BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan memanfaatkan metode machine learning untuk membangun model klasifikasi status kelulusan mahasiswa. Tujuan utama penelitian adalah menghasilkan model yang mampu mengelompokkan mahasiswa ke dalam empat kategori hasil studi, yaitu Lulus Lebih Awal, Lulus Tepat Waktu, Tidak Lulus Tepat Waktu, dan Drop Out. Model klasifikasi yang dikembangkan diarahkan untuk memberikan informasi prediktif terkait capaian studi mahasiswa, sehingga dapat mendukung proses pengambilan keputusan akademik yang bersifat preventif berbasis data.

Objek penelitian dalam studi ini adalah mahasiswa Program Studi Sistem Informasi Universitas Multimedia Nusantara angkatan 2020 hingga 2024. Data yang digunakan merupakan data akademik historis yang menggambarkan perkembangan studi mahasiswa selama masa perkuliahan. Data tersebut mencakup informasi nilai per mata kuliah, Indeks Prestasi Semester, jumlah pengulangan mata kuliah, total Satuan Kredit Semester (SKS) yang telah ditempuh, serta atribut administratif lain yang relevan dengan kondisi akademik mahasiswa. Seluruh data telah melalui proses anonimisasi dengan menghilangkan informasi identitas pribadi, sehingga analisis dilakukan sepenuhnya pada data akademik dan atribut pendukung yang tidak bersifat sensitif.

Metodologi penelitian berfokus pada pembangunan dan pengujian model machine learning untuk memprediksi status kelulusan mahasiswa berdasarkan data akademik yang tersedia. Model dikembangkan dengan mempertimbangkan karakteristik data yang digunakan, termasuk ukuran dataset yang relatif terbatas dan distribusi kelas yang tidak seimbang. Oleh karena itu, dilakukan eksplorasi terhadap konfigurasi model, teknik penyeimbangan data,

serta proses optimasi hyperparameter guna memperoleh kinerja model yang stabil dan representatif.

Selain mengevaluasi performa prediktif, penelitian ini juga memperhatikan aspek interpretabilitas hasil model. Untuk tujuan tersebut, diterapkan pendekatan *Explainable AI* menggunakan metode SHAP (SHapley Additive exPlanations) guna menganalisis kontribusi masing-masing fitur akademik terhadap hasil prediksi status kelulusan mahasiswa. Pendekatan ini memungkinkan hasil prediksi tidak hanya bersifat akurat, tetapi juga dapat dijelaskan secara transparan.

Seluruh rangkaian penelitian disusun berdasarkan kerangka kerja CRISP-DM (Cross Industry Standard Process for Data Mining), yang menyediakan tahapan terstruktur dan bersifat iteratif dalam pengelolaan proyek analitik berbasis data. Kerangka kerja ini digunakan sebagai acuan untuk mengelola proses penelitian mulai dari pemahaman permasalahan hingga evaluasi hasil pemodelan.

3.2 Tahapan Penelitian

Tahapan penelitian ini disusun dengan mengacu pada kerangka kerja CRISP-DM (*Cross Industry Standard Process for Data Mining*) yang membagi proses analisis data ke dalam enam fase utama, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*. Kerangka kerja ini dipilih karena menyediakan alur kerja yang sistematis dan bersifat iteratif, sehingga setiap tahapan dapat dievaluasi dan disesuaikan berdasarkan temuan pada tahap selanjutnya. Dalam penelitian ini, setiap fase diimplementasikan sebagai rangkaian proses yang terstruktur dan terdokumentasi dengan baik, sehingga hasil penelitian dapat dievaluasi secara kuantitatif dan dipertanggungjawabkan dari sisi metodologi.

3.2.1 Business Understanding

Tahap *business understanding* diawali dengan pemahaman terhadap permasalahan yang dihadapi dalam pengelolaan akademik di Program Studi Sistem Informasi Universitas Multimedia Nusantara. Salah

satu tantangan utama yang dihadapi adalah keterbatasan dalam mengidentifikasi mahasiswa yang berpotensi mengalami keterlambatan kelulusan atau berisiko *drop out* secara dini. Proses pembinaan akademik cenderung bersifat reaktif, di mana intervensi dilakukan setelah permasalahan akademik muncul secara nyata, seperti penurunan capaian akademik atau ketidaktercapaian beban studi. Kondisi ini berdampak pada meningkatnya risiko keterlambatan kelulusan dan ketidakefisienan proses pendampingan akademik.

Selain itu, dinamika proses pembelajaran dalam beberapa tahun terakhir, termasuk perubahan pola pembelajaran dan variasi karakteristik mahasiswa antar angkatan, menghasilkan keragaman performa akademik yang semakin kompleks. Hal ini meningkatkan kebutuhan akan pendekatan berbasis data yang mampu memberikan gambaran prediktif terhadap status kelulusan mahasiswa. Oleh karena itu, pemanfaatan data akademik historis menjadi relevan untuk membangun model yang dapat mendeteksi pola-pola risiko secara lebih awal dan objektif.

Pada tahap ini juga dilakukan pemetaan kesenjangan dari penelitian-penelitian sebelumnya. Sejumlah studi terdahulu masih berfokus pada prediksi indikator akademik tunggal, menggunakan fitur agregat yang terbatas, atau menerapkan model yang sulit diinterpretasikan oleh pemangku kepentingan akademik. Selain itu, beberapa penelitian melaporkan nilai evaluasi yang sangat tinggi pada dataset berukuran kecil dan tidak seimbang, yang mengindikasikan potensi *overfitting* dan keterbatasan generalisasi model. Kondisi tersebut menunjukkan perlunya pendekatan yang lebih berhati-hati dalam pemodelan, baik dari sisi pemilihan fitur, penanganan distribusi kelas, maupun transparansi hasil prediksi.

Berdasarkan pemahaman tersebut, kebutuhan utama yang diidentifikasi pada tahap *business understanding* adalah tersedianya model prediksi status kelulusan mahasiswa yang andal, stabil, dan dapat dijelaskan. Model yang dikembangkan diharapkan mampu mengelompokkan mahasiswa

ke dalam empat kategori status kelulusan, serta memberikan informasi mengenai faktor-faktor akademik yang berkontribusi terhadap hasil prediksi. Kebutuhan ini menjadi dasar perumusan tujuan penelitian dan spesifikasi fungsional dari pendekatan analitik yang akan digunakan.

Pada tahap ini juga ditetapkan batasan penelitian yang memengaruhi ruang lingkup analisis. Penelitian dibatasi pada data akademik internal mahasiswa Program Studi Sistem Informasi Universitas Multimedia Nusantara angkatan 2020 hingga 2024 yang telah dianonimkan. Hasil penelitian bersifat kontekstual terhadap karakteristik kurikulum dan kebijakan akademik program studi tersebut. Selain itu, penelitian difokuskan pada pengembangan dan evaluasi model prediksi, sehingga tahap implementasi penuh sistem peringatan dini berada di luar cakupan penelitian dan menjadi peluang pengembangan lanjutan.

Berdasarkan identifikasi permasalahan, kesenjangan penelitian, dan batasan yang telah ditetapkan, tujuan utama pada tahap *business understanding* dirumuskan secara terukur, yaitu mengembangkan model klasifikasi status kelulusan mahasiswa berbasis data akademik yang memiliki kinerja prediktif yang baik serta mampu memberikan penjelasan yang dapat dipahami. Tujuan pendukung meliputi pengelolaan data akademik yang tidak seimbang, optimasi konfigurasi model, dan penyediaan hasil analisis yang dapat dijadikan dasar pengambilan keputusan akademik berbasis data.

3.2.2 Data Understanding

Tahap *data understanding* bertujuan untuk memperoleh pemahaman yang menyeluruh terhadap struktur, karakteristik, dan kualitas data sebelum dilakukan proses pengolahan dan pemodelan. Pada tahap ini, dilakukan identifikasi sumber data, pemetaan peran masing-masing data, serta eksplorasi awal untuk mengenali pola umum dan potensi permasalahan yang dapat memengaruhi kinerja model. Pemahaman ini menjadi fondasi penting agar tahapan selanjutnya dapat dilakukan secara sistematis dan tepat sasaran.

Penelitian ini menggunakan satu sumber data utama yang memiliki fungsi berbeda dalam membangun fitur prediktif dan label keluaran. Sumber pertama berupa data biodata mahasiswa Program Studi Sistem Informasi angkatan 2020 hingga 2024 yang diperoleh dari sistem akademik melalui Biro Administrasi Akademik. Data ini memuat informasi demografis dan atribut administratif, seperti tahun masuk, jenis kelamin, dan status mahasiswa. Seluruh informasi identitas telah dianonimkan untuk menjaga kerahasiaan data. Data biodata digunakan untuk memberikan gambaran awal mengenai populasi mahasiswa serta untuk memetakan distribusi mahasiswa berdasarkan angkatan.

Setelah sumber data diidentifikasi, dilakukan eksplorasi awal untuk memahami distribusi dan pola umum dalam data. Eksplorasi ini dilakukan melalui berbagai visualisasi guna mengidentifikasi kecenderungan, ketidakseimbangan distribusi, serta karakteristik penting lainnya. Visualisasi distribusi jumlah mahasiswa berdasarkan angkatan digunakan untuk melihat persebaran data per tahun masuk, yang berpotensi memengaruhi distribusi kelas kelulusan. Selain itu, visualisasi rata-rata IPK per angkatan digunakan untuk menggambarkan variasi capaian akademik antar cohort.

Analisis visual juga dilakukan untuk melihat dinamika performa akademik secara umum melalui tren rata-rata Indeks Prestasi Semester mahasiswa. Visualisasi ini memberikan gambaran mengenai pola kenaikan atau penurunan performa akademik pada periode tertentu selama masa studi. Selain itu, dilakukan analisis terhadap karakteristik mata kuliah melalui visualisasi mata kuliah dengan jumlah pengambilan terbanyak serta mata kuliah dengan rata-rata nilai tertinggi dan terendah. Analisis ini membantu mengidentifikasi mata kuliah yang cenderung memiliki tingkat kesulitan lebih tinggi atau sebaliknya.

Pada tingkat hasil studi, dilakukan visualisasi distribusi status kelulusan mahasiswa untuk melihat proporsi masing-masing kategori kelulusan dalam dataset. Analisis ini diperluas dengan melihat distribusi status

kelulusan berdasarkan angkatan, sehingga dapat diamati perbedaan pola kelulusan antar cohort mahasiswa.

Melalui keseluruhan proses eksplorasi dan visualisasi tersebut, tahap *data understanding* menghasilkan pemahaman yang komprehensif mengenai struktur dataset, distribusi variabel penting, karakteristik mata kuliah, serta pola status kelulusan mahasiswa. Pemahaman ini menjadi dasar bagi tahap data preparation, khususnya dalam menentukan strategi pembersihan data, rekayasa fitur, dan penanganan ketidakseimbangan kelas sebelum proses pemodelan dilakukan.

3.2.3 Data Preprocessing

Tahap *data preprocessing* merupakan tahapan krusial dalam penelitian ini karena kualitas data pada tahap ini secara langsung memengaruhi kinerja model prediksi serta hasil interpretasi yang dihasilkan. Data akademik memiliki karakteristik khusus, seperti bersifat longitudinal, memiliki distribusi kelas yang tidak seimbang, serta memuat kombinasi fitur numerik dan kategorikal dengan tingkat kompleksitas yang beragam. Oleh karena itu, *preprocessing* dilakukan melalui serangkaian langkah terstruktur yang mencakup pembersihan data, penanganan nilai hilang, pembentukan label, transformasi fitur, rekayasa fitur, pembagian dataset, serta penyeimbangan kelas. Seluruh proses ini dirancang untuk memastikan data yang digunakan memiliki kualitas yang baik dan representatif terhadap kondisi akademik mahasiswa.

Seluruh tahapan *preprocessing* dilakukan secara sistematis dan terdokumentasi dengan baik. Setiap transformasi data dicatat, parameter yang digunakan disimpan, dan hasil antara disimpan sebagai dataset terpisah. Pendekatan ini bertujuan untuk memastikan seluruh proses dapat direplikasi serta memudahkan evaluasi dan pengujian ulang pada tahap pemodelan.

3.2.3.1 Penanganan Missing Values

Penanganan nilai hilang dilakukan secara hati-hati karena nilai kosong dapat memengaruhi perhitungan fitur akademik penting, seperti Indeks Prestasi Semester, total SKS, dan IPK. Pada fitur numerik, nilai

hilang umumnya muncul pada semester awal akibat belum lengkapnya riwayat akademik mahasiswa. Jika proporsi nilai hilang relatif kecil, dilakukan imputasi menggunakan nilai median berdasarkan kelompok angkatan untuk menjaga distribusi data tetap stabil. Pendekatan median dipilih karena lebih *robust* terhadap *outlier* dibandingkan nilai rata-rata.

Untuk fitur kategorikal, dilakukan pemeriksaan awal untuk membedakan nilai hilang yang disebabkan oleh ketidaklengkapan input dengan kondisi mahasiswa yang belum menempuh mata kuliah tertentu. Jika proporsi nilai hilang relatif kecil, imputasi dilakukan menggunakan nilai modus pada fitur tersebut. Namun, apabila suatu mata kuliah memiliki proporsi nilai hilang yang tinggi dan tidak konsisten antar angkatan, data tersebut tidak digunakan dalam proses pemodelan. Seluruh proses imputasi hanya diterapkan pada data pelatihan untuk mencegah terjadinya kebocoran informasi ke data validasi dan data uji.

3.2.3.2 Pembentukan Label dan Definisi Kategori Kelulusan

Label target dalam penelitian ini dibentuk berdasarkan catatan kelulusan resmi mahasiswa. Proses pelabelan dilakukan dengan menghitung durasi studi sejak tahun masuk hingga semester terakhir mahasiswa tercatat aktif atau lulus. Mahasiswa yang menyelesaikan studi dalam waktu tujuh semester atau kurang dikategorikan sebagai Lulus Lebih Awal, mahasiswa yang lulus pada semester kedelapan dikategorikan sebagai Lulus Tepat Waktu, mahasiswa yang menyelesaikan studi lebih dari delapan semester dikategorikan sebagai Tidak Lulus Tepat Waktu, dan mahasiswa yang tidak memiliki catatan kelulusan atau tidak lagi aktif diklasifikasikan sebagai *Drop Out*. Definisi ini mengikuti ketentuan akademik yang berlaku sehingga label antar kelas bersifat konsisten dan terstandarisasi.

Setelah label dibentuk, dilakukan analisis distribusi kelas yang menunjukkan adanya ketidakseimbangan yang cukup signifikan. Kategori Lulus Tepat Waktu mendominasi jumlah data, sementara kategori *Drop Out* memiliki proporsi yang jauh lebih kecil. Kondisi ini menjadi perhatian

penting karena dapat menyebabkan bias prediksi jika tidak ditangani dengan tepat pada tahap selanjutnya.

3.2.3.3 Transformasi dan Pengelolaan Fitur Kategorikal

Pengolahan fitur kategorikal dilakukan dengan mempertimbangkan jumlah kategori dan perannya dalam analisis. Fitur dengan jumlah kategori terbatas, seperti jenis kelamin dan angkatan, dikonversi ke bentuk numerik agar dapat diproses oleh algoritma pembelajaran. Sementara itu, fitur dengan jumlah kategori besar, seperti kode mata kuliah, tidak dilakukan *one-hot encoding* karena dapat menghasilkan dimensi fitur yang sangat tinggi.

Fitur-fitur tersebut dipertahankan dalam bentuk kategorikal karena algoritma CatBoost yang digunakan mendukung pengolahan fitur kategorikal secara langsung melalui mekanisme *encoding* internal. Pendekatan ini membantu menjaga efisiensi komputasi serta menghindari permasalahan sparsity yang umum terjadi pada data pendidikan dengan variasi mata kuliah yang besar. Seluruh transformasi fitur dilakukan hanya pada data pelatihan untuk menjaga integritas proses evaluasi..

3.2.3.4 Rekayasa Fitur dari Data Transkrip

Rekayasa fitur merupakan salah satu proses utama dalam tahap *preprocessing*. Fitur-fitur dibentuk dari data transkrip akademik untuk menangkap informasi mengenai capaian dan pola akademik mahasiswa. Fitur yang dihasilkan meliputi IPK kumulatif, total SKS lulus, jumlah mata kuliah yang diulang, proporsi SKS lulus terhadap SKS yang diambil, serta jumlah nilai rendah yang diperoleh mahasiswa.

Selain fitur statis, dibentuk pula fitur yang mencerminkan dinamika akademik, seperti tren Indeks Prestasi Semester yang dihitung menggunakan pendekatan regresi linier sederhana. Fitur ini digunakan untuk menggambarkan kecenderungan performa akademik mahasiswa, apakah mengalami peningkatan, penurunan, atau stagnasi. Seluruh proses rekayasa fitur dilakukan dengan menjaga urutan waktu agar tidak terjadi

pemanfaatan informasi dari periode akademik yang seharusnya belum tersedia

3.2.3.5 Pembagian Data

Setelah seluruh fitur siap digunakan, dataset dibagi menjadi data pelatihan dan data uji dengan proporsi 80% untuk pelatihan dan 20% untuk pengujian. Pembagian dilakukan menggunakan teknik *stratified sampling* agar distribusi kelas kelulusan tetap proporsional pada setiap subset. Teknik ini penting untuk memastikan bahwa seluruh kategori kelulusan tetap terwakili, terutama kelas minoritas.

Untuk menjaga konsistensi hasil, proses pembagian data dilakukan dengan nilai *seed* acak yang tetap. Seluruh transformasi lanjutan seperti *scaling* dan penyeimbangan data hanya diterapkan pada data pelatihan setelah proses pembagian selesai. Pendekatan ini bertujuan untuk mencegah data *leakage* yang dapat menyebabkan hasil evaluasi menjadi bias.

3.2.3.6 Penyeimbangan Data menggunakan SMOTE dan

Ketidakseimbangan kelas pada data pelatihan ditangani menggunakan teknik oversampling sintetis. Metode utama yang digunakan adalah SMOTE, yang menghasilkan sampel sintetis pada kelas minoritas melalui interpolasi antar sampel terdekat. Parameter jumlah tetangga ditetapkan pada nilai k = 5 untuk menjaga keseimbangan antara variasi dan stabilitas sampel sintetis.

Penyeimbangan data hanya diterapkan pada data pelatihan dan seluruh dataset hasil resampling disimpan sebagai artefak terpisah. Pendekatan ini memastikan bahwa setiap model dilatih pada data yang seimbang tanpa mengorbankan validitas data uji. Dengan demikian, model yang dihasilkan diharapkan mampu mempelajari pola kelas minoritas secara lebih baik dan memiliki kemampuan generalisasi yang lebih stabil.

3.2.4 Modeling

Tahap *modeling* bertujuan untuk membangun, melatih, dan mengevaluasi model *machine learning* yang digunakan dalam memprediksi status kelulusan mahasiswa berdasarkan data akademik yang telah melalui proses *preprocessing*. Pada tahap ini, data pelatihan yang telah dibersihkan, direkayasa, dan diseimbangkan digunakan untuk melatih beberapa algoritma klasifikasi dengan karakteristik yang berbeda. Proses *modeling* dilakukan secara terstruktur untuk memastikan bahwa perbandingan kinerja antar model bersifat adil dan dapat dipertanggungjawabkan.

Penelitian ini menggunakan tiga algoritma utama, yaitu CatBoost, *Logistic Regression* dengan regularisasi L1 (Lasso), dan K-*Nearest Neighbor* (KNN). Ketiga model dipilih untuk merepresentasikan tiga pendekatan berbeda dalam pembelajaran mesin, yaitu model *ensemble* berbasis *boosting*, model linear dengan seleksi fitur, dan model berbasis kedekatan data (*instance-based learning*). Dengan pendekatan ini, analisis tidak hanya berfokus pada performa prediksi, tetapi juga pada stabilitas model dan karakteristik perilaku masing-masing algoritma terhadap data akademik yang dapat dilihat pada Tabel 3.1.

Tabel 3.1 *Hyperaparameter* Tiga Model Pembanding

Model	Parameter Utama	Tujuan Penggunaan	Kelebihan	Kelemahan
CatBoost (Baseline)	• Iterations = 200 • Depth = 4 • Learning Rate = 0,03 • L2 Leaf Reg. = 3,0 • Loss = MultiClass • Eval Metric = TotalF1	Model utama untuk menangkap pola kompleks pada klasifikasi multiclass dengan data tidak seimbang	Mampu menangani fitur kategorikal secara langsung; stabil pada dataset kecil; mampu menangkap hubungan non-linear; optimal pada kasus multiclass	Model relatif kompleks; membutuhkan tuning parameter; interpretasi tidak langsung tanpa XAI

Model	Parameter Utama	Tujuan Penggunaan	Kelebihan	Kelemahan
	• Class Weight = Balanced • Random Seed = 42			
Logistic Regression L1 (Lasso)	• Penalty = L1 • Solver = liblinear • C = 1,0 • Max Iteration = 2000 • Class Weight = Balanced • Random State = 42	Model pembanding linear sebagai baseline yang sederhana dan interpretable	Mudah diinterpretasikan; melakukan seleksi fitur otomatis; stabil dan efisien	Tidak mampu menangkap hubungan non-linear; performa terbatas pada data kompleks
K-Nearest Neighbor (KNN)	• k = 15 • Weights = uniform • Metric = Euclidean	Model pembanding berbasis kemiripan untuk menangkap pola lokal antar data	Konsep sederhana; tidak memerlukan pelatihan eksplisit; efektif untuk pola lokal	Sensitif terhadap skala fitur; performa menurun pada dimensi tinggi; biaya komputasi tinggi saat prediksi

3.2.4.1 Categorical Boosting (Catboost)

CatBoost digunakan sebagai model utama karena kemampuannya dalam menangani fitur kategorikal secara langsung tanpa memerlukan proses *one-hot encoding*. Algoritma ini termasuk dalam keluarga *gradient boosting* yang membangun model secara bertahap dengan mengombinasikan banyak pohon keputusan berukuran kecil. Mekanisme *ordered boosting* pada CatBoost membantu mengurangi risiko *overfitting*, terutama pada dataset berukuran kecil dan tidak seimbang.

Parameter utama yang digunakan pada model CatBoost meliputi jumlah iterasi (*iterations*), kedalaman pohon (*depth*), *learning rate*, dan parameter regularisasi L2. Selain itu, pengaturan *subsample*, *random strength*, dan *grow policy* digunakan untuk meningkatkan generalisasi model. Nilai parameter awal ditetapkan berdasarkan praktik umum dan kemudian dioptimasi menggunakan *Bayesian optimization* untuk memperoleh konfigurasi terbaik.

Tujuan penggunaan CatBoost adalah untuk memperoleh model dengan performa prediktif yang tinggi dan stabil pada data akademik yang kompleks. Kelebihan utama CatBoost terletak pada kemampuannya menangani fitur kategorikal, performa yang baik pada dataset kecil, serta fleksibilitas dalam menangkap hubungan non-linear antar fitur. Namun, kelemahan CatBoost adalah kompleksitas model yang relatif tinggi dan kebutuhan *tuning parameter* yang lebih intensif dibandingkan model linear.

3.2.4.2 Logistic Regression L1 (Lasso)

Logistic Regression dengan regularisasi L1 digunakan sebagai model pembanding yang bersifat linear dan interpretable. Model ini bekerja dengan memodelkan hubungan linier antara fitur input dan probabilitas kelas keluaran menggunakan fungsi logistik. Regularisasi L1 ditambahkan untuk menekan kompleksitas model dengan mendorong sebagian koefisien fitur menjadi nol, sehingga secara otomatis melakukan seleksi fitur.

Parameter utama pada Logistic Regression L1 meliputi jenis penalti (L1), nilai regularisasi (C), *solver* yang mendukung penalti L1, serta skema klasifikasi multikelas. Nilai parameter regularisasi dikontrol untuk menyeimbangkan antara kompleksitas model dan kemampuan generalisasi.

Tujuan penggunaan *Logistic Regression* L1 adalah untuk menyediakan *baseline* model yang sederhana dan mudah

diinterpretasikan. Kelebihan model ini adalah transparansi hasil, stabilitas pelatihan, dan kemampuan seleksi fitur otomatis. Namun, *Logistic Regression* memiliki keterbatasan dalam menangkap hubungan non-linear antar fitur, sehingga performanya dapat menurun ketika pola data bersifat kompleks.

3.2.4.3 K-Nearest Neighbor (KNN)

K-*Nearest Neighbor* digunakan sebagai representasi model berbasis kedekatan data yang tidak membangun model eksplisit pada tahap pelatihan. Prediksi dilakukan dengan menentukan kelas mayoritas dari sejumlah tetangga terdekat berdasarkan metrik jarak tertentu. Algoritma ini sangat bergantung pada struktur data dan representasi fitur.

Parameter utama yang digunakan pada KNN meliputi jumlah tetangga (k), jenis metrik jarak, serta skema pembobotan jarak. Nilai k diuji pada beberapa variasi untuk menghindari sensitivitas terhadap *noise* maupun *over-smoothing*. Sebelum penerapan KNN, data numerik telah melalui proses normalisasi agar perhitungan jarak tidak didominasi oleh fitur tertentu.

Tujuan penggunaan KNN adalah untuk mengevaluasi kemampuan model berbasis instance dalam memanfaatkan kemiripan antar mahasiswa berdasarkan fitur akademik. Kelebihan KNN terletak pada kesederhanaan konsep dan kemampuannya menangkap pola lokal dalam data. Namun, KNN memiliki kelemahan dari sisi efisiensi komputasi, sensitivitas terhadap skala fitur, serta penurunan performa pada data berdimensi tinggi.

3.2.4.4 Optimasi Model Terbaik (Catboost)

Setelah dilakukan evaluasi awal terhadap seluruh model yang digunakan, CatBoost dipilih sebagai model utama karena menunjukkan performa paling stabil pada klasifikasi status kelulusan mahasiswa. Untuk memastikan bahwa konfigurasi model yang digunakan sudah berada pada kondisi yang optimal, dilakukan proses optimasi *hyperparameter*

terhadap model CatBoost. Tahap ini bertujuan untuk mengeksplorasi pengaruh variasi parameter terhadap kinerja model, khususnya pada metrik F1-*score macro* yang menjadi fokus utama penelitian ini.

Optimasi *hyperparameter* dilakukan menggunakan metode *Bayesian Optimization* melalui *framework* Hyperopt dengan algoritma *Tree-structured Parzen Estimator* (TPE). Pendekatan ini dipilih karena mampu mencari kombinasi parameter secara efisien pada ruang pencarian yang relatif besar tanpa harus mengevaluasi seluruh kemungkinan parameter secara menyeluruh. Proses optimasi dilakukan pada data pelatihan yang telah melalui tahap penyeimbangan kelas menggunakan SMOTE.

Fungsi objektif pada proses optimasi ditetapkan untuk memaksimalkan nilai F1-*score macro*, yang dihitung menggunakan 4-*fold cross validation* pada data pelatihan. Penggunaan validasi silang bertujuan untuk memperoleh estimasi kinerja model yang lebih stabil dan mengurangi ketergantungan pada satu pembagian data tertentu. Seluruh parameter yang dioptimasi dibatasi pada rentang nilai tertentu agar tetap sesuai dengan karakteristik data akademik dan mencegah kompleksitas model yang berlebihan.

Parameter yang diikutsertakan dalam proses optimasi mencakup jumlah iterasi, *learning rate*, kedalaman pohon, serta beberapa parameter regularisasi dan randomisasi yang disediakan oleh CatBoost. Selain itu, beberapa parameter bersifat tetap digunakan untuk menjaga konsistensi dengan konfigurasi dasar model, seperti penggunaan skema penyeimbangan kelas otomatis dan metrik evaluasi internal yang relevan untuk permasalahan multikelas.

Tabel 3.2 berikut merangkum parameter yang dioptimasi beserta rentang nilai yang digunakan dalam proses pencarian *hyperparameter*.

Tabel 3.2 Parameter Optimasi Model Catboost

Parameter	Nilai	Keterangan
loss_function	MultiClass	Fungsi loss untuk klasifikasi multikelas
eval_metric	TotalF1	Metrik evaluasi internal CatBoost
auto_class_weights	Balanced	Penyeimbangan kelas otomatis
bootstrap_type	Bernoulli	Skema bootstrap untuk subsampling
od_type	Iter	Tipe *early stopping*
od_wait	50	Jumlah iterasi tanpa perbaikan

Selain parameter yang dioptimasi, beberapa parameter lain ditetapkan secara tetap selama proses pencarian, sebagaimana ditunjukkan pada Tabel 3.3 berikut.

Tabel 3.3 Parameter Tetap Model CatBoost

Parameter	Nilai	Keterangan
loss_function	MultiClass	Fungsi loss untuk klasifikasi multikelas
eval_metric	TotalF1	Metrik evaluasi internal CatBoost
auto_class_weights	Balanced	Penyeimbangan kelas otomatis
bootstrap_type	Bernoulli	Skema bootstrap untuk subsampling
od_type	Iter	Tipe *early stopping*
od_wait	50	Jumlah iterasi tanpa perbaikan

Proses optimasi dijalankan dengan jumlah evaluasi terbatas untuk menjaga efisiensi komputasi dan mencegah eksplorasi parameter yang terlalu agresif. Setelah proses optimasi selesai, kombinasi parameter terbaik digunakan untuk melatih model CatBoost pada data pelatihan dan selanjutnya dievaluasi menggunakan data uji yang tidak terlibat dalam proses pelatihan maupun optimasi.

Hasil evaluasi menunjukkan bahwa meskipun konfigurasi hasil optimasi menghasilkan peningkatan kinerja pada data pelatihan, performa pada data uji tidak melampaui model CatBoost dengan konfigurasi awal. Oleh karena itu, model CatBoost dengan pengaturan awal yang lebih sederhana tetap digunakan sebagai model utama dalam penelitian ini karena menunjukkan keseimbangan performa yang lebih baik pada seluruh kelas kelulusan.

3.2.5 Evaluation

Tahap evaluation bertujuan untuk menilai kinerja model dalam memprediksi status kelulusan mahasiswa serta menentukan model terbaik yang akan dipilih sebagai model akhir. Evaluasi dilakukan secara kuantitatif dengan menggunakan beberapa metrik klasifikasi yang relevan untuk permasalahan multiclass dengan distribusi data yang tidak seimbang. Proses evaluasi difokuskan pada kemampuan model dalam mengklasifikasikan seluruh kelas secara adil, bukan hanya pada kelas mayoritas.

3.2.5.1 Confusion Metrix

Confusion matrix digunakan untuk mengevaluasi kinerja model secara lebih rinci pada tingkat kelas. Matriks ini menunjukkan jumlah prediksi yang benar dan salah untuk setiap kategori status kelulusan, sehingga memungkinkan analisis kesalahan klasifikasi antar kelas. Melalui *confusion matrix*, dapat diamati pola misklasifikasi, misalnya kecenderungan model mengklasifikasikan mahasiswa yang tidak lulus tepat waktu sebagai lulus tepat waktu, atau kesulitan model dalam mengenali kelas dengan jumlah data yang lebih sedikit seperti *Drop Out*.

Penggunaan *confusion matrix* menjadi penting dalam klasifikasi *multiclass* dengan distribusi kelas yang tidak seimbang. Evaluasi tidak hanya berfokus pada jumlah prediksi benar secara keseluruhan, tetapi juga pada kemampuan model dalam membedakan setiap kategori kelulusan secara proporsional. *Confusion matrix* divisualisasikan untuk setiap model sehingga perbedaan perilaku klasifikasi antar algoritma dapat dibandingkan secara langsung.

3.2.5.2 Classification Report

Selain *confusion matrix*, evaluasi dilakukan menggunakan *classification report* yang menyajikan metrik *precision*, *recall*, dan F1-*score* untuk setiap kelas, serta nilai rata-rata secara keseluruhan. *Precision* menggambarkan ketepatan prediksi model pada suatu kelas, *recall* menunjukkan kemampuan model dalam menangkap seluruh data yang benar-benar termasuk dalam kelas tersebut, sedangkan F1-*score* merupakan harmonisasi antara *precision* dan *recall*.

Dalam penelitian ini, perhatian utama diberikan pada nilai F1-*score macro.* Pemilihan F1-*score macro* didasarkan pada karakteristik data yang tidak seimbang, di mana setiap kelas memiliki tingkat kepentingan yang sama. F1-*score macro* menghitung rata-rata F1-*score* dari seluruh kelas tanpa mempertimbangkan proporsi jumlah data, sehingga memberikan gambaran yang lebih adil terhadap performa model pada kelas minoritas. *Classification report* digunakan untuk setiap model guna mengidentifikasi keunggulan dan keterbatasan model dalam memprediksi masing-masing kategori status kelulusan.

3.2.5.3 Model Selection

Proses pemilihan model dilakukan melalui komparasi kinerja tiga algoritma yang digunakan, yaitu CatBoost, *Logistic Regression* L1 (Lasso), dan K-*Nearest Neighbor* (KNN). Evaluasi dilakukan dengan menghitung beberapa metrik kinerja, yaitu akurasi, *precision macro*,

recall macro, dan F1-*score macro*. Seluruh metrik dihitung pada data uji yang sama untuk memastikan perbandingan yang adil antar model.

Hasil evaluasi dirangkum dalam bentuk tabel perbandingan performa dan diurutkan berdasarkan nilai F1-*score macro*. Metrik ini digunakan sebagai dasar utama dalam model selection karena mampu merepresentasikan keseimbangan performa model pada seluruh kelas, termasuk kelas dengan jumlah data yang terbatas. Selain evaluasi numerik, dilakukan pula visualisasi perbandingan F1-*score macro* antar model menggunakan diagram batang untuk mempermudah interpretasi perbedaan performa secara visual.

Dalam proses ini, konfigurasi model pembanding dijaga secara konsisten dengan parameter yang telah ditetapkan pada tahap modeling. Model KNN secara khusus menggunakan nilai jumlah tetangga yang relatif besar dan skema pembobotan uniform, sehingga menghasilkan batas keputusan yang lebih halus dan mengurangi sensitivitas terhadap *noise*. Pendekatan ini bertujuan untuk memastikan bahwa proses pemilihan model dilakukan secara objektif berdasarkan karakteristik alami masing-masing algoritma, bukan karena perbedaan konfigurasi yang tidak seimbang.

Berdasarkan hasil evaluasi kuantitatif dan analisis visual, model dengan nilai F1-*score macro* tertinggi dipilih sebagai model terbaik. Model terpilih dianggap memiliki keseimbangan terbaik antara ketepatan dan kemampuan deteksi pada seluruh kategori status kelulusan, sehingga paling sesuai untuk digunakan pada tahap analisis lanjutan dan interpretasi model.

3.2.6 Deployment

Tahap *deployment* bertujuan untuk menyajikan hasil model prediksi status kelulusan mahasiswa dalam bentuk aplikasi yang dapat digunakan dan dipahami oleh pengguna non-teknis. Pada penelitian ini, *deployment* dilakukan dalam bentuk *prototipe* aplikasi berbasis *web* menggunakan *framework*

Streamlit. Pendekatan ini dipilih karena Streamlit memungkinkan pengembangan aplikasi analitik secara cepat, ringan, dan interaktif, tanpa memerlukan konfigurasi infrastruktur yang kompleks.

Aplikasi Streamlit dirancang untuk mengintegrasikan model terpilih beserta seluruh *pipeline preprocessing* yang telah dibangun pada tahap sebelumnya. Model dan komponen pendukungnya disimpan dalam bentuk objek ter-serialisasi sehingga dapat dimuat kembali tanpa proses pelatihan ulang. Dengan pendekatan ini, proses prediksi dapat dilakukan secara konsisten dengan konfigurasi model yang telah dievaluasi, serta meminimalkan risiko perbedaan hasil antara tahap eksperimen dan tahap implementasi.

Dalam prototipe aplikasi yang dikembangkan, pengguna dapat memasukkan atau memuat data akademik mahasiswa yang telah diformat sesuai dengan struktur dataset penelitian. Aplikasi kemudian memproses data tersebut melalui pipeline preprocessing yang sama, melakukan prediksi status kelulusan, dan menampilkan hasil prediksi secara langsung. Selain hasil klasifikasi, aplikasi juga menyediakan visualisasi pendukung, seperti confusion matrix dan ringkasan metrik evaluasi, untuk memberikan gambaran kinerja model secara ringkas.

Selain aspek prediksi, tahap deployment juga mencakup penyajian hasil interpretabilitas model. Visualisasi berbasis SHAP ditampilkan untuk menunjukkan kontribusi fitur terhadap hasil prediksi, baik secara global maupun pada level individu. Penyajian ini bertujuan untuk membantu pengguna memahami faktor-faktor akademik yang berpengaruh terhadap prediksi yang dihasilkan, sehingga hasil model tidak hanya bersifat informatif, tetapi juga dapat dijelaskan secara transparan.

Deployment dalam penelitian ini dibatasi pada pengembangan prototipe aplikasi dan tidak mencakup implementasi penuh di lingkungan operasional program studi. Aspek seperti integrasi dengan sistem akademik internal, manajemen akses pengguna, dan pemeliharaan sistem berada di luar ruang lingkup penelitian. Meskipun demikian, prototipe yang dikembangkan

telah dirancang agar dapat menjadi dasar bagi pengembangan lanjutan sistem peringatan dini berbasis data di tingkat institusi.

3.3 Teknik Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan dengan memanfaatkan data sekunder berupa data akademik mahasiswa yang diperoleh dari sistem akademik Universitas Multimedia Nusantara melalui Biro Informasi Akademik (BIA). Data yang digunakan merupakan data historis yang mencerminkan perjalanan studi mahasiswa dan telah tersedia dalam sistem akademik institusi. Pendekatan ini dipilih karena data akademik bersifat objektif, terstruktur, serta merepresentasikan kondisi riil proses pembelajaran mahasiswa.

Data yang dikumpulkan mencakup informasi akademik dan administratif mahasiswa Program Studi Sistem Informasi, seperti nilai per mata kuliah, Indeks Prestasi Semester, Indeks Prestasi Kumulatif, jumlah Satuan Kredit Semester (SKS) yang ditempuh, jumlah pengulangan mata kuliah, status akademik, serta atribut pendukung lain yang relevan dengan proses studi. Seluruh data yang digunakan telah melalui proses anonimisasi dengan menghilangkan identitas pribadi mahasiswa, sehingga pemanfaatan data tetap sesuai dengan prinsip etika penelitian akademik dan perlindungan data.

Penggunaan satu sumber data utama dari sistem akademik dipandang memadai untuk mendukung tujuan penelitian, karena seluruh variabel yang digunakan dalam pemodelan prediksi status kelulusan berasal dari rekam jejak akademik mahasiswa. Dengan demikian, penelitian ini tidak melibatkan pengumpulan data primer melalui survei atau kuesioner, dan seluruh analisis didasarkan pada data akademik internal institusi.

3.3.1 Periode Pengambilan Data

Data akademik mahasiswa yang digunakan dalam penelitian ini dikumpulkan dari database Biro Informasi Akademik Universitas Multimedia Nusantara dalam rentang waktu tahun 2020 hingga 2024. Rentang periode ini dipilih untuk mencakup seluruh siklus studi mahasiswa Program Studi Sistem Informasi, mulai dari awal masa perkuliahan hingga status akhir kelulusan.

Periode tersebut juga mencerminkan variasi kondisi pembelajaran, termasuk masa sebelum, selama, dan setelah pandemi Covid-19, yang berpotensi memengaruhi pola akademik mahasiswa. Variasi kondisi ini memberikan keragaman data yang penting bagi proses pembelajaran model, sehingga model dapat mempelajari pola akademik mahasiswa dalam berbagai situasi pembelajaran.

Populasi

Populasi dalam penelitian ini mencakup seluruh mahasiswa Program Studi Sistem Informasi Universitas Multimedia Nusantara yang terdaftar pada angkatan 2020 hingga 2024. Berdasarkan rekapitulasi data akademik dan biodata mahasiswa yang diperoleh dari BIA, total populasi penelitian berjumlah 1.024 mahasiswa.

Populasi ini dipilih karena memiliki kelengkapan data akademik yang memadai, meliputi riwayat nilai per mata kuliah, beban studi, serta status akademik mahasiswa. Selain itu, populasi tersebut merepresentasikan kondisi aktual program studi, sehingga hasil penelitian diharapkan dapat menggambarkan pola risiko akademik yang relevan dengan kebutuhan pengelolaan akademik di lingkungan Program Studi Sistem Informasi UMN.

Sampel

Sampel penelitian ini terdiri dari mahasiswa Program Studi Sistem Informasi yang telah memiliki status kelulusan akhir dan dapat diklasifikasikan ke dalam empat kategori, yaitu Lulus Lebih Awal, Lulus Tepat Waktu, Lulus Terlambat, dan Drop Out. Dari total populasi sebanyak 1.024 mahasiswa, terdapat 421 mahasiswa yang memenuhi kriteria tersebut dan digunakan sebagai sampel penelitian.

Jumlah sampel tersebut diperoleh setelah dilakukan proses penyaringan data terhadap mahasiswa yang belum menyelesaikan studi pada periode pengambilan data. Mahasiswa yang masih berstatus aktif dan belum memiliki status kelulusan akhir dikeluarkan dari dataset

karena belum dapat diberikan label keluaran yang pasti. Berdasarkan hasil penyaringan tersebut, mahasiswa yang termasuk dalam sampel umumnya berasal dari angkatan awal dalam rentang penelitian, yaitu angkatan 2020 dan 2021, yang secara akademik telah mencapai tahap akhir studi atau telah memiliki status kelulusan.

Pemilihan sampel dengan kriteria tersebut dilakukan untuk memastikan bahwa seluruh data yang digunakan dalam pemodelan memiliki label status kelulusan yang valid dan konsisten. Dengan hanya memasukkan mahasiswa yang telah menyelesaikan studi atau memiliki status akhir yang jelas, proses pembelajaran model dapat dilakukan secara lebih akurat dan terhindar dari bias yang disebabkan oleh ketidakpastian *outcome* akademik.

Seluruh data sampel kemudian digunakan sebagai dasar dalam pelatihan dan evaluasi model prediksi status kelulusan mahasiswa. Fitur yang digunakan tidak dibedakan berdasarkan semester tertentu, melainkan merepresentasikan kondisi akademik mahasiswa secara agregat berdasarkan data akademik yang tersedia. Pendekatan ini memungkinkan model mempelajari pola umum yang berkaitan dengan risiko keterlambatan kelulusan dan drop out secara lebih menyeluruh.

3.4 Teknik Analisis Data

Analisis data pada penelitian ini dilakukan untuk membandingkan kinerja tiga model *machine learning*, yaitu CatBoost, *Logistic Regression* L1 (Lasso), dan K-*Nearest Neighbor* (KNN), dalam memprediksi status kelulusan mahasiswa. Seluruh model dilatih dan diuji menggunakan data akademik mahasiswa yang diperoleh dari Biro Informasi Akademik Universitas Multimedia Nusantara dan telah melalui tahapan *preprocessing*. Proses analisis mencakup pelatihan model, pengujian pada data uji, evaluasi kinerja, serta pemilihan model terbaik berdasarkan metrik evaluasi yang telah ditetapkan.

Evaluasi kinerja model dilakukan dengan menggunakan beberapa metrik klasifikasi, yaitu *accuracy*, *precision*, *recall*, dan F1-*score*. Mengingat

permasalahan yang dihadapi bersifat klasifikasi multiclass dengan distribusi kelas yang tidak seimbang, perhatian utama diberikan pada nilai F1-*score macro*. Metrik ini digunakan karena mampu merepresentasikan keseimbangan performa model pada seluruh kelas, termasuk kelas dengan jumlah data yang relatif kecil. Selain evaluasi numerik, analisis juga didukung oleh *confusion matrix* untuk mengidentifikasi pola kesalahan klasifikasi pada masing-masing kategori status kelulusan.

Selain evaluasi performa, analisis data juga mencakup interpretasi hasil prediksi menggunakan pendekatan *Explainable* AI (XAI). Metode SHAP (*Shapley Additive Explanations*) digunakan untuk menganalisis kontribusi masing-masing fitur akademik terhadap hasil prediksi model. Analisis ini dilakukan untuk memberikan penjelasan baik secara global maupun pada tingkat individu, sehingga hasil prediksi tidak hanya bersifat akurat, tetapi juga dapat dipahami secara transparan.

Selain evaluasi performa prediktif, penelitian ini juga mempertimbangkan aspek interpretabilitas model sebagai bagian penting dari analisis data. Dalam prediksi status kelulusan mahasiswa, kemampuan untuk menjelaskan alasan di balik suatu prediksi menjadi krusial agar hasil model dapat dipahami dan digunakan sebagai dasar pengambilan keputusan akademik. Oleh karena itu, pendekatan *Explainable AI* (XAI) digunakan untuk menginterpretasikan hasil prediksi model. Beberapa metode XAI yang umum digunakan dalam penelitian machine learning antara lain SHAP, LIME, dan *Permutation Feature Importance*. Masing-masing metode memiliki karakteristik, kelebihan, dan keterbatasan yang berbeda dalam menjelaskan perilaku model yang dapat dilihat pada Tabel 3.4.

Tabel 3.4 Perbandingan *Explainable AI*

Sumber: [48], [49]

Faktor	**SHAP**	**LIME**	**Permutation Feature Importance**

Prinsip Dasar	Berdasarkan teori Shapley untuk menghitung kontribusi fitur secara adil	Mengaproksimasi model secara lokal dengan model linear sederhana	Mengukur penurunan performa model ketika nilai fitur diacak
Cakupan Interpretasi	Global dan lokal (individu)	Lokal (per *instance*)	Global
Konsistensi Penjelasan	Tinggi, konsisten antar eksperimen	Dapat bervariasi tergantung sampling lokal	Stabil untuk analisis global
Dukungan Model	Model-*agnostic* dan mendukung *tree-based* model secara optimal	Model-*agnostic*	Model-*agnostic*
Kemampuan Menangani Non-linearitas	Sangat baik	Terbatas pada aproksimasi lokal	Tidak menjelaskan hubungan non-linear secara langsung
Kompleksitas Komputasi	Relatif tinggi, tetapi efisien untuk model *tree-based*	Lebih ringan, tetapi sensitif terhadap parameter	Rendah
Kesesuaian untuk Analisis Akademik	Sangat sesuai karena penjelasan komprehensif dan transparan	Cukup sesuai untuk studi kasus individu	Terbatas pada pemeringkatan fitur

Berdasarkan perbandingan tersebut, metode SHAP dipilih sebagai pendekatan utama *Explainable* AI dalam penelitian ini. SHAP memiliki keunggulan dalam memberikan penjelasan yang konsisten dan terukur secara matematis, baik pada tingkat global maupun individu. Kemampuan ini penting dalam akademik karena memungkinkan identifikasi faktor-faktor utama yang memengaruhi status kelulusan mahasiswa secara menyeluruh, sekaligus menjelaskan alasan prediksi pada kasus mahasiswa tertentu. Dibandingkan LIME yang bersifat lokal dan sensitif terhadap proses *sampling*, serta *Permutation Feature Importance* yang hanya memberikan gambaran global tanpa individual, SHAP menawarkan keseimbangan antara kedalaman analisis dan keterbacaan hasil. Oleh karena itu, SHAP dinilai paling sesuai untuk mendukung tujuan penelitian

dalam membangun model prediksi yang tidak hanya akurat, tetapi juga dapat dijelaskan dan ditindaklanjuti secara akademik.

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA