



Implementasi Random Forest dan SMOTE untuk Prediksi Risiko Putus Sekolah Dasar Menuju Indonesia Emas 2045

Muhammad Alfathan Harriz

Universitas Matana, Indonesia

Alamat: Kabupaten Tangerang, Indonesia

Korespondensi penulis: harrizsb@gmail.com

Abstract. *This research investigates the implementation of Random Forest algorithms combined with Synthetic Minority Over-sampling Technique (SMOTE) to predict elementary school dropout rates in Indonesia, supporting the Indonesia Emas 2045 vision. A significant gap was identified in previous studies, which, despite utilizing artificial intelligence for dropout interventions, had not integrated temporal dimensions into data analysis. A temporal data-based classification model was developed using Indonesian Ministry of Education data from 2021-2023, incorporating lag features, delta calculations, and rolling statistics. Two models were implemented: one with SMOTE achieving 99% accuracy with perfect recall for high-risk regions, while the non-SMOTE model reached 100% accuracy. Temporal features were identified as crucial predictors, reflecting external fluctuations and annual changes impacting dropout decisions. This approach enables educational institutions to allocate resources more efficiently by prioritizing operational assistance for high-risk schools. The model's capacity to identify high-risk regions with 100% recall represents a strategic investment in strengthening Indonesia's human resource sustainability. To address the limitations of provincial aggregate data, expansion to include individual-level variables and model validation at district or school scales is recommended for future research.*

Keywords: *Elementary School, Random Forest, SMOTE, Temporal Data*

Abstrak. Dalam penelitian ini, penerapan algoritma Random Forest dengan Synthetic Minority Over-sampling Technique (SMOTE) diinvestigasi untuk memprediksi tingkat putus sekolah dasar di Indonesia, mendukung visi Indonesia Emas 2045. Dimensi temporal yang belum diintegrasikan dalam penelitian sebelumnya ditemukan sebagai kesenjangan signifikan. Model klasifikasi berbasis data temporal dikembangkan menggunakan data Kemendikdasmen periode 2021-2023, dengan fitur lag, perhitungan delta, dan statistik rolling diinkorporasikan. Dua model diimplementasikan: model dengan SMOTE mencapai akurasi 99% dengan recall sempurna untuk wilayah berisiko tinggi, sementara model tanpa SMOTE mencapai akurasi 100%. Fitur temporal diidentifikasi sebagai prediktor krusial yang mencerminkan fluktuasi eksternal dan perubahan tahunan. Melalui pendekatan ini, alokasi sumber daya pendidikan dapat diprioritaskan secara efisien untuk sekolah berisiko tinggi. Kemampuan model mengidentifikasi wilayah berisiko tinggi dengan recall 100% dipandang sebagai investasi strategis untuk keberlanjutan sumber daya manusia Indonesia. Perluasan data tingkat individu dan validasi model dalam skala yang lebih kecil direkomendasikan untuk mengatasi keterbatasan data agregat provinsi.

Kata kunci: Pendidikan Dasar, *Random Forest*, SMOTE, *Temporal Data*

1. LATAR BELAKANG

Indonesia Emas 2045 merupakan visi strategis untuk mentransformasi Indonesia menjadi negara maju pada peringatan 100 tahun kemerdekaan. Realisasi visi ini membutuhkan sinergi seluruh komponen bangsa, termasuk generasi muda sebagai penerus pembangunan (Soleh dkk., 2024). Pada 2045, Indonesia diproyeksikan memiliki jumlah penduduk usia produktif yang besar yang merupakan sebuah peluang demografis yang juga berpotensi menjadi tantangan jika tidak dikelola secara optimal (Nugraha & Amiludin, 2024). Faktor kunci keberhasilan agenda ini terletak pada kualitas sumber daya manusia (Puspa dkk., 2023). Pendidikan dasar menjadi fondasi utama pembentukan

sumber daya manusia (SDM) unggul, namun tingginya angka putus sekolah dasar di berbagai wilayah masih menjadi masalah serius (Alifa, 2023; Fadila dkk., 2024; Marta dkk., 2023; S dkk., 2024; Widiyastuti dkk., 2023; Widyastuti, 2021). Fenomena ini tidak hanya berdampak sosial, tetapi juga ekonomi: setiap kenaikan 1% angka putus sekolah diperkirakan mengurangi Rp12,4 triliun dari perekonomian nasional per tahun (Ramadhanti, 2021).

Kondisi ini mempertegas urgensi pengembangan sistem deteksi dini dan pencegahan dropout berbasis inovasi teknologi. Beberapa studi telah mengeksplorasi pemanfaatan kecerdasan buatan untuk intervensi masalah ini (Fitriana dkk., 2024; Jiménez dkk., 2023; Kumar dkk., 2024; Paput dkk., 2023; Purwanto dkk., 2025; Rofi dkk., 2024; Surip dkk., 2021), namun belum ada yang mengintegrasikan dimensi temporal dalam analisis data. Padahal, integrasi dimensi temporal, seperti yang diungkapkan Cahyani, (2024) dapat mengidentifikasi pola dinamika risiko dropout secara lebih akurat. Penelitian ini mengusulkan model klasifikasi berbasis data temporal menggunakan algoritma Random Forest yang dikombinasikan dengan SMOTE. Pendekatan ini memungkinkan lembaga pendidikan, khususnya di tingkat pusat, mengalokasikan sumber daya secara lebih efisien, misalnya dengan memprioritaskan bantuan operasional bagi sekolah berisiko tinggi. Dengan demikian, implementasi Random Forest tidak hanya menjadi solusi teknis, tetapi juga investasi strategis dalam memperkuat keberlanjutan SDM Indonesia.

2. KAJIAN TEORITIS

Random Forest (RF)

RF merupakan algoritma ensemble learning yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi dan meminimalkan *overfitting* (Dhani dkk., 2024; Nugroho & Rilvani, 2023; Oktaviani dkk., 2024). Keunggulannya mencakup kemampuan menangani data berdimensi tinggi, mengakomodasi hubungan non-linear, serta tahan terhadap *noise* (Ghozali dkk., 2023; Ilham dkk., 2024). Pada data temporal, RF mampu mengidentifikasi pola dinamis dengan memanfaatkan variabel waktu seperti *lag* atau *window* (Sun dkk., 2024).

Data Temporal

Data *temporal* memiliki karakteristik ketergantungan waktu, seperti tren, musiman, atau perubahan struktural (Oktaviani dkk., 2024; Sun dkk., 2024). Proses prapemrosesan data temporal meliputi ekstraksi variabel waktu, seperti timestamp atau deret waktu, serta transformasi data agar sesuai dengan model prediktif berbasis data statis (Sun dkk., 2024).

SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) mengatasi masalah ketidakseimbangan kelas dengan menghasilkan sampel sintetis dari kelas minoritas melalui interpolasi linier antar data (Dhani dkk., 2024; Hussain dkk., 2022; Nugroho & Rilvani, 2023).

Integrasi SMOTE dengan RF terbukti meningkatkan kinerja klasifikasi pada data tidak seimbang, misalnya dalam prediksi kebangkrutan (Nugroho & Rilvani, 2023), *stunting* (Dhani dkk., 2024), dan analisis sentimen (Ilham dkk., 2024).

Imputasi

Imputasi merupakan teknik untuk mengisi nilai yang hilang (*missing values*) pada *dataset*, sehingga proses analisis dan pemodelan data dapat berjalan secara optimal. Metode imputasi dapat berupa pendekatan sederhana, seperti menggantikan nilai hilang dengan rata-rata (*mean*), *median*, atau *modus*, maupun metode yang lebih kompleks, seperti *k-Nearest Neighbors* (KNN), *Multivariate Imputation by Chained Equations* (MICE), dan regresi *kernel smoothing* (Prasetya dkk., 2023).

Encoding

Encoding adalah proses mengubah data kategorikal menjadi format numerik agar dapat digunakan dalam algoritma pembelajaran mesin (Herdian dkk., 2024). Dua teknik encoding yang digunakan pada penelitian ini adalah label encoding dan one-hot encoding.

3. METODE PENELITIAN

Penelitian ini mengimplementasikan RF dengan SMOTE pada data temporal putus sekolah jenjang pendidikan dasar. Berikut kerangka metodologinya:

Desain Penelitian

Penelitian menggunakan desain eksperimen simulasi dengan pendekatan kuantitatif. Variabel dependen berupa kategori jumlah putus sekolah (rendah: <500, sedang: 500-1500, tinggi: >1500), sedangkan variabel independen meliputi fitur *temporal* (Jumlah_Lag1, Delta_Jumlah, Rolling_Mean_3, Rolling_Std_3) dan variabel geografis (provinsi).

Pengumpulan Data

Data diambil dari repositori Kementerian Pendidikan Dasar dan Menengah (Kemendikdasmen) periode 2021–2023, mencakup deret waktu jumlah putus sekolah per provinsi. Format data mentah (*raw data*) dalam bentuk *file Comma Separated Values* (CSV) diunduh langsung dari sumber resmi.

Preprocessing Data

a. Transformasi temporal

Beberapa teknik transformasi temporal diterapkan untuk menghasilkan informasi yang lebih mendalam dan mendukung pemahaman dinamika data. Pertama, Fitur lag digunakan untuk membandingkan jumlah putus sekolah pada tahun berjalan dengan tahun sebelumnya, yang disebut sebagai Jumlah_Lag1. Fitur ini memungkinkan identifikasi pola perubahan tahunan. Kedua, Fitur delta dihitung sebagai selisih antara jumlah putus sekolah pada tahun berjalan dan tahun sebelumnya (Delta_Jumlah). Fitur ini menggambarkan besarnya perubahan antarperiode. Ketiga, Rolling statistics diterapkan untuk menghitung rata-rata (Rolling_Mean_3) dan deviasi standar (Rolling_Std_3) dari jumlah putus sekolah selama tiga tahun terakhir. Pendekatan ini memungkinkan penangkapan tren jangka pendek dan variabilitas data dalam periode waktu tersebut.

b. Imputasi

Untuk menjaga kualitas data dalam analisis, nilai yang hilang pada kolom Rolling_Std_3 diimputasi dengan nilai konstanta kecil, yaitu $1e-5$. Pendekatan ini diterapkan untuk mencegah terjadinya pembagian dengan nol, sehingga tidak terjadi kesalahan saat pemodelan. Kolom numerik lainnya yang mengandung nilai kosong diimputasi dengan nilai 0. Untuk memastikan transparansi, kolom biner tambahan (is_imputed) dibuat sebagai penanda untuk mengidentifikasi data yang telah dimodifikasi. Strategi ini mendukung integritas data dan memungkinkan pelacakan data hasil imputasi.

c. Encoding

Variabel kategorikal Status dan Kategori_Jumlah diubah menjadi representasi numerik melalui teknik label encoding. Pendekatan ini dipilih karena sesuai untuk variabel dengan sifat ordinal atau jumlah kategori yang terbatas, sehingga memudahkan pemrosesan data dalam analisis.

Variabel Provinsi diubah menjadi representasi biner menggunakan teknik one-hot encoding. Teknik ini diterapkan untuk menghindari asumsi hubungan ordinal antar provinsi, sehingga mencegah bias dalam model analisis dan memastikan interpretasi data yang lebih akurat.

d. Analisis Data

Untuk mengatasi ketidakseimbangan kelas dalam data, teknik SMOTE diterapkan. Teknik ini menghasilkan sampel sintetis untuk kelas minoritas, sehingga

distribusi kelas menjadi lebih seimbang dan mendukung performa model yang lebih robust dalam analisis prediktif.

Pemodelan menggunakan algoritma RF dengan optimasi hiperparameter untuk meningkatkan akurasi dan generalisasi. Hyperparameter seperti jumlah pohon (*n_estimators*), kedalaman maksimum (*max_depth*), jumlah minimum sampel untuk pemisahan (*min_samples_split*), dan jumlah minimum sampel pada daun (*min_samples_leaf*) dioptimalkan menggunakan GridSearchCV. Proses validasi dilakukan dengan TimeSeriesSplit sebanyak tiga *fold* untuk memastikan evaluasi model yang sesuai dengan sifat temporal data, sehingga menghasilkan model yang lebih andal untuk prediksi berbasis waktu.

e. Evaluasi Model

Kinerja model dievaluasi menggunakan beberapa metrik performa, yaitu *accuracy*, *precision*, *recall*, dan *F1-score* berbobot. Metrik-metrik ini dipilih untuk memberikan gambaran menyeluruh tentang kemampuan model dalam menangani klasifikasi, terutama pada data dengan distribusi kelas yang tidak seimbang, sehingga memastikan penilaian yang komprehensif terhadap performa prediktif.

Untuk memverifikasi signifikansi peningkatan performa model setelah penerapan SMOTE, analisis confusion matrix dilakukan. Pendekatan ini memungkinkan evaluasi detail terhadap distribusi prediksi yang benar dan salah, sehingga memperjelas dampak SMOTE dalam meningkatkan kemampuan model untuk mengenali kelas minoritas.

Pentingnya fitur (*feature importance*) dianalisis untuk mengidentifikasi sepuluh fitur yang paling berpengaruh dalam proses prediksi. Evaluasi ini memberikan wawasan tentang variabel-variabel kunci yang mendorong performa model, mendukung interpretasi hasil, dan memandu pengambilan keputusan berbasis data dalam konteks penelitian.

4. HASIL DAN PEMBAHASAN

Bagian ini menganalisis efektivitas model RF untuk memprediksi tingkat putus sekolah pada jenjang pendidikan dasar di Indonesia, menggunakan data dari Kemendikdasmen periode 2021–2023. Data mencakup deret waktu jumlah putus sekolah per provinsi, yang diklasifikasikan ke dalam tiga kategori: rendah (<500), sedang (500–1500), dan tinggi (>1500). Dua model diuji: satu dengan SMOTE dan satu tanpa SMOTE,

dengan fitur *temporal* (Jumlah, Jumlah_Lag1, Delta_Jumlah, Rolling_Mean_3, Rolling_Std_3) dan geografis (provinsi).

Kinerja Model

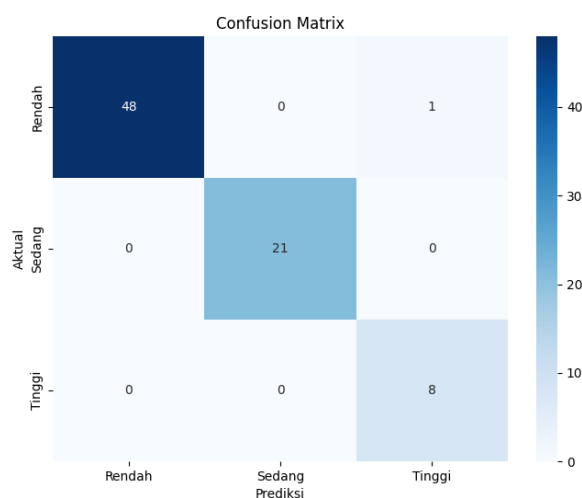
a) Model dengan SMOTE

Model RF dengan SMOTE, dioptimalkan menggunakan GridSearchCV dengan TimeSeriesSplit, mencapai akurasi keseluruhan 99% pada data uji (tahun 2023). Parameter terbaik adalah $\text{max_depth}=5$, $\text{min_samples_leaf}=2$, $\text{min_samples_split}=5$, dan $\text{n_estimators}=50$, dengan pengaturan $\text{class_weight}=\text{'balanced'}$ dan $\text{max_features}=\text{'sqrt'}$. Metrik kinerja, termasuk *presisi*, *recall*, dan *F1-score*, disajikan pada Tabel 1.

Tabel 1. Laporan Klasifikasi Model RF dengan SMOTE

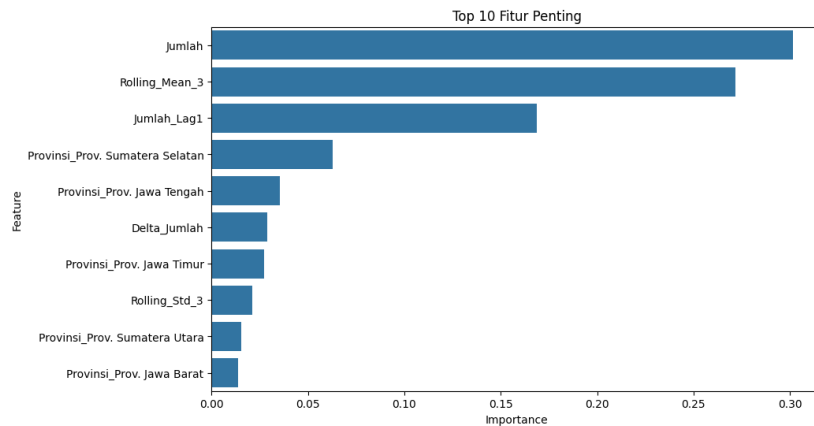
Kategori	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>	<i>Support</i>
Rendah	1.00	0.98	0.99	49
Sedang	1.00	1.00	1.00	21
Tinggi	0.89	1.00	0.94	8
Akurasi				0.99
<i>Macro Avg.</i>	0.96	0.99	0.98	78
<i>Weighted Avg.</i>	0.99	0.99	0.99	78

Confusion matrix yang disajikan pada Gambar 1 menunjukkan bahwa model mengklasifikasikan semua instance kategori "Tinggi" dengan benar (*recall* 1.00), meskipun presisi 0.89 menunjukkan adanya beberapa *false positive*. Untuk kategori "Rendah" dan "Sedang", kinerja hampir sempurna, dengan hanya satu instance "Rendah" yang salah diklasifikasikan.



Gambar 1. Confusion Matrix dengan SMOTE

Feature importance pada Gambar 2 dengan SMOTE menunjukkan bahwa data temporal memiliki andil besar dalam menentukan performa akhir dari model RF. Namun, *feature* Rolling_Std_3 tidak begitu signifikan dibandingkan dengan 2 provinsi yaitu Sumatera Selatan dan Jawa Tengah.



Gambar 2. *Feature Importance* dengan SMOTE

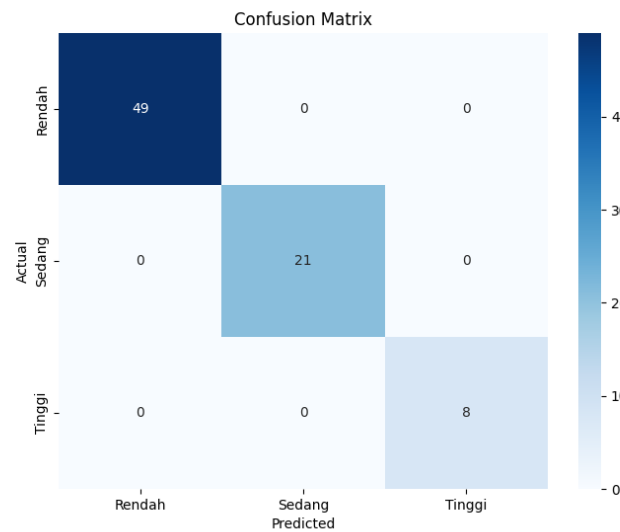
b) Model tanpa SMOTE

Model RF tanpa SMOTE mencapai akurasi 100% pada data uji yang sama, dengan parameter terbaik $\text{max_depth}=20$, $\text{min_samples_leaf}=1$, $\text{min_samples_split}=5$, dan $\text{n_estimators}=200$, hanya menggunakan $\text{random_state}=42$. Laporan klasifikasi disajikan pada Tabel 2.

Tabel 2. Laporan Klasifikasi Model RF tanpa SMOTE

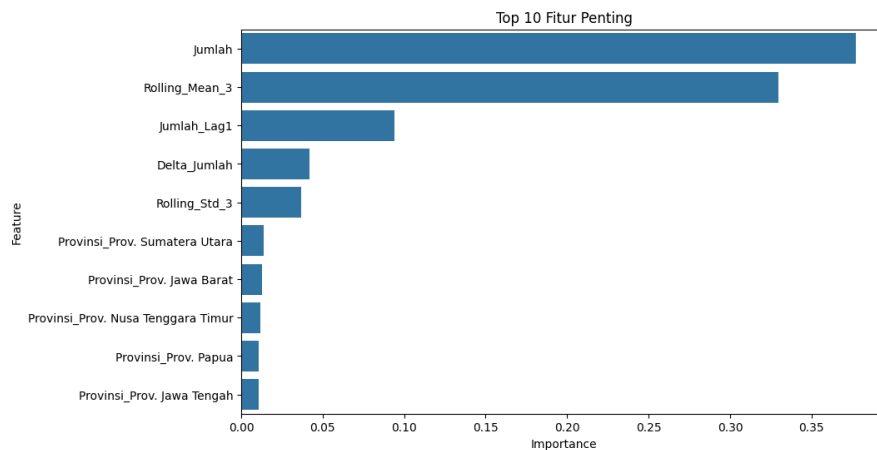
Kategori	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>	<i>Support</i>
Rendah	1.00	1.00	1.00	49
Sedang	1.00	1.00	1.00	21
Tinggi	1.00	1.00	1.00	8
Akurasi			1.00	78
<i>Macro Avg.</i>	1.00	1.00	1.00	78
<i>Weighted Avg.</i>	1.00	1.00	1.00	78

Confusion matrix pada Gambar 2 menunjukkan klasifikasi sempurna untuk semua kategori, tanpa kesalahan prediksi.



Gambar 3. *Confusion Matrix* tanpa SMOTE

Gambar 4 menunjukkan bahwa data temporal sangat menunjukkan bahwa model RF tanpa SMOTE memiliki andil besar dalam menentukan performa akhir dari model RF.



Gambar 4. *Feature Importance* dengan SMOTE

Penelitian ini menunjukkan bahwa model RF, baik dengan maupun tanpa SMOTE, mampu memprediksi tingkat putus sekolah dengan akurasi sangat tinggi. Model dengan SMOTE mencapai akurasi 99%, dengan recall 100% untuk kategori "Tinggi", menjadikannya alat yang sangat efektif untuk mengidentifikasi wilayah berisiko tinggi. Model tanpa SMOTE mencapai akurasi 100%, menunjukkan bahwa

data memiliki pola yang jelas, tetapi SMOTE meningkatkan *robustness* dengan menangani ketidakseimbangan kelas, yang penting untuk generalisasi pada data baru.

Akurasi 99% dari model dengan SMOTE sejalan dengan temuan literatur, di mana RF dilaporkan mencapai akurasi hingga 99% dalam prediksi putus sekolah (Andrade-Girón dkk., 2023). Recall 100% untuk kategori "Tinggi" sangat krusial, karena mendeteksi wilayah berisiko tinggi memungkinkan intervensi dini, seperti alokasi bantuan operasional sekolah atau program dukungan siswa. Penurunan presisi ke 0.89 untuk kategori "Tinggi" menunjukkan adanya beberapa *false positive*, tetapi dalam konteks pencegahan putus sekolah, ini lebih baik daripada melewatkan kasus berisiko tinggi, sebagaimana dicatat dalam studi tentang data tidak seimbang (Villar & de Andrade, 2024).

Model tanpa SMOTE mencapai akurasi 100%, yang mungkin menunjukkan *overfitting*, terutama dengan hanya 8 *instance* dalam kategori "Tinggi". Namun, karena data uji berasal dari periode waktu yang berbeda (2023), kinerja ini menunjukkan bahwa pola putus sekolah cukup konsisten antar tahun. Penggunaan pohon yang lebih dalam ($\text{max_depth}=20$) tanpa SMOTE memungkinkan model menangkap pola kompleks, tetapi tanpa penanganan ketidakseimbangan kelas, model ini mungkin kurang andal pada dataset yang lebih beragam. SMOTE, dengan pohon yang lebih dangkal ($\text{max_depth}=5$), menghasilkan model yang lebih sederhana namun tetap sangat akurat, menunjukkan bahwa sampel sintetis membantu memperjelas batas keputusan.

5. KESIMPULAN DAN SARAN

Efektivitas model Random Forest dengan SMOTE dalam memprediksi tingkat putus sekolah pendidikan dasar di Indonesia, mencapai akurasi 99% dan *recall* 100% untuk wilayah berisiko tinggi. Fitur temporal seperti *Rolling_Std_3* dan *Delta_Jumlah* berperan krusial sebagai prediktor utama, merefleksikan fluktuasi eksternal dan perubahan tahunan yang berdampak pada keputusan putus sekolah. Temuan ini selaras dengan visi Indonesia Emas 2045 yang menekankan pembangunan SDM unggul melalui pengurangan angka putus sekolah, mengingat kerugian ekonomi mencapai Rp12,4 triliun per 1% kenaikan. Meskipun model tanpa SMOTE menunjukkan akurasi 100%, implementasi SMOTE meningkatkan *robustness* dalam menangani ketidakseimbangan kelas, khususnya di wilayah dengan risiko tinggi seperti Jawa Barat dan Sumatera Utara yang memiliki kompleksitas geografis dan ekonomi.

Untuk memaksimalkan dampak strategis, perluasan cakupan data mencakup variabel tingkat individu seperti status ekonomi keluarga dan kinerja akademik, yang secara literatur terbukti memengaruhi keputusan putus sekolah (Banaag dkk., 2024; Nurmalitasari dkk., 2023). Validasi model dalam skala kabupaten atau sekolah akan mengatasi keterbatasan data agregat provinsi sekaligus memitigasi risiko *overfitting* akibat ukuran *dataset* terbatas. Eksperimen dengan algoritma LightGBM yang menunjukkan *F1-score* 0,84 dalam studi komparatif (Cho dkk., 2023) layak dipertimbangkan untuk meningkatkan generalisasi model. Implementasi sistem *monitoring real-time* berbasis temuan ini dapat menjadi landasan kebijakan pendidikan berbasis data, mendukung optimalisasi bonus demografi menuju 2045 melalui intervensi tepat sasaran di sektor pendidikan. Kolaborasi lintas sektor antara pemerintah, akademisi, dan pemangku kepentingan lokal menjadi kunci transformasi temuan akademis menjadi aksi nyata dalam mendukung Indonesia Emas 2045.

DAFTAR REFERENSI

- Alifa, V. N. (2023). Analisis faktor penyebab meningkatnya angka putus sekolah di Indonesia pada tahun 2022. *Jurnal Pendidikan Sultan Agung*, 3(2), Article 2. <https://doi.org/10.30659/jp-sa.3.2.175-182>
- Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodríguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., Villarreal-Torres, H., & Angeles-Morales, J. (2023). Predicting student dropout based on machine learning and deep learning: A systematic review. *EAI Endorsed Transactions on Scalable Information Systems*, 10(5), Article 5. <https://doi.org/10.4108/eetsis.3586>
- Banaag, R., Sumodevilla, J. L., & Potane, J. (2024). Factors affecting student drop out behavior: A systematic review. *International Journal of Educational Management and Innovation*, 5(1), Article 1. <https://doi.org/10.12928/ijemi.v5i1.9396>
- Cahyani, N. L. P. A. (2024). Machine learning approaches for customer churn prediction in the aquaculture technology sector. *International Journal of Current Science Research and Review*, 7(8). <https://doi.org/10.47191/ijcsrr/V7-i8-74>
- Cho, C. H., Yu, Y. W., & Kim, H. G. (2023). A study on dropout prediction for university students using machine learning. *Applied Sciences*, 13(21), Article 21. <https://doi.org/10.3390/app132112004>
- Dhani, A. A., Siswa, T. A. Y., & Pranoto, W. J. (2024). Perbaikan akurasi Random Forest dengan ANOVA dan SMOTE pada klasifikasi data stunting. *Teknika*, 13(2), Article 2. <https://doi.org/10.34148/teknika.v13i2.875>
- Fadila, A., Syafriandi, S., Kurniawati, Y., & Salma, A. (2024). Classification of dropout rates in West Sumatra using the Random Forest algorithm with Synthetic Minority

Oversampling Technique. *UNP Journal of Statistics and Data Science*, 2(3), Article 3. <https://doi.org/10.24036/ujsds/vol2-iss3/183>

- Fitriana, S., Riniyanty, L., Laila, R., Pratama, S. A., & Lamasitudju, C. A. (2024). Prediksi siswa putus sekolah dan keberhasilan akademik menggunakan machine learning: Prediksi siswa putus sekolah dan keberhasilan akademik. *The Indonesian Journal of Computer Science*, 13(6), Article 6. <https://doi.org/10.33022/ijcs.v13i6.4453>
- Ghozali, A., Pratiwi, H., & Handajani, S. S. (2023). Implementasi data mining menggunakan metode Random Forest dan Support Vector Machine dalam klasifikasi penyakit diabetes. *Delta: Jurnal Ilmiah Pendidikan Matematika*, 11(2), 147. <https://doi.org/10.31941/delta.v11i2.2686>
- Herdian, C., Kamila, A., & Agung Musa Budidarma, I. G. (2024). Studi kasus feature engineering untuk data teks: Perbandingan label encoding dan one-hot encoding pada metode linear regresi. *Technologia: Jurnal Ilmiah*, 15(1), 93. <https://doi.org/10.31602/tji.v15i1.13457>
- Hussain, L., Lone, K. J., Awan, I. A., Abbasi, A. A., & Pirzada, J.-R. (2022). Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. *Waves in Random and Complex Media*, 32(3), 1079–1102. <https://doi.org/10.1080/17455030.2020.1810364>
- Ilham, M. F. N., Annurrahma, K. D., Wirayuda, P., & Rudiman, R. (2024). Analisis kepuasan pengguna aplikasi Donorku dengan pendekatan metode Random Forest dengan SMOTE. *Jurnal Informatika Teknologi dan Sains (Jinteks)*, 6(3), 508–513. <https://doi.org/10.51401/jinteks.v6i3.4229>
- Jiménez, O., Jesús, A., & Wong, L. (2023). Model for the prediction of dropout in higher education in Peru applying machine learning algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine. In *2023 33rd Conference of Open Innovations Association (FRUCT)* (pp. 116–124). <https://doi.org/10.23919/FRUCT58615.2023.10143068>
- Kumar, D., Kothiyal, A., Kumar, R., Hemantha, C., & Maranan, R. (2024). Random Forest approach optimized by the Grid Search process for predicting the dropout students. In *2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET)* (pp. 1–6). <https://doi.org/10.1109/ICICET59348.2024.10616372>
- Marta, J. K., Nugraha, A. E., & Anggorowati, K. D. (2023). Analisis penyebab anak putus sekolah pada jenjang pendidikan dasar di Desa Engkurai. *Jurnal Pendidikan dan Pembelajaran Sekolah Dasar*, 1(3), Article 3. <https://doi.org/10.46368/jppsd.v1i3.1398>
- Nugraha, S. A. S., & Amiludin, A. (2024). Inovasi metode pembelajaran sekolah dasar berbasis connection sebagai pengembangan karakter social entrepreneurship dalam mewujudkan Indonesia emas 2045. *Al-Mubtadi: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, 1(2), 92–106. <https://doi.org/10.58988/almubtadi.v1i2.282>

- Nugroho, A., & Rilvani, E. (2023). Penerapan metode oversampling SMOTE pada algoritma Random Forest untuk prediksi kebangkrutan perusahaan. *Techno.Com*, 22(1), Article 1. <https://doi.org/10.33633/tc.v22i1.7527>
- Nurmalitasari, Awang Long, Z., & Faizuddin Mohd Noor, M. (2023). Factors influencing dropout students in higher education. *Education Research International*, 2023(1), 7704142. <https://doi.org/10.1155/2023/7704142>
- Oktaviani, V., Rosmawarni, N., & Muslim, M. P. (2024). Perbandingan kinerja Random Forest dan SMOTE Random Forest dalam mendeteksi dan mengukur tingkat stres pada mahasiswa tingkat akhir. *Informatik: Jurnal Ilmu Komputer*, 20(1), Article 1. <https://doi.org/10.52958/iftk.v20i1.9158>
- Paput, M. J., Suryowati, K., & Jatipaningrum, M. T. (2023). Perbandingan metode Random Forest dan Adaptive Boosting pada klasifikasi indeks pembangunan manusia di Indonesia. *Jurnal Statistika Industri dan Komputasi*, 8(2), 73–83. <https://doi.org/10.34151/statistika.v8i2.4458>
- Prasetya, M. R. A., Priyatno, A. M., & Nurhaeni. (2023). Penanganan imputasi missing values pada data time series dengan menggunakan metode data mining. *Jurnal Informasi dan Teknologi*, 52–62. <https://doi.org/10.37034/jidt.v5i2.324>
- Purwanto, A., Sartono, B., & Notodiputro, K. A. (2025). A comparison of Random Forest and Double Random Forest: Dropout rates of madrasah students in Indonesia. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 19(1), Article 1. <https://doi.org/10.30598/barekengvol19iss1pp227-236>
- Puspa, C. I. S., Rahayu, D. N. O., & Parhan, M. (2023). Transformasi pendidikan abad 21 dalam merealisasikan sumber daya manusia unggul menuju Indonesia emas 2045. *Jurnal Basicedu*, 7(5), 3309–3321. <https://doi.org/10.31004/basicedu.v7i5.5030>
- Ramadhanti, H. D. (2021). Klasifikasi status NEET pada penduduk usia muda di Indonesia dengan SVM dan Random Forest. *Journal of System and Computer Engineering*, 2(1), Article 1. <https://doi.org/10.47650/jsce.v1i2.143>
- Rofi, M. M., Setiawan, F. A., & Riana, F. (2024). Perbandingan metode K-NN dan Random Forest pada klasifikasi mahasiswa berpotensi dropout. *INFOTECH Journal*, 10(1), 84–89. <https://doi.org/10.31949/infotech.v10i1.8856>
- S, G. N., Suryanti, M., & Faridah, H. (2024). Strategi meningkatkan motivasi belajar siswa sekolah dasar sebagai upaya mengatasi putus sekolah. *Jurnal Pengabdian Pendidikan Masyarakat (JPPM)*, 5(1), Article 1. <https://doi.org/10.52060/jppm.v5i1.1500>
- Soleh, N., Fajriah, F., & Rahman, F. (2024). Kontribusi mahasiswa dalam meningkatkan kualitas sumber daya manusia dan mewujudkan visi Indonesia Emas 2045. *Journal of Smart Education and Learning*, 1(1), 22–28. <https://doi.org/10.53088/jsel.v1i1.978>
- Sun, S., Zeng, Z., & Li, Q. (2024). A spatio-temporal evolution analysis framework based on sentiment recognition for temple murals. *Journal of Information Science*, 01655515241293766. <https://doi.org/10.1177/01655515241293766>

- Surip, A., Pratama, M. A., Ali, I., Dikananda, A. R., & Purnamasari, A. I. (2021). Penerapan machine learning menggunakan algoritma C4.5 berbasis PSO dalam menganalisa data siswa putus sekolah. *Informatics for Educators and Professional: Journal of Informatics*, 5(2), 147–155. <https://doi.org/10.51211/itbi.v5i2.1530>
- Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, 4(1), 2. <https://doi.org/10.1007/s44163-023-00079-z>
- Widiasanti, I., Abdul, A. V., Nirwana, A., & Arlita, A. D. (2023). Ancaman melawan putus sekolah dengan dilema kualitas pendidikan Indonesia. *JISIP (Jurnal Ilmu Sosial dan Pendidikan)*, 7(3), Article 3. <https://doi.org/10.58258/jisip.v7i3.5228>
- Widyastuti, N. A. (2021). Analisis tren angka putus sekolah pada pendidikan dasar di Kabupaten Bantul. *Spektrum Analisis Kebijakan Pendidikan*, 10(2), 74–89. <https://doi.org/10.21831/sakp.v10i2.17372>