# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Pendidikan sebagai salah satu pilar utama yang dapat meningkatkan kualitas sumber daya manusia dalam pembangunan dan kemajuan suatu negara. Secara umum, pendidikan mencakup aspek akademik, seperti pengetahuan dan keterampilan, serta pendidikan non-akademik, yang meliputi pengembangan karakter, soft skills, dan keterlibatan dalam kegiatan sosial [1]. Namun, kenyataannya masih banyak kasus yang mana performa akademik tidak memenuhi kritetia dalam membantu menghasilkan pendidikan yang berkualitas.

Demografis sebagai salah satu factor utama yang mempengaruhi rendahnya performa akademik mahasiswa, sepert: faktor keluarga, termasuk kemiskinan dan tingkat pendidikan orang tua, secara signifikan mempengaruhi partisipasi pendidikan di antara para orang tua siswa [2]. Selain itu, manajemen waktu yang efektif berkorelasi kuat dengan kinerja akademik yang lebih tinggi, karena siswa yang merencanakan jadwal mereka cenderung mencapai hasil yang lebih baik [3]. Hal ini menunjukkan bahwa untuk memahami performa akademik secara holistik, kita perlu mempertimbangkan interaksi antara berbagai faktor yang memengaruhi mahasiswa.

Bebarapa studi sebelumnya seperti [4][5][6][7][8], telah melakukan eksperimen dalam menyelesaikan permasalah penelitian ini dibidang performa akademik. Sayangnya dalam penelitian tersebut belum menghasilkan akurasi yang cukup memuaskan. Terdapat beberapa masalah utama yang menyebabkan akurasi tidak memuaskan seperti ketidakseimbangan data dan *outlier* data. Menurut Liu Y. et al. (2024), ketidakseimbangan data secara signifikan dapat menyebabkan menurunnya akurasi klasifikasi *machine learning* khususnya dalam memprediksi kelas minoritas [9]. Masalah outlier juga masih menjadi sebuah tantangan dalam klasifikasi karena dapat mendistorsi distribusi data dan berdampak negatif pada kinerja model [10]. Sehingga bidang penelitian ini masih meninggalkan

permasalahan/tantangan besar untuk meningkatkan kinerja *machine learning* klasifikasi.

Metode *Random Over Sampling* (ROS) merupakan salah satu metode popular yang dapat mengatasi ketidakseimbangan data. Metode ini meningkatkan keseimbangan kelas dengan menduplikasi sampel kelas minoritas, meningkatkan stabilitas pelatihan model [11]. Selain itu, metode ini membantu menyeimbangkan distribusi kelas dengan menduplikasi contoh kelas minoritas, yang dapat meningkatkan kinerja *model* pada kumpulan data yang tidak seimbang [12].

Terdapat beberapa metode *machine learning* popular yang sering digunakan dalam bidang ini, seperti *K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes, Random Forest (RF), Decision Tree (DT), dan Logistic Regression (LG). K-Nearest Neighbor (KNN)* merupakan salah satu metode mac yang sangat sederhana. Sayangnya, metode ini mengalami kesulitan dalam ketidakseimbangan data [13]. *Support Vector Machine* (SVM) menjadi metode *machine learning* yang baik dalam penanganan data dimensi tinggi. Namun, metode ini masih memiliki sensivitas terhadap data outlier hingga ketidakseimbangan data [14]. *Naïve bayes* termasuk dalam metode *machine learning* yang memiliki pelatihan cepat dan kinerja yang baik pada kumpulan data besar. Akan tetapi, metode ini masih memiliki masalah dengan peningkatan bias karena asumsi yang salah [15].

*Random Forest* menjadi satu dari banyak metode *machine learning* dengan akurasi yang tinggi serta penanganan data kompleks. Tetapi, membutuhkan pelatihan yang lebih lama dibandingkan metode lain [16]. *Decision Tree* juga menjadi metode *machine learning* dengan interpretabilitas dan kesederhanaan yang mudah dipahami. Namun, memiliki kecenderungan kurang dalam penanganan outlier [17]. *Logistic Regression* menjadi metode *machine learning* yang sederhana dan interpretabilitas. Walaupun begitu, masih terdapat masalah dalam hubungan yang kompleks [18].

Dalam konteks ini, metode *ensemble* sebagai salah satu metode yang paling popular dan memiliki ketahanan terhadap *outlier*. Salah satunya adalah algoritma *Extreme Gradient Boost (XGBoost)*. *Extreme Gradient Boost (XGBoost)* merupakan metode *machine learning* yang dikenal kuat karena efisiensi dan kinerjanya dalam proses pemodelan [19]. Metode ini memiliki kemampuan untuk menangani interaksi yang kompleks dan hubungan *nonlinier* dalam data [20]. Selain itu, metode ini dapat mencegah terjadinya *overfitting* sehingga efektif bahkan dengan data pelatihan yang terbatas [21].

Dalam upaya untuk meningkatkan kualitas pendidikan, terdapat banyak tantangan yang dihadapi terkait performa akademik siswa yang masih belum optimal. Faktor demografis, ketidakseimbangan data, hingga keberadaan outliers menjadi faktor utama penyebab rendahnya akurasi *model* dalam klasifikasi performa akademik. Meskipun terdapat banyak metode *machine learning* namun akurasi dan ketahanan terhadap ketidakseimbangan data masih kurang memadai. Oleh karena itu, penelitian ini dilakukan untuk mengembangkan metode yang lebih baik menggunakan metode *ensemble* yang kuat dalam menangani data yang kompleks hingga mampu mengatasi *overfitting* data. Dengan menganalisis performa akademik lebih mendalam, penelitian ini tidak hanya akan mendukunng peningkatan kualitas pendidikan tetapi juga membantu institusi pendidikan dalam merancang kebijakan yang lebih adil sehingga dapat meningkatkan potensi mahasiswa dimasa yang akan datang.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang masalah diatas, terdapat beberapa rumusan masalah, antara lain:

1. Bagaimana mengatasi ketidakseimbangan data menggunakan metode *Random Over Sampling* (ROS) dapat meningkatkan akurasi penelitian ini?
2. Bagaimana perbandingan teknik *machine learning* dalam proses evaluasi dari segi akurasi, *presisi, recall, f1-score* dan *AUC*?

## 1.3 Batasan Masalah

Dalam penelitian ini, terdapat Batasan-batasan masalah yang ditentukan oleh peneliti, antara lain:

1. Dataset yang digunakan diperoleh pada website Kaggle secara public dengan nama Students Academic Performance Dataset pada tahun 2016 yang berbentuk CSV (*Comma-separated Values*).
2. Teknik *machine learning* yang digunakan pada penelitian ini ada 2 yaitu *machine learning ensemble* dan *non-ensemble*. Algoritma yang digunakan pada *machine learning ensemble* antara lain *Random Forest dan Extreme Gradient Boost* sedangkan algoritma yang digunakan pada *machine learning non-ensemble* antara lain *Logistic Regression dan Decision Tree*.
3. Teknik untuk mengatasi ketidakseimbangan data yang digunakan pada penelitian ini adalah *Random Over Sampling (ROS)*.
4. Teknik normalisasi data yang digunakan pada penelitian ini adalah *Min Max Scalar*.
5. *Hyperparameter tuning* yang digunakan pada setiap *model* bersifat *default*.
6. Teknik evaluasi validasi *model* dan performa akhir yang digunakan adalah *K-Fold Cross Validation* dan *Confution Matrix*

## 1.4 Tujuan Penelitian

Dalam penelitian ini terdapat beberapa tujuan yang ingin dicapai oleh peneliti, antara lain:

1. Untuk mengatasi ketidakseimbangan data menggunakan metode *Random Over Sampling* (ROS) dapat meningkatkan akurasi penelitian ini
2. Untuk mengetahui perbandingan teknik *machine learning* dalam proses evaluasi dari segi akurasi, presisi, *recall*, f1-score, dan AUC

## 1.5 Manfaat Penelitian

Dalam penelitian ini, terdapat beberapa manfaat yang ingin dicapai oleh peneliti, antara lain:

1. Manfaat praktis dalam penelitian ini adalah dapat meningkatkan akurasi dalam klasifikasi performa akademik khususnya menangani masalah ketidakseimbangan data dan *outlier*. Dengan hasil yang akurat, dapat menjadi acuan untuk Lembaga pendidikan dalam membuat kebijakan kedepannya.
2. Manfaat teoritis dalam penelitian ini adalah memperkaya literatur dibidang pendidikan khususnya terkait efektivitas metode *ensemble* dalam menghadapi data kompleks.

## 1.6 Sistematika Penulisan

Dalam pembuatan naskah penelitian digunakan sistematika penulisan yang disusun sedemikian rupa pada contoh dibawah ini:

BAB I PENDAHULUAN, berisi latar belakang masalah mengenai penelitian, rumusan masalah, batasan masalah, tujuan penelitian, dan manfaat penelitian

BAB II TINJAUAN PUSTAKA, berisi tinjauan pustaka dari penelitian sebelumnya, dasar-dasar teori yang digunakan seperti *machine learning*, dan algoritmanya.

BAB III METODE PENELITIAN, didalamnya terdapat rancangan alur penelitian beserta alat dan bahan yang digunakan dalam penelitian.

BAB IV HASIL DAN PEMBAHASAN, bab ini menjabarkan hasil penelitian dari proses pengumpulan data hingga proses evaluasi *model*

BAB V PENUTUP, berisi kesimpulan dan saran yang dapat peneliti rangkum selama proses penelitian