

PREDIKSI KEPUASAN PENUMPANG TRANSPORTASI DARAT MENGUNAKAN MACHINE LEARNING: RANDOM FOREST DAN XGBOOST

Andrianto Gusti Pradana ¹, Ma'had Wicaksono ²

Program Studi Sistem Informasi, Fakultas Teknik, Universitas Muhammadiyah Gresik ¹²

Jl. Sumatera 101 GKB, Gresik 61121, Indonesia

e-mail : andriantogusti@umg.ac.id

ABSTRAK

Kepuasan penumpang merupakan indikator kritis keberhasilan penyedia layanan transportasi darat, khususnya kereta api, karena berdampak langsung pada loyalitas penumpang, rekomendasi dari mulut ke mulut, dan pendapatan jangka panjang. Penelitian ini bertujuan menganalisis kinerja dua algoritma Machine Learning, yaitu Random Forest dan XGBoost, dalam memprediksi kepuasan penumpang transportasi darat. Penelitian menggunakan dataset hasil survei kepuasan penumpang kereta api sebanyak 12.340 observasi dengan 18 atribut. Tahapan penelitian meliputi preprocessing data (pembersihan data, encoding, normalisasi, penanganan ketidakseimbangan kelas dengan SMOTE), pembagian data training-testing 80:20 secara stratified, pelatihan model, dan evaluasi menggunakan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa XGBoost mencapai akurasi tertinggi sebesar 93,8%, diikuti Random Forest dengan 91,2%. Analisis fitur menunjukkan bahwa ketepatan waktu kedatangan, kenyamanan kursi, dan kebersihan gerbong merupakan prediktor terkuat kepuasan penumpang. Penelitian ini memberikan rekomendasi implementasi model untuk personalisasi layanan dan deteksi dini keluhan penumpang.

Kata kunci : Machine Learning, Random Forest, XGBoost, Kepuasan Penumpang, Transportasi Darat

ABSTRACT

Passenger satisfaction is a critical indicator of success for land transportation service providers, particularly railways, as it directly affects passenger loyalty, word-of-mouth recommendations, and long-term revenue. This study aims to analyze the performance of two Machine Learning algorithms, Random Forest and XGBoost, in predicting passenger satisfaction in land transportation. The study used a railway passenger satisfaction survey dataset of 12,340 observations with 18 attributes. The research stages included data preprocessing (data cleaning, encoding, normalization, handling class imbalance with SMOTE), stratified 80:20 train-test splitting, model training, and evaluation using accuracy, precision, recall, and F1-score metrics. The results showed that XGBoost achieved the highest accuracy of 93.8%, followed by Random Forest with 91.2%. Feature analysis revealed that on-time arrival, seat comfort, and carriage cleanliness were the strongest predictors of passenger satisfaction. This study provides recommendations for model implementation for service personalization and early detection of passenger complaints.

Keywords : Machine Learning, Random Forest, XGBoost, Passenger Satisfaction, Land Transportation

Jejak Artikel

Upload artikel : 30 Juni 2026

Revisi : 15 Juli 2025

Publish : 31 Juli 2026

1. PENDAHULUAN

Industri transportasi darat, khususnya kereta api, memiliki peran strategis dalam mobilitas masyarakat. Dalam lingkungan persaingan yang semakin ketat dengan moda transportasi lain, kepuasan penumpang menjadi indikator kritis keberhasilan penyedia layanan karena berdampak langsung pada tingkat loyalitas, rekomendasi dari mulut ke mulut, dan pendapatan jangka panjang. Penumpang yang puas cenderung menggunakan layanan secara berulang, sedangkan penumpang yang tidak puas dapat berpindah ke moda transportasi lain dan menyebarkan ulasan negatif melalui media sosial. Oleh karena itu, kemampuan untuk memahami dan memprediksi faktor-faktor yang memengaruhi kepuasan penumpang menjadi kebutuhan strategis bagi penyedia layanan transportasi darat.

Machine Learning (ML) hadir sebagai paradigma komputasi yang memungkinkan sistem untuk belajar dari data historis tanpa diprogram secara eksplisit untuk setiap tugas spesifik. Algoritma ML mampu mengidentifikasi pola tersembunyi dalam data berdimensi tinggi dan menghasilkan prediksi yang semakin akurat seiring bertambahnya data pelatihan. Beberapa algoritma klasifikasi yang terbukti efektif untuk prediksi kepuasan pelanggan antara lain Random Forest dan XGBoost, yang keduanya merupakan metode ensemble berbasis pohon keputusan dengan karakteristik yang berbeda.

Penelitian ini bertujuan untuk: (1) mengimplementasikan algoritma Random Forest dan XGBoost untuk memprediksi kepuasan penumpang transportasi darat; (2) membandingkan kinerja kedua algoritma berdasarkan metrik accuracy, precision, recall, dan F1-score; serta (3) mengidentifikasi faktor-faktor yang paling berpengaruh terhadap kepuasan penumpang dan memberikan rekomendasi strategis bagi penyedia layanan. Kajian ini diharapkan dapat memberikan kontribusi akademis sekaligus panduan praktis bagi pemangku kepentingan industri transportasi darat.

2. METODOLOGI PENELITIAN

2.1 Dataset

Penelitian ini menggunakan dataset hasil survei kepuasan penumpang kereta api yang merupakan adaptasi dari dataset publik layanan transportasi publik. Rincian dataset disajikan pada Tabel 1.

Aspek	Keterangan
Sumber dataset	Survei kepuasan penumpang kereta api (adaptasi dataset publik)
Jumlah record	12.340 observasi
Periode data	Januari – Desember 2023
Jumlah atribut	18 atribut (17 fitur + 1 target)
Target prediksi	Kepuasan (1 = puas, 0 = tidak puas)

Tabel 1. Rincian Dataset Penelitian

2.2 Variabel-variabel Penelitian

Variabel independen (fitur) meliputi: ketepatan waktu kedatangan dan keberangkatan, kenyamanan kursi, kebersihan gerbong, kualitas pelayanan petugas, kemudahan pembelian tiket, kebersihan stasiun, keamanan perjalanan, fasilitas pendingin ruangan, kualitas informasi jadwal, dan skor Net Promoter Score (NPS) historis. Variabel dependen (target) adalah tingkat kepuasan penumpang (1 = puas, 0 = tidak puas).

2.3 Tahap Preprocessing

Tahap preprocessing dilakukan melalui enam fase: (1) data cleaning dengan pengisian missing value (median untuk numerik, modus untuk kategorikal); (2) transformasi variabel kategorikal menggunakan label encoding dan one-hot encoding; (3) normalisasi fitur numerik menggunakan StandardScaler; (4) penanganan ketidakseimbangan kelas dengan SMOTE hanya pada data training; (5) feature selection dengan menghilangkan fitur berkorelasi tinggi ($|r| > 0,9$); dan (6) pembagian data secara stratified split.

2.4 Data Penelitian

Pembagian data menggunakan rasio 80:20 (training:testing) dengan metode stratified random sampling. Rincian pembagian data disajikan pada Tabel 2.

Data	Jumlah	Persentase
Data Training	9.872	80%
Data Testing	2.468	20%
Total	12.340	100%

Tabel 2. Pembagian Data Training dan Testing

2.5 Implementasi Algoritma

2.5.1 Random Forest

Random Forest adalah ensemble method yang menggabungkan prediksi dari sejumlah pohon keputusan menggunakan mekanisme bagging dan random feature selection [1]. Pemilihan split menggunakan kriteria gini dengan formula $Gini = 1 - \sum(p_i)^2$. Konfigurasi: $n_estimators = 100$, $max_depth = 20$, $criterion = gini$, $random_state = 42$.

2.5.2 XGBoost

XGBoost adalah algoritma boosting berbasis pohon keputusan yang membangun model secara iteratif melalui teknik gradient boosting [2]. Konfigurasi: $n_estimators = 100$, $learning_rate = 0,1$, $max_depth = 6$, $subsample = 0,8$, $random_state = 42$. Mekanisme regularisasi (L1 dan L2) efektif mengurangi overfitting.

2.6 Metode Evaluasi Model

Evaluasi menggunakan metrik Accuracy, Precision, Recall, dan F1-Score yang diturunkan dari confusion matrix (TP, TN, FP, FN), dengan formula: $Accuracy = (TP + TN) / (TP + TN + FP + FN)$; $Precision = TP / (TP + FP)$; $Recall = TP / (TP + FN)$; $F1-Score = 2 \times (Precision \times Recall) / (Precision + Recall)$.

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Dataset

Dataset terdiri atas 12.340 observasi dengan 2 kelas target. Distribusi target sebelum penanganan ketidakseimbangan kelas adalah 63,5% puas dan 36,5% tidak puas. Setelah penerapan SMOTE pada data training, proporsi kelas menjadi seimbang (50:50).

3.2 Hasil Preprocessing

Seluruh fase preprocessing berhasil diselesaikan: data duplikat dihapus, missing value diisi, variabel kategorikal ditransformasikan, fitur ternormalisasi, ketidakseimbangan kelas diatasi dengan SMOTE, dan data dibagi 80:20 menjadi 9.872 data training dan 2.468 data testing.

3.3 Hasil Training

Random Forest mencapai akurasi training 97,6% dengan indikasi overfitting ringan yang dimitigasi melalui validasi silang 5-fold; XGBoost mencapai akurasi training 96,8% dengan regularisasi yang bekerja baik dalam mengendalikan overfitting.

3.4 Hasil Pengujian

Pengujian dilakukan pada data testing (2.468 observasi). Hasil pengujian setiap model disajikan pada Tabel 3.

Algoritma	Accuracy	Precision	Recall	F1
XGBoost	93,8%	0,94	0,93	0,93
Random Forest	91,2%	0,91	0,90	0,90

Tabel 3. Hasil Pengujian Algoritma pada Data Testing

Analisis confusion matrix menunjukkan kesalahan klasifikasi kedua model relatif kecil. XGBoost menghasilkan jumlah false negative yang lebih sedikit dibandingkan Random Forest, sehingga lebih mampu menangkap penumpang yang tidak puas.

3.5 Perbandingan Kinerja

Tabel 4 menyajikan perbandingan kinerja kedua algoritma. XGBoost unggul pada seluruh metrik dengan akurasi 93,8%, precision 0,94, recall 0,93, dan F1-score 0,93.

Peringkat	Algoritma	Accuracy	F1
1	XGBoost	93,8%	0,93
2	Random Forest	91,2%	0,90

Tabel 4. Perbandingan Kinerja Seluruh Algoritma

3.6 Analisis Hasil dan Pembahasan

XGBoost menghasilkan performa terbaik dibandingkan Random Forest dalam memprediksi kepuasan penumpang transportasi darat. Keunggulan XGBoost dapat dijelaskan oleh mekanisme gradient boosting yang membangun model secara iteratif dengan memperbaiki kesalahan prediksi sebelumnya, serta dilengkapi regularisasi yang efektif mengurangi overfitting [2]. Temuan ini sejalan dengan penelitian sebelumnya [11].

Random Forest menempati posisi kedua dengan akurasi 91,2%. Sebagai metode bagging, Random Forest mampu mengurangi variance, namun kurang optimal dalam menangani pola non-linear yang kompleks. Overfitting ringan pada Random Forest juga menurunkan performa pengujiannya.

Analisis feature importance mengungkapkan bahwa ketepatan waktu kedatangan, kenyamanan kursi, dan kebersihan gerbong merupakan prediktor terkuat kepuasan penumpang. Temuan ini konsisten dengan teori kualitas layanan yang menekankan aspek reliability dan tangible [10][12].

Implementasi model memungkinkan penyedia layanan melakukan personalisasi layanan, deteksi dini keluhan, dan prioritas tindakan perbaikan secara otomatis, sehingga loyalitas penumpang dapat dipertahankan [13].

4. KESIMPULAN

Penelitian ini mendemonstrasikan perbandingan kinerja dua algoritma Machine Learning untuk prediksi kepuasan penumpang transportasi darat pada dataset 12.340 observasi. Hasil penelitian menunjukkan bahwa XGBoost mencapai akurasi tertinggi sebesar 93,8%, diikuti Random Forest dengan 91,2%. XGBoost unggul karena mekanisme gradient boosting dan regularisasi yang efektif mengurangi overfitting.

Saran untuk penelitian selanjutnya: (1) optimasi hyperparameter menggunakan grid search atau random search; (2) menambahkan data ulasan teks penumpang dengan analisis sentimen; (3) mengeksplorasi metode deep learning seperti ANN dan LSTM; dan (4) menerapkan model pada moda transportasi darat lain, seperti bus dan angkutan perkotaan.

DAFTAR PUSTAKA

- [1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [3] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
- [4] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
- [5] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.
- [6] 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.
- [7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer, 2013.
- [8] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.
- [9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
- [10] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality," *Journal of Retailing*, vol. 64, no. 1, pp. 12–40, 1988.
- [11] J. De Oña and R. de Oña, "Quality of service in public transport based on customer satisfaction surveys: A review and assessment of methodological approaches," *Transportation Science*, vol. 49, no. 3, pp. 605–622, 2015.
- [12] C.-F. Chen, "Investigating structural relationships between service quality, perceived value, satisfaction, and behavioral intentions for air passengers," *Transportation Research Part A*, vol. 42, no. 4, pp. 709–717, 2008.
- [13] J. T. Bowen and S.-L. Chen, "The relationship between customer loyalty and customer satisfaction," *International Journal of Contemporary Hospitality Management*, vol. 13, no. 5, pp. 213–217, 2001.
- [14] G. C. Saha and Theingi, "Service quality, satisfaction, and behavioural intentions: A study of low-cost airline carriers in Thailand," *Managing Service Quality*, vol. 19, no. 3, pp. 350–372, 2009.
- [15] E. Nathanail, "Measuring the quality of service for passengers on the Hellenic railways," *Transportation Research Part A*, vol. 42, no. 1, pp. 48–66, 2008.