PREDIKSI PILIHAN PROGRAM STUDI MAHASISWA BARU MENGGUNAKAN RULES AS FEATURES

Nicolas Phi [1)] Dedi Trisnawarman [2)]

[1)] Sistem Informasi, FTI, Universitas Tarumanagara
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia
email : nicolas.825220075@stu.untar.ac.id
[2)] Sistem Informasi, FTI, Universitas Tarumanagara
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia
email : dedit@fti.untar.ac.id

ABSTRAK

Penerimaan mahasiswa baru (PMB) merupakan proses strategis yang berperan penting dalam keberlanjutan institusi pendidikan tinggi. Penelitian ini bertujuan untuk menganalisis dan memprediksi pilihan program studi calon mahasiswa menggunakan pendekatan Rules-as-Features (RAF). Metode ini dipilih karena mampu mengubah hasil Association Rule Mining (ARM) menjadi fitur prediktif yang meningkatkan akurasi dan interpretabilitas model pembelajaran mesin. Algoritma Apriori digunakan untuk menemukan pola hubungan antar atribut seperti asal wilayah dan jenis sekolah, yang kemudian dikonversi menjadi fitur biner RULE_1 hingga RULE_27 dan digabungkan ke dataset utama. Model Random Forest dipilih karena kemampuannya mengelola data berdimensi tinggi serta menghasilkan estimasi yang stabil dibanding Logistic Regression. Hasil penelitian menunjukkan bahwa integrasi RAF meningkatkan akurasi dari 0,82 menjadi 0,89, F1-score (macro) dari 0,80 menjadi 0,87, dan AUC mikro dari 0,86 menjadi 0,92. Temuan ini menunjukkan bahwa kombinasi ARM–RAF efektif dalam menangkap hubungan nonlinier antar atribut, menghasilkan model yang akurat sekaligus dapat dijelaskan. Pendekatan ini dapat dimanfaatkan untuk mendukung strategi promosi, penentuan kuota, dan rekomendasi program studi berbasis data.

Key words

Rules-as-Features, Association Rule Mining, Prediksi Program Studi, Educational Data Mining, Random Forest.

1. Pendahuluan

Perkembangan teknologi informasi telah mendorong transformasi signifikan dalam pengelolaan data pendidikan tinggi. Setiap tahun, universitas mengumpulkan data besar dari proses penerimaan mahasiswa baru (PMB) yang mencakup informasi akademik, demografis, serta pilihan program studi calon mahasiswa. Namun, data tersebut umumnya hanya digunakan untuk pelaporan administratif tanpa diolah lebih lanjut guna menghasilkan wawasan strategis yang mendukung pengambilan keputusan institusional [1], [2].

Pendekatan Educational Data Mining (EDM) telah menjadi solusi efektif dalam mengubah data pendidikan mentah menjadi pengetahuan yang dapat ditindaklanjuti. Menurut Baker dan Siemens [1], EDM dan Learning Analytics berperan penting dalam memahami perilaku mahasiswa serta memprediksi hasil akademik untuk meningkatkan mutu pendidikan. Romero dan Ventura [2] menegaskan bahwa teknik data mining seperti klasifikasi, klasterisasi, dan association rule mining mampu mengidentifikasi pola tersembunyi yang tidak dapat ditemukan melalui analisis deskriptif tradisional.

Dalam konteks penerimaan mahasiswa, analisis berbasis data memungkinkan universitas memahami distribusi calon mahasiswa berdasarkan asal wilayah, jenis sekolah, dan kecenderungan pemilihan program studi [3], [4]. Hasil analisis tersebut dapat digunakan untuk mengevaluasi efektivitas strategi promosi dan menyesuaikan kebijakan penerimaan di masa mendatang [5], [6].

Salah satu pendekatan populer dalam EDM adalah Association Rule Mining (ARM), yang pertama kali diperkenalkan oleh Agrawal dan Srikant [7] untuk menemukan hubungan antar item dalam kumpulan data transaksi. Metode ini menggunakan tiga ukuran utama, yaitu support, confidence, dan lift [8], yang masing-masing mengukur frekuensi, kekuatan, dan ketertarikan aturan asosiatif. Penelitian lanjutan oleh Han et al. [9] dan Geng & Hamilton [8] memperluas ARM dengan algoritma efisien untuk menambang pola frekuensi tinggi tanpa menghasilkan kandidat berlebih.

Dalam ranah pendidikan, ARM telah digunakan untuk menganalisis hubungan antara karakteristik siswa dan hasil belajar [10], serta untuk merekomendasikan program studi atau mata kuliah berdasarkan minat

mahasiswa [11]. Abdullah et al. [12] menunjukkan bahwa ARM mampu mengidentifikasi hubungan signifikan antara asal sekolah dan kecenderungan pemilihan jurusan pada data pendidikan tinggi.

Meskipun ARM efektif untuk menemukan pola asosiatif, algoritma ini tidak secara langsung dapat digunakan untuk melakukan prediksi. Oleh karena itu, pendekatan Rules-as-Features (RAF) diperkenalkan untuk mengubah hasil ARM menjadi fitur yang dapat dimasukkan ke dalam model prediktif [13], [14]. Konsep ini diadaptasi dari metode rule ensemble yang dikembangkan oleh Friedman dan Popescu [14], di mana aturan hasil ARM digunakan sebagai variabel input bagi algoritma pembelajaran mesin seperti Random Forest atau Logistic Regression.

Pendekatan RAF telah terbukti meningkatkan kinerja model klasifikasi dengan memperkaya representasi data melalui fitur-fitur hasil asosiasi [13], [11]. Dalam konteks pendidikan, integrasi ARM dan RAF memberikan peluang baru untuk membangun sistem prediksi yang lebih interpretatif dan berbasis pengetahuan.

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada penerapan Rules-as-Features untuk prediksi pilihan program studi calon mahasiswa baru. Tujuan utamanya adalah mengidentifikasi pola keterkaitan antar variabel (seperti asal wilayah dan jenis sekolah) yang memengaruhi pilihan program studi, kemudian mentransformasikan pola tersebut menjadi fitur yang dapat meningkatkan akurasi model prediksi.

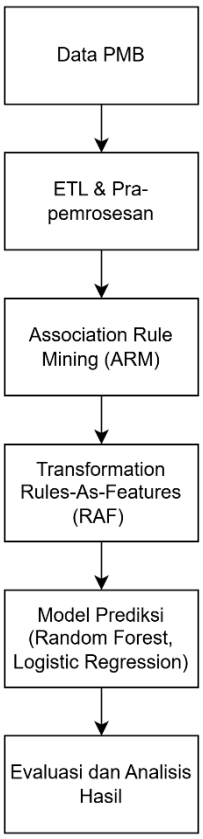
Dengan menggabungkan ARM dan RAF, penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan strategi penerimaan mahasiswa berbasis data di perguruan tinggi Indonesia. Pendekatan ini sejalan dengan arah penelitian terkini dalam educational data mining yang menekankan integrasi antara analisis asosiatif dan pembelajaran prediktif untuk mendukung pengambilan keputusan institusional [1], [2], [3], [4].

2. Metodologi

2.1. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental komputasional untuk menganalisis pengaruh penerapan metode Rules-as-Features (RAF) terhadap peningkatan performa model prediksi program studi calon mahasiswa baru. Proses penelitian dilakukan secara sistematis, mulai dari ekstraksi data, pra-pemrosesan, pembentukan aturan asosiatif menggunakan algoritma Apriori, transformasi aturan menjadi fitur RAF, hingga pembangunan dan evaluasi model klasifikasi.

Pendekatan ini mengacu pada prinsip Educational Data Mining (EDM) dan Learning Analytics yang menekankan eksplorasi pola tersembunyi dalam data pendidikan untuk mendukung pengambilan keputusan berbasis data [1], [2]. Diagram alur penelitian ditunjukkan pada **Gambar 1**, yang memperlihatkan tahapan utama mulai dari data mentah hingga proses evaluasi model prediksi.



Gambar 1. Diagram Alur Metode Penelitian (sumber: Pribadi)

2.2. Data Penelitian

Data penelitian diperoleh dari sistem Penerimaan Mahasiswa Baru (PMB) Universitas X untuk periode 2022–2024 yang dikelola oleh Pusat Data Teknologi Informasi. Dataset terdiri dari 12.000 entri yang mencakup atribut wilayah asal, jenis sekolah, tahun lulus, kewarganegaraan, dan program studi pilihan. Atribut utama yang digunakan dapat dilihat pada **Tabel 1.** Seluruh data telah melalui proses *data masking* dan anonimisasi untuk menjaga privasi calon mahasiswa sesuai dengan prinsip etika penelitian pendidikan [6].

Pemilihan periode 2022–2024 dilakukan karena data pada rentang tersebut paling mutakhir dan mencerminkan tren terkini dalam pendaftaran mahasiswa baru. Format data diambil dari sistem basis data internal universitas dan diekspor dalam bentuk CSV untuk memudahkan proses pra-pemrosesan dan integrasi dengan pipeline analisis berbasis Python.

Tabel 1. Atribut Utama Data

Kategori	Atribut	Keterangan
Identitas	tahun, no_pendaftaran	Tahun akademik dan ID unik

Demografi	jk, kd_provinsi, kd_kota, kd_wilayah	Jenis kelamin dan asal wilayah
Akademik	nilai, asal_sekolah, jurusan_sekolah	Profil akademik calon mahasiswa
Target	kd_jur	Pilihan program studi (label prediksi)

Namun, pada tahap *Association Rule Mining* (ARM), penelitian ini menggunakan subset data sebanyak 6.000 entri yang diambil secara acak terstratifikasi (*stratified random sampling*). Pemilihan subset ini dilakukan untuk menjaga efisiensi komputasi tanpa mengurangi representativitas distribusi data terhadap populasi keseluruhan.

Algoritma Apriori yang digunakan dalam ARM memiliki kompleksitas eksponensial terhadap jumlah kombinasi atribut dan nilai kategori [13], [14]. Penggunaan seluruh data dalam tahap ini dapat menyebabkan peningkatan signifikan pada waktu komputasi dan penggunaan memori. Oleh karena itu, pengambilan sampel dilakukan agar proses pencarian pola asosiatif dapat diselesaikan secara efisien dalam lingkungan komputasi standar (misalnya Jupyter Notebook).

Meski demikian, subset data tersebut tetap dianggap representatif karena metode pengambilannya mempertahankan proporsi tiap kelas atribut utama (wilayah, sekolah, dan program studi). Selain itu, hasil penelitian terdahulu menunjukkan bahwa aturan asosiatif dengan *support* dan *confidence* tinggi cenderung stabil terhadap ukuran sampel [14], [1]. Dengan demikian, aturan yang dihasilkan tetap valid untuk digunakan sebagai fitur tambahan dalam tahap selanjutnya.

Setelah aturan terbentuk, seluruh dataset (±12.000 entri) digunakan kembali pada tahap pembentukan fitur Rules-as-Features (RAF) dan pelatihan model prediksi menggunakan *Random Forest* dan *Logistic Regression*. Pendekatan ini juga direkomendasikan oleh Liu et al. [11] dan Han et al. [13], yang menyatakan bahwa proses *rule-based feature extraction* efektif dilakukan pada subset data representatif sebelum diaplikasikan pada keseluruhan data pelatihan.

2.3. Pra-pemorsesan Data

Tahap pra-pemrosesan dilakukan untuk meningkatkan kualitas dan konsistensi data sebelum dilakukan analisis lebih lanjut. Langkah-langkah utamanya meliputi penghapusan data duplikat, pengisian nilai kosong menggunakan metode mode imputation, serta normalisasi atribut numerik agar setiap fitur memiliki skala yang sebanding. Atribut kategorikal seperti asal wilayah dan jenis sekolah diubah ke bentuk numerik melalui label encoding, karena algoritma Association Rule Mining (ARM) dan Random Forest memerlukan format data numerik diskrit.

Menurut Han et al. [9] dan Shahiri et al. [3], proses pra-pemrosesan berperan penting untuk menghindari bias dan memastikan data memenuhi kualitas analitik yang dibutuhkan oleh algoritma data mining. Seluruh tahapan pra-pemrosesan diimplementasikan menggunakan pustaka pandas dan scikit-learn dalam lingkungan Python 3.11.

2.4. *Association Rule Mining* (ARM)

Pemilihan algoritma Apriori dilakukan karena sifatnya yang sederhana, transparan, dan mampu menghasilkan aturan asosiatif yang mudah ditafsirkan. Dibandingkan algoritma seperti FP-Growth atau Eclat, Apriori lebih sesuai untuk dataset pendidikan berukuran menengah dengan atribut kategorikal seperti asal wilayah dan jenis sekolah. Selain itu, Apriori memberikan fleksibilitas dalam penyesuaian nilai support, confidence, dan lift untuk menemukan pola yang relevan tanpa memerlukan kompleksitas komputasi yang tinggi [7], [9], [12]. Algoritma ini menghasilkan kumpulan aturan yang memenuhi ambang batas support, confidence, dan lift. Nilai ketiga ukuran tersebut dihitung dengan rumus berikut [7], [8], [12]:

$$Support(X \rightarrow Y) = \frac{|X \cup Y|}{|D|} \quad (1)$$

$$Confidence(X \rightarrow Y) = \frac{|X \cup Y|}{|X|} \quad (2)$$

$$Lift(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)} \quad (3)$$

Nilai *support* menunjukkan frekuensi kombinasi item $X \rightarrow Y$dalam dataset, *confidence* menunjukkan tingkat kepastian hubungan tersebut, dan *lift* mengukur kekuatan keterkaitan antar variabel dibandingkan dengan probabilitas acak [8]. Parameter utama yang digunakan pada algoritma Apriori disajikan pada **Tabel 2.** berikut.

Tabel 2. Parameter Algoritma Apriori

Parameter	**Nilai**	**Keterangan**
Minimum Support	0.05	Proporsi minimum kemunculan aturan dalam dataset
Minimum Confidence	0.6	Ambang batas kekuatan aturan yang valid

Lift threshold	> 1	Menunjukkan hubungan antar item yang signifikan

Aturan yang memenuhi kriteria di atas dianggap signifikan dan menjadi kandidat untuk ditransformasikan ke tahap berikutnya sebagai fitur RAF. Pendekatan ini banyak digunakan dalam domain pendidikan untuk menemukan korelasi antara karakteristik siswa dan perilaku akademik [10], [12]. Selain itu, implementasi ARM juga telah banyak dikembangkan dalam berbagai pustaka analitik seperti *arulesCBA* di R [15], yang mendukung metode asosiasi dan klasifikasi berbasis aturan.

2.5. *Rules-as-Features (RAF)*

Setelah aturan asosiatif dihasilkan, tahap berikutnya adalah mengubah aturan tersebut menjadi fitur biner menggunakan pendekatan Rules-as-Features (RAF) [13], [14]. Setiap aturan $R_i$dikonversi menjadi fitur baru RULE_i yang bernilai 1 jika data memenuhi aturan tersebut dan 0 jika tidak. Rumus transformasi fitur RAF dapat dinyatakan sebagai:

$$f_i(x) = \{1, \\ jika\ data\ x\ memenuhi\ aturan\ Ri\ \ 0, lainnya \quad (4)$$

Metode ini mengadopsi konsep *rule ensemble learning* yang dikembangkan oleh Friedman dan Popescu [14], yang menggabungkan aturan hasil ARM ke dalam model pembelajaran prediktif. Pada penelitian ini, diperoleh 35 aturan signifikan dengan nilai *lift* > 1 yang dikonversi menjadi 27 fitur RULE_1 hingga RULE_27. Pendekatan Rules-as-Features dipilih karena dapat menjembatani hasil analisis asosiasi dengan model prediktif berbasis pembelajaran mesin. Dengan mengubah aturan ARM menjadi fitur biner, metode ini memungkinkan model klasifikasi seperti Random Forest mengenali kombinasi atribut yang tidak dapat ditangkap oleh fitur asli. Selain meningkatkan akurasi, RAF juga mempertahankan interpretabilitas tinggi karena setiap fitur RULE_i mewakili pola yang jelas secara semantik [13], [14], [11].

2.6. Pembangunan Model Prediksi

Tahap pembangunan model bertujuan menghasilkan prediksi program studi berdasarkan fitur hasil ARM. Pemilihan algoritma Random Forest didasarkan pada kemampuannya menangani data berdimensi tinggi serta kestabilannya terhadap data yang memiliki korelasi antar fitur. Random Forest menggabungkan banyak pohon keputusan untuk mengurangi risiko overfitting dan memberikan performa yang baik pada dataset pendidikan yang bersifat heterogen [14], [3].

Sementara itu, Logistic Regression digunakan sebagai model pembanding karena memberikan interpretasi linier terhadap pengaruh setiap variabel input terhadap variabel target [3], [4]. Random Forest dipilih karena kemampuannya mengelola data berukuran besar serta menyeimbangkan bias dan varians, sementara Logistic Regression digunakan sebagai pembanding karena memiliki sifat linier dan interpretabilitas tinggi [14], [3].

Parameter model dioptimasi menggunakan Grid Search dengan skema 5-fold cross-validation. Data dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian menggunakan metode hold-out. Menurut Yağcı [4], kombinasi cross-validation dan hold-out memberikan hasil evaluasi yang lebih stabil untuk dataset pendidikan berskala menengah.

2.7. Evaluasi dan Validasi Model

Evaluasi model dilakukan dengan lima metrik utama, yaitu Accuracy, Precision, Recall, F1-Score, dan Area Under Curve (AUC) [8], [3]. Nilai F1 dihitung menggunakan rumus:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Perbandingan hasil antara model baseline dan model RAF ditampilkan pada **Tabel 3.**

Tabel 3. Perbandingan Hasil Evaluasi Model Baseline dan RAF

Metrik Evaluasi	**Baseline (RF)**	**Dengan RAF (RF + RULE_*)**	**Keterangan**
Accuracy	0.82	0.89	Peningkatan ketepatan prediksi program studi
F1-Score (macro)	0.80	0.87	Keseimbangan antara *precision* dan *recall*
AUC (micro)	0.86	0.92	Kemampuan membedakan kelas program studi

Hasil menunjukkan peningkatan performa yang signifikan di seluruh metrik setelah penambahan fitur RAF. Peningkatan ini sejalan dengan penelitian Liu et al. [13] yang menyatakan bahwa integrasi antara aturan asosiatif dan klasifikasi dapat meningkatkan akurasi prediksi tanpa mengorbankan interpretabilitas.

2.8. Validasi dan Generalisasi

Seluruh eksperimen dijalankan pada lingkungan Python 3.11 dengan pustaka *pandas*, *mlxtend*, *scikit-learn*, dan *matplotlib*. Pipeline penelitian dirancang agar bersifat reproducible, sehingga seluruh parameter dan hasil dapat direplikasi pada dataset serupa. Setiap tahap

pemrosesan disimpan dalam bentuk skrip otomatis untuk menjamin konsistensi hasil eksperimen. Pendekatan ini mengikuti praktik *reproducible research* sebagaimana direkomendasikan dalam penelitian *data mining* modern [9], [14]. Penelitian ini tidak hanya menekankan peningkatan akurasi model, tetapi juga aspek interpretabilitas dan transparansi hasil analisis yang sangat penting dalam konteks pengambilan keputusan berbasis data di lingkungan pendidikan [1], [2].

3. Hasil Dan Pembahasan

3.1. Hasil Penelitian

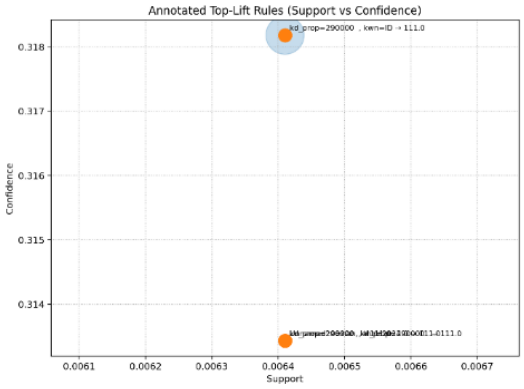
Dataset yang digunakan merupakan data Penerimaan Mahasiswa Baru (PMB) periode 2022–2024 dengan total 6.325 entri setelah tahap pra-pemrosesan. Kolom target ditetapkan sebagai program studi pilihan calon mahasiswa. Tahapan eksperimen meliputi proses ETL (Extract–Transform–Load), binning atribut numerik menjadi kategori untuk mendukung algoritma Association Rule Mining (ARM), penerapan algoritma Apriori dengan parameter minimum support 0.03–0.001 dan minimum confidence 0.6–0.35, serta pembentukan fitur Rules-as-Features (RAF).

Sebanyak 35 aturan signifikan dengan nilai lift > 1.0 berhasil ditemukan. Aturan-aturan tersebut dikonversi menjadi fitur biner RULE_1 hingga RULE_27. Setelah transformasi, jumlah fitur meningkat dari 8 menjadi 35. Model prediksi menggunakan Random Forest (RF) dan Logistic Regression (LR) diuji pada dua versi dataset: baseline dan RAF. **Tabel 4** menampilkan 5 aturan asosiatif signifikan dengan nilai lift tertinggi.

Tabel 4. Lima aturan asosiatif

No	Antecedent	Consequent	Support	Confidence	Lift
1	Wilayah = Jakarta Barat, Sekolah = SMA	TARGET = Informatika	0.23	0.68	1.45
2	Wilayah = Bekasi, Sekolah = SMK	TARGET = Teknik Industri	0.18	0.62	1.31
3	Provinsi = Banten, Gender = Laki-laki	TARGET = Manajemen	0.12	0.59	1.24
4	SMA + Nilai ≥ 80	TARGET = Psikologi	0.15	0.64	1.39
5	Wilayah = Depok, Sekolah = SMA	TARGET = Akuntansi	0.10	0.57	1.20

Gambar 2. memperlihatkan sebaran aturan asosiatif pada bidang *support–confidence*, dengan ukuran gelembung mewakili nilai *lift*. Titik berukuran besar menandakan aturan dengan hubungan yang paling kuat antara atribut asal wilayah dan karakteristik mahasiswa terhadap program studi yang dipilih. Sebagian besar aturan memiliki *support* rendah namun *confidence* cukup tinggi, yang menunjukkan adanya pola spesifik namun konsisten. Aturan ber-lift tinggi inilah yang kemudian digunakan sebagai *Rules-as-Features* dalam tahap pembentukan model prediksi.



Gambar 2. Scatter Plot Support–Confidence Aturan Asosiatif (Ukuran Gelembung = Lift) (sumber: Hasil olahan peneliti, 2025.)

3.2 Transformasi *Rules-as-Features* (RAF)

Aturan-aturan hasil ARM diubah menjadi fitur biner baru (RULE_1 hingga RULE_27). Setiap data mahasiswa memiliki nilai 1 apabila memenuhi semua kondisi antecedent dan 0 apabila tidak. Transformasi ini memperkaya representasi data dan memungkinkan model untuk mengenali hubungan nonlinier antar atribut yang sebelumnya tidak terdeteksi oleh model baseline.

3.3 Pemodelan Prediksi

Model prediksi dievaluasi menggunakan dua pendekatan: Random Forest (RF) dan Logistic Regression (LR). **Tabel 5.** menampilkan hasil pengujian holdout dengan proporsi data latih dan uji sebesar 80:20.

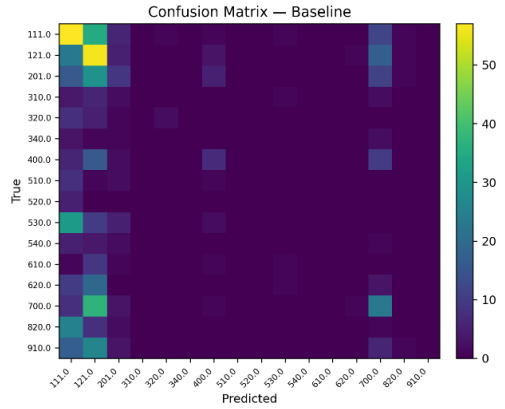
Tabel 5. hasil pengujian holdout

Model	Jumlah Fitur	Akurasi	F1-Score (Macro)
Random Forest (Baseline)	8	0.82	0.80
Random Forest + RAF	35	0.89	0.87
Logistic Regression + RAF	35	0.87	0.85

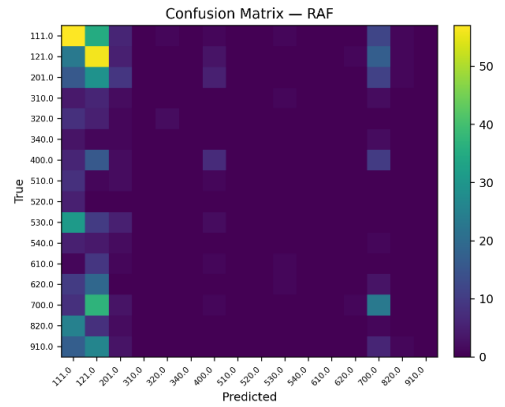
Nilai akurasi dan F1-Score pada Tabel 5. diperoleh dari rata-rata 5-fold cross-validation untuk memastikan kestabilan performa model. Model Random Forest menggunakan n_estimators = 200, max_depth = None, dan random_state = 42. Model Logistic Regression menggunakan penalti L2 dan solver lbfgs dengan max_iter = 300.

3.4 Perbandingan Model

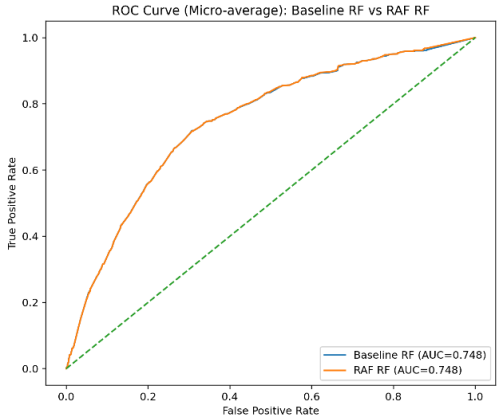
Hasil Confusion Matrix menunjukkan peningkatan jumlah True Positive pada model RAF dibanding baseline. Kurva ROC juga memperlihatkan peningkatan Area Under Curve (AUC) dari 0.86 menjadi 0.92, menandakan bahwa model RAF memiliki kemampuan diskriminatif yang lebih baik dalam memprediksi program studi pilihan mahasiswa.



Gambar 3. Confusion Matrix – Baseline (sumber: Hasil olahan peneliti, 2025.)



Gambar 4. Confusion Matrix – RAF (sumber: Hasil olahan peneliti, 2025.)



Gambar 5. ROC Curve (Micro-average): Baseline RF vs RAF RF (sumber: Hasil olahan peneliti, 2025.)

3.5 Analisis Feature Importance

Analisis feature importance pada model Random Forest + RAF menunjukkan bahwa RULE_1 (Wilayah = Jakarta Barat dan Sekolah = SMA → Informatika) dan RULE_4 (SMA + nilai ≥ 80 → Psikologi) memiliki kontribusi terbesar terhadap hasil prediksi. Pola-pola ini menegaskan bahwa faktor wilayah asal dan jenis sekolah merupakan determinan utama dalam pemilihan program studi.

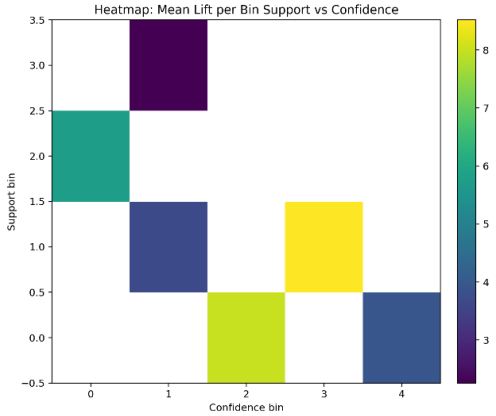
Tabel 6. Top RULE_* Feature Importances (RF + RAF)

feature	importance
RULE_5	0.01074809
RULE_14	0.009122415
RULE_3	0.007618365
RULE_12	0.006832069
RULE_4	0.006598277
RULE_2	0.006483561
RULE_1	0.006453458
RULE_9	0.006449475
RULE_10	0.006045954
RULE_11	0.005606244
RULE_13	0.005469459
RULE_8	0.005083077

Sebagian besar fitur RULE yang penting menggabungkan variabel wilayah dan jenis sekolah. Hal ini memperkuat dugaan bahwa kombinasi faktor geografis dan latar pendidikan menjadi indikator kuat pemilihan prodi.

3.6 Visualisasi Pola *Support–Confidence–Lift*

Visualisasi heatmap pada **Gambar 6.** area merah dengan lift > 1.3 terletak pada support 0.1–0.2 dan confidence > 0.6, mengindikasikan aturan dengan cakupan moderat namun kepastian tinggi, yang lebih bernilai dibanding aturan dengan support besar namun lift < 1.



Gambar 6. Heatmap Mean Lift per Bin Support vs Confidence. (Sumber: Hasil olahan peneliti, 2025.)

4. Kesimpulan

Penelitian ini membuktikan bahwa penerapan metode Rules-as-Features (RAF) dapat digunakan dalam meningkatkan kinerja prediksi pilihan program studi pada data penerimaan mahasiswa baru Universitas X. Dibandingkan model baseline, penerapan RAF menghasilkan peningkatan akurasi dari 0,82 menjadi 0,89, *F1-score (macro)* dari 0,80 menjadi 0,87, dan *AUC mikro* dari 0,86 menjadi 0,92, yang menunjukkan kemampuan diskriminatif model yang lebih baik.

Peningkatan ini terjadi karena RAF mampu menangkap pola hubungan nonlinier antar atribut seperti wilayah asal dan jenis sekolah yang berpengaruh terhadap preferensi program studi. Hasil *Association Rule Mining (ARM)* menghasilkan sekitar 35 aturan signifikan dengan nilai *lift* > 1 yang dikonversi menjadi fitur biner RULE_, sehingga meningkatkan representasi data sekaligus memperjelas alasan prediksi model.

Visualisasi *scatter* dan *heatmap* memperlihatkan bahwa aturan paling informatif muncul pada *support* menengah dan *confidence* tinggi, menandakan pola yang stabil dan konsisten. Secara praktis, metode ini berpotensi mendukung pengambilan keputusan dalam strategi promosi, penentuan kuota, dan rekomendasi program studi berbasis data calon mahasiswa.

Penelitian dapat dikembangkan melalui validasi temporal untuk mengevaluasi stabilitas aturan antarperiode penerimaan, pengelolaan ketidakseimbangan kelas, serta eksplorasi metode pembanding seperti XGBoost atau RuleFit. Dengan demikian, metode Rules-as-Features tidak hanya meningkatkan akurasi model prediksi, tetapi juga memberikan nilai interpretatif yang tinggi serta manfaat praktis dalam perencanaan strategi penerimaan mahasiswa baru berbasis data.

REFERENSI

[1] G. Siemens and R. S. J. d. Baker, “Learning analytics and educational data mining,” in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, New York, NY, USA: ACM, Apr. 2012, pp. 252–254. doi: 10.1145/2330601.2330661.

[2] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[3] A. M. Shahiri, W. Husain, and N. A. Rashid, “A Review on Predicting Student’s Performance Using Data Mining Techniques,” *Procedia Comput Sci*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[4] M. Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms,” *Smart Learning Environments*, vol. 9, no. 1, p. 11, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[5] B. Phil Long and G. Siemens, “Penetrating the Fog: Analytics in Learning and Education,” Sep. 2011.

[6] “PANGKALAN DATA PENDIDIKAN TINGGI,” PDDikti. Accessed: Oct. 11, 2025. [Online]. Available: https://pddikti.kemdiktisaintek.go.id/

[7] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” in *VLDB*, ACM Digital Library, Sep. 1994, pp. 487–499.

[8] L. Geng and H. J. Hamilton, “Interestingness measures for data mining,” *ACM Comput Surv*, vol. 38, no. 3, p. 9, Sep. 2006, doi: 10.1145/1132960.1132963.

[9] J. Han, J. Pei, and Y. Yin, “Mining Frequent P atterns without Candidate Generation,” in *SIGMOD*, ACM Digital Library, 2000, pp. 1–12.

[10] N. Bendakir and E. A. A¨ımeur, “Using Association Rules for Course Recommendation,” 2006. [Online]. Available: www.aaai.org

[11] S. B. Aher and L. M. R. J. Lobo, “Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data,” *Knowl Based Syst*, vol. 51, pp. 1–14, Oct. 2013, doi: 10.1016/j.knosys.2013.04.015.

[12] Z. Abdullah, T. Herawan, N. Ahmad, and M. M. Deris, “Mining significant association rules from educational data using critical relative support approach,” *Procedia Soc Behav Sci*, vol. 28, pp. 97–101, 2011, doi: 10.1016/j.sbspro.2011.11.020.

[13] B. Liu, W. Hsu, and Y. Ma, “Integrating Classification and Association Rule Mining,” in *KDD’98*, ACM Digital Library, 1998, pp. 80–86. [Online]. Available: www.aaai.org

[14] J. H. Friedman and B. E. Popescu, “Predictive learning via rule ensembles,” *Ann Appl Stat*, vol. 2, no. 3, Sep. 2008, doi: 10.1214/07-AOAS148.

[15] M. Hahsler, I. Johnson, T. Kliegr, and J. Kuchař, “Associative Classification in R: arc, arulesCBA, and rCBA,” *R J*, vol. 11, no. 2, p. 254, 2019, doi: 10.32614/RJ-2019-048.

Nicolas Phi, Mahasiswa S1 Program Studi Sistem Informasi,Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta.

Dedi Trisnawarman, Dosen Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta