Studi Komparatif Metode *Boosting* Dalam Pengklasifikasian Penerima Bantuan Program Keluarga Harapan (PKH)

Comparative Study of Boosting Methods in Classifying Recipients of the Family Hope Program (FHP)

Fida Fariha Amatullah[1*], Hadyanti Utami MY[2], Tasya Anisah Rizqi[3], Silvia Tri Wahyuni[4], Bagus Sartono[5], Aulia Rizki Firdawanti[6]

[1,2,3,4,5,6]Prodi Statistika dan Sains Data, Sekolah SMI, IPB University, Bogor, 16680, Indonesia

fidafarihaafida@apps.ipb.ac.id[1*], utamihadyanti@apps.ipb.ac.id[2], tasya.anisah.rizqi@apps.ipb.ac.id[3], silviatriwahyuni@apps.ipb.ac.id[4], bagusco@apps.ipb.ac.id[5], arfirdawanti@gmail.com[6]

Abstrak – *Ensemble Learning adalah paradigma pembelajaran mesin dimana beberapa model (biasanya disebut "weak learners") dilatih untuk memecahkan masalah yang sama dan digabungkan untuk mendapatkan hasil yang lebih baik. Salah satu model Ensemble, yaitu model boosting. Beberapa metode boosting yang digunakan dalam penelitian ini, yaitu Gradient Boosting Machines (GBM), Extreme Gradient Boosting Machine (XGBM), Light Gradient Boosting Machine (LGBM), dan CatBoost. Penelitian ini akan mengklasifikasikan Rumah Tangga (RT) yang menerima bantuan Program Keluarga Harapan (PKH). Pengklasifikasian PKH sangat penting dilakukan, karena saat ini pemberian PKH belum optimal dan masih banyak yang tidak tepat sasaran. Hasil penelitian menunjukkan bahwa metode LGBM menunjukkan performa terbaik ketika jumlah data latih berukuran besar, yaitu 90% dengan akurasi sebesar 67,97%, sedangkan untuk data latih kecil yaitu 60:40, LGBM memiliki performa yang kurang baik, dengan nilai balanced accuracy terendah dibandingkan metode boosting lainnya, yaitu sebesar 54,43%. Keunggulan LGBM ini disebabkan karena kemampuannya dalam mengelola data besar dan kompleks yang sesuai dengan karakteristik data sosial ekonomi rumah tangga penerima PKH. Dua fitur yang memiliki peran penting untuk pengklasifikasian PKH dalam model terbaik yaitu LGBM adalah faktor ekonomi dan jumlah anggota rumah tangga.*

Kata Kunci*: Boosting, Ensemble Learning, Program Keluarga Harapan*

Abstract – *Ensemble Learning is a machine learning paradigm in which multiple models (commonly referred to as "weak learners") are trained to solve the same problem and combined to achieve better results. One of the Ensemble models is the boosting model. Several boosting methods used in this study include Gradient Boosting Machines (GBM), Extreme Gradient Boosting Machine (XGBM), Light Gradient Boosting Machine (LGBM), and CatBoost. This study aims to classify households (RT) that receive assistance from the Program Keluarga Harapan (PKH). The classification of PKH recipients is crucial because the distribution of PKH aid has not been optimal, with many cases of misallocation. The results of the study indicate that the LGBM method demonstrates the best performance when the* latih *dataset is large (90%), achieving an accuracy of 67.97%. However, when the* latih *dataset is small (60:40), LGBM performs poorly, recording the lowest balanced accuracy among the boosting methods, at 54.43%. The superiority of LGBM is attributed to its ability to handle large and complex data, which aligns with the socio-economic characteristics of PKH recipient households. Two key features that play a significant role in PKH classification using the best-performing model, LGBM, are economic factors and the number of household members.*

Keywords*: Boosting, Ensemble Learning, Family Hope Program*

1. Pendahuluan

Pada era digital saat ini teknologi sangat membantu dalam proses pengambilan keputusan. Sebuah keputusan dibuat dengan mempertimbangkan beberapa faktor seperti pembelajaran mesin atau *Machine Learning* (ML) yang dapat membantu proses pengambilan keputusan. *Machine Learning* (ML) melibatkan pemberian kecerdasan pada sistem komputer dengan menerapkan berbagai teknik pemrograman dan statistik. Mengambil konsep dari bidang-bidang seperti statistik, linguistik komputasional, neurosains serta menerapkan statistik modern dan pemrograman dasar [1].

Saat ini telah banyak penelitian yang menggunakan ML, salah satu yang populer yaitu metode *ensemble*. *Ensemble Learning* adalah paradigma pembelajaran mesin yang melatih model tertentu atau biasanya disebut dengan *weak learners* untuk memperoleh hasil yang lebih baik dengan menyelesaikan gabungan beberapa permasalahan yang sama [2]. *Ensemble Learning* terdiri dari beberapa model yaitu *Boosting*, *Bagging*, dan *Stacking*. Pada penelitian yang dilakukan oleh [2], model *boosting* memiliki akurasi yang lebih tinggi dibandingkan dengan model *bagging* dan juga *stacking*.

Berdasarkan pernyataan pada paragraf sebelumnya, paper ini akan melakukan studi komparatif berbagai metode yang menggunakan model *boosting*. Beberapa metode yang digunakan yaitu *Gradient Boosting Machine* (GBM), *Light Gradient Boosting Machine* (LGBM), *Extreme Gradient Boosting Machine* (XGBM), dan *CatBoost*. Penelitian ini akan diterapkan pada data empiris untuk mengklasifikasikan Rumah Tangga (RT) yang memperoleh bantuan dari Program Keluarga Harapan (PKH).

PKH adalah program bantuan rumah tangga miskin dari Kementerian Sosial dengan sasaran seperti balita, anak sekolah, disabilitas, ibu hamil, dan lansia [3]. Pengklasifikasian PKH sangat penting dilakukan, karena saat ini pemberian PKH belum optimal dan masih banyak yang tidak tepat sasaran. Ketidaktepatan dalam pemilihan penerima PKH membuat banyak masyarakat yang sangat membutuhkannya merasa terabaikan [4].

Paper ini mengusulkan penelitian terkait studi komparatif berbagai metode *boosting* pada data PKH, karena belum banyak penelitian yang menerapkan metode *boosting* pada data PKH. Penelitian terdahulu yang menggunakan data PKH dilakukan oleh [5] menggunakan metode *decision tree*, *super vector machine*, *naive bayes*, dan regresi logistik. Selain itu [6] juga melakukan penelitian dengan data PKH menggunakan metode *naive bayes*. Data PKH juga digunakan oleh [3] dalam penelitiannya menggunakan metode *decision tree C.*45. Penelitian terbaru menggunakan data PKH dilakukan oleh [7] dengan metode *support vector machine*. Beberapa penelitian tersebut menerapkan metode klasifikasi konvensional yang umum digunakan dalam analisis data, seperti regresi logistik, *decision tree*, *naive bayes*, dan *support vector machine*. Namun, hingga saat ini belum terdapat sebuah penelitian yang menerapkan metode *boosting* secara spesifik seperti CatBoost, *Gradient Boosting*, XGBoost, atau LightGBM pada data PKH.

Secara karakteristik, data PKH bersifat kompleks dan tidak seimbang karena memuat berbagai atribut sosial ekonomi yang saling berkaitan. Kondisi ini memerlukan metode klasifikasi yang mampu menangani data multidimensi dan distribusi kelas yang tidak seimbang. Metode *boosting* dikenal efektif dalam menangani tantangan tersebut, sebagaimana ditunjukkan dalam penelitian sebelumnya yang mengaplikasikannya pada klasifikasi bantuan sosial [8], sehingga relevan diterapkan pada konteks data PKH.

Penelitian terdahulu yang telah melakukan perbandingan metode *boosting* dilakukan oleh [9] membandingkan metode AdaBoost dan XGBoost/XGBM terhadap resiko kredit yang memperoleh hasil bahwa metode XGBoost yang terbaik dengan nilai AUC sebesar 0,92 sedangkan AdaBoost sebesar 0,89. Selain itu studi komparatif antara XGBoost dan CatBoost juga telah dilakukan pada data polusi udara, penelitian ini memberikan hasil bahwa metode XGBoost memberikan hasil yang lebih baik dengan nilai R^2 sebesar 0,634 sedangkan CatBoost sebesar 0,629 [10]. Penelitian lain melakukan perbandingan antara AdaBoost dan *Gradient Boosting* pada data penyakit jantung yang memberikan hasil bahwa metode *gradient boosting* yang terbaik dengan nilai akurasi sebesar 89,5% , sedangkan AdaBoost hanya sebesar 88,1%.

ISSN (e): 2540-9123
ISSN (p): 2502-1982

Dari uraian diatas, paper ini memiliki tujuan untuk mendapatkan metode *boosting* terbaik dalam proses pengklasifikasian PKH dan diharapkan metode tersebut dapat meminimalisir ketidaktepatan dalam pemberian PKH. Selain itu akan dicari tahu juga fitur penting yang mampu memberikan hasil klasifikasi terbaik pada model terbaik yang dihasilkan.

2. Metode Penelitian

2.1. Data

Penelitian ini menggunakan data Survei Sosial Ekonomi Nasional (SUSENAS) Provinsi Jawa Barat yang diperoleh dari Badan Pusat Statistik (BPS) Provinsi Jawa Barat sebanyak 25.890 rumah tangga. Data dibagi menjadi peubah respon dan peubah penjelas, dimana Program Keluarga Harapan (PKH) akan menjadi peubah respon dan dijelaskan oleh 23 peubah penjelas yang disajikan pada Tabel 1.

Tabel 1. Daftar peubah

Peubah	Keterangan	Skala
Y	Program Keluarga Harapan	Nominal
X1	Status Pekerjaan Kepala Rumah Tangga (KRT)	Nominal
X2	Luas Lantai Tempat Tinggal	Numerik
X3	Jenis Dinding Tempat Tinggal	Nominal
X4	Jenis Lantai Tempat Tinggal	Nominal
X5	Sumber Air Minum Utama	Nominal
X6	Sumber Penerangan Utama	Nominal
X7	Rata-rata Pengeluaran Perkapita	Numerik
X8	Status Kepemilikan Tempat Tinggal	Nominal
X9	Status Kepemilikan Tanah	Nominal
X10	Status Kepemilikan Mobil	Nominal
X11	Jumlah Anggota Rumah Tangga (ART)	Numerik
X12	Status Kepemilikan Tabung Gas 5.5kg	Nominal
X13	Status kepemilikan Kulkas	Nominal
X14	Status Kepemilikan AC	Nominal
X15	Status Kepemilikan Pemanas Air	Nominal
X16	Status Kepemilikan Telepon Rumah	Nominal
X17	Status Kepemilikan Laptop	Nominal
X18	Status Kepemilikan Perhiasan	Nominal
X19	Status Kepemilikan Sepeda Motor	Nominal
X20	Status Kepemilikan Televisi Layar Datar	Nominal
X21	Jumlah Lansia	Numerik
X22	Jumlah Balita	Numerik
X23	Jumlah ART yang Mengalami Disabilitas	Numerik

2.2. Prosedur Analisis

Analisis dilakukan menggunakan metode *boosting* dengan 4 model yaitu *Gradient Boosting Machine* (GBM), *Light Gradient Boosting Machine* (LGBM), *Extreme Gradient Boosting Machine* (XGBM), dan *Categorical Boosting* (CatBoost). Tahapan analisis adalah sebagai berikut:

1. Mempersiapkan data dengan melakukan *cleaning* dan agregat fitur yang diperlukan.
2. Melakukan eksplorasi data untuk mengetahui distribusi peubah Y.
3. Melakukan pembagian data menjadi data latih dan data uji dengan beberapa skenario, yaitu 60%, 70%, 80%, dan 90% yang dialokasikan sebagai data latih dan 40%, 30%, 20%, dan 10% sisanya sebagai data uji.
4. Melakukan *Pra-Processing* data dengan melakukan penanganan pada data kelas tidak seimbang menggunakan *Random Oversampling* (ROS) pada data latih. ROS digunakan untuk mengatasi ketidakseimbangan kelas dengan menghitung selisih jumlah data antara kelas mayoritas dan minoritas, lalu menambahkan data minoritas ke data latih hingga jumlahnya setara dengan kelas mayoritas [11].
5. Melakukan *hyperparameter tuning* untuk mencari *hyperparameter* terbaik dari masing-masing model dengan metode *grid search. Grid search* merupakan metode sederhana

ISSN (e): 2540-9123
ISSN (p): 2502-1982

untuk mengoptimalkan *hyperparameter* dengan mencoba semua kombinasi yang memungkinkan dari subset ruang *hyperparameter* [12].

6. Mengklasifikasikan rumah tangga penerima PKH dengan algoritma *boosting* terhadap data latih dan data uji dengan penanganan dan tanpa penanganan berdasarkan parameter terbaik dengan proses validasi silang K-fold = 10. Beberapa algoritma *boosting* yang digunakan yaitu:
   a. *Gradient Boosting Machine* (GBM)
   GBM adalah metode yang mengadopsi prinsip *boosting* untuk meminimalkan fungsi kerugian dengan cara memperbaiki model awal menggunakan pohon regresi berukuran tetap. Prediksi akhir pada *gradient boosting* dilakukan secara aditif [13]. Adapun elemen penting yang diperlukan meliputi: fungsi kerugian $L(y_i, F(x))$, yang bersifat dapat diturunkan untuk meminimalisasi data berupa peubah bebas dan peubah respon $\{(x_i, y_i)\}_1^n$, jumlah iterasi (M), serta *learning rate* (v). Jumlah iterasi (M) mencerminkan jumlah pohon yang akan dibentuk dalam model dengan nilai *default* sebesar 1000.
   b. *Extreme Gradient Boosting Machine* (XGBoost)
   XGBoost dikenal dengan kemampuannya yang unggul dalam mengatasi permasalahan klasifikasi dan regresi sehingga menghasilkan kinerja yang sangat baik [14]. Algoritma ini dirancang untuk mencegah *overfitting* sekaligus mengoptimalkan efisiensi komputasi. Hal ini dicapai dengan menggabungkan fungsi prediktif dan regularisasi yang mengontrol kompleksitas model dan menyederhanakan fungsi objektif. Fungsi objektif dalam klasifikasi biner menentukan bagaimana XGBoost menghitung kesalahan prediksi dan memperbarui model untuk meningkatkan akurasi, yang mencakup dua komponen utama, yaitu fungsi kerugian dan regularisasi.
   c. *Light Gradient Boosting Machine* (LGBM)
   LGBM dikenal akan efisiensinya dalam mengolah dataset berukuran besar serta kecepatan melakukan pelatihan, terutama dalam konteks lingkungan dengan kebutuhan skala besar [15]. Metode ini menggunakan pendekatan pemisahan berbasis daun (*leaf-wise*) untuk membangun pohon keputusan dan memungkinkan pohon berkembang lebih dalam (*depth-wise*) secara efisien, karena pemisahan dilakukan terlebih dahulu pada *node* yang memberikan penurunan terbesar dalam fungsi kerugian atau peningkatan informasi [16].
   d. *Categorical Boosting* (CatBoost)
   CatBoost memiliki kelebihan dalam mengelola data kategorikal tanpa memerlukan pra-pemrosesan tambahan [17]. Metode ini dirancang untuk meminimalkan perubahan estimasi yang terjadi selama proses pelatihan. Perubahan distribusi ini merujuk pada perbedaan antara $F(x)|x$ untuk data uji x dan muncul karena *gradient boosting* menggunakan sampel yang sama untuk menghitung gradien dan membangun model yang meminimalkan gradien tersebut. Sebagai solusi, *CatBoost* memperkirakan gradien menggunakan serangkaian model dasar yang secara eksplisit mengecualikan sampel tersebut dari himpunan pelatihannya, sehingga mengurangi perubahan estimasi.
7. Mengevaluasi kinerja algoritma *boosting* menggunakan *confusion matrix* dengan nilai *balanced accuracy*, sensitivitas, dan spesifisitas. *Confusion matrix* pada dasarnya berisi informasi yang membandingkan hasil klasifikasi jumlah data uji yang bernilai benar dan salah [18]. Tabel 2 merupakan *confusion matrix* kelas biner yang akan digunakan.

Tabel 2. *Confusion matrix*

Actual Value	**Prediction**	
	Positive	**Negative**
Positive (1)	*True Positive (TP)*	*False Negative (FN)*
Negative (0)	*False Positive (FP)*	*True Negative (TN)*

Perhitungan menggunakan Tabel 2 memungkinkan evaluasi terhadap efektivitas kinerja model dalam melakukan klasifikasi.

Akurasi: Rasio proporsi yang diprediksi dengan benar dibandingkan dengan total keseluruhan prediksi yang dibuat. Akurasi dapat dihitung menggunakan persamaan berikut:

$$Akurasi = \frac{TP + FN}{FP + FN + TP + TN} \quad (1)$$

Sensitivitas: Rasio proporsi data dari kelas positif yang berhasil diprediksi dengan benar sebagai kelasnya. Semakin baik kinerja model, maka nilai sensitivitas semakin mendekati 1. Sensitivitas dapat dihitung menggunakan persamaan berikut:

$$Sensitivitas = \frac{TP}{FN + TP} \quad (2)$$

Spesifisitas: Rasio proporsi data dari kelas negatif yang berhasil diprediksi dengan benar sebagai kelasnya. Semakin baik kinerja model, maka nilai spesifisitas semakin mendekati 1. Spesifisitas dapat dihitung menggunakan persamaan berikut:

$$Spesifisitas = \frac{TN}{FP + TN} \quad (3)$$

8. Membandingkan 4 model algoritma *boosting*, sehingga mendapatkan model terbaik yang dapat digunakan untuk klasifikasi rumah tangga penerima PKH.
9. Analisis fitur penting yang digunakan oleh algoritma *boosting* terbaik.

3. Hasil dan Pembahasan

3.1. Eksplorasi Data

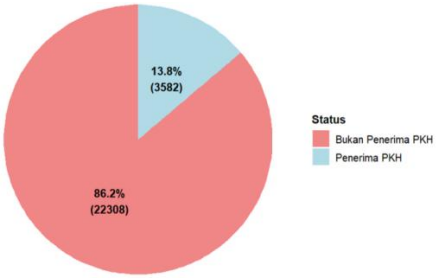
Analisis statistika deskriptif terlebih dahulu dilakukan pada 6 peubah penjelas berskala numerik. Ringkasan statistika deskriptif disajikan pada Tabel 3.

Tabel 3. Statistika deskriptif fitur numerik

Peubah	*Minimum*	*Maximum*	Rata-rata
X2	3	935	71.93
X7	215.471	100.801.869	1.855.078
X11	1	13	3
X21	0	4	0
X22	0	4	0
X23	0	1	0

Berdasarkan Tabel 3, diperoleh luas lantai tempat tinggal memiliki rentang yang sangat lebar merepresentasikan bahwa sebagian besar rumah tangga memiliki luas lantai yang cukup kecil pada populasi dengan pendapatan menengah ke bawah. Pada pengeluaran per kapita terdapat perbedaan yang sangat besar antara pengeluaran terendah dan tertinggi yang menunjukkan adanya kesenjangan ekonomi yang ekstrem. Sebagian besar rumah tangga memiliki anggota yang relatif kecil sehingga mencerminkan bahwa rumah tangga dengan anggota yang banyak merupakan kasus yang jarang didapat. Sebagian besar rumah tangga dengan lansia, balita, dan anggota yang mengalami disabilitas merupakan minoritas. Jumlah penerima dan bukan penerima PKH disajikan pada Gambar 1.

ISSN (e): 2540-9123
ISSN (p): 2502-1982

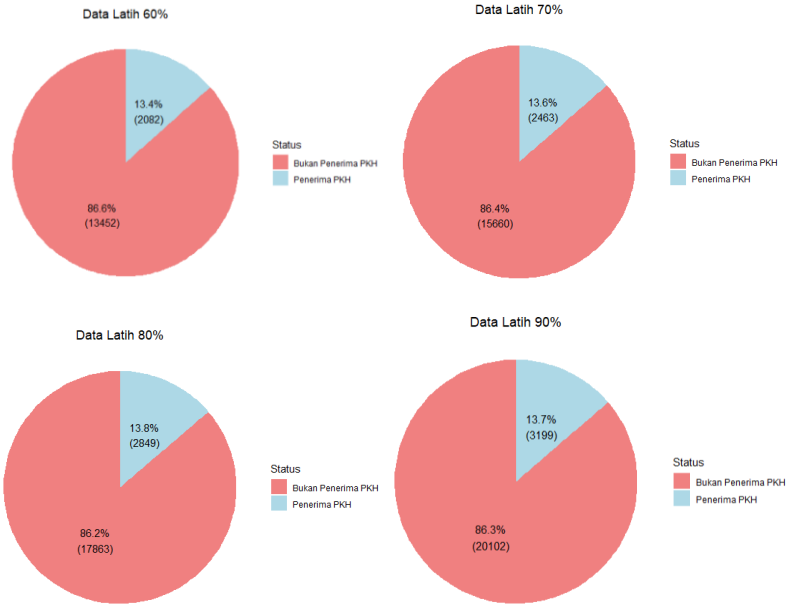


Gambar 1. Persentase rumah tangga penerima PKH

Gambar 1 menampilkan persentase rumah tangga penerima PKH relatif kecil dengan jumlah penerima PKH sebanyak 13,8% yaitu 3.582 rumah tangga dan 86,2% sisanya bukan penerima PKH dengan jumlah rumah tangga sebanyak 22.308.

3.2. Pemodelan Awal

Pada pemodelan awal, data akan dilakukan pembagian menjadi data latih dan data uji dengan perbandingan 60:40, 70:30, 80:20, dan 90:10. Data latih yang diperoleh akan digunakan untuk melatih model sedangkan data uji yang diperoleh akan digunakan untuk mengevaluasi model. Jumlah penerima dan bukan penerima PKH pada data latih disajikan pada Gambar 2.



Gambar 2. Persentase rumah tangga penerima PKH pada data latih

Gambar 2 menunjukkan persentase rumah tangga penerima dan bukan penerima PKH pada data latih yang dihasilkan dari empat skenario pembagian data. Dari keempat skenario tersebut terlihat memiliki kelas yang tidak seimbang, dengan selisih rumah tangga penerima PKH dan bukan penerima PKH cukup besar. Pada pemodelan awal ini data tidak dilakukan penanganan kelas tidak seimbang, namun dilakukan *hyperparameter tuning* untuk memperoleh *hyperparameter* terbaik dari masing-masing model *boosting* dengan proses *gridsearch*. Hasil *balanced accuracy* tanpa penanganan data yang tidak seimbang disajikan pada Tabel 4.

Pemodelan dengan perbandingan 90:10 pada data latih dan data uji tanpa penanganan data tidak seimbang pada Tabel 4 menghasilkan *balanced accuracy* tertinggi menggunakan model CatBoost sebesar 51,32%. Hasil evaluasi model dengan *hyperparameter* terbaiknya disajikan pada Tabel 5.

Tabel 4. Hasil *balanced accuracy* tanpa penanganan data tidak seimbang

Model	*Balanced Accuracy*			
	60:40	**70:30**	**80:20**	**90:10**
GBM	**0,5102**	0,50673	**0,50898**	0,509596
XGBM/XGBoost	0,5046	0,50468	0,50034	0,502402
LGBM	0,5	0,5	0,5	0,5
CatBoost	0,50833	**0,50977**	0,50887	**0,51316**

Tabel 5. Evaluasi model tanpa penanganan data tidak seimbang

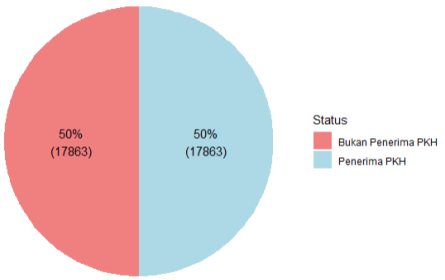
Pembagian data	Evaluasi	Model			
		Catboost	**GBM**	**XGBM**	**LGBM**
60:40	Balanced Accuracy	0,50833	0,5102	0,5046	0,5
	Sensitivity	0,99665	0,99584	0,9976	1
	Specificity	0,02001	0,02456	0,0116	0
	Hyperparameter	depth: 4, learning_rate: 0.05, iterations: 200, l2_leaf_reg: 1, rsm: 0.8, border_count: 64	n_trees: 150, interaction_depth: 3, shrinkage: 0.1, n_minobsinnode: 10	nrounds: 100, max_depth: 3, eta: 0.1, gamma: 1, colsample_bytree: 0.3, min_child_weight: 10, subsample: 0.8	num_leaves: 31, learning_rate: 0.01, n_estimators: 100
70:30	Balanced Accuracy	0,50977	0,50673	0,50468	0,5
	Sensitivity	0,99629	0,99492	0,99731	1
	Specificity	0,02326	0,01855	0,01206	0
	Hyperparameter	depth: 4, learning_rate: 0.1, iterations: 100, l2_leaf_reg: 1, rsm: 0.8, border_count: 32	n_trees: 150, interaction_depth: 3, shrinkage: 0.1, n_minobsinnode: 10	nrounds: 50, max_depth: 3, eta: 0.2, gamma: 1, colsample_bytree: 0.3, min_child_weight: 10, subsample: 0.7	num_leaves: 31, learning_rate: 0.01, n_estimators: 100
80:20	Balanced Accuracy	0,50887	0,50898	0,50034	0,5
	Sensitivity	0,99599	0,99621	0,99955	1
	Specificity	0,02174	0,02174	0,001127	0
	Hyperparameter	depth: 4, learning_rate: 0.05, iterations: 100, l2_leaf_reg: 1, rsm: 1, border_count: 64	n_trees: 150, interaction_depth: 3, shrinkage: 0.1, n_minobsinnode: 10	nrounds: 50, max_depth: 3, eta: 0.2, gamma: 1, colsample_bytree: 0.3 min_child_weight: 10, subsample: 0.7	num_leaves: 31, learning_rate: 0.01, n_estimators: 100
90:10	Balanced Accuracy	0,51316	0,509596	0,502402	0,5
	Sensitivity	0,994181	0,99686	0,999106	1
	Specificity	0,008451	0,02226	0,005698	0

Pembagian data	Evaluasi	Model			
		Catboost	GBM	XGBM	LGBM
	Hyperparameter	depth: 4, learning_rate: 0.05, iterations: 200, l2_leaf_reg: 1, rsm: 0.8, border_count: 64	n_trees: 150, interaction_depth: 3, shrinkage: 0.1, n_minobsinnode: 10	nrounds: 50, max_depth: 3, eta: 0.2, gamma: 5, colsample_bytree: 0.3, min_child_weight: 5, subsample: 0.8	num_leaves: 100, learning_rate: 0.1, n_estimators: 200

Tabel 5 merupakan hasil evaluasi model data sebelum dilakukan penanganan kelas tidak seimbang (data tidak seimbang) dengan proses validasi silang pada data latih pada K = 10, diketahui keempat model *boosting* memiliki akurasi yang tinggi namun nilai sensitivitas yang menunjukkan sebagai derajat keandalan model dalam memprediksi kelas minoritas (1 = penerima PKH) bernilai 99% - 100% atau memprediksi data hampir selalu benar, namun dalam memprediksi kelas negatif (0 = bukan penerima PKH) bernilai kurang dari 1% yang berarti model hampir tidak mampu mengenali kelas negatif, hal ini terjadi untuk setiap pembagian data dengan perbandingan 60:40, 70:30, 80:20, dan 90:10, sehingga menginterpretasikan bahwa pembagian data pada keempat model mengalami *ovefitting* yang disebabkan oleh ketidakseimbangan kelas.

3.3. Model Akhir

Pada pemodelan awal, dihasilkan bahwa model yang terbentuk mengalami *overfitting* yang disebabkan oleh ketidakseimbangan kelas. Oleh karena itu, perlu dilakukan penanganan data tidak seimbang agar dapat meningkatkan sesitivitas terhadap kelas minoritas dan memberikan nilai sensitivitas yang lebih dapat dipercaya. Hasil distribusi status PKH setelah penanganan data yang tidak seimbang pada data latih disajikan pada Gambar 3.



Gambar 3. Jumlah status PKH setelah penanganan data tidak seimbang pada data latih

Gambar 3 menggambarkan distribusi jumlah status PKH setelah dilakukan penanganan data yang tidak seimbang dengan metode *resample*. Metode *resample* yang digunakan adalah *oversampling* sehingga diperoleh jumlah penerima dan bukan penerima PKH masing-masing sebanyak 17.863 rumah tangga sehingga jumlah kelas penerima PKH maupun bukan penerima PKH adalah sama. Metode *oversampling* diketahui mampu memberikan akurasi paling tinggi dibandingkan metode *resample* yang lain pada model [19].

Sama halnya dengan proses pemodelan awal, pada proses pemodelan akhir juga dilakukan *hyperparameter tuning* untuk memperoleh hyperparameter terbaik dari masing-masing model *boosting* dengan metode *gridsearch*. Hasil *balanced accuracy* setelah penanganan data tidak seimbang disajikan pada Tabel 6.

Tabel 6. Hasil *balanced accuracy* setelah penanganan data tidak seimbang data menggunakan *random oversampling*

Model	*Balanced Accuracy*			
	60:40	**70:30**	**80:20**	**90:10**
GBM	**0,6442**	**0,6613**	0,64885	0,6498
XGBM/XGBoost	0,5971	0,6178	0,6043	0,6307
LGBM	0,5443	0,6581	0,6455	**0,6797**
CatBoost	0,6323	0,6448	**0,6656**	0,502211

Pemodelan dengan perbandingan 90:10 pada data latih dan data uji setelah penanganan data tidak seimbang pada Tabel 6 menghasilkan *balanced accuracy* tertinggi menggunakan model LGBM sebesar 67,97%. Hasil evaluasi model dengan *hyperparameter* terbaiknya disajikan pada Tabel 7.

Tabel 7. Evaluasi model setelah penanganan data tidak seimbang data menggunakan *random oversampling*

Pembagian data	Evaluasi	Model			
		Catboost	**GBM**	**XGBM**	**LGBM**
60:40	Balanced Accuracy	0,6323	0,6442	0,5971	0,5443
	Sensitivity	0,7535	0,6949	0,8129	0,9153
	Specificity	0,5111	0,5935	0,3813	0,1733
	Hyperparameter	depth: 8, learning_rate: 0.1, iterations: 200, l2_leaf_reg: 1, rsm: 1, border_count: 32	n_trees: 150, interaction_depth: 5, shrinkage: 0.3, n_minobsinnode: 5	nrounds: 100, max_depth: 9, eta: 0.2, gamma: 1, colsample_bytree: 0.9, min_child_weight: 5, subsample: 0.7	num_leaves: 100, learning_rate: 0.1, n_estimators: 500
70:30	Balanced Accuracy	0,6448	0,6613	0,6178	0,6581
	Sensitivity	0,7537	0,689	0,7986	0,6381
	Specificity	0,5359	0,6336	0,4369	0,6781
	Hyperparameter	depth: 8, learning_rate: 0.1, iterations: 200, l2_leaf_reg: 1, rsm: 1, border_count: 64	n_trees: 150, interaction_depth: 5, shrinkage: 0.3, n_minobsinnode: 5	nrounds: 100, max_depth: 9, eta: 0.2, gamma: 1, colsample_bytree: 0.9, min_child_weight: 5, subsample: 0.8	num_leaves: 31, learning_rate: 0.01, n_estimators: 100
80:20	Balanced Accuracy	0,6656	0,64885	0,6043	0,6455
	Sensitivity	0,637	0,676	0,7897	0,6243
	Specificity	0,6942	0,6217	0,6043	0,6667
	Hyperparameter	Depth = 8, learning_rate = 0.1, iterations = 200, l2_leaf_reg = 1, rsm = 1, border_count = 32	n.trees = 100, interaction.depth = 3, shrinkage = 0.1, n.minobsinnode = 10	nrounds = 100, max_depth = 9, eta = 0.2, gamma = 1, colsample_bytree = 0.9, min_child_weight = 5, subsample = 0.7	Num_leaves = 100, learning_rate = 0.1, n_estimators = 500

ISSN (e): 2540-9123
ISSN (p): 2502-1982

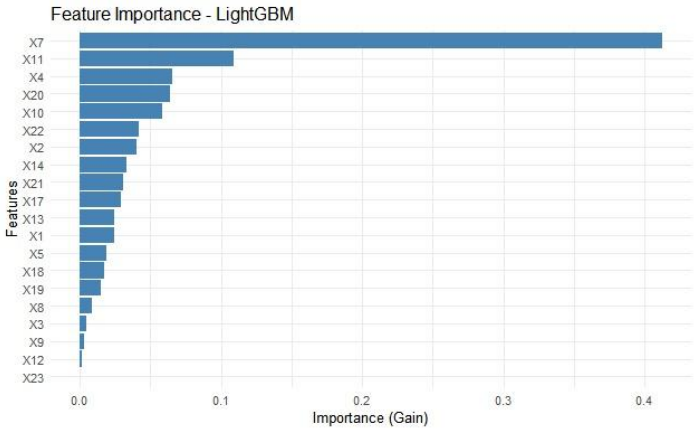
Pembagian data	Evaluasi	Model			
		Catboost	**GBM**	**XGBM**	**LGBM**
90:10	Balanced Accuracy	0,502211	0,6498	0,6307	0,6797
	Sensitivity	0,995971	0,6671	0,7828	0,613
	Specificity	0,008451	0,6325	0,4786	0,7464
	Hyperparameter	depth: 4, learning_rate: 0.1, iterations: 100, l2_leaf_reg: 5, rsm: 0.8, border_count: 32	n_trees: 150, interaction_depth: 5, shrinkage: 0.3, n_minobsinnode: 10	nrounds: 100, max_depth: 9, eta: 0.2, gamma: 1, colsample_bytree: 0.9, min_child_weight: 5, subsample: 0.8	num_leaves: 31, learning_rate: 0.01, n_estimators: 100

Evaluasi model menggunakan penanganan data tidak seimbang pada Tabel 7 menghasilkan akurasi yang lebih tinggi dari sebelumnya. Pada sensitivitas dan spesifisitas menunjukkan nilai yang lebih seimbang yang berarti model dapat memprediksi kelas minoritas maupun mayoritas dengan baik. Penggunaan pembagian data latih dan data uji yang berbeda-beda yaitu 60:40, 70:30, 80:20, dan 90:10 secara berurutan menghasilkan model terbaik yaitu model GBM dengan nilai *balanced accuracy* sebesar 64,42%, GBM 66,13%, CatBoost 66,56%, dan LGBM 67,97%.

Meskipun *balanced accuracy* dan spesifisitas meningkat, terjadi penurunan sensitivitas setelah diterapkannya *Random Oversampling* (ROS). Hal ini disebabkan oleh duplikasi sampel kelas minoritas tanpa penambahan informasi baru yang dapat meningkatkan risiko *overfitting* serta membuat model lebih konservatif dalam mengklasifikasikan kelas positif [20]. Oleh karena itu, meskipun ROS efektif dalam meningkatkan performa model secara keseluruhan, dampaknya terhadap sensitivitas perlu diperhatikan dalam pemilihan strategi penanganan ketidakseimbangan data.

3.4. Analisis Fitur Penting (*Variable Importance*)

Variable importance digunakan sebagai pengukuran seberapa sering peubah tersebut muncul dalam pohon keputusan. Semakin besar nilai *gain* dalam mengurangi ketidakpastian pada pemisahan data, maka semakin besar kontribusi peubah tersebut dalam model pohon keputusan [21]. Hasil analisis fitur penting pada model LGBM disajikan pada Gambar 4.



Gambar 4. *Variable importance*

Berdasarkan gambar 4, diketahui 3 peubah tertinggi yang berperan penting untuk pengklasifikasian PKH dalam model LGBM adalah rata-rata pengeluaran perkapita (X7) dan jumlah anggota rumah tangga (X11) yang memiliki pengaruh sangat besar dalam

mengklasifikasikan rumah tangga penerima PKH di provinsi Jawa Barat, sehingga dapat dikatakan bahwa kedua peubah bebas tersebut menjadi faktor penting dalam menentukan penerima PKH.

4. Kesimpulan

Penerima Program Keluarga Harapan (PKH) di Provinsi Jawa Barat menghadapi tantangan ketidakseimbangan kelas, di mana hanya 13,8% rumah tangga yang menerima PKH. Penangangan data dengan kelas tidak seimbang menggunakan *oversampling* memberikan hasil bahwa metode GBM menunjukkan performa terbaik pada pembagian data latih dan uji 60:40 dan 70:30 dengan nilai *balanced accuracy* secara berturut-turut sebesar 64,42% dan 66,13%. Selanjutnya metode CatBoost memiliki nilai *balanced accuracy* tertinggi pada pembagian data latih dan uji 80:20, yaitu sebesar 66,56%. Sedangkan pada pembagian data latih dan uji 90:10, metode LGBM memiliki *balanced accuracy* tertinggi, yaitu sebesar 67,97%. LGBM menunjukkan performa terbaik ketika jumlah data latih berukuran besar, yaitu 90% dengan akurasi sebesar 67,97%, sedangkan untuk data latih kecil yaitu 60:40, LGBM memiliki performa yang kurang baik, dengan nilai *balanced accuracy* terendah dibandingkan metode *boosting* lainnya, yaitu sebesar 54,43%. Keunggulan LGBM ini disebabkan karena kemampuannya dalam mengelola data besar dan kompleks yang sesuai dengan karakteristik data sosial ekonomi rumah tangga penerima PKH. Analisis lebih lanjut menunjukkan bahwa faktor ekonomi dan jumlah anggota rumah tangga berperan besar dalam penentuan penerima PKH, sehingga program bantuan perlu mempertimbangkan aspek tersebut secara lebih menyeluruh. Namun, akurasi yang masih relatif rendah menunjukkan bahwa pola penerima PKH belum sepenuhnya teridentifikasi dengan baik, sehingga rekomendasi penelitian selanjutnya adalah menerapkan SHAP (*Shapley Additive Explanations*) untuk mengidentifikasi dan mengevaluasi kontribusi masing-masing peubah secara transparan, serta memperkaya data dengan faktor tambahan yang relevan agar hasil klasifikasi lebih akurat dan dapat mendukung kebijakan penyaluran bantuan sosial yang lebih sesuai sasaran. Selain itu, untuk menangani ketidakseimbangan kelas, diperlukan pendekatan yang lebih optimal dalam pemrosesan data guna memastikan model dapat mengenali pola secara lebih seimbang dan akurat.

Referensi

[1] T. C. Nokeri, *Data Science Solutions with Python*. 2022.

[2] L. M. Cendani and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," *J. Masy. Inform.*, vol. 13, no. 1, pp. 33–44, 2022, doi: 10.14710/jmasif.13.1.42912.

[3] N. I. Nella, N. Y. Setiawan, and D. E. Ratnawati, "Klasifikasi Penerima Bantuan Program Keluarga Harapan menggunakan Algoritme Decision Tree C4. 5 (Studi Kasus: Desa Mlirip Kabupaten Mojokerto)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 3, pp. 1332–1339, 2022.

[4] D. Haidar, B. Irawan, and A. Bahtiar, "Penerapan Deep Learning Model Random Forest Untuk Prediksi Penerima Bantuan Program Keluarga Harapan (Pkh)," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3564–3571, 2023, doi: 10.36040/jati.v7i6.8250.

[5] I. A. Sobari and R. A. Zuama, "Pendeketan Machine Learning dalam Memprediksi Keluarga Penerima Program PKH," *J. Tek. Komput. AMIK BSI*, vol. 9, no. 1, pp. 61–64, 2023, doi: 10.31294/jtk.v4i2.

[6] N. Alfiah, "Klasifikasi Penerima Bantuan Sosial Program Keluarga Harapan Menggunakan Metode Naive Bayes," *J. Teknol. Inf.*, vol. 16, no. 1, pp. 32–40, 2021, doi: 10.35842/jtir.v16i1.386.

[7] M. N. Isyam, D. Indrayana, and W. Apriandari, "Klasifikasi Penerima Bantuan Program Keluarga Harapan Menggunakan Support Vector Machine," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 5, pp. 10416–10421, 2024.

[8] W. H. Vidila, R. Kurniawan, and S. Anwar, "Optimization of Social Assistance Recipient Determination using Gradient Boosting Algorithm," vol. 4, no. 2, 2025.

ISSN (e): 2540-9123
ISSN (p): 2502-1982

[9] R. D. Mendrofa, M. H. Siallagan, D. P. Pakpahan, and J. Amalia, "Credit Risk Analysis With Extreme Gradient Boosting and Adaptive Boosting Algorithm," *J. Inf. Syst. Hosp. Technol.*, vol. 5, no. 1, pp. 1–7, 2023, doi: 10.37823/insight.v5i1.233.

[10] E. R. Putri and D. B. Arianto, "Perbandingan Performa Algoritma Metode Bagging dan Boosting pada Prediksi Konsentrasi PM10 di Jakarta Utara," *J. Nas. Teknol. dan Sist. Inf.*, vol. 10, no. 1, pp. 72–81, 2024, doi: 10.25077/teknosi.v10i1.2024.72-81.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[12] P. Liashchynskyi and P. Liashchynskyi, "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS," *arXiv Prepr.*, no. 2017, pp. 1–11, 2019, [Online]. Available: http://arxiv.org/abs/1912.06059.

[13] J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002, doi: 10.1016/S0167-9473(01)00065-2.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[15] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *31st Conf. Neural Inf. Process. Syst.*, pp. 1–9, 2017.

[16] M. J. Sai, P. Chettri, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, "An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, pp. 1–20, 2023, doi: 10.1007/s44196-023-00184-y.

[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *32nd Conf. Neural Inf. Process. Syst.*, pp. 1–11, 2018.

[18] A. Desiani *et al.*, "Penerapan Metode Support Vector Machine Dalam Klasifikasi Bunga Iris," *IJAI (Indoneisan J. Appl. Informatics)*, vol. 7, no. 1, pp. 12–18, 2022.

[19] A. Sharma and W. J. M. I. Verbeke, "Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081)," *Front. Big Data*, vol. 3, no. April, pp. 1–11, 2020, doi: 10.3389/fdata.2020.00015.

[20] C. Yang, E. A. Fridgeirsson, J. A. Kors, J. M. Reps, and P. R. Rijnbeek, "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-023-00857-7.

[21] O. P. Moerdyanto and I. K. D. Nuryana, "Prediksi Kelulusan Tepat Waktu Menggunakan Pendekatan Pohon Keputusan Algoritma Decision Tree," *J. Informatics Comput. Sci.*, vol. 05, no. 1, pp. 90–96, 2023, [Online]. Available: https://ejournal.unesa.ac.id/index.php/jinacs/article/view/55329.