p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1041-1048
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1177

Evaluasi Leakage-Aware dan Imbalance-Sensitive pada Algoritma Klasifikasi untuk Prediksi Keberhasilan Kampanye Bank Marketing

Purwatiningtyas[1*], Retnowati[2]

[1]Teknologi Informasi dan Industri, Teknik Informatika, Universitas Stikubank, Semarang, Indonesia
[2]Teknologi Informasi dan Industri, Magister Teknologi Informasi, Universitas Stikubank, Semarang, Indonesia
Email: [1*]purwati@edu.unisbank.ac.id , [2]retnowati@edu.unisbank.ac.id
(*Email Corresponding Author: purwati@edu.unisbank.ac.id)
Received: June 10,2026 | Revision: June 12,2026 | Accepted: June 24,2026

Abstrak

Prediksi respons nasabah merupakan masalah penting dalam pemasaran perbankan berbasis data, terutama karena kampanye langsung harus menyeimbangkan efisiensi biaya, ketepatan penargetan, dan kemampuan mengenali calon nasabah yang benar-benar berpotensi merespons. Penelitian ini memperkuat evaluasi model prediksi kampanye Bank Marketing dengan dua prinsip metodologis, yaitu leakage-aware evaluation dan imbalance-sensitive evaluation. Atribut duration dikeluarkan dari model karena hanya diketahui setelah panggilan selesai sehingga berpotensi menimbulkan target leakage. Empat algoritma klasifikasi, yaitu Logistic Regression, K-Nearest Neighbor, Decision Tree, dan Random Forest, dievaluasi pada dataset bank.csv UCI yang berisi 4.521 observasi dengan distribusi kelas tidak seimbang, yaitu 4.000 kelas no dan 521 kelas yes. Eksperimen menggunakan train-validation-test stratified split, preprocessing berbasis standardisasi dan one-hot encoding, tuning hyperparameter melalui stratified cross-validation, serta evaluasi dengan accuracy, precision, recall, F1-score, F2-score, balanced accuracy, Matthews correlation coefficient, ROC-AUC, PR-AUC, dan confusion matrix. Selain evaluasi baseline pada threshold 0,50, penelitian ini juga menerapkan threshold tuning berbasis validasi dengan kriteria F2-score untuk meningkatkan sensitivitas terhadap kelas positif. Hasil menunjukkan bahwa Random Forest memiliki performa paling seimbang. Pada threshold 0,50, Random Forest memperoleh ROC-AUC 0,7576, PR-AUC 0,3743, MCC 0,2830, dan recall 0,4231. Setelah threshold dituning menjadi 0,39, recall Random Forest meningkat menjadi 0,7019 dengan F2-score 0,4980 dan balanced accuracy 0,6987. Temuan ini menunjukkan bahwa pemilihan model untuk kampanye pemasaran tidak cukup hanya berdasarkan accuracy, tetapi perlu mempertimbangkan trade-off antara recall, precision, false negative, dan tujuan operasional kampanye.

Kata Kunci: *bank marketing* 1, *klasifikasi* 2, *class imbalance* 3, *target leakage* 4, *threshold tuning* 5, *Random Forest* 6

Abstract

Customer response prediction is a crucial issue in data-driven banking marketing, particularly as direct campaigns must balance cost efficiency, targeting accuracy, and the ability to identify potential customers who are truly likely to respond. This study strengthens the evaluation of Bank Marketing campaign prediction models with two methodological principles: leakage-aware evaluation and imbalance-sensitive evaluation. The duration attribute is excluded from the model because it is only known after the call is completed, potentially leading to target leakage. Four classification algorithms, namely Logistic Regression, K-Nearest Neighbor, Decision Tree, and Random Forest, are evaluated on the UCI bank.csv dataset containing 4,521 observations with an unbalanced class distribution, namely 4,000 no classes and 521 yes classes. The experiment uses a stratified train-validation-test split, standardization-based preprocessing and one-hot encoding, hyperparameter tuning through stratified cross-validation, and evaluations using accuracy, precision, recall, F1-score, F2-score, balanced accuracy, Matthews correlation coefficient, ROC-AUC, PR-AUC, and confusion matrix. In addition to the baseline evaluation at a threshold of 0.50, this study also applied validation-based threshold tuning with the F2-score criterion to increase sensitivity to the positive class. The results showed that Random Forest had the most balanced performance. At a threshold of 0.50, Random Forest obtained an ROC-AUC of 0.7576, a PR-AUC of 0.3743, an MCC of 0.2830, and a recall of 0.4231. After the threshold was tuned to 0.39, Random Forest's recall increased to 0.7019 with an F2-score of 0.4980 and a balanced accuracy of 0.6987. These findings indicate that model selection for marketing campaigns is not sufficient based solely on accuracy, but needs to consider the trade-offs between recall, precision, false negatives, and the campaign's operational objectives.

***Keywords*: :** *bank marketing* 1, *classificationi* 2, *class imbalance* 3, *target leakage* 4, *threshold tuning* 5, *Random Forest* 6

1. PENDAHULUAN

Industri perbankan modern semakin mengandalkan analitik data untuk meningkatkan efektivitas kampanye pemasaran. Dalam konteks direct marketing, keberhasilan kampanye tidak hanya ditentukan oleh banyaknya nasabah yang dihubungi, tetapi oleh kemampuan bank memilih calon nasabah yang paling mungkin merespons secara positif. Strategi penargetan yang terlalu luas dapat meningkatkan biaya operasional, memperbesar risiko gangguan kepada nasabah, dan menurunkan efisiensi tenaga pemasaran. Oleh karena itu, model klasifikasi menjadi salah satu pendekatan yang relevan untuk membantu bank memprioritaskan nasabah dengan probabilitas respons yang lebih tinggi.

Dataset Bank Marketing dari UCI Machine Learning Repository merupakan benchmark yang sering digunakan untuk menguji model prediksi respons kampanye perbankan. Dataset tersebut berasal dari kampanye telemarketing sebuah institusi perbankan di Portugal dan bertujuan memprediksi apakah klien akan berlangganan deposito berjangka.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1041-1048
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1177

Meskipun dataset ini populer, penggunaannya perlu memperhatikan aspek validitas operasional. Salah satu atribut, yaitu duration, sangat prediktif karena berisi durasi kontak terakhir, tetapi nilainya hanya tersedia setelah panggilan berakhir. Jika atribut ini digunakan untuk prediksi sebelum kontak dilakukan, model dapat mengalami target leakage dan menghasilkan estimasi performa yang tidak realistis.

Prediksi respons bank telemarketing telah menjadi salah satu contoh penting penerapan machine learning dalam pemasaran langsung. Moro et al. [1] menunjukkan bahwa data kampanye bank dapat digunakan untuk membangun model prediktif yang membantu meningkatkan efektivitas kampanye deposito. Dataset tersebut kemudian banyak digunakan sebagai benchmark karena memuat karakteristik demografis, finansial, riwayat kontak, dan keluaran kampanye yang dapat dianalisis sebagai masalah klasifikasi biner.

Studi mutakhir memperluas arah penelitian dari sekadar perbandingan algoritma menuju model yang lebih operasional dan interpretabel. Safarkhani dan Moro [8] menekankan pentingnya resampling dan feature selection untuk meningkatkan prediksi perilaku deposan bank. Xie et al. [9] mengembangkan pendekatan hybrid ensemble dan interpretability analysis, serta menegaskan bahwa dalam konteks bank telemarketing, mengurangi false negative sering kali lebih penting daripada sekadar menekan false positive. Guo et al. [10] juga menempatkan ensemble learning sebagai pendekatan penting untuk memperkirakan kompetensi kampanye telemarketing bank.

Selain persoalan leakage, dataset Bank Marketing juga memiliki distribusi kelas yang tidak seimbang. Pada subset bank.csv yang digunakan dalam penelitian ini, terdapat 4.000 observasi kelas no dan hanya 521 observasi kelas yes. Dalam kondisi seperti ini, accuracy dapat terlihat tinggi walaupun model gagal mengenali sebagian besar kelas positif. Oleh karena itu, evaluasi perlu mencakup metrik yang lebih sensitif terhadap kelas minoritas, seperti recall, F1-score, F2-score, balanced accuracy, Matthews correlation coefficient, ROC-AUC, dan PR-AUC.

Ketidakseimbangan kelas merupakan tantangan umum dalam prediksi respons pemasaran karena jumlah nasabah yang merespons positif biasanya jauh lebih sedikit dibandingkan yang tidak merespons. Dalam situasi ini, model dapat mencapai accuracy tinggi dengan memprediksi mayoritas observasi sebagai kelas negatif, tetapi gagal memberikan nilai praktis karena melewatkan calon nasabah yang sebenarnya potensial. Oleh sebab itu, accuracy harus dilengkapi dengan precision, recall, F1-score, F2-score, balanced accuracy, MCC, ROC-AUC, dan PR-AUC.

Nasir et al. [11] menunjukkan bahwa pada dataset target marketing bank yang tidak seimbang, performa model sangat dipengaruhi oleh teknik sampling dan pilihan metrik evaluasi. Peter et al. [12] juga menegaskan pentingnya evaluasi ensemble learning dengan metrik yang lebih luas, termasuk ROC-AUC dan MCC. Dalam penelitian ini, F2-score digunakan pada tahap threshold tuning karena F2 memberi bobot lebih besar pada recall daripada precision. Pendekatan ini relevan ketika kesalahan melewatkan nasabah potensial dipandang lebih merugikan daripada menghubungi sebagian nasabah yang ternyata tidak merespons.

Logistic Regression digunakan sebagai baseline karena sederhana, stabil, dan mudah diinterpretasikan pada masalah klasifikasi biner. KNN mewakili pendekatan instance-based learning yang bergantung pada struktur kedekatan antarobservasi, sehingga sensitif terhadap skala fitur dan dimensi hasil encoding. Decision Tree menawarkan interpretabilitas melalui aturan keputusan, tetapi rentan terhadap overfitting apabila tidak dikendalikan. Random Forest mengatasi kelemahan pohon tunggal dengan membangun banyak decision tree dan menggabungkan hasilnya melalui mekanisme ensemble sehingga lebih stabil pada data tabular.

Penelitian ini diarahkan untuk memperkuat studi komparatif algoritma klasik dengan pendekatan yang lebih relevan secara metodologis. Kontribusi utama penelitian ini adalah: (1) menerapkan evaluasi leakage-aware dengan menghapus atribut duration; (2) menggunakan pipeline preprocessing dan tuning yang konsisten pada seluruh model; (3) menambahkan evaluasi imbalance-sensitive melalui PR-AUC, balanced accuracy, MCC, dan F2-score; serta (4) meningkatkan recall melalui threshold tuning berbasis data validasi sehingga hasil yang dilaporkan lebih sesuai dengan kebutuhan operasional kampanye pemasaran.

2. METODOLOGI PENELITIAN

3.1 Desain penelitian dan dataset

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimen komparatif. Data yang digunakan adalah bank.csv, yaitu subset 10% dari dataset Bank Marketing UCI. Dataset terdiri atas 4.521 observasi dan 17 kolom sebelum penghapusan atribut duration. Variabel target adalah y dengan dua kelas, yaitu no dan yes. Distribusi kelas menunjukkan ketidakseimbangan yang cukup besar, dengan 4.000 observasi no atau 88,48% dan 521 observasi yes atau 11,52%.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1041-1048
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1177

Atribut prediktor mencakup variabel numerik dan kategorikal. Variabel numerik yang digunakan setelah penghapusan duration adalah age, balance, day, campaign, pdays, dan previous. Variabel kategorikal adalah job, marital, education, default, housing, loan, contact, month, dan poutcome. Dengan demikian, setelah kontrol leakage dilakukan, model menggunakan 15 prediktor utama.

Tabel 1. Ringkasan dataset dan pembagian data

Komponen	Nilai
Jumlah observasi	4.521
Jumlah kolom awal	17
Variabel target	y (no/yes)
Kelas no	4.000 (88,48%)
Kelas yes	521 (11,52%)
Atribut leakage yang dihapus	duration
Jumlah prediktor setelah leakage control	15
Data latih	2.712 observasi
Data validasi	904 observasi
Data uji	905 observasi

3.2 Preprocessing dan kontrol target leakage

Preprocessing dilakukan melalui pipeline yang sama untuk menjaga keterbandingan antarmodel. Pertama, atribut duration dihapus sebelum pemisahan data karena atribut ini hanya diketahui setelah panggilan telepon selesai. Penghapusan dilakukan pada awal pipeline agar informasi pasca-kejadian tidak masuk ke model pada tahap pelatihan maupun evaluasi.

Kedua, fitur kategorikal dikonversi menjadi fitur numerik menggunakan one-hot encoding dengan opsi handle_unknown=ignore agar kategori yang tidak muncul pada data latih tidak menyebabkan error pada data validasi atau data uji. Ketiga, fitur numerik distandardisasi menggunakan StandardScaler khusus untuk Logistic Regression dan KNN karena kedua algoritma sensitif terhadap skala fitur. Untuk Decision Tree dan Random Forest, fitur numerik tidak distandardisasi karena model berbasis pohon tidak bergantung pada jarak Euclidean dan relatif tidak sensitif terhadap skala.

3.3 Skema pembagian data dan tuning model

Dataset dibagi menjadi data latih-validasi dan data uji dengan rasio 80:20 menggunakan stratified split agar proporsi kelas no dan yes tetap terjaga. Selanjutnya, data latih-validasi dibagi lagi menjadi data latih dan data validasi. Data latih digunakan untuk melatih dan memilih hyperparameter melalui StratifiedKFold 5-fold cross-validation, data validasi digunakan untuk memilih threshold yang memaksimalkan F2-score, dan data uji digunakan hanya untuk evaluasi akhir.

Random state ditetapkan pada nilai 42 untuk menjaga reproduksibilitas eksperimen. Pada tahap tuning, ROC-AUC digunakan sebagai scoring utama karena metrik ini mengevaluasi kemampuan ranking model terhadap kelas positif dan negatif tanpa bergantung pada satu threshold tertentu. Setelah model terbaik diperoleh, probability score kelas positif dihitung dengan predict_proba. Probability score tersebut digunakan untuk membangun ROC curve, Precision-Recall curve, dan evaluasi threshold.

Tabel 2. Ruang pencarian hyperparameter

Model	Hyperparameter yang diuji
Logistic Regression	C = 0,1; 1; 10; class_weight = None, balanced; solver = liblinear; max_iter = 2000
KNN	n_neighbors = 3; 5; 11; 15; weights = uniform, distance
Decision Tree	criterion = gini, entropy; max_depth = 3; 5; 10; None; min_samples_leaf = 1; 5; 10; class_weight = None, balanced
Random Forest	n_estimators = 150; max_depth = None, 10; min_samples_leaf = 1; 5; max_features = sqrt; class_weight = None, balanced_subsample

3.4 Metrik evaluasi

Evaluasi model dilakukan dengan dua skenario. Skenario pertama adalah baseline evaluation pada threshold 0,50. Skenario kedua adalah recall-oriented evaluation dengan threshold yang dipilih pada data validasi berdasarkan F2-score.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1041-1048
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1177

Evaluasi kedua tidak menggantikan evaluasi baseline, tetapi digunakan untuk mensimulasikan kebutuhan operasional kampanye ketika bank lebih mengutamakan peningkatan deteksi calon nasabah potensial.

Metrik yang digunakan meliputi accuracy, precision, recall, F1-score, F2-score, balanced accuracy, MCC, ROC-AUC, PR-AUC, dan confusion matrix. Accuracy mengukur proporsi prediksi benar secara keseluruhan. Precision menunjukkan proporsi prediksi positif yang benar. Recall menunjukkan proporsi kelas positif yang berhasil ditemukan. F1-score menyeimbangkan precision dan recall, sedangkan F2-score memberi bobot lebih besar pada recall. Balanced accuracy memperhitungkan rata-rata sensitivitas pada kedua kelas. MCC digunakan karena lebih informatif pada distribusi kelas tidak seimbang. ROC-AUC dan PR-AUC dihitung dari probability score model, bukan dari label prediksi biner.

Tabel 3. Rumus metrik evaluasi utama

Metrik	Rumus	Makna
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Proporsi prediksi yang benar
Precision	TP / (TP + FP)	Ketepatan prediksi kelas yes
Recall	TP / (TP + FN)	Kemampuan menemukan kelas yes
F1-score	2 x Precision x Recall / (Precision + Recall)	Keseimbangan precision dan recall
F2-score	5 x Precision x Recall / (4 x Precision + Recall)	Recall diberi bobot lebih besar
Balanced accuracy	(Sensitivity + Specificity) / 2	Akurasi yang mempertimbangkan kedua kelas
MCC	(TP x TN - FP x FN) / sqrt((TP+FP)(TP+FN)(TN+FP)(TN+FN))	Korelasi prediksi dan label aktual

3. HASIL DAN PEMBAHASAN

3.1 Hasil baseline pada threshold 0,50

Tabel 4. Hasil baseline pada threshold 0,50

Model	Acc.	Prec.	Recall	F1	F2	Bal. Acc.	MCC	ROC-AUC	PR-AUC
Logistic Regression	0,7138	0,2261	0,6154	0,3307	0,4578	0,6710	0,2353	0,7265	0,3151
KNN	0,8906	0,8571	0,0577	0,1081	0,0709	0,5282	0,2055	0,6554	0,2534
Decision Tree	0,8133	0,2759	0,3846	0,3213	0,3565	0,6268	0,2204	0,6969	0,2561
Random Forest	0,8365	0,3333	0,4231	0,3729	0,4015	0,6566	0,2830	0,7576	0,3743

Tabel 5. Confusion matrix baseline pada data uji

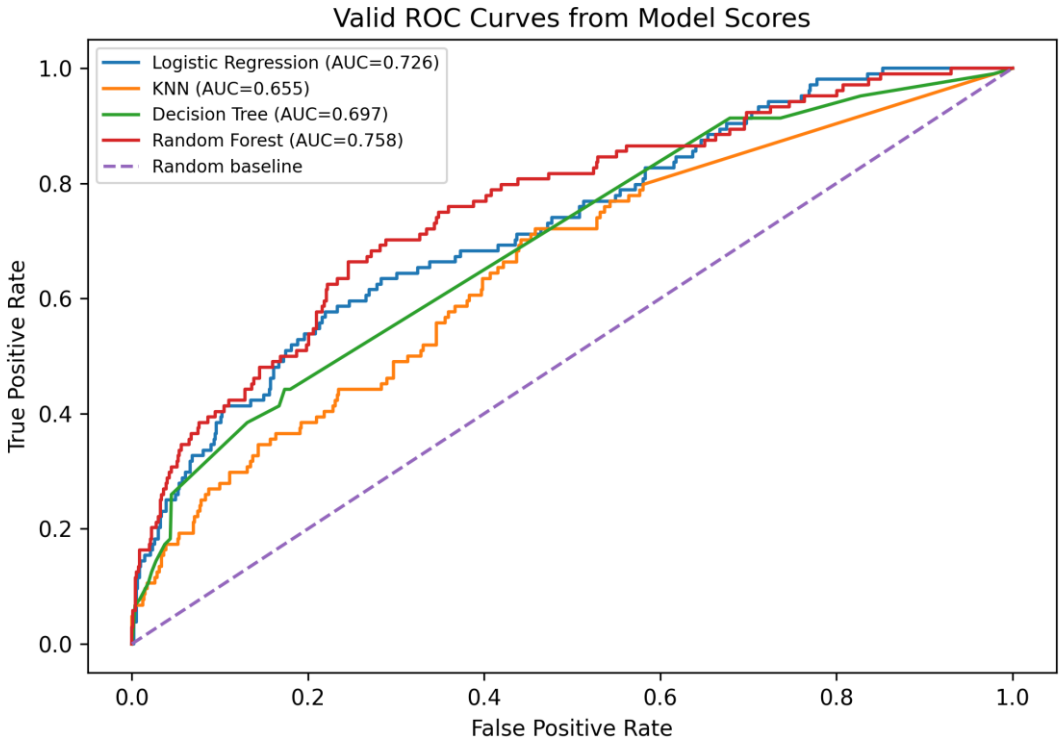
Model	TN	FP	FN	TP
Logistic Regression	582	219	40	64
KNN	800	1	98	6
Decision Tree	696	105	64	40
Random Forest	713	88	60	44

Pada threshold 0,50, KNN menghasilkan accuracy tertinggi sebesar 0,8906, tetapi recall-nya hanya 0,0577. Artinya, KNN hanya mampu mengenali 6 dari 104 nasabah kelas yes pada data uji, sehingga model ini tidak layak dijadikan pilihan utama apabila tujuan kampanye adalah menemukan calon nasabah potensial. Temuan ini menunjukkan keterbatasan accuracy sebagai indikator utama pada data tidak seimbang.

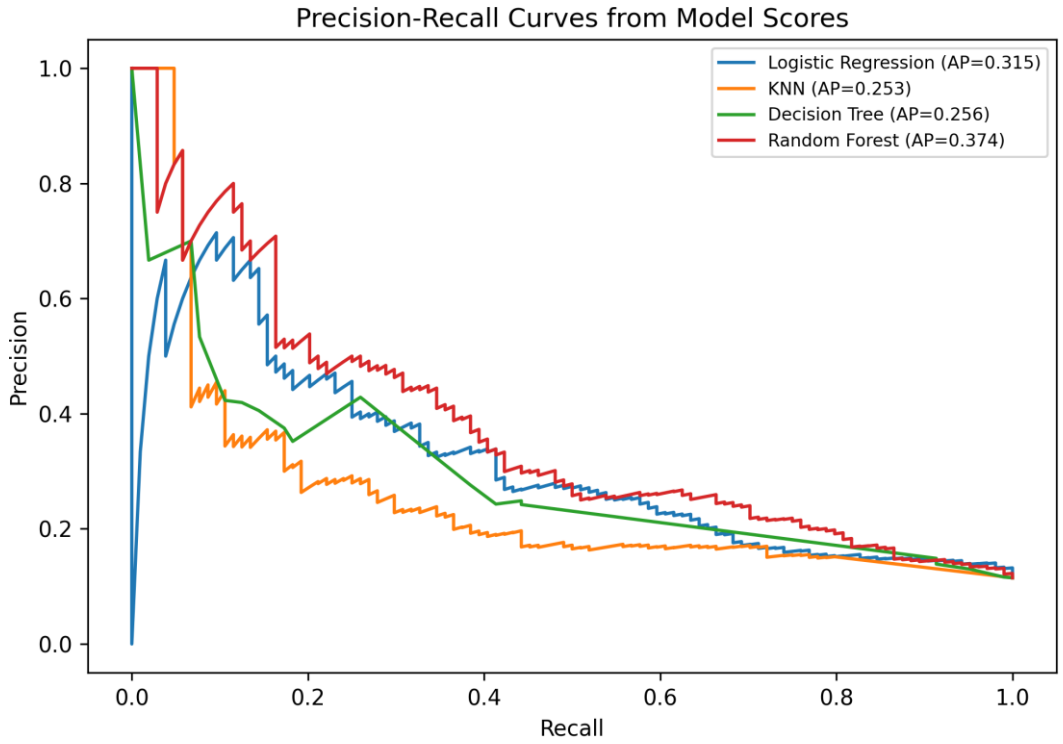
Random Forest menunjukkan profil baseline yang lebih kuat dibandingkan model lain. Model ini memperoleh ROC-AUC tertinggi sebesar 0,7576, PR-AUC tertinggi sebesar 0,3743, MCC tertinggi sebesar 0,2830, dan F1-score tertinggi sebesar 0,3729. Walaupun accuracy Random Forest sedikit lebih rendah daripada KNN, kemampuannya dalam

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1041-1048
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1177

mengenali kelas positif jauh lebih baik. Logistic Regression juga cukup kompetitif dengan recall 0,6154, tetapi precision-nya lebih rendah, yaitu 0,2261.



Gambar 1. Kurva ROC valid yang dihitung dari probability score model.



Gambar 2. Kurva Precision-Recall untuk evaluasi data tidak seimbang.

3.2 Hasil threshold tuning untuk peningkatan recall

Tabel 6. Hasil recall-oriented threshold tuning pada data uji

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1041-1048
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1177

Model	Threshold	Acc.	Prec.	Recall	F1	F2	Bal. Acc.	MCC
Logistic Regression	0,41	0,5536	0,1667	0,7212	0,2708	0,4330	0,6265	0,1614
KNN	0,08	0,5039	0,1515	0,7212	0,2504	0,4116	0,5984	0,1261
Decision Tree	0,37	0,3381	0,1387	0,9135	0,2408	0,4314	0,5884	0,1315
Random Forest	0,39	0,6961	0,2303	0,7019	0,3468	0,4980	0,6987	0,2656

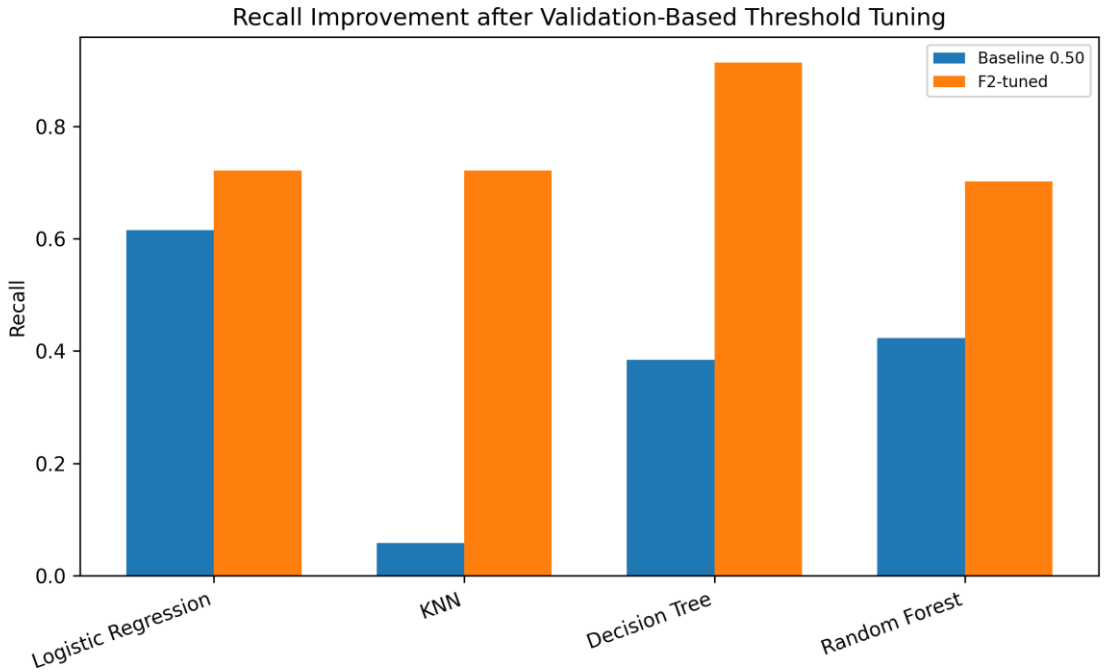
Tabel 7. Confusion matrix setelah threshold tuning

Model	TN	FP	FN	TP
Logistic Regression	426	375	29	75
KNN	381	420	29	75
Decision Tree	211	590	9	95
Random Forest	557	244	31	73

Threshold tuning berhasil meningkatkan recall seluruh model. Logistic Regression naik dari 0,6154 menjadi 0,7212; KNN naik dari 0,0577 menjadi 0,7212; Decision Tree naik dari 0,3846 menjadi 0,9135; dan Random Forest naik dari 0,4231 menjadi 0,7019. Peningkatan ini terjadi karena ambang keputusan diturunkan sehingga model lebih banyak mengklasifikasikan observasi sebagai kelas yes.

Akan tetapi, peningkatan recall selalu memiliki konsekuensi terhadap precision dan false positive. Decision Tree memperoleh recall tertinggi sebesar 0,9135, tetapi precision-nya hanya 0,1387 dan false positive mencapai 590. Kondisi ini menunjukkan bahwa model terlalu agresif memprediksi kelas yes. Random Forest lebih seimbang karena pada threshold 0,39 mampu menemukan 73 dari 104 nasabah kelas yes, menurunkan false negative menjadi 31, dan tetap mempertahankan MCC tertinggi pada skenario tuning, yaitu 0,2656.

Berdasarkan hasil baseline dan threshold tuning, Random Forest dipilih sebagai model yang paling direkomendasikan. Alasan pemilihannya bukan karena accuracy tertinggi, melainkan karena kombinasi ROC-AUC, PR-AUC, F1-score, F2-score, MCC, dan recall yang paling stabil. Dalam konteks kampanye pemasaran, model ini memberi kompromi yang lebih realistis antara menemukan calon responden potensial dan membatasi jumlah nasabah yang salah ditargetkan.



Gambar 3. Peningkatan recall setelah threshold tuning berbasis F2-score.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1041-1048
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1177

3.3 Implikasi ilmiah dan praktis

Secara ilmiah, hasil penelitian ini menunjukkan bahwa evaluasi classifier pada dataset Bank Marketing perlu dibangun sebagai eksperimen yang bebas leakage dan sensitif terhadap kelas minoritas. Penggunaan kurva ROC yang dihitung dari probability score menjadikan visualisasi lebih valid dibandingkan kurva skematik berbasis nilai AUC saja. Penambahan PR-AUC juga penting karena precision-recall curve lebih informatif ketika distribusi kelas tidak seimbang.

Secara praktis, hasil penelitian ini menyarankan bahwa bank tidak sebaiknya memilih model hanya berdasarkan accuracy. Jika tujuan kampanye adalah targeting yang sangat selektif, model dengan precision lebih tinggi dapat diprioritaskan. Namun, jika bank ingin mengurangi risiko melewatkan nasabah potensial, threshold perlu diturunkan dan model harus dievaluasi dengan recall, F2-score, dan jumlah false negative. Pada eksperimen ini, Random Forest dengan threshold 0,39 merupakan pilihan yang lebih sesuai untuk strategi kampanye yang berorientasi pada peningkatan deteksi kelas yes.

4. KESIMPULAN

Penelitian ini membandingkan Logistic Regression, KNN, Decision Tree, dan Random Forest untuk memprediksi keberhasilan kampanye pemasaran pada dataset Bank Marketing. Artikel ini memperkuat rancangan eksperimen dengan menghapus atribut duration untuk menghindari target leakage, menerapkan preprocessing yang konsisten, menggunakan stratified train-validation-test split, membangun ROC curve dan PR curve dari probability score, serta menambahkan evaluasi yang lebih sesuai untuk data tidak seimbang. Hasil baseline menunjukkan bahwa Random Forest memiliki performa paling stabil dengan ROC-AUC 0,7576, PR-AUC 0,3743, F1-score 0,3729, dan MCC 0,2830. Setelah threshold tuning berbasis validasi, Random Forest meningkatkan recall dari 0,4231 menjadi 0,7019, dengan F2-score 0,4980 dan balanced accuracy 0,6987. Decision Tree memang menghasilkan recall tertinggi setelah tuning, tetapi jumlah false positive yang sangat besar membuatnya kurang ideal untuk penerapan kampanye yang membutuhkan keseimbangan biaya dan manfaat. Dengan demikian, kontribusi utama penelitian ini adalah penyajian evaluasi yang lebih valid dan operasional untuk prediksi kampanye bank marketing. Artikel ini menegaskan bahwa model terbaik tidak dapat ditentukan hanya dari accuracy, tetapi harus mempertimbangkan leakage control, ketidakseimbangan kelas, probability-based evaluation, threshold selection, dan prioritas bisnis. Penelitian selanjutnya dapat memperluas eksperimen pada bank-full.csv, menambahkan metode boosting seperti XGBoost dan LightGBM, menerapkan SMOTENC atau Borderline-SMOTE secara terkontrol, serta melakukan interpretability analysis menggunakan feature importance atau SHAP.

Daftar Pustaka

[1] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decision Support Systems, vol. 62, pp. 22-31, 2014, doi: 10.1016/j.dss.2014.03.001.

[2] UCI Machine Learning Repository, "Bank Marketing Data Set," 2014. [Online]. Available: https://archive.ics.uci.edu/dataset/222/bank+marketing

[3] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, An Introduction to Statistical Learning with Applications in R, 2nd ed. New York: Springer, 2021.

[4] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967, doi: 10.1109/TIT.1967.1053964.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. New York: Routledge, 1984.

[6] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[7] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing and Management, vol. 45, no. 4, pp. 427-437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[8] F. Safarkhani and S. Moro, "Improving the Accuracy of Predicting Bank Depositor's Behavior Using a Decision Tree," Applied Sciences, vol. 11, no. 19, p. 9016, 2021, doi: 10.3390/app11199016.

[9] C. Xie, J.-L. Zhang, Y. Zhu, B. Xiong, and G.-J. Wang, "How to improve the success of bank telemarketing? Prediction and interpretability analysis based on machine learning," Computers & Industrial Engineering, vol. 175, p. 108874, 2023, doi: 10.1016/j.cie.2022.108874.

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 1041-1048
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1177

[10] W. Guo, Y. Yao, L. Liu, and T. Shen, "A novel ensemble approach for estimating the competency of bank telemarketing," Scientific Reports, vol. 13, p. 20819, 2023, doi: 10.1038/s41598-023-47177-7.

[11] F. Nasir, A. A. Ahmed, M. S. Kiraz, I. Yevseyeva, and M. Saif, "Data-Driven Decision-Making for Bank Target Marketing Using Supervised Learning Classifiers on Imbalanced Big Data," Computers, Materials & Continua, vol. 81, no. 1, pp. 1703-1728, 2024, doi: 10.32604/cmc.2024.055192.

[12] M. Peter, H. Mofi, S. Likoko, J. Sabas, R. Mbura, and N. Mduma, "Predicting customer subscription in bank telemarketing campaigns using ensemble learning models," Machine Learning with Applications, vol. 19, p. 100618, 2025, doi: 10.1016/j.mlwa.2025.100618.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[14] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," Journal of Machine Learning Research, vol. 18, no. 17, pp. 1-5, 2017.

[15] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[16] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[17] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 233-240, doi: 10.1145/1143844.1143874.

[18] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, 2020, doi: 10.1186/s12864-019-6413-7.