KLASIFIKASI PRIORITAS INTERVENSI SEKOLAH DASAR DI KABUPATEN MADIUN MENGGUNAKAN ALGORITMA RANDOM FOREST

Divia Prisillia Prisca[1], Kartika Maulida Hindrayani[2], Alfan Rizaldy Pratama[3]
[123]*Program Studi Sains Data, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia*
**Penulis korespondensi: 22083010031@student.upnjatim.ac.id*

ABSTRAK

Kesenjangan dalam kualitas fasilitas dan mutu pembelajaran di sekolah dasar masih menghambat pemerataan layanan pendidikan. Kondisi ini memerlukan sistem pemetaan prioritas intervensi yang lebih objektif, terukur, dan dapat diandalkan sebagai dasar pengambilan keputusan. Penelitian ini bertujuan untuk membangun model klasifikasi tingkat prioritas intervensi sekolah dasar menggunakan algoritma *Random Forest* dan memanfaatkan data dari Dapodik pada tahun 2025. Tahap awal mencakup pembersihan dan transformasi data, yang meliputi mengubah variabel dari kategori menjadi nilai numerik dan membuat fitur baru seperti rasio siswa-guru, rata-rata jumlah siswa per kelas, kekurangan toilet, dan total fasilitas yang tersedia. Penetapan label prioritas didasarkan pada indikator kondisi fisik dan capaian kualitas sekolah. Ketidakseimbangan antar kelas diatasi dengan menerapkan teknik *SMOTE*. Model ini dibuat menggunakan *pipeline* yang mengintegrasikan normalisasi data dengan *MinMaxScaler* dan *Random Forest*. Sementara itu, interpretasi dan pemilihan fitur dilakukan dengan *SHAP* untuk memahami peran setiap fitur dalam prediksi. Hasil penelitian menunjukkan bahwa *Random Forest* dapat mengklasifikasikan tingkat prioritas dengan akurasi 0,90, nilai F1-macro 0,90, dan micro AUC 0,9818. Setelah dilakukan rekalibrasi, kinerjanya mengalami sedikit peningkatan, dengan akurasi 0,91, nilai F1-macro 0,91, dan micro AUC 0,9835. Temuan tersebut memperlihatkan bahwa *Random Forest* mampu memberikan klasifikasi prioritas yang akurat dan dapat mendukung pembuat kebijakan intervensi pendidikan yang lebih efektif dan terarah.

Kata kunci: klasifikasi prioritas, *Random Forest*, *SMOTE*, pendidikan dasar, *machine learning*.

1 PENDAHULUAN

Pendidikan dasar memiliki peranan penting dalam membentuk kualitas sumber daya manusia suatu bangsa. Pada tingkat ini, siswa diperkenalkan tidak hanya pada keterampilan literasi dan numerasi dasar, tetapi juga pada nilai-nilai moral dan keterampilan berpikir kritis sebagai persiapan untuk menghadapi berbagai tantangan global. Namun, distribusi pendidikan dasar berkualitas tetap menjadi tantangan di banyak negara. Menurut laporan *UNESCO*, lebih dari 250 juta anak tidak bisa membaca atau melakukan perhitungan dasar meskipun telah bersekolah, hal ini menunjukkan adanya kesenjangan antara akses dan kualitas pendidikan (UNESCO, 2025). Di Indonesia, kondisi serupa terlihat dalam hasil Asesmen Nasional yang menunjukkan perbedaan pencapaian antarwilayah, yang disebabkan oleh kualitas guru, kondisi infrastruktur, dan perbedaan aspek sosial-ekonomi (Wibowo, 2021). Temuan ini menunjukkan pentingnya pendekatan berbasis data dalam merancang kebijakan pendidikan dasar, terutama untuk mendorong kualitas yang merata.

Kesenjangan dalam kualitas pendidikan juga terlihat jelas di Kabupaten Madiun. Dari 401 SD negeri, hanya delapan sekolah yang mampu memenuhi kuota pendaftaran siswa baru,

sementara 397 sekolah lainnya mengalami kekurangan pendaftaran, termasuk 32 sekolah yang hanya menerima kurang dari lima siswa baru dan satu sekolah yang hanya menerima satu hingga dua siswa selama enam tahun terakhir (Rahayu, 2025). Kesenjangan ini menyebabkan sekolah dengan jumlah siswa yang sangat sedikit menghadapi keterbatasan anggaran, fasilitas belajar yang tidak memadai, dan aktivitas pembelajaran yang kurang optimal, sementara sekolah populer di daerah perkotaan menghadapi kelebihan jumlah siswa, yang mengakibatkan rasio guru-siswa yang tinggi dan berkurangnya kualitas interaksi pembelajaran. Masalah ini semakin diperparah oleh kerusakan bangunan-bangunan SD yang memerlukan perbaikan segera. Laporan tersebut, ditambah tuntutan DPRD agar renovasi dilakukan berdasarkan prioritas dan bukan kedekatan, semakin menegaskan perlunya sistem klasifikasi berbasis data untuk memastikan alokasi intervensi dilakukan dengan adil, transparan, dan efisien (Ramadani, 2025). Pemerintah memang telah melaksanakan berbagai program seperti rehabilitasi sekolah melalui DAK fisik, peningkatan sanitasi, peningkatan literasi, pelatihan guru, dan digitalisasi pembelajaran, tetapi pelaksanaannya masih menghadapi tantangan karena pemetaan kebutuhan di lapangan belum sepenuhnya akurat (UNICEF Indonesia, 2023). Oleh karena itu, dibutuhkan mekanisme berbasis data yang dapat secara komprehensif mengidentifikasi kondisi sekolah untuk mendukung penentuan prioritas intervensi secara tepat sasaran.

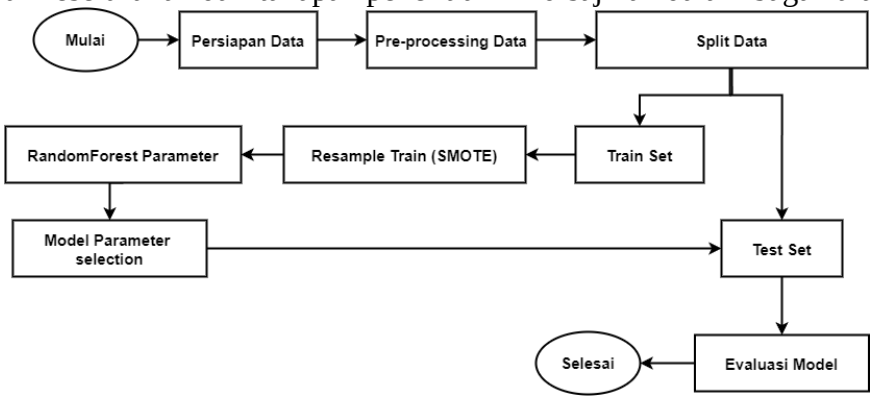
Perkembangan pemanfaatan data pendidikan melalui sistem Dapodik menjadikan *machine learning* sebagai pendekatan yang semakin relevan dalam mendukung pengambilan keputusan berbasis data. Data pendidikan umumnya bersifat multivariat dengan berbagai indikator yang saling terkait, termasuk karakteristik siswa, kondisi sekolah, dan infrastruktur. Meskipun jumlah variabel dalam studi ini relatif terbatas, kompleksitas hubungan antar variabel tetap menimbulkan tantangan dalam pemodelan, seperti potensi redundansi fitur dan risiko *overfitting*. *Random Forest* merupakan salah satu algoritma yang paling banyak digunakan karena mampu menangani variabel yang beragam dan memodelkan hubungan nonlinier dengan baik. Penelitian Adriansyah et al. (2020) menunjukkan bahwa *Random Forest* dapat mencapai akurasi sekitar 91,5% dan terbukti lebih unggul daripada SVM dalam mengklasifikasikan kemampuan adaptasi siswa terhadap pembelajaran (Adriansyah et al., 2022). Penelitian Wibowo (2021) juga melaporkan bahwa *Random Forest* mampu memetakan mutu sekolah secara stabil pada skala nasional (Wibowo, 2021). Sementara itu, penelitian Pratiwi dan Utami (2025) menemukan bahwa penerapan *Random Forest* pada data pendidikan yang telah diseimbangkan kelasnya mampu menghasilkan akurasi hingga 99,1% (Pratiwi & Utami, 2025). Konsistensi hasil tersebut menunjukkan bahwa *Random Forest* tidak hanya efektif dalam mengelola data multivariat dan berdimensi tinggi, tetapi juga mampu memberikan informasi mengenai fitur yang paling berpengaruh dalam proses klasifikasi. Oleh karena itu, *Random Forest* sangat tepat digunakan sebagai dasar penentuan prioritas intervensi pendidikan secara objektif, akurat, dan berbasis data.

Penggunaan data pendidikan melalui sistem Dapodik telah menjadikan *machine learning* sebagai pendekatan yang semakin relevan dalam mendukung pengambilan keputusan berbasis data. Data pendidikan umumnya bersifat multivariat dengan berbagai indikator. Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk mengembangkan model klasifikasi untuk sekolah dasar prioritas di Kabupaten Madiun menggunakan algoritma *Random Forest* dengan memanfaatkan data pendidikan tahun 2025. Model ini dirancang untuk memproses indikator multidimensi seperti jumlah siswa, rasio siswa-guru, kondisi fasilitas, dan variabel terkait lainnya untuk menghasilkan kategori prioritas yang lebih objektif dan sesuai dengan kondisi aktual setiap sekolah. Secara akademis, penelitian ini berkontribusi dalam memperkuat penerapan metode *ensemble learning* dalam konteks pendidikan dasar regional. Secara praktis, hasil penelitian ini diharapkan dapat menjadi dasar bagi pemerintah daerah dalam menentukan alokasi intervensi yang lebih terarah, transparan, dan berkelanjutan. Meskipun jumlah variabel dalam penelitian ini relatif terbatas, kompleksitas hubungan antar variabel masih menimbulkan

tantangan dalam pemodelan, seperti potensi redundansi fitur dan risiko *overfitting*. *Random Forest* merupakan salah satu algoritma yang paling banyak digunakan karena mampu menangani variabel yang beragam dan memodelkan hubungan nonlinier dengan baik.

2 METODE

Penelitian ini dilakukan melalui tahapan-tahapan yang disusun secara sistematis untuk memastikan tercapainya tujuan penelitian. Proses dimulai dengan perumusan dan pemahaman masalah, dilanjutkan dengan pengumpulan dan pengolahan data, kemudian analisis menggunakan metode yang telah ditentukan, dan diakhiri dengan evaluasi terhadap hasil yang diperoleh. Alur keseluruhan dari tahapan penelitian ini disajikan dalam bagan alur berikut.



Gambar 1. Diagram alir penelitian

2.1 Pengumpulan Data

Data yang digunakan adalah data sekolah dasar tahun 2025 yang diperoleh dari Dinas Pendidikan dan Kebudayaan Kabupaten Madiun. Dataset ini terdiri dari 401 entri dengan total 15 variabel yang mencakup jumlah siswa, jumlah guru, jumlah rombel, kondisi bangunan, ketersediaan sarana sekolah, serta indikator literasi, numerasi, karakter, dan kualitas pembelajaran. Variabel target yang digunakan adalah tingkat prioritas intervensi untuk sekolah dasar.

Table 1. Data Sekolah Dasar

No.	Variabel	Tipe Data	Keterangan
1	X1 (Jumlah Murid)	Rasio	Jumlah peserta didik yang terdaftar di sekolah dasar.
2	X2 (Jumlah Guru)	Rasio	Jumlah guru tetap yang mengajar di sekolah dasar.
3	X3 (Jumlah Rombel)	Rasio	Jumlah rombongan belajar (kelas) di sekolah dasar.
4	X4 (Kondisi Gedung)	Ordinal	Kondisi fisik bangunan sekolah (Baik, Rusak Ringan, Rusak Sedang, Rusak Berat).
5	X5 (Perpustakaan)	Nominal	Keberadaan fasilitas perpustakaan di sekolah (Ada/Tidak Ada).
6	X6 (Internet)	Nominal	Akses atau ketersediaan internet di sekolah (Ada/Tidak Ada).
7	X7 (UKS)	Nominal	Keberadaan Unit Kesehatan Sekolah (UKS) di sekolah (Ada/Tidak Ada).

8	X8 (Jumlah Kamar Mandi)	Rasio	Jumlah kamar mandi yang tersedia di sekolah.
9	X9 (Kemampuan literasi)	Rasio	Tingkat kemampuan literasi peserta didik di sekolah dasar (Baik, Sedang, Kurang).
10	X10 (Kemampuan numerasi)	Nominal	Tingkat kemampuan numerasi peserta didik di sekolah dasar (Baik, Sedang, Kurang).
11	X11 (Karakter)	Ordinal	Tingkat pengembangan karakter peserta didik di sekolah dasar (Baik, Sedang, Kurang).
12	X12 (Iklim keamanan satuan pendidikan)	Ordinal	Tingkat keamanan lingkungan sekolah menurut persepsi warga sekolah (Baik, Sedang, Kurang).
13	X13 (Iklim kebinekaan)	Ordinal	Tingkat penerapan nilai toleransi dan keberagaman di sekolah dasar (Baik, Sedang, Kurang).
14	X14 (Kualitas pembelajaran)	Ordinal	Tingkat kualitas proses belajar mengajar di sekolah dasar (Baik, Sedang, Kurang).
15	Y (Level_Prioritas)	Ordinal	Tingkat prioritas intervensi sekolah dasar berdasarkan hasil klasifikasi (Tidak Prioritas, Prioritas Sedang, Prioritas Tinggi).

2.2 Data Preprocessing

1. Data Preparation

Tahap ini dilakukan untuk memastikan bahwa data sekolah dasar yang diperoleh dari Dinas Pendidikan dan Kebudayaan Kabupaten Madiun memenuhi persyaratan penelitian. Pemeriksaan awal dilakukan pada struktur dataset menggunakan .info(), yang mencakup jumlah entri, tipe variabel, format data, dan konsistensi nilai di seluruh kolom. Selain itu, kelengkapan data juga diperiksa untuk memastikan tidak ada nilai yang hilang yang dapat mempengaruhi proses analisis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 401 entries, 0 to 400
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Nama Sekolah                    401 non-null    object
 1   Kecamatan                       401 non-null    object
 2   Jumlah Murid                    401 non-null    int64
 3   Jumlah Guru                     401 non-null    int64
 4   Jumlah Rombel                   401 non-null    int64
 5   Kondisi Gedung                  401 non-null    object
 6   Perpustakaan                    401 non-null    object
 7   Internet                        401 non-null    object
 8   UKS                             401 non-null    object
 9   Jumlah Kamar Mandi              401 non-null    int64
 10  Kemampuan literasi              401 non-null    object
 11  Kemampuan numerasi              401 non-null    object
 12  Karakter                        401 non-null    object
 13  Kualitas pembelajaran           401 non-null    object
 14  Iklim keamanan satuan pendidikan  401 non-null  object
 15  Iklim Kebinekaan                401 non-null    object
dtypes: int64(4), object(12)
memory usage: 50.2+ KB
None
```

Gambar 2. Info Data

2. Data Exploration

Data exploration dilakukan untuk mendapatkan gambaran awal tentang dataset melalui analisis distribusi dan hubungan antar variabel numerik, distribusi variabel kategorikal, serta peta panas korelasi dengan .corr() yang digunakan untuk mengidentifikasi pola, multikolinearitas, dan hubungan antar variabel.

3. Feature Engineering

Feature engineering dilakukan dengan membuat variabel turunan yang lebih representatif, termasuk rasio siswa-guru, rata-rata jumlah siswa per kelas, dan defisit toilet sebagai indikator kecukupan sanitasi. Selain itu, fasilitas sekolah dirangkum dalam variabel Total Fasilitas, dan indikator laporan pendidikan digabungkan untuk mewakili kualitas keseluruhan proses pendidikan di setiap sekolah.

Table 2. Variabel Baru Hasil Feature Engineering

No	Variabel Baru	Tipe Data	Keterangan
1	Rasio Siswa–Guru	Rasio	Indikator beban mengajar
2	Rata-rata Siswa per Kelas	Rasio	Indikator kepadatan kelas
3	Defisit Toilet	Rasio	Kebutuhan sanitasi berdasarkan jumlah siswa
4	Total Fasilitas	Rasio	Akumulasi fasilitas pendukung pembelajaran
5	Skor Mutu Pendidikan	Rasio	Agregasi indikator pembelajaran

4. Data Transformation

Semua variabel kategorikal dalam dataset perlu diubah menjadi nilai numerik agar dapat diproses oleh algoritma *machine learning*. Proses transformasi dilakukan menggunakan metode pemetaan, di mana setiap kategori dalam variabel ordinal atau nominal diberikan nilai numerik sesuai dengan urutan atau kelasnya. Transformasi ini memastikan bahwa semua atribut memiliki format yang sesuai untuk analisis algoritmik.

Table 3. Perbandingan Data sebelum dan sesudah transformasi

Jumlah Murid	Jumlah Guru	Jumlah Rombel	Kondisi Gedung	Perpustakaan	Internet	UKS	Jumlah Kamar Mandi	Kemampua litera
52	6	6	Rusak Ringan	Ada	Ada	Tidak Ada	5	Sedan
85	6	6	Rusak Ringan	Ada	Ada	Ada	8	Ba
95	6	6	Baik	Ada	Ada	Ada	3	Ba
57	7	6	Rusak Sedang	Ada	Ada	Tidak Ada	2	Ba
112	7	6	Rusak Sedang	Ada	Ada	Tidak Ada	4	Ba

Jumlah Murid	Jumlah Guru	Jumlah Rombel	Kondisi Gedung	Perpustakaan	Internet	UKS	Jumlah Kamar Mandi	Kemampuan literas
52	6	6	2	1	1	0	5	2
85	6	6	2	1	1	1	8	3
95	6	6	3	1	1	1	3	3
57	7	6	1	1	1	0	2	3
112	7	6	1	1	1	0	4	3

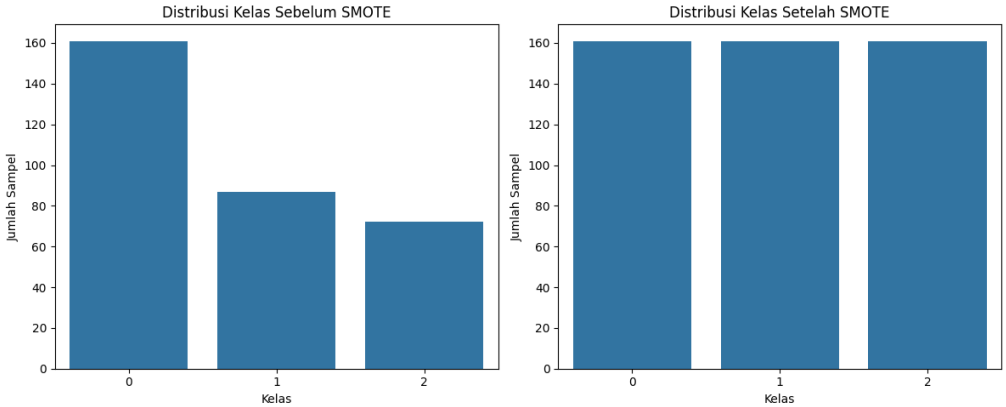
5. Labeling

Proses pelabelan dilakukan untuk membentuk variabel target berupa tingkat prioritas intervensi sekolah yang terdiri atas tiga kelas, yaitu Tidak Prioritas, Prioritas Sedang, dan Prioritas Tinggi. Label pada penelitian ini tidak tersedia secara langsung dalam dataset, sehingga disusun oleh peneliti menggunakan pendekatan berbasis aturan (*rule-based*) berdasarkan pernyataan kebijakan penentuan prioritas pendidikan. Penetapan label didasarkan pada beberapa indikator utama, meliputi kondisi gedung sekolah, rasio siswa per guru, rata-rata jumlah siswa per kelas, jumlah rombongan belajar, ketersediaan fasilitas sanitasi, keberadaan fasilitas pendukung pembelajaran, serta capaian indikator mutu pendidikan. Sekolah dikategorikan sebagai Prioritas Tinggi jika menunjukkan kondisi kritis, seperti

kerusakan bangunan yang parah, jumlah siswa per kelas yang berlebihan, rasio siswa-guru yang tinggi, kekurangan fasilitas sanitasi signifikan, ketiadaan fasilitas pendukung, atau pencapaian kualitas pendidikan yang rendah. Prioritas Sedang diberikan kepada sekolah yang memerlukan perhatian tetapi belum mendesak, sedangkan Tidak Prioritas mencakup sekolah yang telah memenuhi standar pendidikan minimum dan hanya memerlukan pemeliharaan serta pemantauan berkala. Semua aturan pelabelan diterapkan secara konsisten untuk memastikan keterulangan proses dan menghasilkan variabel target yang digunakan dalam tahap pemodelan klasifikasi (UNICEF Indonesia, 2023; Kemendikbudristek, 2023; Direktorat Jenderal Guru dan Tenaga Kependidikan Kementerian Pendidikan dan Kebudayaan, 2016).

6. Data Balancing

Penyeimbangan data dilakukan untuk mengatasi ketidakseimbangan kelas menggunakan teknik *Synthetic Minority Over-sampling Technique* (*SMOTE*). Metode ini bekerja dengan menghasilkan sampel sintetis untuk kelas minoritas berdasarkan kedekatan antar titik data dalam ruang fitur (Rohman et al., 2025). SMOTE memilih sampel minoritas, menentukan tetangga terdekatnya, dan kemudian membuat data baru melalui interpolasi linear dengan faktor acak. Sampel sintetis yang dihasilkan ditambahkan ke dataset sehingga distribusi kelas menjadi lebih seimbang, yang dapat meningkatkan kinerja dan stabilitas model selama fase pemodelan.



Gambar 3. Hasil SMOTE

2.3 Modeling

Pada tahap pemodelan, penelitian ini menggunakan algoritma *Random Forest* sebagai metode klasifikasi untuk menentukan prioritas intervensi di sekolah dasar. *Random Forest* adalah algoritma pembelajaran *ensemble* yang dibangun berdasarkan konsep *bootstrap aggregation* (*bagging*), yaitu penggabungan beberapa pohon keputusan yang dilatih menggunakan sampel data berbeda untuk meningkatkan akurasi dan stabilitas prediksi (Ersa Muliani et al., 2025). Setiap pohon keputusan dibangun dari data pelatihan yang diperoleh melalui pengambilan sampel acak dengan pengembalian (*bootstrap sampling*), sehingga variasi data antar pohon dapat mengurangi risiko *overfitting* (Rahman et al., 2025). Hasil prediksi akhir *Random Forest* ditentukan melalui mekanisme *majority voting*, yaitu kelas yang paling sering diprediksi oleh semua pohon keputusan. Mekanisme ini diformulasikan secara matematis sebagai berikut:

$$\hat{y} = \underset{c \in C}{\operatorname{argmax}} \sum_{i=1}^{m} 1(h_i(x) = C) \quad (1)$$

dengan $\hat{y}$ sebagai kelas prediksi akhir, $h_i(x)$ sebagai hasil prediksi pohon ke-i dan m sebagai jumlah pohon dalam model.

Proses pembentukan pohon keputusan dimulai dari *root node* yang merepresentasikan seluruh data latih, kemudian dilakukan pemilihan subset fitur secara acak pada setiap node untuk menentukan pemisahan terbaik berdasarkan kriteria tertentu hingga mencapai kondisi penghentian (*leaf node*), seperti kedalaman maksimum pohon atau jumlah minimum sampel pada node. Proses ini menghasilkan *leaf node* yang merepresentasikan label kelas. Penerapan unsur keacakan pada pemilihan data dan fitur menyebabkan setiap pohon memiliki struktur yang berbeda, sehingga model menjadi lebih *robust* terhadap variasi data.

Keputusan akhir *Random Forest* diperoleh melalui mekanisme *majority voting*, yaitu kelas yang paling banyak diprediksi oleh seluruh pohon keputusan ditetapkan sebagai hasil klasifikasi. Pendekatan ini efektif dalam meningkatkan akurasi prediksi serta mengurangi bias yang umumnya muncul pada model berbasis pohon tunggal. Dalam penelitian ini, *Random Forest* diimplementasikan dengan beberapa parameter utama, meliputi jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), jumlah fitur yang dipertimbangkan pada setiap pemisahan (*max_features*), serta batas minimum sampel pada node dan daun pohon, yang disesuaikan untuk memperoleh kinerja model yang optimal.

2.4 Evaluation

Evaluasi model dilakukan menggunakan data uji (*testing set*) dengan *confusion matrix*, *Receiver Operating Characteristic* (*ROC*), dan *Area Under the Curve* (*AUC*). *Confusion matrix* memberikan representasi komprehensif terhadap jumlah prediksi yang benar dan salah pada setiap kelas melalui nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) (Imani et al., 2025). Berdasarkan *confusion matrix* tersebut, kinerja model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, yang dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = 2 . \frac{Precision . \ Recall}{Precision + Recall} \tag{5}$$

Selain itu, evaluasi kinerja model juga dilakukan menggunakan *ROC* dan *AUC*. Kurva *ROC* dibentuk berdasarkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai nilai ambang keputusan, dengan TPR dan FPR dirumuskan sebagai berikut:

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

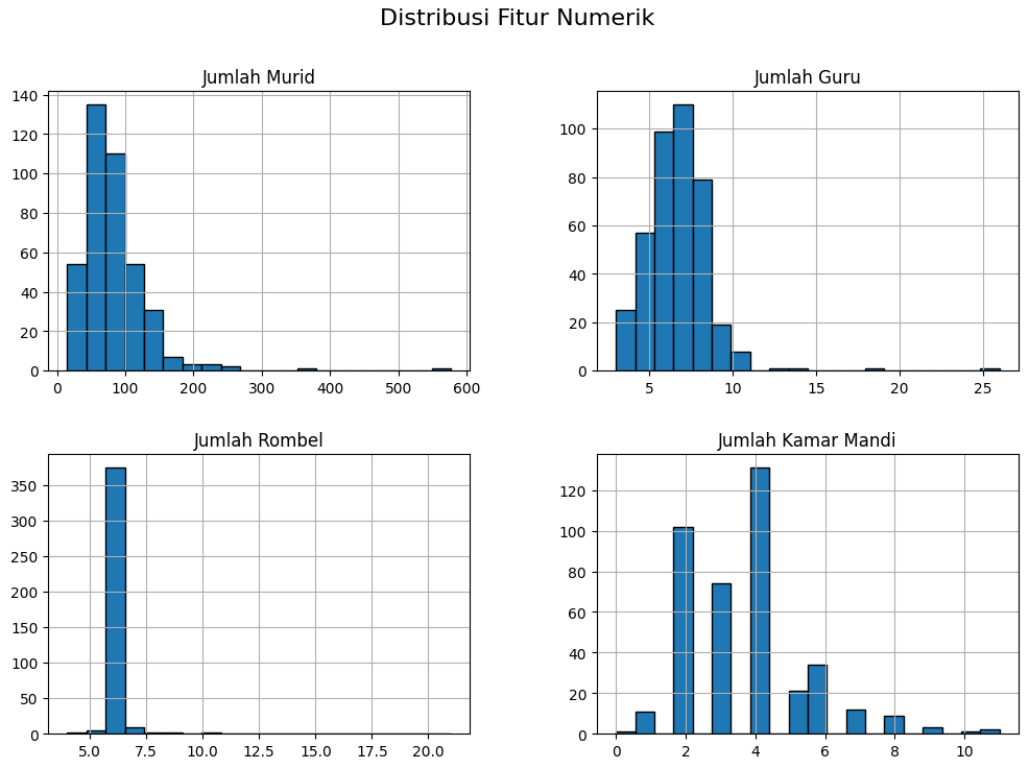
Kurva *ROC* memberikan gambaran sensitivitas dan spesifisitas model pada berbagai ambang batas, sedangkan nilai *AUC* merepresentasikan kemampuan model dalam membedakan kelas secara keseluruhan (Hassanzad & Hajian-Tilaki, 2024).

2.5 Feature Selection and Model Recalibration

Seleksi fitur dilakukan menggunakan *SHAP* (*SHapley Additive exPlanations*) untuk mengidentifikasi kontribusi global masing-masing fitur terhadap keputusan model *Random Forest*. Tingkat kepentingan setiap fitur ditentukan berdasarkan rata-rata nilai *SHAP* mutlak di seluruh sampel, kemudian fitur-fitur tersebut diurutkan dari kontribusi tertinggi hingga terendah (Antonini et al., 2024). Fitur dengan kontribusi rendah dihapus untuk mengurangi kompleksitas dan potensi kebisingan. Selanjutnya, model dikalibrasi ulang dengan melatih kembali *Random Forest* menggunakan fitur-fitur yang terpilih. Kinerja model yang telah direkalibrasi dibandingkan dengan model awal menggunakan metrik evaluasi yang sama, dan model dengan fitur lebih sedikit namun kinerja setara atau lebih baik dipilih sebagai model akhir karena lebih efisien dan lebih mudah untuk diinterpretasikan.

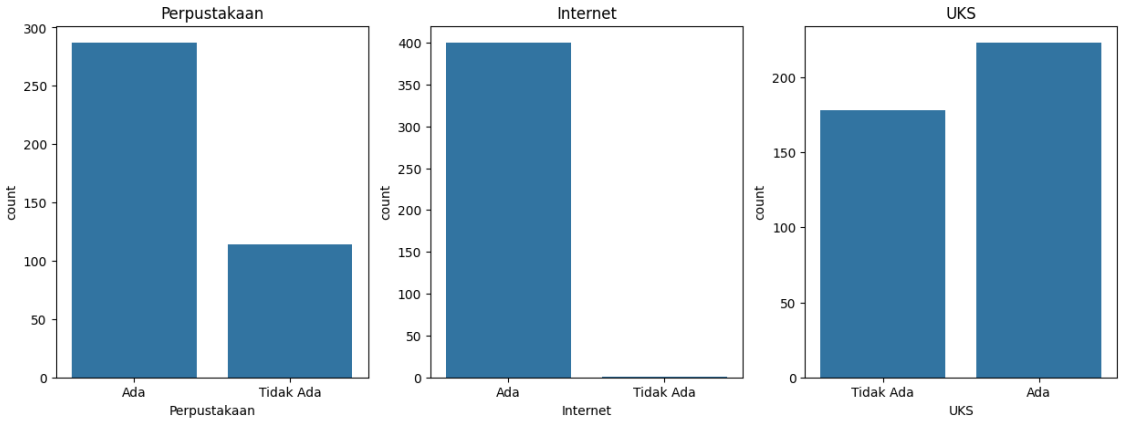
3 HASIL DAN PEMBAHASAN

Sebelum pemodelan, dilakukan analisis deskriptif terhadap data untuk mengenali karakteristik, pola distribusi, dan keterkaitan antara fitur yang terlibat dalam penelitian. Hasil analisis ini disajikan dalam bentuk visualisasi data agar bisa memberikan gambaran awal mengenai keadaan satuan pendidikan berdasarkan aspek numerik, fasilitas, dan indikator kualitas pendidikan.



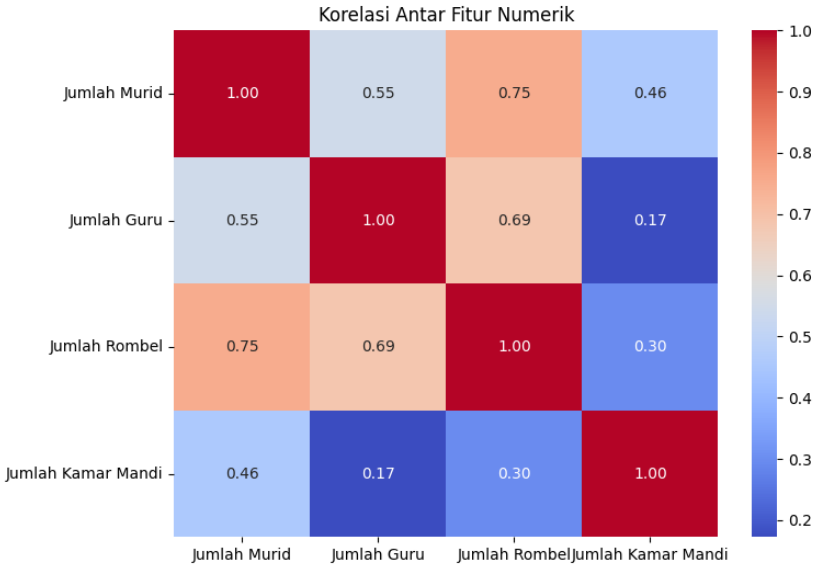
Gambar 4. Distribusi fitur numerik

Visualisasi fitur numerik diatas menunjukkan adanya variasi dan perbedaan skala antar variabel, terutama pada jumlah siswa dan jumlah kelompok belajar yang memiliki rentang nilai yang cukup luas.



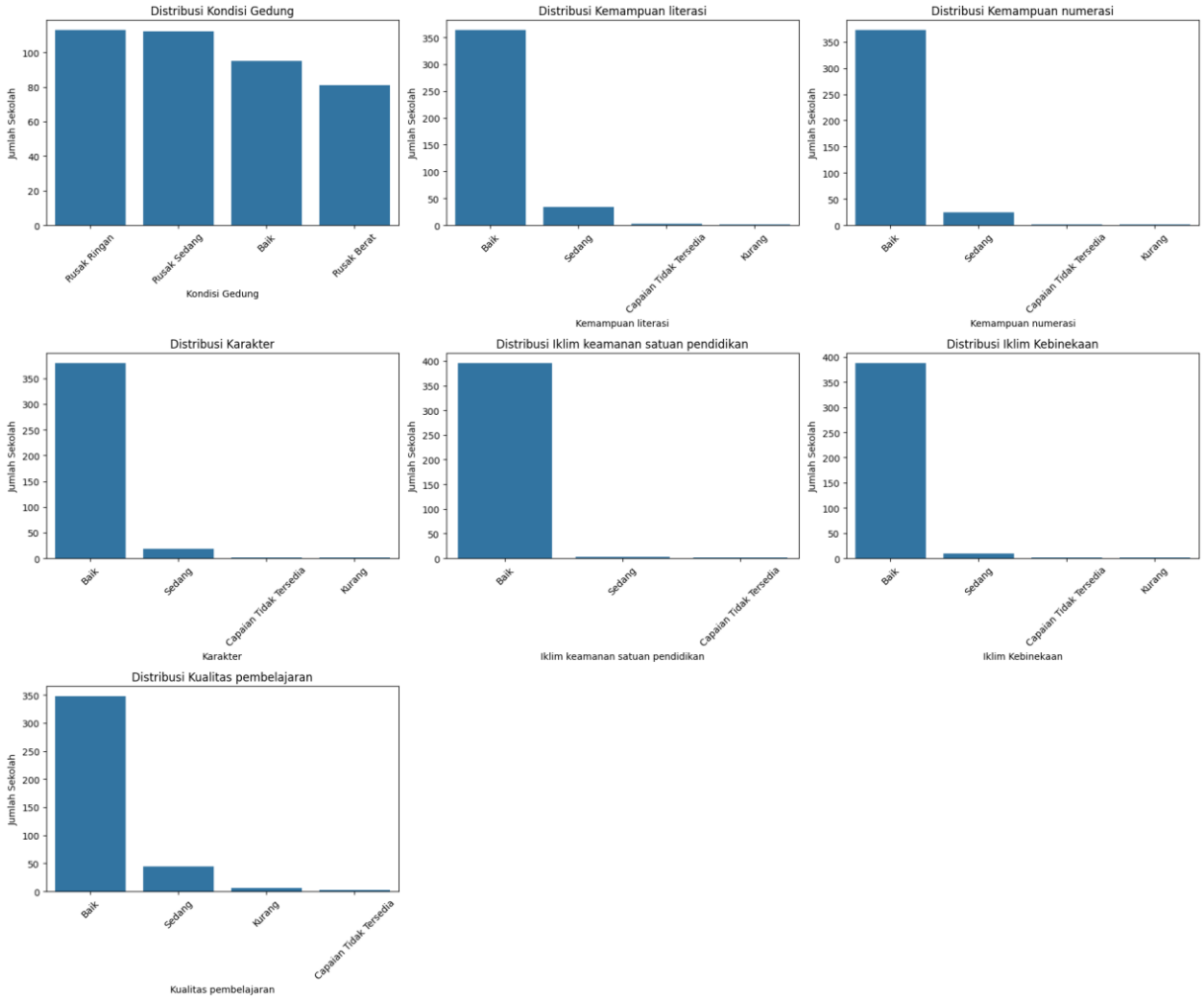
Gambar 5. Distribusi fitur fasilitas

Selain aspek numerik, visualisasi fasilitas sekolah diatas menunjukkan bahwa ada perbedaan ketersediaan sarana pendukung di antara lembaga pendidikan. Perbedaan ini mencerminkan ketidakseimbangan dalam kondisi fasilitas yang bisa berdampak pada tingkat prioritas sekolah.



Gambar 6. Korelasi

Analisis korelasi antar fitur numerik menunjukkan adanya hubungan sedang hingga kuat antara jumlah murid, jumlah guru, dan jumlah rombongan belajar yang mengindikasikan keterkaitan logis antar variabel tanpa adanya korelasi yang berlebihan.



Gambar 7. Distribusi fitur kategorikal

Berkaitan dengan kondisi sarana fisik dan capaian akademik, visualisasi diatas menunjukkan bahwa sekolah memiliki berbagai tingkat kelayakan bangunan yang sebagian besar berada dalam kondisi rusak ringan hingga sedang, sementara hasil literasi dan numerasi umumnya tergolong baik meskipun masih ada yang mencapai tingkat lebih rendah. Hasil ini menunjukkan adanya perbedaan infrastruktur dan hasil belajar di antara sekolah-sekolah yang harus diperhatikan dalam penentuan prioritas.

Pada tahap awal pemodelan, dataset dibagi menjadi data pelatihan dan pengujian dengan rasio 80:20 menggunakan metode *stratified sampling*, dengan 80% untuk data pelatihan dan 20% untuk data pengujian. Pendekatan ini diterapkan untuk memastikan bahwa distribusi kelas dari variabel target Level_Prioritas tetap proporsional dalam data pengujian, sehingga hasil evaluasi kinerja model secara akurat merepresentasikan distribusi data yang sebenarnya. Selain itu, untuk mengatasi masalah ketidakseimbangan kelas dalam data pelatihan, diterapkan *Synthetic Minority Over-sampling Technique* (*SMOTE*). *SMOTE* hanya diterapkan pada data pelatihan untuk meningkatkan kemampuan model dalam mempelajari pola pada kelas minoritas, tanpa mempengaruhi validitas proses evaluasi.

Pengembangan model klasifikasi tingkat prioritas unit pendidikan dilakukan menggunakan algoritma *Random Forest* dengan tiga konfigurasi *hyperparameter* berbeda. Semua model dibangun dalam *pipeline* yang sama, yaitu dengan menerapkan normalisasi fitur menggunakan *MinMaxScaler* sebelum proses pelatihan. Perbedaan konfigurasi model terutama terletak pada jumlah pohon keputusan (*n_estimators*), kedalaman maksimum pohon (*max_depth*), jumlah minimum sampel pada setiap node (*min_samples_split* dan

min_samples_leaf), serta strategi penyeimbangan kelas (*class_weight*). Rincian konfigurasi disajikan pada Tabel 4.

Table 4. Variasi Parameter pada Model Random Forest

Parameter	RF-1	RF-2	RF-3
n_estimators	150	200	300
max_depth	15	20	None
min_samples_split	5	10	8
min_samples_leaf	2	4	3
max_features	log2	sqrt	sqrt
class_weight	balanced	balanced	balanced_subsample

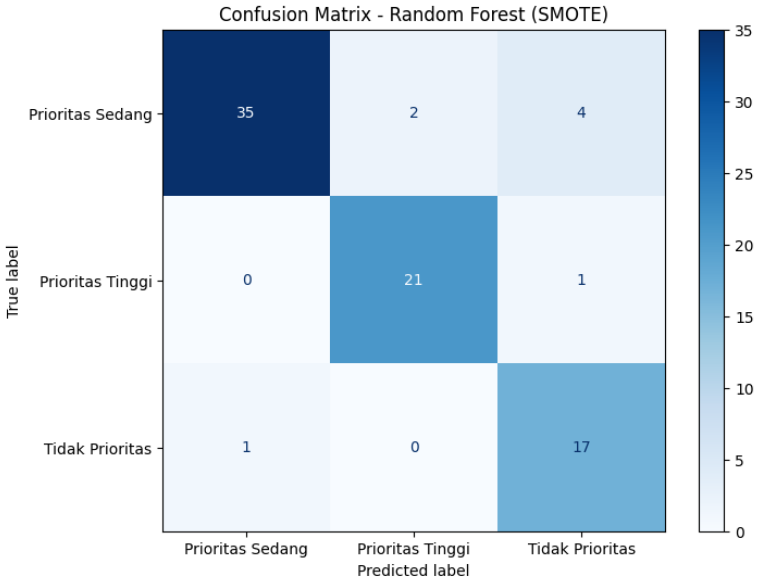
Hasil evaluasi kinerja dari ketiga model ditunjukkan pada Tabel 5, menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* dengan pendekatan rata-rata makro. Pendekatan ini memastikan setiap kelas memiliki kontribusi yang sama dalam menilai kinerja model, terutama mengingat karakteristik data yang tidak seimbang.

Table 5. Perbandingan Kinerja Model Random Forest

Model	Accuracy	Precision	Recall	F1-score
RF-1	0.89	0.88	0.89	0.88
RF-2	0.88	0.86	0.88	0.87
RF-3	0.90	0.89	0.92	0.90

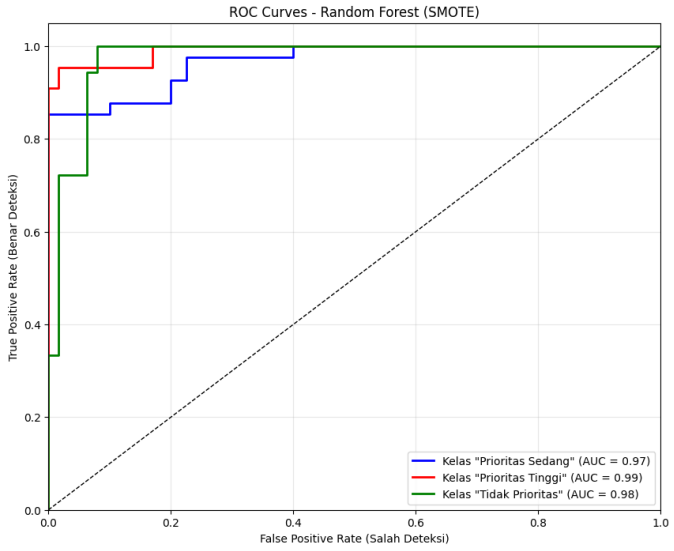
Berdasarkan hasil evaluasi, model RF-3 menunjukkan kinerja terbaik dibandingkan dua konfigurasi lainnya, dengan akurasi sebesar 0,90 dan skor F1 makro sebesar 0,90, yang menunjukkan kinerja seimbang di seluruh kelas prioritas. Peningkatan kinerja RF-3 menunjukkan bahwa peningkatan jumlah pohon keputusan dan penerapan strategi *class_weight=balanced_subsample* dapat meningkatkan kemampuan model dalam mempelajari pola data secara lebih representatif. RF-1 dan RF-2 juga menunjukkan kinerja yang relatif baik, namun RF-2 mengalami sedikit penurunan skor F1 makro dibandingkan RF-1, menunjukkan bahwa nilai lebih tinggi untuk *min_samples_leaf* dan *min_samples_split* cenderung membuat model lebih konservatif dalam membentuk aturan keputusan, sehingga mengurangi sensitivitasnya terhadap kelas minoritas.

Distribusi hasil prediksi model RF-3 pada data uji disajikan pada Gambar 8. *Confussion matrix* menunjukkan bahwa sebagian besar data pada setiap kelas telah diklasifikasikan dengan benar. Pada kelas Prioritas Sedang, 35 dari 41 data diklasifikasikan dengan benar. Kelas Prioritas Tinggi menunjukkan kinerja yang sangat baik dengan 21 dari 22 data diprediksi dengan benar, sementara kelas Tidak Prioritas mencatat 17 dari 18 data diklasifikasikan dengan benar. Kesalahan klasifikasi yang terjadi umumnya terbatas dan cenderung terjadi antar kelas yang berdekatan, sehingga tidak menunjukkan kesalahan prediksi yang signifikan..



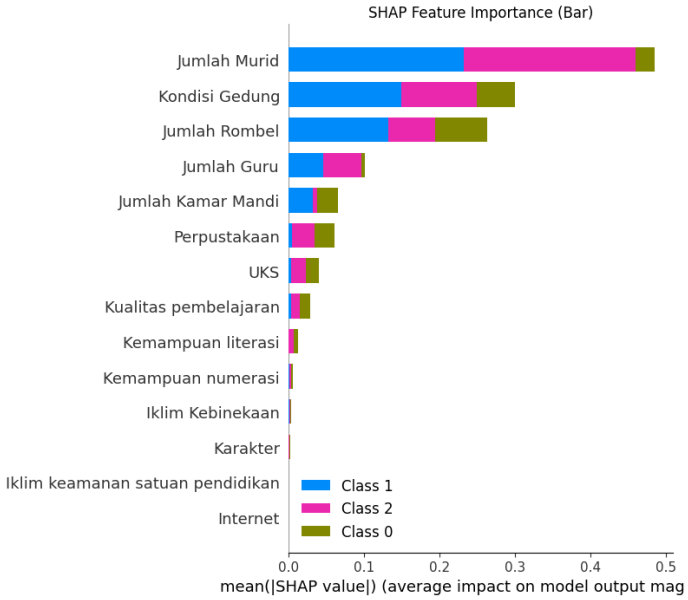
Gambar 8. Confussion matrix

Evaluasi lanjutan dilakukan menggunakan kurva *Receiver Operating Characteristic (ROC)* multi-kelas dengan pendekatan *one-vs-rest*, sebagaimana ditunjukkan pada Gambar 8. Model RF-3 menghasilkan nilai *AUC* mikro sebesar 0,9818 dan *AUC* makro sebesar 0,9780. Selain itu, masing-masing kelas menunjukkan nilai *AUC* yang tinggi, yaitu 0,97 untuk Prioritas Sedang, 0,99 untuk Prioritas Tinggi, dan 0,98 untuk Tidak Prioritas. Nilai *AUC* yang mendekati 1 ini menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dan performa yang relatif seimbang pada seluruh kelas.



Gambar 9. Kurva ROC AUC

Analisis keterinterpretasian menggunakan metode *SHAP (SHapley Additive exPlanations)* menunjukkan bahwa fitur Jumlah Siswa, Kondisi Bangunan, dan Jumlah Rombel merupakan faktor yang paling dominan dalam menentukan tingkat prioritas unit pendidikan. Sebaliknya, fitur Internet, Iklim Keamanan Sekolah, Karakter, dan Iklim Keberagaman memiliki kontribusi yang sangat kecil terhadap prediksi model. Temuan ini menunjukkan bahwa faktor kuantitatif dan kondisi fisik sekolah memiliki peran yang lebih besar dalam menentukan prioritas dibandingkan faktor pendukung lainnya..



Gambar 10. Feature importance

Berdasarkan hasil ini, empat fitur dengan kontribusi terendah dihapus untuk menyederhanakan model. Evaluasi menunjukkan bahwa model tetap mempertahankan kinerja tinggi, dengan akurasi 0,91 dan *F1-score* makro 0,91. Kemampuan model dalam membedakan kelas juga sangat baik, tercermin dari *AUC* mikro 0,9835 dan *AUC* makro 0,9807, serta skor F1 makro pada *cross-validation* sebesar 0,8777. Hasil ini menegaskan bahwa pengurangan fitur berbasis *SHAP* dapat menghasilkan model yang lebih efisien dan mudah diinterpretasikan tanpa mengorbankan kinerja klasifikasi.

4 KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma *Random Forest* mampu mengklasifikasikan secara efektif tingkat prioritas intervensi sekolah dasar menggunakan data pendidikan yang multivariat dan tidak seimbang. Model terbaik mencapai akurasi dan *F1-score* makro sebesar 0,90 dengan nilai *AUC* mikro di atas 0,98. Penerapan seleksi fitur berbasis *SHAP* berhasil mengidentifikasi fitur paling berkontribusi dan memungkinkan pengurangan jumlah fitur tanpa menurunkan kinerja model. Hasil rekalibrasi menunjukkan kinerja setara atau sedikit lebih baik dibandingkan model awal dengan kompleksitas lebih rendah. Interpretasi model menunjukkan bahwa indikator kondisi fisik sekolah dan kepadatan siswa memiliki pengaruh dominan dalam menentukan prioritas. Oleh karena itu, kombinasi *Random Forest* dan *SHAP* menghasilkan model yang akurat, efisien, dan dapat diinterpretasikan, dengan potensi mendukung pengambilan keputusan yang lebih objektif dan berbasis data dalam perencanaan intervensi pendidikan.

UCAPAN TERIMAKASIH

Penulis berterima kasih kepada semua pihak yang telah membantu, baik secara langsung maupun tidak langsung, dalam penyusunan jurnal ini. Dukungan, masukan, dan kerja sama yang diberikan sangat berarti bagi kelancaran penelitian ini.

DAFTAR PUSTAKA

Adriansyah, I., Mahendra, M. D., Rasywir, E., & Pratama, Y. (2022). Perbandingan Metode Random Forest Classifier dan SVM Pada Klasifikasi Kemampuan Level Beradaptasi Pembelajaran Jarak Jauh Siswa. *Bulletin of Informatics and Data Science*, *1*(2), 98. https://doi.org/10.61944/bids.v1i2.49

Antonini, A. S., Tanzola, J., Asiain, L., Ferracutti, G. R., Castro, S. M., Bjerg, E. A., & Ganuza, M. L. (2024). Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task. *Applied Computing and Geosciences*, *23*(February), 100178. https://doi.org/10.1016/j.acags.2024.100178

Direktorat Jenderal Guru dan Tenaga Kependidikan Kementerian Pendidikan dan Kebudayaan. (2016). *Aturan Terbaru Rasio Minimal Jumlah Siswa terhadap Guru*. Panduan Dapodik. https://www.panduandapodik.id/2016/11/aturan-terbaru-rasio-minimal-jumlah-siswa-terhadap-guru.html

Ersa Muliani, et al. (2025). CESS : journal of computer engineering, system and science. *Journal of Computer Engineering, System and Science*, *10*(2), 563–571.

Hassanzad, M., & Hajian-Tilaki, K. (2024). Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. *BMC Medical Research Methodology*, *24*(1), 1–13. https://doi.org/10.1186/s12874-024-02198-2

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies*, *13*(3), 1–40. https://doi.org/10.3390/technologies13030088

Kemendikbudristek. (2023). *Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 47 Tahun 2023 tentang Standar Pengelolaan pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah*. Badan Pembinaan Hukum Nasional Republik Indonesia. https://peraturan.bpk.go.id/Details/289462/permendikbudriset-no-47-tahun-2023

Pratiwi, A. N., & Utami, E. (2025). Prediksi Kinerja Akademik Matematika Siswa berdasarkan Kepribadian Big Five menggunakan Random Forest dengan Teknik Synthetic Minority Over - Sampling Personality Traits using Random Forest with Synthetic Minority Over - Sampling. *Sistemasi: Jurnal Sistem Informasi*, *14*(2), 985–1000. https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/5102/920

Rahayu, D. (2025). *397 SD di Madiun Kekurangan Murid, 6 SMP Kelebihan Pagu!* Radar Madiun (Jawa Pos Group). https://radarmadiun.jawapos.com/kab-madiun/806188903/397-sd-di-madiun-kekurangan-murid-6-smp-kelebihan-pagu

Rahman, A., Mahdiana, D., & Fauzi, A. (2025). Predicting Student On-Time Graduation Using Particle Swarm Optimization and Random Forest Algorithms. *Indonesian Journal of Artificial Intelligence and Data Mining*, *8*(1), 161. https://doi.org/10.24014/ijaidm.v8i1.33577

Ramadani, F. (2025). *DPRD Madiun Soroti Sekolah Rusak, Dorong Renovasi Berdasarkan Prioritas: Bukan Karena Kedekatan*. Tribun Network. https://jatim.tribunnews.com/2025/05/23/dprd-madiun-soroti-sekolah-rusak-dorong-renovasi-berdasarkan-prioritas-bukan-karena-kedekatan

Rohman, M. G., Abdullah, Z., Kasim, S., & Rasyidah. (2025). Hybrid Logistic Regression Random Forest on Predicting Student Performance. *International Journal on Informatics Visualization*, *9*(2), 859–866. https://doi.org/10.62527/joiv.9.2.3487

UNESCO. (2025). *https://www.unesco.org/en/literacy/need-know*. UNESCO. https://www.unesco.org/en/literacy/need-know

UNICEF Indonesia. (2023). *Policy brief: WASH school renovation programme*. https://www.unicef.org/indonesia/media/25116/file/policy-brief-wash-school-renovation-programme.pdf

Wibowo, S. (2021). Building a Classification Model to Predict School Quality in Indonesia. *Proceedings of the International Conference on Educational Assessment and Policy (ICEAP 2020)*, *545*(Iceap 2020), 111–114. https://doi.org/10.2991/assehr.k.210423.074