BAB I
PENDAHULUAN

1.1 Latar Belakang

Perguruan tinggi memiliki peran strategis dalam menghasilkan sumber daya manusia berkualitas melalui proses pendidikan yang berkelanjutan. Namun, salah satu permasalahan yang masih sering terjadi dalam pendidikan tinggi adalah tingginya angka mahasiswa yang tidak menyelesaikan studi atau mengalami dropout. Kondisi ini menjadi indikator penting dalam evaluasi kualitas layanan akademik karena menunjukkan adanya hambatan yang dialami mahasiswa selama proses pendidikan.

Mahasiswa yang mengalami dropout dapat disebabkan oleh berbagai faktor, baik faktor akademik maupun non-akademik. Faktor akademik dapat berupa rendahnya pencapaian nilai, jumlah kredit akademik yang tidak terpenuhi, maupun rendahnya keterlibatan dalam aktivitas pembelajaran. Sementara itu, faktor non-akademik dapat berkaitan dengan kondisi ekonomi, status pembayaran pendidikan, latar belakang sosial, dan karakteristik individu mahasiswa.
Dalam praktiknya, sebagian besar perguruan tinggi masih melakukan identifikasi mahasiswa bermasalah setelah mahasiswa mengalami penurunan performa akademik yang signifikan. Pendekatan tersebut menyebabkan ruang intervensi menjadi lebih terbatas karena masalah sudah berkembang menjadi kondisi yang sulit diperbaiki. Oleh karena itu, diperlukan sistem peringatan dini (Early Warning System/EWS) yang mampu melakukan prediksi risiko secara lebih awal.

Perkembangan sistem informasi akademik menyebabkan perguruan tinggi memiliki banyak data mahasiswa yang dapat dimanfaatkan untuk mendukung pengambilan keputusan. Data mengenai karakteristik mahasiswa, perjalanan akademik, jumlah mata kuliah yang diselesaikan, nilai semester, status pembayaran, dan informasi sosial ekonomi dapat digunakan sebagai sumber informasi dalam membangun model prediksi.
Penelitian sebelumnya menjelaskan bahwa penggunaan machine learning telah banyak diterapkan untuk melakukan klasifikasi risiko dropout mahasiswa. Namun, tidak terdapat satu algoritma yang selalu memberikan performa terbaik pada seluruh kondisi karena hasil prediksi sangat bergantung pada karakteristik data, distribusi fitur, definisi target, serta strategi validasi yang digunakan.

Beberapa algoritma yang banyak digunakan dalam prediksi dropout adalah Logistic Regression, Random Forest, dan XGBoost. Logistic Regression memiliki keunggulan dalam

interpretasi hubungan antara variabel prediktor dan hasil klasifikasi. Random Forest mampu menangani hubungan kompleks antarvariabel melalui pendekatan ensemble learning. Sementara itu, XGBoost dikenal memiliki kemampuan prediksi yang kuat pada data berbentuk tabel (tabular data). Namun, penggunaan machine learning dalam bidang pendidikan memiliki tantangan berupa keterbatasan interpretasi hasil model. Model dengan akurasi tinggi belum tentu mudah dipahami oleh pihak akademik ketika digunakan dalam proses pengambilan keputusan. Oleh sebab itu, penelitian ini menggunakan pendekatan Explainable Artificial Intelligence melalui metode SHAP.

SHAP mampu menjelaskan kontribusi setiap fitur terhadap hasil prediksi model. Melalui pendekatan ini, pihak akademik tidak hanya memperoleh informasi mengenai mahasiswa yang berisiko dropout, tetapi juga memahami faktor-faktor yang berkontribusi terhadap risiko tersebut. Namun, nilai SHAP hanya menunjukkan kontribusi prediktif dan tidak dapat diartikan sebagai hubungan sebab-akibat langsung.

Penelitian dalam jurnal Student Dropout Risk Prediction After Two Semesters: A Comparative Study of Logistic Regression, Random Forest, and XGBoost with SHAP menunjukkan bahwa Logistic Regression menghasilkan PR-AUC dan recall dropout tertinggi, Random Forest menghasilkan F1-score dan balanced accuracy tertinggi, sedangkan XGBoost memiliki performa kompetitif sekaligus dapat dijelaskan menggunakan SHAP. Berdasarkan permasalahan tersebut, penelitian ini dilakukan untuk mengembangkan model prediksi risiko dropout mahasiswa dengan membandingkan beberapa algoritma machine learning dan menganalisis faktor dominan menggunakan SHAP sebagai dasar pengembangan sistem peringatan dini akademik.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana penerapan algoritma Logistic Regression, Random Forest, dan XGBoost dalam melakukan prediksi risiko dropout mahasiswa?
2. Bagaimana perbandingan performa algoritma Logistic Regression, Random Forest, dan XGBoost berdasarkan hasil evaluasi model machine learning?
3. Algoritma manakah yang memberikan performa terbaik dalam melakukan klasifikasi risiko dropout mahasiswa berdasarkan parameter evaluasi yang digunakan?
4. Faktor-faktor apa saja yang menjadi prediktor dominan terhadap risiko dropout mahasiswa berdasarkan hasil analisis SHAP?

5. Bagaimana hasil prediksi machine learning dapat dimanfaatkan sebagai dasar pengembangan sistem peringatan dini (Early Warning System) pada institusi pendidikan tinggi?

1.3 Tujuan Penelitian

1.3.1 Tujuan Umum

Tujuan umum penelitian ini adalah mengembangkan model prediksi risiko dropout mahasiswa menggunakan pendekatan machine learning dengan membandingkan algoritma Logistic Regression, Random Forest, dan XGBoost serta memberikan interpretasi faktor prediktor menggunakan SHAP.

1.3.2 Tujuan Khusus

Tujuan khusus penelitian ini adalah:

1. Membangun model prediksi dropout mahasiswa menggunakan algoritma Logistic Regression, Random Forest, dan XGBoost.
2. Melakukan evaluasi dan perbandingan performa ketiga algoritma berdasarkan metrik PR-AUC, ROC-AUC, recall, precision, F1-score, balanced accuracy, dan Brier score. REVISI FINAL-DO-JMP
3. Mengidentifikasi faktor-faktor yang memiliki kontribusi terbesar terhadap prediksi risiko dropout menggunakan metode SHAP.
4. Mengetahui karakteristik mahasiswa yang memiliki tingkat risiko dropout berdasarkan pola data akademik dan non-akademik.
5. Memberikan rekomendasi pemanfaatan model prediksi sebagai pendukung sistem peringatan dini akademik.

1.4 Batasan Masalah

Agar penelitian lebih terfokus dan tidak mengalami perluasan pembahasan, batasan masalah dalam penelitian ini adalah:

1. Penelitian hanya membahas prediksi risiko dropout mahasiswa menggunakan pendekatan machine learning.
2. Dataset yang digunakan adalah dataset publik Predict Students' Dropout and Academic Success dari UCI Machine Learning Repository.

3. Variabel penelitian terbatas pada fitur yang tersedia dalam dataset, seperti karakteristik mahasiswa, kondisi sosial ekonomi, status pembayaran pendidikan, serta indikator akademik semester awal.
4. Tahapan penelitian mencakup preprocessing data, pelatihan model, optimasi parameter, evaluasi performa, dan interpretasi model menggunakan SHAP.
5. Penelitian tidak membahas implementasi sistem informasi akademik secara langsung maupun pengujian efektivitas intervensi terhadap mahasiswa.
6. Hasil prediksi model hanya digunakan sebagai pendukung pengambilan keputusan akademik dan tidak digunakan sebagai keputusan otomatis terhadap status mahasiswa.

1.5 Manfaat Penelitian

1.5.1 Manfaat Akademik

1. Memberikan kontribusi terhadap pengembangan ilmu pengetahuan pada bidang kecerdasan buatan, machine learning, dan sistem informasi pendidikan.
2. Menambah referensi penelitian terkait penerapan algoritma klasifikasi untuk memprediksi risiko dropout mahasiswa.
3. Memberikan gambaran mengenai penggunaan Explainable AI khususnya SHAP dalam meningkatkan interpretasi model prediksi.

1.5.2 Manfaat Praktis

1. Memberikan model prediksi yang dapat membantu institusi melakukan identifikasi awal mahasiswa yang memiliki risiko dropout.
2. Mendukung pengembangan sistem peringatan dini akademik berbasis data.
3. Membantu pihak akademik memahami faktor-faktor yang berkaitan dengan risiko dropout mahasiswa.
4. Memberikan informasi tambahan dalam menentukan strategi pendampingan mahasiswa.
5. Membantu proses pengambilan keputusan berbasis analisis data.