Sebatik Vol. 30 No. 1 June 2026
ISSN: 1410-3737(p) 2621-069X(e)

DOI:10.46984/sebatik.v29i1.2747

Model Interpretation for Student Major Selection Using Principal Component Analysis and Random Forest

Antoni[1)], Sarjon Defit[2)], dan **Yuhandri[3)]**

[1,2,3]Magister Teknik Informatika, Universitas Putra Indonesia YPTK Padang
[1,2,3]Jl. Raya Lubuk Begalung, Lubuk Begalung Nan XX, Kecamatan Lubuk Begalung Kota Padang Sumbar
E-mail: antonidafri@gmail.com[1)], sarjon_defit@upiyptk.ac.id[2)], yuhandri.yunus@gmail.com[3)]

ABSTRACT

The development of information technology has had a significant impact on the education sector by providing data-driven tools to support the process of major selection. This process often causes confusion among students due to its crucial role in determining their academic and career futures. This study aims to develop an accurate and transparent recommendation system for major selection through the integration of Principal Component Analysis (PCA), Random Forest (RF), and SHAP. The research follows a systematic framework that includes data processing and model evaluation stages. PCA is applied to reduce the dimensionality of complex student data in order to improve computational efficiency and minimize information redundancy. Furthermore, the Random Forest algorithm is employed as a classification model to predict major recommendations such as Science, Social Sciences, and Religious Studies. The SHAP method is integrated to provide both mathematical and visual interpretations of the contribution of each academic feature to the model's prediction results. The research data are obtained from the internal records of MAN 1 Payakumbuh covering the last three academic years (2022/2023–2024/2025). The dataset consists of 571 eleventh-grade students with tenth-grade academic scores and non-academic skill variables. The implementation of this model is able to provide more objective recommendations compared to conventional subjective assessments, achieving an accuracy of 88.70%. Visualization of feature contributions using SHAP enhances transparency and facilitates stakeholders' understanding of the basis for each model decision. This study contributes to improving the efficiency of the major selection process and supports more accurate academic decision-making for students and educators.

***Keywords**: Major Selection, PCA, Random Forest, Recommendation, SHAP.*

Interpretasi Model Pemilihan Jurusan Siswa Menggunakan *Principal Component Analisyt* dan *Random Forest*

ABSTRAK

Pengembangan teknologi informasi berdampak signifikan pada sektor pendidikan melalui penyediaan alat berbasis data untuk mendukung proses pemilihan jurusan. Proses ini seringkali menimbulkan kebingungan bagi siswa karena perannya yang sangat krusial dalam menentukan masa depan akademik dan karir mereka. Penelitian ini bertujuan mengembangkan sistem rekomendasi pemilihan jurusan yang akurat dan transparan melalui integrasi metode Principal Component Analysis (PCA), Random Forest (RF), dan SHAP. Penelitian mengikuti kerangka kerja yang mencakup tahapan pemrosesan data hingga evaluasi model secara sistematis. PCA diterapkan untuk mereduksi dimensi data siswa yang kompleks guna meningkatkan efisiensi komputasi dan meminimalkan redundansi informasi. Selanjutnya, algoritma Random Forest digunakan sebagai model klasifikasi untuk memprediksi rekomendasi jurusan seperti IPA, IPS, dan Keagamaan. Metode SHAP diintegrasikan untuk memberikan interpretasi matematis dan visual mengenai kontribusi setiap fitur akademik terhadap hasil prediksi model. Data penelitian bersumber dari data internal MAN 1 Payakumbuh yang mencakup periode tiga tahun pelajaran terakhir (2022/2023–2024/2025). Dataset tersebut terdiri dari 571 siswa kelas XI dengan data nilai akademik kelas X serta variabel keterampilan non-akademik. Implementasi model ini mampu memberikan rekomendasi yang lebih objektif dibandingkan penilaian subjektif konvensional dengan akurasi 88,70%. Visualisasi kontribusi fitur melalui SHAP memfasilitasi transparansi sehingga memudahkan stakeholder dalam memahami dasar setiap keputusan model. Penelitian ini berkontribusi dalam meningkatkan efisiensi proses penjurusan serta mendukung akurasi pengambilan keputusan akademik bagi siswa dan pendidik.

Kata Kunci: PCA, Pemilihan Jurusan, Random Forest, Rekomendasi, SHAP.

1. PENDAHULUAN

Data mining merupakan proses ekstraksi pengetahuan yang berguna, pola tersembunyi, dan informasi yang tidak diketahui sebelumnya dari kumpulan data berukuran besar (Fajri dkk., 2024). Proses ini melibatkan penggunaan metode otomatis atau semi-otomatis yang mengintegrasikan teknik dari berbagai disiplin ilmu, termasuk statistik, machine learning, sistem basis data, dan visualisasi data (Supriyono dkk., 2025). Data mining juga dikenal dengan istilah *Knowledge Discovery in Databases* (KDD), yang menekankan pada aspek penemuan pengetahuan dari data yang sebelumnya tidak terlihat atau tidak terduga (Rustiyana dkk., 2025).

Data mining dalam pendidikan memiliki beberapa tujuan utama yang berkontribusi terhadap peningkatan kualitas sistem pendidikan. Pertama, prediksi (*prediction*) bertujuan untuk memperkirakan nilai atau kategori target di masa depan berdasarkan data historis, seperti memprediksi kelulusan siswa atau prestasi akademik di semester berikutnya (Maulana dkk., 2025). Penelitian oleh Alboaneen dkk., (2023) menunjukkan bahwa model prediksi berbasis Random Forest dapat mencapai akurasi 92% dalam memprediksi nilai akhir siswa pada data universitas. Kedua, klasifikasi (*classification*) bertujuan untuk mengategorikan data ke dalam kelas-kelas yang telah ditentukan sebelumnya, seperti mengklasifikasikan siswa ke dalam jurusan IPA, IPS, Bahasa, atau Keagamaan berdasarkan profil akademik mereka (Astuti, 2024). Ketiga, *clustering* (pengelompokan) bertujuan untuk mengelompokkan data yang memiliki kesamaan karakteristik tanpa label kelas yang telah ditentukan sebelumnya, misalnya mengelompokkan siswa berdasarkan pola belajar atau gaya pembelajaran (Handayani, 2022). Keempat, *association rule mining* bertujuan untuk menemukan hubungan atau asosiasi antar variabel dalam data, seperti mengidentifikasi mata pelajaran mana yang sering dikuasai secara bersamaan oleh siswa berprestasi tinggi (AS dkk., 2021). Kelima, *anomaly detection* (deteksi anomali) bertujuan untuk mengidentifikasi data yang tidak biasa atau *outliers*, yang dapat mengindikasikan siswa dengan kebutuhan khusus atau masalah pembelajaran (Laksono, 2025).

Aplikasi konkret data mining dalam pendidikan mencakup sistem rekomendasi jurusan atau mata kuliah yang dipersonalisasi (Gustirani, 2024), deteksi dini siswa berisiko dropout untuk intervensi preventif (Al Ghifari, 2021), evaluasi efektivitas metode pengajaran melalui analisis data performa kelas (Aprilya dkk., 2023), optimasi alokasi sumber daya pendidikan berdasarkan kebutuhan siswa (Mulyanti, 2025), serta pengembangan learning path yang adaptif sesuai dengan kemampuan individual siswa (Ananda & Malik, 2025). Data mining dalam pendidikan menawarkan potensi besar, akan tetapi implementasi data mining dalam pendidikan menghadapi beberapa tantangan signifikan. Pertama, kualitas data yang tidak konsisten, termasuk *missing values*, outliers, dan noise, yang dapat mempengaruhi validitas hasil analisis (Rahayu dkk., 2024). Kedua, privasi dan etika terkait penggunaan data siswa yang bersifat sensitif, memerlukan mekanisme anonimisasi dan kepatuhan terhadap regulasi perlindungan data (Fadhilah, 2021). Ketiga, interpretabilitas model, terutama untuk algoritma kompleks seperti *neural networks* atau *ensemble methods*, yang sering dipandang sebagai "*black box*" oleh *stakeholder* pendidikan (Muis dkk., 2025). Keempat, keterbatasan kompetensi teknis dari praktisi pendidikan dalam mengoperasikan *tools* data *mining* dan menginterpretasikan hasil analisis (Rustiyana dkk., 2025). Kelima, bias dalam data historis yang dapat menghasilkan model diskriminatif terhadap kelompok tertentu, misalnya bias gender atau latar belakang sosio-ekonomi (Alifariki, 2022).

Pada konteks pemilihan jurusan di madrasah aliyah, klasifikasi digunakan untuk memprediksi jurusan yang paling sesuai untuk setiap siswa berdasarkan profil akademik dan non-akademik mereka. *Input features* mencakup nilai mata pelajaran numerik (Matematika, IPA, IPS, Bahasa, Agama), hasil TPA, dan portofolio ekstrakurikuler yang *diencode*. Target variable adalah label jurusan kategorikal: IPA, IPS, Bahasa, atau Keagamaan (Sulika, 2024). Tantangan khusus dalam klasifikasi pemilihan jurusan mencakup:

1. *Class imbalance*, dimana distribusi siswa antar jurusan tidak seimbang (misalnya jurusan IPA lebih banyak diminati), memerlukan teknik seperti SMOTE (*Synthetic Minority Over-sampling Technique*) atau *class weighting*.
2. *Multi-dimensional features* dengan potential *multicollinearity* antar mata pelajaran serumpun, memerlukan *feature extraction* seperti PCA.
3. *Interpretability requirements*, karena stakeholder pendidikan perlu memahami alasan di balik rekomendasi, memerlukan *explainable* AI seperti SHAP, serta
4. *Ethical considerations* terkait *fairness* dan bias, memastikan model tidak diskriminatif terhadap gender atau latar belakang sosio-ekonomi (Alifariki, 2022).

Penelitian oleh Xiao dkk., (2022) dalam survey 80 studi EDM menunjukkan bahwa Random Forest unggul dalam akurasi prediksi pemilihan jurusan dengan rata-rata 89%, direkomendasikan sebagai metode terbaik untuk klasifikasi multi-class dalam konteks pendidikan. Studi oleh Albar, (2025) mengonfirmasi bahwa kombinasi PCA dengan Random Forest meningkatkan akurasi hingga 92% dalam prediksi performa dan rekomendasi jurusan siswa.

Explainable AI (XAI) adalah paradigma dalam *artificial intelligence* yang menekankan *transparency, interpretability,* dan *accountability* dari model *machine learning* (Dita dkk., 2025). Pada konteks pendidikan, dimana keputusan AI mempengaruhi masa depan siswa, *explainability* menjadi *critical requirement* untuk membangun *trust* dan memfasilitasi *adoption* oleh *stakeholders* seperti guru, konselor, siswa, dan orang tua (Nirmala, 2025). SHAP berbasis pada *Shapley value*, konsep dari *cooperative game theory* yang dikembangkan

oleh Lloyd Shapley untuk *fairly distribute payoff* antar *players* dalam *coalition* (Victoria dkk., 2024).

Integrasi PCA dengan *Random Forest* (PCA-*Random Forest*) menghasilkan model yang efisien untuk data berdimensi tinggi, sementara SHAP menambahkan lapisan interpretabilitas. Pendekatan ini telah diterapkan dalam berbagai studi pendidikan, seperti yang dilakukan oleh Jufri dkk., (2023), di mana model mencapai akurasi tinggi sambil menyediakan penjelasan yang dapat diverifikasi. Ahamd, (2025) menunjukkan bahwa kombinasi ini mengurangi skeptisisme terhadap AI dengan memberikan insight tentang faktor pengaruh, seperti minat ekstrakurikuler terhadap pilihan jurusan.

Penelitian ini mengusulkan integrasi PCA, Random Forest, dan SHAP untuk menghasilkan model klasifikasi performa siswa yang akurat dan dapat dijelaskan. Kebaruan penelitian terletak pada penerapan metode SHAP untuk menginterpretasikan hasil klasifikasi berbasis PCA dan Random Forest pada data rapor siswa MAN.

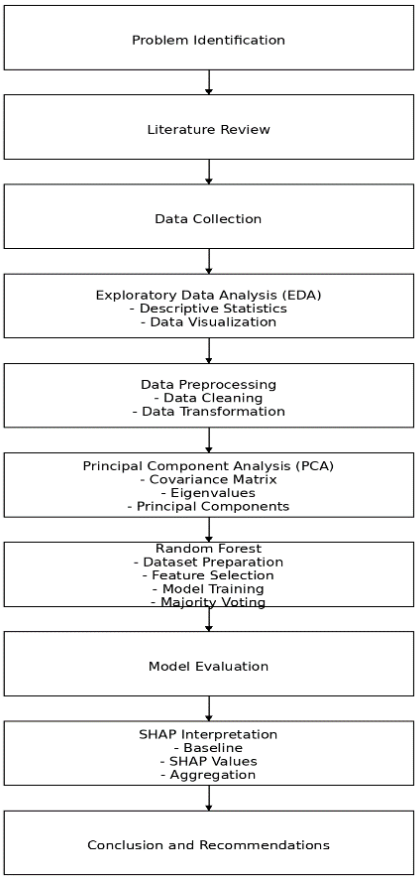
2. RUANG LINGKUP

Ruang lingkup penelitian ini mencakup pemanfaatan data rapor siswa sebagai sumber utama dalam proses klasifikasi performa siswa. Data yang digunakan terdiri atas variabel-variabel akademik yang merepresentasikan capaian belajar siswa pada beberapa mata pelajaran. Penelitian ini difokuskan pada pengolahan data numerik melalui tahapan pra-pemrosesan, standarisasi, serta reduksi dimensi menggunakan *Principal Component Analysis* (PCA) untuk memperoleh representasi fitur yang lebih ringkas dan bebas dari multikolinearitas.

Batasan penelitian ini terletak pada penggunaan variabel akademik tanpa melibatkan faktor non-akademik seperti latar belakang sosial, kondisi ekonomi, maupun aspek psikologis siswa. Model klasifikasi dibangun menggunakan algoritma *Random Forest* dan diinterpretasikan menggunakan metode *SHapley Additive exPlanations* (SHAP) untuk mengidentifikasi komponen utama yang paling berpengaruh terhadap hasil prediksi. Hasil yang diharapkan dari penelitian ini adalah tersusunnya model klasifikasi yang memiliki kinerja prediktif yang baik serta mampu memberikan penjelasan yang transparan mengenai faktor-faktor dominan yang memengaruhi performa siswa.

3. BAHAN DAN METODE

Penelitian ini menggunakan pendekatan eksperimen dengan tahapan pra-pemrosesan data, reduksi dimensi, klasifikasi, dan interpretasi model. PCA digunakan untuk mereduksi 15 fitur numerik awal menjadi 8 komponen utama yang lebih ringkas. Random Forest digunakan sebagai algoritma klasifikasi karena kemampuannya dalam menangani data berdimensi tinggi dan ketahanannya terhadap *overfitting*. Interpretabilitas model dicapai melalui penerapan SHAP yang mengadopsi konsep nilai Shapley dalam teori permainan.



Gambar 1. Kerangka Kerja Penelitian
Figure 1. Research Framework

Gambar 1 menunjukkan kerangka kerja penelitian yang diterapkan dalam pelaksanaan penelitian ini.

3.1 Data

Populasi penelitian adalah siswa kelas XI MAN 1 Kota Payakumbuh (menggunakan data nilai akademik kelas X) dalam rentang 3 tahun terakhir (tahun pelajaran 2022/2023 hingga 2024/2025). Jumlah populasi adalah 571 siswa, berdasarkan catatan internal madrasah. Rincian jumlah data dari populasi penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Rincian Jumlah Sampel Penelitian
Table 1. Details of the Number of Research Samples

No.	Academic Year	Number of Samples
1	2022/2023	185 Students
2	2023/2024	195 Students
3	2024/2025	191 Students

3.2 *Principal Component Analisyt* (PCA)

Implementasi *Principal Component Analysis* (PCA) untuk *feature extraction* dan *dimensionality reduction*. PCA akan diterapkan pada data yang telah dinormalisasi untuk mengidentifikasi komponen utama yang menjelaskan varians maksimum. Jumlah komponen optimal ditentukan berdasarkan:

1. *scree plot (elbow method),*
2. *explained variance ratio* dengan *threshold* kumulatif 85-95%,
3. *Kaiser criterion* (*eigenvalue* >1) (Li et al., 2024).

3.3 *Random Forest* (PCA)

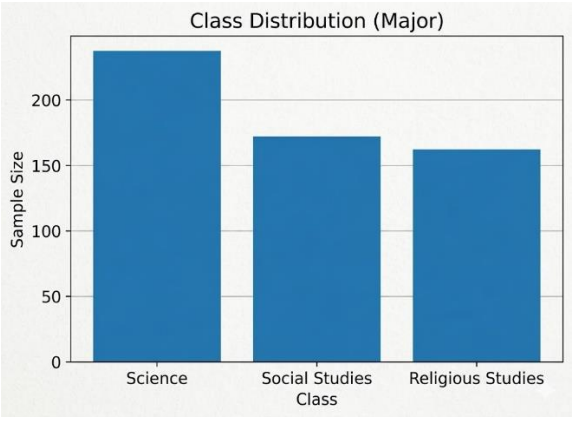
Hasil dari PCA kemudian diolah menggunakan Algoritma Random Forest. Proses penerapan Algoritma RF dilakukan dengan langkah-langkah: (a) inisialisasi model dengan *hyperparameters* default, (b) *hyperparameter tuning* menggunakan *Grid Search* atau *Random Search* untuk mengoptimalkan parameters seperti *n_estimators* (jumlah *trees*: 100-500), *max_depth* (kedalaman maksimum tree: 10-50 atau None), *min_samples_split* (minimum samples untuk split node: 2-20), *min_samples_leaf* (minimum samples di *leaf node*: 1-10), *max_features* (jumlah fitur untuk *best split*: *'sqrt'*, *'log2'*, atau *None*), dan *bootstrap* (*True/False*) (Rahman et al., 2025).

3.4 SHAP

Explainable AI (XAI) adalah paradigma dalam *artificial intelligence* yang menekankan *transparency, interpretability,* dan *accountability* dari model *machine learning* (Victoria dkk., 2024). Pada konteks pendidikan, dimana keputusan AI mempengaruhi masa depan siswa, *explainability* menjadi *critical requirement* untuk membangun *trust* dan memfasilitasi *adoption* oleh *stakeholders* seperti guru, konselor, siswa, dan orang tua (Wang et al., 2024). SHAP berbasis pada *Shapley value*, konsep dari *cooperative game theory* yang dikembangkan oleh Lloyd Shapley untuk *fairly distribute payoff* antar *players* dalam *coalition* (Victoria dkk., 2024).

4. PEMBAHASAN

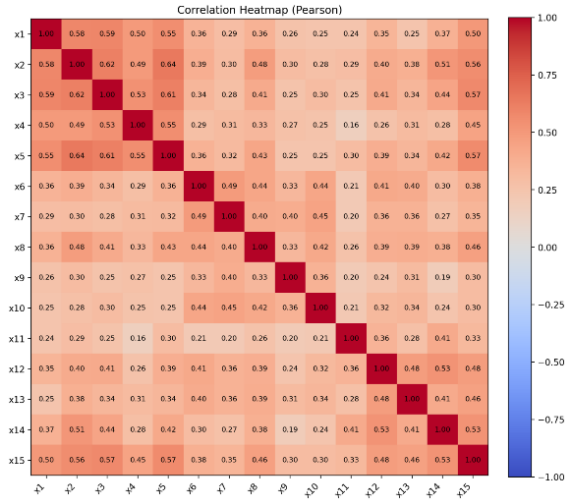
Pembahasan dimulai dari analisis distribusi kelas target, diikuti eksplorasi korelasi antar fitur, konfigurasi *hyperparameter*, evaluasi kinerja model, serta interpretasi model menggunakan SHAP. Langkah pertama sebelum proses pemodelan adalah memahami komposisi data berdasarkan label jurusan. Gambar 2 menampilkan distribusi jumlah sampel siswa pada masing-masing kelas jurusan yang menjadi target klasifikasi dalam penelitian ini.



Gambar 2. Distribusi Kelas
Figure 2. Class Distribution

Gambar 2 menunjukkan distribusi jumlah sampel siswa berdasarkan jurusan, yaitu IPA, IPS, dan Keagamaan. Terlihat bahwa jurusan IPA memiliki jumlah sampel terbanyak dibandingkan dengan jurusan lainnya, diikuti oleh jurusan IPS, sedangkan jurusan Keagamaan memiliki jumlah sampel paling sedikit.

Gambar 3 menampilkan matriks korelasi antar 23 variabel (v1–v23) dalam bentuk *heatmap*. Warna yang semakin terang menunjukkan nilai korelasi yang semakin tinggi, sedangkan warna yang lebih gelap menunjukkan korelasi yang rendah atau mendekati nol. Terlihat bahwa sebagian variabel memiliki korelasi positif sedang hingga kuat, khususnya pada kelompok variabel awal (v1 hingga sekitar v15), yang mengindikasikan adanya keterkaitan linier antar fitur dalam kelompok tersebut. Pola warna yang relatif seragam di sepanjang diagonal utama menunjukkan korelasi sempurna setiap variabel terhadap dirinya sendiri.



Gambar 3. Heatmap Korelasi
Figure 3. Correlation Heatmap

Gambar 3 juga menunukkan hubungan linear antar nilai mata pelajaran siswa secara komprehensif. Secara umum, sebagian besar variabel memiliki korelasi positif

dengan tingkat kekuatan rendah hingga sedang, yang ditunjukkan oleh nilai koefisien korelasi berkisar antara 0,20 hingga 0,65. Kondisi ini mengindikasikan adanya multikolinearitas sedang pada sebagian fitur, yang berpotensi memengaruhi kinerja model klasifikasi apabila digunakan secara langsung. Oleh karena itu, analisis korelasi ini menjadi dasar dalam penerapan *Principal Component Analysis* (PCA) untuk mereduksi dimensi data dan menghilangkan redundansi informasi, sehingga diperoleh representasi fitur yang lebih ringkas dan saling bebas korelasi pada tahap pemodelan selanjutnya.

Tabel 2. Konfigurasi Hyperparameter Tuning
Table 2. Hyperparameter Tuning Configuration

No.	Parameter Tuning	Seting Parameter
1	rf_n_estimators	300
2	rf_min_samples_split	5
3	rf_min_samples_leaf	1
4	rf_max_features	sqrt
5	rf_max_depth	10
6	pca_n_components	8

Tabel 2 menunjukkan hasil proses pencarian parameter terbaik (*hyperparameter tuning*) menggunakan metode validasi silang yang menghasilkan bahwa konfigurasi optimal model diperoleh dengan jumlah pohon (*n_estimators*) sebanyak 300, kedalaman maksimum pohon (*max_depth*) sebesar 10, jumlah minimum sampel untuk pemisahan node (*min_samples_split*) sebesar 5, dan jumlah minimum sampel pada daun (*min_samples_leaf*) sebesar 1. Selain itu, jumlah fitur yang dipertimbangkan pada setiap pemisahan ditetapkan menggunakan pendekatan akar kuadrat (*max_features = sqrt*), serta jumlah komponen utama PCA yang digunakan sebanyak 8. Konfigurasi ini menghasilkan nilai akurasi rata-rata validasi silang (*cross-validation accuracy*) sebesar 0.8949, yang menunjukkan bahwa model memiliki kinerja yang konsisten selama proses pelatihan.

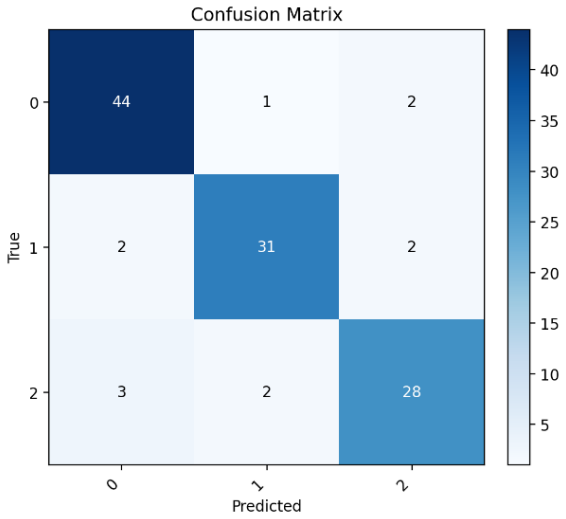
Model dengan parameter terbaik tersebut diuji menggunakan data uji (*test set*) yang tidak terlibat dalam proses pelatihan. Hasil pengujian menunjukkan bahwa model mencapai nilai akurasi sebesar 0.8957. Nilai ini relatif sebanding dengan hasil validasi silang, sehingga mengindikasikan bahwa model tidak mengalami *overfitting* secara signifikan dan mampu melakukan generalisasi dengan baik terhadap data baru. Kondisi ini menunjukkan kombinasi PCA dengan 8 komponen utama dan Random Forest dengan parameter optimal tersebut dapat memberikan kinerja klasifikasi yang stabil dan andal pada data rapor siswa.

Tangkap layar *Classification report* pada Gambar 4 menunjukkan bahwa model memiliki performa yang baik pada ketiga kelas. Pada kelas IPA, model mencapai nilai precision sebesar 0.8984, recall sebesar 0.9362, dan F1-score sebesar 0.91670, yang mengindikasikan bahwa hampir seluruh data IPA berhasil diklasifikasikan dengan benar meskipun masih terdapat sebagian kecil prediksi keliru dari kelas lain.

	precision	recall	f1-score	support
0	0.898	0.9362	0.9167	47
1	0.9118	0.8857	0.8986	35
2	0.875	0.8485	0.8615	33
accuracy	0.8957	0.8957	0.8957	0.8957
macro avg	0.8949	0.8901	0.8923	115
weighted avg	0.8956	0.8957	0.8953	115

Gambar 4. Laporan Klasifikasi
Figure 4. Classification Report

Classification report pada Gambar 4 juga menunjukkan kelas IPS memiliki kinerja paling stabil dengan nilai *precision* sebesar 0.9118, *recall* sebesar 0.8857, dan F1-*score* sebesar 0.8986, yang menandakan bahwa model sangat akurat dalam mengenali kelas ini dengan tingkat kesalahan yang relatif rendah. Sementara itu, pada kelas Keagamaan diperoleh nilai *precision* sebesar 0.875, *recall* sebesar 0.8485, dan F1-*score* sebesar 0.8615, yang menunjukkan bahwa model masih mengalami kesulitan dalam membedakan kelas ini dibandingkan dua kelas lainnya. Secara keseluruhan, model mencapai nilai akurasi sebesar 0.8957 dengan *macro average* F1-*score* sebesar 0.8923 dan *weighted average* F1-*score* sebesar 0.8953, yang menunjukkan bahwa model memiliki kinerja klasifikasi yang baik dan relatif seimbang pada seluruh kelas.



Gambar 5. Matriks Kebingungan
Figure 5. Confusion Matrix

Confusion matrix pada gambar 5 menunjukkan kinerja model klasifikasi dalam memprediksi tiga kelas, yaitu

IPA, IPS, dan Keagamaan pada data uji. Untuk kelas IPA, model mampu mengklasifikasikan dengan benar sebanyak 47 sampel, sementara 1 sampel salah diklasifikasikan sebagai IPS dan 2 sampel salah klasifikasi sebagai Agama. Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam mengenali pola pada kelas IPA.

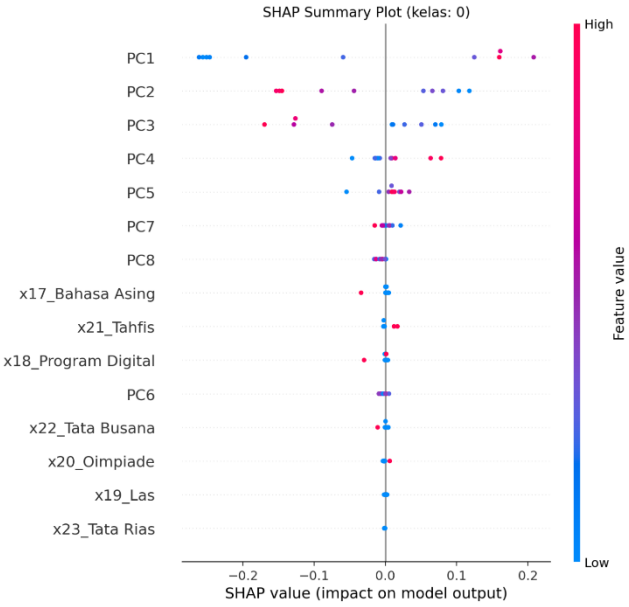
Pada kelas IPS, sebanyak 31 sampel berhasil diprediksi dengan benar, sedangkan 4 sampel salah diklasifikasikan sebagai Agama dan IPA. Sementara itu, pada kelas Agama, model berhasil mengklasifikasikan dengan benar sebanyak 28 sampel, namun masih terdapat kesalahan prediksi, yaitu 3 sampel diklasifikasikan sebagai IPA dan 2 sampel sebagai IPS. Pola kesalahan ini mengindikasikan bahwa model relatif lebih sulit membedakan kelas Keagamaan dengan kelas IPA dibandingkan dengan kelas lainnya. Secara keseluruhan, *confusion matrix* ini menunjukkan bahwa model memiliki performa klasifikasi yang baik, dengan tingkat kesalahan yang masih dapat ditoleransi, serta memberikan gambaran rinci mengenai distribusi prediksi benar dan salah pada masing-masing kelas.

Tabel 3. Kumulatif Varian Principal Component
Table 3. Principal Component Cummulative Varians

PC	Variants Explained	Kumulatif Varian
PC1	42,02%	42,02%
PC2	9,65%	51,68%
PC3	7,55%	59,22%
PC4	5,07%	64,29%
PC5	4,55%	68,84%
PC6	4,11%	72,95%
PC7	3,92%	76,87%
PC8	3,71%	80,59%
PC9	3,49%	84,07%
PC10	3,01%	87,08%
PC11	2,94%	90,02%
PC12	2,83%	92,85%
PC13	2,73%	95,58%
PC14	2,39%	97,97%
PC15	2,03%	100,00%

Tabel 3 menunjukkan hubungan antara jumlah komponen utama (*principal components*) dengan proporsi varians data yang dapat dijelaskan. Terlihat bahwa kontribusi varians terbesar terdapat pada komponen utama pertama dan selanjutnya mengalami penurunan secara bertahap pada komponen berikutnya. Berdasarkan Tabel 3, pemilihan sebanyak 8 komponen utama mampu menjelaskan sekitar 80,59% dari total varians data. Nilai ini menunjukkan bahwa sebagian besar informasi yang terkandung dalam 15 variabel mata pelajaran masih dapat dipertahankan meskipun telah dilakukan reduksi dimensi. Pemilihan 8 komponen utama dianggap optimal karena telah mencapai tingkat representasi varians yang cukup tinggi dengan kompleksitas model yang lebih rendah dibandingkan penggunaan seluruh variabel asli.

Gambar 6 menampilkan *SHAP summary plot* yang menggambarkan kontribusi dan interaksi beberapa komponen utama PCA (PC1, PC2, dan PC3) terhadap hasil prediksi model klasifikasi. Sumbu horizontal menunjukkan nilai SHAP interaction value, yang merepresentasikan besar dan arah pengaruh setiap komponen terhadap output model, sedangkan sumbu vertikal menunjukkan komponen utama yang dianalisis. Setiap titik merepresentasikan satu sampel data, dengan variasi warna menunjukkan perbedaan nilai komponen utama pada masing-masing observasi.

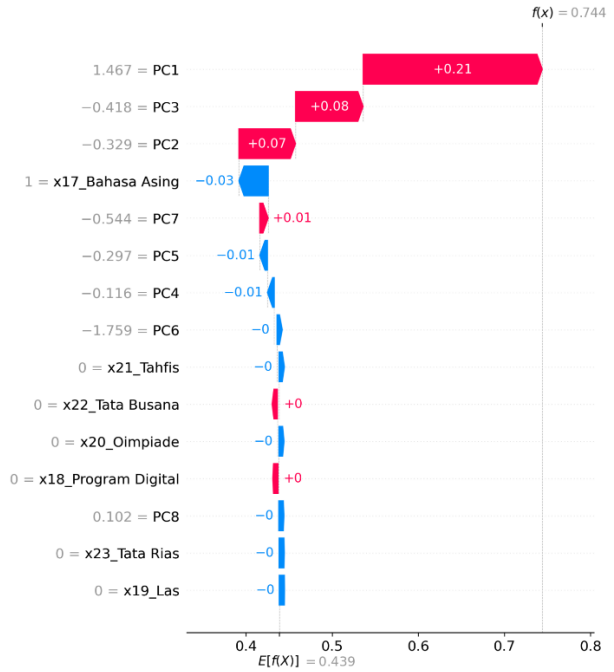


Gambar 6. Diagram Ringkasan SHAP
Figure 6. SHAP summary plot

Berdasarkan sebaran titik pada plot di Gambar 6, terlihat bahwa PC1 dan PC2 memiliki rentang nilai SHAP yang lebih lebar dibandingkan PC3, yang mengindikasikan bahwa kedua komponen tersebut memberikan kontribusi yang lebih signifikan terhadap keputusan model. Terlihat bahwa komponen utama hasil PCA (terutama PC1, PC2, dan PC3) mendominasi urutan teratas, yang menunjukkan bahwa variasi informasi akademik yang telah diringkas oleh PCA menjadi faktor utama dalam mendorong model untuk memprediksi kelas IPA. Arah pengaruh ditunjukkan oleh sumbu horizontal (nilai SHAP positif) berarti fitur tersebut meningkatkan kecenderungan output model menuju kelas 0 (IPA), sedangkan nilai SHAP negatif menurunkan kecenderungan model memilih kelas 0. Gradasi warna merepresentasikan nilai fitur dari rendah (biru) ke tinggi (merah), sehingga dapat diamati pola apakah nilai fitur yang tinggi cenderung mendorong prediksi ke IPA atau justru melemahkannya.

Visualisasi SHAP pada Gambar 6 tidak hanya memberikan informasi mengenai tingkat pengaruh masing-masing komponen utama, tetapi juga memperjelas bagaimana variasi nilai komponen tersebut berinteraksi

dalam memengaruhi hasil prediksi model secara keseluruhan. Setelah gambaran global diperoleh, analisis dilanjutkan ke level lokal (per siswa) untuk melihat fitur apa yang mendorong prediksi pada satu observasi tertentu, sebagaimana ditunjukkan pada Gambar 7.



Gambar 7. Shap Value Impact Siswa Index 1
Figure 7. SHAP Value Impact for Student Index 1

Gambar 7 menjelaskan prediksi untuk siswa index 1 dengan menunjukkan bagaimana setiap fitur menggeser nilai dasar model E[f(X)] menuju nilai keluaran akhir f(X). Batang berwarna merah menunjukkan fitur yang meningkatkan kecenderungan prediksi ke kelas 0 (IPA), sedangkan batang biru menunjukkan fitur yang menurunkan kecenderungan tersebut. Pada kasus ini, kontribusi terbesar berasal dari komponen PCA (PC1, PC3, dan PC2) yang menjadi pendorong utama perubahan skor prediksi, sementara sebagian besar fitur keterampilan memberikan pengaruh kecil atau mendekati nol. Interpretasi lokal ini melengkapi hasil global sebelumnya, karena menunjukkan fitur dominan tidak hanya penting secara rata-rata, tetapi juga benar-benar menjadi penjelas utama pada prediksi individu tertentu. Temuan ini menegaskan bahwa tidak semua komponen utama hasil PCA memiliki peran yang sama dalam proses klasifikasi. Dengan demikian, hasil analisis SHAP *global importance* ini memperkuat kesimpulan bahwa komponen utama tertentu, khususnya PC2 dan PC3, menjadi faktor kunci dalam menentukan performa model, sekaligus memberikan interpretasi yang lebih transparan terhadap hasil reduksi dimensi yang digunakan dalam pemodelan klasifikasi.

Berdasarkan seluruh hasil analisis yang telah dipaparkan, dapat disimpulkan bahwa penerapan PCA mampu mereduksi kompleksitas data tanpa menghilangkan informasi yang signifikan, sementara Random Forest memberikan kinerja klasifikasi yang stabil dan akurat. Analisis interpretabilitas menggunakan SHAP menunjukkan bahwa tidak semua komponen utama memiliki kontribusi yang sama, dimana komponen utama tertentu memiliki pengaruh dominan terhadap hasil prediksi. Dengan demikian, pendekatan yang diusulkan tidak hanya menghasilkan model prediktif yang andal, tetapi juga memberikan pemahaman yang lebih jelas mengenai faktor-faktor utama yang memengaruhi performa siswa, sehingga hasil penelitian ini dapat dijadikan dasar yang kuat untuk pengambilan keputusan akademik berbasis data.

5. KESIMPULAN

Penelitian ini menunjukkan bahwa integrasi Principal Component Analysis (PCA), Random Forest, dan SHapley Additive exPlanations (SHAP) mampu menghasilkan model klasifikasi pemilihan jurusan siswa yang akurat dan dapat dijelaskan. PCA berhasil mereduksi 15 variabel mata pelajaran menjadi 8 komponen utama tanpa menghilangkan informasi yang signifikan, sehingga meningkatkan efisiensi pemodelan dan mengurangi multikolinearitas. Model Random Forest yang dibangun menghasilkan kinerja klasifikasi yang baik dengan nilai akurasi sebesar 0.8957 serta nilai *precision*, *recall*, dan F1-*score* yang relatif seimbang pada seluruh kelas. Analisis SHAP mengungkap bahwa komponen utama tertentu, khususnya PC2 dan PC3, memiliki kontribusi paling dominan terhadap hasil prediksi, sehingga pendekatan ini tidak hanya bersifat prediktif tetapi juga interpretatif dan transparan dalam mendukung pengambilan keputusan akademik.

6. SARAN

Penelitian selanjutnya disarankan untuk memperluas variabel input dengan memasukkan faktor non-akademik, seperti minat siswa, kondisi sosial ekonomi, dan hasil tes bakat, agar rekomendasi jurusan yang dihasilkan menjadi lebih komprehensif. Selain itu, perlu dilakukan perbandingan dengan algoritma klasifikasi lain, seperti Support Vector Machine atau Gradient Boosting, guna memperoleh model dengan kinerja yang lebih optimal. Pengujian pada dataset dari sekolah atau madrasah lain juga penting dilakukan untuk menilai kemampuan generalisasi model. Dari sisi implementasi, model yang dikembangkan dapat diintegrasikan ke dalam sistem pendukung keputusan berbasis aplikasi agar dapat dimanfaatkan secara langsung oleh guru dan pihak sekolah dalam proses penjurusan siswa secara objektif dan berbasis data.

7. REFERENSI

Ahamd, M. (2025). Pengaruh Penggunaan Artificial Intelligence (Ai) Dalam Pembelajaran, Motivasi Belajar, Dan Gaya Belajar, Terhadap Keterampilan Berpikir Kritis Mahasiswa Pendidikan Ekonomi Universitas Lampung.

Al Ghifari, M. G. (2021). Prediksi Dropout Siswa dengan Kecerdasan Buatan yang Dapat Dijelaskan (Explainable AI) Menggunakan SHAP dan Machine Learning.

Albar, F. (2025). Analisis Sentimen dengan Komparasi Random Forest, SVM dan Naive Bayes menggunakan Dataset 20 Aplikasi Edukasi Anak. Universitas Islam Indonesia.

Alboaneen, D., Alqarni, R., Alqahtani, S., Alrashidi, M., Alhuda, R., Alyahyan, E., & Alshammari, T. (2023). Predicting colorectal cancer using machine and deep learning algorithms: Challenges and opportunities. *Big Data and Cognitive Computing*, *7*(2), 74.

Alifariki, L. (2022). Metode Epidemiologi Sosial.

Ananda, A. T., & Malik, M. U. I. (2025). Adaptive learning Islamic education: Literature review model pembelajaran PAI adaptif. *Jurnal Pembelajaran Dan Pengajaran*, *8*(2).

Aprilya, A., Setyani, G. R. T., Pitaloka, N., & Cahyani, S. P. (2023). Menganalisis Efektivitas Metode Evaluasi Pembelajaran di Sekolah Dasar Tinjau terhadap Kinerja Siswa dan Peningkatan Prestasi Belajar. Al-DYAS. *AL-DYAS: Jurnal Inovasi Dan Pengabdian Kepada Masyarakat*, *2*(3), 595–603.

AS, A. H., Anam, K., & Rahman, M. (2021). *Penerapan Data Mining Untuk Menemukan Pola Asosiasi Aktivitas Belajar Dan Prestasi Santri Menggunakan Algoritma Apriori*.

Astuti, M. (2024). Perbandingan Metode Random Forest dan Naive Bayes pada Klasifikasi Perilaku Mahasiswa di LMS SPADA Indonesia= Comparison of Random Forest and Naive Bayes Methods in Student Behavior Classification at LMS SPADA Indonesia. Universitas Hasanuddin.

Dita, O. P., Antara, R. M., & Winarno, A. (2025). Tanggung jawab etis penggunaan artificial intelligence di tanah pendidikan: Formulasi paradigma baru untuk teknologi otonom. *Master Manajemen*, *3*(2), 57–83.

Fadhilah, A. (2021). *Etika Privasi Data dalam Social Network Mining*.

Fajri, I. T. I., Sari, H. L., Kom, S., Kom, M., Dinata, R. K., Hasdyna, N., Retno, S., & Fadhilah, C. (2024). *Data Mining*. Serasi Media Teknologi.

Gustirani, A. (2024). Penerapan Data Mining Untuk Rekomendasi Bidang Studi Menggunakan Algoritma K-Medoids Pada SMA N 9 Kota Jambi. *Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM)*, *4*(2), 1177–1186.

Handayani, F. (2022). Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokan Mahasiswa Berdasarkan Gaya Belajar. *Jurnal Teknologi Dan Informasi*, *12*(1), 46–63.

Jufri, A. P., Asri, W. K., Mannahali, M., & Vidya, A. (2023). *Strategi pembelajaran: Menggali potensi belajar melalui model, pendekatan, dan metode yang efektif*. Ananta Vidya.

Laksono, M. I. A. (2025). Pemanfaatan Algoritma Data Mining Untuk Mendeteksi Anomali Sebagai Red Flag Dalam Audit Data E-Procurement Di Indonesia. Politeknik Keuangan Negara STAN.

Maulana, S., Premana, A., & Irawan, B. (2025). Prediksi Prestasi Akademik Siswa Terbaik Menggunakan Algoritma Decision Tree Berbasis Data Historis. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(5), 7890–7897.

Muis, A., Syafwan, H., Arfianto, A. Z., Simanjuntak, M. S., Riyandi, A., Trisnawan, A. B., Ramdhan, W., Triana, H., Saputra, M. H., & Handoko, W. (2025). *DATA MINING: Konsep, Metode, dan Aplikasi*. Faaslib Serambi Media.

Mulyanti, D. (2025). Strategi Manajemen Pendidikan di Era Digital: Optimalisasi Infrastruktur, SDM, dan Pembelajaran Berbasis Teknologi. *Jurnal Pelita Nusantara*, *2*(4), 376–383.

Nirmala, H. (2025). *AI dan Pendidikan: Peluang, Risiko, dan Strategi Implementasi untuk Guru dan Pendidikan*. PT Indonesia Delapan Kreasi Nusa.

Rahayu, P. W., Sudipa, I. G. I., Suryani, S., Surachman, A., Ridwan, A., Darmawiguna, I. G. M., Sutoyo, M. N., Slamet, I., Harlina, S., & Maysanjaya, I. M. D. (2024). *Buku ajar data mining*. PT. Sonpedia Publishing Indonesia.

Rustiyana, R., Judijanto, L., Mahendra, G. S., Kamil, Z. A., Purba, D. N., Sutoyo, M. N., Hendrayana, I. G., Pasrun, Y. P., & Prayudani, S. (2025). *Data Mining: Algoritma dan Penerapannya*. PT. Sonpedia Publishing Indonesia.

Sulika, S. (2024). *Klasifikasi kemampuan akademik peserta didik mengunakan metode Neural Network dan Metode C4. 5*. Universitas Islam Negeri Maulana Malik Ibrahim.

Supriyono, L. A., Kusumastuti, S. Y., Hartanto, T., Atika, P. D., Kamil, Z. A., Rustiyana, R., Ginting, E. F., Maylani, I., Meilani, B. D., & Arifiyanti, A. A. (2025). *Buku Ajar Big Data dan Data Mining: Konsep, Metodologi, dan Aplikasi*. PT. Sonpedia Publishing Indonesia.

Victoria, A., Vanessa, P.-B., Mensing, S., Stodtmann, S., & Maier, C. S. (2024). *Practical guide to SHAP analysis : Explaining supervised machine learning model predictions in drug development Mathematical background*. *August*, 1–15. https://doi.org/10.1111/cts.70056

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, *4*(5), e12482.