

BAB III

METODE PENELITIAN

3.1. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan tujuan mengembangkan dan mengevaluasi model prediksi kinerja akademik mahasiswa berbasis data akademik. Pendekatan kuantitatif dipilih karena memungkinkan pengukuran performa model secara objektif melalui analisis numerik dan penggunaan metrik evaluasi yang terstandar. Dengan pendekatan ini, proses pengembangan model dapat dilakukan secara sistematis, terukur, serta dapat direplikasi, sehingga hasil penelitian memiliki tingkat keandalan yang tinggi sesuai dengan kaidah penelitian ilmiah di bidang *data mining* dan *machine learning*.

Fokus utama penelitian ini adalah membangun model prediksi yang mampu mengklasifikasikan kinerja akademik mahasiswa ke dalam beberapa kategori, seperti mahasiswa berisiko mengalami penurunan performa, mahasiswa dengan kinerja stabil, maupun mahasiswa dengan performa akademik yang baik. Untuk mencapai tujuan tersebut, penelitian ini menerapkan pendekatan komparatif dan *ensemble*, dengan membandingkan kinerja beberapa algoritma machine learning yang memiliki karakteristik pembelajaran berbeda, yaitu Light Gradient Boosting Machine (LightGBM), Random Forest, dan Multilayer Perceptron (MLP). Pemilihan ketiga algoritma ini didasarkan pada kemampuannya dalam menangani data tabular, memodelkan hubungan non-linear, serta kestabilannya dalam menghadapi variasi data akademik mahasiswa.

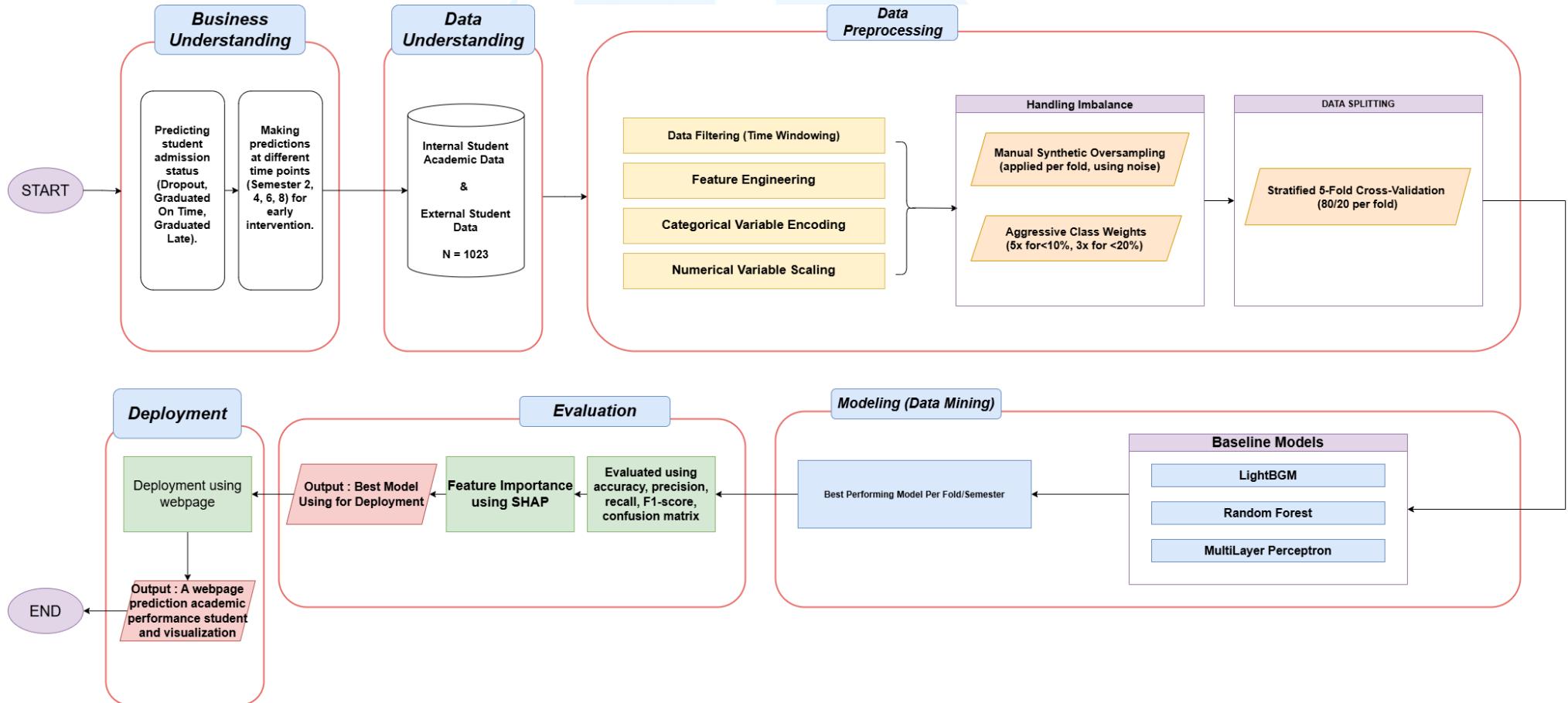
Selain menilai performa prediksi, penelitian ini juga menekankan aspek interpretabilitas model melalui penerapan metode SHapley Additive Explanations (SHAP). Pendekatan ini digunakan untuk menjelaskan kontribusi masing-masing variabel akademik terhadap hasil prediksi yang dihasilkan oleh model. Dengan demikian, model yang dikembangkan tidak hanya menghasilkan prediksi yang

akurat, tetapi juga mampu memberikan pemahaman yang lebih transparan mengenai faktor-faktor yang memengaruhi kinerja akademik mahasiswa. Hal ini penting agar hasil analisis dapat dimanfaatkan sebagai dasar pengambilan keputusan oleh pihak akademik, khususnya dalam mendukung upaya pemantauan dan intervensi dini.

Secara keseluruhan, pelaksanaan penelitian ini mengikuti kerangka kerja Cross-Industry Standard Process for Data Mining (CRISP-DM) yang terdiri atas enam tahapan utama, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Kerangka ini digunakan untuk memastikan bahwa seluruh tahapan penelitian dilakukan secara terstruktur dan konsisten, mulai dari perumusan permasalahan hingga penerapan hasil analisis. Alur penelitian serta hubungan antar tahapan dalam CRISP-DM akan dijelaskan lebih lanjut pada bagian berikutnya dan ditampilkan pada Gambar 3.1 Alur Penelitian.

Melalui penerapan pendekatan komparatif model machine learning dan integrasi metode explainable AI, penelitian ini diharapkan mampu menghasilkan sistem prediksi kinerja akademik mahasiswa yang tidak hanya memiliki tingkat akurasi yang baik, tetapi juga transparan dan mudah dipahami. Hasil penelitian diharapkan dapat memberikan kontribusi dalam mendukung pengelolaan akademik berbasis data, khususnya dalam mengidentifikasi mahasiswa yang memerlukan perhatian lebih awal serta meningkatkan efektivitas pengambilan keputusan di lingkungan perguruan tinggi.

UNIVERSITAS
MULTIMEDIA
NUSANTARA



Gambar 3. 1 Alur Penelitian

Gambar 3.1 menggambarkan alur penelitian yang digunakan dalam pengembangan model prediksi kinerja akademik mahasiswa dengan mengacu pada kerangka kerja Cross-Industry Standard Process for Data Mining (CRISP-DM). Alur penelitian dimulai dari tahap *business understanding* hingga tahap *deployment*, yang disusun secara terstruktur dan saling berkesinambungan untuk memastikan proses analisis data dilakukan secara sistematis dan terarah.

Tahap awal penelitian adalah *Business Understanding*, yang berfokus pada perumusan tujuan analisis dan permasalahan yang ingin diselesaikan. Pada tahap ini ditetapkan sasaran utama penelitian, yaitu memprediksi status akademik mahasiswa ke dalam beberapa kategori, seperti mahasiswa berisiko mengalami penurunan kinerja, mahasiswa dengan performa stabil, maupun mahasiswa dengan performa akademik yang baik. Selain itu, prediksi dilakukan pada beberapa titik waktu akademik untuk mendukung upaya pemantauan dan intervensi dini.

Selanjutnya, tahap *Data Understanding* dilakukan untuk memahami karakteristik data yang digunakan dalam penelitian. Data yang dianalisis berasal dari data akademik mahasiswa yang mencakup data internal dan data pendukung lainnya. Pada tahap ini dilakukan identifikasi struktur data, jumlah data yang tersedia, serta pemahaman awal terhadap variabel-variabel yang berpotensi memengaruhi kinerja akademik mahasiswa.

Tahap berikutnya adalah *Data Preprocessing*, yang bertujuan menyiapkan data agar siap digunakan dalam proses pemodelan. Proses ini meliputi penyaringan data berdasarkan rentang waktu tertentu (*time windowing*), rekayasa fitur (*feature engineering*), pengkodean variabel kategorikal, serta penskalaan variabel numerik. Selain itu, dilakukan penanganan ketidakseimbangan kelas melalui teknik *synthetic oversampling* dan pemberian bobot kelas yang disesuaikan. Setelah data diproses, dilakukan pembagian data menggunakan metode *stratified 5-fold cross-validation* untuk menjaga proporsi kelas pada setiap fold dan menghindari bias dalam pelatihan model.

Tahap *Modeling (Data Mining)* berfokus pada pembangunan dan pengujian model prediksi menggunakan beberapa algoritma machine learning sebagai model dasar, yaitu Light Gradient Boosting Machine (LightGBM), Random Forest, dan Multilayer Perceptron (MLP). Pada tahap ini dilakukan pelatihan model pada setiap fold dan setiap periode akademik, kemudian dipilih model dengan performa terbaik berdasarkan hasil evaluasi.

Setelah model dibangun, tahap *Evaluation* dilakukan untuk menilai kinerja model secara menyeluruh. Evaluasi dilakukan menggunakan berbagai metrik, seperti accuracy, precision, recall, F1-score, serta confusion matrix, guna memastikan model memiliki performa yang seimbang dan andal. Selain itu, metode SHapley Additive Explanations (SHAP) digunakan untuk menganalisis kontribusi setiap fitur terhadap hasil prediksi, sehingga meningkatkan transparansi dan interpretabilitas model.

Tahap terakhir adalah *Deployment*, di mana model terbaik yang telah dievaluasi diimplementasikan ke dalam sebuah aplikasi berbasis web. Aplikasi ini berfungsi sebagai media visualisasi dan sistem pendukung keputusan yang menyajikan hasil prediksi kinerja akademik mahasiswa. Dengan demikian, hasil penelitian tidak hanya bersifat eksperimental, tetapi juga dapat dimanfaatkan secara praktis oleh pihak akademik untuk mendukung proses pemantauan dan pengambilan keputusan berbasis data.

3.2. Tahapan Penelitian

Penelitian ini menerapkan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining) sebagai panduan dalam pelaksanaan proses analisis data. Pemilihan kerangka ini didasarkan pada kemampuannya dalam menyediakan alur kerja yang terstruktur dan sistematis untuk mengelola seluruh tahapan proyek data mining, mulai dari pemahaman permasalahan hingga penerapan hasil model. CRISP-DM mencakup enam tahap utama, yaitu *Business Understanding*, *Data*

Understanding, Data Preparation, Modeling, Evaluation, dan Deployment. Keenam tahapan tersebut saling berkaitan dan bersifat iteratif, sehingga memungkinkan dilakukan penyesuaian dan penyempurnaan pada setiap fase guna menghasilkan model yang akurat, reliabel, dan dapat dimanfaatkan secara praktis. Uraian mengenai masing-masing tahapan disajikan sebagai berikut:

3.2.1 Business Understanding

Tahap *Business Understanding* merupakan tahap awal dalam alur penelitian yang berfokus pada identifikasi permasalahan serta penetapan tujuan analisis secara jelas dan terarah. Pada penelitian ini, permasalahan utama yang diangkat adalah kebutuhan untuk memprediksi status akademik mahasiswa sebagai upaya mendukung pemantauan dan pengambilan keputusan akademik secara dini. Prediksi status akademik dilakukan dengan mengelompokkan mahasiswa ke dalam beberapa kategori, yaitu *dropout*, lulus tepat waktu, lulus tidak tepat waktu, dan lulus lebih awal, serta dianalisis pada beberapa titik waktu studi, yaitu **semester 2, semester 4, dan semester 6**, guna memberikan gambaran perkembangan kinerja akademik mahasiswa secara bertahap.

Tujuan utama dari tahap ini adalah merumuskan kebutuhan analisis yang relevan dengan konteks akademik serta menentukan arah penelitian pada tahap-tahap berikutnya. Melalui pemahaman permasalahan yang komprehensif, penelitian ini diarahkan untuk mengevaluasi kinerja beberapa model machine learning dalam memprediksi status akademik mahasiswa, membandingkan performa model individual dengan pendekatan ensemble, serta mengidentifikasi faktor-faktor yang berpengaruh terhadap hasil prediksi melalui analisis interpretatif. Selain itu, pada tahap *Business Understanding* juga dilakukan kajian terhadap penelitian terdahulu yang relevan serta penentuan metrik evaluasi yang sesuai, seperti *accuracy, precision, recall, F1-score*, dan *confusion matrix*. Tahap ini menjadi fondasi penting karena hasil perumusan tujuan dan kebutuhan analisis akan memengaruhi keseluruhan alur

penelitian, mulai dari pemilihan data, proses pengolahan data, hingga strategi pemodelan yang digunakan.

3.2.2 *Data Understanding*

Tahap *Data Understanding* bertujuan untuk memperoleh pemahaman yang komprehensif terhadap karakteristik data yang digunakan dalam penelitian. Pada tahap ini, dilakukan eksplorasi awal terhadap data mentah yang terdiri dari 1.023 entri mahasiswa Program Studi Sistem Informasi Universitas Multimedia Nusantara, serta data tambahan yang berasal dari kuesioner mengenai faktor-faktor eksternal yang berpotensi memengaruhi kinerja akademik mahasiswa. Data akademik internal mencakup sejumlah atribut utama, seperti Indeks Prestasi Kumulatif (IPK) per semester, jumlah Satuan Kredit Semester (SKS) yang ditempuh, capaian nilai mata kuliah, dan status akademik mahasiswa. Sementara itu, data kuesioner memuat informasi terkait dukungan keluarga, kondisi finansial, kesesuaian pilihan jurusan, serta faktor lingkungan lain yang berkaitan dengan proses belajar mahasiswa.

Dari keseluruhan data mentah tersebut, terdapat sejumlah mahasiswa yang masih berstatus aktif pada tahun akademik terakhir sehingga belum memiliki status akademik akhir. Oleh karena itu, untuk menjaga konsistensi data dengan tujuan penelitian, dilakukan proses identifikasi awal terhadap status mahasiswa. Hasilnya, hanya 416 entri mahasiswa yang telah memiliki status akademik akhir yang digunakan sebagai dataset utama dalam analisis selanjutnya. Tahap ini penting untuk memastikan bahwa data yang dianalisis benar-benar relevan dalam mendukung proses prediksi status kelulusan.

Selain memahami struktur dan jenis variabel, tahap *Data Understanding* juga mencakup analisis awal terhadap kualitas dan distribusi data. Proses ini meliputi pemeriksaan nilai kosong (*missing values*), data ganda, serta analisis keseimbangan kelas pada variabel target. Hasil eksplorasi menunjukkan

adanya ketidakseimbangan distribusi data, khususnya pada kategori dengan jumlah data relatif kecil, seperti mahasiswa yang lulus lebih awal dan mahasiswa yang tidak lulus tepat waktu. Kondisi ini menjadi perhatian penting karena berpotensi memengaruhi performa model prediksi apabila tidak ditangani dengan strategi yang tepat.

Tahap ini juga melibatkan analisis statistik deskriptif untuk memahami sebaran nilai, rentang data, serta hubungan antarvariabel melalui visualisasi seperti histogram, boxplot, dan *correlation heatmap*. Melalui eksplorasi ini, beberapa fitur yang diperkirakan memiliki pengaruh signifikan terhadap status kelulusan mahasiswa dapat diidentifikasi, seperti nilai rata-rata akademik, jumlah semester yang ditempuh, tingkat konsistensi kehadiran, serta indikator perilaku akademik lainnya. Pemahaman mendalam terhadap karakteristik data pada tahap ini menjadi dasar penting dalam menentukan strategi pengolahan data dan pemilihan fitur pada tahap *Data Preparation* berikutnya.

3.2.3 Data Preprocessing

Tahap *Data Preprocessing* merupakan tahapan krusial yang bertujuan untuk menyiapkan data agar layak dan optimal digunakan dalam proses pemodelan machine learning. Pada tahap ini, dilakukan serangkaian proses untuk meningkatkan kualitas data serta memastikan bahwa data yang digunakan mampu merepresentasikan kondisi akademik mahasiswa secara akurat. Proses awal mencakup penanganan nilai yang hilang (*missing values*), baik melalui penghapusan data tertentu maupun pengisian nilai berdasarkan karakteristik masing-masing variabel, sehingga tidak mengganggu proses pelatihan model.

Selanjutnya, variabel kategorikal diubah ke dalam bentuk numerik dengan menggunakan teknik *label encoding* agar dapat diproses oleh algoritma

machine learning. Untuk variabel target, data dengan status “masih aktif” tidak disertakan dalam proses pelatihan karena belum memiliki label status akademik akhir. Selain itu, dilakukan proses *feature engineering* untuk menghasilkan atribut turunan yang lebih informatif, seperti rata-rata nilai per semester, variasi nilai akademik, rata-rata jumlah SKS per semester, tingkat konsistensi kehadiran, serta variasi kehadiran. Fitur-fitur ini dirancang untuk menangkap pola akademik mahasiswa secara lebih komprehensif dibandingkan penggunaan variabel mentah semata.

Permasalahan ketidakseimbangan kelas yang ditemukan pada tahap sebelumnya ditangani melalui pendekatan *manual synthetic oversampling*. Teknik ini dilakukan dengan menambahkan sampel sintetis pada kelas minoritas melalui penambahan gangguan (*noise*) yang terkontrol terhadap data asli, sehingga distribusi kelas menjadi lebih seimbang. Pendekatan ini dipilih karena dinilai lebih sesuai dengan karakteristik data akademik yang bersifat relatif stabil dan sensitif terhadap perubahan nilai, serta mampu menghindari distorsi pola antar fitur yang dapat muncul pada metode sintetik berbasis interpolasi. Proses oversampling diterapkan secara terbatas pada data pelatihan di setiap fold dalam skema *stratified 5-fold cross-validation*, sehingga tidak terjadi kebocoran data (*data leakage*) ke data validasi.

Selain teknik oversampling, penanganan ketidakseimbangan kelas juga diperkuat melalui penerapan *class weight* pada beberapa algoritma, sehingga model dapat memberikan perhatian yang lebih proporsional terhadap kelas dengan jumlah data yang lebih sedikit tanpa harus menambah sampel baru. Tahap preprocessing juga mencakup proses standardisasi fitur numerik menggunakan *StandardScaler*, khususnya untuk algoritma seperti Multilayer Perceptron (MLP) yang sensitif terhadap perbedaan skala antar fitur. Melalui rangkaian tahapan preprocessing ini, data yang digunakan dalam penelitian

menjadi lebih bersih, seimbang, dan stabil, sehingga mampu mendukung proses pemodelan dan evaluasi yang valid, reliabel, serta minim bias.

3.2.4 Modeling (Data Mining)

Tahap *Modeling (Data Mining)* merupakan tahapan utama dalam penelitian ini, di mana data yang telah melalui proses preprocessing digunakan untuk membangun model prediksi kinerja akademik mahasiswa. Pada tahap ini, algoritma machine learning diterapkan untuk mempelajari pola dari data historis dan menghasilkan model yang mampu mengklasifikasikan status akademik mahasiswa secara akurat. Proses pemodelan dilakukan dengan memanfaatkan beberapa algoritma yang memiliki karakteristik pembelajaran berbeda, yaitu Light Gradient Boosting Machine (LightGBM), Random Forest, dan Multilayer Perceptron (MLP). Ketiga algoritma tersebut dipilih untuk merepresentasikan pendekatan *tree-based ensemble* dan *neural network*, sehingga memungkinkan perbandingan performa model dari sudut pandang yang beragam.

Pengembangan model dilakukan secara bertahap dengan menerapkan beberapa konfigurasi parameter yang berbeda untuk masing-masing algoritma. Setiap konfigurasi dirancang dengan tingkat kompleksitas yang meningkat secara bertahap guna mengevaluasi pengaruh pengaturan parameter terhadap performa model, kestabilan hasil prediksi, serta kemampuan model dalam menangani karakteristik data akademik yang tidak seimbang. Pendekatan ini memungkinkan analisis yang lebih sistematis terhadap perilaku model dan membantu dalam mengidentifikasi konfigurasi yang paling sesuai dengan kondisi dataset yang digunakan. Seluruh proses pelatihan model dilakukan menggunakan skema *stratified 5-fold cross-validation* untuk memastikan distribusi kelas pada data pelatihan dan data validasi tetap terjaga. Dengan demikian, evaluasi performa model dapat dilakukan secara lebih objektif dan

mengurangi potensi bias akibat pembagian data yang tidak seimbang. Hasil perbandingan performa dari berbagai konfigurasi model digunakan sebagai dasar dalam menentukan model dan konfigurasi terbaik yang akan digunakan pada tahap evaluasi dan interpretasi selanjutnya. Ringkasan perbandingan konfigurasi model yang diuji pada tahap ini disajikan pada Tabel 3.1.

Tabel 3. 1 Perbandingan Percobaan Hyperparameter Tuning

| Aspek | Model Configuration 1 | Model Configuration 2 | Model Configuration 3 (Terbaik) |
|------------------|---|---|---|
| Strategi | <i>Conservative</i> (S2) / <i>Moderate</i> (S4/6/8) | <i>Ultra-Optimized</i> | <i>Default Configuration</i> |
| Kompleksitas | Rendah - Menengah | Sangat Tinggi | Rendah |
| Parameter | Hampir semua default, hanya objective & eval_metric | n_estimators tinggi (hingga 800), max_depth sampai 12, hidden_layer_sizes besar | Parameter default, hanya hidden_layer_sizes pada MLP dan beberapa setting dasar |
| Regularization | Regularisasi bawaan model (L2 default), tidak diatur manual | Regularization efektif dari tree depth & shrinkage learning rate | Regularization default tanpa penyesuaian tambahan |
| Early Stopping | Tidak digunakan | Ada (pada beberapa model ultra-optimized, XGB/LightGBM jika ada validasi) | Tidak digunakan |
| Tuning Parameter | Minimal, eksplorasi kecil | Tuning signifikan (depth, estimators, hidden layers) | Tanpa grid search, tanpa Bayesian optimization |
| Training Time | Cepat | Lambat (banyak iterasi & parameter besar) | Cepat |
| Performance | Bagus | Baik | Stabil dan Kompetitif (terbaik untuk small dataset) |
| Use Case | Experimentation / Cross-Validation | Production Ready | Deployment |

Tabel 3.1 menyajikan perbandingan tiga skema konfigurasi model yang diterapkan pada tahap *Data Mining*, masing-masing dengan tingkat pengaturan parameter yang berbeda. *Model Configuration 1* menerapkan pengaturan parameter yang relatif sederhana pada beberapa algoritma utama, seperti LightGBM, XGBoost, CatBoost, dan MLP, sehingga waktu pelatihan menjadi lebih efisien dan dapat digunakan sebagai acuan awal (*baseline*) dalam evaluasi performa model. *Model Configuration 2* menggunakan pengaturan parameter yang lebih kompleks serta melibatkan tambahan algoritma Random Forest untuk meningkatkan variasi pendekatan pemodelan, namun konsekuensinya adalah meningkatnya waktu pelatihan serta sensitivitas model terhadap ukuran dan karakteristik dataset.

Selanjutnya, *Model Configuration 3* ditetapkan sebagai konfigurasi akhir yang digunakan dalam implementasi, dengan menerapkan parameter dasar yang disesuaikan secara minimal serta strategi *manual synthetic oversampling* pada setiap fold pelatihan untuk menghindari terjadinya kebocoran data. Hasil perbandingan menunjukkan bahwa peningkatan kompleksitas parameter tidak selalu berbanding lurus dengan peningkatan kinerja model. Dalam konteks dataset berukuran relatif kecil, konfigurasi yang lebih sederhana justru menghasilkan performa yang lebih konsisten dan stabil, sehingga dinilai lebih sesuai untuk tahap penerapan (*deployment*).

3.2.5 Evaluation

Tahap *Evaluation* bertujuan untuk menilai kinerja model machine learning yang telah dibangun pada tahap *Modeling*, serta memastikan bahwa model tersebut mampu memenuhi tujuan penelitian yang telah ditetapkan pada tahap *Business Understanding*. Evaluasi dilakukan untuk mengukur sejauh mana model dapat mengklasifikasikan status akademik mahasiswa secara tepat dan

konsisten, khususnya dalam menghadapi permasalahan ketidakseimbangan kelas pada data akademik.

Penilaian performa model dilakukan dengan menggunakan beberapa metrik evaluasi klasifikasi, yaitu accuracy, precision, recall, dan F1-score, yang masing-masing memberikan sudut pandang berbeda terhadap kualitas prediksi model. Selain itu, confusion matrix digunakan untuk menganalisis distribusi prediksi benar dan salah pada setiap kategori status akademik, sehingga memungkinkan identifikasi jenis kesalahan klasifikasi yang masih sering terjadi. Penggunaan kombinasi metrik ini bertujuan untuk memperoleh evaluasi yang lebih komprehensif dan tidak hanya bergantung pada satu ukuran performa.

Proses evaluasi diterapkan secara konsisten pada seluruh model dan konfigurasi yang diuji dengan menggunakan skema stratified 5-fold cross-validation, sehingga hasil evaluasi mencerminkan performa model yang lebih stabil dan objektif. Nilai evaluasi yang diperoleh dari setiap fold kemudian dirata-ratakan untuk menghasilkan estimasi performa model secara keseluruhan. Pendekatan ini membantu mengurangi bias akibat pembagian data yang tidak seimbang dan memberikan gambaran yang lebih representatif terhadap kemampuan generalisasi model.

Selain evaluasi kuantitatif, tahap ini juga mencakup analisis interpretabilitas model menggunakan metode SHapley Additive Explanations (SHAP). Analisis SHAP digunakan untuk memahami kontribusi masing-masing fitur terhadap hasil prediksi model, sehingga memberikan wawasan mengenai faktor-faktor yang paling berpengaruh terhadap status akademik mahasiswa. Dengan demikian, tahap *Evaluation* tidak hanya berfungsi untuk memilih model dengan performa terbaik, tetapi juga untuk memastikan bahwa model yang dihasilkan bersifat transparan, dapat dipahami, dan relevan untuk mendukung pengambilan keputusan akademik.

3.2.6 Deployment

Tahap *deployment* bertujuan untuk mengimplementasikan model prediksi yang telah dilatih ke dalam sebuah sistem interaktif sehingga dapat digunakan secara langsung oleh pengguna. Pada penelitian ini, proses deployment dilakukan dengan memanfaatkan framework Gradio, yang memungkinkan pembuatan antarmuka berbasis web secara sederhana tanpa memerlukan konfigurasi server yang kompleks. Model machine learning terbaik yang dihasilkan dari tahap evaluasi disimpan dalam bentuk file model dan diintegrasikan ke dalam aplikasi Python sebagai fungsi prediksi utama.

Implementasi sistem dilakukan dengan membangun antarmuka menggunakan modul *gr.Interface()* pada Python. Antarmuka ini dirancang untuk menerima sejumlah input yang merepresentasikan data akademik mahasiswa, seperti Indeks Prestasi Kumulatif (IPK), jumlah SKS yang ditempuh, konsistensi kehadiran, serta variabel pendukung lain yang relevan dengan proses prediksi kinerja akademik. Data input tersebut kemudian diproses oleh fungsi prediksi untuk menghasilkan keluaran berupa kategori status akademik mahasiswa beserta nilai probabilitas prediksi dari masing-masing kelas.

Setelah konfigurasi antarmuka selesai, aplikasi dijalankan melalui lingkungan Python dan menghasilkan alamat URL lokal yang dapat diakses melalui peramban web. Melalui antarmuka ini, pengguna dapat memasukkan data akademik mahasiswa dan memperoleh hasil prediksi secara *real-time* tanpa harus berinteraksi langsung dengan kode program. Sistem secara otomatis melakukan preprocessing ringan pada input, menjalankan model prediksi, dan menampilkan hasil klasifikasi dalam bentuk yang mudah dipahami oleh pengguna.

Dalam kerangka metodologi CRISP-DM, tahap deployment ini berfungsi sebagai fase penerapan dan pengujian akhir, yang bertujuan untuk memastikan bahwa model yang dikembangkan dapat beroperasi secara stabil pada data baru serta menghasilkan keluaran yang konsisten. Selain sebagai sarana validasi fungsional, deployment juga berperan sebagai jembatan antara hasil penelitian eksperimental dan implementasi sistem pendukung keputusan akademik berbasis data. Dengan adanya antarmuka interaktif ini, hasil penelitian tidak hanya terbatas pada evaluasi model, tetapi juga dapat dimanfaatkan secara praktis untuk mendukung pemantauan dan pengambilan keputusan di lingkungan perguruan tinggi.

3.3. Teknik Pengumpulan Data

Teknik pengumpulan data merupakan tahapan penting dalam penelitian ini karena berperan dalam memastikan ketersediaan data yang akurat, relevan, dan sesuai dengan tujuan analisis. Data yang dikumpulkan menjadi dasar utama dalam proses pengembangan dan evaluasi model prediksi kinerja akademik mahasiswa. Oleh karena itu, metode pengumpulan data harus dilakukan secara sistematis dan terkontrol agar data yang diperoleh dapat merepresentasikan kondisi akademik mahasiswa secara nyata serta mendukung proses analisis berbasis data. Dalam penelitian ini, pengumpulan data dilakukan melalui pemanfaatan data akademik internal dan data pendukung lainnya yang diperoleh dari sumber resmi universitas. Proses pengumpulan data mencakup pengambilan, konsolidasi, serta verifikasi data dari berbagai sumber untuk memastikan kelengkapan dan konsistensinya sebelum digunakan pada tahap pengolahan dan pemodelan. Uraian lebih lanjut mengenai populasi, sampel, serta periode pengumpulan data dijelaskan pada subbab berikutnya.

3.3.1 *Populasi dan Sample*

Populasi dalam penelitian ini adalah seluruh mahasiswa Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Multimedia Nusantara

(UMN) angkatan 2020 hingga 2024 yang tercatat dalam sistem administrasi akademik universitas. Populasi ini mencakup mahasiswa dengan berbagai status akademik dan latar belakang performa studi. Dari populasi tersebut, dilakukan proses pemilihan sampel menggunakan teknik purposive sampling, yaitu dengan menetapkan kriteria tertentu yang sesuai dengan tujuan penelitian. Sampel penelitian dipilih dari mahasiswa yang telah memiliki status akademik akhir, sehingga mahasiswa yang masih berstatus aktif pada tahun ajaran terakhir tidak disertakan. Berdasarkan proses seleksi tersebut, diperoleh sebanyak 416 mahasiswa yang digunakan sebagai sampel penelitian dan dijadikan dataset utama dalam proses pelatihan serta pengujian model prediksi.

3.3.2 *Periode Pengumpulan Data*

Pengumpulan data dalam penelitian ini dilakukan secara bertahap dengan memanfaatkan data akademik historis mahasiswa Program Studi Sistem Informasi Universitas Multimedia Nusantara untuk periode angkatan 2020 hingga 2024. Data akademik diperoleh dari rekap per semester yang mencerminkan perkembangan performa mahasiswa selama masa studi, serta dilengkapi dengan data hasil kuesioner terkait faktor eksternal yang memengaruhi kinerja akademik. Proses pengumpulan data dilakukan setelah seluruh data akademik pada periode tersebut tersedia dan diverifikasi oleh pihak fakultas. Dengan cakupan periode yang mencerminkan beberapa angkatan dan tahapan studi, data yang digunakan diharapkan mampu merepresentasikan kondisi akademik mahasiswa secara komprehensif dan mendukung analisis prediksi kinerja akademik secara lebih akurat.

3.4. Teknik Analisis Data

Teknik analisis data dalam penelitian ini bertujuan untuk mengolah dan menganalisis data akademik mahasiswa guna mengidentifikasi pola serta hubungan

antarvariabel yang berpengaruh terhadap kinerja akademik. Proses analisis dilakukan secara terstruktur untuk mendukung pengembangan model prediksi berbasis machine learning yang mampu memberikan hasil prediksi secara akurat dan konsisten. Tahapan analisis mencakup proses persiapan data, pembangunan model, hingga evaluasi performa prediksi yang dihasilkan.

Pada tahap pemodelan, penelitian ini menerapkan pendekatan komparatif dengan memanfaatkan beberapa algoritma machine learning yang memiliki karakteristik pembelajaran berbeda, yaitu Light Gradient Boosting Machine (LightGBM), Random Forest, dan Multilayer Perceptron (MLP). Ketiga algoritma tersebut digunakan untuk membangun model prediksi kinerja akademik mahasiswa dan dibandingkan performanya guna memperoleh model yang paling optimal. Pendekatan ini memungkinkan analisis yang lebih objektif terhadap keunggulan masing-masing algoritma dalam menangani data akademik yang bersifat tabular dan kompleks.

Evaluasi kinerja model dilakukan menggunakan sejumlah metrik klasifikasi, yaitu accuracy, precision, recall, dan F1-score, serta didukung dengan analisis confusion matrix untuk melihat distribusi hasil prediksi pada setiap kategori. Penggunaan kombinasi metrik evaluasi ini bertujuan untuk memberikan gambaran kinerja model secara menyeluruh, baik dari sisi ketepatan prediksi maupun keseimbangan performa antar kelas.

Selain mengevaluasi performa model, penelitian ini juga menekankan aspek interpretabilitas melalui penerapan metode SHapley Additive Explanations (SHAP). Metode ini digunakan untuk menganalisis kontribusi masing-masing variabel terhadap hasil prediksi, sehingga memberikan pemahaman yang lebih transparan mengenai faktor-faktor yang memengaruhi kinerja akademik mahasiswa. Dengan demikian, hasil analisis tidak hanya bersifat prediktif, tetapi juga informatif dan dapat dimanfaatkan sebagai dasar pengambilan keputusan akademik.

Seluruh proses analisis data dalam penelitian ini dilakukan menggunakan bahasa pemrograman Python, yang dipilih karena memiliki ekosistem pustaka yang lengkap untuk analisis data, pemodelan machine learning, serta visualisasi hasil. Beberapa pustaka yang digunakan meliputi *scikit-learn*, *LightGBM*, dan *SHAP* untuk pemodelan dan interpretasi, serta *Matplotlib* untuk visualisasi. Alasan pemilihan Python dibandingkan dengan bahasa pemrograman lain seperti R dan Matlab disajikan secara ringkas pada Tabel 3.4, yang membandingkan keunggulan masing-masing bahasa dalam konteks analisis data akademik.

Tabel 3. 2 Perbandingan Bahasa Pemrograman

Sumber: [91]

| Karakteristik | Phyton | R | Matlab |
|---------------|---|---|---|
| Deskripsi | Python adalah bahasa pemrograman serbaguna dengan fokus pada keterbacaan kode dan dukungan komunitas yang luas. Awalnya dikembangkan untuk tujuan umum, kini banyak digunakan dalam analisis data dan pembelajaran mesin. | R adalah bahasa pemrograman yang dirancang khusus untuk analisis statistik dan visualisasi data. Sangat populer di kalangan akademisi dan praktisi statistik. | Matlab adalah lingkungan komputasi numerik yang dirancang untuk pemrosesan matriks, pengembangan algoritma, dan simulasi teknik. Ideal untuk keperluan teknik dan ilmiah. |
| Biaya | Gratis dan sumber terbuka karena <i>open-source</i> . | Gratis dan sumber terbuka karena <i>open-source</i> . | Berlisensi komersial, dengan biaya tinggi untuk lisensi individual atau institusional. |
| Penggunaan | Digunakan secara luas dalam analisis data, pembelajaran mesin, dan pengembangan perangkat lunak. | Umumnya digunakan untuk statistik, analisis data eksploratif, dan visualisasi. | Sering digunakan dalam teknik, simulasi, dan pengolahan data numerik yang memerlukan presisi |

| Karakteristik | Phyton | R | Matlab |
|---------------|---|---|--|
| | Mendukung berbagai aplikasi di industri, penelitian, dan pendidikan. | Dominan di lingkungan akademik dan penelitian. | tinggi, terutama di bidang teknik dan sains. |
| Library | Memiliki beragam pustaka untuk berbagai kebutuhan, mulai dari analisis data hingga pembelajaran mendalam. | Memiliki banyak pustaka statistik, yang memudahkan analisis data dan pembuatan visualisasi. | Memiliki pustaka bawaan untuk berbagai keperluan teknik dan sains, namun terbatas dibandingkan pustaka sumber terbuka. |

Berdasarkan Tabel 3.4, bahasa pemrograman Python dipandang paling sesuai untuk mendukung pelaksanaan penelitian ini karena kemampuannya dalam mengakomodasi berbagai algoritma *Explainable Artificial Intelligence* dalam satu lingkungan pengembangan yang terpadu. Python menyediakan beragam pustaka pendukung, seperti *scikit-learn*, *XGBoost*, *CatBoost*, *TensorFlow*, dan *PyTorch*, yang memungkinkan proses pelatihan model, pengujian kinerja, serta penyajian hasil analisis dilakukan secara efisien dan terstruktur. Selain itu, ketersediaan pustaka interpretabilitas seperti *SHAP* menjadi nilai tambah karena selaras dengan tujuan penelitian dalam memberikan penjelasan yang transparan terhadap hasil prediksi kinerja akademik mahasiswa. Kemudahan sintaks, dukungan dokumentasi yang luas, serta komunitas pengguna yang besar juga mempermudah proses pengembangan dan replikasi eksperimen, sehingga Python menjadi pilihan yang tepat untuk mendukung penelitian yang akurat, efektif, dan dapat direproduksi.

UNIVERSITAS
MULTIMEDIA
NUSANTARA