

PRINCIPAL COMPONENT ANALYSIS PADA RANDOM FOREST UNTUK DETEKSI DINI DROPOUT MAHASISWA

Rully Ghozali¹, Kevin Ramadhan Oktaviano², Anando Ferdi Setiawan³, Muhammad Rio Ramadani⁴, Heru Saputro⁵

^{1,2,3,4,5} Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Nahdlatul Ulama Jepara; Jalan Taman Siswa, Pekeng, Tahunan, Kec. Jepara, Kabupaten Jepara, Jawa Tengah 59427, Indonesia. Telepon: (0291) 595320

email: rullyghozaly@gmail.com¹, 241250000532@unisnu.ac.id², ferdianando2@gmail.com³, riormdn610@gmail.com⁴, herusaputro@unisnu.ac.id⁵

Abstrak

Penelitian ini menguji pengaruh Principal Component Analysis (PCA) terhadap performa Random Forest dalam deteksi dini risiko dropout mahasiswa. Dua skema dibandingkan, yaitu Random Forest tanpa PCA dan dengan PCA, menggunakan Orange Data Mining (Orange3) pada dataset Kaggle Predict Students' Dropout and Academic Success (4.424 data, 35 atribut, tiga kelas target). PCA diterapkan dengan dua komponen utama (explained variance 26%). Kelas Dropout dievaluasi menggunakan validasi silang 10-lipat bertingkat dengan metrik AUC, Akurasi, F1, Presisi, Recall, dan MCC. Random Forest tanpa PCA memperoleh performa lebih tinggi pada seluruh metrik (AUC 0,905; Accuracy 0,856; F1 0,771; Precision 0,790; Recall 0,754; MCC 0,667) dibandingkan Random Forest dengan PCA (AUC 0,834; Accuracy 0,800; F1 0,676; Precision 0,704; Recall 0,650; MCC 0,533), yang diduga disebabkan oleh rendahnya variansi kumulatif yang dipertahankan.

Kata kunci: Principal Component Analysis; Random Forest; Dropout Mahasiswa; Educational Data Mining; Ekstraksi Fitur

Abstract

This This study examines the effect of Principal Component Analysis (PCA) on the performance of Random Forest in the early detection of student dropout risk. Two approaches were compared: Random Forest without PCA and Random Forest with PCA, using Orange Data Mining (Orange3) on the Kaggle dataset "Predict Students' Dropout and Academic Success" (4,424 data points, 35 features, three target classes). PCA was applied with two principal components (26% explained variance). The "Dropout" class was evaluated using 10-fold hierarchical cross-validation with the following metrics: AUC, Accuracy, F1, Precision, Recall, and MCC. Random Forest without PCA achieved higher performance across all metrics (AUC 0.905; Accuracy 0.856; F1 0.771; Precision 0.790; Recall 0.754; MCC 0.667) compared to Random Forest with PCA (AUC 0.834; Accuracy 0.800; F1 0.676; Precision 0.704; Recall 0.650; MCC 0.533), which is likely due to the low cumulative variance retained.

Keywords: Principal Component Analysis; Random Forest; Student Dropout; Educational Data Mining; Feature Extraction

PENDAHULUAN

Dropout mahasiswa merupakan permasalahan signifikan yang dihadapi institusi pendidikan tinggi di berbagai negara. Fenomena ini tidak hanya berdampak pada capaian akademik individu, tetapi juga menimbulkan kerugian finansial bagi institusi serta berpotensi memengaruhi prospek karier dan kondisi sosial mahasiswa yang bersangkutan. Sejumlah faktor telah diidentifikasi sebagai penyebab dropout, mencakup aspek akademik, sosio-ekonomi, demografis, hingga kondisi keluarga. Institusi pendidikan tinggi menghadapi tantangan tersendiri karena kompleksitas dan keragaman faktor tersebut, yang membuat deteksi dini mahasiswa yang berisiko dropout menjadi tantangan tersendiri [1].

Educational Data Mining (EDM) berkembang sebagai bidang yang memanfaatkan teknik penambangan data dan pembelajaran mesin untuk menganalisis data pendidikan, termasuk dalam memprediksi performa maupun risiko dropout mahasiswa. Berbagai algoritma pembelajaran mesin telah diterapkan untuk keperluan ini, di antaranya Decision Tree, Support Vector Machine, Logistic Regression, hingga metode ensemble seperti Random Forest dan Gradient Boosting [2]. Random Forest secara khusus banyak digunakan karena kemampuannya menangani data dengan atribut yang beragam serta ketahanannya terhadap overfitting melalui mekanisme agregasi banyak pohon keputusan [3][4].

Sebaliknya, masalah dimensi tinggi (high dimensionality) dan redundansi antar fitur dapat muncul karena dataset pendidikan biasanya memiliki jumlah atribut yang besar. Untuk mengurangi dimensi data, analisis komponen utama (PCA) adalah metode ekstraksi fitur yang populer. Metode ini memungkinkan mempertahankan variansi informasi dalam bagian baru yang tidak berkorelasi sebanyak mungkin. [5][6]. Banyak penelitian telah dilakukan mengenai penerapan PCA sebagai tahap praproses sebelum klasifikasi; namun, temuan tersebut tidak selalu konsisten [7]. Beberapa penelitian menunjukkan bahwa kombinasi Random Forest dan PCA dapat meningkatkan kinerja klasifikasi, terutama pada data berskala besar [8]. Studi lain justru menemukan bahwa penggunaan PCA dapat menurunkan performa Random Forest, Karena Random Forest pada dasarnya memiliki mekanisme seleksi fitur internal yang memilih atribut secara acak pada setiap pemisahan pohon, pengurangan dimensi tambahan melalui PCA tidak selalu menguntungkan karena dapat menghapus informasi yang dapat digunakan untuk mengklasifikasikan target.

Hasil yang berbeda-beda tersebut menunjukkan bahwa pengaruh PCA terhadap kinerja Random Forest bersifat kontekstual dan bergantung pada karakteristik kumpulan data serta jumlah komponen utama yang dipertahankan [8]. Dengan menggunakan dataset Predict students' dropout and academic success , yang terdiri dari 4.424 data dengan 35 atribut dan tiga kelas target, yaitu Graduate, Dropout, dan Enrolled, penelitian ini bertujuan untuk menguji secara langsung pengaruh PCA terhadap performa Random Forest dalam konteks deteksi awal dropout mahasiswa.

Dalam penelitian ini, perangkat lunak Orange Data Mining (Orange3) digunakan untuk membandingkan dua skema klasifikasi: Random Forest tanpa PCA dan Random Forest dengan PCA. Orange3 memungkinkan penyusunan workflow analisis data secara visual [9]. Untuk menilai performa, validasi silang berstratifikasi digunakan untuk menilai metrik AUC, Accuracy, F1 Score, Precision, Recall, dan Matthews Correlation Coefficient (MCC), dengan alasan karena keandalannya dalam menilai klasifikasi pada data dengan distribusi kelas yang tidak seimbang [10]. Penelitian ini bertujuan untuk memberikan bukti empiris mengenai dampak penerapan PCA dengan jumlah komponen terbatas terhadap kinerja Random Forest dalam kasus dropout mahasiswa dan memberikan saran praktis tentang bagaimana PCA dapat diterapkan dalam situasi tersebut.

METODE

Tujuan penelitian ini adalah untuk melihat bagaimana penerapan Principal Component Analysis (PCA) berdampak pada kinerja algoritma Random Forest ketika mendeteksi risiko dropout mahasiswa sejak dini. Pengujian dilakukan dengan membandingkan dua skema klasifikasi yang menggunakan perangkat lunak, dataset, dan tahapan praproses yang sama. Hasil dari kedua skema ini dapat digunakan untuk mengevaluasi bagaimana penerapan PCA mempengaruhi kinerja model klasifikasi.

Dataset Penelitian

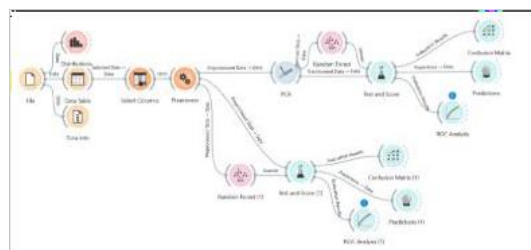
Dataset yang digunakan bersumber dari Kaggle dengan judul Predict Students' Dropout and Academic Success, yang merupakan replikasi dari dataset yang dikumpulkan oleh institusi pendidikan tinggi di Portugal. Kumpulan data tersebut mencakup 4.424 catatan siswa dengan 35 atribut yang berkaitan dengan informasi demografis, akademik, sosioekonomi, dan keluarga. Atribut target (Target) terdiri atas tiga kelas, yaitu Graduate, Dropout, dan Enrolled.

Perangkat Lunak

Seluruh proses analisis dilakukan menggunakan Orange Data Mining (Orange3), sebuah perangkat lunak data mining berbasis pemrograman visual yang memungkinkan analisis data penyusunan workflow dibangun tanpa harus menulis kode pemrograman secara langsung [11]. Orange3 menyediakan widget siap pakai untuk prapemrosesan, ekstraksi fitur, klasifikasi, dan evaluasi model dalam lingkungan kerja terintegrasi.

Rancangan Workflow

Penelitian ini mengembangkan dua proses kerja yang bekerja pada dataset yang sama secara bersamaan :



Gambar 1. Rancangan Workflow

Proses ini menggunakan tahapan praproses, dataset, dan parameter algoritma yang sama. Satu-satunya perbedaan adalah penggunaan widget PCA sebelum algoritma Random Forest pada salah satu cabang. Tujuan dari rancangan ini adalah untuk mengevaluasi dampak PCA terhadap kinerja klasifikasi Random Forest.

Praproses Data

Widget Preprocess melakukan tahap praproses dengan dua perlakuan: imputasi nilai yang hilang (impute missing values) dan standarisasi fitur (normalize features). PCA bergantung pada varians data, sehingga jika data tidak distandarisasi terlebih dahulu, fitur-fitur dengan rentang nilai yang lebih luas dapat mendominasi pembentukan komponen utama. Oleh karena itu, standarisasi fitur merupakan langkah penting yang harus dilakukan sebelum menggunakan PCA [5].

Ekstraksi Fitur dengan Principal Component Analysis

Pada salah satu cabang, PCA diterapkan dengan opsi normalisasi variabel diaktifkan, menggunakan widget PCA yang tersedia di Orange3. Jumlah komponen utama yang akan dipertahankan ditetapkan sebanyak dua komponen, dengan varians kumulatif yang dijelaskan sebesar 26%. Untuk proses klasifikasi, output yang digunakan adalah data transformasi yang merupakan representasi data yang telah sepenuhnya direduksi ke dalam dua komponen utama tersebut. Penting untuk memilih output ini, karena penggunaan data yang salah dapat mengakibatkan model tetap menerima semua atribut asli sebagai metadata tambahan, sehingga membuat pengujian terhadap efek PCA menjadi tidak valid.

Algoritma Klasifikasi

Kedua proses tersebut menggunakan algoritma klasifikasi yang sama, yaitu Random Forest. Parameter-parameternya sama, yaitu jumlah pohon sebanyak 10, tanpa membatasi jumlah atribut yang dipertimbangkan pada setiap pemisahan (split), pelatihan yang dapat direplikasi diaktifkan, tanpa membatasi kedalaman pohon, dan batas minimum subset sebesar 5 sebelum pemisahan dihentikan. Pada kedua model tersebut, penyeimbangan distribusi kelas, atau distribusi kelas yang seimbang, tidak diterapkan.

Evaluasi Model

Evaluasi performa model dilakukan melalui validasi silang (cross validation) 10 lipatan dengan stratifikasi (stratified), yang mempertahankan proporsi kelas pada setiap lipatan agar sebanding dengan distribusi kelas pada dataset asli [10]. Metrik evaluasi yang digunakan meliputi Area Under the Curve (AUC), Accuracy, F1 Score, Precision, Recall, dan Matthews Correlation Coefficient (MCC). Hasil evaluasi difokuskan pada kelas dropout (one-vs-rest) karena hal tersebut merupakan topik utama dalam penelitian ini. Sebagaimana ditunjukkan oleh penelitian-penelitian sebelumnya, MCC dipilih sebagai salah satu metrik utama karena keandalannya dalam menilai kinerja klasifikasi pada data dengan distribusi kelas yang tidak seimbang. Widget ROC Curve, Confusion Matrix, dan tabel Predictions menampilkan visualisasi hasil evaluasi.

HASIL DAN PEMBAHASAN

Hasil Evaluasi Model

Evaluasi performa dilakukan melalui validasi silang 10 lipatan berstratifikasi pada kedua skema klasifikasi. Hasil evaluasi terhadap kelas Dropout untuk Random Forest tanpa PCA dan Random Forest dengan PCA disajikan pada Tabel 1.

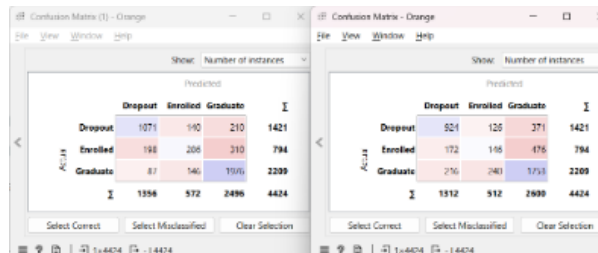
Tabel 1 Perbandingan performa Random Forest tanpa PCA dan dengan PCA

Metrik	Random Forest tanpa PCA	Random Forest dengan PCA
AUC	0,905	0,834
Accuracy	0,856	0,800
F1 Score	0,771	0,676
Precision	0,790	0,704
Recall	0,754	0,650
MCC	0,667	0,533

Random Forest tanpa PCA secara konsisten menerima nilai yang lebih tinggi pada seluruh

metrik evaluasi daripada Random Forest dengan PCA, seperti yang ditunjukkan oleh hasil Tabel 1. Tercatat perbedaan terbesar pada MCC (0,134) dan Recall (0,104), sementara perbedaan AUC sebesar 0,071.

Confosion Matrix



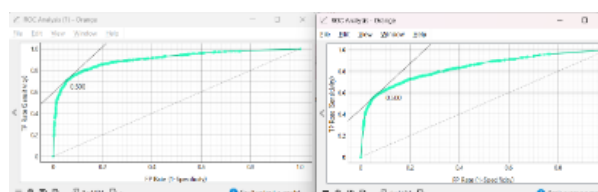
Gambar 2. Confusion Matrix Pada Kedua Model

Confusion matrix untuk kedua menunjukkan matriks kebingungan untuk kedua model pada kelas Dropout, Teraftar, dan Lulus. Pada Random Forest tanpa PCA, dari 1.421 mahasiswa yang sebenarnya berstatus Dropout, sebanyak 1.071 mahasiswa berhasil diklasifikasikan dengan benar, sementara 350 mahasiswa lainnya salah diklasifikasikan ke dalam kelas Enrolled dan Graduate.

Dengan menerapkan PCA pada Random Forest, hanya 924 dari jumlah mahasiswa Dropout yang sama yang diklasifikasikan dengan benar, sementara 497 mahasiswa salah diklasifikasikan ke dalam kelas Enrolled dan Graduate.

Pola kesalahan klasifikasi yang paling menonjol pada model PCA adalah meningkatnya jumlah siswa. Jumlah siswa yang putus sekolah yang diklasifikasikan dengan benar sebagai lulusan meningkat dari 210 pada model tanpa PCA menjadi 371 pada model dengan PCA. Pada kelas “Enrolled”, terdapat pola yang sama, yaitu kesalahan klasifikasi ke kelas “Graduate” meningkat dari 310 menjadi 476. Pola ini menunjukkan bahwa pengurangan dimensi melalui PCA mengakibatkan hilangnya sebagian informasi dalam penelitian ini. Hal ini membantu model membedakan mahasiswa yang berisiko (putus sekolah dan masih terdaftar) dari mahasiswa yang berhasil menyelesaikan studinya (lulus).

ROC Analysis



Gambar 3. Kurva ROC Pada Kedua Model

Pada Gambar 3 Kurva ROC menunjukkan kedua model dalam bentuk kurva yang serupa, dengan kurva Random Forest tanpa PCA kurva tersebut mendekati sudut kiri atas dibandingkan dengan kurva Random Forest yang menggunakan PCA, sejalan dengan nilai AUC yang lebih tinggi pada model tanpa PCA (0,905 dibandingkan dengan 0,834). Meskipun model dengan PCA kehilangan kemampuan diskriminasi untuk kelas Dropout, kedua kurva tersebut masih jauh di atas garis diagonal acak (AUC=0,5), yang berarti bahwa kedua model tersebut masih memiliki kemampuan diskriminasi yang baik untuk kelas Dropout.

Pembahasan

Penelitian ini menunjukkan bahwa penggunaan PCA dengan dua komponen utama yang hanya

mampu menjelaskan 26% dari varians kumulatif secara konsisten menurunkan kinerja Random Forest berdasarkan semua metrik evaluasi. Hasil ini bertolak belakang dengan penelitian sebelumnya yang melaporkan bahwa PCA yang digabungkan dengan Random Forest dapat meningkatkan kinerja klasifikasi, terutama pada data berdimensi sangat tinggi dengan sejumlah komponen yang mempertahankan varians kumulatif yang tinggi. [8][12][13]. Sebaliknya, temuan ini sejalan dengan penelitian yang menemukan bahwa penerapan PCA dapat menurunkan performa Random Forest, karena Random Forest pada dasarnya telah memiliki mekanisme seleksi fitur internal melalui pemilihan atribut secara acak pada setiap pemisahan pohon [14].

Perbedaan hasil dapat dijelaskan oleh jumlah elemen utama yang dipertahankan, serta besarnya varians kumulatif yang dipertahankan. Penelitian yang menunjukkan peningkatan kinerja biasanya mempertahankan varians kumulatif yang jauh lebih besar daripada penelitian ini. Varians kumulatif sebesar 26% dalam penelitian ini menyiratkan bahwa sekitar 74% informasi dari 35 atribut asli tidak diperhitungkan oleh dua komponen utama yang dipertimbangkan, yang menunjukkan bahwa informasi yang hilang kemungkinan besar berisi atribut yang relevan dengan target klasifikasi, tetapi dengan kontribusi yang relatif kecil terhadap varians data secara keseluruhan. Hal ini sejalan dengan fakta bahwa PCA adalah teknik tanpa pengawasan (unsupervised) di mana pemilihan komponen utama ditentukan oleh besarnya varians dalam data, bukan oleh relevansi fitur terhadap label target [5].

Penurunan yang signifikan dalam jumlah recall pada model PCA (dari 0,754 menjadi 0,650) memiliki konsekuensi yang signifikan secara praktis untuk mengidentifikasi dropout dini mahasiswa. Jumlah mahasiswa berisiko dropout yang berhasil terdeteksi berkurang karena jumlah recall yang lebih rendah. Akibatnya, lebih banyak mahasiswa berisiko yang tidak terdeteksi dan tidak menerima intervensi yang diperlukan. Sistem deteksi dini memiliki tujuan utama untuk mengurangi jumlah kasus berisiko yang terlewatkan, oleh karena itu kondisi ini lebih merugikan daripada penurunan pada metrik lain.

Penelitian lain yang menggunakan dataset serupa dari Kaggle melaporkan performa Random Forest dengan Accuracy sebesar 70,16%, yang lebih rendah dari performa model tanpa PCA pada penelitian ini (Accuracy 0,856) [15]. Dalam penelitian ini, kinerja model tanpa PCA tetap berada dalam rentang yang kompetitif dengan penelitian sejenis yang menggunakan dataset dan algoritma yang sama, meskipun perbedaan dalam kinerja dapat disebabkan oleh perbedaan metodologi atau hyperparameter tuning.

Secara keseluruhan, hasil penelitian menunjukkan bahwa penerapan Analisis Komponen Utama (PCA) dengan jumlah komponen yang terbatas tidak meningkatkan kinerja algoritma Random Forest dalam memprediksi siswa mana saja yang berisiko putus sekolah dalam kumpulan data yang digunakan. Sebaliknya, penggunaan PCA justru mengurangi kemampuan model dalam mengidentifikasi siswa yang berisiko putus sekolah.

KESIMPULAN

Berdasarkan hasil evaluasi yang telah dipaparkan, dapat ditarik kesimpulan sebagai berikut:

Pengaruh PCA terhadap Performa Random Forest: Penerapan Principal Component Analysis dengan dua komponen utama (explained variance 26%) tidak menunjukkan adanya dampak peningkatan kinerja pada model Random Forest dalam hal deteksi dini risiko putus sekolah siswa. Sebaliknya, model PCA justru menunjukkan penurunan yang konsisten pada semua metrik evaluasi.

Random Forest tanpa PCA memiliki performa yang lebih baik. Dengan AUC 0,905, Accuracy 0,856, F1 0,771, Precision 0,790, Recall 0,754, dan MCC 0,667, Random Forest dengan PCA memiliki AUC 0,834, Accuracy 0,800, F1 0,676, Precision 0,704, recall 0,650, dan MCC 0,533. MCC (0,134 poin) dan Recall (0,104) mengalami penurunan terbesar dalam pengamatan.

Implikasi bagi Deteksi Dini: Terjadi penurunan yang signifikan pada nilai recall dalam model PCA (dari 0,754 menjadi 0,650), yang memiliki implikasi praktis yang berarti. Nilai recall yang lebih rendah berarti ada lebih banyak siswa yang berisiko putus sekolah yang tidak terdeteksi oleh model tersebut, sehingga mereka tidak menerima intervensi yang diperlukan. Ini merupakan situasi yang tidak diinginkan dalam sistem peringatan dini.

Kelebihan Penelitian: Penggunaan metrik evaluasi yang komprehensif (AUC, Accuracy, F1, Precision, Recall, dan MCC) memberikan gambaran performa model yang lengkap dan kredibel. Rancangan eksperimen terkontrol di mana dua alur kerja dijalankan pada kumpulan data yang identik memastikan bahwa perbedaan kinerja semata-mata disebabkan oleh implementasi PCA. Hasil penelitian sejalan dengan teori bahwa Random Forest memiliki mekanisme seleksi fitur internal yang membuat pengurangan aspek eksternal tidak menguntungkan.

Keterbatasan penelitian ini: Pertama, hasil ini mungkin tidak dapat digeneralisasikan ke jumlah komponen yang lebih besar karena pemilihan jumlah komponen PCA (dua komponen dengan varians yang dijelaskan sebesar 26%) tergolong sangat agresif dalam hal pengurangan dimensi. Selanjutnya, kedua model tersebut tidak menangani ketidakseimbangan kelas, yang dapat memengaruhi akurasi klasifikasi, terutama pada kelas terdaftar. Ketiga, parameter Random Forest menggunakan nilai default dari Orange3 tanpa mengoptimalkan hiperparameter, sehingga pengaturan parameter tersebut berpotensi meningkatkan kinerja kedua model.

Rekomendasi Praktis: Random Forest tanpa PCA lebih direkomendasikan karena memiliki kinerja recall yang lebih baik. Namun, jika dimensi data atau efisiensi komputasi menjadi faktor utama, perlu dilakukan penelitian lebih lanjut dengan jumlah komponen PCA yang lebih besar (varians yang dijelaskan 80–90%) untuk menentukan keseimbangan terbaik antara pengurangan dimensi dan kinerja model.

DAFTAR PUSTAKA

- [1] R. Quincho Apumayta, J. Carrillo Cayllahua, A. Ccencho Pari, V. Inga Choque, J. C. Cárdenas Valverde, and D. Huamán Ataypoma, "University Dropout: A Systematic Review of the Main Determinant Factors (2020-2024)," *F1000Research*, vol. 13, pp. 1–20, 2024, doi: 10.12688/f1000research.154263.2.
- [2] A. Algiffary and T. Sutabri, "Indonesian Journal of Computer Science," *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023, [Online]. Available: <http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135>
- [3] Y. Lin, H. Chen, W. Xia, F. Lin, Z. Wang, and Y. Liu, "A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining," *Data Sci. Eng.*, vol. 10, no. 4, pp. 564–590, 2025, doi: 10.1007/s41019-025-00303-z.
- [4] S. G. A. Utami, H. Setiadi, and A. Rohmadi, "Comparative Analysis of Machine Learning Algorithms with RFE-CV for Student Dropout Prediction," *J. Tek. Inform.*, vol. 6, no. 3, pp. 1319–1338, 2025, doi: 10.52436/1.jutif.2025.6.3.4695.

- [5]O. Rute *et al.*, “Jurnal EurekaMatika,” vol. 13, no. 1, pp. 67–80, 2025.
- [6]A. M. Fajria, A. Faqih, and G. Dwilestari, “The Impact of Principal Component Analysis Dimensionality Reduction on Sentiment Classification Performance Using Support Vector Machine,” *J. Artif. Intell. Eng. Appl.*, vol. 4, no. 2, pp. 764–770, 2025, doi: 10.59934/jaiea.v4i2.744.
- [7]Ridha Afifa, Muhammad Itqan Mazdadi, Triando Hamongan Saragih, Fatma Indriani, and Muliadi, “4015-10989-1-Pb[1],” vol. 13, pp. 1852–1864, 2024.
- [8]E. A. F. Elmuna, T. Chamidy, and F. Nugroho, “Optimization of the Random Forest Method Using Principal Component Analysis to Predict House Prices,” *Int. J. Adv. Data Inf. Syst.*, vol. 4, no. 2, pp. 155–166, 2023, doi: 10.25008/ijadis.v4i2.1290.
- [9]Z. Dobesova, “Evaluation of Orange data mining software and examples for lecturing machine learning tasks in geoinformatics,” *Comput. Appl. Eng. Educ.*, vol. 32, no. 4, pp. 1–18, 2024, doi: 10.1002/cae.22735.
- [10] D. Chicco, N. Tötsch, and G. Jurman, “The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation,” *BioData Min.*, vol. 14, pp. 1–22, 2021, doi: 10.1186/s13040-021-00244-z.
- [11] R. Riaz, M. A. Chatni, T. Rehman, M. Bin Tahir, and S. Khan, “Using Orange Data Mining and Predictive Analytics for Analytical Workflows in Academic Research,” vol. 7, no. 3, pp. 23–30, 2026.
- [12] D. Priyanto, H. Hairani, K. Marzuki, and M. Innuddin, “Optimization of Random Forest for Health Data Classification Using PCA and K-Means SMOTE-ENN,” *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, pp. 27646–27652, 2025, doi: 10.48084/etasr.12976.
- [13] Z. Abidin, T. Suratno, and M. F. Putri, “Penerapan Random Oversampling dan Principal Component Analysis untuk Meningkatkan Akurasi Prediksi Kebangkrutan Perusahaan di Indonesia dengan Model Machine Learning,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 12, no. 5, pp. 1209–1220, 2025, [Online]. Available: <https://jtiik.ub.ac.id/index.php/jtiik/article/view/9972>
- [14] L. F. Voges, L. C. Jarren, and S. Seifert, “Exploitation of surrogate variables in random forests for unbiased analysis of mutual impact and importance of features,” *Bioinformatics*, vol. 39, no. 8, 2023, doi: 10.1093/bioinformatics/btad471.
- [15] L. G. R. Putra, D. D. Prasetya, and M. Mayadi, “Student Dropout Prediction Using Random Forest and XGBoost Method,” *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 9, no. 1, pp. 147–157, 2025, doi: 10.29407/intensif.v9i1.21191.