

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan kecerdasan buatan dan machine learning telah memberikan kontribusi signifikan dalam bidang pendidikan tinggi, khususnya dalam analisis prestasi akademik mahasiswa. Teknologi ini memungkinkan data akademik dimanfaatkan secara lebih sistematis untuk mengidentifikasi pola hubungan antar faktor yang memengaruhi capaian akademik. Berbeda dengan metode statistik konvensional yang umumnya mengandalkan asumsi hubungan linear, algoritma machine learning seperti Random Forest dan XGBoost mampu memodelkan hubungan yang lebih kompleks serta menghasilkan klasifikasi dengan tingkat akurasi yang tinggi (Khosravi & Azarnik, 2024). Kemampuan ini menjadikan machine learning sebagai pendekatan yang efektif dalam mendukung analisis prestasi akademik berbasis data, terutama ketika variabel yang terlibat berasal dari berbagai aspek yang saling berkaitan.

Prestasi akademik mahasiswa merupakan indikator utama dalam mengevaluasi keberhasilan proses pembelajaran di pendidikan tinggi. Indeks Prestasi Kumulatif (IPK) digunakan sebagai ukuran kuantitatif yang merepresentasikan capaian akademik mahasiswa secara menyeluruh selama masa studi. Perbedaan nilai IPK antar mahasiswa menunjukkan adanya variasi capaian akademik yang dipengaruhi oleh berbagai faktor, seperti kualitas bimbingan akademik, kemampuan manajemen waktu, kondisi ekonomi, dukungan keluarga, dan ketersediaan fasilitas pembelajaran. Faktor-faktor tersebut tidak bekerja secara terpisah, melainkan saling berinteraksi dan membentuk pola yang kompleks. Kondisi ini memerlukan pendekatan analisis yang mampu mengidentifikasi hubungan multidimensional secara objektif, sehingga pola yang relevan dapat dipahami dengan lebih akurat.

Sejumlah penelitian sebelumnya menunjukkan bahwa machine learning dapat digunakan secara efektif untuk mengklasifikasikan prestasi akademik mahasiswa. (Huynh-Cam & Nguyen, 2021) menunjukkan bahwa Random Forest mampu

mengidentifikasi pola prestasi akademik melalui analisis variabel akademik dan perilaku belajar secara simultan. (Asselman et al., 2023) menemukan bahwa XGBoost memberikan performa klasifikasi yang lebih baik dibandingkan metode statistik tradisional, terutama pada dataset dengan karakteristik kompleks. Selain itu, (Sánchez-Sánchez & González, 2024) menunjukkan bahwa algoritma ensemble seperti Random Forest dan XGBoost efektif dalam mengidentifikasi variabel yang berkontribusi terhadap perbedaan capaian akademik mahasiswa. Meskipun demikian, sebagian besar penelitian masih berfokus pada performa model, tanpa mengeksplorasi secara mendalam faktor-faktor dominan yang membedakan kelompok mahasiswa dengan tingkat prestasi yang berbeda. Penelitian yang mengintegrasikan faktor akademik, perilaku belajar, kondisi sosial ekonomi, dan fasilitas pembelajaran dalam satu model klasifikasi juga masih terbatas, khususnya dalam konteks pendidikan tinggi di Indonesia. (Cheng et al., 2024) menunjukkan bahwa XGBoost mampu menghasilkan klasifikasi dengan akurasi tinggi, namun belum menggabungkan faktor perilaku dan lingkungan belajar secara komprehensif.

Selain aspek analisis, implementasi model machine learning dalam bentuk sistem klasifikasi berbasis aplikasi masih relatif jarang dilakukan. Padahal, sistem interaktif dapat membantu institusi pendidikan memahami karakteristik mahasiswa secara lebih objektif dan berbasis data. Model klasifikasi yang diintegrasikan ke dalam aplikasi memungkinkan proses identifikasi kategori prestasi akademik dilakukan secara otomatis berdasarkan variabel yang relevan. Sistem ini juga dapat memberikan informasi mengenai faktor yang paling berpengaruh terhadap hasil klasifikasi melalui analisis feature importance. Pendekatan ini mendukung pengambilan keputusan berbasis data dalam pengelolaan pendidikan tinggi, misalnya dalam evaluasi akademik atau penyusunan strategi peningkatan prestasi mahasiswa.

Berdasarkan kebutuhan tersebut, penelitian ini bertujuan untuk membangun model klasifikasi prestasi akademik mahasiswa menggunakan algoritma Random Forest dan XGBoost untuk membedakan mahasiswa ke dalam kategori unggul dan reguler berdasarkan nilai IPK. Penelitian ini juga bertujuan mengidentifikasi faktor-

faktor dominan yang berkontribusi terhadap perbedaan prestasi akademik melalui analisis feature importance. Dataset yang digunakan mencakup berbagai faktor akademik, demografis, ekonomi, perilaku belajar, dan fasilitas pembelajaran, sehingga memungkinkan analisis dilakukan secara komprehensif. Model yang dihasilkan kemudian diintegrasikan ke dalam aplikasi berbasis Streamlit untuk memungkinkan proses klasifikasi dilakukan secara interaktif. Pengguna dapat memasukkan nilai variabel yang relevan dan memperoleh hasil klasifikasi secara langsung, disertai informasi faktor yang berkaitan dengan hasil tersebut. Dengan demikian, penelitian ini tidak hanya berfokus pada pengembangan model klasifikasi, tetapi juga pada implementasi sistem yang dapat digunakan sebagai alat bantu analisis dan pendukung pengambilan keputusan berbasis data di lingkungan pendidikan tinggi.

1.2 Rumusan Masalah

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

1. Bagaimana membangun model klasifikasi yang optimal menggunakan algoritma Random Forest dan XGBoost untuk membedakan mahasiswa unggul ($IPK > 3.5$) dan mahasiswa reguler ($IPK \leq 3.5$) berdasarkan variabel akademik, demografis, ekonomi, dan perilaku mahasiswa?
2. Bagaimana kinerja dan stabilitas model Random Forest dan XGBoost dalam mengklasifikasikan prestasi akademik mahasiswa pada dataset dengan distribusi kelas yang tidak seimbang dan jumlah sampel yang terbatas?
3. Faktor-faktor apa saja yang memiliki kontribusi dominan dalam model Random Forest dan XGBoost terhadap klasifikasi prestasi akademik mahasiswa berdasarkan analisis feature importance?
4. Bagaimana perbandingan kontribusi variabel dan pola karakteristik yang terbentuk dalam membedakan mahasiswa unggul dan mahasiswa reguler berdasarkan hasil interpretasi model Random Forest dan XGBoost?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah disebutkan, tujuan penelitian ini adalah sebagai berikut:

1. Membangun model klasifikasi menggunakan algoritma Random Forest dan XGBoost untuk membedakan mahasiswa unggul ($IPK > 3.5$) dan mahasiswa reguler ($IPK \leq 3.5$) berdasarkan variabel akademik, demografis, ekonomi, dan perilaku mahasiswa.
2. Mengevaluasi kinerja dan stabilitas model Random Forest dan XGBoost dalam mengklasifikasikan prestasi akademik mahasiswa pada dataset dengan distribusi kelas yang tidak seimbang dan jumlah sampel yang terbatas.
3. Mengidentifikasi variabel-variabel yang memiliki kontribusi dominan terhadap klasifikasi prestasi akademik mahasiswa berdasarkan analisis feature importance yang dihasilkan oleh algoritma Random Forest dan XGBoost.
4. Menganalisis kontribusi variabel dan pola karakteristik yang membedakan mahasiswa unggul dan mahasiswa reguler berdasarkan hasil interpretasi model klasifikasi.

1.4 Batasan Masalah

1. Dataset yang digunakan dalam penelitian ini terdiri dari 437 mahasiswa yang berasal dari dua lokasi Perguruan Tinggi, sehingga hasil model klasifikasi dan faktor dominan yang diperoleh terbatas pada karakteristik data dalam penelitian ini dan belum dapat digeneralisasikan secara luas untuk merepresentasikan seluruh populasi mahasiswa dengan karakteristik yang berbeda.
2. Distribusi kelas dalam dataset bersifat tidak seimbang (class imbalance), di mana jumlah mahasiswa dengan kategori IPK tertentu tidak tersebar secara proporsional. Kondisi ini menyebabkan klasifikasi dalam penelitian ini dibatasi menjadi dua kelas, yaitu mahasiswa unggul ($IPK > 3.5$) dan mahasiswa reguler ($IPK \leq 3.5$), karena penggunaan lebih dari dua kelas berpotensi menghasilkan jumlah sampel yang sangat sedikit pada kelas tertentu dan dapat menurunkan kemampuan model dalam mempelajari pola secara optimal.

3. Seluruh data dalam penelitian ini diperoleh melalui pengisian kuesioner oleh responden, sehingga kualitas dataset sepenuhnya bergantung pada keakuratan, kejujuran, dan konsistensi jawaban responden. Kondisi ini berpotensi menimbulkan bias subjektivitas, kesalahan persepsi, atau ketidaksesuaian antara jawaban responden dengan kondisi sebenarnya. Bias tersebut dapat memengaruhi representasi variabel dalam dataset, yang pada akhirnya dapat memengaruhi proses pembelajaran model, performa klasifikasi, serta interpretasi faktor dominan yang dihasilkan.
4. Dataset yang digunakan dalam penelitian ini dikumpulkan dalam satu periode waktu tertentu dan tidak mencerminkan perubahan kondisi mahasiswa secara berkelanjutan. Oleh karena itu, model klasifikasi yang dihasilkan hanya merepresentasikan pola hubungan antara variabel dan prestasi akademik pada waktu pengambilan data, serta tidak dapat menangkap dinamika perubahan prestasi akademik maupun perubahan faktor yang memengaruhinya dari waktu ke waktu. Penelitian ini juga tidak dapat mengevaluasi hubungan temporal atau mengidentifikasi pola perkembangan prestasi akademik mahasiswa secara longitudinal.
5. Klasifikasi prestasi akademik dalam penelitian ini didasarkan pada nilai Indeks Prestasi Kumulatif (IPK) dengan ambang batas tertentu yang ditetapkan sebagai representasi kategori mahasiswa unggul dan reguler. Oleh karena itu, pengukuran prestasi akademik dalam penelitian ini terbatas pada indikator IPK dan tidak mencakup prestasi lain di luar IPK, seperti prestasi non-akademik, kegiatan organisasi, atau pencapaian kompetisi mahasiswa.
6. Penelitian ini menggunakan algoritma Random Forest dan XGBoost sebagai metode klasifikasi utama, sehingga hasil klasifikasi dan identifikasi faktor dominan yang diperoleh bergantung pada karakteristik dan mekanisme pembelajaran dari kedua algoritma tersebut. Penelitian ini tidak mengevaluasi atau membandingkan dengan seluruh algoritma klasifikasi lain, sehingga faktor dominan dan performa model yang dihasilkan terbatas pada pendekatan yang digunakan dalam penelitian ini.

1.5 Manfaat Penelitian

1. Mengidentifikasi faktor dominan yang berkaitan dengan prestasi akademik mahasiswa berdasarkan hasil analisis model Random Forest dan XGBoost, sehingga memberikan pemahaman yang jelas mengenai faktor utama yang membedakan mahasiswa unggul dan mahasiswa reguler.
2. Menyediakan model klasifikasi berbasis machine learning yang mampu memetakan dan mengungkap pola hubungan antara berbagai variabel dengan prestasi akademik mahasiswa secara objektif dan berbasis data.
3. Memberikan dasar analisis yang dapat digunakan oleh Perguruan Tinggi untuk mendukung evaluasi dan perumusan strategi pembinaan akademik yang lebih tepat sasaran berdasarkan faktor-faktor yang teridentifikasi.
4. Mendukung pengambilan keputusan berbasis data dalam memahami karakteristik dan kondisi mahasiswa berdasarkan klasifikasi prestasi akademik yang dihasilkan oleh model.

