BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian ini berfokus pada data akademik internal mahasiswa Program Studi Sistem Informasi Universitas Multimedia Nusantara angkatan 2020 hingga 2024. Data tersebut mencakup variabel-variabel utama yang merepresentasikan kinerja akademik mahasiswa, antara lain Indeks Prestasi Kumulatif (IPK) per semester, jumlah Satuan Kredit Semester (SKS) yang ditempuh, hasil nilai mata kuliah, serta status akademik mahasiswa pada setiap periode studi. Data ini digunakan untuk membangun model prediksi yang bertujuan mengelompokkan mahasiswa ke dalam empat kategori kelulusan, yaitu *dropout*, *lulus tepat waktu*, *tidak lulus tepat waktu*, dan *lulus lebih awal*. Dengan demikian, penelitian ini diarahkan untuk memahami pola-pola akademik mahasiswa dan mengidentifikasi faktor-faktor internal yang berpengaruh terhadap keberhasilan studi.

3.2 Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif yang bertujuan untuk mengembangkan model prediksi kelulusan mahasiswa berdasarkan data akademik internal dengan memanfaatkan metode *machine learning*. Pendekatan kuantitatif dipilih karena mampu mengukur performa model secara objektif melalui penggunaan metrik evaluasi terstandar dan analisis berbasis data numerik. Melalui pendekatan ini, hasil penelitian dapat disajikan secara terukur, sistematis, dan dapat diulang (*reproducible*), sesuai dengan prinsip dasar penelitian ilmiah dalam bidang analitik data.

Penelitian ini berfokus pada pengembangan model prediksi kelulusan mahasiswa yang mengklasifikasikan mahasiswa ke dalam empat kategori, yaitu *dropout*, *lulus tepat waktu*, *tidak lulus tepat waktu*, dan *lulus lebih awal*. Model dibangun menggunakan pendekatan Hybrid Ensemble, yang mengombinasikan beberapa algoritma *machine learning* dengan karakteristik pembelajaran yang berbeda, yaitu XGBoost, LightGBM, CatBoost, Random Forest dan Multilayer

Perceptron (MLPClassifier). Kombinasi ini dipilih karena dapat menggabungkan keunggulan model berbasis *tree boosting* yang efisien dalam memproses data tabular, dengan model berbasis *neural network* yang mampu menangkap hubungan non-linear antar variabel akademik.

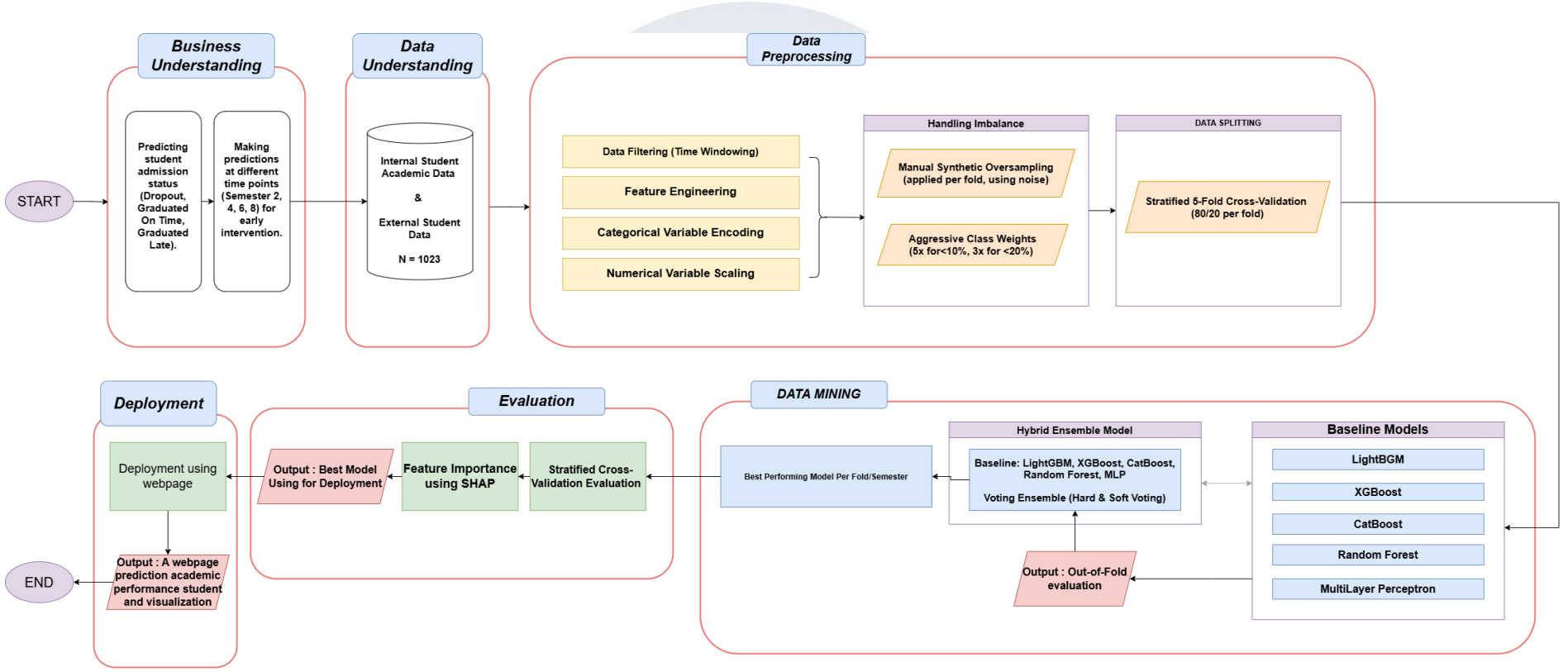
Selain fokus pada akurasi prediksi, penelitian ini juga menekankan aspek interpretabilitas model melalui penerapan metode SHapley Additive Explanations (SHAP). Dengan adanya SHAP, hasil prediksi tidak hanya menghasilkan klasifikasi semata, tetapi juga menjelaskan kontribusi dari setiap fitur akademik terhadap keputusan model. Pendekatan ini memastikan bahwa hasil penelitian dapat digunakan tidak hanya sebagai alat analitik, tetapi juga sebagai bahan pertimbangan strategis bagi pihak universitas dalam melakukan intervensi terhadap mahasiswa yang berpotensi mengalami keterlambatan studi atau dropout.

Secara umum, penelitian ini dilaksanakan berdasarkan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining) yang terdiri dari enam tahapan utama, yaitu *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*. Kerangka ini digunakan untuk memastikan seluruh proses penelitian dilakukan secara sistematis, mulai dari perumusan masalah hingga interpretasi hasil. Tahapan-tahapan dalam CRISP-DM serta hubungan antar proses akan dijelaskan lebih lanjut pada bagian berikutnya bersamaan dengan Gambar 3.1 Alur Penelitian.

Dengan menggunakan kombinasi metode *hybrid ensemble* dan pendekatan *explainable AI*, penelitian ini diharapkan dapat menghasilkan model prediksi kelulusan mahasiswa yang tidak hanya akurat dan efisien, tetapi juga transparan dan mudah diinterpretasikan. Hasil dari penelitian ini diharapkan mampu membantu perguruan tinggi dalam melakukan deteksi dini terhadap mahasiswa berisiko, mendukung proses pembimbingan akademik yang lebih efektif, serta menjadi landasan bagi pengembangan sistem prediktif dalam pengelolaan pendidikan tinggi berbasis data.

3.2.1 Alur Penelitian

Alur penelitian berperan sebagai pedoman untuk memastikan proses penelitian berlangsung secara terencana dan terstruktur Gambar 3.1 menampilkan alur penelitian yang digunakan dalam studi ini, meliputi tahapan penting mulai dari pemilihan data (*selection*), *preprocessing*, transformasi data (*data transformation*), pemodelan (*modeling*), hingga evaluasi (*evaluation*) Penyusunan alur ini bertujuan untuk meminimalkan potensi kesalahan serta menghasilkan temuan yang valid dan relevan bagi penelitian.



Gambar 3. 1 Alur Penelitian

Gambar 3.1 menunjukkan alur penelitian yang digunakan dalam pengembangan model prediksi kelulusan mahasiswa berdasarkan data akademik internal menggunakan pendekatan Hybrid Ensemble. Alur penelitian ini disusun berdasarkan kerangka kerja CRISP-DM (Cross Industry Standard Process for Data Mining) yang terdiri atas enam tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preprocessing*, *Data Mining*, *Evaluation*, dan *Deployment*. Setiap tahapan dalam kerangka ini saling berhubungan secara iteratif, di mana hasil dari satu tahap dapat digunakan untuk memperbaiki atau menyempurnakan tahap lainnya, sehingga proses penelitian dapat berjalan secara sistematis dan terarah.

Tahap Business Understanding berfokus pada penentuan tujuan penelitian, yaitu memprediksi status kelulusan mahasiswa ke dalam empat kategori (dropout, lulus tepat waktu, tidak lulus tepat waktu, dan lulus lebih awal) serta mendukung proses intervensi akademik lebih dini. Tahap Data Understanding dilakukan untuk meninjau struktur dataset, karakteristik fitur akademik mahasiswa, serta identifikasi potensi ketidakseimbangan kelas.

Tahap Data Preprocessing mencakup pembersihan data, feature engineering, encoding, normalisasi, serta penanganan ketidakseimbangan kelas menggunakan kombinasi manual synthetic oversampling dan class weights. Setelah itu, data dibagi menggunakan Stratified 5-Fold Cross-Validation agar evaluasi model lebih adil dan representatif.

Tahap *Data Mining* merupakan inti dari proses penelitian, di mana beberapa model *baseline* dilatih, yaitu LightGBM, XGBoost, CatBoost, Random Forest, dan Multilayer Perceptron (MLP). Pada tahap perancangan metodologi, penelitian ini juga melakukan eksplorasi terhadap beberapa pendekatan *ensemble learning*, termasuk pendekatan stacking dengan meta-learner Logistic Regression, sebagai bagian dari eksperimen awal untuk mengombinasikan prediksi model dasar dan mengevaluasi potensi peningkatan performa melalui mekanisme *meta-learning*. Namun, pendekatan *stacking* tersebut tidak ditetapkan sebagai metode utama dan hanya berfungsi sebagai tahapan eksploratif dalam pengembangan model. Implementasi metodologi yang digunakan secara konsisten pada seluruh eksperimen utama dalam penelitian ini

adalah Hybrid Ensemble Voting, yang mencakup pendekatan Hard Voting dan Soft Voting, guna memperoleh model prediksi yang lebih stabil dan akurat.

Pada tahap Evaluation, performa model dinilai menggunakan metrik klasifikasi seperti akurasi, precision, recall, F1-score, serta confusion matrix. Selain itu, interpretasi model dilakukan menggunakan SHAP untuk memahami kontribusi setiap fitur dalam prediksi.

Tahap terakhir adalah Deployment, di mana model terbaik dari Voting Ensemble diimplementasikan melalui aplikasi berbasis web agar dapat digunakan oleh pihak akademik sebagai alat bantu evaluasi dan pengambilan keputusan. Alur penelitian ini bersifat iteratif, sehingga setiap tahapan dapat dikaji ulang untuk meningkatkan performa model. Berikut disajikan penjelasan pada masing masing tahapan antara lain:

1.4.2.3 Business Understanding

Tahap pertama yaitu Business Understanding berfokus pada identifikasi masalah dan perumusan tujuan penelitian. Penelitian dilatarbelakangi untuk melakukan prediksi status kelulusan mahasiswa meliputi empat kategori yaitu dropout, lulus tepat waktu, tidak lulus tepat waktu, dan lulus lebih awal pada beberapa titik waktu studi (semester 2, 4, 6, dan 8). Tujuan utama dari tahap ini adalah memberikan pemahaman yang jelas terhadap kebutuhan akademik dan strategi pencegahan yang berisiko mengalami keterlambatan atau tidak menyelesaikan studi. Pemahaman yang baik terhadap permasalahan menjadi dasar dalam menentukan arah analisis dan strategi pemodelan yang digunakan pada tahap berikutnya. Pada tahap ini juga ditentukan rumusan masalah penelitian, yaitu: (1) bagaimana performa model individual (XGBoost, LightGBM, CatBoost, Random Forest, dan MLPClassifier) dalam memprediksi status kelulusan mahasiswa, (2) apakah metode hybrid ensemble dapat meningkatkan akurasi dibandingkan model individual, dan (3) faktor-faktor apa yang paling berpengaruh terhadap prediksi status kelulusan berdasarkan interpretasi SHAP. Selain itu, pada tahap ini juga dilakukan studi literatur untuk memahami penelitian-penelitian terdahulu

yang relevan, serta menentukan metrik evaluasi yang sesuai seperti accuracy, precision, recall, F1-score, dan confusion matrix untuk mengukur keberhasilan model. Tahap Business Understanding ini menjadi fondasi yang krusial karena menentukan keseluruhan alur penelitian, mulai dari pemilihan data, teknik preprocessing, hingga pemilihan algoritma machine learning yang akan digunakan.

1.4.2.4 Data Understanding

Tahap Data Understanding berfungsi untuk memahami karakteristik data yang akan digunakan dalam penelitian. Pada tahap ini, dilakukan eksplorasi terhadap dataset akademik internal mahasiswa Program Studi Sistem Informasi Universitas Multimedia Nusantara dengan jumlah total sebanyak 1.023 entri serta data kuesioner faktor eksternal yang mempengaruhi kinerja akademik mahasiswa. Data internal yang dikumpulkan mencakup variabel-variabel utama seperti Indeks Prestasi Kumulatif (IPK) per semester, jumlah SKS yang ditempuh, nilai mata kuliah, dan status akademik mahasiswa, sedangkan data kuesioner mencakup variabel-variabel utama seperti dukungan keluarga terhadap akademik, dukungan finansial dari keluarga, dukungan keluarga terhadap kepribadian/jurusan yang dipilih, dan pilihan jurusan sesuai keinginan. Selain itu, dilakukan pula analisis awal terhadap distribusi data untuk mengidentifikasi adanya nilai kosong (missing values) serta memeriksa keseimbangan antar kelas pada variabel target (status kelulusan). Hasil analisis menunjukkan bahwa terdapat ketidakseimbangan data yang signifikan, terutama pada kelas minoritas seperti "Lulus Lebih Awal" dan "Tidak Lulus Tepat Waktu", yang masing-masing hanya memiliki proporsi kecil dari total dataset. Proses eksplorasi data juga meliputi analisis statistik deskriptif untuk memahami distribusi nilai, rentang data, dan korelasi antar variabel menggunakan visualisasi seperti histogram, boxplot, dan correlation heatmap. Tahap ini juga mengidentifikasi fitur-fitur yang berpotensi penting untuk prediksi, seperti nilai rata-rata (nilai_avg), total semester yang ditempuh (TOTAL_SEMESTER), konsistensi kehadiran

(hadir_consistency), dan variabel-variabel behavioral lainnya yang dapat menjadi indikator kuat terhadap status kelulusan mahasiswa.

Dalam penelitian ini, penanganan potensi bias kuesioner diutamakan melalui beberapa strategi metodologis, terutama karena data *self-report* rentan terhadap *nonresponse bias* dan *social desirability bias*. Untuk mengurangi bias representatif, distribusi responden dibandingkan dengan populasi akademik yang relevan berdasarkan atribut seperti angkatan dan tingkat progres, sehingga dapat diidentifikasi apakah terdapat kesenjangan representasi yang memengaruhi interpretasi hasil; pendekatan semacam ini direkomendasikan dalam survei pendidikan untuk menilai kelayakan data *self-report* sebelum digunakan dalam analisis lanjut [62]. Desain kuesioner dibuat anonim dan tidak mengevaluatif, serta diadministrasikan tanpa konsekuensi akademik, karena studi empiris menunjukkan bahwa anonimitas secara signifikan mengurangi *social desirability bias* dalam respons *self-report* mahasiswa [63]. Peran variabel *self-report* dibatasi pada analisis deskriptif dan interpretasi berbasis SHAP, bukan sebagai fitur utama dalam pemodelan prediktif, sehingga potensi bias persepsi tidak secara langsung memengaruhi keakuratan model; pendekatan ini sejalan dengan praktik dalam *learning analytics* terbaru yang memisahkan data subjektif dari fitur prediktif inti [64]. Dengan kombinasi desain kuesioner yang mempertimbangkan anonimitas, evaluasi representativitas responden, serta pembatasan penggunaan *self-reports* dalam model, penelitian ini berupaya meminimalkan dampak bias kuesioner secara metodologis, sesuai dengan temuan studi kasus nyata pada literatur 2020 -2025.

Selain pemahaman terhadap kualitas dan karakteristik data, tahap *Data Understanding* juga mencakup validasi pemberian label target yang digunakan dalam penelitian. Validasi pemberian label untuk setiap kategori S2, S4, dan S6 tidak didasarkan secara langsung pada wawancara dengan BIA, melainkan dilakukan melalui pendekatan administratif dan akademik yang mengacu pada ketentuan resmi universitas. Berdasarkan *Handbooks Information System Study Program Curriculum Guidebook Universitas*

Multimedia Nusantara, struktur masa studi mahasiswa dirancang untuk diselesaikan dalam delapan semester pada jalur reguler dengan opsi fast-track tujuh semester, serta ketentuan kelulusan yang bergantung pada pemenuhan beban SKS dan penyelesaian tugas akhir. Meskipun dokumen tersebut tidak secara eksplisit mendefinisikan label S2, S4, dan S6, informasi mengenai durasi studi dan mekanisme kelulusan digunakan sebagai dasar dalam pemetaan kategori hasil studi, yaitu lulus lebih awal, lulus tepat waktu, tidak lulus tepat waktu, dan dropout. Penetapan label tersebut kemudian dikonfirmasi dan divalidasi oleh dosen pembimbing penelitian, yaitu Bapak Iwan Prasetiawan, S.Kom., M.M., Ibu Ririn Ikana Desanti, S.Kom., M.Kom., dan Ibu Suryasari, S.Kom., M.T., agar selaras dengan kebijakan akademik UMN dan konteks penelitian, sehingga validitas label yang digunakan dapat dipertanggungjawabkan secara akademik. Secara keseluruhan, tahap *Data Understanding* memberikan landasan yang kuat bagi tahap selanjutnya, khususnya dalam menentukan strategi *data preprocessing* dan pemodelan yang sesuai. Pemahaman mendalam terhadap karakteristik data, distribusi kelas, potensi bias, serta relevansi fitur memastikan bahwa data yang digunakan tidak hanya valid secara statistik, tetapi juga dapat dipertanggungjawabkan secara metodologis dalam konteks penelitian *learning analytics* dan *educational data mining*.

1.4.2.5 Data Preprocessing

Tahap *Data Preprocessing* merupakan tahap penting dalam mempersiapkan data sebelum digunakan dalam pemodelan *machine learning*. Pada tahap ini dilakukan penanganan *missing values* melalui penghapusan atau imputasi sesuai dengan karakteristik masing-masing variabel. Variabel kategorikal kemudian ditransformasikan ke dalam bentuk numerik menggunakan Label Encoder agar dapat diproses oleh algoritma pembelajaran mesin. Untuk variabel target, kategori “Masih Aktif” dikeluarkan dari data latih karena belum memiliki label kelulusan

akhir, sehingga berpotensi menimbulkan ambiguitas dalam proses supervisi model.

Selanjutnya, dilakukan proses feature engineering untuk membentuk fitur-fitur turunan yang lebih representatif dibandingkan fitur mentah. Proses ini bertujuan untuk mengekstraksi informasi akademik laten yang tidak dapat ditangkap secara langsung oleh variabel asli. Melalui feature engineering, jumlah fitur meningkat dari 15 fitur awal menjadi 45 fitur, yang mencerminkan berbagai aspek performa akademik mahasiswa secara longitudinal, termasuk rata-rata capaian, tren perkembangan, stabilitas nilai, intensitas beban studi, serta interaksi antar variabel akademik.

Tabel 3. 1 Fitur Turunan Hasil Feature Engineering

No	**Nama Fitur**	**Makna Utama**	**Rumus**
1	avg_ips	Performa akademik rata-rata mahasiswa sepanjang studi	avg_ips = mean(IPS)
2	ips_trend	Arah perkembangan performa akademik (meningkat atau menurun)	ips_trend = IPS_terakhir − IPS_pertama
3	ips_std	Stabilitas nilai IPS antar semester	ips_std = std(IPS)
4	sks_per_semester	Intensitas beban studi per semester	sks_per_semester = TOTAL_SKS / TOTAL_SEMESTER
5	ipk_sks_interaction	Interaksi antara capaian akademik dan beban studi	ipk_sks_interaction = IPK × sks_per_semester

Peningkatan jumlah fitur dari 15 fitur awal menjadi 45 fitur dilakukan melalui proses *feature engineering* untuk mengekstraksi informasi akademik yang tidak dapat ditangkap oleh fitur mentah secara langsung. Lima contoh fitur turunan yang paling penting, sebagaimana ditunjukkan pada Tabel 3.x, meliputi avg_ips, yang dihitung sebagai rata-rata nilai IPS seluruh semester *(avg_ips = mean(IPS))* untuk merepresentasikan performa akademik jangka panjang; ips_trend, yang mengukur arah perubahan performa akademik mahasiswa dari awal hingga akhir studi *(ips_trend = IPS_terakhir − IPS_pertama)*; ips_std, yang merepresentasikan stabilitas nilai IPS antar semester melalui simpangan baku *(ips_std = std(IPS))*; sks_per_semester, yang menggambarkan intensitas beban studi mahasiswa *(sks_per_semester = TOTAL_SKS / TOTAL_SEMESTER)*; serta ipk_sks_interaction, yaitu fitur interaksi yang menangkap hubungan antara capaian akademik dan beban studi *(ipk_sks_interaction = IPK × sks_per_semester)*. Penambahan fitur-fitur turunan ini memungkinkan model mempelajari performa akademik mahasiswa secara lebih komprehensif, mencakup aspek rata-rata, tren, stabilitas, serta interaksi antar variabel, sehingga meningkatkan kemampuan model dalam mengenali pola risiko akademik yang kompleks.

Selain feature engineering, penanganan ketidakseimbangan kelas dilakukan menggunakan pendekatan manual synthetic oversampling, yaitu dengan menambahkan sejumlah sampel baru pada kelas minoritas melalui penambahan *noise* terkontrol pada sampel asli. Pemilihan teknik oversampling ini didasarkan pada hasil uji coba empiris yang telah dilakukan, di mana metode SMOTE (Synthetic Minority Over-sampling Technique) juga dievaluasi sebagai pembanding. Hasil pengujian menunjukkan bahwa penerapan SMOTE pada data latih tidak menghasilkan peningkatan performa yang konsisten, serta pada beberapa konfigurasi justru meningkatkan variansi model dan indikasi *overfitting*, terutama akibat terbentuknya sampel sintetis yang kurang representatif pada kelas minoritas yang sangat terbatas. Temuan ini sejalan dengan penelitian

terdahulu yang menyatakan bahwa SMOTE berpotensi menghasilkan sampel sintetis yang mengaburkan batas keputusan kelas dan menurunkan kemampuan generalisasi model, khususnya pada dataset dengan distribusi kompleks dan ketidakseimbangan ekstrem [65], [66]. Sebaliknya, beberapa studi komparatif terbaru melaporkan bahwa manual synthetic oversampling dapat memberikan performa yang lebih stabil dan kompetitif dibandingkan SMOTE, terutama ketika dikombinasikan dengan class weighting untuk menekan dominasi kelas mayoritas tanpa mendistorsi ruang fitur asli [67], [68]. Oleh karena itu, penelitian ini secara eksplisit memilih random oversampling sebagai pendekatan utama dalam tahap preprocessing karena lebih sederhana, lebih terkontrol, dan secara empiris lebih *robust* terhadap karakteristik data yang digunakan.

Proses oversampling dilakukan di dalam setiap fold pada stratified 5-fold cross-validation, sehingga sampel hasil augmentasi hanya berada pada data pelatihan dalam fold tersebut dan tidak pernah muncul pada data validasi. Dengan mekanisme ini, proses pelatihan model terhindar dari risiko *data leakage*. Selain oversampling, beberapa algoritma juga memanfaatkan class weight untuk memperkuat penanganan ketidakseimbangan kelas tanpa menambah sampel baru.

Tahap preprocessing juga mencakup standardisasi fitur numerik menggunakan StandardScaler, terutama untuk algoritma yang sensitif terhadap perbedaan skala antar fitur, seperti Multilayer Perceptron (MLP). Melalui keseluruhan rangkaian proses preprocessing ini, data yang digunakan dalam pelatihan model menjadi lebih bersih, seimbang, stabil, dan siap mendukung evaluasi model yang valid, reliabel, serta bebas dari bias akibat kebocoran data.

1.4.2.6 Data Mining

Tahap *Data Mining* merupakan bagian inti dari penelitian ini, di mana algoritma *machine learning* diterapkan untuk membangun model

prediksi status kelulusan mahasiswa. Pada tahap ini digunakan lima model individual, yaitu LightGBM, XGBoost, CatBoost, Random Forest, dan MLPClassifier, dengan masing-masing model dikonfigurasi menggunakan parameter dasar atau penyesuaian ringan sesuai kebutuhan karakteristik dataset. Proses pengembangan model dilakukan secara bertahap melalui tiga *model configuration* dengan tingkat kompleksitas yang berbeda untuk mengevaluasi kestabilan model, pengaruh parameter, serta kesesuaian model terhadap kondisi data yang memiliki ketidakseimbangan kelas. Pendekatan ini memungkinkan peneliti membandingkan performa antar konfigurasi secara sistematis dan memilih konfigurasi terbaik yang konsisten untuk implementasi akhir. Perbandingan dari ketiga model configuration tersebut ditunjukkan pada Tabel 3.2.

Tabel 3. 2 Perbandingan Percobaan Hyperparameter Tuning

Aspek	Model Configuration 1	Model Configuration 2	Model Configuration 3 (Terbaik)
Strategi	*Conservative* (S2) / *Moderate* (S4/6/8)	*Ultra-Optimized*	*Default Configuration*
Kompleksitas	Rendah - Menengah	Sangat Tinggi	Rendah
Jumlah Model	4 (LGBM, XGB, CatBoost, MLP)	5 (+ Random Forest)	5 (+ Random Forest)
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		Cepat

		Lambat (banyak iterasi & parameter besar)	
Performance	Bagus	Baik	Stabil dan Kompetitif (terbaik untuk small dataset)
Use Case	Experimentation / Cross-Validation	Production Ready	Deployment

Tabel 3.2 membandingkan tiga model configuration yang diterapkan pada tahap Data Mining dengan tingkat kompleksitas parameter yang berbeda. Model Configuration 1 menggunakan pengaturan parameter yang sederhana dengan empat algoritma utama (LightGBM, XGBoost, CatBoost, dan MLP), sehingga waktu pelatihan relatif cepat dan berfungsi sebagai baseline awal. Model Configuration 2 menggunakan parameter yang lebih kompleks serta menambahkan Random Forest untuk meningkatkan keragaman model, namun memerlukan waktu pelatihan lebih lama dan menunjukkan sensitivitas lebih tinggi terhadap ukuran dataset. Sementara itu, Model Configuration 3 merupakan konfigurasi final yang digunakan dalam implementasi kode, memanfaatkan parameter dasar dengan penyesuaian minimal serta manual synthetic oversampling di dalam setiap fold untuk mencegah data leakage. Berdasarkan hasil perbandingan, peningkatan kompleksitas parameter tidak selalu menghasilkan peningkatan performa yang signifikan, dan konfigurasi yang lebih sederhana justru memberikan hasil yang lebih stabil dan sesuai untuk deployment pada dataset berukuran kecil.

1.4.2.7 Evaluation

Tahap Evaluation dilakukan untuk menilai performa model dalam memprediksi status kelulusan mahasiswa berdasarkan data akademik dan faktor eksternal. Pada tahap ini, model diuji menggunakan metrik evaluasi yang umum digunakan dalam penelitian klasifikasi, yaitu accuracy, precision, recall, F1-score, dan confusion matrix, dengan teknik stratified 5-fold cross-validation yang memastikan bahwa proses evaluasi berlangsung

secara adil dan konsisten dengan mempertahankan proporsi kelas pada setiap fold, terutama karena adanya ketidakseimbangan kelas dalam dataset. Evaluasi dilakukan tidak hanya pada setiap baseline model (LightGBM, XGBoost, CatBoost, Random Forest, MLP) secara individual, tetapi juga pada model Hybrid Ensemble, sehingga performa setiap model dapat dibandingkan secara objektif untuk menentukan model terbaik. Hasil evaluasi menunjukkan bahwa pada Semester 2 (binary classification), semua model individual mencapai performa yang sangat baik dengan F1-score macro berkisar 89.75%-92.70%, di mana XGBoost menunjukkan performa terbaik dengan F1-score 92.70% dan akurasi 94%, diikuti oleh LightGBM dengan F1-score 91.64%. Pada Semester 4 (3-class classification), terjadi penurunan performa yang signifikan dengan F1-score macro turun menjadi sekitar 74%-80%, di mana XGBoost dan LightGBM tetap unggul dengan F1-score sekitar 80%, namun semua model mengalami kesulitan dalam memprediksi kelas minoritas "Tidak Lulus Tepat Waktu" dengan F1-score hanya sekitar 50-54%. Pada Semester 6 (4-class classification), performa turun lebih lanjut dengan F1-score macro berkisar 58%-72%, di mana LightGBM mencatat F1-score 72% dan XGBoost 70%, sementara model Hybrid Ensemble berhasil mencapai F1-score macro 73% dengan akurasi 80%, mengungguli semua model individual dan menunjukkan peningkatan yang signifikan terutama pada kelas minoritas seperti "Lulus Lebih Awal" dan "Tidak Lulus Tepat Waktu". Analisis confusion matrix pada setiap semester menunjukkan bahwa mayoritas kesalahan prediksi terjadi pada kelas-kelas yang memiliki overlap fitur tinggi, seperti antara "Lulus Lebih Awal" dan "Lulus Tepat Waktu", serta antara "Tidak Lulus Tepat Waktu" dan "Lulus Tepat Waktu". Evaluasi juga mencakup perbandingan dengan tiga penelitian terdahulu yang relevan, menunjukkan bahwa penelitian ini lebih kuat dalam hal kompleksitas multi-stage prediction, penanganan data imbalance yang comprehensive, dan implementasi SHAP analysis untuk interpretability, dengan pencapaian performa yang kompetitif atau superior meskipun menghadapi dataset yang lebih challenging dengan ketidakseimbangan ekstrem.

1.4.2.8 Deployment

Tahap Deployment merupakan tahap akhir dalam kerangka CRISP-DM di mana model terbaik yang telah dilatih dan dievaluasi diimplementasikan untuk penggunaan praktis dalam sistem early warning akademik. Pada tahap ini, model DEPLOYMENT yang telah dioptimalkan dengan konfigurasi ultra-optimized hyperparameters (n_estimators=800, max_depth=12, learning_rate=0.015) dipilih sebagai best model karena terbukti memberikan performa terbaik dan paling robust dalam menangani kompleksitas multi-class classification dengan extreme imbalance. Model final disimpan dalam format yang dapat digunakan kembali menggunakan library pickle atau joblib untuk serialization, sehingga dapat di-load dan digunakan untuk prediksi pada data baru tanpa perlu melatih ulang model dari awal. Selain model prediksi, tahap deployment juga mencakup implementasi SHAP (SHapley Additive Explanations) analysis untuk memberikan interpretability terhadap prediksi model, yang mengidentifikasi top 20 fitur paling berpengaruh dengan TOTAL_SEMESTER, ANGKATAN, nilai_avg, TOTAL_SKS, dan hadir_consistency sebagai lima faktor teratas yang mempengaruhi prediksi status kelulusan. Berdasarkan hasil SHAP analysis, disusun strategi intervensi bertingkat (tier-based intervention) yang mencakup: (1) Tier 1 - High Priority Monitoring untuk top 5 features seperti flagging mahasiswa dengan semester >8, alert untuk GPA <2.5, monitoring pace <15 SKS/semester, dan tracking penurunan kehadiran >20%; (2) Tier 2 - Medium Priority Intervention seperti academic counseling untuk mahasiswa dengan performa tidak konsisten dan balance program untuk keterlibatan organisasi ekstrem; serta (3) Tier 3 - Background Support seperti financial aid dan peer support program untuk mahasiswa dari background socioeconomic yang kurang menguntungkan. Implementasi sistem early warning ini dirancang untuk dapat diintegrasikan dengan sistem informasi akademik universitas, di mana prediksi dilakukan secara otomatis pada setiap akhir semester untuk mengidentifikasi mahasiswa yang berisiko

dan memberikan rekomendasi intervensi yang sesuai kepada akademik advisor atau dosen pembimbing akademik. Model deployment juga dilengkapi dengan dashboard monitoring yang menampilkan visualisasi prediksi, distribusi risiko mahasiswa per kelas, feature importance, dan historical trend untuk membantu stakeholder dalam pengambilan keputusan. Tahap deployment ini memastikan bahwa hasil penelitian tidak hanya berhenti pada level akademis, tetapi dapat memberikan dampak praktis yang nyata dalam meningkatkan kualitas pendidikan dan mengurangi tingkat dropout mahasiswa melalui sistem deteksi dini dan intervensi yang tepat waktu.

3.2.2 Metode Data Mining

Metode data mining dalam penelitian ini berfokus pada proses penggalian pola dan pengetahuan tersembunyi dari data akademik mahasiswa untuk tujuan prediksi kelulusan. Pendekatan ini dilakukan dengan menerapkan teknik analisis dan pemodelan prediktif yang mampu mengidentifikasi hubungan kompleks antara variabel akademik, seperti Indeks Prestasi Kumulatif (IPK), jumlah SKS, serta status perkuliahan mahasiswa. Melalui metode ini, data mentah diolah secara bertahap mulai dari pemahaman dan eksplorasi data, pembersihan dan transformasi, hingga tahap pemodelan dan evaluasi hasil prediksi. Dalam penelitian ini, diterapkan pendekatan Hybrid Ensemble yang mengombinasikan beberapa algoritma seperti LightGBM, XGBoost, CatBoost, Random Forest, dan Multilayer Perceptron (MLP) untuk meningkatkan akurasi dan ketahanan model terhadap variasi data. Dalam bidang *data mining*, terdapat beberapa kerangka kerja yang dapat digunakan, antara lain Knowledge Discovery in Databases (KDD), Cross-Industry Standard Process for Data Mining (CRISP-DM), dan Sample, Explore, Modify, Model, and Access (SEMMA). Masing-masing metode ini memberikan kerangka kerja sistematis mulai dari pemahaman data, pengolahan, pemodelan, hingga evaluasi, yang membantu memastikan proses penelitian berjalan terstruktur. Perbandingan ketiga kerangka kerja tersebut disajikan pada Tabel 3.3.

Tabel 3. 3 Perbandingan ketiga kerangka kerja

Karakteristik	KDD	CRISP-DM	SEMMA
Tujuan Utama	Penemuan pola dan pengetahuan baru dalam dataset akademik secara mendalam untuk memahami faktor-faktor yang memengaruhi kinerja mahasiswa.	Penerapan proses standar yang fleksibel untuk beragam kebutuhan analisis, termasuk evaluasi kinerja akademik.	Fokus pada pembangunan model prediktif dengan tahapan eksplorasi dan modifikasi data yang minimal.
Tahapan Utama	Seleksi data, pra-pemrosesan, transformasi, penambangan data, dan evaluasi hasil.	Pemahaman tujuan analisis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi.	Sampling, eksplorasi data, modifikasi data, pembangunan model, dan penilaian hasil.
Fokus Framework	Menekankan eksplorasi data mendalam, analisis hubungan antar variabel, dan penemuan pola kompleks yang relevan dengan kinerja akademik mahasiswa.	Berorientasi pada siklus analitik terstruktur yang dapat diterapkan pada berbagai kasus bisnis dan pendidikan.	Lebih terfokus pada pembangunan model cepat tanpa eksplorasi mendalam terhadap karakteristik data akademik.
Kesesuaian dengan Penelitian Ini	Sangat sesuai karena mendukung eksplorasi dan integrasi fitur kompleks dari berbagai sumber data akademik untuk mengoptimalkan kinerja algoritma XAI.	Relatif sesuai, namun kurang menekankan eksplorasi mendalam yang diperlukan untuk penggabungan fitur kompleks.	Kurang sesuai karena lebih mengutamakan model prediktif cepat dibandingkan proses eksplorasi yang komprehensif.

Fleksibilitas dalam Penelitian Akademik	Tinggi, dapat disesuaikan dengan kebutuhan penelitian berbasis eksplorasi data dan pengujian berbagai algoritma pembelajaran mesin.	Cukup tinggi, tetapi lebih optimal untuk proyek praktis yang membutuhkan standar industri.	Rendah, karena lebih mengutamakan hasil cepat daripada proses penelitian yang mendalam.
Keunggulan Utama	Mendukung analisis data akademik yang kompleks, integrasi berbagai sumber data, dan pengujian model inovatif untuk prediksi kinerja mahasiswa.	Memberikan kerangka kerja terstandar yang memudahkan implementasi proyek dengan tahapan jelas.	Cepat dalam menghasilkan model prediktif sederhana tanpa memerlukan eksplorasi data yang rumit.

Berdasarkan Tabel 3.3, kerangka kerja CRISP-DM (Cross Industry Standard Process for Data Mining) dinilai paling sesuai untuk penelitian ini karena mampu memberikan alur proses analisis yang sistematis, terstruktur, dan iteratif dalam pengembangan model prediksi kelulusan mahasiswa. CRISP-DM terdiri atas enam tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Setiap tahap saling terhubung dan dapat diulang untuk memastikan hasil akhir yang akurat dan relevan dengan tujuan penelitian. Kerangka ini dipilih karena sejalan dengan fokus penelitian yang bertujuan membangun model prediksi berbasis *Hybrid Ensemble* dengan mengombinasikan algoritma LightGBM, XGBoost, CatBoost, Random Forest, dan Multilayer Perceptron (MLP). Melalui penerapan CRISP-DM, proses analisis dilakukan secara menyeluruh mulai dari identifikasi kebutuhan institusi, eksplorasi data akademik mahasiswa, pembersihan dan transformasi data, pembangunan serta evaluasi model, hingga penerapan model ke dalam sistem berbasis web yang aplikatif. Pendekatan ini memastikan bahwa model yang dihasilkan tidak hanya memiliki

tingkat akurasi tinggi, tetapi juga dapat digunakan secara langsung sebagai sistem pendukung keputusan akademik yang transparan dan berorientasi pada peningkatan efektivitas pembimbingan mahasiswa.

3.3 Teknik Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari hasil rekap akademik yang bersumber langsung dari Fakultas Ilmu Komputer Universitas Multimedia Nusantara (UMN). Data tersebut mencakup informasi akademik mahasiswa Program Studi Sistem Informasi angkatan 2020 hingga 2024 dan data hasil kuesioner faktor eksternal yang mempengaruhi prestasi kinerja mahasiwa, yang digunakan sebagai dasar dalam pengembangan model prediksi kelulusan mahasiswa. Dataset ini dikumpulkan melalui beberapa lembar kerja (sheet) yang berbeda, di mana masing-masing sheet merepresentasikan data akademik per semester serta variabel-variabel pendukung lain yang relevan untuk proses analisis.

Proses pengumpulan data dilakukan secara internal dengan izin dari pihak fakultas, sehingga data yang diperoleh bersifat primer, karena bersumber dari sistem administrasi akademik resmi universitas dan bersumber dari hasil kuesioner yang disebar. Setelah dilakukan proses konsolidasi dari beberapa sheet, total data awal mencakup 1.023 entri mahasiswa. Namun, untuk memastikan konsistensi dan relevansi dengan tujuan penelitian, dilakukan proses penyaringan (filtering) dengan mengecualikan mahasiswa yang masih berstatus *"aktif"* pada tahun ajaran terakhir. Hasil akhir dari proses seleksi ini menghasilkan 416 mahasiswa yang digunakan sebagai dataset utama dalam tahap pelatihan dan pengujian model.

Data yang dikumpulkan bersifat tabular dan terstruktur, mencakup beberapa atribut penting seperti Indeks Prestasi Kumulatif (IPK) per semester, jumlah Satuan Kredit Semester (SKS) yang ditempuh, nilai mata kuliah, total sks, serta angkatan.Variabel-variabel tersebut dipilih karena memiliki korelasi langsung terhadap performa akademik dan tingkat kelulusan mahasiswa.

Seluruh data yang digunakan dalam penelitian ini telah melalui tahap validasi awal untuk memastikan kelengkapan dan integritasnya, termasuk pemeriksaan data

ganda, nilai kosong (*missing values*), dan konsistensi antar sheet. Langkah ini penting dilakukan agar dataset yang digunakan representatif dan dapat memberikan hasil yang akurat dalam proses pemodelan. Dengan demikian, data yang terkumpul mencerminkan kondisi akademik nyata mahasiswa Universitas Multimedia Nusantara dan dapat digunakan secara andal dalam pengembangan sistem prediksi kelulusan berbasis *machine learning*.

3.4 Teknik Analisis Data

Analisis data dalam penelitian ini bertujuan untuk mengidentifikasi pola dan hubungan antarvariabel akademik yang memengaruhi kelulusan mahasiswa, serta membangun model prediksi berbasis pendekatan Hybrid Ensemble. Proses analisis dilakukan dengan menggabungkan hasil dari beberapa algoritma pembelajaran mesin, yaitu LightGBM, XGBoost, CatBoost, Random Forest, dan Multilayer Perceptron (MLP), untuk meningkatkan akurasi dan stabilitas hasil prediksi. Data yang digunakan dianalisis melalui tahapan utama, mulai dari pembersihan dan persiapan data hingga pemodelan dan evaluasi hasil.

Evaluasi model dilakukan dengan menggunakan metrik klasifikasi seperti *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix* untuk menilai performa model dalam mengenali setiap kategori kelulusan mahasiswa. Selain itu, metode SHAP (SHapley Additive Explanations) digunakan untuk memberikan interpretasi terhadap hasil prediksi dan mengetahui faktor-faktor yang paling berpengaruh terhadap kelulusan. Dengan pendekatan ini, analisis data tidak hanya menghasilkan model prediksi yang akurat, tetapi juga memberikan pemahaman yang lebih transparan mengenai kinerja akademik mahasiswa. Seluruh proses analisis dilakukan menggunakan bahasa pemrograman Python karena memiliki dukungan pustaka komputasi ilmiah dan *machine learning* yang luas, seperti scikit-learn, XGBoost, LightGBM, CatBoost, dan TensorFlow untuk pemodelan, serta SHAP, Matplotlib, dan Seaborn untuk interpretasi dan visualisasi hasil. Kombinasi pustaka tersebut memungkinkan pengolahan data secara efisien, pemodelan dengan berbagai algoritma, serta penyajian hasil yang informatif dan interaktif. Tabel 3.4

berikut membandingkan Python dengan dua bahasa pemrograman populer lainnya, yaitu R dan Matlab, dalam analisis data akademik.

Tabel 3. 4 Perbandingan Bahasa Pemrograman

No	Bahasa Pemrograman	Kelebihan dalam Penelitian	Kekurangan dalam Penelitian
1	Python	Memiliki pustaka machine learning dan explainable AI yang sangat lengkap (Scikit-learn, XGBoost, CatBoost, TensorFlow, PyTorch, SHAP, LIME). Sintaks sederhana, komunitas besar, dukungan visualisasi kuat (Matplotlib, Seaborn, Plotly). Cocok untuk integrasi berbagai algoritma sekaligus.	Performa eksekusi bisa lebih lambat dibanding bahasa yang terkompilasi seperti C++ jika tidak dioptimalkan.
2	R	Kuat dalam analisis statistik, data exploration, dan visualisasi. Banyak paket khusus untuk statistical modeling.	Kurang optimal untuk deep learning dan integrasi explainable AI skala besar, sintaks lebih kompleks untuk pemula.
3	Matlab	Andal dalam perhitungan numerik dan signal processing, dokumentasi resmi lengkap, stabil.	Berbayar, dukungan pustaka machine learning dan deep learning terbatas dibanding Python, kurang fleksibel untuk open-source integration.

Berdasarkan Tabel 3.4, Python dinilai sangat cocok untuk penelitian ini karena kemampuannya dalam mengintegrasikan berbagai algoritma *Explainable Artificial*

Intelligence seperti Random Forest, XGBoost, FNN, LSTM, dan CatBoost dalam satu ekosistem pemrograman yang efisien. Dukungan pustaka seperti Scikit-learn, XGBoost, CatBoost, TensorFlow, dan PyTorch memungkinkan proses pelatihan, evaluasi, serta visualisasi hasil dilakukan secara terstruktur dan cepat. Selain itu, Python memiliki pustaka *explainability* seperti SHAP yang relevan dengan fokus penelitian ini, yaitu memberikan interpretasi yang transparan terhadap hasil prediksi kinerja akademik mahasiswa. Sintaks yang sederhana dan komunitas pengguna yang besar juga mempermudah pengembangan, debugging, dan replikasi eksperimen, sehingga Python menjadi pilihan yang tepat untuk memastikan penelitian ini berjalan efektif, akurat, dan dapat direproduksi