

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pendidikan tinggi berperan penting dalam membentuk kualitas sumber daya manusia dan pembangunan sosial, ekonomi suatu negara karena menghasilkan individu yang memiliki kompetensi akademik dan keterampilan profesional yang relevan dengan kebutuhan industri modern. Performa akademik pada tahap awal perkuliahan, dapat menjadi indikator awal keberhasilan studi selanjutnya, karena kemampuan untuk memprediksi performa akademik mahasiswa sejak dini menjadi kebutuhan strategis bagi institusi pendidikan. Perguruan tinggi dapat mengidentifikasi mahasiswa yang berpotensi mengalami kesulitan akademik serta merancang intervensi berbasis data guna meningkatkan kualitas pembelajaran dan keberhasilan studi melalui prediksi yang akurat. (Wang & Luo, 2024).

Pendekatan *machine learning* telah banyak digunakan untuk memprediksi performa akademik mahasiswa dalam beberapa tahun terakhir dengan memanfaatkan berbagai faktor seperti latar belakang sosial, ekonomi, riwayat akademik, serta kebiasaan belajar. Penelitian sebelumnya menunjukkan bahwa pendekatan *stacking ensemble* dengan label *encoding* mampu menghasilkan performa prediksi yang lebih baik dibandingkan model tunggal. Pendekatan ini mampu meningkatkan akurasi klasifikasi dalam konteks prediksi performa akademik mahasiswa dengan menggabungkan beberapa algoritma (Johora dkk., 2025) (Ayulani dkk., 2023).

Meskipun demikian arsitektur *stacking ensemble* memiliki beberapa keterbatasan, antara lain meningkatnya kompleksitas model serta kebutuhan komputasi yang lebih tinggi dibandingkan model tunggal. Selain itu, struktur *stacking ensemble* yang melibatkan beberapa *base learner* dan *meta learner* juga berpotensi meningkatkan risiko *overfitting* serta menurunkan tingkat interpretabilitas model (Widiyatmoko dkk., 2025). Permasalahan ini semakin kompleks karena dataset performa akademik juga sering menunjukkan distribusi kelas yang tidak seimbang. Ketidakseimbangan distribusi kelas dapat menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas sehingga menurunkan kemampuan model dalam mengenali kelas minoritas. Berbagai teknik *data balancing* telah dikembangkan untuk mengatasi permasalahan tersebut, salah satunya adalah *Synthetic Minority Oversampling Technique* (SMOTE) dan variannya. Namun, SMOTE memiliki keterbatasan karena menghasilkan data sintetis tanpa mekanisme pembersihan *noise*. Pendekatan yang lebih sesuai dengan karakteristik dataset penelitian ini adalah SMOTEN (*Synthetic Minority Oversampling Technique for Nominal Data*), yaitu metode *oversampling* yang dirancang khusus untuk menangani ketidakseimbangan kelas pada data kategorikal. SMOTEN menghasilkan sampel sintetis berdasarkan pola kategori dari tetangga terdekat kelas minoritas, sehingga distribusi kelas menjadi lebih seimbang tanpa menghasilkan nilai numerik sintetis yang tidak sesuai dengan makna kategori asli (García-Vicente dkk., 2023).

Meskipun berbagai penelitian telah menggunakan pendekatan *stacking ensemble* untuk meningkatkan akurasi prediksi performa akademik mahasiswa,

masih terbatas penelitian yang secara langsung mengevaluasi model *boosting* tunggal Catboost yang memiliki mekanisme *ordered boosting* dan struktur *symmetric trees* yang berpotensi mendukung kestabilan model, khususnya pada dataset berukuran kecil hingga menengah serta data yang mengandung sampel sintetis (Ramaswami dkk., 2022) (Tirumanadham dkk., 2024).

Selain menilai kinerja klasifikasi, penelitian ini turut menganalisis interpretabilitas CatBoost dan *stacking ensemble* melalui SHAP dan LIME. Penelitian sebelumnya telah menggunakan kedua metode tersebut untuk menjelaskan prediksi *stacking ensemble*. Penelitian ini menerapkan SHAP dan LIME pada CatBoost sehingga pola penjelasan dari kedua model dapat ditelaah dalam kondisi eksperimen yang sama. SHAP digunakan untuk mengidentifikasi kontribusi fitur secara global, sedangkan LIME digunakan untuk menjelaskan kontribusi fitur pada prediksi individual mahasiswa. Hasil dari kedua metode tersebut kemudian digunakan untuk melihat pola fitur yang dominan serta karakteristik penjelasan yang dihasilkan oleh masing-masing model (Salih dkk., 2023)

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan penggunaan CatBoost sebagai model *boosting* tunggal yang memiliki mekanisme *ordered boosting* dan *symmetric tree* yang dirancang untuk mengurangi risiko *overfitting* serta meningkatkan stabilitas pembelajaran model (Aldania dkk., 2023) Sejumlah penelitian menunjukkan bahwa CatBoost memiliki performa yang kompetitif dalam klasifikasi data tabular, khususnya pada dataset berukuran kecil hingga menengah yang didominasi oleh fitur kategorikal (Ramaswami dkk., 2022;

Tirumanadham dkk., 2024). Penelitian ini juga mempertimbangkan pentingnya interpretabilitas model dalam konteks pendidikan. Pendekatan *Explainable Artificial Intelligence* (XAI) digunakan untuk memahami kontribusi fitur terhadap hasil prediksi. Metode SHAP digunakan untuk mengidentifikasi kontribusi fitur secara global, sedangkan LIME digunakan untuk memberikan penjelasan pada level individu terhadap hasil prediksi model (Salih dkk., 2023).

Penelitian ini bertujuan mengimplementasikan CatBoost dengan SMOTEN sebagai alternatif model tunggal dengan arsitektur yang lebih sederhana, mengkaji kinerja klasifikasi multikelas CatBoost dan *stacking ensemble* berdasarkan metrik *accuracy*, *precision*, *recall*, *F1-score* dan *confusion matrix*, serta menganalisis kontribusi fitur yang dihasilkan oleh kedua model secara global menggunakan SHAP dan secara lokal menggunakan LIME.

1.2 Rumusan Masalah

Berdasarkan latar belakang dan permasalahan penelitian, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana implementasi dan kinerja CatBoost dengan SMOTEN pada model prediksi performa akademik mahasiswa?
2. Bagaimana kontribusi fitur secara global dan lokal terhadap hasil prediksi performa akademik mahasiswa yang dihasilkan oleh model CatBoost berdasarkan analisis SHAP dan LIME?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dipaparkan, penelitian ini memiliki tujuan sebagai berikut:

1. Mengimplementasikan dan mengevaluasi kinerja CatBoost dengan SMOTEN pada model prediksi performa akademik mahasiswa.
2. Menganalisis kontribusi fitur secara global menggunakan SHAP dan secara lokal menggunakan LIME terhadap hasil prediksi model CatBoost.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat baik dari sisi teoritis, metodologis, maupun praktis diantaranya:

1. Penelitian ini diharapkan memberikan kontribusi pada bidang *educational data mining* melalui kajian kinerja CatBoost dalam klasifikasi multikelas performa akademik mahasiswa, serta analisis kontribusi fitur secara global dan lokal menggunakan SHAP dan LIME.
2. Penelitian ini diharapkan menyediakan referensi empiris mengenai penggunaan CatBoost sebagai model *boosting* tunggal pada dataset pendidikan dengan fitur kategorikal.
3. Hasil penelitian ini diharapkan dapat menjadi referensi awal bagi institusi pendidikan dalam mempertimbangkan pengembangan model prediksi performa akademik yang tidak hanya memperhatikan kinerja klasifikasi, tetapi juga menyediakan informasi mengenai faktor-faktor yang berkontribusi terhadap hasil prediksi mahasiswa.

1.5 Batasan Masalah

Ruang lingkup dalam penelitian ini dibatasi dengan beberapa hal sebagai berikut:

1. Penelitian ini menggunakan satu dataset publik mengenai performa akademik mahasiswa semester pertama yang telah digunakan pada penelitian sebelumnya dan diperoleh melalui *platform* GitHub. Penelitian tidak menambahkan data dari sumber eksternal maupun melakukan validasi menggunakan dataset dari institusi pendidikan lain.
2. Penelitian difokuskan pada klasifikasi multikelas performa akademik mahasiswa dengan target klasifikasi dibatasi menjadi empat kelas performa akademik dan tidak mencakup prediksi regresi, perubahan performa antarsemester, maupun analisis longitudinal mahasiswa.
3. Proses evaluasi kinerja model dibatasi pada metrik evaluasi *accuracy*, *precision*, *recall*, *F-1 Score* dan *confusion matrix*.
4. Penanganan ketidakseimbangan kelas dibatasi pada penggunaan SMOTEN untuk menentukan teknik *resampling* yang paling sesuai dengan karakteristik dataset.
5. Analisis kontribusi fitur dibatasi pada SHAP untuk penjelasan global dan LIME untuk penjelasan lokal secara deskriptif, serta penelitian tidak membahas efisiensi komputasi dan implementasi model dalam sistem produksi.

1.6 Sistematika Penulisan

Proses penulisan penelitian ini disesuaikan dengan pedoman penulisan TA Universitas Siliwangi, yang terdiri dari lima BAB sebagai berikut.

BAB I PENDAHULUAN

BAB I membahas mengenai latar belakang, rumusan masalah, tujuan, manfaat, batasan masalah, dan sistematika penulisan penelitian.

BAB II LANDASAN TEORI

BAB II menjelaskan definisi *Educational Data Mining*, *supervised learning*, klasifikasi multikelas, SMOTEN, *stacking ensemble*, Random Forest, Gradient Boosting, Support Vector Classifier, CatBoost, SHAP, LIME, evaluasi model, *state of the art*, dan matriks penelitian.

BAB III METODOLOGI

BAB III menjelaskan kontribusi terhadap *Artificial Intelligence* Siliwangi (AIS) dan metodologi penelitian beserta penjelasannya secara lengkap.

BAB IV HASIL DAN PEMBAHASAN

BAB IV menjelaskan hasil yang telah didapatkan dari proses penelitian, seperti hasil pengumpulan data, hasil pra-proses data, hasil pengembangan model, hasil informasi model dan hasil evaluasi model.

BAB V KESIMPULAN DAN SARAN

BAB V menjelaskan mengenai kesimpulan yang didapatkan dari semua rangkaian penelitian dan memberikan saran untuk memperbaiki kekurangan dari penelitian yang telah dilakukan.