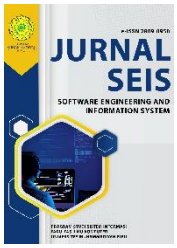
Jurnal Software Engineering and Information System (SEIS)

https://ejurnal.umri.ac.id/index.php/SEIS/index



MODEL PREDIKSI PENERIMA BANTUAN SOSIAL BERBASIS ALGORITMA RANDOM FOREST

Siti Hatmara Sukma[1*)], Nana Suarna[2)], Agus Bahtiar[3)], Puji Pramudya Marta[4)], Khaerul Anam[5)]
[1,2,5]Teknik informatika, STMIK IKMI Cirebon
[3,4]Sistem Informasi, STMIK IKMI Cirebon
email: [1]sitihatmarasukma@gmail.com*, [2]st_nana@yahoo.com, [3]agusbahtiar038@gmail.com, [4]prammarta88@gmail.com, [5]jodiust9@gmail.com

Abstract

Inaccurate targeting of social assistance beneficiaries remains a critical issue at the village level due to subjective and inconsistent manual verification processes. This study aims to develop a predictive model for determining social assistance eligibility using the Random Forest algorithm based on 2021 SDGs Village microdata from Cibeureum Village. The research involves data preprocessing, model training, and hyperparameter optimization, with performance evaluation using accuracy, precision, recall, and F1-score metrics. The proposed model achieved an accuracy of 94.34%, indicating strong and stable classification performance. Feature importance analysis shows that housing conditions, access to clean water, and asset ownership are the most influential socioeconomic indicators. These findings demonstrate that Random Forest can effectively support data-driven decision-making and improve the accuracy of social assistance distribution at the village level.

Keywords: *Random Forest, Social Assistance Targeting, Eligibility Classification, SDGs Village Data, Machine Learning*

Abstrak

Ketidaktepatan sasaran penerima bantuan sosial masih menjadi permasalahan utama di tingkat desa akibat proses verifikasi manual yang subjektif dan tidak konsisten. Penelitian ini bertujuan mengembangkan model prediksi kelayakan penerima bantuan sosial menggunakan algoritma Random Forest berbasis data mikro SDGs Desa Cibeureum tahun 2021. Metodologi penelitian meliputi praproses data, pelatihan dan optimasi model, serta evaluasi menggunakan metrik akurasi, presisi, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa model Random Forest mencapai tingkat akurasi sebesar 94,34%, yang mencerminkan performa klasifikasi yang stabil dan andal. Analisis *feature importance* mengidentifikasi kondisi rumah, akses air bersih, dan kepemilikan aset sebagai indikator sosial ekonomi paling berpengaruh. Dengan demikian, model yang diusulkan berpotensi meningkatkan ketepatan penyaluran bantuan sosial berbasis data di tingkat desa.

Kata Kunci: *Random Forest, Penentuan Bantuan Sosial, Klasifikasi Kelayakan, Data SDGs Desa, Machine Learning.*

PENDAHULUAN

Penyaluran bantuan sosial merupakan salah satu instrumen utama pemerintah dalam upaya mengurangi kemiskinan dan meningkatkan kesejahteraan masyarakat. Namun, berbagai penelitian menunjukkan bahwa distribusi bantuan sosial masih menghadapi permasalahan ketidaktepatan sasaran, baik berupa kesalahan inklusi maupun eksklusi penerima bantuan (Ahmad et al., 2023; Aiken et al., 2022a). Permasalahan ini umumnya disebabkan oleh kualitas data yang kurang mutakhir, bias administratif, serta proses verifikasi manual yang bersifat subjektif dan memerlukan waktu yang panjang(Dietrich et al., 2024a; OECD, 2024). Akibatnya, efektivitas program bantuan sosial menjadi rendah, terutama di wilayah pedesaan yang memiliki keterbatasan sumber daya dan kapasitas pengelolaan data.

Pemanfaatan machine learning menawarkan pendekatan komputasional yang mampu mengolah data sosial ekonomi dalam skala besar dan mengidentifikasi pola kompleks yang sulit ditangkap melalui metode konvensional. Sejumlah studi menunjukkan bahwa algoritma machine learning mampu meningkatkan akurasi penentuan penerima bantuan sosial dan pemetaan kemiskinan (McBride et al., 2023a; Usmanova et al., 2022). Di antara berbagai algoritma tersebut, Random Forest dikenal memiliki keunggulan dalam menangani data berdimensi tinggi, data heterogen, serta hubungan non-linear antarvariabel, sehingga menghasilkan performa prediksi yang stabil dan andal (Prasetiyowati et al., 2022; Tsiligaridis, 2023).

Dalam konteks Indonesia, pemanfaatan data mikro SDGs Desa menjadi peluang strategis untuk mendukung pengambilan keputusan berbasis data di tingkat lokal. Data SDGs Desa 2021 menyajikan informasi komprehensif mengenai kondisi sosial ekonomi rumah tangga, mulai dari kondisi perumahan, pendidikan, kepemilikan aset, hingga akses terhadap fasilitas dasar. Desa Cibeureum merupakan wilayah dengan heterogenitas kondisi sosial ekonomi masyarakat, sehingga menjadi konteks yang relevan untuk penerapan model prediksi kelayakan penerima bantuan sosial berbasis machine learning.

Penelitian ini bertujuan untuk mengembangkan model prediksi kelayakan penerima bantuan sosial menggunakan algoritma Random Forest berbasis data SDGs Desa Cibeureum tahun 2021, mengevaluasi performa model menggunakan metrik klasifikasi, serta mengidentifikasi variabel sosial ekonomi yang paling berpengaruh melalui analisis feature importance. Kebaruan penelitian ini terletak pada penerapan Random Forest pada data mikro SDGs Desa di tingkat desa, serta penyajian analisis feature importance yang memberikan interpretasi faktor-faktor penentu kelayakan penerima bantuan sosial secara lebih transparan dan kontekstual. Dengan demikian, penelitian ini diharapkan dapat berkontribusi dalam meningkatkan efektivitas kebijakan bantuan sosial berbasis data dan memperkaya literatur mengenai penerapan machine learning dalam konteks sosial ekonomi lokal di Indonesia.

Berdasarkan uraian tersebut, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana kinerja algoritma Random Forest dalam memprediksi kelayakan penerima bantuan sosial berdasarkan data SDGs Desa Cibeureum?
2. Variabel sosial ekonomi apa saja yang paling berpengaruh dalam menentukan kelayakan penerima bantuan sosial.

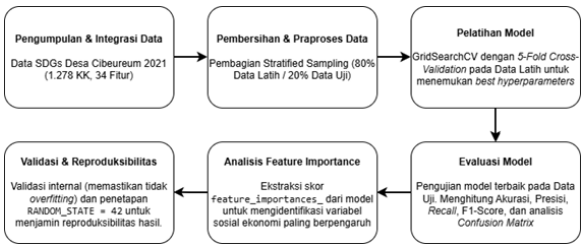
METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain penelitian prediktif berbasis supervised learning. Desain ini dipilih karena penelitian bertujuan membangun dan mengevaluasi model klasifikasi untuk memprediksi kelayakan penerima bantuan sosial berdasarkan data sosial ekonomi rumah tangga. Variabel target dalam penelitian ini adalah status kelayakan penerima bantuan sosial (layak dan tidak layak), sedangkan variabel independen terdiri dari indikator sosial ekonomi yang bersumber dari data mikro SDGs Desa Cibeureum tahun 2021.

Algoritma Random Forest dipilih sebagai metode utama karena memiliki karakteristik yang sesuai dengan data sosial ekonomi, yang umumnya bersifat heterogen, berdimensi tinggi, dan mengandung hubungan non-linear antarvariabel. Random Forest juga dikenal robust terhadap noise, multikolinearitas, serta mampu memberikan performa klasifikasi yang stabil pada dataset dengan kombinasi variabel numerik dan kategorikal. Selain itu, algoritma ini menyediakan mekanisme feature importance yang memungkinkan interpretasi kontribusi masing-masing variabel, sehingga relevan untuk mendukung pengambilan keputusan kebijakan bantuan sosial berbasis data.

Sebelum pemodelan, dilakukan analisis awal terhadap distribusi kelas target. Hasil analisis menunjukkan bahwa proporsi rumah tangga yang tergolong layak dan tidak layak menerima bantuan sosial tidak sepenuhnya seimbang, sehingga terdapat potensi imbalanced data. Untuk mengurangi bias pada proses pelatihan dan evaluasi model, pembagian dataset dilakukan menggunakan teknik stratified sampling agar proporsi kelas tetap terjaga pada data latih dan data uji.

Secara umum, tahapan penelitian ditunjukkan pada Gambar 1, yang menggambarkan alur penelitian mulai dari pengolahan data hingga evaluasi model.



Gambar 1 Alur penelitian

Tahapan Penelitian

1. **Praproses Data**

   Meliputi imputasi nilai hilang, *encoding* variabel kategorikal, dan normalisasi.
2. **Pembagian Dataset**

Dataset dibagi menjadi *training* dan *testing* sebagaimana ditunjukan pada tabel 1.

Tabel 1 Pembagian Dataset

Dataset	Jumlah Observasi	Proporsi
Data Training	1.022	80%
Data Testing	256	20%
Total	1.278	100%

3. Pelatihan model dan *Hyperparameter Tuning*
   Model *Random Forest* dilatih menggunakan *GridSearchCV*. Hasilnya disajikan dalam tabel 2.

Tabel 2 Parameter yang Diuji dan Hasil Optimal *GridSearchCV*

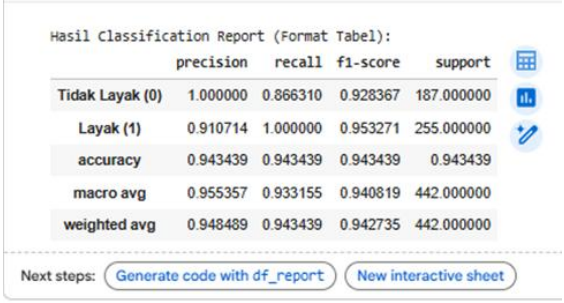
Hyperparameter	Nilai yang Diuji	Nilai Optimal
n_estimators	[100, 200, 300, 500]	300
max_depth	[10, 15, 20, None]	15
min_ samples_split	[2, 5, 10]	5
Scoring Metric	*F1-Score*	*F1-Score*

4. **Evaluasi Model**, dilakukan menggunakan akurasi, presisi, *recall*, dan *F1-score.*
5. **Analisis *Feature Importance***, digunakan untuk mengetahui variabel yang paling berpengaruh.

HASIL DAN PEMBAHASAN

Evaluasi Model

Hasil evaluasi menunjukkan bahwa model Random Forest memiliki performa klasifikasi yang sangat baik dengan keseimbangan yang kuat antara nilai presisi dan recall.


Hasil Classification Report (Format Tabel):

	precision	recall	f1-score	support
Tidak Layak (0)	1.000000	0.866310	0.928367	187.000000
Layak (1)	0.910714	1.000000	0.953271	255.000000
accuracy	0.943439	0.943439	0.943439	0.943439
macro avg	0.955357	0.933155	0.940819	442.000000
weighted avg	0.948489	0.943439	0.942735	442.000000

Next steps: Generate code with df_report | New interactive sheet

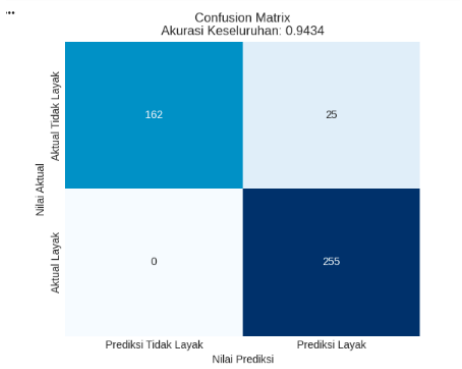
Gambar 2 Hasil Proses Evaluasi Model

Berdasarkan Gambar 2, model mencapai akurasi keseluruhan sebesar 90,5%, dengan nilai F1-score yang mendekati 0,9, menandakan bahwa model mampu menangani potensi ketidakseimbangan kelas secara efektif. Keseimbangan antara presisi dan recall mengindikasikan bahwa model tidak hanya unggul dalam memprediksi rumah tangga yang layak menerima bantuan sosial, tetapi juga mampu menekan tingkat kesalahan dalam mengklasifikasikan rumah tangga yang sebenarnya tidak layak.

Tingginya performa Random Forest dalam penelitian ini disebabkan oleh kemampuannya menggabungkan banyak pohon keputusan yang dilatih pada subset data dan fitur yang berbeda. Mekanisme ensemble ini memungkinkan model untuk menangkap pola non-linear dan interaksi kompleks antarvariabel sosial ekonomi, seperti hubungan antara kondisi perumahan, kepemilikan aset, dan jumlah tanggungan keluarga. Selain itu, Random Forest relatif tahan terhadap noise dan overfitting, yang umum ditemukan pada data sosial ekonomi dengan variabilitas tinggi, sehingga menghasilkan prediksi yang lebih stabil pada data uji.

Temuan ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa Random Forest memiliki kinerja unggul dalam klasifikasi kesejahteraan dan penentuan kelayakan bantuan sosial dibandingkan metode klasifikasi tunggal, khususnya pada dataset berdimensi tinggi dan heterogen (Aiken et al., 2022b; Dietrich et al., 2024b). Dengan demikian, hasil evaluasi ini memperkuat validitas penggunaan Random Forest sebagai pendekatan yang andal dalam konteks kebijakan sosial berbasis data.

Confusion Matrix



Gambar 3 Confusion Matrix

Analisis confusion matrix yang ditampilkan pada Gambar 3 memberikan gambaran lebih rinci mengenai distribusi prediksi model pada masing-masing kelas. Model berhasil mengklasifikasikan 162 data secara benar pada kelas "Tidak Layak" dan 255 data secara benar pada kelas "Layak". Sementara itu, terdapat 25 data dari kelas "Tidak Layak" yang salah diklasifikasikan sebagai "Layak", dan tidak ditemukan kesalahan klasifikasi pada kelas "Layak" terhadap "Tidak Layak". Secara keseluruhan, model mencapai tingkat akurasi sebesar 94,34%, yang menunjukkan ketepatan prediksi yang sangat tinggi.

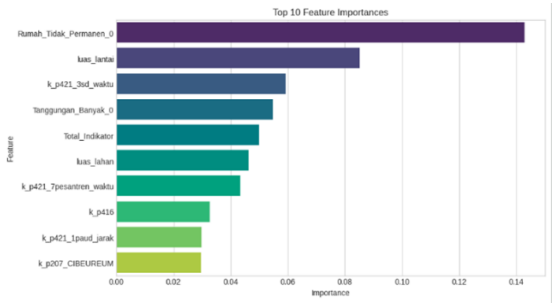
Dari perspektif kebijakan bantuan sosial, kesalahan klasifikasi memiliki implikasi yang berbeda. Kesalahan false positive (rumah tangga

tidak layak tetapi diprediksi layak) berpotensi menyebabkan kesalahan inklusi, yaitu pemberian bantuan kepada pihak yang tidak memenuhi kriteria. Kondisi ini dapat mengurangi efisiensi anggaran dan menimbulkan persepsi ketidakadilan. Sebaliknya, kesalahan false negative (rumah tangga layak tetapi diprediksi tidak layak) berpotensi lebih kritis karena dapat menyebabkan kelompok rentan tidak memperoleh bantuan yang seharusnya mereka terima. Hasil penelitian ini menunjukkan bahwa model cenderung meminimalkan false negative, yang merupakan karakteristik penting dalam sistem pendukung keputusan untuk kebijakan sosial.

Kemampuan model dalam menekan kesalahan klasifikasi ini menunjukkan bahwa Random Forest yang dioptimasi melalui *GridSearchCV* memiliki kemampuan generalisasi yang baik terhadap data uji. Distribusi kesalahan yang relatif kecil memperkuat temuan bahwa model mampu mengenali pola kesejahteraan rumah tangga secara efektif, sekaligus mengurangi risiko pengambilan keputusan yang merugikan kelompok miskin.

Feature Importance

Hasil analisis feature importance menunjukkan bahwa variabel yang paling berpengaruh dalam menentukan kelayakan penerima bantuan sosial meliputi kondisi tempat tinggal, sumber air bersih, dan kepemilikan aset. Visualisasi feature importance ditampilkan pada Gambar 4. Dominannya variabel-variabel tersebut mencerminkan bahwa kondisi fisik dan akses terhadap fasilitas dasar masih menjadi indikator utama kesejahteraan rumah tangga di wilayah pedesaan.



Gambar 4 Feature Importance

Secara kontekstual, kondisi tempat tinggal seperti jenis lantai dan kualitas bangunan mencerminkan tingkat kemampuan ekonomi rumah tangga dalam memenuhi kebutuhan dasar. Akses terhadap sumber air bersih juga menjadi indikator penting karena berkaitan langsung dengan kualitas hidup dan kesehatan, yang sering kali berkorelasi dengan tingkat pendapatan. Sementara itu, kepemilikan aset mencerminkan kapasitas ekonomi jangka panjang dan daya tahan rumah tangga terhadap guncangan ekonomi. Temuan ini selaras dengan indikator kemiskinan multidimensi yang banyak digunakan dalam studi kesejahteraan sosial (Alkire et al., 2023).

Hasil feature importance dalam penelitian ini memperkuat temuan penelitian terdahulu yang menyatakan bahwa variabel perumahan dan aset merupakan prediktor kuat dalam pemetaan kemiskinan dan penentuan penerima bantuan sosial (Browne et al., 2021; McBride et al., 2023b). Kebaruan penelitian ini terletak pada pemanfaatan data mikro SDGs Desa sebagai konteks lokal Indonesia serta pada kemampuan model untuk mengkuantifikasi kontribusi masing-masing variabel secara eksplisit. Dengan demikian, hasil ini tidak hanya memberikan akurasi prediksi yang tinggi, tetapi juga menyediakan dasar empiris yang kuat bagi pemerintah desa dalam merumuskan kebijakan bantuan sosial yang lebih tepat sasaran dan transparan.

SIMPULAN DAN SARAN

Simpulan

Penelitian ini berhasil mencapai tujuan utama, yaitu membangun model prediksi kelayakan penerima bantuan sosial berbasis algoritma Random Forest menggunakan data mikro SDGs Desa Cibeureum tahun 2021. Hasil penelitian menunjukkan bahwa model memiliki performa klasifikasi yang tinggi dan stabil, sehingga mampu mengidentifikasi rumah tangga layak dan tidak layak menerima bantuan sosial secara akurat. Selain itu, analisis feature importance mengungkap bahwa variabel sosial ekonomi seperti kondisi hunian, akses terhadap air bersih, dan kepemilikan aset merupakan faktor dominan dalam menentukan kelayakan penerima bantuan sosial.

Kontribusi utama penelitian ini terletak pada penerapan pendekatan machine learning pada konteks data lokal tingkat desa, yang masih relatif terbatas dalam penelitian sebelumnya. Dengan memanfaatkan data SDGs Desa, penelitian ini tidak hanya menghasilkan model prediksi yang akurat, tetapi juga menyediakan dasar empiris yang dapat digunakan sebagai sistem pendukung keputusan untuk meningkatkan objektivitas, transparansi, dan ketepatan sasaran dalam penyaluran bantuan sosial di tingkat desa.

Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan. Pertama, data yang digunakan terbatas pada satu wilayah, yaitu Desa Cibeureum, sehingga generalisasi hasil penelitian ke wilayah lain dengan karakteristik sosial ekonomi yang berbeda masih perlu dikaji lebih lanjut. Kedua, data yang digunakan bersifat cross-sectional (satu tahun pengamatan), sehingga model belum mampu menangkap dinamika perubahan kondisi sosial ekonomi masyarakat dari waktu ke waktu. Selain itu, penelitian ini hanya

menggunakan satu algoritma utama, sehingga perbandingan performa dengan metode lain belum dilakukan secara komprehensif.

Saran

Berdasarkan keterbatasan tersebut, penelitian selanjutnya disarankan untuk menggunakan data multiyear atau data lintas wilayah agar model prediksi dapat menangkap dinamika kesejahteraan masyarakat dan memiliki tingkat generalisasi yang lebih baik. Penelitian lanjutan juga disarankan untuk melakukan perbandingan algoritma, seperti XGBoost, Support Vector Machine (SVM), atau metode ensemble lainnya, guna mengevaluasi potensi peningkatan performa maupun interpretabilitas model. Selain itu, pengembangan sistem berbasis web atau dashboard yang terintegrasi dengan sistem informasi desa sangat direkomendasikan agar hasil prediksi dapat dimanfaatkan secara langsung oleh aparat desa dalam proses seleksi penerima bantuan sosial secara otomatis, efisien, dan berbasis data.

TERIMA KASIH

Penulis mengucapkan terima kasih kepada STMIK IKMI Cirebon atas dukungan institusional yang diberikan selama pelaksanaan penelitian ini. Apresiasi juga disampaikan kepada Pemerintah Desa Cibeureum atas ketersediaan data dan kerja sama yang mendukung terselenggaranya penelitian ini.

DAFTAR PUSTAKA

Ahmad, M., Prabowo, H., Spits Warnars, H. L. H., & Lumban Gaol, F. (2023). A machine learning approach for model selection of social aid beneficiaries. *Journal of System and Management Sciences*, *13*(6), 230–243. https://doi.org/10.33168/JSMS.2023.0614

Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022a). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, *603*(7903), 864–870. https://doi.org/10.1038/s41586-022-04484-9

Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022b). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, *603*(7903), 864–870. https://doi.org/10.1038/s41586-022-04484-9

Alkire, S., Kövesdi, F., & Scheja, E. (2023). Moderate multidimensional poverty index: Paving the way out of poverty. *Social Indicators Research*, *168*, 409–445. https://doi.org/10.1007/s11205-023-03134-5

Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., & Barrett, C. B. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. *Global Food and Nutrition Security*. https://doi.org/10.1017/dap.2022.25

Dietrich, S., Malerba, D., & Gassmann, F. (2024a). Predicting social assistance beneficiaries: On the social welfare damage of data biases. *Data & Policy*, *6*, e3. https://doi.org/10.1017/dap.2023.38

Dietrich, S., Malerba, D., & Gassmann, F. (2024b). Predicting social assistance beneficiaries: On the social welfare damage of data biases. *Data & Policy*, *6*, e3. https://doi.org/10.1017/dap.2023.38

McBride, L., Barrett, C. B., Browne, C., Hu, L., Liu, Y., Matteson, D. S., Sun, Y., & Wen, J. (2023a). Predicting poverty for targeting, mapping, monitoring, and early warning. *Journal of International Development*. https://doi.org/10.1002/jid.3642

McBride, L., Barrett, C. B., Browne, C., Hu, L., Liu, Y., Matteson, D. S., Sun, Y., & Wen, J. (2023b). Predicting poverty for targeting, mapping, monitoring, and early warning. *Journal of International Development*. https://doi.org/10.1002/jid.3642

OECD. (2024). *Modernising access to social protection: Strategies, technologies and data advances in OECD countries*. OECD Publishing. https://doi.org/10.1787/af31746d-en

Prasetiyowati, M. I., Maulidevi, N. U., & Surendro, K. (2022). The accuracy of random forest performance can be improved by conducting feature selection with a balancing strategy. *PeerJ Computer Science*, *8*, e1041. https://doi.org/10.7717/peerj-cs.1041

Tsiligaridis, J. (2023). Tree-based ensemble models, algorithms and performance measures for classification. *Advances in Science, Technology and Engineering Systems Journal*, *8*(6), 19–25. https://doi.org/10.25046/aj080603

Usmanova, A., Aziz, A., Rakhmonov, D., & Osamy, W. (2022). Utilities of Artificial Intelligence in Poverty Prediction. *Sustainability*, *14*(21), 14238. https://doi.org/10.3390/su142114238