 ISSN:2302-8149
e-ISSN:2540-9719

Prediksi Kinerja Akademik Matematika Siswa berdasarkan Kepribadian Big Five menggunakan Random Forest dengan Teknik *Synthetic Minority Over-Sampling*

Predicting Students' Academic Performance in Mathematics based on Big Five Personality Traits using Random Forest with Synthetic Minority Over-Sampling Technique

[1]Annisa Nurul Pratiwi*, [2]Ema Utami
[1,2]Informatika Program Magister, Universitas Amikom Yogyakarta
[1,2]Jl Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281, Indonesia
*e-mail: annisanpratiwi@students.amikom.ac.id

(***received*:** 24 February 2025, ***revised*:** 1 March 2025, ***accepted*:** 1 March 2025)

Abstrak

Masa sekolah menengah merupakan periode penting untuk perkembangan kinerja akademik dan sosial siswa. Educational data mining (EDM) menjadi salah satu metode strategis yang mampu mengeksplorasi pola dalam data pendidikan untuk memprediksi kinerja akademik berdasarkan berbagai faktor, termasuk kepribadian siswa. Namun, ketidakseimbangan data pendidikan masih menjadi masalah yang dapat menyebabkan bias pada model prediksi. Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor yang berkontribusi terhadap kinerja akademik matematika siswa sekolah menengah pertama, seperti faktor akademik, demografis, dan kepribadian model Big Five. Metode Random Forest dan teknik oversampling SMOTE digunakan untuk mengidentifikasi komponen yang berkontribusi terhadap kinerja akademik siswa, serta meningkatkan performa model prediksi. Penelitian ini menunjukkan bahwa faktor akademik menjadi faktor penting, sementara faktor sosial-ekonomi dan kepribadian kurang signifikan terhadap kinerja akademik. Selain itu, penerapan teknik SMOTE terbukti efektif dalam mengatasi ketidakseimbangan data, serta model Random Forest memiliki performa optimal dengan tuning yang tepat. Kombinasi antara Random Forest, hyperparameter tuning GridSearchCV dan SMOTE berhasil mengembangkan model dengan tingkat akurasi mencapai 99%

Kata kunci: *Big Five*, *Educational* Data Mining, Kinerja Siswa, Random *Forest*, SMOTE

Abstract

The secondary school period is a crucial time for the development of students' academic and social performance. Educational data mining (EDM) has emerged as a strategic method capable of exploring patterns in educational data to predict academic performance based on various factors, including students' personalities. However, the imbalance in educational data remains an issue that can lead to bias in predictive models. This study aims to identify the factors contributing to the academic performance in mathematics of junior high school students, such as academic, demographic, and Big Five personality factors. The Random Forest method and SMOTE oversampling technique are employed to identify components that contribute to students' academic performance and to enhance the performance of the predictive model. The research indicates that academic factors are significant, while socio-economic and personality factors are less significant in relation to academic performance. Additionally, the application of the SMOTE technique proves effective in addressing data imbalance, and the Random Forest model demonstrates optimal performance with appropriate tuning. The combination of Random Forest, hyperparameter tuning using GridSearchCV, and SMOTE successfully develops a model with an accuracy rate of 99%.

Keywords: *Big Five, Educational Data Mining, Student Performance, Random Forest, SMOTE*

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Educational Data Mining (EDM) merupakan suatu disiplin ilmu yang memfokuskan diri dalam pengembangan sistem pendidikan, termasuk sekolah, universitas, dan sistem pembelajaran cerdas [1]. EDM memanfaatkan teknik berbasis komputer, seperti *data mining*, *machine learning*, dan metode statistik, untuk mengungkap pola-pola dalam data pendidikan yang cukup kompleks [2]. Penggunaan EDM dapat membantu dalam memodelkan proses pembelajaran, meningkatkan kualitas pengambilan keputusan, serta membantu sekolah dalam mengidentifikasi informasi untuk pemahaman siswa ke arah yang lebih baik [2]. Prediksi kinerja akademik menjadi salah satu elemen penting dalam EDM, karena memiliki peran dalam meningkatkan proses pembelajaran dan pemahaman siswa [2]. Topik prediksi ini menitikberatkan pada pemanfaatan data pembelajaran untuk mendeteksi siswa yang memiliki kinerja akademik rendah, memperbaiki metode pengajaran, serta meningkatkan prestasi akademik siswa [3][4][5]. Pada dasarnya dengan memprediksi kinerja siswa, institusi pendidikan dapat mengatasi kelemahan dalam sistem pendidikan dengan mengembangkan materi yang lebih responsif untuk meningkatkan prestasi siswa, serta memotivasi dan membangun kepercayaan diri siswa [4][6][7]. Walaupun prediksi kinerja akademik sudah sering dilakukan, penelitian mengenai prediksi kinerja akademik siswa di tingkat sekolah menengah masih berada pada tahap awal [6], terutama dalam mata pelajaran matematika [8]. Matematika menjadi tantangan bagi para siswa, karena para siswa sulit dalam memahami dan menerapkan konsep-konsep matematika. Oleh karena itu, mengidentifikasi faktor-faktor yang mempengaruhi kinerja akademik siswa menjadi langkah penting, terutama selama masa sekolah menengah khususnya yang berusia antara 10 hingga 18 tahun, yang merupakan periode kritis dalam perkembangan akademik dan sosial [9].

Banyak penelitian telah mengidentifikasi sejumlah faktor yang berfungsi sebagai indikator utama dalam kinerja akademik. Faktor akademik dianggap sebagai elemen utama [1], namun kombinasi antara faktor akademik dan variabel lainnya telah menunjukkan potensi untuk menghasilkan prediksi yang lebih tepat. Penelitian yang dilakukan oleh [10] merekomendasikan penggunaan model kepribadian Big Five sebagai indikator kinerja akademik. Model ini telah terbukti konsisten dan dapat diterapkan untuk menganalisis berbagai karakteristik kepribadian. Kepribadian Big Five mengelompokkan kepribadian ke dalam lima faktor atau domain utama, yaitu *Neuroticism* (N), *Extraversion* (E), *Agreeableness* (A), *Conscientiousness* (C), dan *Openness* (O), di mana setiap variabel ini diyakini memiliki pengaruh terhadap kinerja akademik [11][12][13]. Selain faktor kepribadian, kinerja akademik siswa juga diyakini dapat dipengaruhi oleh berbagai faktor pribadi lain, termasuk latar belakang keluarga, status sosial-ekonomi, dan lingkungan di sekitar mereka [9][13][14]. Sehingga, penting untuk memperhatikan faktor-faktor tersebut dalam memprediksi kinerja akademik siswa di sekolah menengah.

Para peneliti telah mengeksplorasi berbagai faktor yang dianggap mempengaruhi kinerja akademik dengan menggunakan berbagai teknik, seperti teknik *data mining* yang paling sering digunakan dalam penelitian EDM [5]. Algoritma Decision Tree, Naïve Bayes, dan Random Forest telah terbukti efektif dalam memprediksi kinerja siswa dengan tingkat akurasi yang cukup tinggi [1][10][14][15][16]. Namun, mengingat prediksi kinerja akademik yang melibatkan berbagai fitur yang mungkin saling terikat, metode Naïve Bayes dianggap kurang tepat karena ketidakmampuannya dalam menganalisis data secara efektif, di mana algoritma ini memiliki asumsi bahwa fitur-fitur bersifat independen [16][17]. Sementara itu, meskipun Decision Tree mampu mengidentifikasi pola-pola kompleks, metode ini juga tidak menjadi pilihan yang ideal karena masalah *overfitting* yang seringkali terjadi, terutama ketika berhadapan dengan dataset yang besar [18]. Oleh karena itu, Random Forest muncul sebagai pendekatan yang lebih sesuai untuk membangun model prediksi kinerja akademik, di mana Random Forest dapat meminimalisir risiko *overfitting* melalui teknik *boostrap sampling* dan pemilihan fitur secara random, serta menerapkan mekanisme *voting* untuk membuat keputusan di antara titik keputusan pohon tunggal, yang pada gilirannya menghasilkan prediksi yang lebih akurat [16][18]. Random Forest dianggap menjadi pilihan tepat dalam membangun model prediksi, karena memiliki jumlah *hyperparameter* yang lebih banyak dibandingkan dengan motede lainnya, yang secara signifikan dapat meningkatkan kinerja model [19]. Setiap parameter dalam Random Forest memiliki pengaruh yang signifikan terhadap kinerja model prediksi. Untuk mendapatkan parameter yang optimal, Random Forest dapat menerapkan proses *hyperparameter tuning* seperti Grid Search, yang berfungi untuk menyeimbangkan antara *overfitting*

 ISSN:2302-8149
e-ISSN:2540-9719

dan *underfitting*, serta meningkatkan akurasi dan generalisasi model [19][20][21]. Proses penyesuaian ini sangat penting untuk dilakukan, mengingat setiap parameter memiliki dampak yang signifikan terhadap hasil prediksi, sehingga tanpa penyesuaian yang tepat model berisiko tidak mencapai performa optimal [20].

Meskipun penelitian di bidang EDM telah menunjukkan kemajuan yang signifikan, masih terdapat tantangan mengenai ketidakseimbangan data [7]. Serta, meskipun Random Forest telah menunjukkan performa yang baik dalam pengembangan model prediksi, algortima ini memiliki keterbatasan ketika dihadapkan pada data yang tidak seimbang, karena dapat mengakibatkan bias serta model tidak tergeneralisasi dengan baik [1][15][16]. Ketika jumlah data dalam kelas mayoritas sangat dominan, model cenderung memberikan perhatian lebih pada kelas tersebut dan mengabaikan kelas minoritas, yang sering kali memiliki informasi yang tidak memadai [22]. Sehingga, penting untuk menggunakan dataset yang komprehensif, serta menangani ketidakseimbangan data, seperti menerapkan teknik sampling [7]. Metode *oversampling* seperti *Synthetic Minority Over-Sampling Technique* (SMOTE) telah terbukti efektif dalam mengatasi masalah ketidakseimbangan data dalam model prediksi kinerja siswa [22]. Penerapan SMOTE pada model Random Forest secara signifikan dapat meningkatkan akurasi, presisi, dan recall [23][24]. Secara garis besar, SMOTE merupakan teknik yang dapat mengatasi ketidakseimbangan data dalam EDM, karena dapat meningkatkan akurasi algoritma *supervised* [24] dan memperbaiki kinerja algoritma Random Forest dalam memprediksi kinerja siswa [23].

Penelitian ini bertujuan untuk mengidentifikasi berbagai faktor yang berkontribusi terhadap prestasi akademik siswa sekolah menengah pertama dengan fokus utama pelajaran Matematika, di mana kinerja akademik pada tingkat ini memiliki signifikasi yang cukup tinggi dalam upaya peningkatan kualitas pendidikan. Penelitian ini akan mengeskplorasi faktor akademik, demografis serta karakteristik kepribadian berdasarkan model Big Five dalam memprediksi kinerja siswa. Mengingat bahwa kinerja akademik siswa dipengaruhi oleh berbagai faktor pribadi [9], maka dataset yang diperlukan dalam penelitian ini mencakup informasi sensitif dan privasi. Sehingga, penggunaan dataset *private* dianggap lebih sesuai karena sifatnya lebih personal dan kontekstual, sedangkan dataset publik cenderung bersifat umum. Oleh karena itu, penelitian ini bertujuan untuk mengumpulkan data yang diperlukan melalui suatu lembaga pendidikan menengah pertama. Metode klasifikasi Random Forest juga akan diterapkan untuk mengklasifikasikan siswa berdasarkan kinerja akademik mereka ke dalam tiga kategori, yaitu cukup, baik dan sangat baik. Proses *hyperparameter tuning* akan dilakukan untuk menemukan kombinasi yang optimal, dengan menerapkan Grid Search. Untuk mengatasi ketidakseimbangan data, teknik *oversampling* SMOTE akan digunakan, yang pada gilirannya dapat mengatasi ketidakseimbangan serta meningkatkan akurasi model. Pendekatan ini diharapkan dapat memberikan pemahaman yang lebih baik mengenai faktor-faktor yang dapat diandalkan, yang bermanfaat bagi institusi pendidikan dalam mengidentifikasi siswa berisiko mengalami kegagalan dan meningkatan strategi pembelajaran. Serta, penerapan beberapa metode juga diharapkan dapat menciptakan sistem pendukung keputusan yang lebih baik dalam merancang pendekatan yang tepat untuk meningkatkan tingkat keakuratan hasil prediksi.

2 Tinjauan Literatur

Educational Data Mining (EDM) menggunakan berbagai metode analisis data untuk memprediksi kinerja siswa berdasarkan aktivitas belajar, dengan tujuan untuk meningkatkan kualitas pendidikan, yang berkontribusi pada peningkatan kinerja akademik siswa [2][3][4][6]. Namun, penelitian mengenai prediksi kinerja akademik di tingkat menengah masih berada pada tahap awal [6]. Adapun jenjang ini merupakan periode penting dalam perkembangan individu, di mana perubahan fisik, emosional dan sosial dapat mempengaruhi hasil akademik [9]. Oleh karena itu, dalam penelitian ini fokus utama berada pada prediksi kinerja akademik siswa dalam mata pelajaran matematika di tingkat sekolah menengah pertama yang diyakini memiliki peran signifikan terhadap perkembangan akademik siswa di masa depan. Khususnya di Indonesia, prestasi matematika siswa masih tergolong rendah [8].

Beberapa penelitian telah menunjukkan bahwa berbagai faktor telah mempengaruhi kinerja akademik, termasuk faktor akademik, demografis dan kepribadian. Sebuah penelitian [1] mengidentifikasi bahwa faktor akademik menjadi indikator utama yang mempengaruhi keberhasilan

 ISSN:2302-8149
e-ISSN:2540-9719

akademik. Penelitian lain menunjukkan bahwa selain faktor akademik, terdapat faktor kepribadian model Big Five yang dapat digunakan untuk meningkatkan kinerja akademik [10]. Penelitian tersebut menunjukkan dimensi *conscientiousness* dari model *Big Five* dan fitur akademik memiliki hubungan yang paling erat dengan kinerja akademik, serta menghasilkan kinerja model prediksi yang lebih optimal pada semua metriks evaluasi (akurasi, presisi, *recall* dan *f1-measure*). Penelitian lain menekankan bahwa dimensi-dimensi pada kepribadian Big Five, seperti *conscientiousness* [11][12][13], *openness* [12][13], dan *agreeableness* [12][13] memiliki hubungan paling signifikan dengan pencapaian akademik siswa. Namun, penelitian [13][14] telah menyoroti signifikasi untuk mempertimbangkan variabel demografis, termasuk jenis kelamin dan latar belakang sosial-ekonomi, yang berdampak pada kinerja akademik siswa. Sehingga, dalam penelitian ini penerapan faktor akademik, demografis, kepribadian model Big Five akan dilakukan guna memberikan kontribusi dalam melakukan analisis yang lebih dalam mengenai hubungan antara variabel-variabel tersebut dalam memprediksi kinerja akademik.

Penelitian [13] telah menekankan perlunya pengembangan metode baru yang lebih mutakhir dengan tetap memanfaatkan fitur yang relevan dengan kinerja akademik. Dalam konteks ini, teknik *data mining* dengan metode klasifikasi menjadi sangat relevan, mengingat temuan dari [5] telah menunjukkan bahwa metode klasifikasi merupakan pendekatan yang paling umum dan mudah dipahami, serta memiliki tingkat akurasi prediksi yang cukup baik. Selain itu, penelitian tersebut juga mengidentifikasi bahwa algoritma Naïve Bayes (NB), Decision Tree (DT) dan Random Forest (RF) menjadi yang paling umum digunakan untuk melakukan prediksi akademik.

Penelitian [10] yang berfokus pada pengembangan model prediksi kinerja akademik mahasiswa sarjana ke dalam tiga kategori (rendah, sedang, dan tinggi), menunjukkan bahwa akurasi model yang dihasilkan masing-masing mencapai 86% untuk NB, 83% untuk RF, dan 90% untuk DT, dengan fitur akademik dan dimensi kepribadian Big Five *conscientiousness* berkontribusi secara signifikan terhadap kinerja akademik. Selain itu, penelitian tersebut juga menekankan bahwa penghilangan fitur kepribadian menyebabkan penurunan performa model secara signifikan. Penelitian [16] juga berhasil menunjukkan bahwa model prediksi waktu kelulusan menggunakan RF memiliki tingkat akurasi yang lebih baik dibandingkan dengan NB, dengan akurasi masing-masing sebesar 76% (RF) dan 64% (NB), di mana fitur akademik menjadi indikator utama. Penelitian [15] juga menunjukkan bahwa RF dan DT menjadi model dengan performa terbaik dalam memprediksi kelulusan mahasiswa, namun RF tetap lebih unggul dibandingkan DT. Hasil penelitian tersebut menunjukkan bahwa model DT mengalami penurunan dari 99% menjadi 96% dalam *f-measure* saat proses validasi, sementara RF menunjukkan konsistensi dalam semua aspek, dengan tingkat akurasi stabil di tingkat 99%.

Kekonsistenan performa model Random Forest juga diungkapkan oleh [1], di mana akurasi model prediksi nilai suatu mata kuliah mencapai 90,33%. Hal ini juga ditunjukkan oleh [16][18], bahwa Random Forest dianggap sebagai solusi yang efektif untuk mengatasi atau setidaknya meminimalisir kekurangan dari Naive Bayes yang tidak efektif dalam menangani interaksi antar fitur [16][17], dan Decision Tree yang cenderung mengalami *overfitting* [18]. Apalagi diketahui bahwa Random Forest memiliki lebih banyak *hyperparameter* dibandingkan metode lainnya, yang dapat berdampak positif terhadap peningkatan kinerja model [19]. Maka, dalam penelitian ini Random Forest dipilih untuk mengembangkan model prediksi kinerja akademik, karena dianggap menjadi model prediksi dengan performa yang konsisten. Namun, salah satu penelitian menekankan perlunya peningkatan performa model Random Forest dengan mengatur parameter secara tepat [10]. Oleh karena itu, dalam penelitian ini penerapan teknik *hyperparameter tuning*, seperti *grid search* [19][20][21] pada Random Forest diterapkan guna mengidentifikasi kombinasi parameter yang paling efektif, sehingga pada gilirannya dapat meningkatkan performa model prediksi.

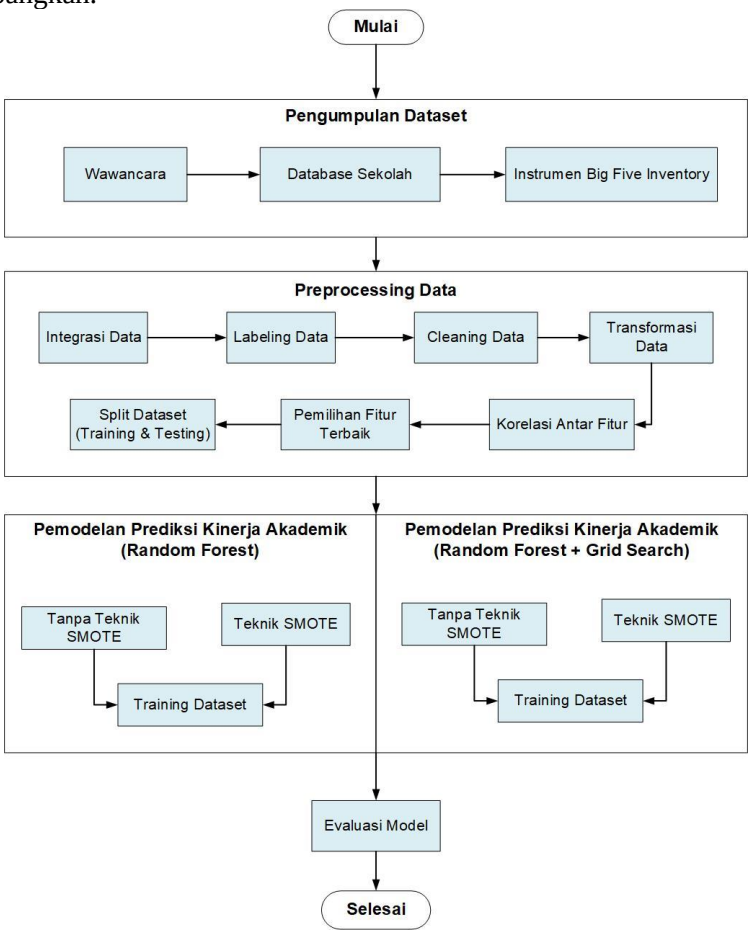
Beberapa penelitian telah menyoroti perlunya penanganan ketidakseimbangan data pada dataset agar performa model prediksi Random Forest menjadi lebih optimal [1][15][16]. Ketidakseimbangan dalam data perlu ditangani, karena dapat menyebabkan bias dan rendahnya akurasi pada kelas minoritas, yang pada akhrinya berdampak pada kinerja model [7][22]. Metode *oversampling* seperti SMOTE (*Synthetic Minority Over-sampling Technique*), bisa menjadi salah satu solusi untuk mengatasi masalah ketidakseimbangan data dengan menambah jumlah *instance* dari kelas minoritas dengan menciptakan *incstance* baru [22]. Penelitian [23] menunjukkan bahwa penerapan SMOTE dalam model prediksi kepribadian berhasil meningkatkan nilai presisi dan recall dari 72% dan 58% menjadi 79% dan 70%. Penelitian [24] juga menyoroti peranan penting SMOTE dalam meningkatkan

 ISSN:2302-8149
e-ISSN:2540-9719

akurasi Random Forest, di mana performa mencatat nilai presisi, *recall*, dan *f1-measur*e tertinggi. Selain itu, penerapan SMOTE pada Random Forest yang di-*tuning* menunjukkan peningkatan akurasi antara 1-3% dalam memprediksi mahasiswa yang berisiko gagal [19]. Berdasarkan temuan-temuan yang sudah dijelaskan, dengan tujuan untuk menangani ketidakseimbangan data pada model prediksi kinerja siswa, pendekatan menggunakan teknik SMOTE akan dilakukan agar tercipta titik data tambahan dalam dataset pelatihan guna mencapai keseimbangan kelas data dan performa model menjadi lebih baik.

3 Metode Penelitian

Penelitian ini melakukan eksperimen pada dataset yang telah dikumpulkan untuk mengidentifikasi karakteristik yang berkaitan dengan kinerja akademik siswa. Penelitian ini bertujuan untuk mengeksplorasi hubungan antara variabel yang digunakan, seperti fitur akademik, demografis dan fitur kepribadian terhadap kinerja akademik siswa. Penelitian ini juga bertujuan untuk mengevaluasi bagaimana performa model Random Forest dalam memprediksi kinerja siswa, serta mengevaluasi dampak penggunaan teknik SMOTE dalam mengatasi ketidakseimbangan dataset terhadap performa model prediksi menggunakan Random Forest. Pengembangan model terdiri dari empat tahapan utama (Gambar 1), mencakup pengumpulan data, *preprocessing data*, pemodelan prediksi dengan dua skenario utama di mana masing-masing skenario terdiri dari dua skenario yang sama, serta tahap akhir ialah melakukan evaluasi performa untuk mengevaluasi efektivitas dari model prediksi yang dikembangkan.



Gambar 1. Alur penelitian pemodelan prediksi kinerja akademik

3.1 Pengumpulan Dataset

Pengumpulan data dilakukan secara berjenjang, seperti melakukan wawancara dengan guru matematika sekolah menengah pertama. Langkah ini bertujuan untuk memperolah pemahaman menyeluruh tentang konteks dan dinamika kelas dalam lingkungan akademik. Sebanyak 793 sampel data siswa/siswi dari angkatan 2022 dan 2023 suatu lembaga pendidikan sekolah menengah pertama,

 ISSN:2302-8149
e-ISSN:2540-9719

berhasil dikumpulkan dari *database* pendidikan melalui operator yang bertugas. Data mencakup informasi demografis serta nilai akademik dalam mata pelajaran matematika.

Pengumpulan informasi mengenai kepribadian siswa dilakukan melalui penyebaran kuesioner secara bertahap kepada siswa/siswi, dengan memanfaatkan *Google Form*. Penyebaran kuesioner dipilih karena memberikan kontrol yang lebih besar terhadap data, sehingga memungkinkan peneliti untuk mengatur format dan standar pengisian sesuai dengan kebutuhan, sehingga menghasilkan data yang lebih relevan dan bersih [25]. Kuesioner yang digunakan dalam penelitian ini mengacu pada instrumen *Big Five Inventory* (BFI) yang dikembangkan oleh [26][27], terdiri dari 44 item versi bahasa indonesia yang digunakan untuk mengukur lima dimensi kepribadian pada model Big Five (*Neuroticism*, *Extraversion*, *Agreeableness*, *Conscientiousness*, dan *Openness*) dengan pengukuran skala *likert* 1-5. Penggunaan instrumen ini dipilih karena instrumen 44 big five item telah berulang kali divalidasi oleh komunitas peneliti [12]. Meskipun begitu, uji validitas dan reliabilitas [28][29] akan tetap dilakukan untuk memastikan kevalidan insturmen kepribadian Big Five yang digunakan. Validasi konstruk dilakukan setelah proses penyebaran selesai, dengan tujuan untuk menilai sejauh mana data dapat mengukur dimensi Big Five secara efektif, menggunakan metode analisis korelasi. Selanjutnya, pengujian reliabilitas juga akan dilakukan untuk memastikan konsistensi pengukuran setiap dimensi, menggunakn metode *Cronbach's Alpha*. (CA). Metode korelasi dan CA akan diimplempentasikan menggunakan *python*, dengan kriteria valid, nilai korelasi (r) harus di atas 0,3 serta nilai signifikasi atau *p_value* harus dibawah 0,05 dan CA harus di atas 0,6.

Data yang berhasil dikumpulkan dari database sekolah dan penyebaran kuesioner memiliki banyak atribut, seperti nama, jenis kelamin, jenis tinggal, pendidikan ayah, pekerjaan ayah, penghasilan ayah, pendidikan ibu, pekerjaan ibu, penghasilan ibu, anak ke, jumlah saudara kandung, status pernikahan orang tua, nilai harian 1, nilai harian 2, nilai harian 3, nilai PAS, nilai akhir. dan lima domain kepribadian model big five.

3.2 Preprocessing Data

Berdasarkan hasil penyebaran kuesioner, dari 793 siswa/siswi, 763 sampel berhasil dikumpulkan. Sehingga, 763 sampel akan diintegrasikan untuk menjadi satu dataset tunggal. Proses intgerasi dilakukan dengan menggabungkan nilai rata-rata jawaban responden di setiap dimensi pada data kuesioner kepribadian yang tervalidasi dengan data demografis dan akademik siswa yang diperoleh dari database sekolah. Proses pelabelan atau labeling data dilakukan dengan mengklasifikasikan data menjadi tiga kategori, yaitu cukup atau C (75-82), baik atau B (83-91) dan sangat baik atau SB (92-100). Proses klasifikasi atau *labeling* data dilakukan berdasarkan penilaian yang dilakukan langsung oleh guru matematika selaku seseorang yang mengamati dan berinteraksi saat proses pembelajaran berlangsung. Berdasarkan proses ini diketahui bahwa label B memiliki 425 data (55,7%), label SB memiliki 183 data (23,98%), dan label C memiliki 155 data (20,31%).

Pembersihan data atau *cleaning data* dilakukan dengan tujuan untuk menangani data yang tidak lengkap melalui berbagai proses, seperti menghapus atribut yang tidak relevan, mengisi data yang hilang, menghapus data yang tidak konsisten, serta mencari dan menghapus data duplikat [30]. Pada tahapan ini terdapat satu fitur yang dihapus karena tidak relevan, yaitu fitur nama. Transformasi data dilakukan terhadap beberapa atribut (Tabel 1), di mana data yang bersifat kategorikal (biner dan multi) akan diubah menjadi data numerik melalui metode label *encoding* [16][30]. Label *encoding* merupakan teknik pengolahan data yang mengkonversi data kategorikal menjadi data numerik dengan cara memberikan label angka (0-9) untuk masing-masing kategori [30].

Tabel 1. Informasi data siswa

Kategori	Fitur	Type	Value
Demografis	Jenis Kelamin	Kategorikal (Biner)	Perempuan (P) = 0 Laki-laki (L) = 1
	Jenis Tinggal	Kategorikal (Multi)	Bersama orang tua = 0 Wali = 1 Asrama = 2 Lainnya = 3
	Pendidikan Ayah	Kategorikal	Tidak Sekolah = 0

 ISSN:2302-8149
e-ISSN:2540-9719

	Pendidikan Ibu	(Multi)	SD/Sederajat = 1 SMP/Sederajat = 2 SMA/Sederajat = 3 Diploma = 4 S1 = 5 S2 = 6 S3 = 7 Putus Sekolah 8
	Pekerjaan Ayah Pekerjaan Ibu	Kategorikal (Multi)	Tidak Bekerja = 0 Sudah Meninggal = 1 Karyawan Swasta = 2 PNS/TNI/Polri = 3 Wiraswasta/Wirausaha = 4 Pedagang = 5 Lainnya = 6
	Penghasilan Ayah Penghasilan Ibu	Kategorikal (Multi)	Tidak Berpenghasilan = 0 Kurang dari Rp.500000 = 1 Rp.500000 s.d Rp.999999 = 2 Rp.1000000 s.d Rp.1999999 = 3 Rp.2000000 s.d Rp.4999999 = 4 Rp.5000000 s.d Rp. 20000000 = 5
	Anak Ke	Numerik	0-9
	Jumlah Saudara Kandung		0-9
	Status Pernikahan Orang Tua	Kategorikal (Multi)	Menikah = 0 Orang tua tidak tinggal bersama lagi (bercerai) = 1 Salah satu orang tua sudah tiada (meninggal) = 2
Akademik	Nilai Harian 1	Numerik	75-100
	Nilai Harian 2		75-100
	Nilai Harian 3		75-100
	Nilai PAS		75-100
	Nilai Akhir (Dependent)		Baik = 0 Cukup = 1 Sangat Baik = 2
Kepribadian Big Five	Extraversion	Numerik	1-5
	Agreebleness		1-5
	Conscientiousness		1-5
	Neuroticism		1-5
	Openess		1-5

Pengukuran korelasi fitur dilakukan dengan membangun matriks korelasi *pearson*. Korelasi dilakukan untuk mengukur hubungan antar fitur yang ada pada dataset. Sehingga, dapat diketahui fitur-fitur yang berpengaruh signifikan terhadap kinerja akademik siswa. Pemilihan fitur dilakukan berdasarkan nilai korelasi yang dimiliki antar fitur. Berdasarkan hasil analisis matriks korelasi terhadap fitur yang ada, di mana fitur-fitur yang menunjukkan nilai korelasi tertinggi akan dipilih untuk diterapkan dalam pemodelan prediksi yang akan dilakukan. Data yang telah dibersihkan akan dibagi dalam dua kategori, yaitu untuk melatih model mengenali pola (data latih/*training*) sebesar 70% dan mengevaluasi kinerja model (data uji/*testing*) sebesar 30%.

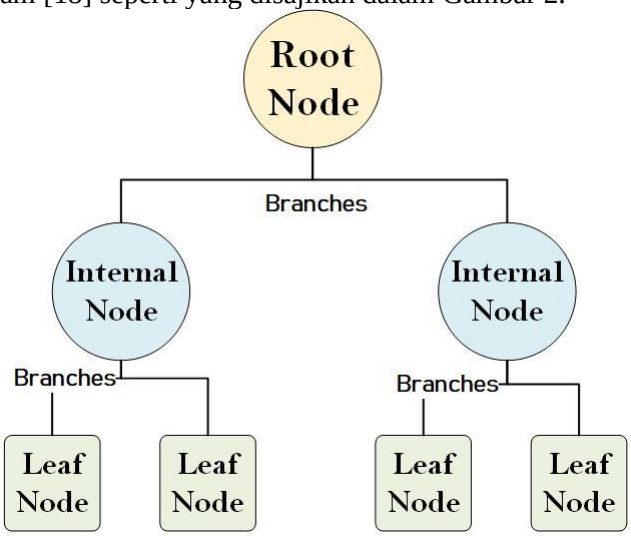
3.3 Pemodelan

Model dikembangkan menggunakan metode Random Forest dengan beberapa skenario. Skenario pertama akan menggunakan dataset asli yang tidak dimodifikasi, yang menunjukkan

 ISSN:2302-8149
e-ISSN:2540-9719

ketidakseimbangan kelas. Pada skenario kedua, akan diterapkan teknik SMOTE untuk menyeimbangkan dataset tersebut. Skenario ketiga dan keempat akan melibatkan penerapan *hyperparameter tuning* melalui grid search pada model yang menggunakan teknik SMOTE dan yang tidak, dengan tujuan untuk meningkatkan keakuratan prediksi dan performa dari model Random Forest. Skenario-skenario tersebut dilakukan dengan menggunakan dataset yang sama, serta performanya akan diuji secara terpisah, di mana hasil evaluasi akan dibandingkan untuk menentukan pendekatan yang paling efektif dalam menangani ketidakseimbangan kelas dan meningkatkan kinerja model prediksi.

Random Forest (RF) merupakan algoritma *ensemble*, yang juga dikenal sebagai *bootsrap aggregation* atau *bagging* [31]. Teknik statistik tersebut digunakan untuk memperkirakan parameter dari sebuah dataset berdasarkan sekumpulan sampel. *Boostrap* berarti memastikan bahwa setiap pohon keputusan dalam Random Forest dibangun dari kumpulan data yang bervariasi [32]. Proses *bagging* berarti melakukan penggabungan prediksi berbagai algoritma *machine learning* untuk meningkatkan akurasi, terutama pada algoritma dengan varians yang signifikan, di mana implementasi nya melibatkan pembuatan subsampel acak dari dataset dengan penggantian (*replacement*), sehingga elemen yang sama dapat terpilih lebih dari satu kali [32]. Algortima Random Forest menggabungkan beberapa Decision Tree [33] dengan serangkaian langkah, serta data dipecah menjadi subset yang seragam [18] seperti yang disajikan dalam Gambar 2.



Gambar 2. Prosedur decision tree

Pertama, terdapat *root node* yang merupakan titik awal dalam struktur pohon yang menggambarkan dataset. Kedua, algoritma akan mengidentifikasi fitur dan batas yang menghasilkan pembagian optimal berdasarkan kriteria tertentu, proses akan berlangsung secara berulang di mana setiap subset data akan dibagi lebih lanjut di node *internal node* sampai mencapai kriteria penghentian. Keputusan untuk pembagian di setiap *node* didasarkan pada rumus matematis seperti *information gain* yang merupakan rumus matematis yang didasarkan pada prinsip *entropy* dalam teori informasi [18]. Terakhir ada *node leaf* yang menggambarkan hasil atau label kelas. Konsep random forest melibatkan penggunaan sekumpulan probabilitas yang bervariasi untuk meningkatkan akurasi prediksi dan mengurangi risiko *overfitting*, dengan memanfaatkan sejumlah pohon keputusan yang memanfaatkan elemen keacakan selama proses pengambilan keputusan agar tidak bias dalam proses pengembangan [32]. Pada keputusan akhir, algoritma ini mengintegrasikan data dari berbagai titik menggunakan *majority voting*, untuk memastikan label dengan probabilitas lebih tinggi untuk dipilih sebagai keputusan akhir [33]. Secara sederhana terdapat empat langkah utama membentuk algoritma Random Forest, yaitu [33]:

1. Menggunakaan sampel acak sebanyak n dari kumpulan data latih dengan menerapkan *replacement (boostrap);*
2. Menggunakan sampel boostrap untuk membangun pohon keputusan, di mana fitur d dipilih secara acak tanpa *replacement* pada setiap simpul (*node*) pohon dan bagi simpul berdasarkan fitur yang memberikan pemisahan terbaik;

 ISSN:2302-8149
e-ISSN:2540-9719

3. Ulangi langkah pertama dan kedua sebanyak N kali;
4. Menggabungkan prediksi dari setiap pohon untuk menghasilkan label kelas dengan mayoritas suara terbanyak.

Grid Search merupakan metode optimasi *hyperparameter* yang digunakan untuk menentukan kombinasi nilai *hyperparameter* yang paling optimal. Metode *grid search* pada dasarnya melibatkan pengujian semua kombinasi parameter yang relevan guna menemukan parameter yang paling efektif untuk meningkatkan kinerja model [32]. Pendekatannya pun tergolong sederhana, karena melibatkan pencarian menyeluruh melalui metode *brute-force*, di mana pengguna menetapkan daftar nilai untuk berbagai *hyperparameter*, dan komputer akan menguji setiap kombinasi nilai untuk mengidentifikasi yang paling optimal [33]. Random Forest sendiri memiliki sejumlah parameter penting yang dapat disesuaikan untuk meningkatkan efektivitas model prediksi [32]. Parameter yang akan diterapkan dalam penelitian ini sendiri ialah *n_estimators*, yang menentukan jumlah pohon dalam hutan, *max_depth* mengatur kedalaman maksimum setiap pohon, *max_features* yang mengatur jumlah fitur yang dipertimbangkan pada setiap *split* untuk mengendalikan keragaman pohon, *min_samples_leaf* mengatur jumlah minimum sampel untuk memecah sebuah *node* dan *min_samples_split* yang mengatur jumlah minimum sampel pada setiap daun pohon.

SMOTE (*Synthetic Minority Over-sampling Technique*) merupakan metode yang dikembangkan untuk mengatasi ketidakseimbangan kelas dalam dataset dengan cara mengubah bias algoritma untuk kelas minoritas. Pada dasarnya metode ini menggunakan pendekatan berbasis *k-nearest neighbors* dengan cara menghitung jarak *euclidean* dan memilih tetangga terdekat dari contoh minoritas yang dipilih [34]. Proses pembuatan data sintetis dilakukan dengan mengidentifikasi kepadatan contoh dalam kelas minoritas, sehingga SMOTE dapat menemukan contoh nyata yang memberikan informasi penting mengenai label kelas minoritas serta outlier yang berpotensi mempengaruhi hasil secara negatif [35]. Dalam penelitian ini, pendekatan SMOTE dilakukan dengan menambahkan data tambahan pada kumpulan data latih yang minoritas [19]. Secara garis besar, data sintesis baru (D_{new}) untuk kelas minoritas dihasilkan dengan cara memilih sampel dari kelas tersebut (D_i) dan menghasilkan sampel baru di antara pilihan yang ada serta tetangga terdekat (D_l), tanpa menghapus sampel yang sudah ada, sehingga memastikan bahwa informasi tetap utuh dan tidak ada yang hilang. Persamaan terkait prosedur teknik SMOTE dapat dilihat dalam Persamaan 1 berikut [35] :

$$D_{new} = D_i + (D_l - D_i) \times w \quad (1)$$

Keterangan :
w = angka acak antara 0 dan 1

3.4 Evaluasi Model

Confusion matrix digunakan untuk menghitung seberapa sering sampel dari kelas A salah diklasifikasikan sebagai kelas B [36]. Artinya metriks ini memberikan representasi yang komprehensif terkait jumlah prediksi yang benar dan salah untuk setiap kelas [32]. *Confusion matrix* digambarkan sebagai sebuah tabel persegi yang mencatat jumlah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) pada prediksi klasifikasi. Ada beberapa cara untuk merangkum hasil dari *confusion matrix*, seperti *accuracy, precision, recall*, dan *F1-Score* [32]. *Accuracy* adalah jumlah prediksi yang benar, *Precision* menghitung berapa banyak sampel yang diprediksi positif ternyata positif, *Recall* menghitung semua sampel positif untuk menghindari negative palsu, dan *F1-Score* menggabungkan rata-rata harmonik dari *precision* dan *recall*. Metrik-metrik evaluasi tersebut dapat dinyatakan sebagai bentuk Persamaan 2, 3, 4, dan 5 berikut :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

 ISSN:2302-8149
e-ISSN:2540-9719

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

Receiver Operating Characteristic (ROC) merupakan metode evaluasi yang menilai kinerja model dengan menggambarkan tingkat positif asli terhadap tingkat positif palsu (*recall*) [32][33]. Metode ini tidak dipengaruhi oleh jumlah sampel dalam dataset, sehingga metode ini lebih akurat dalam mempred-iksi berbagai nilai ambang diskriminasi [37]. AUC (*Area Under ROC Curve*) merupakan alat ukur yang mengukur luas area bahwa kurva dengan interpretasi yang salah akan memiliki nilai ROC-AUC = 1, sementara interpretasi yang salah akan memiliki nilai ROC-AUC = 0,5 [32][33]. Semakin tinggi tingkat *True Positive* (TP), semakin banyak pula *False Positive* (FP) yang dihasilkan oleh pengklasifikasi. Metode ini akan menggambarkan bahwa kurva yang padat atau berada jauh dari garis putus-putus merupakan pengklasifikasi yang menunjukkan kinerja lebih baik. Perhitungan dari kurva ini dilakukan dengan menerapkan Persamaan 6 berikut :

$$F = \frac{FP\ (False\ Positive)}{TN\ (True\ Negative) + FP\ (False\ Positive)} \quad (6)$$

Cross Validation (CV) merupakan metodologi dasar dalam *machine learning* yang membagi data menjadi k-blok untuk pengujian dan pelatihan [38]. Artinya data akan dipartisi menjadi k-blok dan menggunakan k-blok tersebut untuk evaluasi. *K-folds validation* berarti membagi set pelatihan menjadi k-folds, kemudian melakukan prediksi dan evaluasi pada setiap *folds* menggunakan model yang dilatih pada setiap folds yang tersisa [36]. Metode ini memastikan tidak ada tumpang tindih antara set pengujian selama proses pengambilan sampel, di mana setiap *folds* merujuk pada jumlah subset yang dihasilkan [38]. Dalam penelitian ini *k-folds cross validation* yang akan digunakan dalam penelitian ini ialah *k-folds* = 5. Model akan menjalani proses pelatihan dan pengujian sebanyak lima kali, di mana pada setiap sesi, dua bagian akan digunakan untuk pelatihan dan satu bagian untuk pengujian, sehingga setiap bagian akan berfungsi sebagai data uji satu kali.

4 Hasil dan Pembahasan

Pengembangan model prediksi kinerja akademik dilakukan dengan menerapkan beberapa motede, guna menghasilkan performa model yang optimal. Sebelum melakukan evaluasi terhadap model secara keseluruhan, penting untuk menilai dataset yang diperoleh dari siswa dan siswi guna memastikan validitas serta reliabilitasnya. Dalam menganalisis data kuesioner yang telah dikumpulkan, beberapa metode statistik diterapkan [28][29]. Hasil analisis deskriptif dan pengujian reliabilitas menggunakan *Cronbach's Alpha* dapat dilihat pada Tabel 2 dan 3, yang menunjukkan bahwa data kuesioner dinyatakan valid.

Tabel 2. Statistik deskriptif data kuesioner kepribadian big five

Dimensi	Mean	Standar Deviasi	Min	Max	Median	Modus
Neuroticism	3,206	1,002	1	5	3	3
Extraversion	3,172	0,916	1	5	3	3
Agreeableness	3,554	0,867	1	5	4	3
Conscientiousness	3,152	0,862	1	5	3	3
Openness	3,337	0,909	1	5	3	3

Hasil analisis statistik dekriptif (Tabel 2) terhadap lima dimensi kepribadian Big Five, yaitu *neuroticism*, *extraversion*, *agreeableness*, *conscientiousness*, dan *openness*, menunjukkan bahwa setiap dimensi memperoleh nilai antara 1-5. Hal ini mengindikasikan distribusi yang merata pada skala yang diterapkan. Terdapat variasi yang signifikan dalam respon peserta, di mana tercermin dari nilai standar deviasi yang berkisar antara 0,8 hingga 1. Nilai modus dan median yang serupa untuk *neuroticism*, *extraversion*, *conscientiousness*, dan *openness* mengindikasikan bahwa mayoritas responden memiliki distribusi yang relative merata di sekitar nilai tengah. Nilai median 4 pada agreeableness, menunjukkan bahwa sebagian besar peserta cenderung mencerminkan sifat kooperatif dan kemampuan bergaul yang baik di antara responden. Secara keseluruhan, responden cenderung

 ISSN:2302-8149
e-ISSN:2540-9719

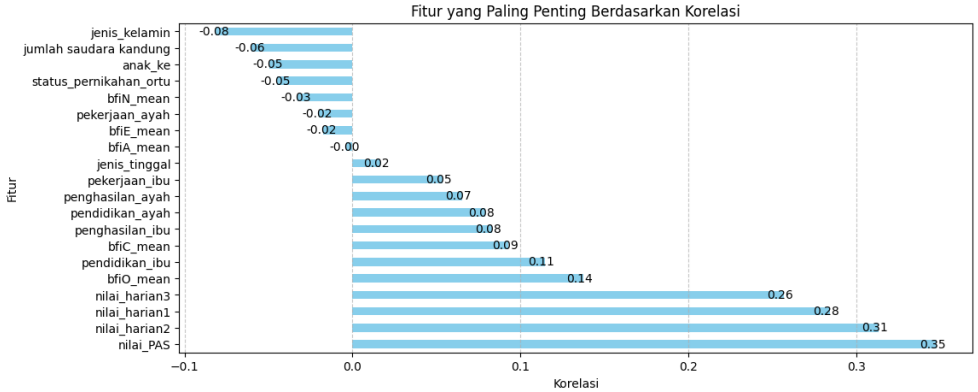
memperoleh skor yang berada pada tingkat sedang, dengan sedikit pergeseran menuju skor tinggi atau rendah.

Hasil dari analisis validitas juga menunjukkan bahwa seluruh item dalam setiap dimensi kepribadian dapat dinyatakan valid dengan nilai p=0,00, ini menunjukkan bahwa semua pernyataan dalam setiap dimensi memiliki signifikasi yang telah teruji kevalidannya. Untuk dimensi *extraversion* (E) yang terdiri dari 8 item, memiliki nilai korelasi (r) berkisar antara 0,55 hingga 0,67. Dimensi *agreeableness* (A) yang terdiri dari 9 item menunjukkan nilai r antara 0,52 hingga 0,57. Sementara itu, dimensi *conscientiousness* (C) yang terdiri dari 9 item memiliki nilai r antara 0,56 hingga 0,66. Pada dimensi *neuroticism* (N), yang terdiri 8 item memiliki nilai r berkisar antara 0,49 hingga 0,75. Terakhir, dimensi *openness* (O) yang terdiri dari 10 item menunjukkan nilai r berkisar antara 0,52 hingga 0,70. Temuan ini menunjukkan bahwa setiap item mampu merepresentasikan aspek kepribadian dengan cukup akurat. Hasil analisis reliabilitas menggunakan Cronbach's Alpha (CA), yang disajikan dalam Tabel 3 juga menunjukkan bahwa setiap dimensi memperoleh nilai yang cukup baik secara keseluruhan, di mana nilai reliabilitas untuk instrumen ini secara keseluruhan memperoleh nilai 0,81. Hal ini mengindikasikan bahwa instrument yang digunakan memenuhi kriteria yang dapat diterima. Nilai 0,6 dianggap memadai untuk tujuan eksplorasi, sedangkan nilai 0,7-0,8 dianggap cukup baik untuk tujuan konfirmasi [29]. Temuan ini semakin menegaskan bahwa instrumen yang diterapkan dalam penelitian ini memiliki kemampuan yang signifikan dalam mengukur berbagai aspek kepribadian model big five.

Tabel 3. Validasi konstruk dan realibility

Realibilitas	Neuroticism	Extra-version	Agree-ableness	Conscien-tiousness	Openness	Keseluruhan
Cronbach's Alpha	0,75	0,75	0,69	0,78	0,79	0,81

Membangun matriks korelasi dan ketergantungan antara fitur dilakukan untuk mengukur seberapa besar pengaruh satu fitur terhadap fitur lainnya, seperti yang ditunjukkan oleh Gambar 3. Korelasi yang kuat terjadi ketika dua fitur memiliki nilai positif maupun negatif yang menjauhi nol [10]. Sebaliknya, keitka nilainya mendekati nol maka kedua fitur tersebut berkorelasi rendah. Korelasi positif terjadi ketika peningkatan nilai pada suatu fitur diikuti oleh peningkatan nilai pada fitur yang berkorelasi. Sementara itu, korelasi negatif terjadi ketika peningkatan nilai pada satu fitur disertai dengan penurunan nilai pada fitur yang berkorelasi.



Gambar 3. Matriks korelasi

Berdasarkan hasil eksperimen yang telah dilakukan, penelitian ini menemukan bahwa fitur-fitur seperti, jenis_kelamin, jumlah_saudara_kandung, pendidikan_ibu, pendidikan_ayah, penghasilan_ayah, penghasilan_ibu, pekerjaan_ibu, nilai_PAS, nilai_harian1, nilai_harian2, nilai_harian3, bfiC_mean (*conscientiousness*) dan bfiO_mean (*openness*) dianggap sebagai variabel yang memiliki pengaruh terhadap model prediksi kinerja akademik siswa. Hal ini juga dapat dilihat pada hasil analisis hubungan yang disajikan dalam Gambar 3. Berdasarkan hubungan yang ada, dapat disimpulkan bahwa faktor akademik menunjukkan korelasi yang paling signifikan. Hal ini juga sejalan dengan temuan [1][10] yang menunjukkan bahwa diantara fitur lainnya, fitur akademik

 ISSN:2302-8149
e-ISSN:2540-9719

memiliki hubungan yang paling erat dengan kinerja akademik Dalam penelitian ini fitur nilai_PAS mencatat nilai korelasi positif tertinggi, diikuti oleh nilai_harian2, nilai_harian1, dan nilai_harian3. Sementara itu, dalam faktor kepribadian tidak semua dimensi memiliki korelasi yang signifikan terhadap kinerja akademik. Diketahui bahwa dimensi bfiC_mean (*conscientiousness*) dan bfiO_mean (*openness*) menunjukkan korelasi positif yang lebih baik dibandingkan dengan fitur kepribadian lainnya. Penelitian sebelumnya juga mendukung temuan ini, dengan menunjukkan bahwa siswa dengan nilai *conscientiousness* tinggi [11][12][13], menunjukkan individu yang lebih teratur dan disiplin cenderung mencapai prestasi akademik yang lebih baik. Selain itu, siswa dengan nilai *openness* yang tinggi [12][13], menunjukkan bahwa individu yang lebih terbuka terhadap pengalaman cenderung meraih prestasi akademik yang lebih baik. Untuk faktor demografis, ditemukan bahwa fitur jenis kelamin dan aspek sosial-ekonomi seperti pendidikan orang tua dan penghasilan orang tua memiliki dampak yang lebih besar dibandingkan dengan fitur demografis lainnya. Variabel seperti jenis_kelamin, pendidikan_ibu, pendidikan_ayah, penghasilan_ayah, penghasilan_ibu, pekerjaan_ibu menunjukkan korelasi yang cukup baik terhadap kinerja akademik. Temuan ini juga didukung oleh penelitian [13][14], yang menunjukkan bahwa faktor ekonomi dan pendidikan orang tua serta jenis kelamin memiliki korelasi paling signifikan dalam keberhasilan akademik Sementara itu, jenis_tinggal, pekerjaan_ayah, status_pernikahan_orangtua, jumlah saudara kandung dan anak_ke memiliki korelasi rendah, ini menunjukkan tidak ada pengaruh signifikan terhadap kinerja akademik. Temuan ini juga didukung oleh beberapa penelitian [10][14], di mana hasil penelitian [10] menunjukkan bahwa fitur jumlah saudara kandung memiliki nilai korelasi yang cukup rendah, sehingga ketika digunakan dalam pemodelan mengakibatkan penurunan kinerja model. Selain itu, penelitian [14] menemukan bahwa dalam model yang dikembangkan, status pernikahan orang tua hanya teridentifikasi satu kali dari delapan prediksi, sehingga dianggap tidak signifikan. Fitur pekerjaan ayah juga dianggap tidak terlalu signifikan, di mana penelitian [14] mengungkapkan bahwa tingkat pendidikan orang tua memiliki peranan yang lebih penting dibandingkan dengan status pekerjaan dalam menentukan prestasi akademik siswa.

Pengembangan model prediksi dilakukan dengan melakukan pembagian dataset dengan rasio data pelatihan sebesar 70% dan data pengujian sebesar 30%. Pemodelan Random Forest dilakukan dalam 4 skenario, yaitu model tanpa *tuning* dengan menerapkan SMOTE dan tanpa menerapkan SMOTE. Kemudian, model tuning (*GridSearchCV*) dengan menerapkan SMOTE, dan tanpa menerapkan SMOTE. Hasil eksperimen pada skenario yang berbeda, telah menunjukkan bahwa fitur nilai_PAS, nilai_harian1, nilai_harian3, dan nilai_harian2 memiliki bobot kepentingan (*feature importance*) yang konsisten di seluruh percobaan. Sementara itu, fiur lain seperti bfiO_mean, bfiC_mean, dan jenis kelamin menunjukkan dampak yang cenderung minimal. Di sisi lain, faktor sosial-ekonomi, termasuk pendidikan dan penghasilan orang tua, memiliki pengaruh yang paling rendah. Hal ini menunjukkan bahwa faktor akademik memiliki peran yang paling signifikan dalam menentukan hasil prediksi kinerja akademik dibandingkan dengan faktor lainnya. Temuan ini sejalan dengan beberapa penelitian yang menyatakan bahwa faktor akademik menjadi fitur paling signifikan [1][10][14] disusul oleh faktor demografis [1] dan kepribadian [10][14].

Berdasarkan hasil evaluasi model yang disajikan dalam Tabel 4, model Random Forest (RF) yang dikombinasikan dengan *GridSearchCV* dan SMOTE menunjukkan hasil terbaik, di mana nilai akurasi, precision, recall dan F1-Score tertinggi mencapai 0,99127, serta nilai ROC-AUC sebesar 0,99959. Model mendapatkan performa optimal melalui pengaturan parameter seperti, *n_estimators*=400, *max_depth*=15, *max_features*=sqrt, *min_samples_leaf*=1, dan *min_samples_split*=2, serta penerapan *k-folds*=5 dalam proses *cross validation*. Sementara itu, model Random Forest yang tidak menggunakan teknik SMOTE, menunjukkan bahwa pengaturan *tuning* saja justru sedikit menurunkan performa *cross validation* menjadi 0,97939. Namun, ketika teknik SMOTE diterapkan, performa *cross validtaion* kembali meningkat menjadi 0,98875. Sebagaimana terlihat dalam Tabel 4, penerapan teknik SMOTE tanpa *tuning* yang tepat ataupun pengaturan *tuning* tanpa SMOTE dalam dataset tidak seimbang, tidak menjamin model akan mencapai performa optimal. Hal ini sejalan dengan temuan [23], yang menunjukkan bahwa penggunaan SMOTE dalam model Random Forest pada data tidak seimbang, tidak meningkatkan performa model secara signifikan. Maka, menerapkan teknik SMOTE pada data tidak seimbang berperan penting dalam meningkatkan generalisasi model, terutama ketika dikombinasikan dengan *GridSearchCV* yang berfungsi untuk mengoptimalkan pemilihan *hyperparameter* pada Random Forest. Temuan ini juga didukung oleh

 ISSN:2302-8149
e-ISSN:2540-9719

penelitian [19] yang mengungkapkan bahwa penerapan SMOTE pada Random Forest dengan *tuning* menggunakan *GridSearchCV* dapat meningkatkan akurasi antara 1-3%. Dengan demikian, proses tuning sangat penting untuk mencapai keseimbangan performa model [21], tanpa penyesuaian yang tepat, Random Forest tidak dapat mencapai kinerja optimalnya [20].

Tabel 4. Evaluasi pemodelan kinerja akademik

Evaluasi	RF	RF+SMOTE	RF+ GridSearchCV	RF+GridSearchCV+ SMOTE
CV (k-folds=5) %	0,98690	0,98603	0,97939	0,98875
Accuracy %	0,98689	0,98689	0,99127	0,99127
Precision %	0.98704	0,98704	0,99127	0,99127
Recall %	0,98689	0,98689	0,99127	0,99127
F1 Score %	0,98694	0,98694	0,99127	0,99127
ROC-AUC %	0,99959	0,99942	0,99934	0,99959

Pada dasarnya penelitian ini berhasil mengatasi permasalahan ketidakseimbangan data [1][15][16], serta meningkatkan kinerja model Random Forest [10] melalui penerapan *hyperparameter tuning* menggunakan *GridSearchCV*. Selain itu, hasil penelitian ini menunjukkan bahwa memprediksi kinerja akademik siswa dapat memberikan manfaat signifikan bagi pendidik dan siswa dalam upaya meningkatkan efektivitas proses pembelajaran. Ini disebabkan oleh fakta bahwa model yang dikembangkan mampu mengidentifikasi hubungan kompleks antara variabel yang digunakan dalam memprediksi kinerja siswa. Sehingga, dengan menerapkan *educatioanl data mining*, institusi pendidikan dapat dengan mudah menemukan pola dalam data pendidikan yang tidak dapat dianalisis secara manual [2], yang pada gilirannya dapat membantu dalam mengidentifikasi siswa yang berpotensi memiliki kinerja rendah sejak awal, serta mencegah terjadinya kekurangan dalam sistem pendidikan [4][6]. Pada dasarnya, model yang dikembangkan akan memfasilitasi lembaga pendidikan dalam memahami kebutuhan siswa dengan lebih efektif, yang pada gilirannya dapat meningkatkan hasil belajar dan kinerja akademik mereka.

5 Kesimpulan

Penelitian ini menunjukkan bahwa faktor kepribadian dan sosial-ekonomi memiliki kontribusi terhadap kinerja akademik siswa, meskipun pengaruhnya tidak terlalu signifikan. Sehingga, untuk meningkatkan keakuratan prediksi, analisis mendalam terhadap faktor akademik yang lebih spesifik atau eksplorasi terhadap faktor-faktor tambahan lainnya yang lebih relevan dapat dilakukan. Penelitian ini juga menemukan bahwa kombinasi Random Forest, *GridSearchCV*, dan SMOTE merupakan metode yang paling efektif dalam kondisi di mana terdapat ketidakseimbangan kelas dalam dataset. Dengan demikian, teknik SMOTE terbukti efektif dalam mengatasi ketidakseimbangan data, dan penerapan *tuning* yang tepat memungkinkan model prediksi Random Forest untuk mencapai performa optimal. Meskipun pemodelan ini menunjukkan kinerja yang baik, penelitian ini juga memiliki beberapa keterbatasan, seperti ukuran dataset yang terbatas mengakibatkan hasil yang kurang dapat digeneralisasi. Hal ini disebabkan oleh fakta bahwa dataset yang digunakan hanya berasal dari satu sekolah negeri, yang memiliki jumlah siswa dan karakteristik pembelajaran yang terbatas. Sehingga, untuk meningkatkan generalisasi hasil kinerja model, disarankan agar pengambilan sampel dilakukan dengan tingkat keberagaman yang lebih tinggi. Selain itu, penyesuaian *hyperparameter* telah secara signifikan meningkatkan kinerja model, tetapi proses ini memakan waktu yang cukup lama, terutama dengan penggunaan *GridSearchCV* yang melibatkan berbagai kombinasi parameter. Oleh karena itu, untuk meningkatkan efisiensi tanpa mengorbankan performa, penting untuk mempertimbangkan penerapan teknik *tuning* lainnya yang dapat mempercepat pencarian parameter optimal. Secara keseluruhan, penelitian ini menyoroti bahwa masih ada banyak peluang untuk mengembangkan model prediksi kinerja akademik siswa, sehingga memerlukan penelitian lebih lanjut, yang pada gilirannya dapat membantu meningkatkan kualitas kinerja model prediksi dan kualitas sistem pendidikan. Penelitian di masa depan sebaiknya mempertimbangkan pendekatan yang

 ISSN:2302-8149
e-ISSN:2540-9719

lebih inovatif, seperti penggunaan metode deep learning atau penggabungan random forest dengan teknik ensemble lainnya untuk meningkatkan performa model secara keseluruhan.

Referensi (Reference)

[1] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "*Student Course Grade Prediction using the Random Forest Algorithm: Analysis Of Predictors' Importance*," *Trends Neurosci. Educ.*, vol. 33, p. 100214, 2023, doi: 10.1016/j.tine.2023.100214.

[2] O. Ozyurt, H. Ozyurt, and D. Mishra, "*Uncovering the Educational Data Mining Landscape and Future Perspective: A Comprehensive Analysis*," *IEEE Access*, vol. 11, no. October, pp. 120192–120208, 2023, doi: 10.1109/ACCESS.2023.3327624.

[3] S. M. Dol and P. M. Jawandhiya, "*Systematic Review and Analysis of EDM for Predicting the Academic Performance of Students"*, no. 0123456789. Springer India, 2024. doi: 10.1007/s40031-024-00998-0.

[4] I. Issah, O. Appiah, P. Appiahene, and F. Inusah, "*A Systematic Review of the Literature on Machine Learning Application of Determining the Attributes Influencing Academic Performance*," *Decis. Anal. J.*, vol. 7, no. February, p. 100204, 2023, doi: 10.1016/j.dajour.2023.100204.

[5] M. H. bin Roslan and C. J. Chen, "*Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021)*," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 5, pp. 147–179, 2022, doi: 10.3991/ijet.v17i05.27685.

[6] L. S. Rodrigues, M. Dos Santos, I. Costa, and M. A. L. Moreira, "*Student Performance Prediction on Primary and Secondary Schools-A Systematic Literature Review*," *Procedia Comput. Sci.*, vol. 214, no. C, pp. 680–687, 2022, doi: 10.1016/j.procs.2022.11.229.

[7] S. M. Dol and P. M. Jawandhiya, "*A Review of Data Mining in Education Sector*," *J. Eng. Educ. Transform.*, vol. 36, no. Special Issue 2, pp. 13–22, 2022, doi: 10.16920/jeet/2023/v36is2/23003.

[8] Wawan and H. Retnawati, "*Empirical Study of Factors Affecting the Students' Mathematics Learning Achievement*," *Int. J. Instr.*, vol. 15, no. 2, pp. 417–434, 2022, doi: 10.29333/iji.2022.15223a.

[9] A. Costa, D. Moreira, J. Casanova, Â. Azevedo, A. Gonçalves, Í. Oliveira, R. Azevedo, and P. C. Dias, "*Determinants of Academic Achievement from the Middle to Secondary School Education: A Systematic Review*," *Social Psychology of Education*, vol. 27, pp. 3533–3572, Jul. 2024, doi*:* 10.1007/s11218-024-09941-z.

[10] S. El-Keiey, D. ElMenshawy, and E. Hassanein, "*Student's Performance Prediction based on Personality Traits and Intelligence Quotient using Machine Learning*," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 292–299, 2022, doi: 10.14569/IJACSA.2022.0130934.

[11] J. Meyer, T. Jansen, N. Hübner, and O. Lüdtke, *Disentangling the Association between the Big Five Personality Traits and Student Achievement: Meta-Analytic Evidence on the Role of Domain Specificity and Achievement Measures*, vol. 35, no. 1. Springer US, 2023. doi: 10.1007/s10648-023-09736-2.

[12] J. R. Rico-Juan, C. Cachero, and H. Macià, "*Study Regarding the Influence of a Student's Personality and an LMS usage Profile on Learning Performance using Machine Learning Techniques*," *Appl. Intell.*, vol. 54, no. 8, pp. 6175–6197, 2024, doi: 10.1007/s10489-024-05483-1.

[13] F. S. E. Shaninah and M. H. Mohd Noor, "*The Impact of Big Five Personality Trait in Predicting Student Academic Performance*," *J. Appl. Res. High. Educ.*, vol. 16, no. 2, pp. 523–539, 2024, doi: 10.1108/JARHE-08-2022-0274.

[14] M. H. Bin Roslan and C. J. Chen, "*Predicting Students' Performance in English and Mathematics using Data Mining Techniques*," *Educ. Inf. Technol.*, vol. 28, no. 2, pp. 1427–1453, 2023, doi: 10.1007/s10639-022-11259-2.

[15] D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, and R. A. Abougalala, "*Prediction of Student Exam Performance using Data Mining Classification Algorithms*," *Educ. Inf. Technol.*, no. 0123456789, 2024, doi: 10.1007/s10639-024-12619-w.

[16] A. Santoso, H. Retnawati, Kartianom, E. Apino, I. Rafi, and M. N. Rosyada, "*Predicting Time*

 ISSN:2302-8149
e-ISSN:2540-9719

to Graduation of Open University Students: An Educational Data Mining Study," *Open Educ. Stud.*, vol. 6, no. 1, 2024, doi: 10.1515/edu-2022-0220.

[17] P. J. B. Pajila, B. G. Sheena, A. Gayathri, J. Aswini, M. Nalini, and R. Siva Subramanian, "*A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications*," *Proc. 4th Int. Conf. Smart Electron. Commun. ICOSEC 2023*, pp. 1228–1234, 2023, doi: 10.1109/ICOSEC58147.2023.10276274.

[18] I. D. Mienye and N. Jere, "*A Survey of Decision Trees: Concepts, Algorithms, and Applications*," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.

[19] J. Pecuchova and M. Drlik, "*Predicting Students at Risk of Early Dropping Out from Course using Ensemble Classification Methods*," *Procedia Comput. Sci.*, vol. 225, pp. 3223–3232, 2023, doi: 10.1016/j.procs.2023.10.316.

[20] F. Arden and C. Safitri, "*Hyperparameter Tuning Algorithm Comparison with Machine Learning Algorithms*," *Proceeding - 6th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. Appl. Data SCI. Artif. Intell. Technol. Environ. Sustain. ICITISEE 2022*, pp. 183–188, 2022, doi: 10.1109/ICITISEE57756.2022.10057630.

[21] Y. Rimal, N. Sharma, and A. Alsadoon, "*The Accuracy of Machine Learning Models Relies on Hyperparameter Tuning: Student Result Classification using Random Forest, Randomized Search, Grid Search, Bayesian, Genetic, And Optuna Algorithms*," *Multimed. Tools Appl.*, vol. 83, no. 30, pp. 74349–74364, 2024, doi: 10.1007/s11042-024-18426-2.

[22] S. D. A. Bujang, A. Selamat, O. Krejcar, F. Mohamed, L. K. Cheng, and P. C. Chiu, "*Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review*," *IEEE Access*, vol. 11, pp. 1970-1989, 2023, doi: 10.1109/ACCESS.2022.3225404.

[23] N. G. Ramadhan and Adiwijaya, "*Data Mining Techniques in Handling Personality Analysis for Ideal Customers*," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 8, no. 2, pp. 175–181, 2022, doi: 10.20473/jisebi.8.2.175-181.

[24] A. K. Hamoud, M. B. M. Kamel, A. S. Gaafar, A. S. Alasady, A. M. Humadi, W. A. Awadh, and J. M. Dahr, "*A Prediction Model based Machine Learning Algorithms with Feature Selection Approaches Over Imbalanced Dataset*," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 2, pp. 1105–1116, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp1105-1116.

[25] H. Jatnika, A. Waluyo, and A. Azis, "*A Comparative Study on Data Collection Methods : Investigating Optimal Datasets for Data Mining Analysis*," vol. 5, no. 1, pp. 16–23, 2024.

[26] O. P. John, E. M. Donahue, and R. Kentle, "*The Big Five Inventory--Versions 4a and 54*." CA: University of California,Berkeley, Institute of Personality and Social Research, Berkeley, 1991.

[27] O. P. John, L. P. Naumann, and C. J. Soto, "*Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues*," in *Handbook of personality: Theory and research*, O. P. John, R. W. Robins, and L. A. Pervin, Eds. New York, NY: Guilford Press, 2008, pp. 114–158.

[28] D. Budiastuti and A. Bandur, *Validitas dan Reliabilitas Penelitian*. Penerbit Mitra Wacana Media, 2018.

[29] G. D. Garson, *Validity & Reliability*. Statistical Publishing Associates, 2013.

[30] T. Gori, A. Sunyoto, and H. Al Fatta, "*Preprocessing* Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 215–224, 2024, doi: 10.25126/jtiik.20241118074.

[31] J. Brownlee, *Master Machine Learning Algorithms Discover how They Work and Implement Them from Scratch*. Machine Learning Mastery, 2016. [Online]. Available: https://machinelearningmastery.com/master-machine-learning-algorithms/

[32] A. C. Muller and S. Guido, *Introduction to Machine Learning with Python*, 1st Editio. O'Reilly Media, Inc, 2016. [Online]. Available: https://www.nrigroupindia.com/e-book/Introduction to Machine Learning with Python (PDFDrive.com)-min.pdf

[33] S. Raschka and V. Mirjalili, *Python Machine Learning*, 2nd Editio. Packt Publishing Ltd, 2017. [Online]. Available: http://radio.eng.niigata-u.ac.jp/wp/wp-content/uploads/2020/06/python-machine-learning-2nd.pdf

[34] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "*SMOTE : Synthetic Minority Over-sampling Technique*," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi:

 ISSN:2302-8149
e-ISSN:2540-9719

10.1613/jair.953.

[35] N. A. Azhar, M. S. Mohd Pozi, A. M. Din, and A. Jatowt, "*An investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis*," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6651–6672, 2023, doi: 10.1109/TKDE.2022.3179381.

[36] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd Editio. O'Reilly Media, Inc., 2019. [Online]. Available: https://powerunit-ju.com/wp-content/uploads/2021/04/Aurelien-Geron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-OReilly-Media-2019.pdf

[37] G. Hackeling, *Mastering Machine Learning with Scikit-Learn*. Packt Publishing Ltd, 2014. [Online]. Available: https://www.amazon.com/Mastering-Machine-Learning-Scikit-learn-Algorithms/dp/1783988363

[38] S. Rajvanshi, G. Kaur, A. Dhatwalia, Arunima, A. Singla, and A. Bhasin, *Research on Problems and Solutions of Overfitting in Machine Learning*, vol. 1191 LNEE. Springer Nature Singapore, 2024. doi: 10.1007/978-981-97-2508-3_47.