Evaluasi Support Vector Machine dan Random Forest untuk Klasifikasi Kemiskinan di Indonesia

1st Sandy Yopa Boangmanalu
Teknik Informatika
Universitas Telkom Purwokerto
Purwokerto, Indonesia
sandyyopa@student.telkomuniversity.ac.id

2nd Amalia Beladinna Arifa
Teknik Informatika
Universitas Telkom Purwokerto
Purwokerto, Indonesia
amaliabela@telkomuniversity.ac.id

3rd Nenen Isnaeni
Teknik Informatika
Universitas Telkom Purwokerto
Purwokerto, Indonesia
nenisna@telkomuniversity.ac.id

***Abstrak*—Kemiskinan merupakan persoalan multidimensi yang menuntut pendekatan analitik presisi. Penelitian ini mengevaluasi kinerja algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF) untuk mengklasifikasikan status kemiskinan kabupaten/kota di Indonesia. Menggunakan dataset sekunder indikator sosial-ekonomi dari Kaggle (514 observasi, 9 fitur) dengan distribusi kelas tidak seimbang (87,94% "Tidak Miskin" dan 12,06% "Miskin"), penelitian ini menerapkan *Synthetic Minority Over-sampling Technique* (SMOTE) untuk menangani ketidakseimbangan kelas. Evaluasi dilakukan melalui skenario pembagian data (90:10, 80:20, 70:30, 50:50) dan *hyperparameter tuning*. Hasil menunjukkan model terbaik SVM adalah pendekatan *baseline* pada skenario 70:30 (Akurasi 0,9419; F1 0,7273), sedangkan RF terbaik pada *baseline* 80:20 (Akurasi 0,9417; F1 0,7273; PR-AUC 0,7628). Analisis *feature importance* RF menunjukkan IPM dan pengeluaran per kapita sebagai indikator paling dominan.**

***Kata kunci*—Kemiskinan, Klasifikasi, *Support Vector Machine*, *Random Forest*, SMOTE, *Confusion Matrix, Feature Importance*.**

I. PENDAHULUAN

Kemiskinan masih menjadi isu strategis pembangunan nasional. Meskipun data BPS September 2024 menunjukkan penduduk miskin sebesar 8,57%, disparitas antarwilayah tetap tinggi, terutama di kawasan Maluku-Papua (18,62%). Kemiskinan dipengaruhi oleh kompleksitas faktor seperti pendidikan, kesehatan (IPM), pengeluaran per kapita, serta infrastruktur dasar.[1].

Secara konseptual, kemiskinan dipengaruhi oleh keterkaitan kompleks antarvariabel. Faktor pendidikan, seperti Rata-rata Lama Sekolah, berperan penting dalam meningkatkan produktivitas dan peluang kerja [2]. Selain itu, aspek kesehatan dan standar hidup layak yang tercermin dalam Indeks Pembangunan Manusia (IPM) [3] dan pengeluaran per kapita juga menjadi determinan utama [4]. Faktor makroekonomi seperti PDRB serta infrastruktur dasar (sanitasi dan air bersih)[5] turut berkontribusi signifikan terhadap status kesejahteraan suatu wilayah [6].Kompleksitas tersebut membuat pemetaan status kemiskinan membutuhkan metode yang mampu menangani hubungan nonlinier antarvariabel dan variasi antarwilayah yang besar. Dalam praktik kebijakan, klasifikasi status kemiskinan wilayah penting untuk mendukung perencanaan pembangunan, penentuan prioritas intervensi, dan peningkatan ketepatan sasaran program bantuan. Klasifikasi yang tidak akurat berisiko menimbulkan *exclusion error* (wilayah miskin tidak terdeteksi) maupun *inclusion error* (wilayah tidak miskin salah dikategorikan miskin), yang pada akhirnya dapat mengurangi efektivitas kebijakan. Namun, tantangan utama dalam pemodelan kemiskinan sering kali terkait dengan distribusi data yang tidak seimbang (*imbalanced data*). Ketika jumlah kelas minoritas (miskin) jauh lebih sedikit dibanding kelas mayoritas (tidak miskin), model cenderung "belajar" untuk mengoptimalkan prediksi pada kelas mayoritas [7]. Akibatnya, metrik seperti akurasi dapat terlihat tinggi, tetapi kemampuan model mendeteksi kelas minoritas rendah [8].

Sejumlah penelitian sebelumnya telah menggunakan pendekatan *data mining* dan *machine learning* untuk pemetaan kesejahteraan/kemiskinan, misalnya pemanfaatan RF untuk penentuan penerima bantuan [9] atau klasifikasi penduduk miskin dengan RF dan KNN [10], serta kajian perbandingan SVM dan RF pada berbagai domain [11]..

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada evaluasi kinerja SVM dan RF untuk klasifikasi kemiskinan kabupaten/kota di Indonesia mempertimbangkan penanganan data tidak seimbang menggunakan SMOTE [12], serta menganalisis variabel dominan melalui *feature importance*. Kajian ini diharapkan memberikan kontribusi metodologis dan praktis dalam menghasilkan model yang terukur pada metrik kelas minoritas.

II. KAJIAN TEORI

Bagian ini membahas landasan teori yang digunakan sebagai dasar penelitian, khususnya terkait klasifikasi kemiskinan kabupaten/kota di Indonesia menggunakan metode *machine learning*. Kajian teori mencakup indikator kemiskinan yang digunakan, konsep *supervised learning*, algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF), teknik penanganan data tidak seimbang dengan *SMOTE*, serta metrik evaluasi seperti akurasi, presisi, recall, F1-score, ROC-AUC, dan PR-AUC. Uraian ini bertujuan

memberikan pemahaman konseptual dan memperkuat justifikasi pemilihan metode dalam penelitian.

A. Kemiskinan dan Indikator Sosial Ekonomi

Penelitian ini menggunakan 9 fitur indikator sosial-ekonomi sebagai proksi kualitas hidup [13], meliputi: Rata-rata Lama Sekolah, Pengeluaran per Kapita, PDRB Harga Konstan, IPM, Umur Harapan Hidup, Sanitasi Layak [5], Air Minum Layak [14], Tingkat Pengangguran Terbuka (TPT) [15], dan Tingkat Partisipasi Angkatan Kerja (TPAK) [16].

B. Support Vector Machine

Support Vector Machine (SVM): Bertujuan mencari *hyperplane* terbaik untuk memisahkan dua kelas data. SVM efektif pada ruang dimensi tinggi menggunakan fungsi kernel (Linear, RBF) [17]. Adapun formulasi matematis *Support Vector Machine* untuk klasifikasi dinyatakan dalam persamaan (2.1).

$$f(x) = sign(W^T x + b) \quad (2.1)$$

Keterangan :

$f(x)$ = hasil prediksi kelas
W = *vektor* bobot
x = *vektor input*
b = bias yang mengatur posisi
$sign$ = fungsi yang menentukan kelas

C. Random Forest

Random Forest merupakan metode *ensemble learning* yang menggabungkan beberapa pohon keputusan guna meningkatkan akurasi serta memperbaiki stabilitas model [12]. Cara kerja *Random Forest* melibatkan pembuatan pohon keputusan secara acak dengan menggunakan subset data dan fitur yang berbeda-beda, kemudian hasil prediksi dari setiap pohon tersebut dikombinasikan dengan menggunakan teknik *voting* mayoritas [18].

Rumus umum dari *Random Forest* dinyatakan dalam persamaan (2.2)

$$\hat{y} = mode\ \{T_1(x), T_2(x), \dots, T_n(x)\} \quad (2.2)$$

Keterangan :

y = prediksi kelas akhir
$T_1(x)$ = hasil klasifikasi dari pohon ke
N = jumlah total pohon dalam hutan
$mode$ = fungsi yang mengembalikan kelas yang paling sering muncul dari hasil klasifikasi $T_1(x), T_2(x), \dots, T_n(x)$

Selain berfungsi sebagai model klasifikasi, *Random Forest* juga menyediakan mekanisme interpretasi model melalui pengukuran feature importance. Nilai *feature importance* menggambarkan kontribusi relatif setiap variabel dalam proses pengambilan keputusan model dan umumnya dihitung menggunakan pendekatan *Mean Decrease Impurity* (MDI), yaitu pengurangan tingkat impuritas (misalnya indeks Gini) yang dihasilkan oleh suatu fitur pada seluruh pohon keputusan. Semakin besar nilai *feature importance* suatu variabel, semakin besar pula perannya dalam menentukan hasil klasifikasi [19]. Analisis ini memungkinkan identifikasi variabel yang paling berpengaruh, sehingga model *Random Forest* tidak hanya bersifat *prediktif*, tetapi juga interpretatif dalam menjelaskan faktor-faktor yang memengaruhi status kemiskinan.

D. SMOTE

SMOTE adalah teknik *oversampling* untuk data tidak seimbang dengan membangkitkan contoh sintetis pada kelas minoritas. Secara teknis, untuk setiap sampel minoritas x, dicari *k*tetangga terdekat (misalnya *k-nearest neighbors*) dalam ruang fitur.

Kemudian dipilih salah satu tetangga x_{nn}, dan sampel sintetis dibangkitkan dengan interpolasi linier:

$$x_{new} = x + \lambda(x_{nn} - x), \lambda \sim U\ (0,1) \quad (2.3)$$

Keterangan:

x_{new} = data
x = sampel asli
x_{nn} = salah satu tetangga terdekat
$\lambda \sim U\ (0,1)$ = bilangan acak
$x_{nn} - x$ = vektor arah dari x menuju x_{nn}

Dengan cara ini, SMOTE menambah kepadatan kelas minoritas di sekitar manifold data minoritas, sehingga mengurangi bias model terhadap kelas mayoritas. Studi yang memadukan RF dengan SMOTE menunjukkan peningkatan sensitivitas deteksi kelas minoritas, namun sering muncul *trade-off* berupa peningkatan *false positive* [20]

E. Evaluasi Model

Model dievaluasi menggunakan *Confusion Matrix* (TP, TN, FP, FN) dan metrik turunan: Akurasi, Presisi, *Recall*, F1-Score, ROC-AUC, dan PR-AUC. Pada data tidak seimbang, PR-AUC dan F1-Score lebih representatif dibandingkan akurasi semata. Adapun rumus dari *Confusion Matrix* adalah :

Akurasi (Accuracy) : Akurasi mengukur seberapa besar porsi prediksi yang benar dari total prediksi.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.4)$$

Presisi (*Precision*) : Presisi mengukur seberapa besar proporsi prediksi positif yang benar.

$$Presisi = \frac{TP}{TP+FP} \quad (2.5)$$

Recall (Sensitivity) : Mengukur kemampuan model menemukan kembali data positif

$$Recall = \frac{TP}{TP+FN} \quad (2.6)$$

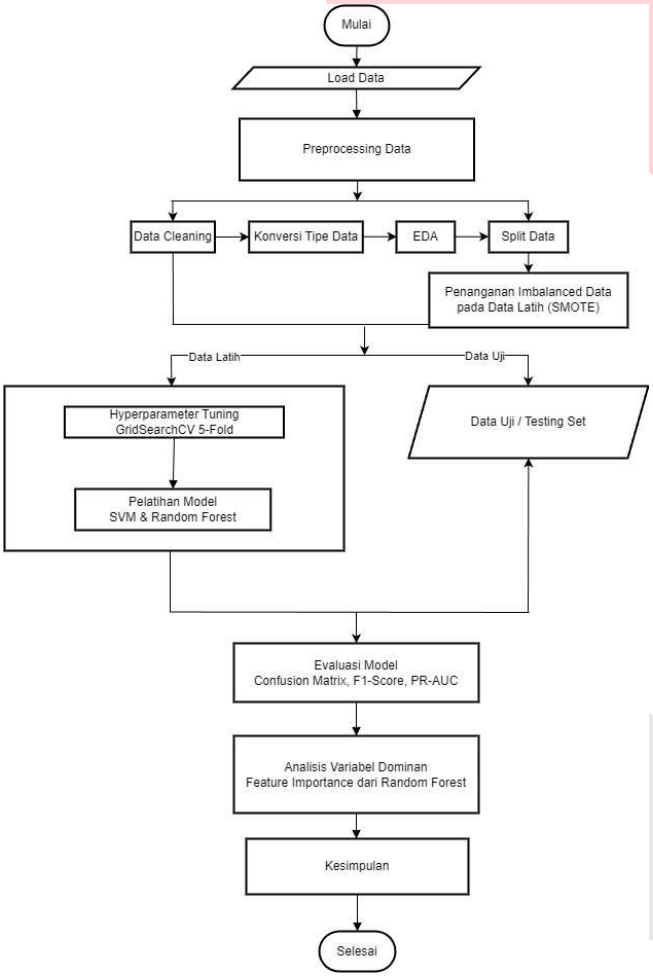
F-1 Score : rata-rata harmonik dari Presisi dan *Recall*, sangat berguna untuk data yang tidak seimbang.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Reccall} \quad (2.7)$$

Selain itu digunakan *ROC-AUC* dan *PR-AUC*. Pada kondisi data tidak seimbang, *PR-AUC* sering lebih representatif karena menekankan kinerja pada kelas positif (minoritas) dibanding sekadar pemisahan ranking keseluruhan.

III. METODE

Penelitian ini mengevaluasi kinerja *Support Vector Machine* (SVM) dan *Random Forest* (RF) dalam mengklasifikasikan status kemiskinan kabupaten/kota di Indonesia pada kondisi data tidak seimbang.Alur penelitian secara keseluruhan dirangkum dalam diagram alir pada gambar berikut.



GAMBAR 1
(Diagram Alir Penelitian)

A. Pengumpulan Dataset

Tahap pengumpulan data dilakukan dengan mengunduh data sekunder dari penyedia *dataset* publik, yaitu Kaggle. *Dataset* yang dikumpulkan memiliki 514 baris data, di mana setiap baris merepresentasikan satu kabupaten/kota. Data memuat indikator makro yang mencakup aspek pendidikan, kesehatan, ekonomi, infrastruktur, dan ketenagakerjaan. Dalam penelitian ini, indikator yang digunakan sebagai variabel masukan berjumlah 9 *fitur*, sedangkan Status Kemiskinan digunakan sebagai variabel *target*.

B. *Preprocessing* Data

Tahap selanjutnya di penelitian ini adalah tahap *preprocessing* untuk menyiapkan *dataset* yang akan dianalisis lebih lanjut. Berdasarkan karakteristik data yang diperoleh, proses ini meliputi beberapa langkah sistematis, antara lain:

1. Pembersihan Data (*Data Cleaning*)
2. Konversi Tipe Data (*Type Casting*) dan Penanganan *Missing Values*
3. Penghapusan Fitur (Feature Dropping)
4. Penetapan dan Definisi *Label* Target

Sebelum dilakukan pemrosesan data lebih lanjut, tahap *Exploratory Data Analysis* (*EDA*) dilakukan untuk memahami karakteristik dasar dan pola distribusi data. Proses ini bertujuan untuk mendapatkan wawasan awal mengenai kualitas data serta hubungan antar variabel. Kegiatan yang dilakukan dalam tahap ini meliputi:

1. Analisis Statistik Deskriptif: Melihat gambaran umum data melalui nilai rata-rata (*mean*), median, dan standar deviasi untuk mengetahui sebaran nilai pada setiap *fitur*
2. Visualisasi Distribusi Kelas: Memeriksa proporsi jumlah data antara kelas *target* "Miskin" dan "Tidak Miskin" untuk mengidentifikasi adanya ketidakseimbangan kelas (*imbalanced data*).
3. Deteksi *Outlier*: Menggunakan visualisasi *boxplot* untuk mendeteksi adanya pencilan nilai (*outlier*) pada variabel numerik yang berpotensi memengaruhi kinerja model.
4. Analisis Korelasi: Menggunakan *heatmap* korelasi untuk melihat hubungan keeratan antar variabel independen maupun hubungan variabel independen terhadap variabel *target*.

Tahap berikutnya adalah pembagian data (*data splitting*) yang bertujuan untuk memecah *dataset* menjadi dua *subset* terpisah, yaitu data latih (*training data*) dan data uji (*testing data*). Skenario pembagian data yang diterapkan dalam penelitian ini mencakup empat rasio perbandingan, yaitu 90:10, 80:20, 70:30, dan 50:50. Penerapan variasi skenario ini bertujuan untuk mengevaluasi model pada kondisi yang berbeda-beda guna mengetahui skenario mana yang menghasilkan evaluasi terbaik [7].

C. Penanganan Ketidaseimbangan Kelas (SMOTE)

Ketidakseimbangan kelas terjadi karena jumlah data “Tidak Miskin” lebih dominan dibandingkan “Miskin”. Untuk mengatasi kondisi tersebut, diterapkan SMOTE sebagai teknik *oversampling* dengan membangkitkan sampel sintetis pada kelas minoritas menggunakan konsep tetangga terdekat (*k-nearest neighbors*). Penerapan SMOTE dilakukan hanya pada data latih setelah proses pembagian data, sedangkan data uji dibiarkan dalam kondisi asli agar evaluasi tetap objektif dan terhindar dari *data leakage*.

D. Hyperparameter Tuning

Setelah tahap seleksi dan standarisasi *fitur*, dilakukan proses penalaan parameter (*hyperparameter tuning*) untuk mendapatkan konfigurasi model yang optimal. Tujuan utama dari tahap ini adalah untuk meningkatkan kinerja algoritma sekaligus mencegah terjadinya *overfitting* atau *underfitting*. Penelitian ini menggunakan metode *Grid Search* yang dikombinasikan dengan *Stratified K-Fold Cross-Validation* sebanyak 5 lipatan (*5-fold*). Adapun parameter yang diuji untuk masing-masing algoritma adalah sebagai berikut:

1. *Support Vector Machine* (SVM): Parameter yang dioptimasi meliputi tipe *kernel* (*RBF* dan *Linear*), nilai regularisasi(*C*) untuk mengontrol kompleksitas *margin*, serta nilai *gamma* untuk menentukan jangkauan pengaruh data latih.
2. *Random Forest* (RF): Parameter yang diuji meliputi jumlah pohon (*n_estimators*), kedalaman maksimal pohon (*max_depth*), serta parameter pembatas percabangan seperti *min_samples_split* dan *min_samples_leaf* untuk mengontrol struktur pohon agar tidak terlalu kompleks.

E. Pemodelan

Pada penelitian ini, tahap pemodelan dilakukan menggunakan dua algoritma *machine learning*, yaitu *Support Vector Machine* (SVM) dan *Random Forest*. Tujuan utamanya adalah membangun model klasifikasi yang mampu memprediksi status kemiskinan suatu wilayah ("Miskin" atau "Tidak Miskin") secara akurat berdasarkan *fitur* sosial-ekonomi yang telah diproses. Model dilatih menggunakan data latih (*training data*) dengan konfigurasi parameter terbaik yang diperoleh dari tahap *hyperparameter tuning*.

F. Evaluasi Model

Tahapan berikutnya adalah evaluasi model, di mana dilakukan penilaian terhadap hasil klasifikasi untuk mengukur efektivitas algoritma *Support Vector Machine* dan *Random Forest*. Evaluasi ini menggunakan *Confusion Matrix* sebagai alat ukur utama untuk membandingkan hasil prediksi model dengan data aktual.

Melalui matriks ini, performa model dalam memprediksi status kemiskinan dapat dianalisis secara mendalam melalui berbagai indikator kinerja seperti *Accuracy*, *Recall*, *Presisi*, dan *F1-Score*.Evaluasi akurasi semata dapat menyesatkan pada data tidak seimbang [21].

G. Feature Importance

Analisis *Feature Importance* dilakukan untuk memahami kontribusi setiap variabel independen (fitur) terhadap hasil prediksi model *Machine Learning*. Dalam penelitian ini, analisis kepentingan fitur diekstraksi utamanya dari model *Random Forest* yang memiliki kapabilitas *embedded feature selection* melalui perhitungan *Mean Decrease Impurity* (Gini Importance). Nilai ini merepresentasikan seberapa besar suatu fitur mampu mengurangi ketidakmurnian (*impurity*) dalam pembentukan *decision tree*.

IV. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil eksperimen pembagian data dan penerapan teknik penanganan data tidak seimbang pada model *Support Vector Machine* (SVM) dan *Random Forest* (RF) untuk klasifikasi kemiskinan kabupaten/kota di Indonesia. Eksperimen dilakukan untuk membandingkan kinerja masing-masing model pada beberapa skenario pembagian data latih dan data uji, serta dua pendekatan pelatihan, yaitu *baseline* (tanpa penyeimbangan kelas) dan *SMOTE* pada data latih.

A. Load Data

Pada tahap ini dilakukan eksplorasi awal data (*Exploratory Data Analysis*/EDA) untuk memahami karakteristik dataset yang digunakan sebelum proses pemodelan dilakukan.

TABEL 1
(INFORMASI DATASET)

Kab/Kota	RLS	Pengeluaran	IPM	UHH	Sanitasi	Air Minum	TPT	TPAK	PDRB
Simeulue	9,48	7148	66,41	65,28	71,56	87,45	5,71	71,15	1648096
Aceh Singkil	8,68	8776	69,22	67,43	69,56	78,58	8,36	62,85	1780419
Aceh Selatan	8,88	8180	67,44	64,4	62,55	79,65	6,46	60,85	4345784
Aceh Tenggara	8,88	8030	69,44	68,22	62,71	86,71	6,43	69,62	3487157
Aceh Timur	9,67	8577	67,83	68,74	66,75	83,16	7,13	59,48	8433526

B. Preprocessing Data

Tahap *preprocessing* dilakukan untuk memastikan dataset bersih, konsisten, dan siap digunakan pada pemodelan. Proses diawali dengan *data cleaning* berupa penghapusan spasi berlebih (*trimming*) serta penyesuaian format numerik (koma menjadi titik) agar seluruh nilai dapat diproses sebagai bilangan.

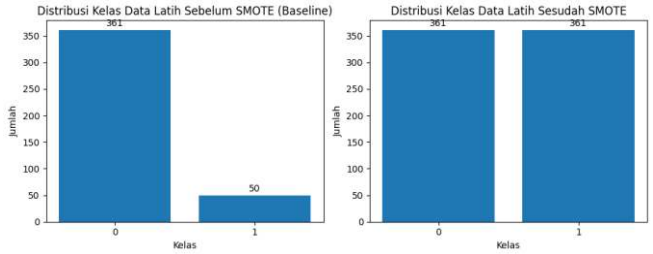
Selanjutnya, seluruh fitur dikonversi ke tipe numerik menggunakan *pd.to_numeric(errors='coerce')* sehingga nilai tidak valid otomatis menjadi NaN, kemudian baris yang memiliki nilai hilang pada fitur maupun target dihapus. Setelah itu dilakukan EDA untuk memahami karakteristik data. Hasil EDA menunjukkan adanya ketidakseimbangan kelas yang signifikan pada variabel target, di mana kelas 0 (tidak miskin) mendominasi dibanding kelas 1 (miskin)



GAMBAR 3
(Distribusi kelas target pada dataset)

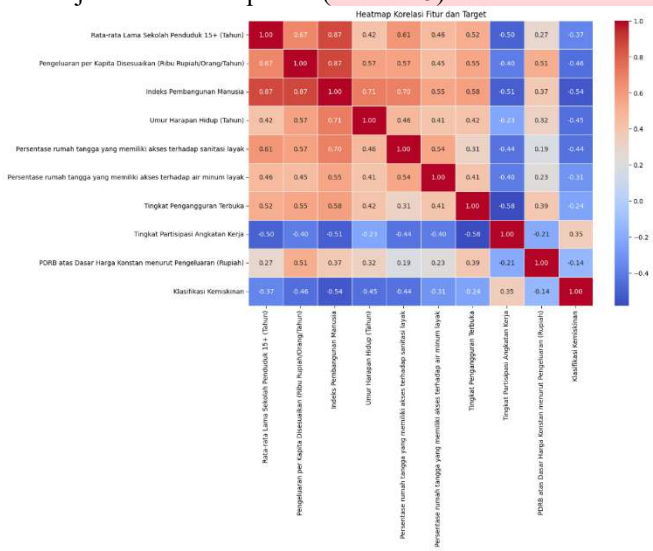
Kondisi ini berpotensi menyebabkan model bias ke kelas mayoritas, sehingga penelitian menerapkan strategi

penanganan *imbalanced* menggunakan *SMOTE* pada data latih saja untuk menghindari *data leakage*. Perubahan distribusi kelas data latih sebelum dan sesudah penerapan *SMOTE* ditunjukkan pada Gambar 4.



GAMBAR 4
(Distribusi kelas data latih sebelum dan sesudah penerapan *SMOTE*)

Selain itu, analisis korelasi digunakan untuk melihat hubungan antar fitur dan target, yang menunjukkan beberapa indikator kesejahteraan berkorelasi negatif terhadap kemiskinan seperti IPM, pengeluaran per kapita, umur harapan hidup, dan akses sanitasi layak, sedangkan TPAK menunjukkan korelasi positif (Gambar 5).



GAMBAR 5
(*Heatmap* korelasi Pearson antar fitur dan target klasifikasi kemiskinan)

Selanjutnya, dataset dibagi menjadi data latih dan data uji menggunakan *stratified split* agar proporsi kelas tetap terjaga, dengan empat skenario pembagian yaitu 90:10, 80:20, 70:30, dan 50:50 untuk menguji konsistensi performa model pada variasi ukuran data uji. Pada tahap pelatihan dan penalaan parameter, digunakan *StratifiedKFold* 5-*fold* dengan *shuffle* untuk memperoleh evaluasi yang lebih stabil. Tahap akhir *preprocessing* adalah standarisasi fitur menggunakan *StandardScaler* agar setiap variabel memiliki skala yang sebanding, terutama karena SVM sensitif terhadap perbedaan skala.

C. Hasil Eksperimen Skenario Split Data

Pada tahap ini, dataset dipisahkan menjadi variabel independen (X) dan variabel dependen (y), kemudian dibagi menjadi data latih dan data uji menggunakan fungsi train_test_split dari pustaka *scikit-learn*. Hasil dari masing-masing skenario kemudian dibandingkan untuk menentukan konfigurasi pembagian data yang paling optimal dan digunakan sebagai model akhir.

TABEL 2
(SPLIT DATA 90%:10%)

No	Jenis	Presentase	Total
1	Training	90%	462
2	Testing	10%	52

TABEL 3
(SPLIT DATA 80%:20%)

No	Jenis	Presentase	Total
1	Training	80%	411
2	Testing	20%	103

TABEL 4
(SPLIT DATA 70%:30%)

No	Jenis	Presentase	Total
1	Training	70%	359
2	Testing	30%	155

TABEL 5
(SPLIT DATA 50%:50%)

No	Jenis	Presentase	Total
1	Training	50%	257
2	Testing	50%	257

Berdasarkan Tabel 2 sampai Tabel 5, pembagian data menghasilkan proporsi data latih dan data uji yang berbeda pada setiap skenario. Variasi proporsi ini memungkinkan analisis lebih lanjut mengenai pengaruh jumlah data latih dan data uji terhadap performa SVM dan RF, khususnya dalam mendeteksi kelas minoritas (miskin). Hasil evaluasi dari tiap skenario kemudian dibandingkan untuk menentukan skenario pembagian data yang paling stabil dan optimal untuk digunakan pada tahap pemodelan serta penetapan model akhir.

D. Hyperparameter Tuning

Berdasarkan hasil penalaan parameter (*hyperparameter tuning*), diperoleh kombinasi parameter terbaik (*best parameters*) untuk masing-masing algoritma pada setiap skenario pembagian data. Parameter optimal ini merepresentasikan konfigurasi yang paling sesuai dalam mempelajari pola pada data latih sekaligus menjaga keseimbangan antara risiko *overfitting* dan *underfitting*. Proses penalaan dilakukan menggunakan *GridSearchCV* dengan *Stratified K-Fold cross-validation* 5-*fold* agar proporsi kelas tetap terjaga pada setiap lipatan, mengingat dataset bersifat *imbalanced*. Setelah parameter terbaik diperoleh, langkah berikutnya adalah membangun model final dengan melakukan pelatihan ulang (*retraining*) pada keseluruhan data latih (*full training set*) menggunakan parameter terpilih tersebut, baik pada pendekatan *baseline* (tanpa penanganan ketidakseimbangan) maupun pendekatan dengan *SMOTE*. Model final inilah yang kemudian digunakan untuk memprediksi data uji sehingga kinerja klasifikasi kemiskinan dapat dievaluasi secara objektif menggunakan metrik evaluasi yang telah ditetapkan.

Secara lebih spesifik, hasil optimasi menunjukkan karakteristik konfigurasi yang berbeda pada kedua algoritma. Pada SVM, model cenderung memilih kernel *linear* atau RBF dengan nilai parameter regularisasi C yang bervariasi antar skenario. Variasi nilai Cmencerminkan penyesuaian tingkat penalti kesalahan dan lebar margin keputusan, terutama ketika distribusi data latih berubah akibat penerapan *SMOTE*. Sementara itu, pada RF, konfigurasi terbaik relatif stabil, di mana model sering memilih kedalaman pohon yang tidak

dibatasi (*max_depth = None*) pada beberapa skenario. Pilihan ini mengindikasikan bahwa RF berupaya menangkap kompleksitas pola data secara lebih rinci, sementara mekanisme *bagging* membantu menjaga stabilitas dan mengurangi varians model.

E. Hasil Model *Support Vector Machine*

Pada pendekatan baseline, SVM mencapai performa terbaik di skenario 90:10 (F1 0,80; Akurasi 0,96). Namun, skenario 70:30 dinilai lebih representatif untuk generalisasi data dengan kinerja yang tetap konsisten (F1 0,73), meskipun PR-AUC lebih rendah. Sebaliknya, penerapan SMOTE cenderung menurunkan F1-score di berbagai skenario. Walaupun SMOTE pada skenario 70:30 mampu meningkatkan deteksi kelas minoritas secara signifikan (PR-AUC 0,83; ROC-AUC 0,97), F1-score yang dihasilkan (0,58) tetap lebih rendah dibanding baseline. Secara keseluruhan, pendekatan baseline terbukti lebih stabil dibandingkan SMOTE dalam menjaga keseimbangan presisi dan recall pada dataset ini.

TABEL 5
(HASIL EVALUASI *SUPPORT VECTOR MACHINE*)

Scenario	Approach	Accuracy	F1	ROC	PR
90:10	Baseline	0,96	0,80	0,92	0,81
90:10	SMOTE	0,84	0,60	0,96	0,79
80:20	Baseline	0,93	0,66	0,95	0,69
80:20	SMOTE	0,82	0,57	0,92	0,48
70:30	Baseline	0,94	0,72	0,90	0,66
70:30	SMOTE	0,82	0,58	0,97	0,83
50:50	Baseline	0,93	0,69	0,91	0,77
50:50	SMOTE	0,83	0,58	0,94	0,80

F. Hasil Model Random Forest

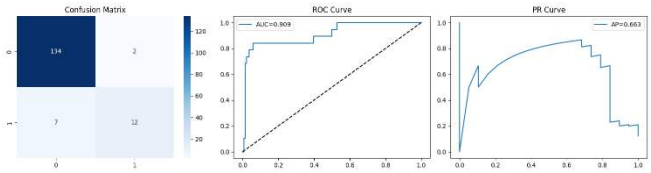
Evaluasi Random Forest (RF) menunjukkan bahwa pendekatan *baseline* lebih unggul dan stabil, mencatatkan F1-score berkisar 0,6786–0,7273. Performa terbaik dicapai pada skenario 90:10 dan 80:20 (F1 0,7273), didukung oleh nilai ROC-AUC (>0,96) dan PR-AUC (>0,86) yang tinggi. Sebaliknya, penerapan SMOTE cenderung menurunkan stabilitas model; meskipun sempat meningkatkan PR-AUC menjadi 0,8965 pada skenario 90:10, nilai F1-score justru turun menjadi 0,6000. Penurunan serupa terjadi pada skenario 70:30 (F1 0,5538 vs *baseline* 0,7222), menegaskan bahwa pendekatan *baseline* lebih efektif dalam menjaga keseimbangan presisi dan *recall* untuk dataset ini.

TABEL 6
(HASIL EVALUASI *RANDOM FOREST*)

Scenario	Approach	Accuracy	F1	ROC	PR
90:10	Baseline	0,94	0,72	0,96	0,86
90:10	SMOTE	0,84	0,60	0,97	0,89
80:20	Baseline	0,94	0,72	0,97	0,76
80:20	SMOTE	0,85	0,61	0,96	0,71
70:30	Baseline	0,93	0,72	0,97	0,86
70:30	SMOTE	0,81	0,55	0,96	0,78
50:50	Baseline	0,92	0,67	0,92	0,74
50:50	SMOTE	0,86	0,63	0,94	0,73

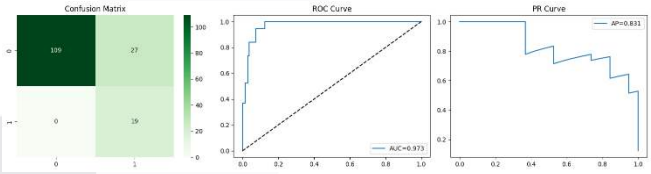
G. Analisis *Confusion Matrix*, ROC *Curve*, dan *PR Curve* pada *model Support Vector Machine*

Pada tahap ini dilakukan analisis terhadap performa model *Support Vector Machine* (SVM) pada skenario pembagian data 70:30, baik pada pendekatan *Baseline* maupun SMOTE. Analisis difokuskan pada *confusion matrix*untuk melihat jenis kesalahan klasifikasi yang terjadi, serta diperkuat dengan *ROC curve* dan *PR curve* untuk mengevaluasi kemampuan pemisahan kelas pada kondisi data tidak seimbang.



GAMBAR 6
(*Confusion Matrix*, Kurva ROC, dan Kurva *PR SVM Baseline*)

Berdasarkan *confusion matrix* pada model SVM *baseline*, prediksi yang benar untuk kelas 0 (tidak miskin) tercatat sebanyak 134 data sebagai *true negative* (TN), sedangkan terdapat 2 data kelas 0 yang keliru diprediksi sebagai kelas 1 (*false positive*/FP). Untuk kelas 1 (miskin), model mampu mengklasifikasikan dengan tepat sebanyak 12 data sebagai *true positive* (TP), namun masih terdapat 7 data miskin yang diprediksi sebagai tidak miskin (*false negative*/FN). Keberadaan FN ini perlu mendapat perhatian karena menunjukkan adanya wilayah yang seharusnya termasuk kategori miskin tetapi tidak teridentifikasi oleh model.Dari sisi evaluasi kurva, nilai ROC-AUC pada SVM *baseline* sebesar 0,909 mengindikasikan kemampuan pemisahan kelas yang cukup baik. Sementara itu, nilai PR-AUC (*average precision*) sebesar 0,663 menunjukkan performa dalam mendeteksi kelas minoritas masih berada pada tingkat moderat, sehingga masih terdapat ruang untuk meningkatkan ketepatan deteksi wilayah miskin.



GAMBAR 7
(*Confusion Matrix*, Kurva ROC, dan Kurva *PR SVM SMOTE*)

Pada pendekatan *SMOTE*, hasil *confusion matrix* menunjukkan bahwa model memprediksi benar kelas 0 (tidak miskin) sebanyak 109 data sebagai TN, namun kesalahan prediksi kelas 0 menjadi kelas 1 meningkat menjadi 27 data (FP). Untuk kelas 1 (miskin), model berhasil mengklasifikasikan dengan tepat sebanyak 19 data sebagai TP dan tidak ditemukan kesalahan FN (0 data). Temuan ini mengindikasikan bahwa penerapan SMOTE pada skenario 70:30 mampu meningkatkan kemampuan model dalam menangkap seluruh data miskin pada data uji (FN = 0), tetapi dengan konsekuensi meningkatnya jumlah wilayah yang sebenarnya tidak miskin namun diprediksi miskin (FP tinggi).

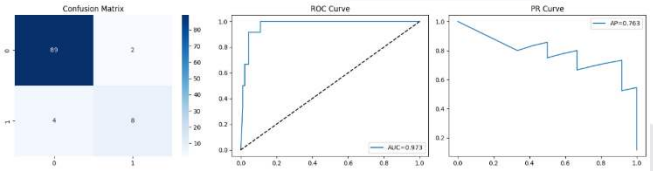
Ditinjau dari kurva evaluasi, nilai ROC-AUC pada pendekatan *SMOTE* mencapai 0,973, lebih tinggi dibanding *baseline*, yang menunjukkan kemampuan pemisahan kelas sangat baik. Selain itu, nilai PR-AUC sebesar 0,831 juga

lebih tinggi daripada *baseline*, menandakan peningkatan kemampuan model dalam mendeteksi kelas minoritas. Namun, kenaikan FP yang cukup besar memperlihatkan adanya *trade-off*: model menjadi lebih sensitif terhadap kelas miskin sehingga lebih banyak wilayah tidak miskin ikut terklasifikasi sebagai miskin.

Secara keseluruhan pada skenario 70:30,pendekatan *baseline* menghasilkan FP yang sangat rendah (2 data), tetapi masih menyisakan FN (7 data) sehingga terdapat wilayah miskin yang belum terdeteksi. Sebaliknya, pendekatan *SMOTE* menghilangkan FN (0 data) sehingga seluruh wilayah miskin teridentifikasi, tetapi meningkatkan FP menjadi 27 data yang berarti lebih banyak wilayah tidak miskin salah dikategorikan miskin. Oleh karena itu, pemilihan pendekatan dapat disesuaikan dengan tujuan: apabila prioritasnya meminimalkan FN, maka *SMOTE* lebih unggul; sedangkan jika prioritasnya menekan FP agar ketepatan sasaran lebih terjaga, maka *baseline* lebih sesuai. Pada penelitian ini, pemilihan model akhir tetap mempertimbangkan metrik F1-score pada data uji serta keseimbangan antara presisi dan recall.

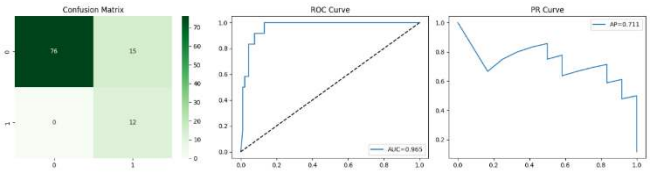
H. Analisis *Confusion Matrix*, ROC *Curve*, dan *PR Curve* pada *model Random Forest*

Berdasarkan *confusion matrix* pada RF *Baseline*, model memprediksi benar kelas 0 (tidak miskin) sebanyak 89 data (*true negative*), dan terdapat 2 data kelas 0 yang salah diprediksi sebagai kelas 1 (*false positive*). Pada kelas 1 (miskin), model berhasil memprediksi benar sebanyak 8 data (*true positive*/TP), namun masih terdapat 4 data miskin yang diprediksi tidak miskin (*false negative*/FN). Kesalahan *FN* menunjukkan masih ada beberapa wilayah miskin yang belum terdeteksi, namun kesalahan *FP* relatif kecil sehingga model tidak terlalu banyak mengklasifikasikan wilayah tidak miskin sebagai miskin.



GAMBAR 8
(*Confusion Matrix*, Kurva ROC, dan Kurva *PR RF Baseline*)

Dari sisi evaluasi kurva, nilai *ROC-AUC* sebesar 0,973 menunjukkan kemampuan pemisahan kelas yang sangat baik. Sementara itu, nilai *PR-AUC* (*Average Precision*) sebesar 0,763 menandakan bahwa performa model cukup baik dalam mendeteksi kelas minoritas, meskipun masih terdapat *FN* pada hasil klasifikasi.



GAMBAR 9
(*Confusion Matrix*, Kurva ROC, dan Kurva *PR RF SMOTE*)

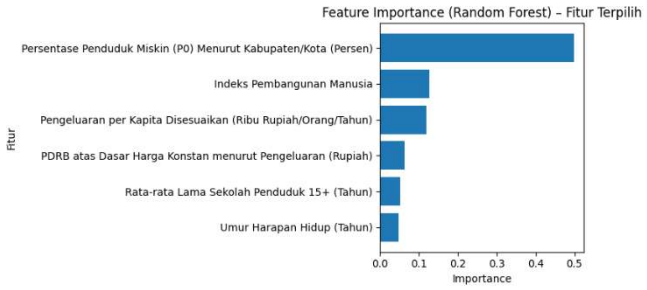
Pada pendekatan SMOTE, *confusion matrix* menunjukkan model memprediksi benar kelas 0 (tidak miskin) sebanyak 76 data (*TN*), sedangkan kesalahan *FP* meningkat menjadi 15 data. Untuk kelas 1 (miskin), model memprediksi benar sebanyak 12 data (*TP*) dan tidak terdapat kesalahan *FN* (0 data). Hasil ini menunjukkan bahwa penerapan SMOTE membuat model lebih sensitif terhadap kelas miskin sehingga seluruh data miskin pada data uji berhasil terdeteksi (*FN* = 0). Namun demikian, peningkatan *FP* menunjukkan adanya konsekuensi berupa lebih banyak wilayah yang seharusnya tidak miskin tetapi diprediksi miskin.

Nilai ROC-AUC pada pendekatan *SMOTE* sebesar 0,965 masih menunjukkan kemampuan pemisahan kelas yang sangat baik, meskipun sedikit lebih rendah dibanding *baseline*. Sementara itu, PR-AUC sebesar 0,711 lebih rendah daripada *baseline*, yang mengindikasikan bahwa peningkatan *recall* pada kelas miskin terjadi bersamaan dengan penurunan *precision* akibat bertambahnya jumlah FP.

Secara keseluruhan pada skenario 80:20, RF *baseline* menghasilkan FP yang kecil (2 data) namun masih menyisakan FN (4 data). Sebaliknya, RF + *SMOTE* mampu menurunkan FN hingga 0 sehingga seluruh kelas miskin terdeteksi, tetapi konsekuensinya FP meningkat menjadi 15 data. Temuan ini menegaskan adanya *trade-off* antara upaya memaksimalkan deteksi kelas miskin dan risiko salah klasifikasi pada kelas tidak miskin. Dalam penelitian ini, pemilihan pendekatan terbaik tetap mempertimbangkan keseimbangan performa yang tercermin pada metrik F1-score dan PR-AUC pada data uji, sehingga pendekatan *baseline* dinilai lebih stabil untuk dataset yang digunakan.

I. Hasil Analisis Variabel Dominan (*Feature Importance*)

Analisis ini bertujuan untuk mengidentifikasi indikator yang paling berpengaruh dalam klasifikasi status kemiskinan kabupaten/kota. Pengukuran dilakukan menggunakan nilai *feature importance* dari algoritma *Random Forest*, yang merepresentasikan kontribusi setiap fitur dalam menurunkan tingkat ketidakmurnian (*impurity*) saat pembentukan pohon keputusan. Dengan demikian, dapat diketahui variabel yang paling dominan dalam membedakan kelas “Miskin” dan “Tidak Miskin”. Semakin tinggi nilai *importance* suatu fitur, semakin besar pengaruhnya terhadap keputusan model. Peringkat fitur berdasarkan nilai kepentingannya ditampilkan pada Gambar 10.



GAMBAR 10
(Hasil *Feature Importance Random Forest*)

V. KESIMPULAN

Berdasarkan hasil pengujian, algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF) dapat diterapkan secara efektif untuk klasifikasi kemiskinan kabupaten/kota di

Indonesia pada kondisi data tidak seimbang. Model yang dipilih sebagai konfigurasi terbaik adalah SVM *baseline* pada skenario *split* 70:30 serta RF *baseline* pada skenario *split* 80:20, karena memberikan performa yang relatif stabil pada data uji. Secara umum, kedua model menunjukkan kinerja tinggi dengan akurasi di atas 94%, namun penilaian utama difokuskan pada F1-score dan PR-AUC agar lebih representatif terhadap kelas minoritas. Hasil evaluasi menunjukkan RF *baseline* (80:20) memiliki PR-AUC 0,7628, lebih tinggi dibanding SVM *baseline* (70:30) sebesar 0,6629, sehingga RF dinilai lebih baik dalam menjaga keseimbangan presisi–recall pada data uji. Selain itu, analisis *feature importance* mengindikasikan bahwa indikator ekonomi dan kualitas hidup merupakan faktor dominan dalam membedakan status kemiskinan wilayah, dengan kontribusi terbesar berasal dari pengeluaran per kapita disesuaikan, IPM, dan rata-rata lama sekolah.

REFERENSI

[1] Badan Pusat Statistik, "Profil Kemiskinan di Indonesia September 2024," Jan. 2025.

[2] A. Rodiyah, Destifa Prastiti Kusuma, Nia Bunga Ramadani, Rayya Malik Ibrahim, Rian Ahmadul Huda, and Anggi Rahajeng, "Pengaruh Rata Rata Lama Sekolah, Garis Kemiskinan, dan Usia Harapan Hidup terhadap IPM di Provinsi Jawa Tengah," *JURNAL ILMIAH EKONOMI DAN MANAJEMEN*, vol. 3, no. 5, pp. 345–357, May 2025.

[3] Kurnia Mauludiyah, "Klasifikasi Indeks Pembangunanan Manusia Kabupaten/Kota di Indonesia Menggunakan Metode Random Forest," Semarang, 2020. [Online]. Available: http://repository.unimus.ac.id

[4] S. Putri Wardyana and P. Setiawan Prabowo, "Analisis Pengeluaran Perkapita dalam Upaya Peningkatan Pembangunan Manusia di Sulawesi Tenggara pada Tahun 2021," 2022.

[5] A. Ananta, Y. Sabilu, Z. Hikmawati, J. Kesehatan, K. Masyarakat, and U. H. Oleo, "Gambaran Sanitasi Dasar Pada Masyarakat Bajo di Pulau Masudu Desa Terapung Kecamatan Poleang Tenggara Kabupaten Bombana," *JKL-UHO*, vol. 6, no. 2, 2025.

[6] A. Dwi Satria and A. Habibi, "Pengaruh Sektor Produk Domestik Regional Bruto (Basis dan Non Basis) Terhadap Pertumbuhan Ekonomi dalam Perspektif Ekonomi Islam," 2023, [Online].

[7] M. Faisal and I. Budi Santoso, "Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE) untuk Deteksi Diabetes," MEI, 2025.

[8] F. Elfaladonna, D. Sartika, and A. M. Putra, "Exploratory Data Analysis on the Process of Determining the Relationship between Student Interest and Talent Variables," *Jurnal Sains, Teknologi dan Industri*, vol. 21, no. 2, pp. 418–424, Jun. 2024.

[9] I. Kurniawan, D. Cahya Putri Buani, W. Apriliah, R. Amegia Saputra, and P. Korespondensi, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin," vol. 10, no. 2, pp. 421–428, 2023.

[10] A. Ernawati and Z. Sitorus, "Penerapan Data Mining Untuk Klasifikasi Penduduk Miskin Di Kabupaten Labuhanbatu Menggunakan Random Forest (RF) Dan K-Nearest Neighbors (KNN)," *Bulletin of Information Technology (BIT)*, vol. 6, no. 1, pp. 23–29, 2025.

[11] M. Samantri, "Perbandingan Algoritma Support Vector Machine dan Random Forest untuk Analisis Sentimen Terhadap Kebijakan Pemerintah Indonesia Terkait Kenaikan Harga BBM Tahun 2022," *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 1, p. 2024, 2024, [Online].

[12] A. Sabat-Tomala, E. Raczko, and B. Zagajewski, "Comparison of Support Vector Machine and Random Forest Algorithms for Invasive and Expansive Species Classification Using Airborne Hyperspectral Data," *Remote Sens. (Basel).*, vol. 12, no. 3, Feb. 2020.

[13] J. Kause and F. Fithriyah, "Analisis Determinan Kemiskinan Multidimensi di Indonesia," *Jurnal Riset Ilmu Ekonomi*, vol. 4, no. 2, pp. 115–127, Aug. 2024.

[14] R. A. Putri and A. Jatnika Effendi, "Analisis Keberlanjutan Sistem Penyediaan Air Minum Berbasis Masyarakat pada Program PAMSIMAS di Kecamatan Tanjungsari, Kabupaten Sumedang," vol. X, no. 4, 2025.

[15] I. D. Lestari and A. Nilasari, "Pengaruh Laju Pertumbuhan Penduduk, Tingkat Partisipasi Angkatan Kerja, Dan Upah Minimum Kabupaten/Kota Terhadap Tingkat Pengangguran Terbuka Di Kabupaten/Kota Provinsi Jawa Barat Tahun 2018 - 2022," *Jurnal Ekuilnomi*, vol. 7, no. 2, pp. 483–493, May 2025.

[16] Al Ghifari Ismar Muhamamd, "Hubungan Tingkat Partisipasi Angkatan Kerja (TPAK) dan Tingkat Pengangguran Terbuka (TPT) Wilayah Kab/Kota Provinsi Sumatera Barat Tahun 2022-2023," 2025.

[17] W. Kurniawan and U. Indahyanti, "Prediksi Angka Harapan Hidup Penduduk Menggunakan Metode XGBoost," *Indonesian Journal of Applied Technology*, vol. 1, no. 2, p. 18, Jul. 2024.

[18] V. R. Saragih, A. Arnita, Z. Indra, I. Taufik, and M. S. Sinaga, "Comparison of Supervised Machine Learning Methods in Predicting the Prevalence of Stunting in North Sumatra province," *Journal of Soft Computing Exploration*, vol. 5, no. 4, pp. 370–379, Dec. 2024.

[19] S. NURADILLA, K. SADIK, C. SUHAENI, and A. M. SOLEH, "Klasifikasi Halaman SEO Berbasis Machine Learning Melalui Mutual Information dan Random Forest Feature Importance," *MIND Journal*, vol. 10, no. 1, pp. 114–129, Jun. 2025.

[20] J. B. Wang, C. A. Zou, and G. H. Fu, "AWSMOTE: An SVM-Based Adaptive Weighted SMOTE for Class-Imbalance Learning," *Sci. Program.*, vol. 2021, 2021.

[21] Indra Griha Tofik Isa and Febie Elfaladonna, "Penilaian Kinerja Akurasi Metode Klasifikasi dalam Dataset Penerimaan Mahasiswa Baru Universitas

XYZ," *Jurnal Edukasi dan Penelitian Informatika*, Aug. 2022.