UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 3 Nomor 2 Tahun 2025 Halaman 232-239.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Application of Extreme Gradient Boosting Algorithm with ADASYN for Classification of Households Receiving Program Keluarga Harapan in West Sumatra Province

Amelia Susrifalah, Dodi Vionanda*, Yenni Kurniawati, Dwi Sulistiowati
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: dodi_vionanda@fmipa.unp.ac.id

Submitted : 15 Mei 2025
Revised : 23 Mei 2025
Accepted : 30 Mei 2025

ABSTRACT

Program Keluarga Harapan (PKH) is a form of social protection provided by the government to overcome poverty in Indonesia. However, challenges remain in accurately predicting eligible households. Therefore, a data-based classification method is needed to identify PKH recipients based on their factors. This research was conducted in West Sumatra Province using variables from the Data Terpadu Kesejahteraan Sosial (DTKS) variable group contained in SUSENAS 2024. Based on data from Badan Pusat Statistik (BPS) of West Sumatera Province, there are 1.790 PKH recipient households and 9.810 non-recipient households, indicating a class imbalance. Considering the large amount of data and complex variables, PKH can be analyzed using the Extreme Gradient Boosting (XGBoost) algorithm because of its ability to handle large-scale data and produce high classification performance. To address data imbalance, Adaptive Synthetic (ADASYN) was applied before analysis. The application of XGBoost with the scale_pos_weight parameter shows low classification performance, with sensitivity value of 12.3% and balanced accuracy of 55.2%. To overcome this, unbalanced data was handled using the ADASYN method. The application of XGBoost after data balancing with ADASYN showed significant performance improvement, with sensitivity value 80% and balanced accuracy 87.6%. In classifying PKH recipient households, the variables that make an important contribution are the age of the head of household, floor area, diploma of the head of household, and floor material. This research shows that the combination of XGBoost and ADASYN is effective in overcoming data imbalance and improving PKH recipient classification performance.

Keywords: *Adaptive Synthetic, Extreme Gradient Boosting, Program Keluarga Harapan.*

I. PENDAHULUAN

Extreme Gradient Boosting (XGBoost) merupakan bagian *machine learning* yang menerapkan teknik *ensamble boosting,* dikembangkan sebagai optimalisasi *gradient boosting* dirancang efisien dan fleksibel (Boehmke & Greenwell, 2020). Algoritma XGBoost diperkenalkan oleh Chen & Guestrin (2016) dimana *week learner* diproses secara berurutan, dan setiap iterasi meminimalkan fungsi objektif dengan pendekatan *taylor expansion* hingga turunan kedua, sehingga dapat menghasilkan *strong learner* yang sangat akurat. Algoritma XGBoost memiliki berbagai fitur yang dapat digunakan untuk data berskala besar dan dapat menyelesaikan masalah klasifikasi. Pada penelitian Wang dkk., (2021) telah membandingkan algoritma XGBoost, *Decision tree*, dan KNN pada klasifikasi resiko kredit dengan 26 variabel pada 10.000 nasabah Bank diperoleh XGBoost memiliki akurasi terbaik mencapai 87%.

Masalah ketimpangan data menjadi tantangan dalam klasifikasi. Ketimpangan data terjadi ketika jumlah kelas minoritas jauh lebih sedikit dibandingkan dengan kelas mayoritas. Hal ini menimbulkan kecenderungan bias pada kelas mayoritas, sehingga kesalahan klasifikasi pada kelas minoritas menjadi lebih tinggi (Fernández dkk., 2018:14). Algoritma XGBoost memiliki parameter pembobotan scale_pos_weight yang dapat digunakan untuk menangani data tidak seimbang (Bischl dkk., 2023). Namun, beberapa kasus data tidak seimbang, parameter ini belum mampu mengatasi permasalahan ketepatan antar kelasnya. Merujuk penelitian Perdeck (2024) menerapkan parameter scale_pos_weight pada XGBoost, hal ini dapat meningkatkan nilai AUC, namun perlu mengorbankan ketepatan salah satu kelasnya. Maka, penanganan data tidak seimbang perlu dilakukan jika perbandingan ketepatan kelas cukup jauh.

DOI: https://doi.org/10.24036/ujsds/vol3-iss2/369

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

Metode penanganan data tidak seimbang dapat diterapkan dengan meningkatkan sampel data pada kelas minoritas. Salah satu metode yang dapat digunakan adalah *Adaptive Synthetic* (ADASYN), dengan menghasilkan sampel-sampel minoritas secara adaptif sesuai dengan distribusinya, sampel sintetis dibuat berdasarkan kelas minoritas yang sulit dipelajari, sehingga dapat mengurangi pembelajaran yang bias (He dkk., 2008). Sejalan dengan penelitian Kaope & Pristyanto (2023) yang membandingkan kinerja SMOTE, ADASYN dan SMOTE-ENN dalam mengatasi data tidak seimbang pada metode klasifikasi diperoleh ADASYN memberikan kinerja yang lebih baik.

Klasifikasi dapat diterapkan dalam identifikasi rumah tangga penerima Program Keluarga Harapan (PKH). Menurut Direktorat Jaminan Sosial Keluarga (2020) PKH adalah salah satu bentuk perlindungan sosial yang diberikan pemerintah untuk mengatasi kemiskinan. PKH secara signifikan telah menurunkan angka kemiskinan sebesar 2,2% dan berkontribusi meningkatkan kualitas hidup di Indonesia (Kementerian Sosial RI, 2024). Namun, terdapat tantangan dalam penyaluran bantuan sosial ke depan, terletak pada kemampuan untuk memprediksi rumah tangga yang berhak menerima PKH. Menteri Sosial Tri Rismaharini menyatakan pada tahun 2023 sebanyak 2,2 juta KPM di Indonesia tidak berhak menerima bantuan sosial dan berpotensi kerugian mencapai Rp 140 miliar per bulan (Gandhawangi, 2023). Maka, perlu metode klasifikasi dengan pendekatan prediktif berbasis data untuk memprediksi penerima PKH berdasarkan faktor-faktornya.

Berbagai penelitian terdahulu telah membahas terkait klasifikasi penerima PKH. Menurut Meilaniwati & Fauzan (2022) terdapat 5 variabel yang berperan penting dalam pengklasifikasian PKH yaitu usia, status kehamilan, pendidikan tertinggi kepala rumah tangga, kepemilikan aset bergerak, dan kepemilikan aset tidak bergerak. Penerpan algoritma KNN menghasilkan akurasi yang cukup baik mencapai 76,695%. Pada penelitian Dina dkk., (2023) yang melakukan perbandingan antara algoritma NBC, KNN, dan C4.5 diperoleh algoritma C4.5 menghasilkan akurasi lebih baik yaitu 80,16% dan berhasil mereduksi dari 33 variabel menjadi 8 variabel yaitu jumlah ART, fasilitas BAB, rumah lain, ada emas, ada lemari es, jenis dinding, dan pembuangan tinja.Merujuk pada penelitian terdahulu, variabel yang memberikan kontribusi penting dalam klasifikasi penerima PKH adalah kelompok variabel yang berasal dari variabel Data Terpadu Kesejahteraan Sosial (DTKS). Dimana salah satu syarat penerima program PKH harus terdaftar dalam DTKS. DTKS terdiri atas beberapa kelompok variabel dan yang akan digunakan dalam penelitian ini adalah kelompok identitas rumah tangga, perumahan rumah tangga, aset rumah tangga, demografi anggota rumah tangga (ART), dan pendidikan ART (Helmizar dkk., 2021).

Provinsi Sumatera Barat dipilih sebagai lokasi penelitian. Berdasarkan data Badan Pusat Statistik (BPS) persentase penduduk miskin Provinsi Sumatera Barat mengalami peningkatan pada tahun 2023-2024, hal ini berbanding terbalik dengan tren nasional yang mengalami penurunan. Berdasarkan data Survey Sosial Ekonomi Nasional (SUSENAS) maret 2024 oleh BPS Provinsi Sumatera Barat, jumlah rumah tangga penerima PKH sebanyak 1.790 sedangkan yang tidak menerima sebanyak 9.810. Kondisi ini menunjukkan ketidakseimbangan data yang cukup besar. Sehingga, perlu dilakukan penanganan data tidak seimbang sebelum proses analisis.

Dengan mempertimbangkan jumlah data yang besar dan variabel kompleks, PKH cocok dianalisis menggunakan algoritma XGBoost karena kemampuannya menangani data berskala besar dan menghasilkan performa klasifikasi yang tinggi. Metode ADASYN digunakan untuk menganani data tidak seimbang sebelum proses analisis. Penelitian ini penerapan algoritma XGBoost dengan ADASYN dalam klasifikasi rumah tangga penerima PKH di Provinsi Sumatera Barat. Penelitian ini diharapkan dapat mengetahui performa algoritma XGBoost dan XGBoost dengan ADASYN dalam melakukan klasifikasi, sehingga dapat memperoleh performa terbaik dalam klasifikas dan dapat mengindikasikan variabel yang memberikan kontribusi penting dalam melakukan klasifikasi penerima PKH.

II. METODE PENELITIAN

A. *Extreme Gradient Boosting (XGBoost)*

Extreme Gradient Boosting (XGBoost) merupakan optimalisasi dari *gradient boosting* yang dirancang efisien, dan fleksibel. Algoritma XGBoost menggunakan implementasi paralel untuk mempercepat pencarian *split* terbaik, algoritma ini juga menggabungkan regularisasi dalam fungsi objektif untuk mengurangi *overfitting*. Algoritma XGBoost dapat menangani data *sparse* secara efisien (Chen & Guestrin, 2016). Jika diberikan dataset dengan n amatan dan m variabel, $D = \{(x_i, y_i)\}$ $(|D| = n,\ x_i \in \mathbb{R}^m,\ y_i \in \mathbb{R})$. Pohon algoritma *ensambel* menggunakan k fungsi aditif yang masing-masing mewakili *Classification and Regression Trees* (CART) digunakan untuk memprediksi output. Output yang diprediksi diberikan oleh jumlah dari setiap fungsi prediksi individu berikut :

DOI: https://doi.org/10.24036/ujsds/vol3-iss2/369

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

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Dalam pelatihan model, menemukan parameter terbaik yang sesuai dengan data *training* x_i dan y_i merupakan tugas yang penting. Dalam pelatihan model perlu mendefinisikan fungsi objektif, yang bertujuan untuk mengetahui seberapa baik model tersebut sesuai dengan data *training* dan perlu dilakukannya proses meminimumkan fungsi objektif di setiap iterasinya. Pada XGboost ditambahkan regularisasi untuk mencegah *overfitting* dan fungsi objektif diminimumkan dengan pendekatan *taylor expansion* dari *loss function* hingga *second orde. Package* yang dapat digunakan pada R adalah xgboost.

B. *Adaptive Synthetic (ADASYN)*

Teknik *resampling* dapat digunakan untuk menyeimbangkan kelas data, salah satu metode *resampling* yaitu *oversampling* untuk meningkatkan sampel kelas minoritas. Salah satu metode *oversampling* adalah *Adaptive Synthetic* (ADASYN) yang merupakan pengembangan dari metode (SMOTE) dimana menghasilkan sampel-sampel minoritas secara adaptif sesuai dengan distribusinya, dimana membentuk sampel sintetis dari kelas minoritas yang sulit dipelajari. ADASYN dapat mengurangi pembelajaran bias. *Package* yang dapat digunakan pada R adalah UBL.

Menurut He dkk., (2008) berikut persamaan untuk membangkitkan sampel sintetis dari ADASYN :

$$x_{new} = x_i + (x_{zi} - x_i) \times \lambda$$

Keterangan :

x_{new} : sampel sintetis yang baru
x_i : sampel minoritas asli
x_{zi} : sampel dari k tetangga yang memiliki jarak terdekat dengan x_i
λ : bilangan acak antara 0 dan 1

C. Sumber Data dan Variabel Penelitian

Data penelitian berasal dari data sekunder SUSENAS Provinsi Sumatera Barat Maret 2024. Populasi dalam penelitian yaitu rumah tangga yang berada di Provinsi Sumatera Barat. Sampel penelitian ini adalah rumah tangga yang terpilih menjadi sampel SUSENAS di Provinsi Sumatera Barat, dengan metode *two stages one phase stratified sampling* yaitu sebanyak 11.600 amatan. Variabel yang digunakan pada penelitian ini mengacu pada variabel dari kelompok variabel DTKS yang tersedia di data SUSENAS yaitu Penerima PKH (Y), Jumlah keluarga (X_1) , Jumlah ART(X_2), Luas Lantai Rumah (X_3), Status Kepemilikan Rumah (X_4), Bahan Atap Rumah (X_5), Bahan Dinding Rumah (X_6), Bahan Lantai Rumah (X_7), Fasilitas Sanitasi (X_8), Sumber Penerangan (X_9), Daya Listrik (X_{10}), Bahan Bakar Masak (X_{11}), Memiliki Kulkas (X_{12}), Memiliki AC (X_{13}), Memiliki Motor (X_{14}), Status Nikah (X_{15}), Umur KRT (X_{16}), dan Ijazah (X_{17}).

D. Teknik Analisis Data

Adapun langkah-langkah pada penelitian ini sebagai berikut :

1. Melakukan eksplorasi data, untuk memperoleh gambaran umum dari data.
2. Pembagian data *training* dan *testing*, dengan perbandingan data *training* 70% dan *testing* 30%.
3. Melakukan pembelajaran dengan menggunakan Algoritma XGboost
   a. Inisialisasi prediksi awal $\widehat{y_0} = 0$, gradien (g_i) dengan persamaan $g_i = \partial_{\widehat{y_i}^{(t-1)}}\left(y_i, \hat{y}_i^{(t-1)}\right) = p_i - y_i$ dan *hessian* (h_i) dengan persamaan $h_i = \partial^2{}_{\widehat{y_i}^{(t-1)}}\left(y_i, \hat{y}_i^{(t-1)}\right) = p_i(1-p_i)$
   b. Tentukan *split node* terbaik berdasarkan nilai *gain* maksimum, nilai *gain* dihitung dengan persamaan $Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$, *split node* terbaik memiliki nilai *gain* terbesar, jika *gain* < γ maka dilakukan proses *pruning*.
   c. Tentukan set simpul biner, membagi data menjadi dua subset sisi kanan dan kiri berdasarkan *threshold* terpilih dari langkah (b).

DOI: https://doi.org/10.24036/ujsds/vol3-iss2/369

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

d. Hitung nilai prediksi seluruh simpul daun, berdasarkan *gradien* dan *hessian* dengan persamaan $w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$, Prediksi pohon kedua dihitung dengan persamaan $\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$.

e. Ulangi langkah yang sama seperti langkah a-d untuk membuat lebih banyak pohon hingga mencapai jumlah yang ditentukan, prediksi akan diperbarui secara iteratif dengan persamaan $\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$.

f. Tentukan hasil klasifikasi sampel dengan mengkonversi hasil prediksi menjadi probabilitas kelas. Probabilitas (p_i) dihitung menggunakan rumus fungsi sigmoid $p_i = \frac{1}{1+e^{-\hat{y}_i^{(t)}}}$.

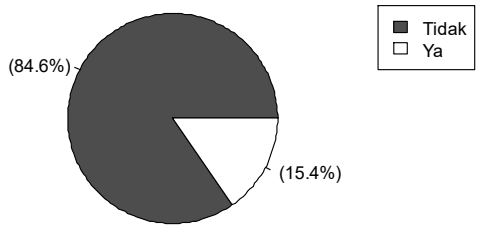
1) Jika $p_i \geq 0.5$, sampel diklasifikasikan sebagai 1
2) Jika $p_i < 0.5$, sampel diklasifikasikan sebagai 0

4. Temukan *hyperparameter* terbaik, menggunakan *gridsearch* dari kombinasi parameter yang telah ditentukan.
5. Evaluasi kinerja algoritma, menggunakan *matrix evaluation* yaitu $Akurasi = \frac{TP+TN}{P+N}$, $Sensitivity = \frac{TP}{P}$, $Specificity = \frac{TN}{N}$, $Balanced\ accuracy = \frac{sensitivity+specivici}{2}$
6. Selanjutnya lakukan pembelajaran XGBoost dengan penanganan data tidak seimbang ADASYN terlebih dahulu.
   a. Hitung perbandingan kelas minoritas dan mayoritas $d = m_s/m_l$.
   b. Jika $d < dth$ (ambang batas toleransi yang ditetapkan, *deafult* 0.95), maka:
      1) Hitung jumlah data sintetis yang perlu dibangkitkan dari kelas minoritas $G = (m_l - m_s) \times \beta$.
      2) Setiap x_i *minority class*, tentukan k data terdekat pada n *dimensional space*, dan hitung rasio $r_i = \frac{\Delta_i}{K}$, $i = 1, \ldots, m_s$
      3) Normaslisasi r_i dengan $\hat{r_i} = \frac{r_i}{\sum_{i=1}^{m_s} r_i}$ sehingga membentuk *density distibution.*
      4) Hitung jumlah sampel sintetis yang dibutuhkan dari masing-masing kelas minoritas menggunakan $g_i = \hat{r_i} \times G$.
      5) Bangkitkan data sampel sintetis menggunakan $x_{new} = x_i + (x_{zi} - x_i) \times \lambda$ dengan memilih acak satu sampel minoritas dari K tetangga terdekat dari x_i.
7. Membagi data menjadi data *training* dan data *testing*, dengan pembagian data *training* 70% dan *testing* 30%.
8. Pembelajaran XGBoost kembali, dan temukan parameter terbaik model.
9. Evaluasi kinerja algoritma XGBoost dengan ADASYN, dan bandingkan dengan XGBoost tanpa ADASYN
10. Tentukan *Variable Importance* untuk mengindentifikasi variabel penting yang berkontribusi dalam klasifikasi.

III. HASIL DAN PEMBAHASAN

A. Eksplorasi Data

Eksplorasi data digunakan untuk mendapatkan gambaran terkait variabel yang digunakan dalam penelitian. Penelitian ini menggunakan 11.600 amatan rumah tangga dengan 1 variabel target dan 17 variabel prediktor. Penelitian ini akan mengklasifikasikan rumah tangga penerima PKH. Gambar 1 menunjukkan visualisasi untuk variabel prediktor yaitu penerima PKH, terdiri atas 2 kategori.

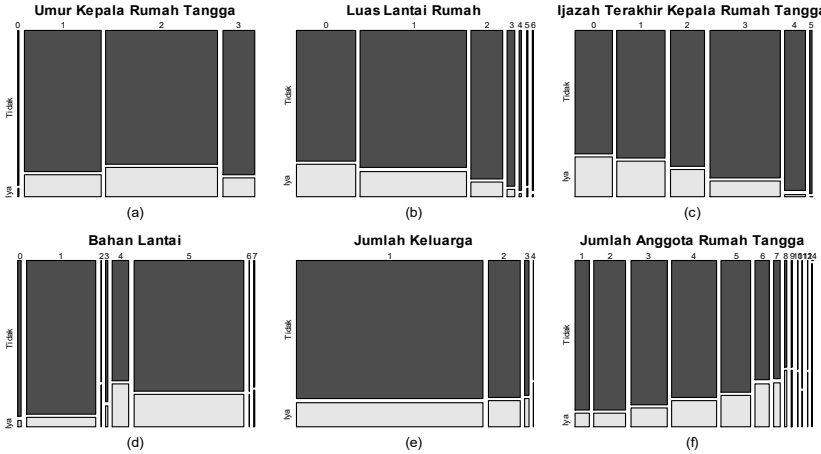


Gambar 1. Persentase Penerima PKH

Rumah tangga penerima bantuan PKH di Provinsi Sumatera Barat sebanyak 15.4% sedangkan 84.6% lainnya tidak menerima PKH, persentase penerima PKH ini jauh lebih sedikit dibandingkan yang tidak menerima, hal ini menunjukkan adanya ketimpangan data pada variabel target PKH. Selanjutnya eksplorasi dilakukan untuk variabel prediktor terhadap variabel target. Eksplorasi ditampilkan dalam *mosaic plot* untuk melihat hubungan kedua variabel yang ditunjukkan Gambar 2.

DOI: https://doi.org/10.24036/ujsds/vol3-iss2/369

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



Gambar 2. Penerima PKH berdasarakan (a) Umur Kepala Rumah Tangga, (b) Luas Lantai Rumah, (c) Ijazah Kepala Rumah Tangga, (d) Bahan Lantai, (e) Jumlah Keluarga, (f) Jumlah Anggota Rumah Tangga

Keterangan kriteria :

(a) 0: ≤25 tahun, 1:26-45 tahun, 2:46-65 tahun, 3:>65 tahun

(b) 0: ≤50 m2, 1:51-100 m^2, 2:101-150 m^2, 3:151-200 m^2, 4:201-250 m^2, 5:251-300 m^2,6: >300 m^2

(c) 0:Tidak Tamat SD, 1:SD/Sederajat, 2:SLTP/Sederajat, 3:SLTA/Sederajat, 4:D1/D2/D3/D4/S1, 5:Profesi/S2/S3

(d) 0:Marmer/Granit, 1:Keramik, 2:Parket/Vinil/Karpet, 3:Ubin/Tegel/Teraso, 4:Kayu/Papan, 5:Semen/Bata Merah, 6:Bambu, 7:Tanah

Rumah tangga yang menerima PKH memiliki proporsi yang lebih tinggi pada kondisi tertentu. Usia Kepala Rumah Tangga (KRT) sekitar 46-65 tahun, luas lantai rumah yang kecil kurang dari 50 m^2, jenis bahan lantai rumah yang digunakan kayu/papan/parket/vinil/karpet, jumlah keluarga dan jumlah anggota rumah tangga yang lebih banyak. Hal ini menunjukkan bahwa terdapat hubungan antara variabel prediktor dengan variabel targetnya penerima PKH.

B. XGBoost Tanpa Penanganan Data Tidak Seimbang

Klasifikasi rumah tangga penerima Program Keluarga Harapan (PKH) menggunakan algoritma XGBoost dapat dilakukan dengan membagi data terlebih dahulu. Penelitian ini membagi secara proporsional kelas PKH 70% untuk data *training* dan 30% untuk data *testing*. Pada penerapan XGBoost data perlu dikonversi ke dalam bentuk Dmatrix yang merupakan struktur data internal dari XGBoost yang dioptimalkan untuk efisiensi memori dan kecepatan pelatihan model. Pembelajaran dilakukan pada data *traininng* dengan algoritma *exact greedy* untuk mempercepat pencarian *split* terbaik. Variabel target yang digunakan berkategori biner maka fungsi objektifnya logistik biner dan dievaluasi dengan logloss.

Dalam pemodelan, pemilihan parameter yang digunakan sangatlah penting. Maka, perlu penyetelan *hyperparameter* sehingga memperoleh performa yang optimal, penerapan *grid search* digunakan untuk menemukan *hyperparameter* dengan melakukan percobaan untuk keseluruhan kombinasi parameter, lalu mengevaluasi dan memilih parameter terbaik berdasarkan kombinasi yang telah ditentukan. Tabel 1, menunjukkan kombinasi dan hasil dari penyetelan *hyperparameter*.

Tabel 1. *Hyperpameter* Terbaik XGBoost Tanpa ADASYN

Parameter	Percobaan	Terbaik
Eta	0.1	0.1
Max_depth	3,4,5,6,7,8	4
Min_child_weight	1,5	5
Subsample	0.6,0.8,1	0.6
Colsample_bytree	0.7,0.8,1	1
Gamma	1	1
Alpha	0.1,1	1
Lambda	0.1,1	0.1
Scale_post_weight	5.5	5.5
nrounds	400	109

DOI: https://doi.org/10.24036/ujsds/vol3-iss2/369

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

Nilai parameter yang digunakan ditetapkan untuk learning_rate 0.1, gamma 1, dan scale_post_weight dapat dihitung dari perbandingan kelas positif dengan negatif yaitu 5.5. Parameter yang dilakukan penyetelan yaitu max_ depth, min_child_weight, subsample, colsample_bytree, alpha, lambda, dan nrounds. Setelah menemukan *hyperparameter* terbaik dilakukan pelatihan menggunakan kombinasi tersebut dan dievaluasi kinerja model pada data *testing* menggunakan kriteria *confussion matrix*. Hasil *confussion matrix* ditunjukkan pada Tabel 2.

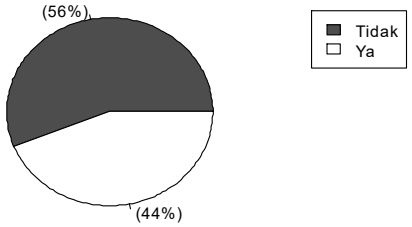
Tabel 2. Evaluasi Model XGBoost Tanpa ADASYN

Evaluasi	Nilai
Accuracy	0.848
Sensitivity	0.123
Specificity	0.981
Balanced Accuracy	0.552

Penerapan algoritma XGBoost, diperoleh sebesar 84.8% model dapat mengklasifikasikan penerima PKH secara keseluruhan, dan sebesar 98.1% model berhasil mengklasifikasikan rumah tangga yang tidak menerima PKH. Hasil ini menunjukkan performa yang sangat baik. Namun, model mengalami kesulitan dalam mengklasifikasikan rumah tangga penerima PKH, dengan keberhasilan 12.3%. sehingga rata-rata ketepatan kedua kelas juga menurun, yaitu sebesar 55.2%. nilai menunjukkan performa yang buruk dalam mengklasifikasikan kelas yang menjadi fokus dalam penelitian. Hal ini dapat terjadi karena data kelas yang tidak seimbang, sehingga model terlalu fokus pada kelas mayoritas yaitu tidak menerima PKH. Sehingga model bias dalam mengklasifikasikan penerima PKH.

C. XGBoost dengan Penanganan Data Tidak Seimbang

Ketimpangan data dalam klasifikasi merupakan suatu tantangan, hal ini dapat mengakibatkan model bias pada salah satu kelas. Jika perbedaan ketepatan kelas cukup jauh maka perlu dilakukan penanganan terlebih dahulu. Penelitian ini menggunakan metode *Adaptive Synthetic* (ADASYN) untuk menambah sampel pada kelas minoritas dengan membuat sampel sintetis berdasarkan distribusi data, sampel yang dibuat merupakan sampel yang sulit dipelajari. *Oversampling* dengan ADASYN menerapkan jarak *Heterogeneous Eucledian-Overlap Metric* (HEOM), beta 1 yang artinya data akan seimbang antara minoritas dan mayoritas, dan tingkat kesulitannya dth = 0.5. Gambar 3 menunjukkan hasil *oversampling* yang telah dilakukan.



Gambar 3. Hasil Oversampling dengan ADASYN

Pada variabel numerik terdapat hasil sampel sintetis yang tidak realistis, sampel tidak realistis ini tidak digunakan dalam penelitian, sampel dibersihkan dengan dideteksi oleh metode *Interquartile Range* (IQR). Penambahkan sampel sintetis pada data, membuat proporsi kelas menjadi lebih seimbang. Setelah dilakukan penanganan terhadap kelas yang tidak seimbang, selanjutnya dilakukan pembagian data secara proposional, *training* 70% dan *testing* 30%. Pembelajaran dilakukan pada data *training* untuk memperoleh model XGBoost perlu menyesuaikan *hyperparameter*nya terlebih dahulu. Untuk menemukan *hyperparameter* terbaik digunakan metode *grid search,* hasil penyetelan *hyperparameter* ditunjukkan pada Tabel 3.

Tabel 3. *Hyperparameter* Terbaik XGBoost dengan ADASYN

Parameter	Percobaan	Parameter Terbaik
Eta	0.1	0.1
Max_depth	3,4,5,6,7,8	6
Min_child_weight	1,5	1
Subsample	0.6,0.8,1	1

DOI: https://doi.org/10.24036/ujsds/vol3-iss2/369

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

Colsample_bytree	0.7,0.8,1	0.7
Gamma	1	1
Alpha	0.1,1	0.1
Lambda	0.1,1	0.1
nrounds	400	375

Kombinasi yang digunakan pada XGBoost dengan ADASYN sama dengan XGBoost tanpa penanganan data tidak seimbang kecuali pada kombinasi ini tidak menggunakan scale_post_weight, karena datanya telah seimbang sehingga tidak perlu pembobotan. Untuk mengukur kinerja model, dapat dilakukan pengujian menggunakan data *testing*. Evaluasi dapat dilihat dari nilai *accuracy, specificity, sensitifity*, dan *balanced accuracy*. Hasil evaluasi dapat dilihat pada Tabel 4.

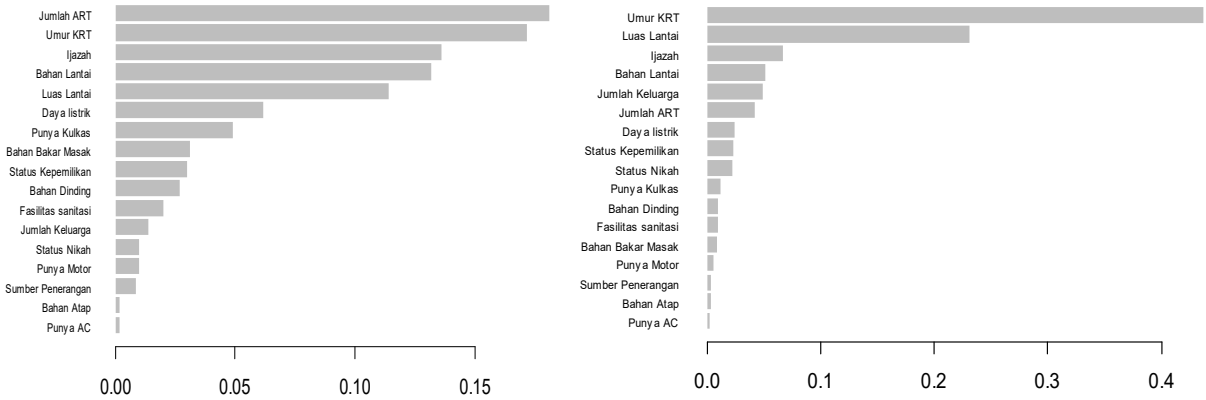
Tabel 4. Evaluasi Model XGBoost dengan ADASYN

Evaluasi	Nilai
Accuracy	0.885
Sensitivity	0.800
Specificity	0.952
Balanced Accuracy	0.876

Pemodelan XGBoost dalam mengklasifikasikan rumah tangga PKH memperoleh performa yang baik setelah dilakukan penanganan terhadap data tidak seimbang. Model ini mampu mengklasifikasikan rumah tangga penerima PKH dengan ketepatan 80% sedangkan pada rumah tangga yang tidak menerima PKH sebesar 95.2%. Sehingga, memperoleh ke akuratan dalam mengklasifikasi status penerima PKH sebesar 88.5% dan rata-rata ketepatan untuk kedua kelasnya adalah 87.6%. ini menunjukkan performa model baik dalam melakukan klasifikasi dan mengalami peningkatan setelah dilakukan ADASYN.

D. *Variable Importance*

Variable importance digunakan untuk mengukur kontribusi relatif dari setiap variabel prediktor dalam memprediksi hasil, pada algoritma XGBoost *variable importance* dapat dilihat dari rata-rata nilai *gain* dari semua pohon pada prediktor tertentu. Berikut *variable importance* pada pemodelan XGBoost dengan ADASYN yang ditunjukkan pada Gambar 4.



Gambar 4. *Variable Importance* pada model (a) XGBoost, (b) XGBoost dengan ADASYN

Dalam mengklasifikasikan Rumah Tangga Penerima PKH di Provinsi Sumatera Barat terdapat 4 variabel teratas yang berkontribusi penting dalam klasifikasi pada kedua model yaitu umur KRT, luas lantai rumah, ijazah terakhir KRT, bahan lantai rumah. Pada model XGboost jumlah ART menjadi variabel yang paling penting, sedangkan setelah dilakukan ADASYN, umur KRT menjadi variabel paling penting dalam klasifikasi rumah tangga penerima PKH.

IV. KESIMPULAN

Klasifikasi Rumah Tangga Penerima PKH di Provinsi Sumatera Barat dengan menggunakan XGBoost tanpa dilakukan penanganan terhadap data tidak seimbang, dengan memanfaatkan pembobotan parameter scale_post_weight

DOI: https://doi.org/10.24036/ujsds/vol3-iss2/369

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

sebagai solusi. Namun, XGBoost dengan parameter scale_post_weight masih belum mampu memberikan performa yang baik dalam mengklasifikasikan penerima PKH. Hal dilihat dari nilai ketepatan dalam mengklasifikasikan rumah tangga penerima PKH (*sensitivity*) sebesar 12.3% dan *balaced accuracy* 55.2%. Untuk mengatasi hal tersebut, digunakan metode ADASYN untuk penanganan data tidak seimbang. XGBoost dengan ADASYN dapat meningkatkan performa model dengan meningkatkan nilai *sensitivity* sebesar 80% dan *balanced accuracy* 87.6%. Ini menunjukkan performa model yang baik dalam melakukan klasifikasi. Dalam klasifikasi terdapat variabel yang memberikan kontribusi penting yaitu Umur KRT, Luas Lantai, Ijazah KRT, dan Bahan Lantai.

DAFTAR PUSTAKA

Badan Pusat Statistik Sumatera Barat. (2024). *Persentase Penduduk Miskin Menurut Kabupaten/Kota di Sumatera Barat (Persen) 2022-2024*. Badan Pusat Statistik. https://sumbar.bps.go.id/id/statistics-table/2/MzQjMg==/persentase-penduduk-miskin-menurut-kabupaten-kota-di-sumatera-barat--persen-.html

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A. L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(2), 1–43.

Boehmke, B., & Greenwell, B. (2020). *Hands-On Machine Learning with R*. CRC Press Taylor & Francis Group.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794.

Dina, A., Permana, I., Muttakin, F., & Maita, I. (2023). Perbandingan Algoritma NBC, KNN, dan C4.5 Untuk Klasifikasi Penerima Bantuan Program Keluarga Harapan. *Jurnal Media Informatika Budidarma*, *7*(3), 1079.

Direktorat Jaminan Sosial Keluarga. (2021). *Pedoman Pelaksanaan Program Keluarga Harapan*. Kementerian Sosial RI.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. In *Springer*.

Gandhawangi, S. (2023). *Bansos Tidak Tepat Sasaran, Negara Merugi Ratusan Miliar Rupiah Per Bulan*. Kompas. https://www.kompas.id/baca/humaniora/2023/09/06/bansos-tidak-tepat-sasaran-negara-merugi-ratusan-miliar-per-bulan

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, *3*, 1322–1328.

Kaope, C., & Pristyanto, Y. (2023). The Effect of Class Imbalance Handling on Datasets Toward Classification Algorithm Performance. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *22*(2), 227–238.

Kementerian Sosial RI. (2024). *PKH Bantu Turunkan Angka Kemiskinan hingga 2,2%*. Kementrian Sosial Republik Indonesia. https://kemensos.go.id/infografis/sekretariat-jenderal/pkh-bantu-turunkan-angka-kemiskinan-hingga-22

Meilaniwati, E. R., & Fauzan, D. M. (2022). Klasifikasi penduduk miskin penerima PKH menggunakan metode naïve bayes dan KNN Classification. *Jurnal Kajian Dan Terapan Matematika*, *8*(2), 75–84.

Perdeck, J. (2024). *XGBoost Performance in Blood Culture Prediction : A Study on Re-sampling Techniques for Imbalanced Data*. Utrecht University.

Wang, K., Li, M., Cheng, J., Zhou, X., & Li, G. (2021). Research on personal credit risk evaluation based on XGBoost. *Procedia Computer Science*, *199*, 1128–1135.

DOI: https://doi.org/10.24036/ujsds/vol3-iss2/369