu dalam mencapai hasil belajar selama mengikuti proses pendidikan di perguruan tinggi. Kinerja ini umumnya diukur melalui nilai akademik, seperti Indeks Prestasi Kumulatif (IPK), nilai mata kuliah, maupun pencapaian kompetensi tertentu [23]. Menurut teori pendidikan, kinerja akademik tidak hanya dipengaruhi oleh kemampuan kognitif, tetapi juga oleh faktor afektif dan psikomotorik yang mencerminkan motivasi belajar, kedisiplinan, serta kemampuan manajemen waktu mahasiswa [24]. Oleh karena itu, kinerja

akademik dapat menjadi indikator penting dalam menilai efektivitas proses pembelajaran dan kualitas pendidikan secara keseluruhan.

Selain faktor internal seperti kemampuan intelektual, motivasi, dan kebiasaan belajar, kinerja akademik juga dipengaruhi oleh faktor eksternal seperti lingkungan sosial, dukungan keluarga, metode pengajaran dosen, serta fasilitas kampus. Dalam penelitian modern, terutama dengan pendekatan data-driven, analisis kinerja akademik mahasiswa mulai melibatkan penggunaan data akademik dan non-akademik untuk memahami pola yang memengaruhi hasil belajar. Dengan bantuan teknologi seperti machine learning, berbagai faktor tersebut dapat dimodelkan secara kuantitatif untuk memprediksi serta menjelaskan faktor-faktor yang paling signifikan terhadap keberhasilan akademik mahasiswa [25].

2.2.2 Motivasi Belajar

Motivasi belajar merupakan salah satu aspek psikologis internal yang berperan penting dalam menentukan keberhasilan akademik mahasiswa. Motivasi ini dapat dipahami sebagai kekuatan pendorong yang memengaruhi kesiapan individu untuk memulai, mengarahkan, serta mempertahankan keterlibatannya dalam proses pembelajaran. Mahasiswa yang memiliki tingkat motivasi belajar tinggi umumnya menunjukkan sikap aktif, disiplin, dan berorientasi pada pencapaian tujuan akademik, sehingga mampu mengelola tuntutan perkuliahan dengan lebih baik [26]. Sebaliknya, rendahnya motivasi belajar sering dikaitkan dengan menurunnya partisipasi akademik, lemahnya komitmen terhadap tugas perkuliahan, serta meningkatnya risiko keterlambatan studi atau kegagalan akademik [27].

Pada pendidikan tinggi, motivasi belajar tidak hanya dipengaruhi oleh ketertarikan terhadap materi perkuliahan, tetapi juga oleh faktor-faktor kontekstual seperti dukungan dari dosen, kualitas sistem pembelajaran,

lingkungan akademik, serta persepsi mahasiswa terhadap manfaat jangka panjang dari studinya [28]. Motivasi dapat bersumber dari dorongan intrinsik, yaitu keinginan untuk memahami dan menguasai pengetahuan [29]. Sedangkan dorongan ekstrinsik, seperti target pencapaian nilai, kelulusan tepat waktu, atau prospek karier di masa depan [30]. Mahasiswa dengan tujuan akademik yang jelas dan motivasi yang terjaga cenderung menunjukkan performa yang lebih konsisten dan perkembangan akademik yang positif, sehingga menjadikan motivasi belajar sebagai indikator penting dalam analisis dan prediksi keberhasilan studi mahasiswa.

2.2.3 Peringatan Sistem Dini (Early System Warning)

Peringatan sistem dini (*Early Warning System/EWS*) merupakan pendekatan yang digunakan untuk mengidentifikasi secara awal potensi risiko akademik yang dialami mahasiswa, seperti penurunan capaian belajar, keterlambatan penyelesaian studi, atau kemungkinan putus kuliah. Sistem ini berfungsi sebagai sarana pemantauan berkelanjutan terhadap perkembangan akademik mahasiswa dengan memanfaatkan indikator-indikator tertentu yang mencerminkan performa belajar. Melalui mekanisme peringatan dini, pihak akademik dapat memperoleh sinyal awal mengenai mahasiswa yang memerlukan perhatian khusus, sehingga langkah pencegahan dapat dilakukan sebelum permasalahan akademik berkembang lebih lanjut dan berdampak signifikan terhadap keberhasilan studi mahasiswa [31].

Dalam lingkungan pembelajaran di perguruan tinggi yang bersifat kompleks dan dinamis, implementasi EWS menjadi semakin relevan mengingat mahasiswa dihadapkan pada berbagai tuntutan akademik dan non-akademik secara simultan. Indikator seperti tingkat kehadiran, keterlambatan pengumpulan tugas, fluktuasi nilai, serta menurunnya keterlibatan dalam aktivitas akademik sering kali mencerminkan adanya risiko penurunan kinerja [32]. Dengan memanfaatkan sistem peringatan dini, institusi pendidikan dapat

melakukan intervensi yang tepat sasaran, seperti pemberian bimbingan akademik, layanan konseling, atau pendampingan belajar secara intensif. Selain mendukung upaya pencegahan kegagalan studi, EWS juga berkontribusi dalam meningkatkan kualitas pengelolaan akademik, memperkuat komunikasi antara mahasiswa dan pihak institusi, serta mendorong pendekatan yang lebih proaktif dan berbasis data dalam mendukung keberhasilan studi mahasiswa [33].

2.2.4 Faktor Keluarga

Faktor keluarga merupakan salah satu elemen penting yang memengaruhi kinerja akademik mahasiswa. Lingkungan keluarga berperan sebagai fondasi awal dalam pembentukan nilai, sikap, motivasi, serta kebiasaan belajar individu [34]. Teori *Social Learning* oleh Albert Bandura menjelaskan bahwa perilaku seseorang terbentuk melalui proses observasi dan interaksi dengan lingkungan sekitarnya, termasuk keluarga [35]. Dukungan emosional, perhatian orang tua terhadap pendidikan, serta komunikasi yang baik antara anggota keluarga dapat meningkatkan motivasi intrinsik mahasiswa untuk berprestasi [36]. Sebaliknya, kurangnya dukungan atau adanya konflik dalam keluarga dapat menurunkan fokus, kepercayaan diri, dan semangat belajar mahasiswa, yang pada akhirnya berdampak negatif terhadap hasil akademik.

Selain dukungan emosional, faktor ekonomi keluarga juga memiliki pengaruh signifikan terhadap pencapaian akademik. Berdasarkan teori *Ecological Systems* dari Bronfenbrenner, lingkungan mikro seperti keluarga secara langsung memengaruhi perkembangan individu [37]. Kondisi sosial ekonomi yang stabil memungkinkan mahasiswa memiliki akses terhadap sumber belajar yang memadai, fasilitas pendidikan, serta kesempatan untuk mengembangkan kemampuan diri secara optimal. Sebaliknya, tekanan ekonomi atau ketidakstabilan keluarga dapat menimbulkan stres dan mengganggu konsentrasi belajar. Oleh karena itu, peran keluarga bukan hanya

sebatas pemberi dukungan material, tetapi juga sebagai sumber motivasi, bimbingan, dan stabilitas emosional yang berkontribusi besar terhadap keberhasilan akademik mahasiswa.

2.2.5 Faktor Ekonomi dan Pekerjaan

Faktor ekonomi dan pekerjaan memiliki pengaruh yang signifikan terhadap kinerja akademik mahasiswa karena keduanya menentukan kemampuan individu dalam memenuhi kebutuhan pendidikan serta mengatur keseimbangan antara tanggung jawab akademik dan non-akademik. Berdasarkan *Maslow's Hierarchy of Needs*, kebutuhan fisiologis dan keamanan (seperti makanan, tempat tinggal, dan kestabilan ekonomi) merupakan dasar sebelum seseorang dapat fokus pada kebutuhan tingkat lebih tinggi, termasuk aktualisasi diri melalui prestasi akademik [38]. Mahasiswa yang berasal dari keluarga dengan kondisi ekonomi stabil cenderung memiliki kesempatan lebih besar untuk mengakses sumber belajar tambahan, mengikuti pelatihan, dan belajar dalam lingkungan yang mendukung. Sebaliknya, kesulitan ekonomi dapat menimbulkan tekanan psikologis, stres, dan penurunan motivasi belajar karena mahasiswa harus membagi perhatian antara studi dan pemenuhan kebutuhan hidup [39].

Selain faktor ekonomi, pekerjaan juga menjadi variabel penting yang memengaruhi performa akademik, terutama bagi mahasiswa yang bekerja sambil kuliah. Berdasarkan *Role Strain Theory*, individu yang menjalankan dua peran sekaligus sebagai mahasiswa dan pekerja berpotensi mengalami konflik peran dan kelelahan emosional yang dapat menurunkan kinerja akademik [40]. Namun, pekerjaan tidak selalu berdampak negatif; pengalaman kerja dapat meningkatkan kedisiplinan, manajemen waktu, dan tanggung jawab, yang pada gilirannya mendukung prestasi belajar apabila mahasiswa mampu mengelola waktu dengan baik [41]. Oleh karena itu, keseimbangan antara kebutuhan

ekonomi dan tuntutan akademik menjadi faktor penting dalam menentukan keberhasilan mahasiswa selama menempuh pendidikan tinggi.

2.2.6 Faktor Pilihan Studi

Faktor pilihan studi merupakan salah satu faktor eksternal yang memengaruhi keberhasilan akademik mahasiswa, khususnya terkait kesesuaian antara minat, kemampuan, dan karakteristik program studi yang dipilih. Pilihan studi tidak hanya ditentukan oleh preferensi pribadi mahasiswa, tetapi juga dipengaruhi oleh berbagai faktor eksternal seperti rekomendasi orang tua, pengaruh lingkungan sosial, reputasi institusi, prospek karier, serta kebijakan penerimaan mahasiswa [42]. Ketidaksesuaian antara latar belakang akademik, minat belajar, dan tuntutan kurikulum program studi berpotensi menimbulkan kesulitan adaptasi, menurunkan motivasi belajar, serta berdampak negatif pada kinerja akademik mahasiswa [43].

Selain itu, faktor pilihan studi juga berkaitan erat dengan persepsi mahasiswa terhadap peluang masa depan dan relevansi bidang studi dengan kebutuhan dunia kerja. Mahasiswa yang memilih program studi berdasarkan pertimbangan eksternal seperti tekanan sosial atau tren pasar tanpa pemahaman yang memadai mengenai karakteristik bidang studi cenderung mengalami ketidakpuasan akademik dan ketidakstabilan performa belajar [44]. Sebaliknya, pemilihan studi yang didukung oleh informasi yang jelas mengenai kurikulum, beban akademik, serta prospek lulusan dapat membantu mahasiswa membangun ekspektasi yang realistis dan meningkatkan komitmen terhadap proses pembelajaran [45]. Oleh karena itu, faktor pilihan studi menjadi aspek penting yang perlu diperhatikan dalam analisis kinerja akademik mahasiswa, terutama dalam upaya mengidentifikasi potensi risiko akademik sejak dini.

2.3 Framework dan Algoritma

2.3.1 Data Mining

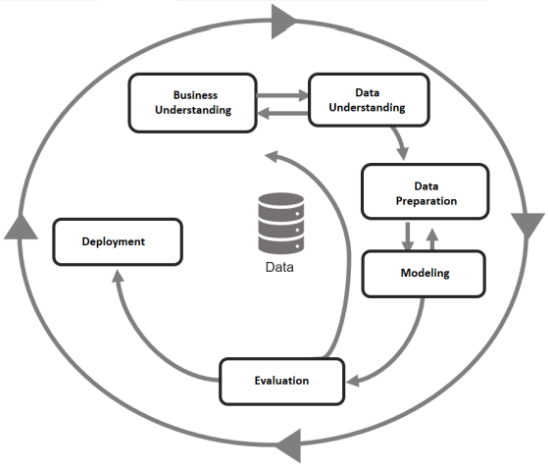
Data mining merupakan proses sistematis untuk mengekstraksi pola, hubungan, dan pengetahuan yang bermakna dari kumpulan data berukuran besar dan kompleks. Proses ini mencakup serangkaian tahapan mulai dari pengumpulan data, pembersihan data, transformasi, hingga penerapan teknik analisis untuk menemukan informasi tersembunyi yang tidak dapat diperoleh melalui pengamatan langsung [46]. Data mining berperan sebagai jembatan antara data mentah dan pengetahuan yang dapat digunakan sebagai dasar pengambilan keputusan, terutama dalam konteks analisis prediktif dan eksploratif.

Dalam kerangka kerja analisis data, data mining memanfaatkan berbagai pendekatan algoritmik seperti klasifikasi, regresi, klastering, dan asosiasi untuk mengidentifikasi pola perilaku dan tren tertentu [47]. Penerapan data mining memungkinkan pengolahan data secara efisien serta penyederhanaan kompleksitas data menjadi informasi yang lebih mudah dipahami. Dalam bidang pendidikan, data mining banyak digunakan untuk menganalisis data akademik mahasiswa guna mendukung pemantauan kinerja, prediksi hasil belajar, serta pengembangan sistem pendukung keputusan [48]. Dengan demikian, data mining menjadi komponen penting dalam framework analisis berbasis data yang mendasari penerapan algoritma machine learning pada penelitian ini.

2.3.2 Cross Industry Standard Process for Data Mining

Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan kerangka kerja yang banyak diadopsi dalam pengembangan proyek data mining dan analisis data karena menyediakan alur kerja yang jelas, terstruktur, dan mudah diadaptasi lintas sektor. Framework ini dirancang untuk bersifat umum namun tetap fleksibel, sehingga dapat diterapkan pada berbagai domain penelitian, termasuk bidang pendidikan, bisnis, dan kesehatan [49]. Dalam

penelitian berbasis data, CRISP-DM berperan sebagai panduan metodologis yang membantu peneliti memastikan bahwa proses pengolahan dan analisis data berjalan secara sistematis serta selaras dengan tujuan penelitian yang telah ditetapkan. CRISP-DM terdiri dari enam tahapan utama yang saling terhubung dan bersifat iteratif, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment* [50]. Keenam tahap tersebut membentuk suatu siklus kerja yang memungkinkan dilakukannya evaluasi dan penyempurnaan secara berulang pada setiap fase, sehingga kualitas hasil analisis dapat terus ditingkatkan. Setiap tahapan memiliki peran strategis dalam menjamin ketepatan proses, keandalan model, serta relevansi hasil terhadap permasalahan yang diteliti. Hubungan antar tahapan dalam framework CRISP-DM tersebut disajikan secara visual pada Gambar 2.1.



Gambar 2. 1 *Framework CRISP-DM*

Sumber : [51]

Gambar 2.1 menggambarkan alur kerja dalam framework CRISP-DM yang terdiri dari beberapa tahapan inti dalam proses pemahaman bisnis dan analisis data. Framework ini menyajikan pendekatan terstruktur yang memandu peneliti mulai dari perumusan masalah hingga penerapan hasil analisis. Adapun tahapan-tahapan dalam CRISP-DM dijelaskan sebagai berikut [51]:

1) *Business Understanding*

Tahap ini berfokus pada pemahaman tujuan dan kebutuhan analisis secara menyeluruh. Pada fase ini, permasalahan yang ingin diselesaikan didefinisikan dengan jelas, serta ditentukan sasaran dan indikator keberhasilan yang akan digunakan sebagai acuan dalam proses analisis data.

2) *Data Understanding*

Tahap data understanding bertujuan untuk mengenali karakteristik data yang tersedia. Proses ini meliputi eksplorasi awal data, identifikasi pola distribusi, pemeriksaan kelengkapan data, serta deteksi anomali yang berpotensi memengaruhi kualitas analisis.

3) *Data Preparation*

Pada tahap ini, data disiapkan agar layak digunakan dalam proses pemodelan. Kegiatan yang dilakukan mencakup pembersihan data dari kesalahan atau inkonsistensi, transformasi data ke format yang sesuai, serta pemilihan atribut yang relevan dengan tujuan analisis.

4) *Modeling*

Tahap modeling berfokus pada pembangunan model analitik dengan menggunakan teknik atau algoritma yang sesuai. Pada fase ini juga dilakukan penyesuaian parameter model untuk memperoleh hasil yang optimal sesuai dengan karakteristik data yang digunakan.

5) *Evaluation*

Tahap evaluasi bertujuan untuk menilai kinerja dan kualitas model yang telah dibangun. Evaluasi dilakukan untuk memastikan bahwa model mampu memenuhi tujuan analisis serta memberikan hasil yang valid dan dapat diandalkan.

6) *Deployment*

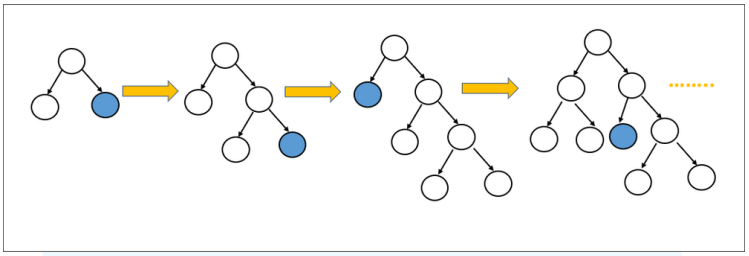
Tahap deployment merupakan fase penerapan hasil analisis ke dalam sistem atau media yang dapat dimanfaatkan oleh pengguna. Pada tahap ini, model atau hasil analisis disajikan dalam bentuk yang mudah dipahami sehingga dapat digunakan sebagai dasar pengambilan keputusan.

2.3.2 LightGBM

Light Gradient Boosting Machine (LightGBM) merupakan algoritma pembelajaran mesin berbasis *ensemble* yang dikembangkan oleh Microsoft dan termasuk ke dalam kelompok metode *gradient boosting* berbasis pohon keputusan. Algoritma ini dirancang untuk meningkatkan efisiensi proses pelatihan model, khususnya pada dataset yang memiliki ukuran besar dan jumlah fitur yang tinggi, tanpa mengorbankan kualitas prediksi [52]. LightGBM membangun model secara bertahap dengan mengombinasikan sejumlah *decision tree* yang disusun untuk meminimalkan kesalahan prediksi secara iteratif.

Salah satu karakteristik utama LightGBM adalah penggunaan strategi pertumbuhan pohon secara *leaf-wise*, di mana cabang dengan potensi penurunan kesalahan terbesar diprioritaskan untuk dikembangkan. Pendekatan ini berbeda dari strategi *level-wise* yang umum digunakan pada algoritma boosting lainnya dan memungkinkan proses pembelajaran berjalan lebih cepat serta menghasilkan model yang lebih efisien [53]. Selain itu, LightGBM mengintegrasikan berbagai teknik optimisasi, seperti *Gradient-based One Side Sampling* dan *Exclusive Feature Bundling*, yang bertujuan untuk mengurangi beban komputasi dan kebutuhan memori [54]. Berkat kombinasi strategi tersebut, LightGBM dikenal memiliki kinerja yang stabil, waktu pelatihan yang singkat, serta kemampuan yang baik dalam menangani data berskala besar dan

berdimensi tinggi, sehingga banyak digunakan dalam berbagai aplikasi analisis data dan pembelajaran mesin.



Gambar 2. 2 Arsitektur LightBGM

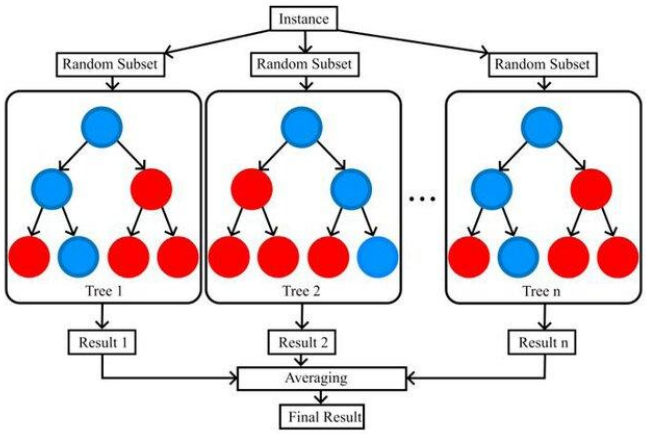
Sumber : [55]

Ilustrasi pada Gambar 2.2 menunjukkan mekanisme pertumbuhan pohon keputusan yang diterapkan oleh LightGBM melalui pendekatan *leaf-wise*. Strategi ini menjadi karakteristik utama LightGBM dan membedakannya dari metode boosting lainnya yang umumnya mengembangkan pohon secara bertingkat atau *level-wise*. Menambah cabang secara seragam pada setiap tingkat, LightGBM melakukan evaluasi terhadap seluruh daun yang telah terbentuk untuk menentukan bagian mana yang paling berpotensi memberikan peningkatan kinerja apabila dilakukan pemisahan lebih lanjut [56].

Pada setiap iterasi, algoritma akan memilih satu daun dengan kontribusi penurunan kesalahan paling besar untuk dikembangkan, sehingga struktur pohon yang dihasilkan cenderung tidak simetris [57]. Pola pertumbuhan ini memungkinkan model memusatkan proses pembelajaran pada area data yang memiliki tingkat kesulitan atau informasi paling tinggi. Dengan demikian, sumber daya komputasi dimanfaatkan secara lebih efektif, yang berdampak pada percepatan waktu pelatihan sekaligus peningkatan kualitas prediksi [58]. Pendekatan *leaf-wise* ini menjadikan LightGBM mampu menghasilkan model yang efisien dan kompetitif dalam berbagai skenario analisis data.

2.3.3 Random Forest

Random Forest merupakan algoritma *ensemble learning* yang menggabungkan sejumlah besar *decision tree* untuk menghasilkan prediksi yang lebih akurat dan stabil. Setiap pohon dalam Random Forest dibangun menggunakan subset data dan fitur yang dipilih secara acak, sehingga mampu mengurangi risiko *overfitting* dan meningkatkan kemampuan generalisasi model [59]. Proses pengambilan keputusan dilakukan dengan cara menggabungkan hasil dari seluruh pohon, baik melalui voting untuk klasifikasi maupun rata-rata untuk regresi [60]. Pendekatan ini membuat Random Forest unggul dalam menangani data berukuran besar, bersifat non-linear, serta mampu memberikan interpretasi terhadap tingkat kepentingan setiap fitur (*feature importance*), yang berguna untuk memahami faktor-faktor yang paling berpengaruh terhadap hasil prediksi [61].



Gambar 2. 3 Arsitektur Random Forest

Sumber : [62]

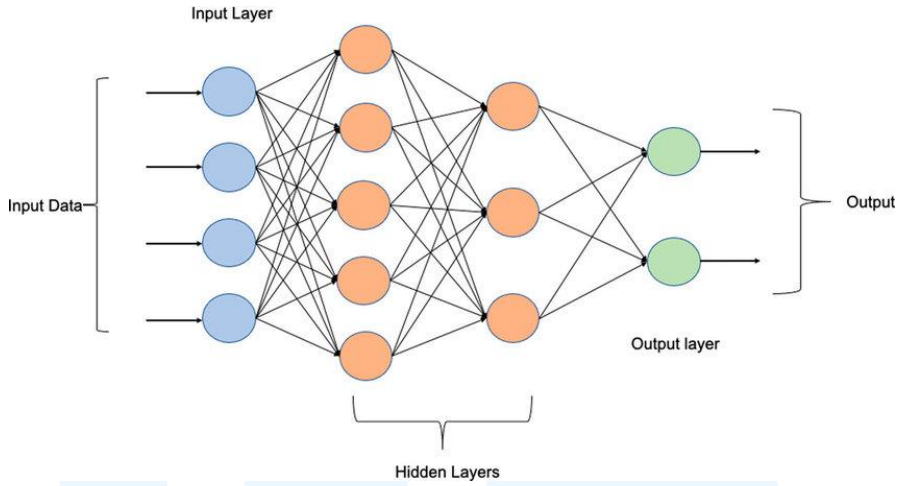
Gambar 2.3 Arsitektur Random Forest menggambarkan bagaimana algoritma ini bekerja melalui kombinasi beberapa *decision tree* untuk menghasilkan prediksi akhir yang lebih akurat. Setiap pohon keputusan (*tree*) dibangun dari *subset* data dan fitur yang dipilih secara acak dari *dataset* utama. Setiap pohon menghasilkan hasil prediksi masing-masing (*Result 1, Result 2,*

..., Result n), yang kemudian digabungkan menggunakan proses *averaging* (untuk regresi) atau *voting* (untuk klasifikasi) guna memperoleh hasil akhir (*Final Result*) [63]. Pendekatan ini memungkinkan Random Forest mengurangi kesalahan prediksi dari satu pohon tunggal serta meningkatkan kestabilan model secara keseluruhan. Dengan mekanisme *random subset* pada tahap pembentukan pohon, algoritma ini mampu meminimalkan *overfitting* dan memberikan hasil prediksi yang lebih andal pada berbagai jenis data [64].

2.3.4 Multilayer Perceptron

Multilayer Perceptron (MLP) adalah salah satu bentuk jaringan saraf tiruan yang bekerja dengan pola aliran data satu arah, di mana informasi diproses dari lapisan masukan hingga menghasilkan keluaran tanpa adanya mekanisme umpan balik [65]. Struktur MLP umumnya tersusun atas beberapa lapisan, yaitu lapisan input, satu atau lebih lapisan tersembunyi, serta lapisan output. Setiap lapisan terdiri dari unit pemroses atau neuron yang saling terhubung dan memiliki bobot sebagai pengatur kekuatan hubungan antar neuron [66].

Proses pembelajaran pada MLP dilakukan dengan menyesuaikan bobot koneksi secara bertahap melalui mekanisme pembaruan berbasis kesalahan antara hasil prediksi dan nilai target. Mekanisme ini memungkinkan jaringan memperbaiki performanya secara iteratif hingga mencapai tingkat kesalahan yang minimal. Untuk mendukung kemampuan pemodelan yang lebih kompleks, MLP menggunakan fungsi aktivasi *non-linier*, seperti sigmoid atau ReLU, yang memungkinkan jaringan menangkap hubungan yang tidak bersifat linier antar variabel input [67]. Dengan karakteristik tersebut, MLP banyak digunakan dalam berbagai tugas analisis data, termasuk klasifikasi dan prediksi, terutama ketika pola data yang dihadapi bersifat kompleks dan tidak dapat direpresentasikan secara sederhana oleh model linier.



Gambar 2. 4 Arsitektur Multilayer Perceptron

Sumber : [68]

Gambar 2.4 menampilkan struktur Multilayer Perceptron (MLP) sebagai salah satu jenis jaringan saraf tiruan berarsitektur *feedforward* dengan alur pemrosesan satu arah. Jaringan ini terdiri atas tiga jenis lapisan utama, yaitu lapisan masukan, satu atau lebih lapisan tersembunyi, dan lapisan keluaran [69]. Pada ilustrasi, lapisan masukan menerima sejumlah fitur data yang kemudian diteruskan ke lapisan-lapisan tersembunyi untuk diproses secara bertahap sehingga hubungan dan pola non-linier yang lebih kompleks dapat ditangkap. Setiap neuron pada suatu lapisan terhubung dengan seluruh neuron pada lapisan berikutnya melalui bobot sinaptik, sehingga membentuk arsitektur *fully connected* [67]. Pada contoh yang ditampilkan, terdapat dua lapisan tersembunyi yang masing-masing berisi beberapa neuron, sedangkan lapisan keluaran menghasilkan nilai prediksi akhir. Karakteristik utama arsitektur ini adalah aliran informasi yang hanya bergerak maju dari lapisan masukan menuju lapisan keluaran tanpa adanya koneksi umpan balik, serta kepadatan koneksi antarneuron antar-lapisan yang memungkinkan MLP memodelkan fungsi pemetaan yang kompleks [70].

2.3.5 Evaluation Metrics

Evaluation metrics merupakan kumpulan ukuran yang digunakan untuk menilai sejauh mana suatu model analisis data atau algoritma machine learning bekerja dengan baik. Metrik ini berfungsi sebagai alat untuk mengukur kemampuan model dalam mengenali pola, menghasilkan prediksi, serta mencapai tujuan analisis yang telah ditetapkan. Dalam *data mining* dan *machine learning*, *evaluation metrics* memberikan informasi penting mengenai tingkat ketepatan, keandalan, dan efektivitas model ketika diterapkan pada suatu dataset tertentu.

Penggunaan *evaluation metrics* sangat krusial, khususnya pada proses pelatihan model klasifikasi, karena hasil evaluasi menjadi dasar dalam membandingkan performa antar model dan menentukan model yang paling sesuai. Pemilihan metrik evaluasi juga perlu disesuaikan dengan karakteristik data dan jenis permasalahan yang dihadapi, seperti klasifikasi atau regresi, agar penilaian kinerja model dapat dilakukan secara objektif dan relevan. Oleh karena itu, pemahaman terhadap *evaluation metrics* yang digunakan menjadi langkah penting dalam memastikan bahwa model yang dikembangkan mampu memberikan hasil yang optimal dan dapat diandalkan. Berikut adalah penjelasan evaluation metrics yang digunakan:

1) Accuracy

Accuracy adalah alat ukur proporsi prediksi yang benar terhadap total jumlah data. *Accuracy* bekerja dengan baik jika data seimbang, tetapi kurang efektif pada dataset yang memiliki ketidakseimbangan kelas (*imbalance*). Rumus *accuracy* dapat dilihat pada rumus (2.1) [71].

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \quad (2.1)$$

Rumus 2. 1 Rumus Accuracy

Berdasarkan Persamaan (2.1), *True Positive (TP)* merepresentasikan jumlah data yang termasuk dalam kelas positif dan berhasil dikenali dengan benar oleh model. Sementara itu, *True Negative (TN)* menunjukkan jumlah data dari kelas negatif yang juga diklasifikasikan secara tepat. Di sisi lain, *False Positive (FP)* menggambarkan kondisi ketika data yang seharusnya termasuk kelas negatif justru diprediksi sebagai positif, sedangkan *False Negative (FN)* menunjukkan jumlah data positif yang keliru diidentifikasi sebagai kelas negatif oleh model.

2) *Precision*

Precision digunakan untuk mengukur akurasi prediksi model terhadap kelas positif, dengan menghitung proporsi data positif yang benar dari semua data yang diprediksi sebagai positif. *Precision* penting untuk digunakan dalam kasus di mana kesalahan *false positive* harus diminimalkan. Rumus Precision dapat dilihat pada rumus (2.2) [72]

$$Precision = \frac{tp}{tp+fp} \quad (2.2)$$

Rumus 2. 2 Rumus Precision

Berdasarkan Persamaan (2.2), *True Positive (TP)* merupakan jumlah data yang diprediksi sebagai positif dan sesuai dengan kondisi sebenarnya, sedangkan *False Positive (FP)* adalah data yang diperkirakan positif namun tidak sesuai dengan kelas aslinya. Nilai *precision* menggambarkan tingkat ketepatan prediksi positif yang dihasilkan oleh model, di mana nilai precision yang lebih tinggi menunjukkan kemampuan model yang lebih baik dalam meminimalkan kesalahan prediksi positif.

3) Recall (Sensitivity)

Recall mengukur kemampuan model untuk mendeteksi semua data positif yang sebenarnya. Recall penting dalam situasi di mana kesalahan *false negative* memiliki konsekuensi yang besar. Berikut merupakan rumus dari recall pada (2.3) [73].

$$Recall = \frac{tp}{tp + fn} \tag{2.3}$$

Rumus 2. 3 Rumus Recall

Berdasarkan Persamaan (2.3), *True Positive (TP)* menunjukkan jumlah data positif yang berhasil diprediksi dengan tepat oleh model, sedangkan *False Negative (FN)* merepresentasikan data yang seharusnya termasuk dalam kelas positif tetapi keliru diklasifikasikan sebagai negatif.

4) F1-Score

F1-Score adalah rata-rata harmonis dari *precision* dan recall. Metrik ini berguna untuk distribusi data yang tidak seimbang. F1-Score memberikan keseimbangan antara *precision* dan *recall*, sehingga menjadi metrik yang ideal untuk evaluasi model secara keseluruhan. Berikut merupakan rumus dari F1-Score (2.4) [74].

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.4}$$

Rumus 2. 4 Rumus F1-Score

Berdasarkan Persamaan (2.4), nilai tersebut diperoleh dari perhitungan rata-rata harmonik antara *precision* dan *recall*, sehingga mampu merepresentasikan keseimbangan kinerja model dalam menghasilkan prediksi yang tepat dan lengkap.

5) Confusion Matrix

Confusion Matrix adalah salah satu teknik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi dengan cara membandingkan

hasil prediksi yang dihasilkan model dengan label sebenarnya. Penyajian dalam bentuk tabel memungkinkan pemetaan antara prediksi yang tepat dan prediksi yang keliru pada setiap kelas, sehingga memudahkan pemahaman terhadap jenis kesalahan yang terjadi. Melalui matriks ini, dapat diketahui kemampuan model dalam mengklasifikasikan setiap kategori secara benar serta bagian-bagian di mana model masih mengalami kekeliruan. Data yang dihasilkan dari Confusion Matrix selanjutnya digunakan sebagai dasar perhitungan berbagai metrik evaluasi, seperti *accuracy*, *precision*, *recall*, dan F1-score, yang secara bersama-sama memberikan penilaian menyeluruh terhadap kualitas performa model klasifikasi. Berikut rumus dari dari Confusion Matrix pada (2.5) [75].

$$Confusion\ Matrix = \begin{matrix} TP & FP \\ FN & TN \end{matrix} \quad (2.5)$$

Rumus 2. 5 Rumus Confusion Matrix

Rumus 2.5 menggambarkan hubungan antara hasil prediksi model dan kondisi aktual dalam bentuk matriks dua dimensi. Baris pada matriks menunjukkan kelas aktual, sedangkan kolom merepresentasikan kelas hasil prediksi. Nilai pada diagonal utama menunjukkan prediksi yang benar, sedangkan nilai di luar diagonal menunjukkan kesalahan klasifikasi. Matriks ini menjadi dasar untuk menghitung metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score*.

2.3.6 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) merupakan bidang dalam kecerdasan buatan yang menitikberatkan pada pengembangan pendekatan agar mekanisme pengambilan keputusan pada model machine learning dapat dijelaskan dan dipahami oleh manusia [76]. Tujuan utama XAI adalah

meningkatkan transparansi model dengan mengungkap alasan di balik suatu prediksi, termasuk peran dan pengaruh masing-masing fitur terhadap hasil yang dihasilkan [77]. Selain meningkatkan tingkat kepercayaan pengguna, XAI juga berperan penting dalam memastikan bahwa keputusan model selaras dengan prinsip etika, keadilan, serta akuntabilitas dalam penerapannya [78].

Secara garis besar, pendekatan XAI terbagi menjadi dua kategori, yaitu *intrinsic interpretability* dan *post-hoc explanation*. *Intrinsic interpretability* merujuk pada model yang sejak awal memiliki struktur yang mudah dipahami, seperti *linear regression* dan *decision tree*. Sementara itu, *post-hoc explanation* digunakan untuk menjelaskan model yang bersifat kompleks, seperti *deep learning* atau *ensemble models*, dengan memanfaatkan teknik interpretasi tambahan, salah satunya adalah metode SHAP (*SHapley Additive exPlanations*) [79].

2.3.7 SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) merupakan metode interpretabilitas model yang dikembangkan untuk menjelaskan hasil prediksi dari algoritma pembelajaran mesin secara lebih transparan dan akurat. SHAP didasarkan pada teori *Shapley value* dari bidang teori permainan (*game theory*), yang bertujuan untuk menentukan kontribusi masing-masing fitur terhadap hasil prediksi model [80]. Dengan pendekatan ini, setiap fitur dianggap sebagai "pemain" dalam sebuah permainan kolaboratif yang berkontribusi terhadap "keuntungan" akhir, yaitu nilai prediksi. Melalui pembagian kontribusi yang adil, SHAP mampu menjelaskan seberapa besar pengaruh setiap fitur terhadap output yang dihasilkan oleh model.

Metode SHAP memberikan penjelasan yang bersifat aditif, artinya hasil prediksi suatu model dapat dianggap sebagai penjumlahan dari kontribusi setiap fitur ditambah dengan nilai dasar (*base value*) [10]. Nilai dasar ini

merepresentasikan rata-rata output model ketika tidak ada fitur yang digunakan, sedangkan kontribusi setiap fitur menunjukkan seberapa jauh fitur tersebut menggeser prediksi dari nilai dasar [81]. Pendekatan ini membuat SHAP unggul dibandingkan teknik interpretasi lainnya karena mampu memberikan penjelasan yang konsisten dan mudah dipahami secara matematis maupun visual.

Salah satu kelebihan utama SHAP adalah kemampuannya untuk diterapkan pada berbagai algoritma, baik model sederhana seperti *Linear Regression* maupun model kompleks seperti *Random Forest* dan *XGBoost* [82]. Melalui visualisasi seperti *SHAP summary plot* atau *dependence plot*, peneliti dapat melihat fitur mana yang paling berpengaruh terhadap prediksi serta arah pengaruhnya antara positif atau negatif [10]. Dengan demikian, SHAP tidak hanya menjawab pertanyaan "seberapa besar" pengaruh suatu fitur, tetapi juga "bagaimana" fitur tersebut memengaruhi hasil prediksi.

Selain meningkatkan transparansi, SHAP juga membantu pengambilan keputusan berbasis data yang lebih tepat. Misalnya, pada analisis kinerja akademik mahasiswa, SHAP dapat mengidentifikasi faktor-faktor utama yang berkontribusi terhadap pencapaian akademik seperti nilai ujian, kehadiran, atau aktivitas pembelajaran mandiri [83]. Melalui interpretasi ini, pengambil kebijakan dapat merancang strategi pembelajaran yang lebih efektif dan personal. Dengan demikian, SHAP berperan penting dalam menjembatani kesenjangan antara performa tinggi model pembelajaran mesin dan kebutuhan akan interpretabilitas dalam pengambilan keputusan yang bertanggung jawab.

2.4 *Tools* Penelitian

Dalam penelitian ini digunakan beberapa perangkat bantu (tools) yang memiliki peran penting dalam tahapan pengolahan data, analisis, hingga pembangunan model prediksi. Pemilihan tools yang tepat membantu meningkatkan efisiensi serta akurasi

dalam proses penelitian. Adapun tools utama yang digunakan meliputi Jupyter Notebook dan Python, yang akan dijelaskan lebih lanjut pada subbagian berikut.

2.4.1 *Google Colabs*

Google Colab atau Google Collaboratory merupakan platform komputasi berbasis cloud yang disediakan oleh Google dan digunakan secara luas untuk menjalankan kode Python, terutama pada bidang data science dan machine learning. Platform ini memungkinkan pengguna untuk menulis, menjalankan, serta membagikan notebook interaktif tanpa perlu melakukan instalasi perangkat lunak tambahan di komputer lokal [84]. Google Colab terintegrasi langsung dengan Google Drive, sehingga memudahkan pengguna dalam menyimpan, mengakses, dan berbagi proyek secara online [85].

Salah satu keunggulan utama Google Colab adalah kemampuannya menyediakan sumber daya komputasi gratis seperti CPU, GPU, dan TPU, yang sangat membantu dalam mempercepat proses pelatihan model machine learning berskala besar [86]. Dengan fitur ini, peneliti dan pengembang dapat menjalankan eksperimen secara efisien tanpa harus memiliki perangkat keras berkapasitas tinggi. Selain itu, Google Colab juga mendukung berbagai library populer seperti NumPy, Pandas, Matplotlib, Scikit-learn, dan TensorFlow, yang mempermudah proses analisis data dan pengembangan model.

Google Colab juga mendukung kolaborasi secara real-time, di mana beberapa pengguna dapat mengedit dan meninjau notebook secara bersamaan, mirip dengan fitur yang terdapat pada Google Docs. Fasilitas ini menjadikannya alat yang sangat efektif untuk penelitian, pembelajaran, serta pengembangan proyek kolaboratif [87]. Dengan kemudahan akses, kemampuan komputasi tinggi, serta integrasi dengan ekosistem Google, Google Colab menjadi salah satu platform utama yang digunakan oleh akademisi, mahasiswa, dan praktisi data untuk melakukan eksperimen dan analisis berbasis machine learning [84].

2.4.2 *Python*

Python merupakan bahasa pemrograman tingkat tinggi yang populer dan banyak digunakan di berbagai bidang, terutama pada data science, machine learning, artificial intelligence, serta pengembangan web. Bahasa ini dikenal karena sintaksnya yang sederhana dan mudah dibaca, sehingga memudahkan pengguna dari berbagai latar belakang untuk mempelajarinya. Python juga bersifat open-source dan memiliki komunitas pengguna yang sangat luas, menjadikannya bahasa yang terus berkembang dengan berbagai pembaruan serta dukungan dari ribuan kontributor di seluruh dunia [88].

Salah satu keunggulan utama Python adalah ketersediaan berbagai pustaka (library) dan framework yang mendukung analisis data dan pengembangan model machine learning. Beberapa library yang paling sering digunakan antara lain NumPy untuk komputasi numerik, Pandas untuk pengolahan data, Matplotlib dan Seaborn untuk visualisasi, serta Scikit-learn dan TensorFlow untuk pemodelan dan pembelajaran mesin [89]. Dengan ekosistem pustaka yang kaya, Python memungkinkan pengguna untuk melakukan berbagai tahapan analisis data, mulai dari pembersihan hingga prediksi, secara efisien dalam satu lingkungan kerja yang terintegrasi.

Selain fleksibilitasnya, Python juga mendukung integrasi dengan berbagai bahasa dan sistem lain seperti C, C++, dan Java, serta dapat dijalankan di berbagai platform seperti Windows, macOS, dan Linux. Hal ini menjadikan Python sebagai bahasa pemrograman yang serbaguna dan dapat diandalkan untuk berbagai keperluan riset maupun implementasi industri [90]. Karena kemudahan penggunaan, efisiensi, dan dukungan komunitas yang luas, Python telah menjadi standar utama dalam pengembangan solusi berbasis data dan kecerdasan buatan di berbagai institusi pendidikan, penelitian, dan perusahaan teknologi.