Volume 6 Nomor 2, Oktober 2023
e-ISSN : 2621-234X

Analisis Perilaku Pelanggan menggunakan Metode Ensemble Logistic Regression

***Jeffry[1], Syahrul Usman[2], Firman Aziz[3]**

Address : Universitas Pancasakti Makassar, Fakultas Matematika dan Ilmu Pengetahuan Alam, Program Studi Ilmu Komputer, Indonesia[1,2,3]

Email : jeffry@unpacti.ac.id[1], syahrul.usman@unpacti.ac.id[2], firman.aziz@unpacti.ac.id[3]

Abstrak

Analisis perilaku pelanggan menjadi hal yang sangat penting bagi perusahaan khususnya Usaha Mikro, Kecil dan Menengah (UMKM) untuk mengambil keputusan strategis. Melalui analisis perilaku pelanggan, dapat diketahui pola-pola perilaku pelanggan dalam membeli suatu produk atau jasa. Sehingga dapat memberikan wawasan yang lebih mengenai preferensi dan strategi yang tepat untuk mempertahankan atau meningkatkan loyalitas pelanggan. Penelitian ini bertujuan untuk menganalisis perilaku pelanggan menggunakan metode ensemble logistic regression. Data dikumpulkan dari basis pelanggan perusahaan selama 5 tahun terakhir. Perilaku pelanggan ditandai oleh jenis kelamin, usia, dan preferensi transmisi kendaraan yang dipilih. Metode ensemble logistic regression diimplementasikan untuk meningkatkan akurasi model. Hasil penelitian menunjukkan bahwa Teknik ensemble mampu meningkatkan akurasi dari metode Logistic Regression. Akurasi model meningkat secara signifikan dengan teknik boosting mencapai akurasi sebesar 76%, sementara model dengan teknik bagging mencapai akurasi sebesar 75%.

Kata Kunci – *Ensemble, Logistic Regression, Bagging, Boosting, Perilaku Pelanggan*

Abstract

Customer behavior analysis is crucial for companies, especially Small and Medium Enterprises (SMEs), to make strategic decisions. Through customer behavior analysis, patterns of customer behavior in purchasing a product or service can be identified. This provides insights into preferences and the right strategies to maintain or increase customer loyalty. This research aims to analyze customer behavior using the ensemble logistic regression method. Data was collected from the company's customer database over the last 5 years. Customer behavior is characterized by gender, age, and the preferred vehicle transmission type. The ensemble logistic regression method was implemented to improve model accuracy. The results show that the ensemble technique can enhance the accuracy of the Logistic Regression method. Model accuracy significantly increased with boosting, achieving an accuracy of 76%, while the model with bagging achieved an accuracy of 75%.

Keywords *– Ensemble, Logistic Regression, Bagging, Boosting, Customer Behavior*

1. Latar Belakang

Prediksi perilaku pelanggan merupakan suatu teknik yang digunakan untuk menganalisis dan memahami bagaimana perilaku pelanggan dalam berinteraksi dengan produk atau layanan suatu perusahaan. Teknik ini dapat membantu perusahaan dalam merencanakan strategi pemasaran yang tepat untuk meningkatkan kepuasan pelanggan dan meningkatkan profitabilitas perusahaan.

Dalam era persaingan bisnis yang semakin ketat, pemahaman mengenai perilaku pelanggan menjadi hal yang sangat penting bagi perusahaan khususnya UMKM untuk mengembangkan strategi pemasaran yang tepat dan meningkatkan kepuasan pelanggan. Dalam hal ini, analisis perilaku pelanggan menggunakan teknik-teknik machine learning dapat membantu pelaku usaha dalam memprediksi kebutuhan pelanggan dan menyesuaikan strategi pemasaran yang efektif.

Jika perusahaan tidak mengetahui analisis perilaku pelanggan, perusahaan akan kesulitan dalam menentukan strategi pemasaran dan hilangnya pelanggan karena adanya kompetitor atau pesaing yang

memiliki penawaran lebih yang dikenal sebagai churn pelanggan[1]. Hal ini dapat mengakibatkan kesulitan dalam menjual produk atau jasa, mengalami penurunan penjualan, dan bahkan kebangkrutan perusahaan. Dalam jangka panjang, perusahaan yang tidak melakukan analisis perilaku pelanggan dapat kehilangan daya saingnya di pasar dan sulit untuk bertahan dalam persaingan yang semakin ketat. Oleh karena itu, penting bagi perusahaan, terutama UMKM, untuk memahami perilaku pelanggan dan menerapkan teknik analisis yang tepat untuk meningkatkan penjualan dan daya saingnya di pasar.

Dalam melakukan prediksi perilaku pelanggan, salah satu teknik yang sering digunakan adalah machine learning. Machine learning merupakan cabang dari ilmu kecerdasan buatan yang memungkinkan komputer untuk belajar dari data yang diberikan dan melakukan prediksi berdasarkan data yang telah dipelajari sebelumnya [2].

Beberapa metode machine learning yang sering digunakan untuk prediksi perilaku pelanggan antara lain decision tree, random forest, logistic regression, neural network, dan KNN [3]. Masing-masing metode memiliki tingkat akurasi yang berbeda-beda tergantung pada karakteristik data yang digunakan.

Telah banyak penelitian terkait yang telah dilakukan dan bisa dijadikan sebagai state-of-the-art pada penelitian ini. Diantaranya yaitu penelitian yang dilakukan oleh Choudhury, dkk [4] dengan mengusulkan pendekatan machine learning untuk mengidentifikasi pelanggan potensial untuk superstore ritel. Hasil yang diperoleh menunjukkan bahwa metode Logistic Regression memiliki akurasi tertinggi dibandingkan dengan metode SVM, Decision Tree, Random Forest dan Perceptron dengan menggunakan fitur native yaitu sebesar 56.78%.

Suryadi, dkk. [5] mengusulkan metodologi berbasis data untuk secara otomatis mengidentifikasi konteks penggunaan produk dari ulasan pelanggan online. Penelitian ini menggunakan klasifikasi Naive Bayes dengan membagi tiga kategori penggunaan produk. Kategori "aktivitas sehari-hari" memiliki akurasi tertinggi sebesar 94.8%, diikuti oleh kategori "bisnis" dengan akurasi 86.8%, dan kategori "hiburan" dengan akurasi 85.2%.

Al-Mashraie, dkk [6] membandingkan kinerja berbagai model prediksi churn berdasarkan data nyata yang diperoleh dari perusahaan mitra. Analisis komparatif metode yaitu SVM sebesar 91.58%, decision tree sebesar 86.88%, random forest sebesar 91.46%, dan logistic regression sebesar 91.20%.

Penelitian yang disajikan oleh Mohbey [7] dengan melakukan perbandingan metode Random Forest, Naive Bayes, Logistic Regression, SVM, dan Decision Tree untuk prediksi pengurangan karyawan. Hasil penelitian menunjukkan bawah Algoritma Logistic Regression memiliki akurasi yang lebih baik dengan akurasi 86%.

Bahrami, dkk [8] mengusulkan model penilaian perilaku baru yang digunakan sebagai variabel masukan untuk model prediktif. Hasil penelitian menunjukkan bahwa Logistic Regression memiliki akurasi hingga 97%.

Penelitian yang dilakukan oleh Wu, dkk [9] yaitu menganalisis perilaku pelanggan dengan membandingkan metode Random Guess, Logistic Regression, Random Forest, LSTM dan DNN. Hasil penelitian menunjukkan bahwa kombinasi antara Algoritma LSTM+DNN memiliki akurasi yang lebih baik dengan 72%. Metode-metode machine learning juga digunakan pada penelitian Mun, dkk [10] dengan melakukan analisis terhadap perilaku penggunaan AC, dimana algoritma yang digunakan yaitu Logistic Regression dengan akurasi 71%, Random Forest 77% dan SVM 74%.

Berdasarkan state-of-the-art di atas, dapat disimpulkan bahwa Logistic regression adalah salah satu algoritma machine learning yang sering digunakan dalam analisis perilaku pelanggan. Hal ini disebabkan karena logistic regression dapat memprediksi apakah suatu pelanggan akan melakukan tindakan tertentu atau tidak berdasarkan beberapa faktor yang dapat diukur. Namun pada beberapa penelitian di atas belum menerapkan teknik ensemble pada algoritma yang digunakan. Ensemble logistic regression adalah metode yang menggabungkan beberapa model logistic regression untuk meningkatkan kinerja prediksi. Sehingga penelitian ini akan melakukan analisis pelanggan menggunakan Algoritma Logistic Regression menggunakan teknik ensemble untuk meningkatkan akurasi prediksi.

2. Kajian Teori

2.1 Logistic Regression

Regresi Logistik adalah metode statistik yang digunakan untuk menganalisis dataset dengan variabel target yang bersifat kategoris[11]. Metode ini sering digunakan dalam kasus klasifikasi biner, di mana kita ingin memprediksi hasil yang termasuk dalam satu dari dua kategori yang mungkin [12], seperti "ya" atau "tidak", "spam" atau "bukan spam", dan sejenisnya.

Tujuan utama dari Regresi Logistik adalah untuk membangun model yang dapat memahami hubungan antara variabel input (biasanya disebut fitur atau prediktor) dan kemungkinan hasil yang termasuk dalam kategori tertentu. Model ini menggunakan fungsi logistik (atau sigmoid) untuk mengubah nilai-nilai prediktor menjadi probabilitas yang berkisar antara 0 dan 1.

Rumus matematis yang mendasari Regresi Logistik adalah sebagai berikut [13]:

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\cdots+b_pX_p}} \quad (1)$$

- $P(Y = 1)$ adalah probabbilitas bahwa hasil Y masuk dalam kategori 1
- e adalah dasar logaritma natural
- $b_0, b_1, b_2, \dots, b_p$ adalah koefisien yang ditemukan melalui proses pelatihan model
- $X_1, X_2, \dots, X_p$ adalah nilai-nilai predictor.

Proses pelatihan model melibatkan menemukan koefisien yang optimal untuk meminimalkan perbedaan antara probabilitas yang diprediksi oleh model dengan hasil aktual dalam data pelatihan. Model yang telah dilatih dapat digunakan untuk memprediksi probabilitas hasil tertentu berdasarkan nilai-nilai prediktor pada data baru.

2.2 Boosting

Teknik Boosting adalah salah satu metode dalam machine learning yang bertujuan untuk meningkatkan kinerja model prediksi dengan cara menggabungkan beberapa model lemah atau classifier menjadi satu model yang lebih kuat[14]. Ide dasar di balik teknik ini adalah bahwa dengan menggabungkan model-model lemah yang berfokus pada titik data yang sulit diklasifikasikan, kita dapat mengurangi kesalahan prediksi dan meningkatkan akurasi keseluruhan.

Proses Boosting dimulai dengan membangun model lemah pertama menggunakan data pelatihan. Kemudian, kesalahan dari model pertama dihitung, dan data yang salah diklasifikasikan oleh model tersebut diberi bobot lebih tinggi. Model lemah kedua kemudian dibangun dengan fokus pada data yang memiliki bobot tinggi, dengan tujuan memperbaiki kesalahan yang dilakukan oleh model pertama. Hasil prediksi dari kedua model tersebut digabungkan dengan memberikan bobot pada masing-masing prediksi. Iterasi ini dilakukan beberapa kali, dengan setiap iterasi membangun model lemah baru dan menggabungkannya dengan model-model sebelumnya. Hasil akhir adalah gabungan dari semua model lemah, yang menghasilkan model ensemble yang lebih kuat[15].

2.3 Bagging

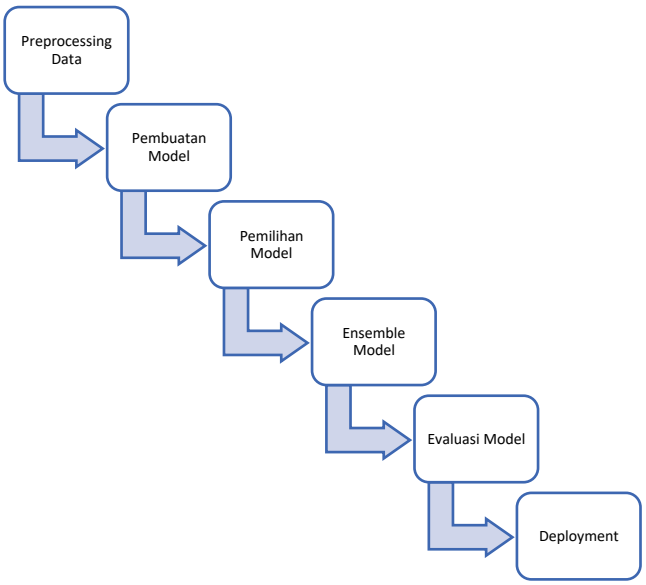
Teknik Bagging (Bootstrap Aggregating) adalah salah satu metode ensemble dalam machine learning yang bertujuan untuk meningkatkan kinerja model prediksi dengan cara menggabungkan beberapa model base (biasanya model simpel) menjadi satu model ensemble yang lebih kuat. Bagging bekerja dengan cara mengambil sampel acak (bootstrap) dari data pelatihan dengan penggantian, sehingga menghasilkan sejumlah dataset pelatihan yang berbeda-beda. Setiap dataset pelatihan ini digunakan untuk melatih model base yang sama[16].

Ide utama di balik Bagging adalah bahwa dengan melatih model pada dataset yang bervariasi, kita dapat mengurangi varians dari model, yang dapat menghasilkan prediksi yang lebih stabil dan akurat. Setelah semua model base terlatih, hasil prediksi dari setiap model diambil, dan hasilnya biasanya diambil berdasarkan mayoritas (mode) dalam klasifikasi atau rata-rata dalam regresi[17].

Jadi, secara singkat, teknik Bagging adalah pendekatan ensemble yang menggabungkan hasil dari beberapa model base yang sama untuk menghasilkan prediksi yang lebih stabil dan kuat. Ia sering digunakan dalam situasi di mana kita ingin mengurangi varians model dan meningkatkan akurasi.

3. Metode

Penelitian ini menggunakan teknik machine learning dengan Metode Ensemble Logistic Regression yang tergambar pada diagram berikut:



Gambar 1. Diagram Teknik Ensemble Logistic Regression

3.1 Preprocessing Data

Tahap preprocessing data merupakan langkah awal dalam analisis perilaku pelanggan menggunakan teknik ensemble logistic regression. Pada tahap ini, data penjualan mobil akan dipersiapkan dengan melakukan pembersihan dari nilai kosong (missing value), mengidentifikasi dan menangani outliers yang dapat mempengaruhi hasil analisis, serta menghapus data yang tidak relevan atau tidak diperlukan [18]. Proses ini bertujuan untuk memastikan bahwa data yang digunakan dalam pemodelan memiliki kualitas yang baik dan siap untuk digunakan dalam pembuatan model dengan Teknik ensemble.

3.2 Pembuatan Model

Setelah data bersih dan siap, langkah selanjutnya adalah pembuatan model. Pada tahap ini, beberapa model logistic regression akan dibangun dengan variasi parameter seperti learning rate, jumlah iterasi, dan regularisasi. Model-model ini akan menjadi model dasar yang akan digunakan dalam teknik ensemble logistic regression. Pembuatan model ini bertujuan untuk

menghasilkan beragam model yang memiliki karakteristik berbeda-beda, sehingga dapat diambil manfaat dari variasi model *dalam* proses ensemble.

3.3 Pemilihan Model

Memilih model-model yang memiliki performa terbaik berdasarkan evaluasi model seperti akurasi, precision, recall, dan F1-score.

3.4 Ensemble Model

Menggabungkan model-model terbaik menjadi satu model yang lebih kuat dengan teknik boosting dan bagging.

3.5 Evaluasi Model

Melakukan evaluasi terhadap model ensemble logistic regression yang telah dibuat. Evaluasi dilakukan dengan menggunakan data uji dan menghitung akurasi yaitu dengan mengukur sejauh mana model mengklasifikasikan secara benar seluruh instance atau sampel dalam dataset, precision yaitu mengukur sejauh mana prediksi positif yang dibuat oleh model adalah benar, recall (Sensitivity atau True Positive Rate) yaitu mengukur sejauh mana model dapat mengidentifikasi secara benar semua instance positif dalam dataset., dan F1-score adalah adalah metrik gabungan yang mengkombinasikan precision dan recall [19].

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Jumlah\ Total\ Sampel}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1 - Score = \frac{2.precision.recall}{precision + recall}$$

Selain itu metrik evaluasi yang yang digunakan adalah Precision-Recall Curve (PRC) dan Receiver Operating Characteristic Curve (ROC). Precision-Recall Curve (PRC) adalah alat visualisasi yang digunakan untuk mengukur kinerja model klasifikasi, terutama pada kasus di mana kelas minoritas dalam dataset sangat tidak seimbang. PRC menggambarkan hubungan antara precision dan recall pada berbagai ambang batas (threshold) yang berbeda dalam model klasifikasi[20]. PRC memberikan wawasan yang lebih baik tentang seberapa baik model dapat mengidentifikasi instance dari kelas minoritas dengan tingkat precision yang dapat diandalkan pada berbagai tingkat recall. Area di bawah kurva PRC (AUC-PRC) adalah ukuran kinerja model, di mana semakin besar nilainya, semakin baik model tersebut dalam membedakan antara kelas positif dan negative [21]. Sedangkan Karakteristik Penerima Operasi (Receiver Operating Characteristic Curve atau ROC Curve) adalah sebuah grafik yang digunakan untuk mengevaluasi kinerja model klasifikasi biner. ROC Curve mengilustrasikan hubungan antara Tingkat True Positive (True Positive Rate atau TPR) dan Tingkat False Positive (False Positive Rate atau FPR) pada berbagai ambang batas (threshold) yang berbeda [22]. Area Under the Curve (AUC-ROC) adalah ukuran yang menggambarkan kinerja keseluruhan model. Ini mengukur area di bawah kurva ROC, dengan nilai maksimum 1 yang menunjukkan kinerja sempurna dan nilai 0,5 yang menunjukkan kinerja yang sama dengan pemilihan acak[23], [24].

3.6 Deployment

Tujuan dari tahap ini adalah untuk mengintegrasikan model ke dalam proses bisnis atau sistem yang sesungguhnya agar dapat digunakan untuk membuat prediksi atau pengambilan keputusan yang lebih akurat.

4. Hasil dan Pembahasan

4.1 Deskripsi Data

Dalam penelitian ini, fokus diberikan pada analisis perilaku pelanggan terkait penjualan mobil pada suatu perusahaan otomotif di Makassar dalam rentang waktu Januari 2018 hingga Juni 2023. Data yang diolah termasuk jenis kelamin, usia, dan tipe transmisi pelanggan.

4.2 Preprocessing Data

Preprocessing data dalam konteks penelitian ini menjadi langkah penting untuk menormalisasi data serta penanganan atribut yang hilang (missing value), serta mempersiapkan data agar siap untuk dimodelkan dengan teknik ensemble logistic regression. Berikut adalah tahapan-tahapan preprocessing data yang dapat dilakukan:

a. Normalisasi dan Standarisasi

Melakukan normalisasi atau standarisasi pada fitur-fitur numerik seperti usia. Normalisasi dapat membantu menghindari masalah skala yang berlebihan dan menghasilkan representasi data yang seragam. Metode normalisasi yang umum digunakan dan paling sederhana adalah Min-Max Scaling [25] dengan rumus sebagai berikut:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (6)$$

b. Encoding Fitur Kategorikal

Fitur jenis kelamin (gender) merupakan data kategorikal yang perlu diubah menjadi representasi numerik agar dapat dimasukkan ke dalam model. Teknik encoding yang umum digunakan adalah one-hot encoding, di mana setiap kategori akan diubah menjadi kolom terpisah dengan nilai biner 0 dan 1[26].

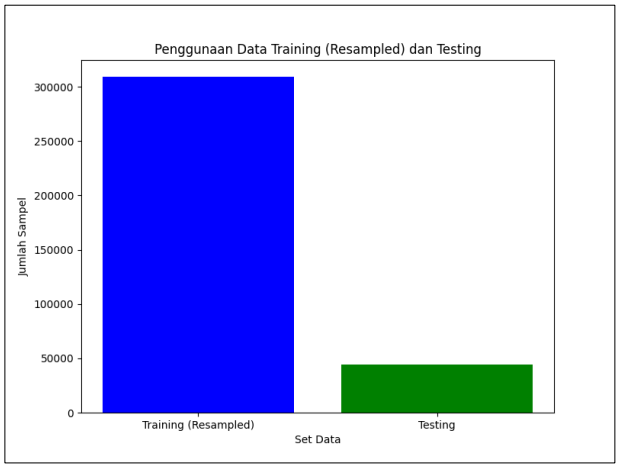
c. Pemisahan Data

Data perlu dipisahkan menjadi fitur (X) dan target (y) yang akan digunakan dalam pemodelan. Fitur terdiri dari jenis kelamin dan usia yang telah di-preprocess, sedangkan target adalah tipe transmisi (manual atau otomatis).

d. Pembagian Data

Terakhir, data dapat dibagi menjadi data pelatihan (training data) dan data uji (testing data) untuk evaluasi model. Pembagian ini dilakukan untuk mengukur kinerja

model pada data yang belum pernah dilihat sebelumnya. Pada penelitian ini dibagi menjadi 80:20 yaitu 80% untuk data training dan 20% untuk data testing.



Gambar 2. Pembagian Data

4.2 Implementasi Ensemble Logistic Regression

Penelitian ini akan membandingkan penggunaan metode logistic regression dengan atau tanpa Teknik Ensemble

a. Logistic Regression

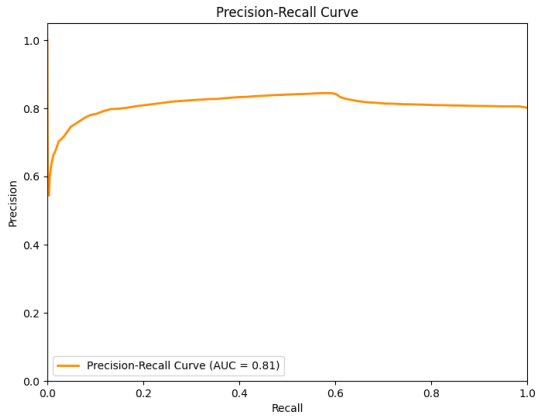
Dari hasil pengukuran yang terlihat pada Tabel 1 mencerminkan bahwa kinerja suatu model dalam memprediksi dua kelas berbeda yaitu kelas 0 dan kelas 1 dimana kelas 0 merupakan transmisi otomatis dan kelas 1 adalah transmisi manual. Dalam hal ini, presisi (precision) untuk kelas 0 adalah 0.25, yang mengindikasikan bahwa ketika model memprediksi kelas 0, hanya sekitar 25% prediksinya yang benar-benar relevan dengan kelas 0. Recall untuk kelas 0 adalah 0.54, yang menunjukkan bahwa model mampu mengidentifikasi sekitar 54% dari keseluruhan instance kelas 0. F1-score untuk kelas 0 adalah 0.34, yang merupakan ukuran gabungan dari presisi dan recall.

Di sisi lain, kelas 1 memiliki hasil yang lebih baik, dengan presisi sekitar 0.84, yang artinya sebagian besar prediksi positif model relevan dengan kelas 1. Recall untuk kelas 1 adalah 0.60, yang mengindikasikan bahwa model dapat mengidentifikasi sekitar 60% dari keseluruhan instance kelas 1. F1-score untuk kelas 1 adalah 0.70, yang mencerminkan kualitas keseluruhan prediksi positif untuk kelas 1.

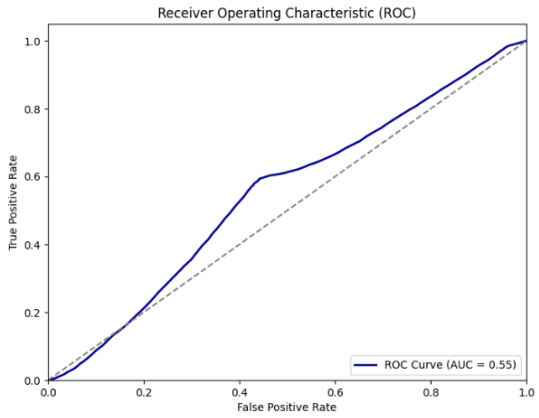
Akurasi model secara keseluruhan adalah sekitar 59%, yang mengindikasikan sejauh mana model dapat memprediksi kelas dengan benar. Rata-rata tingkat makro (macro avg) adalah rata-rata sederhana dari metrik-metrik (presisi, recall, dan F1-score) untuk kedua kelas, dan rata-rata tingkat berbobot (weighted avg) adalah rata-rata bobot yang memperhitungkan seberapa besar kelas masing-masing dalam data. Dalam kasus ini, rata-rata tingkat berbobot memberikan gambaran tentang kualitas keseluruhan prediksi model dan menunjukkan bahwa model memiliki kinerja yang lebih baik dalam mengklasifikasikan kelas 1.

Tabel 1. Metrik Klasifikasi Metode Logistic Regression

Class	Precision	Recall	F1-Score
0	0.25	0.54	0.34
1	0.84	0.60	0.70
Macro Avg	0.55	0.57	0.52
Weighted Avg	0.72	0.59	0.63



Gambar 3. Logistic Regression PRC



Gambar 4. Logistic Regression ROC

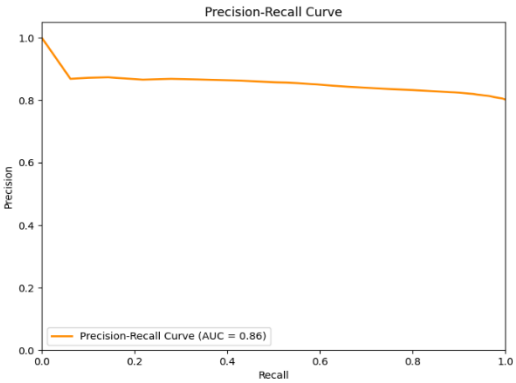
b. Logistic Regression dengan Teknik Boosting

Hasil evaluasi klasifikasi pada table 2 menggambarkan bahwa kinerja model dalam memprediksi dua kelas berbeda: kelas 0 dan kelas 1. Presisi (precision) kelas 0 yaitu 0.34, yang berarti sekitar 34% dari prediksi positif model relevan dengan kelas 0. Namun, recall untuk kelas 0 adalah 0.25, menunjukkan bahwa model hanya dapat mengidentifikasi sekitar 25% dari keseluruhan instance kelas 0. F1-score untuk kelas 0 adalah 0.29. Sementara itu, kelas 1 memiliki kinerja yang lebih baik, dengan presisi sekitar 0.83, yang mengindikasikan bahwa

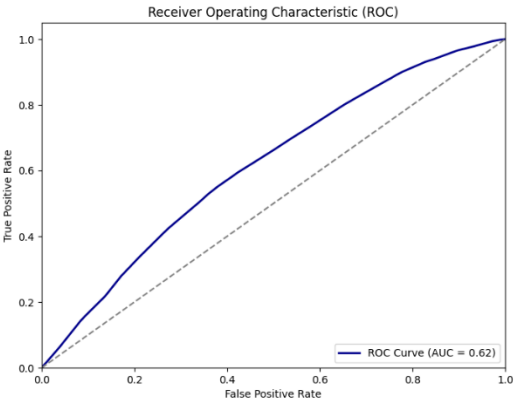
sebagian besar prediksi positif model relevan dengan kelas 1. Recall untuk kelas 1 adalah 0.88, yang artinya model dapat mengidentifikasi sekitar 88% dari keseluruhan instance kelas 1. F1-score untuk kelas 1 adalah 0.85, yang menunjukkan keunggulan prediksi positif untuk kelas 1.
Secara keseluruhan akurasi model adalah sebesar 76%, yang menunjukkan bahwa model memiliki kinerja lebih baik.

Tabel 2. Metrik Klasifikasi Metode Logistic Regression dengan Teknik Boosting

Class	Precision	Recall	F1-Score
0	0.34	0.25	0.29
1	0.83	0.88	0.85
Macro Avg	0.58	0.56	0.57
Weighted Avg	0.73	0.76	0.74



Gambar 5. Logistic Regression dengan Teknik Boosting PRC



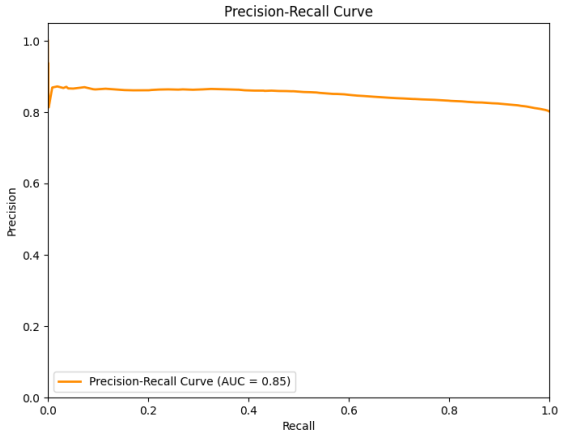
Gambar 6. Logistic Regression dengan Teknik Boosting ROC

c. Logistic Regression dengan Teknik Bagging

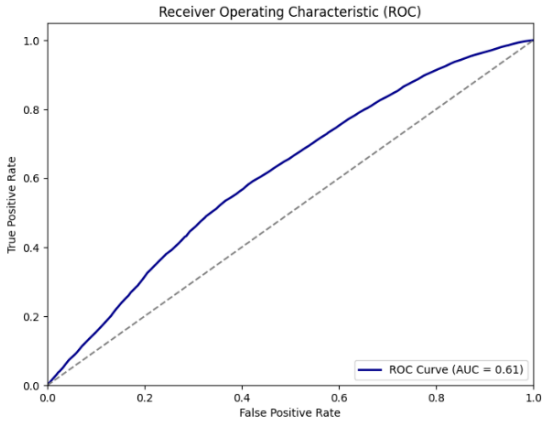
Hasil evaluasi klasifikasi dengan Teknik Bagging ditunjukkan pada tabel 3. Presisi (precision) untuk kelas 0 adalah 0.33, yang berarti dari semua prediksi yang diklasifikasikan sebagai kelas 0, hanya 33% di antaranya yang benar-benar relevan dengan kelas 0. Recall untuk kelas 0 adalah 0.27, yang menandakan bahwa model mampu mengidentifikasi sekitar 27% dari seluruh instance kelas 0. F1-score untuk kelas 0 adalah 0.29, yaitu sebuah ukuran gabungan dari presisi dan recall untuk kelas 0.
Sebaliknya, kelas 1 memiliki hasil yang lebih baik dengan presisi sekitar 0.83, menunjukkan bahwa mayoritas prediksi positif yang diberikan oleh model untuk kelas 1 memang relevan dengan kelas 1. Recall untuk kelas 1 adalah 0.87, mengindikasikan bahwa model mampu mengidentifikasi sekitar 87% dari seluruh instance kelas 1. F1-score untuk kelas 1 adalah 0.85, mencerminkan kualitas keseluruhan prediksi positif untuk kelas 1.
Akurasi model secara keseluruhan adalah sekitar 75%, yang mengukur seberapa baik model dalam memprediksi dengan benar seluruh data. Dalam kasus ini, rata-rata tingkat berbobot menunjukkan bahwa model memiliki kinerja yang lebih baik dalam mengklasifikasikan kelas 1 dibandingkan kelas 0, meskipun masih terdapat ruang untuk perbaikan dalam kedua kelas tersebut.

Tabel 3. Metrik Klasifikasi Metode Logistic Regression dengan Teknik Bagging

Class	Precision	Recall	F1-Score
0	0.33	0.27	0.29
1	0.83	0.87	0.85
Macro Avg	0.58	0.57	0.57
Weighted Avg	0.73	0.75	0.74



Gambar 7. Logistic Regression dengan Teknik Bagging PRC



Gambar 8. Logistic Regression dengan Teknik Bagging ROC

Tabel 4. Evaluasi Kualitas Model

Metode	Precision	Recall
Logistic Regression	0.81	0.55
Logistic Regression dengan Teknik boosting	0.86	0.62
Logistic Regression dengan Teknik bagging	0.85	0.61

Tabel 5. Perbandingan Akurasi

Metode	Akurasi
Logistic Regression	0.59
Logistic Regression dengan Teknik boosting	0.76
Logistic Regression dengan Teknik bagging	0.75

5. Kesimpulan

Dari hasil penelitian diperoleh bahwa perbandingan akurasi dari tiga metode yang berbeda dalam melakukan klasifikasi dengan menggunakan model Logistic Regression tanpa menggunakan teknik ensemble memiliki tingkat akurasi sekitar 0.59, menunjukkan performa model dasar tanpa peningkatan tambahan. Namun, dengan menerapkan teknik ensemble, baik dengan menggunakan metode boosting maupun bagging, akurasi model meningkat secara signifikan. Model yang ditingkatkan dengan teknik boosting mencapai akurasi sekitar 0.76, sementara model dengan teknik bagging mencapai akurasi sekitar 0.75. Teknik boosting dan bagging telah terbukti efektif dalam meningkatkan performa model Logistic Regression, dan pilihan antara keduanya dapat disesuaikan dengan kebutuhan khusus dan karakteristik dataset yang digunakan.

Selain itu, Evaluasi model-model ini dilakukan dengan menggunakan dua metrik penting, yaitu Precision-Recall Curve (PRC) dan Receiver Operating Characteristic (ROC). PRC memberikan kita wawasan tentang sejauh mana model dapat mengklasifikasikan dengan benar kelas minoritas, sedangkan ROC mengukur kinerja keseluruhan model dalam mengklasifikasikan data. Hasil menunjukkan bahwa Logistic Regression dengan Teknik Boosting memiliki nilai tertinggi pada PRC (AUC=0.86) dan ROC (AUC=0.62), diikuti oleh Logistic Regression dengan Teknik Bagging (PRC AUC=0.85, ROC AUC=0.61). Sedangkan Logistic Regression biasa memiliki performa yang lebih rendah dengan PRC AUC=0.81 dan ROC AUC=0.55. Oleh karena itu, model Logistic Regression dengan Teknik Boosting dapat dianggap sebagai pilihan terbaik untuk tugas klasifikasi ini karena mampu memberikan hasil yang lebih baik dalam mengidentifikasi kelas minoritas dan secara keseluruhan lebih baik dalam mengklasifikasikan data.

References

[1] V. Kavitha, G. H. Kumar, S. M. Kumar, and M. Harish, “Churn prediction of customer in telecom industry using machine learning algorithms,” Int. J. Eng. Res. Technol. IJERT, vol. 9, no. 5, pp. 181–184, 2020.

[2] G. R. Raajini S. P. Preethi, R. Shanmuga Priya, L. Rajesh, X. Mercilin, “Customer Behavior Prediction for E-Commerce Sites Using Machine Learning Techniques: An Investigation,” in Reinventing Manufacturing and Business Processes Through Artificial Intelligence, CRC Press, 2021.

[3] A. Roihan, P. A. Sunarya, and A. S. Rafika, “Pemanfaatan Machine Learning dalam Berbagai Bidang,” J. Khatulistiwa Inform., vol. 5, no. 1, p. 490845, 2020.

[4] A. M. Choudhury and K. Nur, “A machine learning approach to identify potential customer based on purchase behavior,” in 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), IEEE, 2019, pp. 242–247.

[5] D. Suryadi and H. M. Kim, “A data-driven approach to product usage context identification from online customer reviews,” J. Mech. Des., vol. 141, no. 12, p. 121104, 2019.

[6] M. Al-Mashraie, S. H. Chung, and H. W. Jeon, “Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach,” Comput. Ind. Eng., vol. 144, p. 106476, 2020.

[7] K. K. Mohbey, “Employee’s attrition prediction using machine learning approaches,” in Machine Learning and Deep Learning in Real-Time Applications, IGI Global, 2020, pp. 121–128.

[8] M. Bahrami, B. Bozkaya, and S. Balcisoy, “Using behavioral analytics to predict customer invoice payment,” Big Data, vol. 8, no. 1, pp. 25–37, 2020.

[9] Q. Wu et al., "Speaking with actions-learning customer journey behavior," in 2019 IEEE 13th International conference on semantic computing (ICSC), IEEE, 2019, pp. 279–286.

[10] S.-H. Mun, Y. Kwak, and J.-H. Huh, "A case-centered behavior analysis and operation prediction of AC use in residential buildings," Energy Build., vol. 188, pp. 137–148, 2019.

[11] E. Bisong, "Logistic Regression," in Building Machine Learning and Deep Learning Models on Google Cloud Platform, Berkeley, CA: Apress, 2019, pp. 243–250. doi: 10.1007/978-1-4842-4470-8_20.

[12] S. Sreejesh, S. Mohapatra, and M. R. Anusree, "Binary Logistic Regression," in Business Research Methods: An Applied Orientation, S. Sreejesh, S. Mohapatra, and M. R. Anusree, Eds., Cham: Springer International Publishing, 2014, pp. 245–258. doi: 10.1007/978-3-319-00539-3_11.

[13] L. B. C. Tanujaya, B. Susanto, and A. Saragih, "The Comparison of Logistic Regression Methods and Random Forest for Spotify Audio Mode Featurre Classification," Indones. J. Data Sci., vol. 1, no. 3, Art. no. 3, Dec. 2020, doi: 10.33096/ijodas.v1i3.16.

[14] A. Shahraki, M. Abbasi, and Ø. Haugen, "Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost," Eng. Appl. Artif. Intell., vol. 94, p. 103770, Sep. 2020, doi: 10.1016/j.engappai.2020.103770.

[15] P. Bahad and P. Saxena, "Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics," in International Conference on Intelligent Computing and Smart Communication 2019, G. Singh Tomar, N. S. Chaudhari, J. L. V. Barbosa, and M. K. Aghwariya, Eds., in Algorithms for Intelligent Systems. Singapore: Springer, 2020, pp. 235–244. doi: 10.1007/978-981-15-0633-8_22.

[16] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, "Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation," Remote Sens., vol. 13, no. 21, Art. no. 21, Jan. 2021, doi: 10.3390/rs13214405.

[17] E. Yaman and A. Subasi, "Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification," BioMed Res. Int., vol. 2019, p. e9152506, Oct. 2019, doi: 10.1155/2019/9152506.

[18] S. García, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining, vol. 72. in Intelligent Systems Reference Library, vol. 72. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-10247-4.

[19] R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," in Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Online: Association for Computational Linguistics, Nov. 2020, pp. 79–91. doi: 10.18653/v1/2020.eval4nlp-1.9.

[20] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLOS ONE, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[21] J. Miao and W. Zhu, "Precision-Recall Curve (PRC) Classification Trees," Evol. Intell., vol. 15, no. 3, pp. 1545–1569, Sep. 2022, doi: 10.1007/s12065-021-00565-2.

[22] S. Macskassy and F. Provost, "Confidence Bands for Roc Curves." Rochester, NY, 2004. Accessed: Aug. 23, 2023. [Online]. Available: https://papers.ssrn.com/abstract=1281313

[23] A. R. Rachakonda and A. Bhatnagar, "ARatio: Extending area under the ROC curve for probabilistic labels," Pattern Recognit. Lett., vol. 150, pp. 265–271, Oct. 2021, doi: 10.1016/j.patrec.2021.06.023.

[24] S. Narkhede, "Understanding auc-roc curve," Data Sci., vol. 26, no. 1, pp. 220–227, 2018.

[25] D. Borkin, A. Némethová, G. Michaľčonok, and K. Maiorov, "Impact of Data Normalization on Classification Model Accuracy," Res. Pap. Fac. Mater. Sci. Technol. Slovak Univ. Technol., vol. 27, no. 45, pp. 79–84, Sep. 2019, doi: 10.2478/rput-2019-0029.

[26] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation?," Emerg. Mark. Finance Trade, vol. 58, no. 2, pp. 472–482, Jan. 2022, doi: 10.1080/1540496X.2020.1825935