UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 3 Nomor 4 Tahun 2025 Halaman 507-517.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Classification of Recipients of the Family Hope Program in West Sumatra Province Using the Random Forest Algorithm

Nini Erdiani, Dwi Sulistiowati*, Nonong Amalita, Zamahsary Martha
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dwisulistiowati@fmipa.unp.ac.id

Submitted : 09 Oktober 2025
Revised : 04 November 2025
Accepted : 01 Desember 2025

ABSTRACT

According to the Central Statistics Agency (BPS), the percentage of poor people in West Sumatra Province increased by 0.02% in 2024. One of the government's efforts to overcome poverty is a social assistance program issued by the government to help people who are economically disadvantaged. The targeted distribution of social assistance is an important challenge in improving community welfare, especially for families receiving PKH benefits. This study aims to classify households receiving the Family Hope Program (PKH) in West Sumatra Province using a random forest algorithm with Synthetic Minority Oversampling Technique (SMOTE). This study uses data on PKH recipient households in West Sumatra Province in 2024, which has a significant class imbalance. Therefore, the SMOTE method was applied to balance the data. The data was divided into training and testing data with a ratio of 80%:20%, then parameter tuning was performed to optimize mtry and ntree. The model was evaluated using a confusion matrix to compare model performance. The results show that the accuracy obtained is 74%. The precision value is 71%, the recall is 82%, and the f1-score is 76%. Based on the Mean Decrease Gini value, the head of household's diploma became the main attribute in determining whether a household received PKH or not. This study concluded that the use of SMOTE in the random forest algorithm performed well in classifying PKH recipients in West Sumatra Province, where the model performed well and was quite reliable in identifying PKH recipients.

Keywords: *Family Hope Program, Random Forest, Synthetic Minority Oversampling Technique*

I. PENDAHULUAN

Salah satu tantangan yang masih dihadapi oleh negara Indonesia ialah masalah kemiskinan. Permasalahan ini melibatkan berbagai aspek, sehingga menjadi fokus utama dalam upaya pembangunan. Hingga saat ini, Pemerintah Republik Indonesia sudah merancang serta mengimplementasikan berbagai strategi dan program yang terarah dalam rangka menanggulangi permasalahan kemiskinan yang masih menjadi tantangan utama di tingkat nasional (Ferezagia, 2018). Berdasarkan data yang dikeluarkan oleh Badan Pusat Statistik (BPS) tahun 2024, tercatat bahwa persentase penduduk miskin di Indonesia mencapai 9,03%, yang menunjukkan adanya penurunan sebesar 0,33% dibanding dengan tahun 2023. Namun, kondisi ini berbanding terbalik dengan situasi di Provinsi Sumatera Barat yang justru mengalami peningkatan angka kemiskinan sebesar 0,02% pada tahun 2024. Dalam menghadapi kondisi tersebut, pemerintah berupaya menekan tingkat kemiskinan melalui berbagai kebijakan bantuan sosial yang ditujukan untuk membantu masyarakat yang berada dalam kondisi ekonomi kurang mampu. Salah satu program utama yang dijalankan ialah Program Keluarga Harapan (PKH), yaitu bentuk bantuan sosial bersyarat yang sudah dijalankan sejak tahun 2007 oleh pemerintah Indonesia.

Program ini termasuk bentuk intervensi sosial berupa pemberian bantuan tunai bersyarat pada keluarga penerima manfaat yang sudah ditetapkan secara resmi oleh pemerintah. Sasaran utama dari PKH ialah keluarga miskin atau rentan yang memenuhi sedikitnya satu dari beberapa kriteria yang sudah ditentukan, antara lain mempunyai ibu hamil, ibu nifas, atau anak balita; mempunyai anak usia 5–7 tahun yang belum menempuh pendidikan dasar; mempunyai anak usia 7–12 tahun yang sedang bersekolah di jenjang SD/MI/Paket A/SDLB; mempunyai anak usia 12–15 tahun yang menempuh pendidikan SLTP/MTs/Paket B/SMLB; atau mempunyai anak berusia 15–18 tahun yang belum selesai pendidikan dasar, termasuk anak dengan disabilitas. Dengan demikian, seluruh keluarga miskin yang memenuhi kriteria tersebut berhak memperoleh bantuan melalui Program Keluarga Harapan sebagai salah satu langkah strategis

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

pemerintah dalam meningkatkan kesejahteraan masyarakat serta menekan angka kemiskinan di Indonesia (Anggraeni & Nugroho, 2022).

Berdasarkan BPS (2024) persentase rumah tangga yang pernah menerima PKH di Provinsi Sumatera Barat tercatat sebesar 16,93% pada tahun 2024. Hal ini mencerminkan bahwa PKH tetap menjadi program penting bagi kelompok rumah tangga miskin dan rentan di Provinsi Sumatera Barat dalam mengatasi kemiskinan. Salah satu faktor yang di duga berkontribusi terhadap meningkatnya persentase penduduk miskin ialah ketidaktepatan sasaran dalam penyaluran PKH. Beberapa penelitian yang mendukung dugaan tersebut yaitu, merujuk pada pengamatan yang di lakukan Ekardo dkk, (2014) di Nagari Lagan Hilir Punggasan, Kabupaten Pesisir Selatan ditemukan bahwa penetapan penerima PKH tidak hanya melihat kondisi ekonomi yang kurang mampu, namun melihat adanya hubungan antar kader yang mendata dengan masyarakat setempat. Sehingga penetapan target PKH masih dianggap belum tepat sasaran, karena masih ditemukan di lapangan masyarakat yang dikategorikan ke dalam ekonomi menengah ke atas yang mendapatkan bantuan. Menurut Oktarina dkk, (2022) menunjukkan bahwa pelaksanaan PKH di Kecamatan Padang Ganting, Kabupaten Tanah Datar belum berjalan dengan optimal, dikarenakan masih banyak rumah tangga sangat miskin yang belum terdaftar pada Data Terpadu Kesejahteraan Sosial (DTKS), sehingga penerima PKH pada setiap tahunnya sama. Hal ini dikarenakan calon penerima PKH berasal dari data DTKS, sejak tahun 2020 RTSM dapat mendaftarkan mandiri secara online atau mendaftarkan diri melalui Kepala Desa/Wali Nagari dengan menyerahkan KTP dan KK. Namun di Kecamatan Padang Ganting informasi pendaftaran ini belum merata, sehingga penerima program cenderung tidak berubah karena minimnya pembaruan data. Selanjutnya menurut Wulandari dkk, (2024) menunjukkan bahwa PKH di Gunung Sarik Kota Padang tidak efektif dalam pelaksanaannya. Hal ini dipengaruhi oleh beberapa faktor penghambat seperti kesalahan pendataan dan penerapan penerima, rendahnya kesadaran peserta PKH, dan pemanfaatan dana yang kurang produktif.

Mengacu pada beberapa observasi yang di lakukan di Kabupaten/ Kota Provinsi Sumatera Barat, bisa disimpulkan bahwa masalah utama dalam pelaksanaan PKH di Provinsi Sumatera Barat terletak pada akurasi pendataan dan ketepatan sasaran penerima. Oleh karena itu, penerapan statistika seperti klasifikasi dalam *machine learning,* dapat dimanfaatkan untuk memprediksi atribut dalam mengidentifikasi keluarga miskin yang menerima PKH, agar penerima PKH bisa tepat sasaran. *Machine learning* termasuk kecerdasan buatan yang mampu melakukan pembelajaran dari data untuk melakukan prediksi, estimasi, klasterisasai, maupun klasifikasi secara otomatis (Mursalim dkk, 2024). Namun, permasalahan yang sering muncul dalam klasifikasi ialah ketidakseimbangan data, yaitu kondisi saat suatu ketegori mempunyai jumlah jauh lebih besar (mayoritas) dibanding kategori lain sehingga dapat menurunkan kinerja model pada kategori yang lebih sedikit (Fitriani dkk, 2021).

Berdasarkan data Survey Sosial Ekonomi Nasional (SUSENAS) maret 2024 oleh BPS Provinsi Sumatera Barat, jumlah rumah tangga penerima PKH sebanyak 1.795 sedangkan yang tidak menerima sebanyak 9.847 rumah tangga. Kondisi ini menunjukkan ketidakseimbangan data yang cukup besar. Sehingga, perlu dijalankan penanganan data tidak seimbang sebelum proses analisis yaitu memakai metode *Synthetic Minority Oversampling Technique* (SMOTE) yang bekerja dengan memperbanyak data minoritas hingga seimbang dengan data mayoritas. Penelitian sebelumnya oleh Siboro dkk, (2024) menunjukkan bahwa penggunaan SMOTE mampu meningkatkan akurasi model klasifikasi dibandingkan tanpa SMOTE.

Beberapa metode klasifikasi yang sering digunakan seperti *Naïve Bayes, Logistic Regression, K-Nearest Neighbors* (KNN)*, Support Vector Machine* (SVM)*, Decision Tree, Random Forest, Xgboot, Gradient Bossting* dan lainnya. Pemilihan algoritma klasikasi yang tepat dapat meningkatkan keakuratan keputusan yang di ambil. Menurut Jaya & Kadyanan (2023) membandingkan empat algoritma (*gradient boosting, logistic regression, decision tree,* dan *random forest*) pada klasifikasi penyakit jantung, dan menemukan bahwa *random forest* memberikan hasil terbaik dengan nilai *recall* sebesar 80,6%. Penelitian lain oleh Adriansyah dkk, (2022) juga menunjukkan bahwa *random forest* mempunyai akurasi lebih tinggi dibandingkan *support vector machine* dalam mengklasifikasikan kemampuan adaptasi siswa pada pembelajaran jarak jauh, dengan akurasi mencapai 91,5%. Berdasarkan kedua sumber acuan tersebut, proses klasifikasi terhadap penerima Program Keluarga Harapan (PKH) di Provinsi Sumatera Barat dijalankan dengan menerapkan algoritma Random Forest. Algoritma ini termasuk bentuk pengembangan dari metode *Classification and Regression Trees* (CART) yang mengintegrasikan teknik *Random Feature Selection* serta *Bootstrap Aggregating* (*bagging*) dalam penerapannya. Pada prinsip kerjanya, *random forest* membangun sejumlah pohon keputusan (*decision trees*) yang kemudian digabungkan menjadi satu himpunan besar yang disebut "hutan" (*forest*). Himpunan pohon tersebut selanjutnya dianalisis secara kolektif untuk menghasilkan keputusan yang lebih stabil dan akurat. Pendekatan ini terbukti mampu meningkatkan tingkat ketepatan (*accuracy*) melalui proses pembangkitan simpul turunan (*child nodes*) pada setiap simpul induk (*parent node*) secara acak, sehingga mengurangi kemungkinan *overfitting* (Hadi & Benedict, 2024).

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Ketepatan penyaluran PKH sangat penting bagi efektifitas program dalam mengatasi masalah kemiskinan. Penerapan algoritma *random forest* dalam mengklasifikasikan PKH di Provinsi Sumatera Barat diharapkan dapat memberikan saran dan membantu pemerintah untuk mengambil kebijakan dalam memperbaiki sistem pendataan dan validasi penerima PKH di Provinsi Sumatera Barat agar penyaluran PKH bisa tepat sasaran.

II. METODE PENELITIAN

Penelitian ini tergolong dalam kategori penelitian terapan, yang berfokus pada implementasi praktis konsep dan metode ilmiah untuk memperoleh solusi terhadap permasalahan nyata. Dalam studi ini, algoritma *random forest* dipadukan dengan metode SMOTE (*Synthetic Minority Oversampling Technique*) diterapkan sebagai pendekatan analitis untuk melakukan klasifikasi pada rumah tangga penerima Program Keluarga Harapan (PKH) di Provinsi Sumatera Barat tahun 2024.

Data yang dimanfaatkan termasuk data sekunder mengenai rumah tangga penerima PKH di Provinsi Sumatera Barat tahun 2024, dengan total 11.642 unit observasi yang merepresentasikan karakteristik sosial ekonomi penerima bantuan setelah dilakukan *preprocessing* data.

Selanjutnya, penelitian ini melibatkan satu variabel dependen sebagai objek utama klasifikasi dan dua belas variabel independen sebagai faktor penjelas yang berpotensi memengaruhi hasil klasifikasi. Rincian setiap variabel tersebut tercantum secara sistematis pada Tabel 1 sebagai acuan analisis data.

Tabel 1. Variabel Penelitian

No	Variabel	Skala	Kategori
(1)	(2)	(3)	(4)
1	Menerima PKH	Nominal	Iya dan Tidak
2	Status kepemilikan tempat tinggal	Nominal	Milik sendiri; Dinas; Bebas Sewa; Kontrakan/Sewa; dan lainnya
3	Bahan atap rumah	Nominal	Jerami/Ijuk; Kayu/Sirap; Bambu; Asbes; Seng; Genteng; Beton
4	Bahan dinding rumah	Nominal	Tembok; Plasteran; Batang Kayu; Anyaman Bambu; Kayu/Papan; Bambu; Jerami/Ijuk
5	Bahan lantai rumah	Nominal	Kayu/ Papan; Ubin/Tegel/Teraso; Paket/Vinil/Karpet; Keramik; Marmer; Bambu; Semen/Bata merah; Tanah
6	Mempunyai fasilitas tempat buang air besar	Nominal	Ada, digunakan bersama rumah tangga tertentu; Ada, di MCK komunal; Ada di MCK umum; Ada, digunakan ART sendiri; Ada, ART tidak menggunakan; Tidak ada
7	Sumber utama penerangan rumah	Nominal	Listrik PLN dengan meteran; Listrik PLN non meteran; Non PLN; Bukan Listrik
8	Daya listrik	Nominal	Tidak ada; 450 watt; 900 watt; 1300 watt
9	Bahan Bakar Memasak	Nominal	Tidak memasak dirumah, Minyak tanah, Listrik, Elpiji 3 kg, Elpiji 12 kg, Elpiji 5,5 kg/bluegaz, Kayu bakar
10	Memiliki Kulkas	Nominal	Iya dan Tidak
11	Memiliki AC	Nominal	Iya dan Tidak
12	Memiliki Sepeda Motor	Nominal	Iya dan Tidak
13	Ijazah	Nominal	Tidak tamat SD, SD/Sederajat, SLTP/Sederajat, SLTA/sederajat, D1/D2/D3/D4/S1, Profesi/S2/S3

A. Tahapan Analisis

Tahapan analisis data sebagai berikut.

1. Input data kemudian melakukan deskriptif data.
2. Melakukan penanganan data tidakseimbang dengan metode SMOTE pada data *training*.

   Teknik *Synthetic Minority Oversampling Technique* (SMOTE) termasuk suatu pendekatan yang digunakan untuk menyeimbangkan distribusi data dengan cara melakukan oversampling terhadap kelas minoritas melalui pembentukan sampel buatan (*synthetic samples*). Prosedur kerja metode SMOTE dimulai dengan mengidentifikasi sejumlah *k nearest neighbors* yakni himpunan data yang mempunyai jarak terdekat sebanyak k terhadap setiap data dalam kelas minoritas. Selanjutnya, sistem menghasilkan data sintetis baru berdasarkan

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

persentase penggandaan yang sudah ditentukan, dengan cara mengombinasikan karakteristik data minoritas asli dan tetangga terdekatnya yang dipilih secara acak (Siringoringo, 2018). Langkah-langkah metode SMOTE sebagai berikut:

a. Menghitung jarak antar kategori tetangga terdekat kelas minoritas memakai persamaan *Value Difference Metric* (VDM) berikut:

$$\delta(V_1, V_2) = \sum_{i=0}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|$$

dengan n ialah jumlah kelas dalam data, C_{1i} ialah jumlah V_1 yang masuk ke dalam kelas ke- i, C_{2i} ialah jumlah V_2 yang masuk ke dalam kelas ke- i. C_1, C_2 ialah jumlah total kelas V_1, V_2.

b. Menghitung jarak antar amatan antara kelas minoritas memakai persamaan berikut:

$$\Delta(A, B) = \sum_{i=1}^{N} \delta(V_1, V_2)$$

dengan $\Delta(A, B)$ ialah jarak antara amatan A dan B, N ialah jumlah variabel, dan $\delta(V_1, V_2)$ ialah jarak kategori V_1 dan V_2.

c. Membuat sampel sintetis baru antara dua amatan berdasarkan jarak terdekat dari langkah sebelumnya memakai persamaa berikut:

$$S_{syn} = r(S_{kNN} - S_f) + S_f \quad (1)$$

dengan S_{syn} ialah sampel sintetis yang dihasilkan, r termasuk bilangan acak bernilai 0 dan 1, S_{kNN} ialah tetangga terdekat k sampel fitur yang di pertimbangkan, S_f ialah sampel fitur.

3. Data penelitian dipecah menjadi dua kelompok utama, yakni data pengujian (*testing* data) dan data pelatihan (*training* data). Data pelatihan berfungsi sebagai sarana untuk melatih sistem dalam proses pembentukan model prediktif, sehingga model tersebut mampu mengenali pola dan hubungan antarvariabel secara optimal. Sementara itu, data pengujian digunakan untuk mengevaluasi kemampuan model dalam melakukan prediksi serta menilai tingkat akurasi hasil yang dihasilkan dari proses pelatihan sebelumnya. Dalam konteks penelitian ini, proporsi pembagian data ditetapkan sebesar 20% untuk data pengujian dan 80% untuk data pelatihan, guna memastikan keseimbangan antara proses pembelajaran model dan pengujian kinerjanya secara objektif.
4. Melakukan klasifikasi memakai algoritma *random forest*.

Metode *random forest* termasuk salah satu teknik dari sekian banyak pendekatan *ensemble learning* yang diperkenalkan dan dikembangkan oleh Leo Breiman pada tahun 2001. Menurut Breiman (2001), *random forest* termasuk penyempurnaan dari metode CART melalui penerapan prinsip *random feature selection* serta B*ootstrap Aggregating* (*bagging*).

Secara konseptual, algoritma ini beroperasi dengan membangun sejumlah pohon keputusan (*decision tree*) secara acak, di mana setiap pohon dilatih memakai subset data dan variabel yang dipilih secara acak. Hasil prediksi dari seluruh pohon tersebut kemudian dikombinasikan melalui proses agregasi untuk menghasilkan keputusan akhir yang lebih konsisten, stabil, serta mempunyai tingkat akurasi yang melebihi penggunaan satu pohon keputusan tunggal.

Tahapan pembentukan pohon dalam proses klasifikasi memakai algoritma *random forest* dapat dijelaskan secara sistematis sebagai berikut:

a. Melakukan proses pengambilan sampel acak berukuran tertentu dengan metode pengembalian pada himpunan data, tahapan ini dikenal sebagai *Bootstrap Aggregating* (*bagging*).
b. Melakukan *tuning* parameter untuk menentukan nilai parameter optimum. *Tuning* parameter yang digunakan *mtry* $\sqrt{p}$ dan *ntree* 100, 250, 500, dan 1000 yang di kombinasikan.
c. Membangun pohon ke- k berdasarkan sampel *bagging* dan tahapan *random feature selection* (*mtry*). Tahapan ini terdiri dari:

1). Menghitung nilai indeks gini ($i(t)$) simpul kanan dan simpul kiri memakai persamaan berikut:

Simpul Kiri: $i(t_L) = 1 - \sum_{j=0}^{1} P^2(j|t_L)$ (2)

Simpul Kanan: $i(t_R) = 1 - \sum_{j=0}^{1} P^2(j|t_R)$ (3)

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$i(t) = 1 - \sum_j P^2(j|t) \quad (4)$$

dengan:
$P_L = \frac{P(t_L)}{P(t)}$, $P_R = \frac{P(t_R)}{P(t)}$ dimana P_L, P_R ialah proporsi banyaknya objek yang masuk pada t_L atau t_R, untuk $P(t_L)$ dan $P(t_R)$ ialah banyaknya observasi yang tergolong dalam kelas L dan R pada node t, sedangkan P(t) menggambarkan total jumlah observasi yang terdapat pada *node* t.

2). Menentukan pemilah terbaik oleh pemilah s pada simpul t berdasarkan kriteria *goodness of split memakai* persamaan berikut:

$$\Delta i(s,t) = i(t) - P_L i((t_L) - P_R i((t_R) \quad (5)$$

5. Menghitung tingkat kesalahan klasifikasi untuk menentukan pemilihan *forest* terbaik berdasarkan laju galat *Out of Bag* (OOB) terkecil memakai persamaan berikut:

$$Laju\ Galat\ OOB_i = \frac{1}{n}\sum_{i=1}^{n} 1_{y_i \neq \hat{y}_i}$$

dengan, $\sum_{i=1}^{n} 1_{y_i \neq \hat{y}_i}$ ialah jumlah data hasil prediksi salah, y_i ialah nilai aktual sebenarnya ke-i, $\hat{y}_i$ ialah hasil amatan prediksi ke-i, dan n ialah jumlah OOB ke-i yang menjadi sampel OOB.
Kemudian menghitung rataan tingkat kesalahan klasifikasi OOB dengan persamaan berikut:

$$Laju\ Galat\ OOB = \frac{\sum OOB_i laju\ galat}{k} \times 100\% \quad (6)$$

6. Mengidentifikasi variabel penting.

Tingkat kepentingan peubah penjelas (variabel *importance*) yang didapat oleh *random forest* dapat diukur dengan *Mean Decrease Impurity (MDI)* atau yang juga dikenal sebagai *Mean Decrease Gini (MDG)*. Misalkan ada p peubah penjelas dengan $h = 1,2,\ldots,p$, maka MDG mengukur tingkat kepentingan peubah penjelas X_h dengan cara berikut (Kristanaya dkk, 2025).

$$MDG_h = \frac{1}{k}\sum_t [d(h,t)I(h,t)] \quad (7)$$

Dimana, k ialah jumlah pohon yang dibentuk, $d(h,t)$ penurunan indeks gini variabel penjelas pada simpul t, $I(h,t)$ fungsi indikator yang bernilai 1 saat variabel penjelas memilih simpul t dan 0 yang lain.

7. Melakukan evaluasi model memakai *confussion matrix*.

Berdasarkan pandangan Romadloni dkk, (2022), penggunaan *confusion matrix* berfungsi sebagai instrumen evaluatif yang menyajikan perbandingan antara hasil klasifikasi yang dihasilkan oleh algoritma dengan hasil klasifikasi aktual atau sesungguhnya. Tingkat keakuratan performa klasifikasi dapat diketahui melalui nilai akurasi yang diperoleh dari hasil pengolahan *confusion matrix* tersebut. Adapun bentuk penyajian *confusion matrix* dapat dilihat pada Tabel 2.

Tabel 2. *Confussion Matrix*

classification		***Predicted Class***	
		Positive	***Negative***
Actual Class	***Positive***	*True Positive (TP)*	*False Negative (FN)*
	Negative	*False Positive (FP)*	*True Negative (TN)*

Dari Tabel 2, bisa dihitung nilai berikut:

a. *Accuracy*

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Accuracy termasuk ukuran ketepatan yang menunjukkan sejauh mana suatu model mampu melakukan klasifikasi dengan benar. Nilai *accuracy* tersebut dapat dihitung memakai rumus atau persamaan seperti berikut.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (8)$$

b. *Precision*
Precision termasuk perbandingan diantara jumlah prediksi positif yang benar dengan keseluruhan data yang diprediksi sebagai positif. Nilai *precision* dapat diperoleh melalui penggunaan rumus atau persamaan berikut.

$$Precision = \frac{TP}{(TP + FP)} \quad (9)$$

c. *Recall* atau *Sensitivity*
Recall, atau yang dikenal juga sebagai *sensitivity*, termasuk ukuran yang membandingkan jumlah prediksi positif yang benar dengan keseluruhan data aktual yang memang tergolong ke dalam kelas positif. Dengan kata lain, *recall* menunjukkan tingkat kemampuan model dalam mengenali seluruh data yang benar-benar positif. Nilai *recall* dapat ditentukan melalui rumus atau persamaan berikut.

$$Recall = \frac{TP}{(TP + FN)} \quad (10)$$

d. *Specificity*
Specificity termasuk perbandingan antara jumlah prediksi negatif yang benar dengan total semua data yang termasuk dalam kategori negatif. Nilai s*pecificity* dapat diperoleh melalui penggunaan rumus atau persamaan berikut.

$$Specificity = \frac{TN}{(TN + FP)} \quad (11)$$

e. F1-*Score*
F1-*score* termasuk ukuran yang menyajikan keseimbangan antara *precision* dan *recall*, sehingga sangat bermanfaat ketika data yang digunakan mempunyai distribusi kelas yang tidak merata. Nilai F1-*score* dapat diperoleh melalui rumus atau persamaan (12).

$$\text{F1} - Score = 2 \times \frac{Precision \times recall}{Precision + recall} \quad (12)$$

8. Menginterpretasikan hasil atau menarik kesimpulan.

III. HASIL DAN PEMBAHASAN

A. Deskriptif Data

Data yang digunakan sebanyak 11.642 amatan dengan satu variabel dependen sebagai objek utama klasifikasi dan dua belas variabel independen sebagai faktor penjelas yang berpotensi memengaruhi hasil klasifikasi. Maka diperoleh tabel frekuensi dari kategori masing masing variabel independen. Tabel frekuensi untuk tiap-tiap variabel terdapat pada Tabel 3.

Tabel 3. Frekuensi Masing-Masing Variabel

Variabel	Kategori	Menerima	Tidak menerima
Status kepemilikan tempat tinggal	Milik sendiri	**1377**	7235
	Kontrakan/Sewa	105	746
	Bebas Sewa	311	1668
	Dinas	2	182
	Lainnya	0	2
Bahan atap rumah	Beton	8	124
	Genteng	32	248
	Seng	**1707**	9332
	Asbes	20	69
	Bambu	4	16
	Kayu/Sirap	1	7

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

	Jerami/Ijuk	23	51
Bahan dinding rumah	Tembok	**1193**	7850
	Plasteran	34	125
	Kayu/Papan	552	1809
	Anyaman Bambu	8	26
	Batang Kayu	3	21
	Bambu	1	7
	Lainnya	4	9
Bahan lantai rumah	Marmer	10	215
	Keramik	232	3744
	Paket/Vinil/Karpet	6	17
	Ubin/Tegel/Teraso	15	96
	Kayu/ Papan	246	682
	Semen/Bata merah	**1271**	5038
	Bambu	6	24
	Tanah	9	31
Memiliki fasilitas tempat buang air besar	Ada, digunakan ART sendiri	**1389**	8544
	Ada, digunakan bersama rumah tangga tertentu	92	562
	Ada, di MCK komunal	17	35
	Ada di MCK umum	71	196
	Ada, ART tidak menggunakan	4	15
	Tidak ada	222	495
Sumber utama penerangan rumah	Listrik PLN dengan meteran	**1601**	9235
	Listrik PLN non meteran	130	409
	Non PLN	8	78
	Bukan Listrik	56	125
Daya listrik	Tidak ada	194	612
	450 watt	517	2182
	900 watt	**1061**	6005
	1300 watt	23	1048
Bahan bakar memasak	Tidak memasak dirumah	6	74
	Listrik	2	37
	Elpiji 5,5 kg/bluegaz	2	57
	Elpiji 12 kg	11	335
	Elpiji 3 kg	**1353**	8157
	Minyak tanah	37	186
	Kayu bakar	384	1001
Memiliki kulkas	Iya	**942**	7085
	Tidak	853	2762
Memiliki AC	Iya	2	482
	Tidak	**1793**	9365
Memiliki sepeda motor	Iya	**1419**	8158
	Tidak	376	1689
Ijazah kepala rumah tangga	Tidak tamat SD	525	1613
	SD/Sederajat	**560**	1983
	SLTP/Sederajat	321	1565
	SLTA/Sederajat	368	3466
	D1/D2/D3/D4/S1	21	1087
	Profesi/S2/S3	0	133

Berdasarkan Tabel 3 diperoleh sebagian besar rumah tangga penerima PKH di Provinsi Sumatera Barat mempunyai rumah dengan status milik sendiri, beratap seng, diding tembok, lantai semen/bata merah, mempunyai

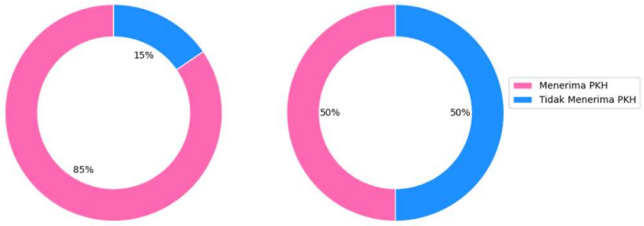
DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

fasilitas tempat buang air besar yang dipergunakan ART sendiri, penerangan rumah memakai listrik PLN dengan meteran, daya listrik 900 watt, bahan bakar memasak menggunakan gas elpiji dengan berat 3 kg, mempunyai kulkas, tidak mempunyai AC, mempunyai sepeda motor, serta pendidikan terakhir kepala rumah tangga adalah tamat SD/sederajat. .

B. *Synthetic Minority Oversampling Technique*

Mengacu pada Tabel 3 diketahui bahwa proporsi data penerima PKH dan tidak penerima PKH tidak seimbang, dimana amatan cenderung berada di kelas tidak penerima PKH. Hal ini mengakibatkan kategori penerima PKH menjadi kelas minoritas dan kategori tidak penerima PKH menjadi kelas mayoritas. Oleh karena itu, sebelum melakukan analisis data harus di seimbangkan terlebih dahulu dengan menerapkan metode SMOTE pada persamaan (3). Visualisasi kelas sebelum dan sesudah dijalankan proses penyeimbangan terdapat pada Gambar 1.



(a) (b)

Gambar 1. (a) Data Sebelum Diseimbangkan dan (b) Data Setelah Diseimbangkan

Gambar 1 (a) memperlihatkan jika jumlah penerima PKH jauh lebih sedikit dibanding dengan tidak penerima PKH. Kelas penerima PKH termasuk kelas minoritas dengan proporsi 15% atau sebanyak 1.795 amatan, sedangkan kelas tidak penerima PKH menjadi kelas mayoritas dengan proporsi 85% atau sebanyak 9.847 amatan. Setelah dijalankan penerapan metode SMOTE, diperoleh distribusi data yang seimbang sebagaimana terlihat pada Gambar 1 (b). Hasil penyeimbangan ini memperoleh proporsi yang sama, yaitu 50% untuk penerima PKH dengan jumlah 9.847 amatan dan 50% untuk kelas tidak penerima PKH dengan jumlah 9.847 amatan.

C. *Random Forest*

Pada proses klasifikasi dengan algoritma *random forest,* tahapan pertama yang dijalankan untuk membangun model ialah dengan melakukan *tuning* parameter, yaitu penentuan banyaknya variabel (*mtry*) yang akan digunakan dalam pembentukan pohon. Pada kajian ini *mtry* yang akan digunakan ialah $\sqrt{p} = \sqrt{12} \approx 3$. *Mtry* tersebut akan dikombinasikan dengan jumlah pohon (*ntree*) yang berbeda yaitu 100, 250, 500, dan 1000. Proses pembentukan pohon dijalankan memakai persamaan (2), (3), (4), dan (5). Selanjutnya menentukan pemilihan *forest* terbaik berdasarkan laju galat OOB memakai persamaan (6). Semakin kecil laju galat OOB yang dihasilkan, maka prediksi pada *forest* akan semakin akurat serta bisa dipercaya. Nilai laju galat OOB terhadap kasus penerima PKH di Provinsi Sumatera Barat tahun 2024 bisa dipahami dari Tabel 4.

Tabel 4. *Tuning* Parameter *Random Forest*

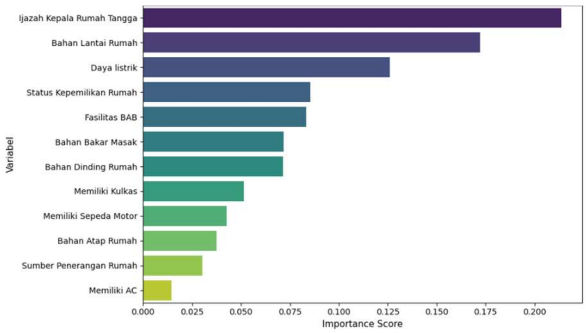
Jumlah *Ntree*	**Laju Galat OOB**
100	25,42%
250	25,23%
500	**25,22%**
1000	25, 35%

Berdasarkan Tabel 4 diketehui bahwa *random forest* yang optimal yakni *forest* dengan kombinasi *mtry* = 3 dan *ntree* = 500. Nilai laju galat OOB ini digunakan untuk mengetahui besarnya kesalahan klasifikasi yang dijalankan oleh algoritma *random forest* dalam membedakan status penerima bantuan PKH dan tidak penerima bantuan PKH di Provinsi Sumatera Barat pada tahun 2024. Selain itu, hasil pengukuran ini juga menjadi dasar dalam evaluasi model memakai *confussion matrix*.

Setelah diperoleh *random forest* yang optimal, langkah berikutnya ialah melihat tingkat kepentingan peubah penjelas dengan menghitung nilai MDG pada persamaan (7). Tingkat kepentingan atribut ini menggambarkan variabel-variabel yang berpengaruh terhadap kasus penerima PKH di Provinsi Sumatera Barat pada tahun 2024 yang diurutkan mulai dari nilai tertinggi hingga terendah. Nilai MDG dapat dilihat pada Gambar 2.

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Gambar 2. MDG Penerima PKH

Berdasarkan Gambar 2 bisa dilihat jika urutan atribut yang paling penting ialah Ijazah kepala rumah tangga (X_{12}), di susul dengan bahan lantai rumah (X_4), dilanjutkan dengan daya listrik (X_7), status kepemilikan rumah (X_1), fasilitas BAB (X_5), bahan bakar memasak (X_8), bahan dinding rumah (X_3), mempunyai kulkas (X_9), mempunyai sepeda motor (X_{11}), bahan atap rumah (X_2), sumber utama penerangan rumah (X_6), dan mempunyai AC (X_{10}). Sehingga bisa disimpulkan bahwa faktor yang paling berpengaruh dalam klasifikasi penerima bantuan PKH ialah Ijazah kepala rumah tangga (X_{12}). Dengan kata lain, tingkat pendidikan kepala rumah tangga berperan penting dalam menentukan status ekonomi rumah tangga, yang berdampak pada kelayakan penerima bantuan sosial. Dimana sebagian besar rumah tangga penerima PKH berpendidikan rendah yaitu tidak tamat SD dan tamat SD/sederajat. Hanya sedikit yang menempuh pendidikan menengah, dan hampir tidak ada yang berpendidikan tinggi. Kondisi ini menunjukkan bahwa rendahnya pendidikan berhubungan dengan kerentanan ekonomi penerima PKH.

D. *Confussion Matrix*

Setelah dijalankan proses analisis algoritma *random forest* menggunakan SMOTE, data *testing* yang digunakan sebanyak 3.939 amatan dari 19.694 data keseluruhan. *Confussion matrix* dari algoritma *random forest* menggunakan SMOTE dapat dilihan pada Tabel 5.

Tabel 5. Nilai *Confussion Matrix Random Forest*

Classification		*Predicted Class*	
		Penerima PKH	Tidak Penerima PKH
Actual Class	Penerima PKH	1624	345
	Tidak Penerima PKH	661	1309

Berdasarkan Tabel 5 dapat dijalankan perhitungan nilai *accuracy, precision, recall, specificity* serta F1-*Score* dari klasifikasi penerima bantuan PKH memakai algoritma *random forest* dengan SMOTE dapat dilihat pada Tabel 6.

Tabel 6. Nilai *Acuracy, Precision, Recall, Specitificity, dan F1-Score*

Kinerja Model	**Hasil Klasifikasi**
Accuracy	74%
Precision	71%
Recall	82%
Specificity	66%
F1-Score	76%

Berdasarkan Tabel 6 hasil pengujian model *random forest* dengan jumlah pohon terbaik sebanyak 500 estimator menunjukkan bahwa penerapan teknik SMOTE mampu mengklasifikasikan rumah tangga penerima PKH dengan *accuracy* sebesar 74%, dimana hasil prediksi rumah tangga yang benar-benar menerima PKH diperoleh sebesar 82%. Nilai *precision* sebesar 71% memperlihatkan jika mayoritas rumah tangga yang diprediksi oleh model sebagai penerima PKH memang benar-benar termasuk dalam kategori penerima. Nilai f1-*score* menunjukkan bahwa rata-rata untuk ketepatan dua kelas sebesar 76%. Secara keseluruhan, penggunaan SMOTE pada algoritma *random forest* menunjukkan kinerja yang baik dalam mengatasi ketidakseimbangan kelas serta meningkatkan kemampuan model dalam mendeteksi rumah tangga penerima PKH di Provinsi Sumatera Barat. Hasil ini sejalan dengan temuan Fernandez dkk, (2018) yang menunjukkan bahwa kombinasi metode *ensemble learning* seperti *random forest* dengan teknik *oversampling* mampu

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

mengurangi bias terhadap kelas mayoritas dan menghasilkan akurasi yang lebih stabil. Sehingga, penerapan SMOTE terbukti dapat meningkatkan kemampuan model dalam mengenali kelas minoritas pada data.

IV. KESIMPULAN

Model optimal yang dihasilkan oleh algoritma *random forest* untuk mengidentifikasi penerima PKH di Provinsi Sumatera Barat tahun 2024 ialah memakai *mtry* = 3 dan *ntree* = 500. Sesuai dengan hasil *forest* optimal diperoleh ijazah kepala rumah tangga yang menjadi atribut utama dalam menentukan rumah tangga menerima PKH atau tidak. Temuan ini mengindikasikan bahwa tingkat pendidikan kepala rumah tangga dapat dijadikan salah satu indikator penting dalam proses verifikasi dan validasi data kesejahteraan masyarakat dalam Data Terpadu Kesejahteraan Sosial (DTKS). Pemerintah daerah dapat mempertimbangkan untuk memberikan bobot lebih besar pada variabel pendidikan dalam penyusunan kebijakan penentuan penerima sasaran program sosial khususnya PKH agar penyaluran bantuan lebih tepat sasaran. Model dengan SMOTE bisa untuk mengklasifikasi kategori kelas minoritas yakni penerima PKH dengan baik, dilihat dari nilai persentase *precision, recall,* dan f1-*score*. Sehingga model mempunyai kinerja yang baik dan cukup andal dalam mengidentifikasi penerima PKH di Provinsi Sumatera Barat pada tahun 2024. Namun demikian, penelitian ini memiliki keterbatasan karena hanya menggunakan data dari satu provinsi, sehingga generalisasi hasil ke wilayah lain di Indonesia perlu dilakukan dengan hati-hati. Penelitian selanjutnya disarankan untuk memperluas cakupan data antarprovinsi serta mempertimbangkan variabel sosial-ekonomi lain seperti kondisi pekerjaan dan jumlah tanggungan rumah tangga guna meningkatkan ketepatan model prediksi penerima PKH.

UCAPAN TERIMA KASIH

Terimaksih kepada instansi penyedia data pada kajian ini yakni Badan Pusat Statistik (BPS) Provinsi Sumatera Barat.

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

Anggraeni, A. P., & Nugroho, A. A. (2022). Evaluasi Kebijakan Pkh (Program Keluarga Harapan) Di Indonesia. *Journal of Public Policy and Applied Administration*, *4*(2), 39–54. https://doi.org/10.32834/jplan.v4i2.529

BPS. (2024). *Statistika Kesejahteraan Rakyat 2024* (Vol. 39, Issue 1). Badan Pusat Statistik. https://doi.org/10.1016/j.earlhumdev.2006.05.022

Breiman, L. (2001). *"Random Forest" Machine Learning*. *45*, 5–32. https://doi.org/10.1007/978-3-030-62008-0_35

Ekardo, A., Firdaus, F., & Elfemi, N. (2014). Efektifitas Program Keluarga Harapan (Pkh) Dalam Upaya Pengentasan Kemiskinan Di Nagari Lagan Hilir, Kab. Pesisir Selatan. *Jurnal Ilmu Sosial Mamangan*, *3*(1), 1–9. https://doi.org/10.22202/mamangan.v3i1.1345

Ferezagia, D. V. (2018). Analisis Tingkat Kemiskinan di Indonesia. *Jurnal Sosial Humaniora Terapan*, *1*(1). https://doi.org/10.7454/jsht.v1i1.6

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets.* Springer. https://doi.org/10.1007/978-3-319-98074-4

Fitriani, R. D., Yasin, H., & Tarno, T. (2021). Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Oversampling Pada Naive Bayes (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal). *Jurnal Gaussian*, *10*(1), 11–20. https://doi.org/10.14710/j.gauss.v10i1.30243

Hadi, N., & Benedict, J. (2024). Implementasi Machine Learning Untuk Prediksi Harga Rumah memakai Algoritma

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Random Forest. *Computatio : Journal of Computer Science and Information Systems*, *8*(1), 50–61. https://doi.org/10.24912/computatio.v8i1.15173

Jaya, M. K. ., & Kadyanan, G. A. . (2023). Perbandingan Random Forest, Decision Tree, Gradient Boosting, Logistic Regression untuk Klasifikasi Penyakit Jantung. *Jnatia*, *2*(November), 1–5.

Kristanaya, M., Azzahra, P. M., Trimono, & Idhom, M. (2025). Klasifikasi Status Rujukan Pasien Poliklinik Bandara Berbasis Random Forest dan Interpretabilitas Model memakai SHAP. *Data Sciences Indpnesia*, *5*(1), 60–74.

Mursalim, M., Aprilia, T., Samas, M. A., Rahmawati, A., & Mufidah, I. F. (2024). Pengenalan Machine Learning Untuk Mahasiswa memakai Metode Service Learning. *Abdimasku : Jurnal Pengabdian Masyarakat*, *7*(2), 493. https://doi.org/10.62411/ja.v7i2.1959

Oktarina, V., Karlina, N., & Candradewini, C. (2022). Evaluasi Konteks Program Keluarga Harapan (Pkh) Di Kecamatan Padang Ganting Kabupaten Tanah Datar. *JANE - Jurnal Administrasi Negara*, *14*(1), 361. https://doi.org/10.24198/jane.v14i1.41324

Romadloni, P., Adhi Kusuma, B., & Maulana Baihaqi, W. (2022). Komparasi Metode Pembelajaran Mesin Untuk Implementasi Pengambilan Keputusan Dalam Menentukan Promosi Jabatan Karyawan. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *6*(2), 622–628. https://doi.org/10.36040/jati.v6i2.5238

Siboro, O., Pricilia Banjarnahor, Y., Gultom, A., Antonius Siagian, N., & Silitonga, P. D. (2024). Penanganan Data Ketidakseimbangan dalam Pendekatan SMOTE Guna Meningkatkan akurasi Algoritma K-NN 1). *SNISTIK : Seminar Nasional Inovasi Sains Teknologi Informasi Komputer*, *1*(2), 473–478. https://ejournal.ust.ac.id/index.php/SNISTIK/article/view/3705

Siringoringo, R. (2018). Klasifikasi Data Tidak Seimbang memakai Algoritma. *Jurnal ISD*, *3*(1), 44–49.

Wulandari, T. N., Chandra, D., Ramadhan, R., & Padang, U. (2024). Efektifitas Pelaksanaan Program Keluarga Harapan di Kelurahan Gunung Srik Kecamatan Kuranji Kota Padang Sumatra Barat. *Jurnal Pendidikan Dan Sosial Budaya*, *4*, 1095–1101.

DOI: https://doi.org/10.24036/ujsds/vol3-iss4/431