IJAI (Indonesian Journal of Applied Informatics)
Vol. 9 No. 1 Tahun 2024 pISSN: 2548-3846, eISSN: 2598-5981

Explainable Machine Learning dalam Analisis Risiko Akademis Mahasiswa Fakultas Vokasi Institut Teknologi Sepuluh Nopember

Lovinki Fitra Ananda[1], **Mukti Ratna Dewi[1]***, Mochammad Reza Habibi[1]
[1]Departemen Statistika Bisnis, Fakultas Vokasi, Institut Teknologi Sepuluh Nopember
*Email: mukti_ratna@its.ac.id

Info Artikel

Kata Kunci :
pembelajaran mesin yang dapat dijelaskan, performa mahasiswa, *random forest*, risiko akademis, SHAP

Keywords :
explainable machine learning, student performance, random forest, academic risk, SHAP

Tanggal Artikel
Dikirim : 31 Oktober 2024
Direvisi : 15 November 2024
Diterima : 17 November 2024

Abstrak

Mahasiswa dengan performa akademis yang buruk dan tingkat *drop out* yang relatif tinggi dapat memengaruhi akreditasi dan citra institusi pendidikan tinggi. Hal tersebut dapat diantisipasi dengan cara mengevaluasi kondisi akademik mahasiswa, khususnya pada mahasiswa yang menunjukkan penurunan performa akademis. Penelitian ini bertujuan memberikan informasi mengenai faktor-faktor yang memengaruhi risiko akademis mahasiswa menggunakan *explainable machine learning*. Persentase mahasiswa yang berisiko akademis hanya sebesar 7,3% sehingga kasus *imbalance* ini perlu ditangani menggunakan SMOTE untuk mengoptimalkan kinerja model klasifikasi. Model *random forest* pada data yang telah seimbang memiliki kemampuan prediksi dengan tingkat akurasi 96,4%, *specificity* mencapai 95%, dan nilai *recall* atau *sensitivity* sebesar 98%. Selanjutnya, SHAP diimplementasikan untuk mengetahui kontribusi masing-masing faktor terhadap potensi risiko akademik. Hasil dari SHAP menunjukkan bahwa skor TPKA kuantitatif, diikuti oleh jenis kelamin dan jalur masuk memiliki kontribusi paling tinggi terhadap risiko akademis mahasiswa.

Abstract

Students with poor academic performance and relatively high dropout rates can affect the accreditation and image of higher education institutions. This can be anticipated by evaluating the academic conditions of students, especially those who show a decline in academic performance. This study uses explainable machine learning to provide information on the factors that influence students' academic risk. The percentage of students at academic risk is only 7.3%, so this imbalance case needs to be handled using SMOTE to optimize the performance of the classification model. The random forest model on balanced data has a predictive ability with an accuracy level of 96.4%, specificity reaching 95%, and a recall or sensitivity value of 98%. Furthermore, SHAP is implemented to determine the contribution of each factor to the potential academic risk. The results of SHAP show that the three most significant contributing factors to students' academic risk are the quantitative TPKA score, followed by gender and type of student admission.

1. PENDAHULUAN

Fakultas Vokasi Institut Teknologi Sepuluh Nopember (ITS) yang berdiri sejak 26 Januari 2017 terus melakukan perbaikan dan evaluasi untuk menjaga akreditasi dan citra positifnya guna menarik minat mahasiswa baru. Kualitas mahasiswa menjadi indikator utama kesuksesan institusi perguruan tinggi. Mahasiswa yang gagal memenuhi kriteria akademik berpotensi mengalami *drop out* (DO) yang dapat berdampak negatif pada akreditasi dan citra fakultas.

Untuk mengurangi risiko DO, penting bagi pihak universitas melakukan deteksi dini terhadap kondisi akademik mahasiswa yang menunjukkan penurunan performa.

Fakultas Vokasi ITS menerapkan sistem paket dan kebijakan Tidak Naik Semester (TNS) ketika mahasiswa gagal memenuhi kriteria evaluasi dalam satu semester. Mahasiswa yang TNS diharuskan mengambil cuti studi selama satu semester atau disebut cuti TNS. Pada Semester Gasal 2022/2023, tercatat sebanyak 42 mahasiswa yang dinyatakan TNS, meningkat 23,5% dari semester sebelumnya yang berjumlah 34 mahasiswa. Hal ini secara tidak langsung juga meningkatkan potensi seorang mahasiswa mengalami DO karena cuti TNS akan memperpanjang masa studi mahasiswa. Apabila masa studi mahasiswa melebihi empat belas semester maka mahasiswa tersebut dinyatakan gagal studi. Oleh karena itu, mahasiswa yang menunjukkan penurunan performa dengan TNS, menempuh lebih dari delapan semester atau memiliki IPK ≤2,75 dianggap berisiko akademis. Untuk mengantisipasi risiko akademis, perlu dikenali faktor-faktor yang memengaruhi performa akademis mahasiswa sejak dini. Penelitian-penelitian yang telah dilakukan di berbagai negara menunjukkan bahwa IPK, jalur masuk perguruan tinggi, jenis kelamin, asal sekolah, usia memulai studi serta biaya kuliah memengaruhi performa akademis mahasiswa dan kecenderungan DO [1], [2], [3], [4].

Kasus yang umum terjadi padapemodelan *machine learning* adalah ketidakseimbangan antar kelas (*imbalanced data*). Umumnya mahasiswa berisiko akademis, seperti pada [5] berjumlah sangat sedikit dibandingkan dengan mahasiswa yang tidak berisiko akademis. Adanya ketidakseimbangan data dapat memengaruhi kebaikan model dalam melakukan prediksi, di mana model akan cenderung baik dalam memprediksi kelas mayoritas dan buruk dalam memprediksi kelas minoritas [6]. Penanganan *imbalanced data* telah terbukti secara konsisten meningkatkan performa dan akurasi model, terutama pada dataset dengan ketidakseimbangan tinggi [7], [8]. Lebih lanjut, [9] menyatakan bahwa metode *oversampling* menghasilkan performa model klasifikasi yang lebih baik dibandingkan *undersampling.* Berdasarkan hal tersebut, penelitian ini mengaplikasikan *Synthetic Minority Oversampling Technique* (SMOTE) dalam penanganan *imbalanced data*.

Prediksi yang dilakukan oleh [5] menunjukkan bahwa *decision tree* lebih unggul dibandingkan *Artificial Neural Network* (ANN) dan *Bayesian Networks* dalam memprediksi kecenderungan *drop-out* mahasiswa. Namun, penelitian tersebut kurang menangani masalah ketidakseimbangan data, yang berdampak pada rendahnya precision dan F1 score untuk kelas minoritas. Meskipun *decision tree* efektif dalam klasifikasi, metode ini memiliki cenderung *overfit* dan terbatas dalam memprediksi di luar *training data*. *Random forest*, sebagai pengembangan dari *decision tree*, dapat memberikan prediksi yang lebih akurat, mengukur pentingnya fitur, serta mengukur kedekatan pasangan sampel dalam data pelatihan [10]. Temuan [8] menunjukkan penggunaan *Random Forest* yang dioptimalkan dengan teknik *resampling*, seperti SMOTE, dapat menghasilkan model yang lebih akurat, terutama pada dataset dengan ketidakseimbangan tinggi. Namun, penelitian-penelitian ini belum banyak mengeksplorasi penggunaan SHAP untuk memberikan penjelasan lebih mendalam terhadap hasil prediksi model *Random Forest* dalam konteks risiko akademis mahasiswa. Di sinilah SHAP (*Shapley Additive Explanations*) memainkan peran penting. SHAP membantu menjelaskan mengapa model memprediksi seorang mahasiswa berisiko atau tidak. Dengan kata lain, SHAP dapat menduga apa saja faktor-faktor yang dominan untuk memprediksi risiko akademis. Penelitian ini bertujuan untuk mengidentifikasi risiko akademis mahasiswa Fakultas Vokasi ITS dan mengatasi kekurangan penelitian sebelumnya dengan menerapkan SMOTE untuk penanganan data tidak seimbang, *Random Forest* untuk membangun model prediksi, serta interpretasi model menggunakan SHAP untuk menjelaskan faktor-faktor yang mempengaruhi keputusan model [11], [12].

Secara spesifik, objektif dari penelitian ini adalah sebagai berikut:

1. Membentuk dan mengevaluasi kebaikan model prediksi risiko akademis menggunakan *random forest*.
2. Mengetahui faktor-faktor utama yang memengaruhi performa dan risiko akademis mahasiswa Fakultas Vokasi ITS.

2. TINJAUAN PUSTAKA

2.1. Risiko Akademis

Berdasarkan Peraturan Rektor ITS Nomor 13 Tahun 2019 tentang Peraturan Akademik Program Vokasi ITS Pasal 14 ayat (6) menyatakan bahwa mahasiswa dinyatakan tidak lulus atau gagal studi apabila masa tempuh studi lebih dari empat belas semester. Keterlambatan studi memperbesar peluang seseorang untuk gagal studi. Pada

mahasiswa Fakuktas Vokasi ITS, risiko keterlambatan studi dapat diperparah bila mahasiswa tersebut mengalami Tidak Naik Semester (TNS). Pada Pasal 10 ayat (12) disebutkan bahwa mahasiswa dinyatakan TNS apabila memiliki nilai D atau E dan pada semester selanjutnya, mahasiswa tersebut wajib mengambil cuti TNS. Selain itu, pada Pasal 10 ayat (9) disebutkan bahwa mahasiswa dengan IPK ≤2,75 lulus tanpa predikat. Oleh karena itu, pada penelitian ini mahasiswa dikatakan memiliki risiko akademis apabila memenuhi setidaknya satu dari ketiga kriteria berikut:

1. Pernah tidak naik semester (TNS)
2. Menempuh masa studi lebih dari 8 semester
3. Memiliki IPK ≤ 2,75.

2.2. Penanganan *Imbalanced Data*

Imbalanced data adalah kondisi data yang tidak seimbang dengan jumlah data suatu kelas jauh melebihi jumlah data kelas yang lain [13]. Dalam *pembelajaran mesin*, data yang tidak seimbang akan menghasilkan model klasifikasi yang bias dan memiliki kecenderungan menghasilkan akurasi yang buruk pada kelas minoritas [14]. Penanganan data tak seimbang dapat dilakukan dengan mengurangi data sampel pada kelas mayoritas (*undersampling*) atau menambah jumlah data pada kelas minoritas (*oversampling*) [15]. Pengurangan jumlah data akan mengurangi keseluruhan informasi yang ada, sedangkan replikasi data cenderung akan mengakibatkan model menjadi *overfitting.*

Pada penelitian ini, penanganan *imbalanced data* dilakukan dengan metode *Synthetic Minority Oversampling Technique* (SMOTE) dikarenakan memiliki efektifitas yang lebih baik dibandingkan metode lain seperti ADASYN [16]. *SMOTE* merupakan metode *oversampling* dengan prinsip mencari *k-nearest neighbor* (kedekatan data) untuk setiap data di kelas minoritas. Metode ini membuat *synthetic data* atau replikasi data sebanyak persentase duplikasi yang diinginkan antara data kelas minoritas dan *k-nearest neighbor* yang dipilih secara *random* dengan formulasi pada persamaan (1) [17], [18].

$$x_{syn} = x_i + (x_{knn} - x_i)\delta \quad (1)$$

di mana x_i adalah data yang akan direplikasi, x_{knn} adalah data yang paling dekat dengan x_i, yang ditentukan oleh jarak *Euclidean* dengan rumus $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ serta δ merupakan bilangan *random* yang bernilai antara 0 sampai 1.

2.3. *Random Forest*

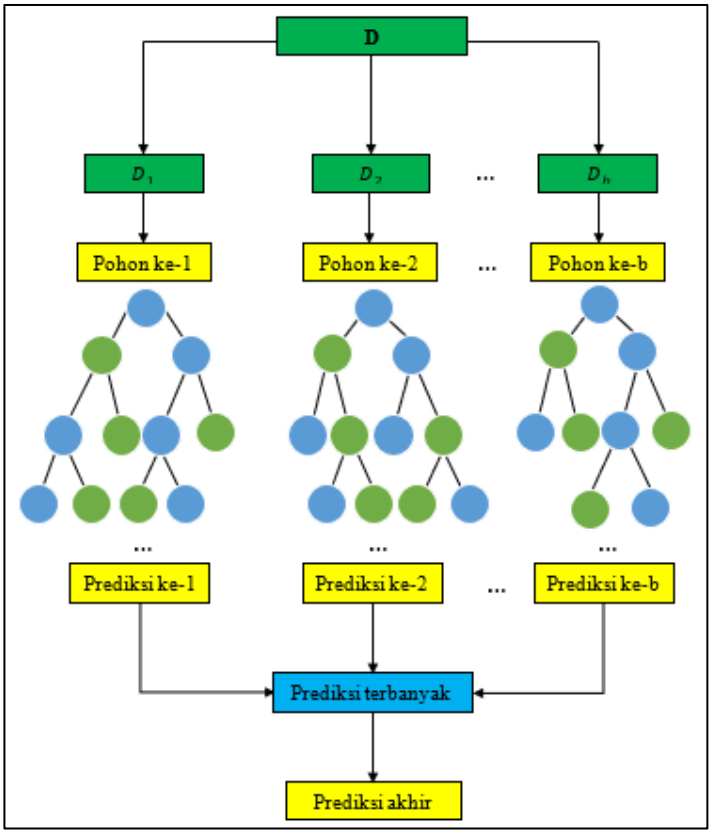
Random forest merupakan metode pembelajaran mesin yang mengombinasikan beberapa *decision tree*. Metode ini mengambil keputusan berdasarkan mayoritas keputusan dari *decision tree*. Model *random forest* dapat terdiri dari puluhan bahkan ratusan *decision tree* sehingga memiliki distribusi tingkat kesalahan yang kecil dengan waktu proses yang lebih lama [19], [20]. Tahapan penyusunan model *pembelajaran mesin* menggunakan algoritma *random forest* adalah sebagai berikut:

1. *Bootstrapping*
   Random Forest dimulai dari tahap *bootstrapping*, di mana dataset D dibagi menjadi beberapa dataset yakni D_1, D_2 hingga D_b. Pengambilan sampel dilakukan secara acak dengan pengembalian.
2. *Random Feature Selection*
   Pada masing-masing dataset yang baru dibentuk kemudian dilakukan pemilihan *m* variabel yang dilakukan secara acak dari *p* variabel dimana $m \leq p$. Pada tahapan ini, *decision tree* dibangun hingga mencapai ukuran maksimum tanpa pemangkasan.
3. Pembentukan *Decision Tree*
   Atribut sebagai *root node* dipilih sebagai langkah awal pembentukan struktur *decision tree*. Pemilihan root node didasarkan pada nilai indeks gini terbesar dari masing-masing variabel yang ada. Node akan terus bercabang hingga hingga indeks gini telah bernillai 0. Hal ini kemudian akan membentuk suatu *decision tree*. Nilai indeks gini dapat dihitung menggunakan persamaan (2).

$$Gini(H) = 1 - \sum_{i-1}^{k} (p_i)^2 \quad (2)$$

   di mana *H* merupakan himpunan dataset, p_i adalah proporsi kelas ke-*i* dalam dataset, sedangkan *k* adalah jumlah kelas dalam dataset. Rentang dari indeks Gini adalah 0 (*purity*) sampai 1 (*impurity*). Indeks Gini bernilai 0 jika seluruh anggota dari termasuk dalam kelas yang sama.

4. Langkah 2 dan 3 diulang sebanyak *b* (sejumlah dataset yang terbentuk pada langkah 1). Setiap *decision tree* akan menghasilkan suatu nilai prediksi, dari *decision tree* sejumlah *b*, prediksi terbanyak lah yang akan dipilih menjadi prediksi akhir. Tahapan ini dijelaskan secara sederhana pada Gambar 1.



Gambar 1. Tahapan *Random Forest*

2.4. *Shapley Additive Explanation* (SHAP)

Shapley Additive Explanation (SHAP) bertujuan untuk menghitung kontribusi kepentingan dari setiap variabel terhadap model. SHAP merupakan suatu metode yang digunakan untuk menginterpretasikan model pembelajaran mesin. Konsep SHAP berasal dari teori Shapley yang diperkenalkan pada tahun 1953 oleh Lloyd Shapley yang digunakan untuk memberi solusi saat menentukan hadiah yang adil bagi setiap pemain pada suatu permainan dengan tujuan mencari nilai terbaik diantara pemain, tergantung seberapa penting kontribusi yang telah dilakukan [12]. Rumus dari SHAP dapat dilihat pada persamaan (3).

$$\emptyset_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!\,(|M| - |S| - 1)!}{|M|!} \left(v(S \cup \{i\}) - v(S)\right), i = 1,2,\dots,|M| \quad (3)$$

di mana M merupakan koalisi dari semua atribut serta *S* merupakan koalisi tanpa kehadiran atribut ke-*i*. Sementara itu $v(S \cup \{i\})$ adalah performa model dengan atribut ke-*i* dan $v(S)$ adalah performa model tanpa kehadiran atribut ke-*i* dengan *i* bernilai 1 sampai |M|. Nilai SHAP ($\emptyset_j$) menggambarkan rata-rata kontribusi marginal dari atribut *i* pada semua kemungkinan himpunan bagian *S*.

3. METODE PENELITIAN

3.1. Sumber Data dan Variabel Penelitian

Berdasarkan penelitian terdahulu serta mempertimbangkan kondisi mahasiswa Fakultas Vokasi ITS, penulis merumuskan variabel-variabel yang digunakan dalam penelitian ini dalam Tabel 1. Data penelitian ini merupakan data

sekunder yang diperoleh dari Direktorat Pendidikan ITS, Subdirektorat Koordinasi Perkuliahan Bersama (SKPB ITS) serta Direktorat Perencanaan dan Pengembangan ITS. Ruang lingkup penelitian ini dibatasi hanya pada mahasiswa aktif program sarjana terapan Fakultas Vokasi Institut Teknologi Sepuluh Nopember angkatan 2017 hingga 2022. Pengambilan data dilakukan pada semester genap tahun akademik 2022/2023 dengan jumlah total data yang digunakan berjumlah 2.496 mahasiswa.

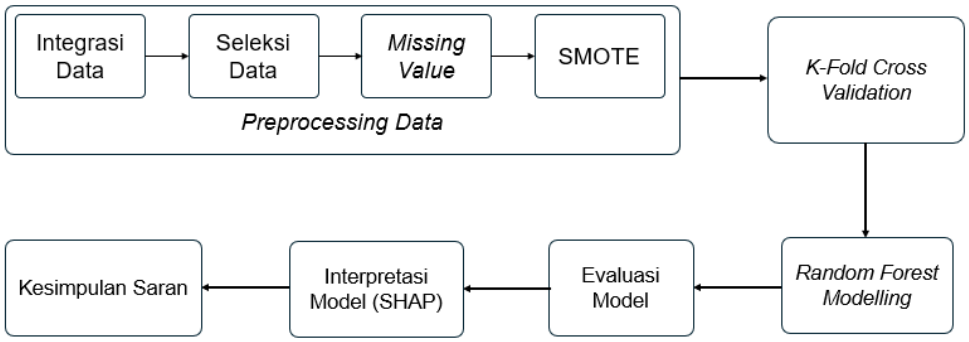
Tabel 1. Variabel Penelitian

Variabel	Satuan/ Kategori/ Rentang
Risiko Akademis	0: Tidak berisiko, 1: Berisiko
Jenis Kelamin	0: Laki-laki, 1: Perempuan
Lama *Gap Year*	0: Diterima di ITS pada tahun kelulusan SMA 1: Diterima di ITS satu tahun setelah lulus SMA 2: Diterima di ITS dua tahun setelah lulus SMA
Jalur Masuk	0: Bidikmisi / KIPK, 1: Reguler, 2: Mandiri, 3: Prestasi
Asal Sekolah	0: SMA Negeri 1: SMA Swasta 2: SMK Negeri 3: SMK Swasta 4: Madrasah Aliyah / Pondok Pesantren
Asal Daerah	0: Kota Surabaya 1: Kawasan Penyangga Surabaya / Gerbangkertosusilo (Gresik, Bangkalan, Mojokerto, Sidoarjo, Lamongan) 2: Provinsi Jawa Timur, di luar kawasan Gerbangkertosusilo 3: Pulau Jawa / Bali, di luar Provinsi Jawa Timur 4: Lainnya (Luar Jawa)
Program Studi	0: Teknologi Rekayasa Pengelolaan dan Pemeliharaan Bangunan Sipil (TRPBS) 1: Teknologi Rekayasa Konstruksi Bangunan Air (TRKBA) 2: Teknologi Rekayasa Manufaktur (TRM) 3: Teknologi Rekayasa Konversi Energi (TRKE) 4: Teknologi Rekayasa Otomasi (TRO) 5: Teknologi Rekayasa Kimia Industri (TRKI) 6: Teknologi Rekayasa Instrumentasi (TRI) 7: Statistika Bisnis (SB)
Usia Masuk	Bulan
Skor Logika	200 – 800
Skor Verbal	200 – 800
Skor Kuantitatif	200 – 800
Skor Spasial	200 – 800
Golongan UKT	0: Penerima Bidikmisi / KIPK 1: Golongan 1 - Rp500.000 2: Golongan 2 - Rp1.000.000 3: Golongan 3 - Rp2.500.000 4: Golongan 4 - Rp4.000.000 5: Golongan 5 - Rp5.000.000 6: Golongan 6 - Rp6.000.000 7: Golongan 7 - Rp7.500.000 8: Golongan 8 - Rp10.000.000 9: Golongan 9 - Rp12.500.000

3.2. Langkah Analisis

Penelitian ini diawali dengan eksplorasi variabel yang diperkirakan memengaruhi risiko akademis mahasiswa. Langkah berikutnya adalah melakukan *pre-processing data*, yang dimulai dengan integrasi dan pemrosesan beberapa dataset dari berbagai sumber untuk menyusun dataset yang utuh. Proses ini dilanjutkan dengan mengeliminasi

mahasiswa yang telah lulus, lanjut jenjang, mengundurkan diri, dan kelompok yang tidak relevan, sehingga hanya mahasiswa aktif program sarjana terapan angkatan 2017 hingga 2022 yang dianalisis. Risiko akademis mahasiswa kemudian diidentifikasi berdasarkan IPK, status per semester, dan lama studi. Data yang memiliki *missing value* ditangani menggunakan *imputasi,* yakni mengganti nilai data yang hilang menggunakan *median* untuk data numerik (skor logika, verbal, kuantitatif, spasial), mengganti nilai yang hilang dengan kategori terbanyak (modus) untuk data kategori, serta eliminasi baris yang memiliki nilai hilang di lebih dari empat kolom. Setelah itu, data kategori diubah menjadi format numerik dan dilakukan penanganan ketidakseimbangan data menggunakan teknik SMOTE. Dataset yang sudah bersih ini kemudian dibagi menjadi *training set* dan *testing* set melalui k-fold *cross validation* untuk memastikan model yang lebih *robust*. Pemodelan dilakukan menggunakan algoritma *Random Forest*, dengan evaluasi kinerja model berdasarkan nilai akurasi, *sensitivity*, dan *specificity* karena metrik tersebut dapat memberikan gambaran menyeluruh baik pada kelas positif maupun kelas negatif [21]. Interpretasi model dilakukan menggunakan SHAP, yang bertujuan untuk mengidentifikasi kontribusi masing-masing fitur terhadap prediksi model. Terakhir, penelitian ini diakhiri dengan penarikan kesimpulan serta pemberian rekomendasi berdasarkan temuan yang diperoleh.

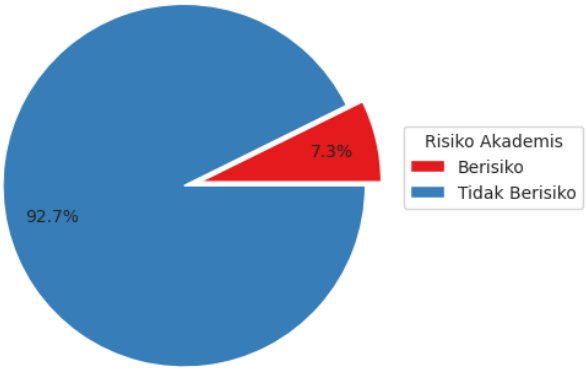


Gambar 2. *Flow chart* analisis

4. HASIL DAN PEMBAHASAN

4.1. Karakteristik Mahasiswa Fakultas Vokasi ITS

Untuk memahami karakteristik mahasiswa Fakultas Vokasi ITS maka dilakukan eksplorasi data terlebih dahulu sebelum masuk ke pemodelan. Eksplorasi data menggunakan statistik deskriptif dan visualisasi data membantu memberikan informasi awal mengenai faktor-faktor yang diduga memengaruhi performa mahasiswa.



Gambar 3. Risiko Akademis Mahasiswa FV ITS

Gambar 3 menunjukkan bahwa pada semester genap tahun akademik 2022/2023, sebanyak 182 dari 2.496 (7,3%) mahasiswa sarjana terapan Fakultas Vokasi ITS berisiko akademis. Bila ditelusuri lebih lanjut, program studi dengan persentase mahasiswa berisiko akademis paling banyak ditemukan adalah Program Studi Teknologi Rekayatasa Otomasi (TRO), yaitu sebesar 10,6% dari 385 mahasiswa. Bila dijabarkan untuk setiap variabel kategorik, persentase mahasiswa berisiko akademis paling besar pada kelompok laki-laki, diterima kuliah melalui jalur bidik

misi/ KIPK, lulusan SMK swasta, dan berasal dari Kota Surabaya. Fakta lainnya adalah mereka yang berisiko akademik secara rata-rata memiliki nilai TPKA semua bidang lebih rendah daripada mereka yang tidak berisiko.

4.2. Pemodelan Risiko Akademis Mahasiswa

Klasifikasi mahasiswa berisiko akademis dilakukan menggunakan algoritma *Random Forest* dengan data yang sudah dibagi menjadi *data training* dan *data testing* melalui *k-Fold Cross Validation*. Fokus utama analisis ini adalah pada evaluasi kinerja model. Pengukuran kinerja model dilakukan menggunakan metrik akurasi, *sensitivity*, dan *specificity*. Pendekatan *k-Fold Cross Validation* digunakan untuk memastikan hasil evaluasi yang komprehensif dan mencakup variasi data secara menyeluruh.

Pemodelan dilakukan menggunakan bahasa pemrograman *python* dengan media *Google Colaboratory*. Dalam membangun model *random forest*, beberapa parameter telah diatur dengan nilai tertentu seperti pada Tabel 2. Parameter pertama, *random_state*, diatur ke nilai 42 yang berfungsi sebagai kunci untuk menjaga hasil agar tetap konsisten pada setiap eksekusi model dalam program *python*. Selanjutnya, *criterion* diatur ke "Gini", menunjukkan penggunaan kriteria *Gini impurity* untuk mengukur kualitas pembagian simpul dalam pohon keputusan. Parameter *max_features* menggunakan log2 dan *max_depth* diatur ke "None" yang berarti pohon keputusan dalam *ensemble* tidak memiliki batasan kedalaman sehingga baru akan berhenti saat Gini bernilai 0 (*pure*). Terakhir, *n_estimators* diatur ke 100 yang menandakan bahwa model *random forest* ini terdiri dari 100 pohon keputusan yang dihasilkan secara independen dan kemudian hasil prediksinya digabungkan untuk memberikan prediksi akhir.

Tabel 2. Parameter Model

Parameter	*Nilai*
random_state	42
criterion	Gini
max_depth	None
max_feature	log2
n_estimators	100

Tabel 3 menunjukkan bahwa hasil terbaik terjadi pada fold ke-2. Berdasarkan nilai akurasi, model mampu mengklasifikan potensi risiko akademis mahasiswa secara tepat sebesar 94,2%. Selain itu, *specificity* yang mencapai 100% menandakan kemampuan model dalam mengidentifikasi mahasiswa tanpa risiko akdemik dengan sempurna. Namun, nilai *sensitivity* yang sangat kecil menunjukkan bahwa model yang terbentuk memiliki keterbatasan dalam melakukan prediksi pada mahasiswa yang berisiko akademik. Hal ini dapat disebabkan karena jumlah mahasiswa berisiko akademis hanya sebesar 7,3% dari keseluruhan mahasiswa Fakultas Vokasi ITS sebagaimana terlihat pada Gambar 3. Oleh karena itu, penanganan ketidakseimbangan data diperlukan untuk meningkatkan kemampuan prediksi pada kelas minoritas.

Tabel 3. Hasil Pemodelan

Fold	*Akurasi*	*Specificity*	*Sensitivity*
Fold 1	0,946	0,998	0,037
Fold 2	**0,942**	**1,000**	**0,147**
Fold 3	0,934	0,996	0,061
Fold 4	0,906	1,000	0,096
Fold 5	0,932	0,994	0,139

Perbandingan hasil penanganan *imbalance* menggunakan SMOTE dapat dilihat pada Tabel 4. Sebelumnya, terdapat 182 mahasiswa berisiko akademi dan 2.341 mahasiswa yang tidak berisiko akademis. Jumlah mahasiswa yang tidak berisiko hampir 13 kali lipat dari jumlah mahasiswa berisiko. Melalui penerapan metode SMOTE, terdapat sebanyak 2.132 data sintesis yang terbentuk sehingga jumlah mahasiswa berisiko menjadi 2.314. Data yang telah seimbang ini kemudian dimodelkan ulang menggunakan *decision tree* dan hasilnya dapat dilihat pada Tabel 5.

Tabel 4. Jumlah Data

Kategori	*Sebelum SMOTE*	*Sesudah SMOTE*
Tidak Berisiko	2314	2314
Berisiko	182	2314

Hasil pemodelan pada data yang telah seimbang menunjukkan performa terbaik berada pada *fold* ke-2 yang memiliki *confusion matrix* pada

Tabel 6. Secara rinci, model *decision tree* pada *fold-2* memiliki tingkat akurasi 96,4%, *specificity* 95%, dan *recall* atau *sensitivity* sebesar 98%. Jika dibandingkan dengan hasil yang ditunjukkan pada Tabel 3, terjadi peningkatan *sensitivity* dari 14,7% menjadi 98%. Implementasi SMOTE berhasil meningkatkan kinerja model dalam mengklasifikasikan mahasiswa berisiko akademis.

Tabel 5. Hasil Pemodelan Setelah SMOTE

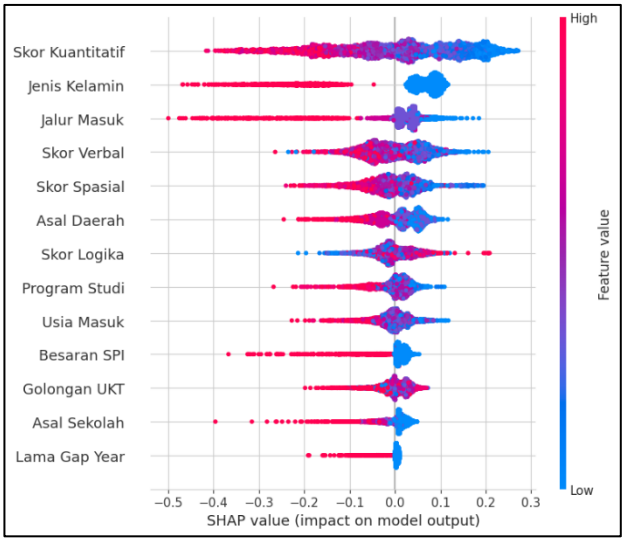
Fold	*Akurasi*	*Specificity*	*Sensitivity*
Fold 1	0,940	0,927	0,953
Fold 2	**0,964**	**0,950**	**0,980**
Fold 3	0,936	0,905	0,968
Fold 4	0,942	0,926	0,956
Fold 5	0,935	0,933	0,937

Tabel 6. *Confusion Matrix*

Prediksi	*Aktual*		*Total*
	Tidak Berisiko	**Berisiko**	
Tidak Berisiko	455	9	464
Berisiko	24	436	462
Total	479	447	926

4.3. Analisis Kontribusi Faktor Terhadap Risiko Akademik

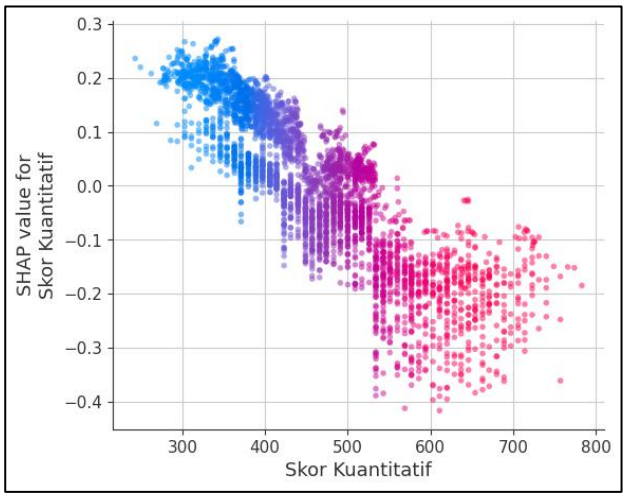
Analisis kontribusi masing-masing fitur dihitung menggunakan SHAP untuk mengetahui bagaimana model bekerja melakukan prediksi. SHAP *summary plot* pada Gambar 4 merupakan visualisasi kontribusi masing-masing fitur terhadap prediksi yang diurutkan berdasarkan rata-rata dari nilai SHAP mutlak. Setiap poin mewakili satu mahasiswa. Posisi poin menunjukkan arah dan pengaruh sementara warna menggambarkan nilai dari tiap variabel. Sebagai contoh, skor kuantitatif tinggi cenderung memiliki nilai SHAP rendah. Artinya mahasiswa dengan skor kuantitatif tinggi cenderung tidak berisiko.



Gambar 4. SHAP *Summary Plot*

Berdasarkan Gambar 4, skor TPKA kuantitatif memiliki kontribusi paling tinggi dalam menentukan hasil prediksi risiko akademik mahasiswa, diikuti oleh jenis kelamin dan jalur masuk ITS. Skor kuantitatif yang semakin rendah cenderung meningkatkan peluang mahasiswa terhadap risiko akademis, hal ini didukung oleh [22], yang menyatakan

bahwa kemampuan spasial, numerikal dan verbal berpengaruh terhadap prestasi akademik mahasiswa program STEM. Selanjutnya Laki-laki lebih mungkin berisiko akademis dibandingkan perempuan, ini sejalan dengan penelitian [4] bahwa perempuan memiliki performa lebih baik dibandingkan laki-laki dalam STEM. Faktor tertinggi selanjutnya adalah jalur masuk, didapatkan bahwa mahasiswa yang masuk melalui jalur prestasi memiliki peluang berisiko akademis lebih rendah dibandingkan mahasiswa yang masuk dari jalur lainnya, hal ini diduga disebabkan oleh performa yang baik semasa jenjang pendidikan sebelumnya. Sebaliknya, fitur dengan kontribusi paling rendah adalah lama *gap year*, asal sekolah, dan golongan UKT. Nilai SHAP untuk variabel skor TPKA kuantitatif dari masing-masing mahasiswa dapat digambarkan lebih detail melalui SHAP *dependency plot* pada Gambar 5.



Gambar 5. SHAP *Dependency Plot*

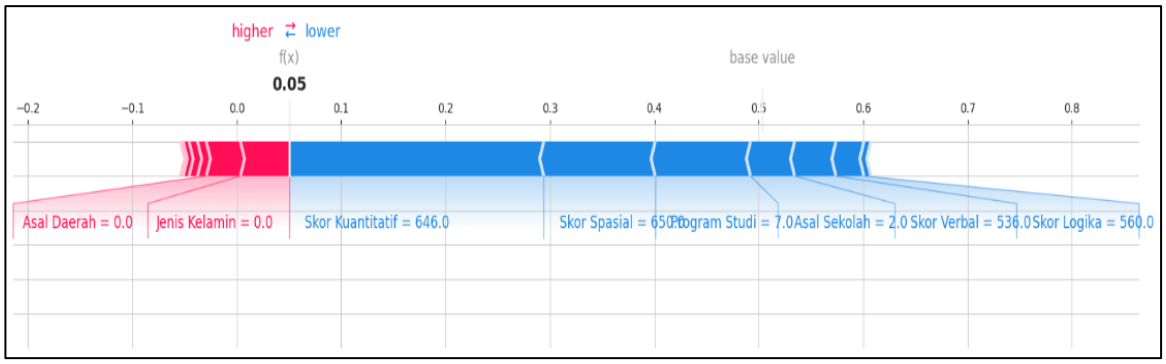
Dependency Plot pada dasarnya merupakan *scatter plot* antara nilai variabel dengan nilai SHAP yang merupakan penjabaran detail dari *summary plot.* Gambar 5 menunjukkan bahwa semakin tinggi skor TPKA kuantitatif, nilai SHAP-nya akan semakin rendah. Hal ini berarti bahwa nilai TPKA kuantitatif yang tinggi cenderung menghasilkan prediksi kelas tidak berisiko. Dalam kata lain, mahasiswa yang memiliki skor TPKA kuantitatif yang tinggi cenderung tidak berisiko akademis. Sebaliknya, mahasiswa dengan skor TPKA kuantitatif yang rendah cenderung berisiko akademis.

Tabel 7. Profil Mahasiswa

Variabel	***Nilai***
Risiko Akademis	Tidak berisiko
Jenis Kelamin	Laki-laki
Lama *Gap Year*	0 tahun
Jalur Masuk	Reguler
Asal Sekolah	SMA Negeri
Asal Daerah	Kota Surabaya
Program Studi	Statistika Bisnis
Usia Masuk	225 Bulan
Skor Logika	560
Skor Verbal	536
Skor Kuantitatif	646
Skor Spasial	650
Golongan UKT	Golongan 2 - Rp1.000.000

Data pada Tabel 7 merupakan profil salah satu mahasiswa Fakultas Vokasi ITS yang digunakan sebagai contoh untuk interpretasi SHAP terhadap hasil prediksi menggunakan model *random forest* pada data individu. Gambar 6 merupakan nilai SHAP hasil pemodelan pada mahasiswa Fakultas Vokasi ITS dengan profil pada Tabel 7. Asal daerah dan jenis kelamin (panah merah) merupakan fitur dominan yang memperbesar peluang mahasiswa untuk dikelompokkan menjadi berisiko akademis. Sedangkan skor kuantitatif, skor spasial, program studi, asal sekolah,

skor verbal dan skor logika (panah biru) berkontribusi mengurangi peluang mahasiswa diprediksi berisiko akademis. Nilai SHAP sebesar 0,05 menunjukkan bahwa peluang mahasiswa tersebut berisiko akademis adalah sebesar 0,05. Oleh karena peluang mahasiswa tidak berisiko akademis adalah 0,95 maka model yang terbentuk akan mengelompokkan mahasiswa tersebut ke dalam kategori tidak berisiko akademis. Hasil prediksi ini sesuai dengan kondisi asli dari mahasiswa tersebut yang memang tidak berisiko akademis. Artinya model klasifakasi *random forest* yang terbentuk mengelompokkan mahasiswa tersebut secara tepat.



Gambar 6. SHAP *Force Plot*

5. KESIMPULAN

Model *Random Forest* terbaik yang menggunakan data seimbang setelah proses SMOTE memiliki akurasi prediksi 96,4% dengan *specificity* 95% dan *sensitivity* 98%. Kombinasi algoritma *Random Forest* dan SHAP tidak hanya menghasilkan prediksi yang lebih akurat, tetapi juga memberikan wawasan yang mendalam mengenai faktor-faktor utama yang memengaruhi risiko akademik. Hal ini penting karena model prediksi sering kali dianggap sebagai *black box*, sehingga sulit bagi para pemangku kepentingan di bidang pendidikan untuk memahami dasar keputusan yang diambil. Melalui interpretasi yang jelas, penelitian ini membantu mengidentifikasi fitur-fitur kunci yang berkontribusi signifikan terhadap risiko akademik. Tiga faktor utama yang mempengaruhi potensi risiko akademis mahasiswa adalah skor kuantitatif, jenis kelamin, dan jalur masuk, sementara lama *gap year*, asal sekolah, dan golongan UKT memiliki kontribusi paling rendah. Dengan demikian, hasil penelitian ini memberikan dasar yang kuat bagi dosen dan administrator untuk melakukan intervensi yang lebih tepat sasaran, merancang kebijakan dukungan yang lebih efektif, dan meningkatkan upaya pencegahan risiko akademik di kalangan mahasiswa.

Terdapat beberapa rekomendasi untuk studi lanjutan. Pertama, penanganan ketidakseimbangan data secara lebih komprehensif melalui teknik seperti *oversampling* atau metode kombinasi *oversampling-undersampling*, guna meningkatkan prediksi pada kelas minoritas yang berisiko akademis namun juga menghindari *overfit*. Selain itu, eksplorasi lebih lanjut terhadap fitur-fitur bar yang berpotensi mempengaruhi kinerja akademis, seperti keterlibatan mahasiswa dalam kegiatan ekstrakurikuler atau aspek psikologis, dapat memperkaya hasil analisis. Penelitian selanjutnya juga sebaiknya mempertimbangkan metode klasifikasi alternatif, seperti *gradient boosting*, untuk mengevaluasi performa prediksi yang lebih baik dibandingkan *random forest*. Terakhir, diperlukan pengembangan strategi yang dapat meningkatkan motivasi dan keterlibatan mahasiswa pada kelompok yang rentan terhadap risiko akademis, sehingga memungkinkan pendekatan yang lebih proaktif dalam mengurangi risiko mengulang mata kuliah, terlambat studi, ataupun memiliki IPK yang cenderung rendah.

DAFTAR PUSTAKA

[1] R. H. Tambunan, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Kinerja Akademis Mahasiswa Menggunakan Algoritma Naïve Bayes dengan Implementasi Data Mining Studi Kasus : Departemen Teknik Industri USU," Universitas Sumatera Utara, 2020.

[2] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates," *Comput Educ*, vol. 53, no. 3, pp. 563–574, 2009, doi: 10.1016/j.compedu.2009.03.013.

[3] S.-Y. Hwang, D.-J. Shin, J.-K. Oh, Y.-S. Lee, and J.-J. Kim, "A Regression Analysis of Factors Affecting Dropout of College Students," *The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)*, vol. 20, no. 4, pp. 187–193, 2020, doi: 10.7236/JIIBC.2020.20.4.187.

[4] B. Bloodhart, M. M. Balgopal, A. M. A. Casper, L. B. Sample McMeeking, and E. V. Fischer, "Outperforming yet undervalued: Undergraduate women in STEM," *PLoS One*, vol. 15, no. 6, pp. 1–13, 2020, doi: 10.1371/journal.pone.0234685.

[5] M. Tan and P. Shao, "Prediction of student dropout in E-learning program through the use of machine learning method," *International Journal of Emerging Technologies in Learning*, vol. 10, no. 1, pp. 11–17, 2015, doi: 10.3991/ijet.v10i1.4189.

[6] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Comput Surv*, vol. 52, no. 4, 2019, doi: 10.1145/3343440.

[7] J. Brandt and E. Lanzén, "A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification," Uppsala University, 2020.

[8] G. L. Pritalia, "Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 2, no. 1, pp. 43–55, 2022, doi: 10.24002/konstelasi.v2i1.5630.

[9] V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowl Based Syst*, vol. 25, no. 1, pp. 13–21, Feb. 2012, doi: 10.1016/j.knosys.2011.06.013.

[10] Prajwala, "A Comparative Study on Decision Tree and Random Forest Using R Tool," *Ijarcce*, no. January 2015, pp. 196–199, 2015, doi: 10.17148/ijarcce.2015.4142.

[11] S. Yang, "Who will dropout from university? Academic risk prediction based on interpretable machine learning," vol. 1, pp. 0–2, 2021, doi: 10.3969/j.issn.1673-4807.2012.01.018.

[12] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions Scott," *Nips*, vol. 16, no. 3, pp. 426–430, 2012.

[13] G. L. Pritalia, "Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 2, no. 1, pp. 43–55, 2022, doi: 10.24002/konstelasi.v2i1.5630.

[14] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Comput Surv*, vol. 52, no. 4, 2019, doi: 10.1145/3343440.

[15] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[16] J. Brandt and E. Lanzén, "A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification," Uppsala University, 2020.

[17] D. L. Pratiwi, "Penerapan Metode Combine Sampling Pada Klasifikasi Imbalanced Data Biner," 2018.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Deep Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research 16*, vol. 16, pp. 321–357, 2002.

[19] Prajwala, "A Comparative Study on Decision Tree and Random Forest Using R Tool," *Ijarcce*, no. January 2015, pp. 196–199, 2015, doi: 10.17148/ijarcce.2015.4142.

[20] L. Breiman, "Random Forests," 2001. [Online]. Available: https://link.springer.com/content/pdf/10.1023/a:1010933404324.pdf

[21] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations.," *Northeast SAS Users Group 2010: Health Care and Life Sciences*, pp. 1–9, 2010.

[22] M. Berkowitz and E. Stern, "Which Cognitive Abilities Make the Difference? Predicting Academic Achievements in Advanced STEM Studies," *J Intell*, vol. 6, no. 4, p. 48, Oct. 2018, doi: 10.3390/jintelligence6040048.