

# Prediksi Preferensi Kelas Layanan Penumpang Bus pada Periode Mudik Lebaran Menggunakan Machine Learning

Rahmatika Yuandra<sup>1</sup>, Mart Kellin<sup>2</sup>, and Patimah Pulungan<sup>3</sup>

**\*Correspondent :**

Email: tikayuandra30@gmail.com

**Affiliation :**

<sup>1, 2, 3</sup> UIN Sjech M. Djamil Djambek  
Bukittinggi , Indonesia

**Submit Article :**

Submission : June 16, 2026

Revision : June 17, 2026

Accepted : June 17, 2026

Published : June 30, 2026

**Keywords :**

Machine Learning, Random Forest,  
XGBoost, Pilihan Penumpang,  
Transportasi Bus, Mudik Lebaran.

**Abstrak**

Perkembangan data transportasi membuka peluang pemanfaatan machine learning untuk mendukung pengambilan keputusan dalam layanan transportasi publik. Namun, penelitian mengenai klasifikasi pilihan kelas layanan bus antar kota selama periode mudik Lebaran masih relatif terbatas. Penelitian ini bertujuan untuk menganalisis dan mengklasifikasikan pilihan penumpang terhadap kelas layanan bus menggunakan algoritma Random Forest dan XGBoost. Dataset yang digunakan terdiri dari 3.942 data penumpang dari enam perusahaan bus antar kota selama periode 1–31 Maret 2026. Fitur yang digunakan meliputi hari relatif terhadap Hari Raya Idulfitri, kategori akhir pekan (weekend), wilayah tujuan, status pembayaran, dan perusahaan bus. Data diproses melalui tahapan data cleaning, feature engineering, exploratory data analysis, dan pemodelan machine learning. Hasil penelitian menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi 79,21%, sedikit lebih tinggi dibandingkan XGBoost sebesar 78,96%. Analisis feature importance menunjukkan bahwa wilayah tujuan (35,24%) dan hari relatif terhadap Idulfitri (28,02%) merupakan faktor yang paling berpengaruh terhadap pilihan kelas layanan penumpang. Penelitian ini memberikan wawasan mengenai perilaku penumpang selama periode mudik serta mendukung pengambilan keputusan berbasis data bagi perusahaan bus antar kota.

**Abstract**

The growth of transportation data has created opportunities to leverage machine learning for supporting decision-making in public transportation services. However, research on classifying intercity bus service class preferences during the Eid al-Fitr homecoming (Mudik) period remains relatively limited. This study aims to analyze and classify passengers' service class preferences using the Random Forest and XGBoost algorithms. The dataset consists of 3,942 passenger records collected from six intercity bus companies during the period of March 1–31, 2026. The features used include the relative day to Eid al-Fitr, weekend category, destination region, payment status, and bus company. The data were processed through data cleaning, feature engineering, exploratory data analysis, and machine learning modeling. The results indicate that the Random Forest algorithm achieved the best performance with an accuracy of 79.21%, slightly outperforming XGBoost, which achieved an accuracy of 78.96%. Feature importance analysis revealed that the destination region (35.24%) and the relative day to Eid al-Fitr (28.02%) were the most influential factors affecting passengers' service class preferences. This study provides insights into passenger behavior during the Eid al-Fitr homecoming period and supports data-driven decision-making for intercity bus companies.

## PENDAHULUAN

Transportasi darat, khususnya bus antar kota, merupakan tulang punggung mobilitas masyarakat Indonesia selama periode mudik Lebaran. Berdasarkan data Kementerian Perhubungan, pada tahun 2025 pergerakan pemudik diproyeksikan mencapai 146,48 juta orang dengan 24,76 juta di antaranya (16,9%) menggunakan moda bus (Halim, 2025). Memasuki tahun 2026, meskipun total pergerakan diprediksi sedikit menurun menjadi 143,91 juta orang, bus tetap menjadi moda angkutan umum yang paling diminati dengan kontribusi sebesar 16,22% atau sekitar 23,34 juta penumpang (Tim Survei Angleb Badan Kebijakan Transportasi, 2026). Fenomena nasional ini berdampak langsung pada Kota Bukittinggi sebagai salah satu hub transportasi utama di Sumatera Barat, di mana terminal dan loket bus selalu mengalami lonjakan penumpang setiap tahunnya. Dinamika ini menuntut perusahaan otobus seperti NPM, ANS, SAN, TAM, P3, dan RJT untuk memahami pola pilihan kelas layanan penumpang (Reguler vs Premium). Ketidakmampuan dalam memprediksi pilihan kelas ini dapat menyebabkan ketidakefisienan dalam alokasi armada dan kegagalan dalam memenuhi ekspektasi

kenyamanan penumpang pada hari-hari puncak menjelang Lebaran. Oleh karena itu, klasifikasi pilihan kelas layanan berbasis data transaksi menjadi krusial untuk mendukung strategi operasional yang lebih tepat sasaran.

Perkembangan teknologi informasi telah mendorong pemanfaatan data operasional transportasi sebagai sumber informasi untuk memahami perilaku penumpang secara lebih objektif. Berbeda dengan pendekatan survei yang bergantung pada persepsi responden, data transaksi mampu merekam aktivitas perjalanan aktual sehingga dapat digunakan untuk mengidentifikasi karakteristik, pola perjalanan, loyalitas, hingga kebiasaan penggunaan layanan transportasi. Informasi tersebut menjadi dasar dalam memahami perilaku penumpang serta mendukung pengambilan keputusan operasional berbasis data (Daneshvar et al., 2023; Wang et al., 2023). Data transaksi penumpang dapat dimanfaatkan untuk membangun profil perjalanan individu secara akurat guna mendukung perencanaan jaringan transportasi dan personalisasi layanan (Wang et al., 2023). Selain itu, pendekatan data mining pada data transaksi mampu mengungkap pola perilaku penumpang yang bermanfaat dalam penyusunan jadwal operasional, pengelolaan layanan, serta peningkatan kualitas sistem transportasi (Daneshvar et al., 2023).

Seiring meningkatnya ketersediaan data transportasi, penerapan machine learning juga berkembang pesat dalam berbagai permasalahan transportasi publik. Berdasarkan tinjauan sistematis, algoritma machine learning telah banyak digunakan untuk prediksi waktu kedatangan bus, prediksi jumlah penumpang, prediksi permintaan layanan, serta analisis perilaku perjalanan (Alexandre et al., 2023). Dibandingkan metode statistik konvensional, machine learning mampu menangani hubungan yang kompleks dan nonlinier antarvariabel serta mengolah data dalam jumlah besar dengan tingkat akurasi yang lebih (Ruiz et al., 2024). Kemampuan tersebut menjadikan machine learning sebagai salah satu pendekatan yang potensial dalam mendukung pengambilan keputusan berbasis data pada sektor transportasi.

Penelitian mengenai transportasi publik yang telah dilakukan sebelumnya umumnya dapat dikelompokkan ke dalam tiga fokus utama. Kelompok pertama berfokus pada evaluasi kualitas pelayanan dan kepuasan penumpang untuk mengidentifikasi faktor-faktor yang memengaruhi kualitas layanan transportasi. Berbagai penelitian menunjukkan bahwa aspek seperti ketepatan waktu, fasilitas, kenyamanan, serta kualitas pelayanan memiliki pengaruh yang signifikan terhadap tingkat kepuasan pengguna transportasi publik (Ajeng et al., 2023; Buran, 2025; Fadhilah & Amalia, 2021). Kelompok kedua menitikberatkan pada analisis pilihan moda transportasi (travel mode choice) dengan mempertimbangkan faktor sosial ekonomi, biaya perjalanan, keamanan, kenyamanan, maupun waktu tempuh sebagai dasar pengambilan keputusan perjalanan (Hillel et al., 2021; MAULANA, 2024; Zhang et al., 2023). Sementara itu, kelompok ketiga lebih banyak menerapkan algoritma machine learning untuk memprediksi jumlah penumpang, permintaan layanan, maupun arus penumpang guna mendukung perencanaan operasional transportasi (Alexandre et al., 2023; Jumlah Penumpang Bus Rahmawati et al., 2025; Li et al., 2023; Stadler et al., 2021).

Berdasarkan sintesis penelitian terdahulu, masih terdapat beberapa kesenjangan penelitian yang perlu mendapat perhatian. Sebagian besar penelitian masih berfokus pada evaluasi kepuasan penumpang, analisis pemilihan moda transportasi, atau prediksi jumlah penumpang sebagai dasar perencanaan layanan transportasi. Penelitian yang menerapkan machine learning juga umumnya diarahkan untuk memprediksi permintaan perjalanan, arus penumpang, maupun pilihan moda transportasi, sedangkan penerapannya untuk mengklasifikasikan pilihan kelas layanan bus antarkota masih sangat terbatas. Selain itu, sebagian besar penelitian memanfaatkan data survei, data smart card, atau data operasional transportasi perkotaan, sementara penggunaan data transaksi aktual yang berasal dari proses reservasi perusahaan otobus antarkota masih jarang dilaporkan. Penelitian yang secara khusus mengkaji pilihan kelas layanan bus pada periode mudik Lebaran juga masih terbatas, padahal periode tersebut memiliki karakteristik perjalanan yang berbeda dibandingkan hari biasa akibat meningkatnya permintaan penumpang, perubahan pola perjalanan, serta keterbatasan kapasitas armada.

Berdasarkan kesenjangan tersebut, penelitian ini menawarkan beberapa kebaruan. Pertama, penelitian memanfaatkan data transaksi operasional nyata yang diperoleh dari pencatatan reservasi enam perusahaan otobus antarkota sehingga mampu merepresentasikan pilihan aktual (revealed choice) penumpang, bukan preferensi yang diperoleh melalui pendekatan survei. Kedua, penelitian difokuskan pada klasifikasi pilihan kelas layanan bus antarkota (REGULER dan PREMIUM) selama periode mudik Lebaran tahun 2026 yang hingga saat ini masih belum banyak dikaji dalam penelitian sebelumnya. Ketiga, penelitian membandingkan kinerja algoritma Random Forest dan XGBoost dalam mengklasifikasikan pilihan kelas layanan sekaligus mengidentifikasi faktor-faktor yang paling berpengaruh melalui analisis feature importance. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan penerapan machine learning pada sektor transportasi serta menjadi dasar bagi perusahaan otobus dalam mengoptimalkan alokasi armada, menyusun strategi operasional, dan meningkatkan kualitas layanan berdasarkan pola pilihan penumpang.

## METODOLOGI PENELITIAN

### 2.1. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode klasifikasi berbasis machine learning untuk menganalisis pilihan kelas layanan bus antar kota pada periode mudik Lebaran. Proses penelitian mengadopsi tahapan Cross Industry Standard Process for Data Mining (CRISP-DM), yang merupakan salah satu metodologi yang banyak digunakan dalam proyek data science karena menyediakan alur kerja yang sistematis mulai dari pemahaman masalah hingga evaluasi model (Rianti et al., n.d.). Dalam penelitian ini, tahapan CRISP-DM disesuaikan dengan kebutuhan penelitian sehingga difokuskan pada proses pengumpulan data, pra-pemrosesan data, rekayasa fitur, analisis data eksploratif, pemodelan machine learning, dan evaluasi model.

Secara umum, penelitian diawali dengan proses pengumpulan data transaksi pemesanan tiket bus antar kota dari enam perusahaan otobus yang melayani perjalanan selama periode mudik Lebaran. Data yang telah diperoleh kemudian melalui tahapan data wrangling untuk membersihkan data yang tidak lengkap, tidak konsisten, maupun duplikasi data. Selanjutnya dilakukan proses feature engineering untuk membentuk variabel baru yang dianggap mampu meningkatkan kemampuan model dalam mengenali pola pilihan kelas layanan penumpang. Setelah itu, dilakukan Exploratory Data Analysis (EDA) guna memahami karakteristik dataset melalui statistik deskriptif dan visualisasi data sebelum proses pemodelan dilakukan.

Tahap berikutnya adalah pembangunan model klasifikasi menggunakan algoritma Random Forest dan XGBoost. Kedua algoritma dipilih karena merupakan metode ensemble learning berbasis pohon keputusan (decision tree) yang memiliki kemampuan baik dalam menangani data tabular serta mampu memodelkan hubungan nonlinier antarvariabel. Performa kedua model kemudian dievaluasi menggunakan beberapa metrik, yaitu accuracy, precision, recall, F1-score, dan confusion matrix. Selain itu, dilakukan pengujian menggunakan 5-fold cross validation untuk mengukur kestabilan model pada beberapa subset data, serta analisis feature importance untuk mengidentifikasi variabel yang paling berpengaruh terhadap hasil klasifikasi.

### 2.2. Sumber Data

Data yang digunakan dalam penelitian ini merupakan data primer yang diperoleh dari pencatatan transaksi pemesanan tiket pada enam perusahaan otobus (PO), yaitu NPM, ANS, SAN, TAM, P3, dan RJT, yang melayani perjalanan bus antar kota selama periode mudik Lebaran. Data dikumpulkan melalui admin loket masing-masing perusahaan dalam bentuk pencatatan manual pada buku transaksi selama periode 1–31 Maret 2026. Selanjutnya, data tersebut ditranskripsikan ke dalam format digital sehingga dapat diproses pada tahapan analisis. Penggunaan data operasional berupa rekaman transaksi dinilai mampu memberikan gambaran perilaku penumpang yang lebih objektif dibandingkan data survei karena merekam aktivitas perjalanan aktual tanpa dipengaruhi persepsi responden (Daneshvar et al., 2023; Wang et al., 2023).

Penggunaan data dalam penelitian ini telah memperoleh izin dari pihak perusahaan otobus untuk keperluan penelitian akademik. Data yang diperoleh selanjutnya dipersiapkan melalui tahapan pra-pemrosesan sebelum digunakan pada proses analisis dan pemodelan machine learning. Tahapan tersebut bertujuan untuk memastikan kualitas data sehingga layak digunakan dalam proses pembangunan model klasifikasi.

### 2.3. Pra-pemrosesan Data

Tahap pra-pemrosesan data (data preprocessing) dilakukan untuk meningkatkan kualitas dataset sebelum digunakan dalam proses analisis dan pemodelan machine learning. Tahapan ini bertujuan menghasilkan data yang konsisten, lengkap, dan siap dianalisis sehingga mampu meningkatkan performa model klasifikasi. Proses data preprocessing meliputi berbagai tahapan seperti data cleaning, transformasi data, integrasi data, dan validasi untuk mengurangi inkonsistensi serta meningkatkan kualitas data (Bala & Behal, 2024). Selain itu, penanganan missing value menjadi salah satu tahapan penting karena data yang tidak lengkap dapat memengaruhi akurasi model apabila tidak ditangani dengan tepat (Emmanuel et al., 2021).

Data penelitian diperoleh dalam enam dataset yang terpisah sesuai dengan masing-masing perusahaan otobus, yaitu NPM, ANS, TAM, SAN, P3, dan RJT. Sebelum dilakukan proses penggabungan, seluruh dataset memiliki struktur atribut yang berbeda sehingga dilakukan standarisasi nama kolom agar memiliki format yang seragam. Selanjutnya ditambahkan atribut `pt_bus` sebagai identitas perusahaan otobus untuk mempertahankan informasi asal data setelah proses integrasi dataset. Tahap berikutnya meliputi standarisasi format teks pada atribut kategorikal, seperti jenis bus, tujuan, dan status pembayaran, serta konversi atribut tanggal ke dalam tipe data `datetime` agar dapat digunakan pada proses rekayasa fitur.

Tahap data cleaning dilakukan melalui penanganan missing value dan penghapusan data duplikat. Penanganan missing value dilakukan dengan pendekatan yang disesuaikan dengan karakteristik masing-masing atribut. Nilai kosong pada atribut kategorikal, seperti nama, nomor kursi, tujuan, jenis bus, plat bus, dan nomor telepon, diganti menggunakan label yang merepresentasikan informasi tidak tersedia, seperti "NAMA TIDAK DIKETAHUI", "TIDAK DIKETAHUI", dan "TUJUAN TIDAK DIKETAHUI". Sementara itu, atribut numerik seperti uang muka yang tidak memiliki nilai diisi dengan nilai 0 sesuai kondisi transaksi. Selain penanganan missing value, dilakukan penghapusan data duplikat untuk memastikan setiap transaksi hanya direpresentasikan satu kali pada dataset.

Sebelum proses pra-pemrosesan dilakukan, data yang diperoleh dari enam perusahaan otobus berjumlah 3.962 transaksi. Hasil identifikasi menunjukkan terdapat 2.106 missing value yang tersebar pada beberapa atribut di masing-masing dataset, serta 20 data duplikat yang ditemukan pada dataset NPM dan ANS. Setelah dilakukan proses data cleaning, seluruh missing value berhasil ditangani sehingga tidak terdapat lagi nilai kosong pada dataset dan seluruh data duplikat berhasil dihapus. Hasil prapemrosesan dataset adalah 3.942 transaksi yang valid untuk analisis selanjutnya.

Dataset hasil pra-pemrosesan selanjutnya memiliki struktur atribut yang konsisten, tidak mengandung data duplikat, serta telah memenuhi kebutuhan analisis sehingga siap digunakan pada tahap feature engineering, exploratory data analysis, dan pemodelan menggunakan algoritma Random Forest maupun XGBoost.

### 2.4. Feature Engineering

Tahap feature engineering dilakukan untuk membentuk atribut yang lebih representatif sehingga mampu meningkatkan kemampuan model dalam mengenali pola pada data. Proses ini bertujuan menghasilkan fitur yang memiliki informasi lebih bermakna dibandingkan atribut asli sehingga dapat meningkatkan performa dan interpretabilitas model klasifikasi. Feature engineering merupakan salah satu tahapan penting dalam machine learning karena mampu meningkatkan akurasi model melalui pembentukan maupun transformasi fitur tanpa harus mengubah algoritma yang digunakan (Verdonck et al., 2021).

Pada penelitian ini dilakukan beberapa proses rekayasa fitur yang meliputi pembentukan variabel Hari ke Lebaran, Kategori Hari, Wilayah Tujuan, dan Kelas Layanan. Fitur Hari ke Lebaran diperoleh dari perhitungan selisih antara tanggal keberangkatan dengan Hari Raya Idulfitri tahun 2026 sehingga menunjukkan kedekatan waktu keberangkatan terhadap puncak arus mudik. Nilai yang semakin mendekati nol menunjukkan jadwal keberangkatan yang semakin dekat dengan Hari Raya, sedangkan nilai negatif menunjukkan jumlah hari sebelum Lebaran.

Selanjutnya, atribut tanggal ditransformasikan menjadi fitur Kategori Hari dengan mengelompokkan setiap tanggal keberangkatan ke dalam kategori Weekday dan Weekend. Pengelompokan ini dilakukan untuk mengetahui pengaruh jenis hari terhadap pilihan kelas layanan. Selain itu, atribut tujuan yang memiliki banyak variasi penulisan dan nama kota dikelompokkan ke dalam beberapa wilayah, yaitu Jabodetabek, Jawa Barat, Banten, Lampung, Sumatera Utara, Aceh, Riau, Sumatera Selatan, Bengkulu, Lainnya, dan Tidak Diketahui. Pengelompokan wilayah bertujuan mengurangi jumlah kategori sekaligus merepresentasikan karakteristik geografis tujuan perjalanan.

Rekayasa fitur juga dilakukan pada atribut jenis bus dengan mengelompokkan berbagai tipe bus menjadi dua kategori kelas layanan, yaitu REGULER dan PREMIUM. Proses ini menghasilkan variabel target yang digunakan pada proses klasifikasi. Ringkasan fitur hasil rekayasa ditunjukkan pada Tabel 1.

**Tabel 1.** Hasil Feature Engineering

| Fitur           | Tipe Data   | Cara Pembentukan                                                                   | Keterangan                                                                                                                                             |
|-----------------|-------------|------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| hari_ke_lebaran | Integer     | Menghitung selisih antara tanggal keberangkatan dan Hari Raya Idulfitri tahun 2026 | Menunjukkan kedekatan waktu keberangkatan terhadap Hari Raya. Nilai yang semakin mendekati nol menunjukkan keberangkatan semakin dekat dengan Lebaran. |
| weekend         | Kategorikal | Mengelompokkan tanggal menjadi Weekday dan Weekend                                 | Digunakan untuk mengetahui pengaruh jenis hari terhadap pilihan kelas layanan.                                                                         |
| wilayah_tujuan  | Kategorikal | Mengelompokkan berbagai kota tujuan ke dalam wilayah geografis                     | Mengurangi variasi kategori tujuan serta merepresentasikan lokasi tujuan perjalanan.                                                                   |
| kelas_layanan   | Kategorikal | Mengelompokkan berbagai jenis bus menjadi REGULER dan PREMIUM                      | Digunakan sebagai variabel target (target variable) pada proses klasifikasi.                                                                           |

**Sumber :** Hasil pengolahan data peneliti (2026).

## 2.5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) merupakan tahapan awal analisis data yang bertujuan untuk memahami karakteristik dataset sebelum proses pemodelan machine learning dilakukan. Tahap ini dilakukan dengan menyajikan ringkasan statistik deskriptif serta berbagai bentuk visualisasi data sehingga pola, distribusi, hubungan antarvariabel, maupun potensi anomali pada data dapat diidentifikasi. Informasi yang diperoleh dari EDA membantu peneliti memahami struktur data dan menjadi dasar dalam menentukan langkah analisis serta pemodelan yang tepat (Hodeghatta & Nayak, 2023).

Pada penelitian ini, EDA dilakukan terhadap dataset hasil pra-pemrosesan dan rekayasa fitur untuk memperoleh gambaran mengenai distribusi kelas layanan, karakteristik variabel kategorikal, hubungan antarvariabel, serta pola data yang berkaitan dengan pilihan kelas layanan bus. Hasil eksplorasi tersebut kemudian digunakan sebagai dasar dalam proses pembangunan model klasifikasi menggunakan algoritma Random Forest dan XGBoost.

Tahapan EDA yang dilakukan pada penelitian ini meliputi distribusi kelas layanan, distribusi wilayah tujuan, distribusi hari ke lebaran, distribusi status pembayaran, distribusi kategori hari, distribusi kelas layanan berdasarkan wilayah tujuan, distribusi kelas layanan berdasarkan hari ke lebaran, distribusi kelas layanan berdasarkan perusahaan bus, distribusi kelas layanan berdasarkan

status pembayaran, distribusi kelas layanan berdasarkan kategori hari (weekday dan weekend), proporsi kelas layanan berdasarkan wilayah tujuan.

## 2.6. Pemodelan Machine Learning

Pemodelan machine learning dilakukan untuk mengklasifikasikan pilihan kelas layanan bus berdasarkan karakteristik transaksi penumpang. Untuk memperoleh evaluasi yang lebih komprehensif, penelitian ini menggunakan tiga algoritma dengan tingkat kompleksitas yang berbeda, yaitu Logistic Regression sebagai baseline model, Random Forest sebagai algoritma ensemble berbasis bagging, dan Extreme Gradient Boosting (XGBoost) sebagai algoritma ensemble berbasis boosting.

Logistic Regression digunakan sebagai baseline untuk memberikan titik pembandingan terhadap model berbasis pohon keputusan. Penggunaan baseline diperlukan untuk mengetahui apakah peningkatan performa yang diperoleh dari Random Forest dan XGBoost benar-benar memberikan keuntungan dibandingkan model klasifikasi yang lebih sederhana. Random Forest digunakan karena mampu membangun sejumlah decision tree melalui proses bootstrap aggregating dan menggabungkan hasil prediksi melalui majority voting. Pendekatan tersebut memungkinkan model menangkap hubungan nonlinier antarvariabel serta mengurangi risiko overfitting. Selanjutnya, XGBoost digunakan sebagai model berbasis gradient boosting yang membangun pohon keputusan secara bertahap dengan memperbaiki kesalahan prediksi dari model sebelumnya. Ketiga model dilatih menggunakan data training yang sama dan dievaluasi menggunakan data testing yang sama sehingga perbandingan performa dapat dilakukan secara adil. Kedua model dipilih karena memiliki kemampuan yang baik dalam menangani hubungan nonlinier antarvariabel, menghasilkan akurasi yang tinggi, serta memiliki kemampuan generalisasi yang baik pada berbagai permasalahan klasifikasi (Lan et al., 2025, 2026)

Seluruh model menggunakan pembagian data training dan testing dengan rasio 80:20 serta stratified sampling untuk mempertahankan proporsi kelas REGULER dan PREMIUM pada kedua subset data. Evaluasi dilakukan menggunakan accuracy, precision, recall, F1-score, ROC-AUC, area under the precision-recall curve (PR-AUC), dan Matthews Correlation Coefficient (MCC). Selain itu, 5-fold stratified cross-validation digunakan untuk memperoleh estimasi performa yang lebih stabil.

### Pemodelan Random Forest

Random Forest merupakan algoritma ensemble learning yang membangun sejumlah decision tree secara acak (bootstrap aggregating) kemudian menggabungkan hasil prediksi setiap pohon menggunakan mekanisme majority voting. Pendekatan tersebut mampu meningkatkan akurasi klasifikasi sekaligus mengurangi risiko overfitting dibandingkan penggunaan satu pohon keputusan (Lan et al., 2026).

Pada penelitian ini, model Random Forest dibangun menggunakan 100 decision tree ( $n\_estimators = 100$ ) dengan  $random\_state = 42$  untuk memastikan hasil pelatihan dapat direproduksi. Parameter lainnya menggunakan nilai bawaan (default parameter) dari pustaka Scikit-learn.

**Tabel 2.** Hyperparameter Random Forest

| Hyperparameter    | Nilai          |
|-------------------|----------------|
| $n\_estimators$   | 100            |
| criterion         | gini(default)  |
| max_depth         | None (default) |
| min_samples_split | 2 (default)    |
| min_samples_leaf  | 1 (default)    |
| bootstrap         | True (default) |
| random_state      | 42             |

**Sumber :** Hasil pengolahan data peneliti (2026).

## Pemodelan XGBoost

XGBoost (Extreme Gradient Boosting) merupakan algoritma berbasis gradient boosting yang membangun pohon keputusan secara bertahap (sequential learning) dengan tujuan memperbaiki kesalahan prediksi dari pohon sebelumnya. Pendekatan tersebut memungkinkan model menghasilkan kemampuan generalisasi yang tinggi serta efektif dalam menangani data klasifikasi dengan kompleksitas yang tinggi (Lan et al., 2025).

Pada penelitian ini, model XGBoost dibangun menggunakan `random_state = 42` untuk menjaga konsistensi hasil pelatihan. Seluruh parameter lainnya menggunakan konfigurasi bawaan (default parameter) dari pustaka XGBoost.

**Tabel 3.** Hyperparameter XGBoost

| Hyperparameter   | Nilai                     |
|------------------|---------------------------|
| objective        | binary:logistic (default) |
| n_estimators     | 100 (default)             |
| learning_rate    | 0.3 (default)             |
| max_depth        | 6 (default)               |
| subsample        | 1.0 (default)             |
| colsample_bytree | 1.0 (default)             |
| random_state     | 42                        |

**Sumber :** Hasil pengolahan data peneliti (2026).

Setelah proses pelatihan selesai, kedua model digunakan untuk melakukan prediksi terhadap data uji. Evaluasi performa model dilakukan menggunakan nilai accuracy, serta classification report yang menyajikan metrik precision, recall, F1-score, dan support. Selain itu, confusion matrix digunakan untuk memvisualisasikan hasil klasifikasi berdasarkan prediksi benar dan salah pada masing-masing kelas. Untuk menguji konsistensi performa model, dilakukan 5-fold cross validation, sedangkan analisis feature importance digunakan untuk mengidentifikasi kontribusi setiap fitur terhadap hasil klasifikasi.

### 2.7. Evaluasi Model

Kinerja model pada penelitian ini dievaluasi menggunakan beberapa metrik klasifikasi yang diperoleh melalui classification report, yaitu accuracy, precision, recall, F1-score, dan support, serta divisualisasikan menggunakan confusion matrix. Penggunaan beberapa metrik evaluasi diperlukan agar performa model tidak hanya dinilai berdasarkan tingkat akurasi secara keseluruhan, tetapi juga berdasarkan kemampuan model dalam mengidentifikasi masing-masing kelas secara lebih komprehensif. Selain itu, penelitian ini juga menerapkan k-fold cross validation untuk menguji konsistensi performa model pada beberapa pembagian data, serta analisis feature importance untuk mengidentifikasi tingkat kontribusi setiap fitur terhadap hasil klasifikasi (Gorritz et al., 2026; Shirdel et al., 2026).

Nilai accuracy menunjukkan proporsi seluruh data yang berhasil diklasifikasikan dengan benar terhadap keseluruhan data uji. Semakin tinggi nilai accuracy, semakin baik kemampuan model dalam melakukan klasifikasi secara umum. Accuracy dihitung menggunakan Persamaan (1).

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

dengan TP (True Positive) adalah data positif yang diprediksi benar, TN (True Negative) adalah data negatif yang diprediksi benar, FP (False Positive) adalah data negatif yang diprediksi positif dan FN (False Negative) adalah data positif yang diprediksi negatif.

Nilai precision digunakan untuk mengukur ketepatan model dalam memberikan prediksi pada suatu kelas, yaitu perbandingan antara jumlah data yang diprediksi benar sebagai suatu kelas dengan seluruh data yang diprediksi berada pada kelas tersebut. Precision yang tinggi menunjukkan bahwa

model memiliki tingkat kesalahan prediksi positif (false positive) yang rendah. Precision dihitung menggunakan Persamaan (2).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Selanjutnya, recall digunakan untuk mengukur kemampuan model dalam menemukan seluruh data yang benar-benar termasuk ke dalam suatu kelas. Nilai recall yang tinggi menunjukkan bahwa model mampu meminimalkan jumlah false negative. Recall dihitung menggunakan Persamaan (3).

$$Recall = \frac{TP + TN}{TP + FN} \quad (3)$$

Sementara itu, F1-score merupakan rata-rata harmonik antara precision dan recall sehingga mampu memberikan ukuran performa yang lebih seimbang, terutama ketika distribusi data antar kelas tidak sepenuhnya seimbang. Nilai F1-score dihitung menggunakan Persamaan (4).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Selain keempat metrik tersebut, support digunakan untuk menunjukkan jumlah data aktual pada masing-masing kelas yang digunakan selama proses evaluasi. Meskipun support bukan merupakan ukuran performa model, informasi ini membantu dalam memahami distribusi jumlah sampel pada setiap kelas sehingga hasil evaluasi dapat diinterpretasikan dengan lebih tepat (Shirdel et al., 2026).

Hasil evaluasi selanjutnya divisualisasikan menggunakan confusion matrix, yaitu matriks yang memperlihatkan jumlah prediksi benar maupun prediksi salah untuk setiap kelas. Melalui confusion matrix dapat diketahui jumlah True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) sehingga kesalahan klasifikasi yang dilakukan model dapat dianalisis secara lebih rinci (Sanchez-Marquez & Jabaloyes Vivas, 2026).

Untuk memperoleh estimasi performa model yang lebih stabil dan mengurangi ketergantungan terhadap satu kali pembagian data, penelitian ini juga menerapkan 5-fold cross validation. Pada metode ini, dataset dibagi menjadi lima subset, kemudian proses pelatihan dan pengujian dilakukan secara bergantian sehingga setiap subset berperan sebagai data uji sebanyak satu kali. Nilai performa akhir diperoleh dari rata-rata hasil evaluasi pada seluruh fold, sehingga memberikan gambaran yang lebih representatif mengenai kemampuan generalisasi model (Gorriz et al., 2026).

Selain mengevaluasi performa model, penelitian ini juga melakukan analisis feature importance pada model Random Forest dan XGBoost untuk mengetahui tingkat kontribusi masing-masing fitur dalam proses klasifikasi pilihan kelas layanan bus. Analisis ini membantu mengidentifikasi fitur-fitur yang paling berpengaruh terhadap keputusan model sekaligus meningkatkan interpretabilitas hasil penelitian (Iranzad & Liu, 2024; Takefuji, 2025).

## HASIL DAN PEMBAHASAN

### 3.1 Hasil Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) dilakukan untuk memahami karakteristik data sebelum proses pemodelan machine learning. Analisis ini bertujuan untuk mengidentifikasi distribusi kelas layanan, wilayah tujuan, waktu perjalanan relatif terhadap Hari Raya Idulfitri, serta status pembayaran penumpang.

#### 3.1.1 Distribusi Kelas Layanan

Tabel 1 menunjukkan distribusi kelas layanan yang digunakan sebagai variabel target pada penelitian ini.

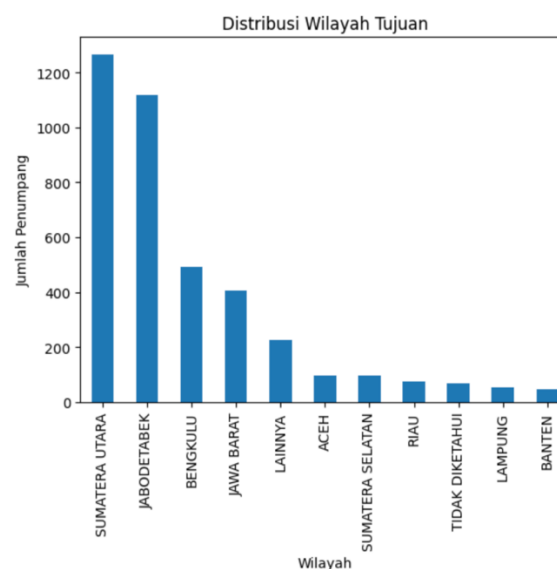
**Tabel 4.** Distribusi Kelas Layanan

| Kelas Layanan | Jumlah |
|---------------|--------|
| REGULER       | 2489   |
| PREMIUM       | 1453   |

**Sumber :** Hasil pengolahan data peneliti (2026)

Berdasarkan Tabel 1, mayoritas penumpang memilih kelas layanan REGULER dengan jumlah 2.489 data (63,14%), sedangkan kelas layanan PREMIUM berjumlah 1.453 data (36,86%). Hal ini menunjukkan bahwa layanan reguler masih menjadi pilihan utama penumpang selama periode mudik Lebaran.

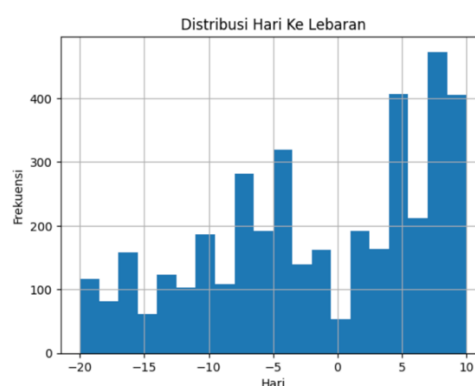
### 3.1.2 Distribusi Wilayah Tujuan



**Gambar 1.** Distribusi Wilayah Tujuan

Berdasarkan distribusi wilayah tujuan, sebagian besar perjalanan penumpang terkonsentrasi pada wilayah Sumatera Utara, Jabodetabek, dan wilayah lainnya. Hal ini menunjukkan bahwa mobilitas penumpang selama periode penelitian didominasi oleh rute-rute utama yang menghubungkan Pulau Sumatera dengan Pulau Jawa.

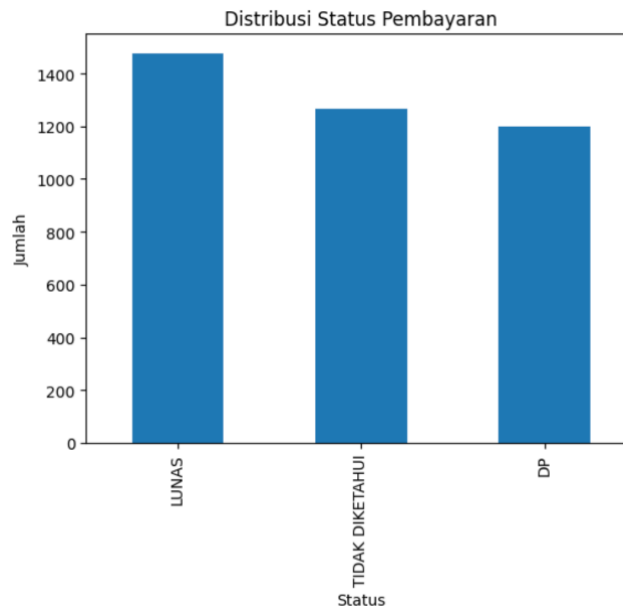
### 3.1.3 Distribusi Hari ke Lebaran



**Gambar 2.** Histogram Hari ke Lebaran

Histogram hari ke Lebaran menunjukkan persebaran waktu keberangkatan penumpang relatif terhadap Hari Raya Idulfitri. Terlihat adanya peningkatan jumlah keberangkatan pada periode mendekati Lebaran yang mengindikasikan tingginya aktivitas mudik pada waktu tersebut.

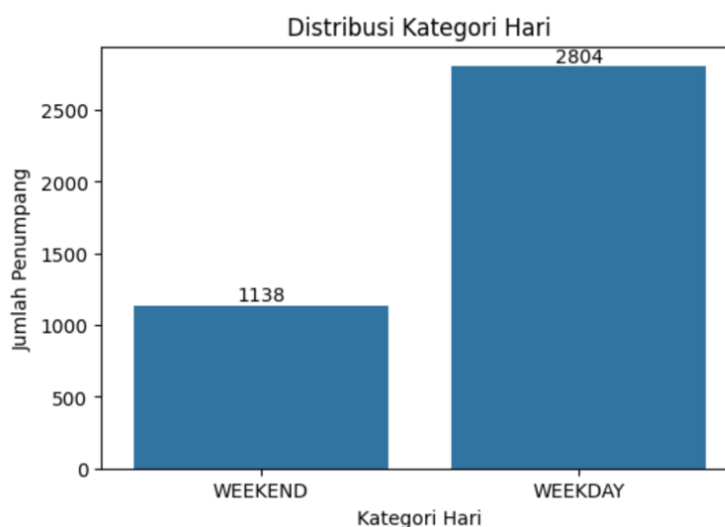
### 3.1.4 Distribusi Status Pembayaran



**Gambar 3.** Distribusi Status Pembayaran

Distribusi status pembayaran menunjukkan bahwa sebagian besar penumpang telah melakukan pelunasan tiket sebelum keberangkatan. Informasi ini menjadi salah satu variabel yang digunakan dalam proses klasifikasi karena dapat merepresentasikan perilaku pemesanan tiket penumpang.

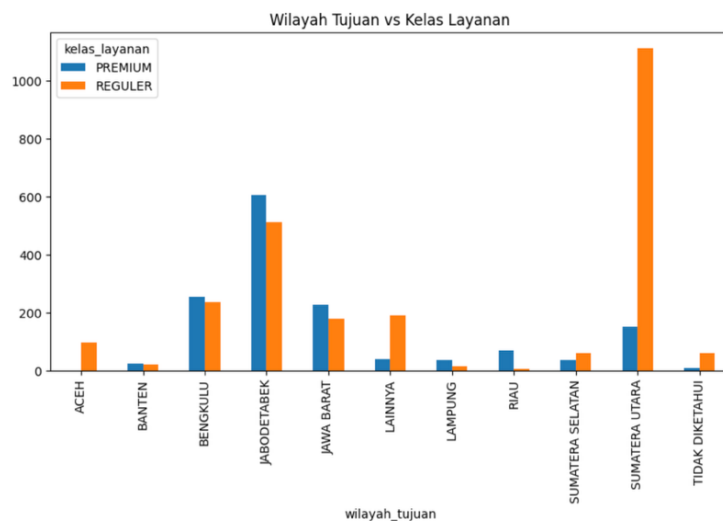
### 3.1.5 Distribusi Kategori Hari



**Gambar 4.** Distribusi Kategori hari

Distribusi kategori hari menunjukkan jumlah data yang termasuk ke dalam kategori weekday dan weekend. Pengelompokan ini dilakukan berdasarkan tanggal keberangkatan penumpang untuk mengetahui sebaran data pada hari kerja dan akhir pekan. Visualisasi ini memberikan gambaran mengenai proporsi perjalanan yang terjadi pada masing-masing kategori hari, yang selanjutnya digunakan sebagai salah satu fitur dalam proses klasifikasi preferensi kelas layanan bus.

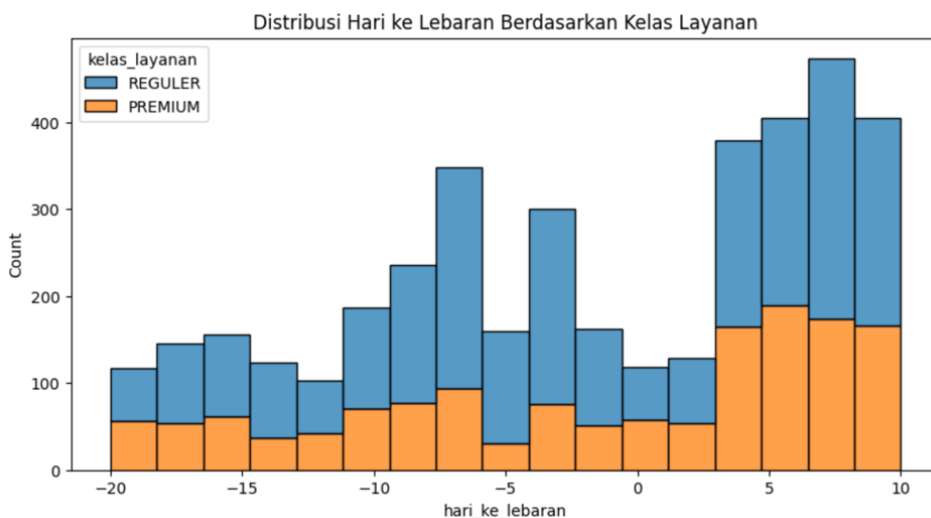
### 3.1.6 Distribusi Kelas Layanan Berdasarkan Wilayah Tujuan



**Gambar 5.** Distribusi Kelas Layanan Berdasarkan Wilayah Tujuan

Distribusi kelas layanan bus berdasarkan wilayah tujuan perjalanan menunjukkan bahwa jumlah penumpang pada setiap wilayah tujuan memiliki komposisi kelas layanan yang berbeda. Wilayah dengan jumlah penumpang terbesar juga cenderung memiliki jumlah pemilihan kelas layanan yang lebih tinggi dibandingkan wilayah lainnya. Perbedaan distribusi ini mengindikasikan bahwa karakteristik tujuan perjalanan dapat memengaruhi preferensi penumpang dalam memilih kelas layanan bus. Oleh karena itu, variabel wilayah tujuan diperkirakan memiliki kontribusi yang cukup besar dalam proses klasifikasi yang dilakukan oleh model machine learning.

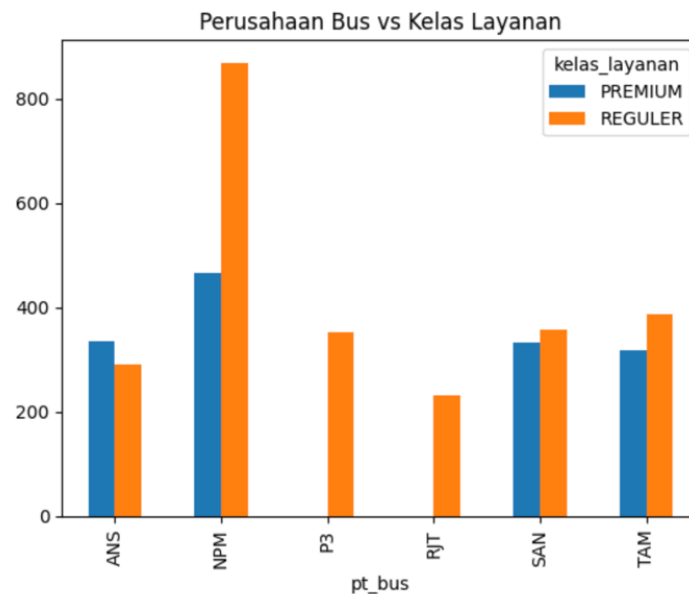
### 3.1.7 Distribusi Kelas Layanan Berdasarkan Hari ke Lebaran



**Gambar 6.** Distribusi Kelas Layanan Berdasarkan Hari ke Lebaran

Distribusi kelas layanan berdasarkan jumlah hari sebelum dan sesudah Hari Raya Idulfitri menunjukkan bahwa jumlah penumpang mengalami perubahan pada periode mendekati Hari Raya Idulfitri, yang menunjukkan adanya peningkatan aktivitas perjalanan selama musim mudik. Kondisi tersebut menyebabkan pola pemilihan kelas layanan juga berubah mengikuti waktu keberangkatan. Temuan ini menunjukkan bahwa variabel hari ke Lebaran berpotensi menjadi salah satu faktor penting dalam membedakan preferensi kelas layanan penumpang.

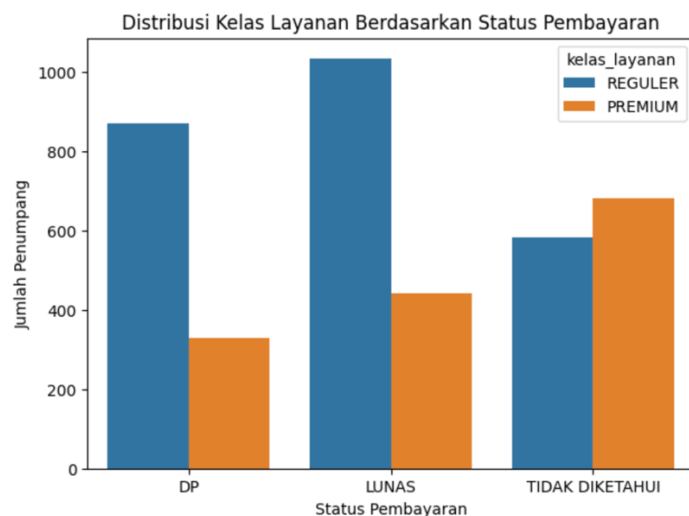
### 3.1.8 Distribusi Kelas Layanan Berdasarkan Perusahaan Bus



**Gambar 7.** Distribusi Kelas Layanan Berdasarkan Perusahaan Bus

Distribusi kelas layanan pada masing-masing perusahaan otobus yang menjadi objek penelitian menunjukkan bahwa setiap perusahaan memiliki jumlah penumpang dan komposisi kelas layanan yang berbeda. Perbedaan tersebut dapat dipengaruhi oleh variasi fasilitas, rute perjalanan, maupun strategi pelayanan yang diterapkan oleh masing-masing perusahaan. Dengan demikian, variabel perusahaan bus diperkirakan memiliki pengaruh terhadap preferensi penumpang dalam memilih kelas layanan tertentu.

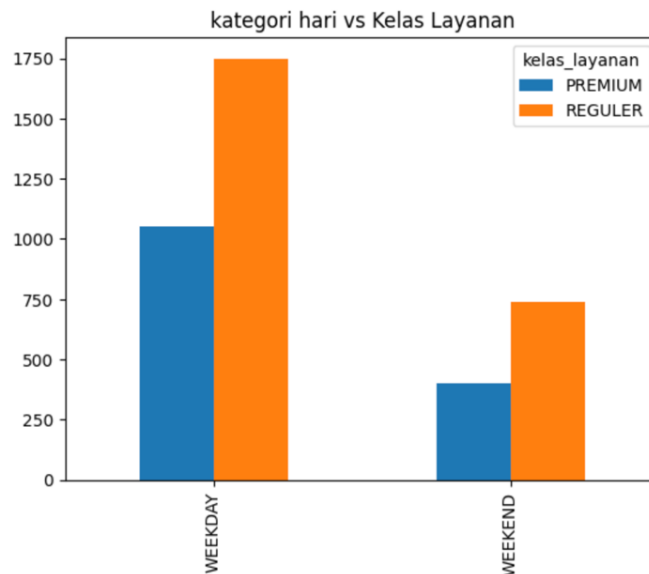
### 3.1.9 Distribusi Kelas Layanan Berdasarkan Status Pembayaran



**Gambar 8.** Distribusi Kelas Layanan Berdasarkan Status Pembayaran

Hasil visualisasi menunjukkan adanya perbedaan jumlah penumpang pada masing-masing kategori status pembayaran untuk setiap kelas layanan. Hal ini mengindikasikan bahwa status pembayaran memiliki hubungan dengan pola pemilihan kelas layanan bus. Oleh karena itu, variabel ini dipertimbangkan sebagai salah satu fitur yang digunakan dalam proses pembangunan model klasifikasi.

### 3.1.10 Distribusi Kelas Layanan Berdasarkan Kategori Hari (Weekday dan Weekend)

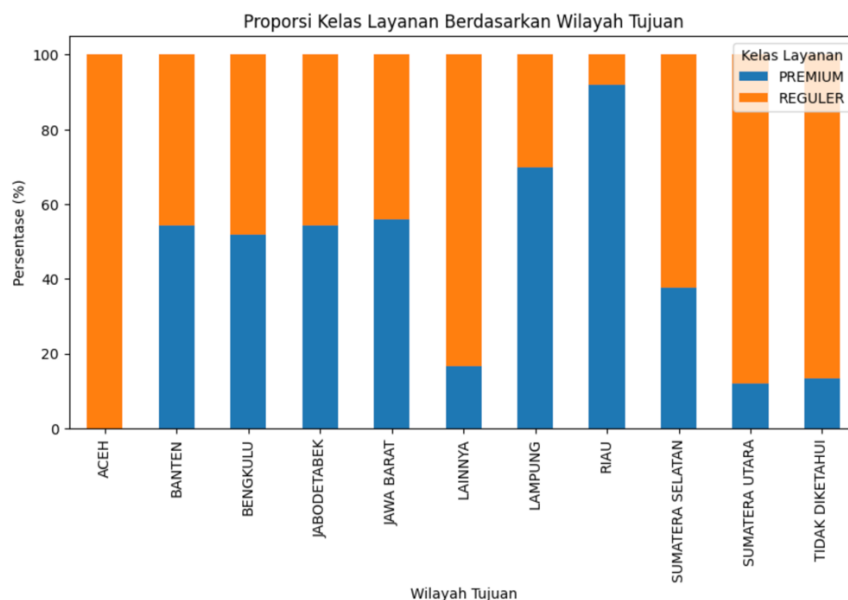


**Gambar 9.** Distribusi Kelas Layanan Berdasarkan Kategori Hari (Weekday dan Weekend)

Berdasarkan hasil visualisasi, jumlah keberangkatan penumpang lebih banyak terjadi pada kategori weekday dibandingkan weekend. Pada kedua kategori hari tersebut, kelas layanan reguler tetap mendominasi dibandingkan kelas premium, sehingga menunjukkan bahwa preferensi terhadap kelas reguler lebih tinggi baik pada hari kerja maupun akhir pekan.

Selain itu, pola distribusi antara weekday dan weekend terlihat relatif serupa, di mana tidak terdapat perbedaan proporsi kelas layanan yang terlalu mencolok. Hal ini mengindikasikan bahwa kategori hari bukan merupakan faktor utama yang membedakan preferensi penumpang dalam memilih kelas layanan bus. Temuan ini juga konsisten dengan hasil analisis feature importance, yang menunjukkan bahwa variabel weekend memiliki tingkat kepentingan paling rendah dibandingkan fitur lainnya dalam proses klasifikasi.

### 3.1.11 Proporsi Kelas Layanan Berdasarkan Wilayah Tujuan



**Gambar 10.** Distriporsi Kelas Layanan Berdasarkan Wilayah Tujuan

Berbeda dengan grafik distribusi sebelumnya yang menampilkan jumlah absolut, visualisasi ini menunjukkan persentase masing-masing kelas layanan sehingga memudahkan dalam

membandingkan kecenderungan pemilihan kelas layanan pada setiap wilayah. Dari hasil tersebut terlihat bahwa beberapa wilayah memiliki proporsi penumpang kelas PREMIUM yang lebih tinggi dibandingkan wilayah lainnya, sedangkan wilayah lain masih didominasi oleh kelas REGULER. Perbedaan proporsi ini memperkuat indikasi bahwa wilayah tujuan merupakan salah satu faktor yang berpengaruh terhadap preferensi kelas layanan penumpang.

### 3.2 Hasil Pemodelan Random Forest

Model Random Forest dibangun menggunakan data latih dan diuji menggunakan data uji untuk mengklasifikasikan preferensi kelas layanan bus.

**Tabel 5.** Nilai Accuracy Model Random Forest

| Model         | Accuracy |
|---------------|----------|
| Random Forest | 0.7921   |

**Sumber :** Hasil pengolahan data peneliti (2026).

Berdasarkan Tabel 5, model Random Forest memperoleh nilai accuracy sebesar 0.7921. Nilai tersebut menunjukkan bahwa model mampu mengklasifikasikan preferensi kelas layanan bus dengan tingkat ketepatan yang cukup baik, sehingga model dinilai mampu mengenali pola pada data yang digunakan dalam penelitian ini.

**Tabel 6.** Classification Report Random Forest

| Kelas   | Precision | Recall   | F1-Score | Support |
|---------|-----------|----------|----------|---------|
| Reguler | 0.736059  | 0.680412 | 0.707143 | 291.0   |
| Premium | 0.821154  | 0.857430 | 0.838900 | 498.0   |

**Sumber :** Hasil pengolahan data peneliti (2026).

Model random forest menunjukkan performa yang lebih baik dalam mengklasifikasikan kelas premium dibandingkan kelas reguler. Hal tersebut terlihat dari nilai precision, recall, dan f1-score kelas premium yang masing-masing mencapai 0,82, 0,86, dan 0,84, lebih tinggi dibandingkan kelas reguler yang memperoleh nilai 0,74, 0,68, dan 0,71. Selain itu, nilai support menunjukkan bahwa data uji terdiri atas 291 data reguler dan 498 data premium, sehingga distribusi kelas pada proses pengujian masih didominasi oleh kelas premium.

**Tabel 7.** Confussion Matrix Random Forest

| Actual/Predicted | Predicted<br>Reguler | Predicted<br>Premium |
|------------------|----------------------|----------------------|
| Actual Reguler   | 198                  | 93                   |
| Actual Premium   | 71                   | 427                  |

**Sumber :** Hasil pengolahan data peneliti (2026).

Berdasarkan tabel 7, model random forest berhasil mengklasifikasikan 198 data reguler dan 427 data premium dengan benar. Namun, masih terdapat 93 data reguler yang diprediksi sebagai premium (false positive) serta 71 data premium yang diprediksi sebagai reguler (false negative). Jumlah kesalahan klasifikasi tersebut menunjukkan bahwa meskipun model memiliki kemampuan prediksi yang baik, masih terdapat beberapa karakteristik data yang menyebabkan kedua kelas layanan memiliki pola yang saling beririsan sehingga sulit dibedakan secara sempurna.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa algoritma Random Forest memiliki performa yang baik dalam mengklasifikasikan preferensi kelas layanan bus antar kota selama periode mudik Lebaran. Nilai accuracy sebesar 0.7921 menunjukkan bahwa model mampu memberikan prediksi yang cukup akurat terhadap data uji. Selain itu, model memiliki kemampuan yang lebih baik dalam mengenali kelas PREMIUM dibandingkan kelas REGULER, sebagaimana ditunjukkan oleh nilai precision, recall, dan F1-score yang lebih tinggi pada kelas tersebut.

### 3.3 Hasil Pemodelan XGBoost

Model XGBoost digunakan sebagai model pembandingan untuk mengetahui algoritma yang memberikan performa terbaik pada dataset penelitian.

**Tabel 8.** Nilai Accuracy Model XGBoost

| Model         | Accuracy |
|---------------|----------|
| Random Forest | 0.7896   |

**Sumber :** Hasil pengolahan data peneliti (2026).

Berdasarkan Tabel 8, model XGBoost memperoleh nilai accuracy sebesar 0,7896. Nilai tersebut menunjukkan bahwa model mampu mengklasifikasikan preferensi kelas layanan bus dengan tingkat ketepatan yang baik. Hasil ini menunjukkan bahwa algoritma XGBoost memiliki kemampuan yang baik dalam mempelajari pola data dan menghasilkan prediksi yang cukup akurat pada kasus klasifikasi preferensi kelas layanan bus.

**Tabel 9.** Classification Report XGBoost

| Kelas   | Precision | Recall   | F1-Score | Support |
|---------|-----------|----------|----------|---------|
| Reguler | 0.730627  | 0.680412 | 0.704626 | 291     |
| Premium | 0.820463  | 0.853414 | 0.836614 | 498     |

**Sumber :** Hasil pengolahan data peneliti (2026).

Berdasarkan Tabel 9, model XGBoost menunjukkan performa yang lebih baik dalam mengklasifikasikan kelas PREMIUM dibandingkan kelas REGULER. Hal tersebut ditunjukkan oleh nilai precision sebesar 0,82, recall sebesar 0,85, dan F1-score sebesar 0,84 pada kelas PREMIUM. Sementara itu, kelas REGULER memperoleh nilai precision sebesar 0,73, recall sebesar 0,68, dan F1-score sebesar 0,70. Nilai support menunjukkan bahwa data uji terdiri atas 291 data REGULER dan 498 data PREMIUM, sehingga distribusi data pengujian masih didominasi oleh kelas PREMIUM.

**Tabel 10.** Confussion Matrix XGBoost

| Actual/Predicted | Predicted<br>Reguler | Predicted<br>Premium |
|------------------|----------------------|----------------------|
| Actual Reguler   | 198                  | 93                   |
| Actual Premium   | 73                   | 425                  |

**Sumber :** Hasil pengolahan data peneliti (2026).

Berdasarkan Tabel 10, model XGBoost berhasil mengklasifikasikan 198 data REGULER dan 425 data PREMIUM dengan benar. Namun demikian, masih terdapat 93 data REGULER yang diprediksi sebagai PREMIUM (False Positive) serta 73 data PREMIUM yang diprediksi sebagai REGULER (False Negative). Hasil tersebut menunjukkan bahwa model masih mengalami beberapa kesalahan klasifikasi, terutama pada data yang memiliki karakteristik serupa antara kedua kelas layanan, sehingga menyebabkan model belum mampu membedakan seluruh data secara sempurna.

Secara keseluruhan, algoritma XGBoost menunjukkan performa yang baik dalam mengklasifikasikan preferensi kelas layanan bus antar kota pada periode mudik Lebaran dengan nilai accuracy sebesar 0.7896. Nilai tersebut menunjukkan bahwa sebagian besar data uji berhasil diprediksi dengan benar. Selain itu, model juga memiliki kemampuan yang lebih baik dalam mengenali kelas PREMIUM dibandingkan kelas REGULER, yang terlihat dari nilai precision, recall, dan F1-score yang lebih tinggi pada kelas PREMIUM.

### 3.4 Analisis Stabilitas Model Menggunakan Cross Validation

**Tabel 11.** Accuracy Setiap Fold Hasil Cross Validation

| Fold   | Accuracy |
|--------|----------|
| Fold 1 | 0.806084 |
| Fold 2 | 0.794677 |

|        |          |
|--------|----------|
| Fold 3 | 0.789340 |
| Fold 4 | 0.804569 |
| Fold 5 | 0.769036 |

**Sumber :** Hasil pengolahan data peneliti (2026).

**Tabel 12.** Ringkasan Hasil Cross Validation (Mean Accuracy dan Standard Deviation)

| Parameter          | Accuracy |
|--------------------|----------|
| Mean Accuracy      | 0.792741 |
| Standard Deviation | 0.013380 |

**Sumber :** Hasil pengolahan data peneliti (2026).

Berdasarkan hasil 5-Fold Cross Validation, model Random Forest memperoleh nilai accuracy yang relatif konsisten pada setiap fold, yaitu 0,8061, 0,7947, 0,7893, 0,8046, dan 0,7690. Nilai tersebut menunjukkan bahwa performa model tidak mengalami perbedaan yang signifikan meskipun data dibagi ke dalam lima subset yang berbeda.

Ringkasan hasil cross validation menunjukkan bahwa model memperoleh rata-rata accuracy sebesar 0,7927 dengan standar deviasi sebesar 0,0134. Nilai standar deviasi yang relatif kecil menunjukkan bahwa variasi performa model antar fold tergolong rendah, sehingga model memiliki tingkat stabilitas yang baik dalam melakukan klasifikasi pada berbagai pembagian data.

Selain itu, nilai rata-rata accuracy hasil cross validation (0,7927) sangat mendekati nilai accuracy pada data uji (0,7921). Kesamaan hasil tersebut menunjukkan bahwa model Random Forest memiliki kemampuan generalisasi yang baik dan tidak mengalami indikasi overfitting yang berarti. Dengan demikian, model dinilai mampu memberikan performa yang konsisten dalam mengklasifikasikan preferensi kelas layanan bus antar kota selama periode mudik Lebaran.

### 3.5 Perbandingan Kinerja Model

Perbandingan performa kedua algoritma dilakukan menggunakan nilai accuracy sebagai indikator utama.

**Tabel 12.** Perbandingan Kinerja Model

| Model         | Accuracy |
|---------------|----------|
| Random Forest | 0.7921   |
| XGBoost       | 0.7896   |

**Sumber :** Hasil pengolahan data peneliti (2026).

Berdasarkan Tabel 12, algoritma Random Forest memperoleh nilai accuracy sebesar 0,7921, sedangkan XGBoost memperoleh nilai accuracy sebesar 0,7896. Perbedaan nilai accuracy antara kedua model hanya sebesar 0,0025, sehingga keduanya menunjukkan performa klasifikasi yang relatif sebanding pada dataset penelitian ini.

Meskipun Random Forest memperoleh nilai accuracy yang sedikit lebih tinggi, selisih tersebut belum cukup untuk menyimpulkan bahwa Random Forest secara signifikan lebih unggul dibandingkan XGBoost. Oleh karena itu, evaluasi model tidak hanya mempertimbangkan nilai accuracy, tetapi juga memperhatikan hasil cross validation yang menunjukkan bahwa Random Forest memiliki rata-rata accuracy sebesar 0,7927 dengan standar deviasi sebesar 0,0134. Nilai tersebut mengindikasikan bahwa performa model relatif konsisten pada berbagai pembagian data.

Berdasarkan keseluruhan hasil evaluasi, baik Random Forest maupun XGBoost mampu memberikan performa klasifikasi yang baik dalam memprediksi preferensi kelas layanan bus antar kota selama periode mudik Lebaran. Dalam penelitian ini, Random Forest dipilih sebagai model yang digunakan karena memberikan nilai accuracy yang sedikit lebih tinggi serta menunjukkan performa yang stabil pada proses cross validation.

### 3.6 Analisis Feature Importance

Analisis feature importance dilakukan untuk mengetahui faktor-faktor yang paling memengaruhi preferensi kelas layanan penumpang.

**Tabel 13.** Feature Importance

| Fitur             | Importance |
|-------------------|------------|
| Wilayah Tujuan    | 0.352430   |
| Hari ke Lebaran   | 0.280165   |
| PT Bus            | 0.212963   |
| Status Pembayaran | 0.136825   |
| Weekend           | 0.017617   |

**Sumber :** Hasil pengolahan data peneliti (2026).

Berdasarkan Tabel 13, wilayah tujuan merupakan fitur dengan nilai feature importance tertinggi, yaitu sebesar 0.3524. Hal ini menunjukkan bahwa lokasi tujuan perjalanan memberikan kontribusi paling besar dalam proses klasifikasi preferensi kelas layanan bus. Perbedaan wilayah tujuan umumnya berkaitan dengan variasi jarak tempuh, durasi perjalanan, serta kebutuhan kenyamanan selama perjalanan. Penumpang yang melakukan perjalanan dengan rute yang lebih jauh cenderung memiliki preferensi yang lebih tinggi terhadap kelas layanan PREMIUM karena menawarkan fasilitas dan tingkat kenyamanan yang lebih baik dibandingkan kelas REGULER.

Fitur hari\_ke\_lebaran memiliki nilai importance sebesar 0.2802, sehingga menjadi faktor kedua yang paling berpengaruh. Semakin mendekati Hari Raya Idulfitri, permintaan perjalanan biasanya meningkat sehingga memengaruhi ketersediaan kursi, pola pemesanan, dan pilihan kelas layanan oleh penumpang. Kondisi tersebut menyebabkan waktu keberangkatan relatif terhadap Hari Raya menjadi salah satu variabel penting dalam proses klasifikasi.

Selanjutnya, fitur pt\_bus memperoleh nilai importance sebesar 0.2130. Hal ini menunjukkan bahwa perusahaan otobus juga memengaruhi preferensi kelas layanan karena masing-masing perusahaan memiliki karakteristik armada, fasilitas, kualitas pelayanan, serta segmentasi pelanggan yang berbeda. Perbedaan tersebut memungkinkan penumpang memiliki kecenderungan memilih kelas layanan tertentu sesuai dengan perusahaan bus yang digunakan.

Fitur status\_pembayaran memiliki nilai importance sebesar 0.1368, yang menunjukkan bahwa pola pembayaran juga memberikan kontribusi terhadap proses klasifikasi. Status pembayaran dapat mencerminkan perilaku pemesanan penumpang, misalnya penumpang yang melakukan pembayaran lebih awal memiliki peluang lebih besar memperoleh kelas layanan yang diinginkan dibandingkan penumpang yang melakukan pemesanan mendekati tanggal keberangkatan.

Sementara itu, fitur weekend memiliki nilai importance paling kecil, yaitu sebesar 0.0176. Nilai tersebut menunjukkan bahwa kategori akhir pekan maupun hari kerja memberikan pengaruh yang relatif rendah terhadap preferensi kelas layanan dibandingkan fitur lainnya. Hal ini mengindikasikan bahwa selama periode mudik Lebaran, keputusan penumpang dalam memilih kelas layanan lebih banyak dipengaruhi oleh tujuan perjalanan, waktu keberangkatan relatif terhadap Hari Raya, perusahaan bus, dan pola pembayaran dibandingkan kategori hari.

## KESIMPULAN

Penelitian ini berhasil menerapkan algoritma Random Forest dan XGBoost untuk mengklasifikasikan preferensi kelas layanan bus antar kota pada periode mudik Lebaran tahun 2026. Dataset yang digunakan terdiri dari 3.942 data penumpang yang berasal dari enam perusahaan otobus dan telah melalui tahapan data wrangling, feature engineering, validasi data, serta pemodelan machine learning. Hasil pengujian menunjukkan bahwa kedua algoritma mampu memberikan performa klasifikasi yang baik, dengan nilai accuracy sebesar 0.7921 pada Random Forest dan 0.7896 pada XGBoost. Perbedaan performa kedua model relatif kecil, sehingga keduanya dapat dianggap memiliki kemampuan klasifikasi yang sebanding pada dataset penelitian ini. Namun demikian, Random Forest menunjukkan performa yang sedikit lebih tinggi serta hasil cross validation yang konsisten, sehingga dipilih sebagai model yang digunakan dalam penelitian ini.

Analisis feature importance menunjukkan bahwa wilayah tujuan merupakan faktor yang paling berpengaruh terhadap preferensi kelas layanan penumpang, diikuti oleh hari ke Lebaran,

perusahaan otobus, dan status pembayaran. Temuan ini menunjukkan bahwa karakteristik perjalanan, khususnya tujuan dan waktu keberangkatan relatif terhadap Hari Raya Idulfitri, memiliki peranan penting dalam menentukan pilihan kelas layanan. Meskipun demikian, penelitian ini masih memiliki keterbatasan, yaitu menggunakan data yang hanya mencakup periode 1–31 Maret 2026 dan melibatkan lima variabel prediktor utama, sehingga hasil yang diperoleh belum tentu dapat digeneralisasikan pada periode atau kondisi operasional yang berbeda. Oleh karena itu, penelitian selanjutnya disarankan menggunakan data dengan rentang waktu yang lebih panjang, menambahkan variabel yang lebih beragam, seperti tarif, durasi perjalanan, atau fasilitas bus, serta menerapkan teknik optimasi model dan pengujian statistik untuk memperoleh hasil klasifikasi yang lebih komprehensif.

### Ucapan Terimakasih

Penulis mengucapkan terima kasih kepada Program Studi Informatika Universitas Islam Negeri (UIN) Sjech M. Djamil Djambek Bukittinggi yang telah memberikan dukungan dalam pelaksanaan penelitian ini. Penulis juga mengucapkan terima kasih kepada seluruh pihak yang telah membantu dalam proses pengumpulan data, pengolahan data, serta penyusunan penelitian hingga dapat diselesaikan dengan baik. Penulis menyadari bahwa penelitian ini masih memiliki keterbatasan. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan untuk pengembangan penelitian selanjutnya.

### REFERENCES

- Ajeng, D., Febriyanti, A., Mustofa, A., & Fatah, Z. (2023). Kualitas Pelayanan Publik Pada Bus Trans Jatim. *Soetomo Administrasi Publik*, 381–394. <https://ejournal.unitomo.ac.id/index.php/sap/article/view/7240>
- Alexandre, T., Bernardini, F., Viterbo, J., & Pantoja, C. E. (2023). Machine Learning Applied to Public Transportation by Bus: A Systematic Literature Review. *Transportation Research Record*, 2677(7), 639–660. <https://doi.org/10.1177/03611981231155189;WGROU:STRING:PUBLICATION>
- Bala, B., & Behal, S. (2024). A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques. *8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2024 - Proceedings*, 1755–1762. <https://doi.org/10.1109/I-SMAC61858.2024.10714767>
- Buran, B. (2025). Passenger satisfaction modeling in public bus transportation based on business model approach: Ten city case studies. *Case Studies on Transport Policy*, 21, 101472. <https://doi.org/10.1016/J.CSTP.2025.101472>
- Daneshvar, A., Salahi, F., Ebrahimi, M., & Nahavandi, B. (2023). Analyzing Behavioral Patterns of Bus Passengers Using Data Mining Methods (Case Study: Rapid Transportation Systems). *Journal of Applied Research on Industrial Engineering*, 10(1), 11–24. <https://doi.org/10.22105/IARIE.2021.302560.1373>
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data* 2021 8:1, 8(1), 140-. <https://doi.org/10.1186/S40537-021-00516-9>
- Fadhilah, M. W. (Maulida), & Amalia, S. (Sholihati). (2021). Pengaruh Kualitas Pelayanan Bus Kota DAMRI terhadap Kepuasan Pelanggan (Studi pada Penumpang Bus Kota DAMRI Bandung). *Jurnal Riset Bisnis Dan Investasi*, 7(3), 150–162. <https://doi.org/10.35313/JRBI.V7I2.3254>
- Gorriz, J. M., Clemente, R. M., Segovia, F., Ramirez, J., Ortiz, A., & Suckling, J. (2026). Is K-fold cross validation the best model selection method for machine learning? *Information Fusion*, 135, 104404. <https://doi.org/10.1016/J.INFFUS.2026.104404>
- Halim, M. R. (2025, May 12). *Hadapi Arus Mudik 2025, Kemenhub Siapkan Langkah Strategis untuk 146 Juta Pemudik*. Badan Kebijakan Transportasi Kementerian Perhubungan. <https://baketrans.kemenhub.go.id/publikasi/view/hadapi-arus-mudik-2025-kemenhub-siapkan-langkah-strategis-untuk-146-juta-pemudik>

- Hillel, T., Bierlaire, M., Elshafie, M. Z. E. B., & Jin, Y. (2021). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*, 38, 100221. <https://doi.org/10.1016/J.JOCM.2020.100221>
- Hodeghatta, U. R., & Nayak, U. (2023). Exploratory Data Analysis. *Practical Business Analytics Using R and Python*, 139–165. [https://doi.org/10.1007/978-1-4842-8754-5\\_5](https://doi.org/10.1007/978-1-4842-8754-5_5)
- Iranzad, R., & Liu, X. (2024). A review of random forest-based feature selection methods for data science education and applications. *International Journal of Data Science and Analytics* 2024 20:2, 20(2), 197–211. <https://doi.org/10.1007/S41060-024-00509-W>
- Lan, B., Chen, Y., & Wu, K. (2025). A granular XGBoost classification algorithm. *Applied Intelligence* 2025 55:13, 55(13), 895-. <https://doi.org/10.1007/S10489-025-06762-1>
- Lan, B., Wu, K., Chen, Y., & Li, J. (2026). PCA-GRF: A novel granular random forest algorithm for enhancing classification of small sample data. *Information Sciences*, 746, 123446. <https://doi.org/10.1016/J.INS.2026.123446>
- Li, P., Wu, W., & Pei, X. (2023). A separate modelling approach for short-term bus passenger flow prediction based on behavioural patterns: A hybrid decision tree method. *Physica A: Statistical Mechanics and Its Applications*, 616, 128567. <https://doi.org/10.1016/J.PHYSA.2023.128567>
- MAULANA, H. (2024). PEMILIHAN MODA TRANSPORTASI MUDIK LEBARAN BAGI MASYARAKAT KABUPATEN LEBAK DENGAN MENGGUNAKAN METODE ANALYTIC HIERARCHY PROCESS. <http://eprints.untirta.ac.id/id/eprint/34705>
- Rahmawati, P., Estriana, R., Praptiwi, R., Salsabila, B., Rafa Oktafian, N., & Zakiyudin, T. (2025). Analisis Perbandingan XGBoost, Random Forest, dan LSTM untuk Prediksi Jumlah Penumpang Bus. *Journal of Informatics and Interactive Technology*, 2(3), 405–413. <https://doi.org/10.63547/JIITE.V2I3.73>
- Rianti, A., Wachid, N., Majid, A., Fauzi, A., Sistem, P., Informasi, T., Telekomunikasi, S., Upi Di Purwakarta, K., & Edu, A. (n.d.). CRISP-DM: Metodologi Proyek Data Science.
- Ruiz, E., Yushimito, W. F., Aburto, L., & de la Cruz, R. (2024). Predicting passenger satisfaction in public transportation using machine learning models. *Transportation Research Part A: Policy and Practice*, 181, 103995. <https://doi.org/10.1016/J.TRA.2024.103995>
- Sanchez-Marquez, R., & Jabaloyes Vivas, J. (2026). A statistical approach to the confusion matrix for classification problems using machine learning. *Neurocomputing*, 695, 133962. <https://doi.org/10.1016/J.NEUCOM.2026.133962>
- Shirdel, M., Di Mauro, M., & Liotta, A. (2026). A framework for binary classification evaluation metrics. *Information Sciences*, 749, 123540. <https://doi.org/10.1016/J.INS.2026.123540>
- Stadler, T., Sarkar, A., & Dünnweber, J. (2021). Bus Demand Forecasting for Rural Areas Using XGBoost and Random Forest Algorithm. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12883 LNCS, 442–453. [https://doi.org/10.1007/978-3-030-84340-3\\_36/SAVE-RESEARCH](https://doi.org/10.1007/978-3-030-84340-3_36/SAVE-RESEARCH)
- Takefuji, Y. (2025). Beyond XGBoost and SHAP: Unveiling true feature importance. *Journal of Hazardous Materials*, 488, 137382. <https://doi.org/10.1016/J.IHAZMAT.2025.137382>
- Tim Survei Angleb Badan Kebijakan Transportasi. (2026, March 6). 143,9 JUTA ORANG AKAN LAKUKAN PERJALANAN PADA ANGKUTAN LEBARAN TAHUN 2026. Badan Kebijakan Transportasi Kementerian Perhubungan. <https://baketrans.kemenhub.go.id/publikasi/view/143-9-juta-orang-akan-lakukan-perjalanan-pada-angkutan-lebaran-tahun-2026>
- Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2021). Special issue on feature engineering editorial. *Machine Learning* 2021 113:7, 113(7), 3917–3928. <https://doi.org/10.1007/S10994-021-06042-2>
- Wang, L., Wang, Y., Sun, X., Wu, Y., Peng, F., Chen, C.-H. P., & Song, G. (2023). Public Transit Passenger Profiling by Using Large-Scale Smart Card Data. *Journal of Transportation Engineering, Part A:*

*Systems*, 149(4), 04023013. <https://doi.org/10.1061/JTEPBS.TEENG-7634;WGROUP:STRING:PUBLICATION>

Zhang, H., Zhang, L., Liu, Y., & Zhang, L. (2023). Understanding Travel Mode Choice Behavior: Influencing Factors Analysis and Prediction with Machine Learning Method. *Sustainability* 2023, Vol. 15, Page 11414, 15(14), 11414. <https://doi.org/10.3390/SU151411414>