

## MODEL RANDOM FOREST PREDIKSI KUNJUNGAN ULANG BERDASARKAN POLA PENGELUARAN

Yannanta Laga Kusuma\*, Joko Handoyo

Program Studi Informatika, Sekolah Tinggi Teknologi Ronggolawe Cepu.

\*[yannantalagakusuma@gmail.com](mailto:yannantalagakusuma@gmail.com)

Jl. Kampus Ronggolawe Blok B No 1 Mentul, Blora, Jawa Tengah, Indonesia.

*Article history:* Received: 15 March 2026; Revised: 25 May 2026; Accepted: 30 June 2026

### Abstract

Noyo Gimbal View Blora tourism has experienced a significant decline in visits, from 500,552 (2024) to a projected 300,000 (2025), with revenue dropping from IDR 3 billion to IDR 1.5 billion. This research aims to build a predictive Model to identify potential tourist disloyalty using the Random Forest algorithm based on spending patterns and service satisfaction. The research method utilizes 3,000 visit records (June 2023–December 2025) from ticket transactions and Google Maps reviews. The target variable is loyalty status derived from revisit history. Modeling employs Random Forest with Hyperparameter optimization through GridSearchCV and 5-fold cross-validation, along with Model interpretation using feature importance and SHAP. Results show the Model achieves 87.3% Accuracy, 83.5% Precision, 78.2% Recall, 80.8% F1-Score, and 0.91 AUC. The three most influential features are review sentiment score (0.26), spending variation (0.20), and number of facility complaints (0.17). Characteristics of at-risk tourists include negative sentiment, at least 2 complaints, rating  $\leq 2$  stars, spending  $<$  IDR 20,000, high spending variation, and high Recency days. In conclusion, service satisfaction factors are more dominant than spending patterns in determining tourist loyalty. This Model can be implemented as an early warning system to design intervention strategies for increasing tourist retention.

**Keywords:** Churn Prediction; Data Mining; Expenditure Patterns; Noyo Gimbal View; Random Forest; Tourist Loyalty.

### Abstrak

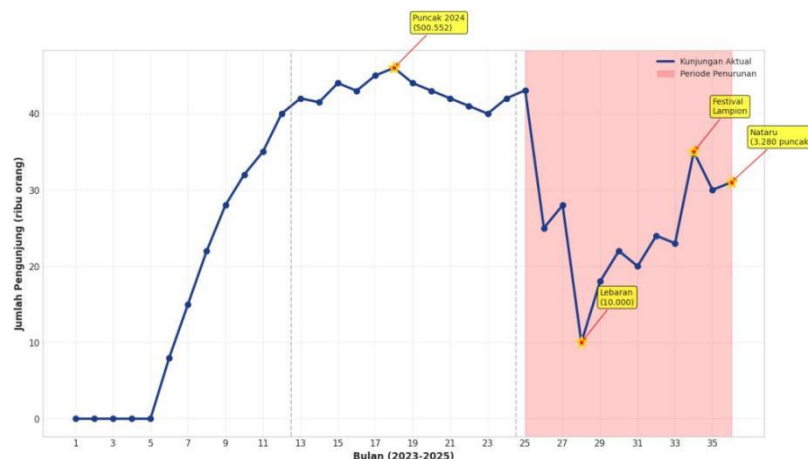
Wisata Noyo Gimbal View Blora mengalami penurunan kunjungan signifikan dari 500.552 (2024) menjadi diproyeksikan 300.000 (2025), dengan pendapatan turun dari Rp3 miliar menjadi Rp1,5 miliar. Penelitian ini bertujuan membangun Model prediktif untuk mengidentifikasi potensi ketidakloyalan wisatawan menggunakan algoritma *Random Forest* berbasis pola pengeluaran dan kepuasan layanan. Metode penelitian menggunakan 3.000 data rekam kunjungan (Juni 2023–Desember 2025) dari transaksi tiket dan ulasan *Google Maps*. Variabel target adalah status loyalitas yang diturunkan dari riwayat kunjungan ulang. Pemodelan menggunakan *Random Forest* dengan optimasi *Hyperparameter* melalui *GridSearchCV* dan *5-fold cross-validation*, serta interpretasi Model menggunakan *feature importance* dan *SHAP*. Hasil penelitian menunjukkan Model mencapai akurasi 87,3%, presisi 83,5%, *Recall* 78,2%, *F1-Score* 80,8%, dan *AUC* 0,91. Tiga fitur paling berpengaruh adalah skor sentimen ulasan (0,26), variasi pengeluaran (0,20), dan jumlah keluhan fasilitas (0,17). Karakteristik wisatawan berisiko ditandai dengan sentimen negatif, minimal 2 keluhan, rating  $\leq 2$  bintang, pengeluaran  $<$  Rp20.000, variasi pengeluaran tinggi, dan *Recency* hari tinggi. Kesimpulannya, faktor kepuasan layanan lebih dominan daripada pola pengeluaran dalam menentukan loyalitas wisatawan. Model ini dapat diimplementasikan sebagai sistem peringatan dini untuk merancang strategi intervensi meningkatkan retensi wisatawan.

**Kata Kunci:** Data Mining; Loyalitas Wisatawan; Noyo Gimbal View; Pola Pengeluaran; Prediksi Churn; Random Forest.



## Pendahuluan

Pariwisata berbasis desa di Indonesia mengalami perkembangan yang signifikan dalam dekade terakhir. Namun, keberhasilan awal dalam menarik pengunjung tidak selalu menjamin keberlanjutan destinasi dalam jangka panjang, seperti yang terlihat pada fluktuasi kunjungan di berbagai objek wisata [1]. Noyo Gimbal View, sebuah destinasi wisata yang dikelola oleh BUMDes Bangsri di Kabupaten Blora, menjadi studi kasus yang menarik karena mengalami fluktuasi kunjungan yang tajam sejak mulai beroperasi pada Juni 2023 [2], [3]. Fluktuasi kunjungan di Noyo Gimbal View tergambar jelas pada Gambar 1. Grafik *line* tersebut menunjukkan tren jumlah pengunjung per bulan dengan puncak kunjungan terjadi sepanjang tahun 2024, kemudian mengalami penurunan drastis mulai awal tahun 2025.



Gambar 1. Grafik Tren Kunjungan Noyo Gimbal View  
Periode Juni 2023 – Desember 2025

Gambar 1 menyajikan grafik tren kunjungan Noyo Gimbal View selama periode Juni 2023 hingga Desember 2025. Grafik berbentuk *Line chart* ini menggambarkan fluktuasi jumlah pengunjung setiap bulannya dengan sangat jelas. Terlihat bahwa puncak kunjungan terjadi sepanjang tahun 2024 di mana destinasi wisata ini mencapai masa kejayaannya dengan jumlah pengunjung yang sangat tinggi. Namun, grafik juga menunjukkan penurunan yang sangat drastis mulai awal tahun 2025, mengonfirmasi adanya masalah serius dalam retensi wisatawan [4]. Visualisasi data ini menjadi fondasi penting bagi penelitian karena membuktikan secara empiris bahwa terjadi fenomena penurunan kunjungan yang signifikan, yang kemudian memotivasi dilakukannya penelitian prediksi minat kunjungan ulang menggunakan algoritma *Random Forest*.

Data resmi dari pengelola menunjukkan bahwa hingga akhir tahun 2025, jumlah kunjungan diproyeksikan mencapai sekitar 300.000, hampir 40 persen lebih rendah dibandingkan tahun 2024 yang mencapai 500.552. Penurunan drastis ini secara langsung berdampak pada pendapatan yang anjlok dari lebih Rp3 miliar menjadi sekitar Rp1,5 miliar. Pengelola, Kasidin, mengidentifikasi musim hujan yang panjang sebagai faktor utama penyebab kemerosotan ini. Dampak lanjutan dari penurunan jumlah wisatawan ini terasa pada sektor ketenagakerjaan, di mana jumlah tenaga kerja berkurang dari 90 orang menjadi hanya 40–50 orang. Berdasarkan kondisi tersebut, penelitian ini merancang tiga rumusan masalah utama. Pertama, bagaimana karakteristik pola pengeluaran dan tingkat kepuasan wisatawan yang berpotensi tidak kembali berkunjung ke Noyo Gimbal View. Kedua, seberapa efektif algoritma *Random Forest* dalam memprediksi minat kunjungan ulang wisatawan. Ketiga, fitur apa yang paling berpengaruh dalam model prediksi minat kunjungan ulang tersebut. Sejalan dengan rumusan masalah, tujuan dari penelitian ini adalah untuk mengidentifikasi karakteristik pola pengeluaran dan tingkat kepuasan wisatawan yang berisiko tidak kembali. Selanjutnya, penelitian ini bertujuan untuk mengukur efektivitas algoritma *Random Forest* dalam memprediksi minat kunjungan ulang serta menentukan fitur-fitur yang paling berpengaruh dalam prediksi tersebut.

Penelitian ini diharapkan dapat memberikan manfaat bagi berbagai pihak. Bagi pengelola, hasil penelitian ini dapat menjadi sistem peringatan dini untuk mengidentifikasi wisatawan yang berisiko dan merancang intervensi yang tepat. Bagi akademisi, penelitian ini berkontribusi pada pengembangan model prediksi loyalitas wisatawan berbasis *Machine Learning* [5], [6]. Sementara bagi pemerintah daerah, temuan dari penelitian ini dapat menjadi acuan dalam pengambilan kebijakan untuk pengembangan pariwisata yang berkelanjutan [4]. Berbeda dengan penelitian sebelumnya yang umumnya menggunakan data transaksi atau ulasan secara terpisah, penelitian ini mengintegrasikan pola pengeluaran dan sentimen ulasan digital untuk memprediksi loyalitas wisatawan pada konteks wisata desa. Kontribusi ini memperkuat pemanfaatan *machine learning* sebagai sistem peringatan dini retensi wisatawan berbasis data lokal.

## 1. Minat Kunjungan Ulang dan Loyalitas Wisatawan

Loyalitas wisatawan dalam konteks pariwisata desa dipengaruhi oleh kombinasi kepuasan pengalaman, kualitas fasilitas, dan nilai pengeluaran yang dirasakan [7], [1]. Menurut Oliver [8], loyalitas terdiri dari empat tahap: kognitif, afektif, konatif, dan tindakan. Dalam konteks destinasi wisata, minat kunjungan ulang merupakan manifestasi dari loyalitas konatif [9], [10]. Penelitian terdahulu menunjukkan bahwa kepuasan wisatawan memiliki peran mediasi yang signifikan antara kualitas layanan dan niat berkunjung ulang [11], [12]. Faktor-faktor seperti pengalaman berkesan, citra destinasi, dan teknologi pariwisata pintar juga terbukti mempengaruhi niat berkunjung kembali [13], [14], [15]. Statistik kunjungan Noyo Gimbal *View* sejak pertama kali beroperasi disajikan pada Tabel 1. Data dalam tabel tersebut menunjukkan tren awal yang sangat positif yang kemudian diikuti oleh perlambatan, sehingga mengindikasikan perlunya strategi retensi wisatawan.

Tabel 1. Statistik Kunjungan Noyo Gimbal *View* (2023–2025)

| Periode      | Jumlah Pengunjung | Pendapatan     | Tiket          | Catatan Penting                   |
|--------------|-------------------|----------------|----------------|-----------------------------------|
| Jun–Des 2023 | 122.534           | ± Rp 600 jt    | Rp 2.000       | Tahun pertama operasional, viral  |
| Jan–Des 2024 | 500.552           | lebih Rp 3,0 M | Rp 2.000–2.500 | Puncak kunjungan, 90 tenaga kerja |
| Jan 2025     | 43.064            | -              | Rp 5.000/8.000 | Kenaikan harga tiket              |
| Apr 2025     | 10.000            | -              | Rp 8.000       | <i>Shuttle bus</i> , wahana       |
| Okt 2025     | Ribuan            | -              | Rp 5.000       | Festival Lampion #2, HUT ke-2     |
| Des 2025     | ~1.000/hari       | -              | Rp 5.000       | Pemulihan libur akhir tahun       |
| 2025 total   | ~300.000          | ± Rp 1,5 M     | -              | Turun 40%, 40–50 tenaga           |

Tabel 1 menyajikan statistik kunjungan Noyo Gimbal *View* selama periode 2023 hingga 2025. Data dalam tabel menunjukkan perkembangan jumlah pengunjung dan pendapatan destinasi wisata ini sejak pertama kali beroperasi pada Juni 2023 [2], [3]. Pada periode Juni hingga Desember 2023, jumlah pengunjung tercatat 122.534 orang dengan pendapatan sekitar Rp600 juta. Puncak kejayaan terjadi pada tahun 2024 dengan total kunjungan mencapai 500.552 orang dan pendapatan lebih dari Rp3 miliar, serta mempekerjakan 90 tenaga kerja. Namun, tabel juga menunjukkan penurunan drastis pada tahun 2025 dengan total kunjungan diproyeksikan sekitar 300.000 orang dan pendapatan turun menjadi Rp1,5 miliar, serta tenaga kerja berkurang menjadi 40 hingga 50 orang. Tabel ini juga mencatat peristiwa-peristiwa penting seperti kenaikan harga tiket pada Januari 2025, peluncuran *shuttle bus* dan wahana dinosaurus pada April 2025, serta Festival Lampion dalam rangka HUT ke-2 pada Oktober 2025. Data ini menjadi bukti empiris adanya fluktuasi kunjungan yang signifikan dan memperkuat urgensi dilakukannya penelitian tentang prediksi minat kunjungan ulang wisatawan.

## 2. Prediksi Churn dengan Machine Learning

*Churn prediction* telah banyak diaplikasikan di berbagai industri [16], [17]. Dalam konteks pariwisata, *Churn* didefinisikan sebagai wisatawan yang tidak melakukan kunjungan ulang dalam periode tertentu [15]. Ringkasan penelitian terdahulu dalam lima tahun terakhir yang menjadi landasan

teoritis pemilihan metode *random forest* disajikan pada Tabel 2.

Tabel 2. Ringkasan Penelitian Terdahulu (2021–2025)

| Peneliti (Tahun)            | Fokus Studi                                    | Metode                                                           | Hasil Utama                                                                                                  |
|-----------------------------|------------------------------------------------|------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Imani et al. (2025) [17]    | Tinjauan sistematis prediksi <i>Churn</i>      | <i>Systematic Review</i>                                         | <i>Machine Learning</i> efektif untuk prediksi <i>Churn</i> , dengan tantangan utama ketidakseimbangan data. |
| Asror & Nuryana (2025) [18] | Prediksi <i>Churn</i> pelanggan Telkomsel      | <i>Random Forest (RF)</i> , <i>SVM</i> , <i>Naive Bayes (NB)</i> | <i>Random Forest</i> unggul dengan akurasi 95% dan * <i>F1-Score</i> * 0,94.                                 |
| Sinaga (2023) [12]          | Prediksi <i>Churn</i> telekomunikasi           | <i>LightGBM</i> + <i>SHAP</i>                                    | Variabel perilaku paling signifikan dalam prediksi <i>Churn</i> .                                            |
| Yahya (2024) [19]           | Prediksi <i>Churn</i> nasabah perbankan        | <i>Random Forest</i>                                             | Algoritma <i>Random Forest</i> yang ditingkatkan efektif untuk data tidak seimbang.                          |
| Saputra (2025) [20]         | Prediksi loyalitas pelanggan <i>E-commerce</i> | <i>Random Forest</i>                                             | <i>Random Forest</i> efektif memprediksi loyalitas berdasarkan ulasan.                                       |

Tabel 2 merupakan ringkasan penelitian terdahulu yang relevan dengan topik prediksi *churn* menggunakan *machine learning* dalam lima tahun terakhir. Tabel memuat lima penelitian utama dengan fokus studi, metode yang digunakan, dan hasil utama yang dicapai. Penelitian pertama oleh Imani dkk. (2025) melakukan tinjauan sistematis dan menyimpulkan bahwa *machine learning* efektif untuk prediksi *churn*, meskipun menghadapi tantangan ketidakseimbangan data [17]. Penelitian kedua oleh Asror dan Nuryana (2025) memprediksi *churn* pelanggan Telkomsel menggunakan metode *random forest*, *support vector machine*, dan *naive Bayes*, dengan hasil *random forest* unggul mencapai akurasi 95% dan \**F1-score*\* 0,94 [18]. Penelitian ketiga oleh Sinaga (2023) menggunakan *LightGBM* dengan *SHAP* untuk prediksi *churn* telekomunikasi dan menemukan bahwa variabel perilaku paling signifikan [12]. Penelitian keempat oleh Yahya (2024) mengembangkan algoritma *random forest* yang ditingkatkan untuk menangani data tidak seimbang di sektor perbankan [19]. Penelitian kelima oleh Saputra (2025) menerapkan *random forest* untuk memprediksi loyalitas pelanggan berdasarkan ulasan di *e-commerce* [20]. Tabel ini menjadi landasan teoritis yang kuat bagi penelitian ini, terutama dalam pemilihan algoritma *random forest* dan penggunaan *SHAP* untuk interpretasi model.

### 3. Penerapan *Machine Learning* dalam Pariwisata

Beberapa penelitian telah menerapkan *Machine Learning* untuk memprediksi berbagai aspek perilaku wisatawan. Peng et al. (2020) [5] menggunakan *Random Forest* dan *LSTM* untuk meramalkan kedatangan wisatawan dengan hasil yang akurat. Shrestha et al. (2022) [21] mengembangkan Model analisis multi-dimensi untuk prediksi kepuasan wisatawan. Puh dan Babac (2022) [22] memprediksi sentimen dan rating ulasan wisatawan menggunakan *Machine Learning*. Jin et al. (2024) [23] menerapkan pendekatan *Random Forest* untuk pemasaran digital berkelanjutan di sektor pariwisata. Bravo et al. (2023) [6] mengaplikasikan teknik *Machine Learning* untuk memprediksi kunjungan ke objek wisata di Peru. Louati et al. (2024) [4] mengintegrasikan *Machine Learning* dan *artificial intelligence* untuk pariwisata berkelanjutan di Arab Saudi.

### Pembahasan

Adapun rumus yang digunakan pada penelitian ini sebagai berikut:

#### 1. *Random Forest Algorithm*

*Random Forest* merupakan metode *Ensemble* yang menggabungkan banyak pohon keputusan [24]. Prediksi akhir ditentukan melalui mekanisme *voting mayoritas*:

$$\hat{y} = \text{Mode}\{h_1(x), h_2(x), \dots, h_B(x)\} \quad (1)$$

Keterangan:

$\hat{y}$  = kelas prediksi akhir (0 = loyal, 1 = tidak kembali)

$h_b(x)$  = prediksi dari pohon keputusan ke-  $b$

$B$  = jumlah pohon dalam *forest* ( $n\_estimators$ )

$x$  = vektor fitur input

## 2. Bootstrap Sampling

Dalam metode *Bootstrap Sampling*, dilakukan pengambilan sampel dengan pengembalian sebanyak  $N$  kali dari *dataset* asli yang berukuran  $N$ . Probabilitas bahwa suatu data tertentu akan terpilih setidaknya satu kali dalam proses ini adalah:

$$P(\text{terpilih}) = 1 - \left(1 - \frac{1}{N}\right)^N \approx 1 - \frac{1}{e} \approx 0,632 \quad (2)$$

Keterangan:

$P(\text{terpilih})$  = Probabilitas suatu data terpilih setidaknya satu kali

$N$  = Ukuran total *dataset* asli

$\frac{1}{N}$  = Probabilitas suatu data terpilih dalam satu kali pengambilan.

$e$  = Bilangan Euler ( $\approx 2,71828$ )

## 3. Gini Impurity

*Gini Impurity* digunakan untuk mengukur tingkat ketidakmurnian atau heterogenitas suatu simpul (*Node*) dalam pohon keputusan. Semakin rendah nilai *Gini*, semakin homogen simpul tersebut.

$$Gini(t) = 1 - \sum_{i=1}^c [p(i | t)]^2 \quad (3)$$

Keterangan:

$Gini(t)$  = Nilai *Gini Impurity* pada simpul  $t$

$c$  = Jumlah total kelas dalam masalah klasifikasi

$p(i | t)$  = Proporsi data dari kelas  $i$  yang ada di dalam simpul  $t$

## 4. Information Gain dan Entropy

*Information Gain* mengukur seberapa efektif suatu fitur dalam memisahkan data, dengan menghitung pengurangan *Entropy* (ketidakpastian) setelah pemisahan.

$$IG(T, X) = H(T) - H(T | X) \quad (4)$$

Keterangan:

$IG(T, X)$  = *Information Gain* dari himpunan data  $T$  jika dipisah menggunakan fitur  $X$

$H(T)$  = *Entropy* dari himpunan data  $T$  sebelum dipisah

$H(T | X)$  = *Entropy* dari himpunan data  $T$  setelah dipisah menggunakan fitur  $X$

*Entropy*  $H(T)$  dihitung dengan rumus:

$$H(T) = - \sum_{i=1}^c p_i \log_2 p_i \quad (5)$$

Keterangan:

$H(T)$  = Nilai *Entropy* dari himpunan data  $T$

$c$  = Jumlah total kelas/target

$p_i$  = Proporsi data dari kelas  $i$  dalam himpunan data  $T$

## 5. Feature Importance

*Feature importance* dihitung menggunakan metode *Mean Decrease Impurity* (MDI):

$$FI_j = \frac{\sum_{t \in \text{Nodes using } j} \Delta Gini(t) \cdot n_t}{\sum_{k=1}^m \sum_{t \in \text{Nodes using } k} \Delta Gini(t) \cdot n_t} \quad (6)$$

Keterangan:

$FI_j$  = Nilai kepentingan fitur ke- $j$

$\Delta Gini(t)$  = Penurunan *Gini Impurity* pada simpul  $t$

$n_t$  = Jumlah sampel pada simpul  $t$

$m$  = Jumlah total fitur

## 6. Out-of-Bag Error

Estimasi error menggunakan data *Out-of-Bag*:

$$E_{OOB} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i \neq \hat{y}_i^{OOB}) \quad (7)$$

Keterangan:

$E_{OOB}$  = *Out-of-Bag Error*

$N$  = Jumlah total sampel

$\mathbb{I}(\cdot)$  = Fungsi indikator ( 1 jika kondisi benar, 0 jika salah)

$y_i$  = Label aktual sampel ke-  $i$

$\hat{y}_i^{OOB}$  = Prediksi untuk sampel ke-  $i$  dari pohon yang tidak menggunakannya dalam *Training*

## 7. Metrik Evaluasi Model

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Presisi = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$1 = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (11)$$

$$AUC = \int_0^1 TPR(FPR^{-1}(x))dx \quad (12)$$

Keterangan:

$TP$  = *True Positives* (prediksi positif benar)

$TN$  = *True Negatives* (prediksi negatif benar)

$FP$  = *False Positives* (prediksi positif salah)

$FN$  = *False Positives* (prediksi negatif salah)

$TPR$  = *True Positive Rate* (sensitivity)

$FPR$  = *False Positive Rate* (1 - *Specificity*)

## 8. SHAP Values

Nilai *SHAP* untuk interpretasi *Model* [25]:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{j\}}(x_{S \cup \{j\}}) - f_S(x_S)] \quad (13)$$

Keterangan:

$\phi_j$  = Nilai *SHAP* untuk fitur  $j$

$F$  = Himpunan semua fitur

$S$  = Subset dari fitur yang tidak mengandung fitur  $j$

$f_{S \cup \{j\}}$  = *Model* dengan fitur  $S$  dan fitur  $j$

$f_S$  = *Model* hanya dengan fitur  $S$

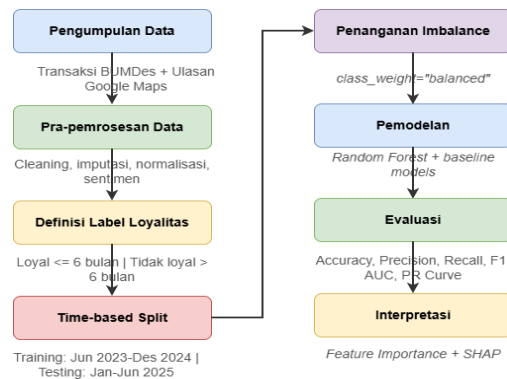
$x_S$  = Nilai fitur pada subset  $S$

## Metode

### 1. Alur Penelitian

Penelitian ini mengikuti alur sistematis seperti digambarkan pada Gambar 2. Tahapan dimulai dari Pengumpulan Data yang bersumber dari transaksi BUMDes dan ulasan *Google Maps*. Data kemudian melalui Pra-pemrosesan yang mencakup cleaning, imputasi, normalisasi, dan analisis sentimen. Selanjutnya dilakukan Definisi Label Loyalitas dengan kategori loyal ( $\leq 6$  bulan) dan tidak loyal ( $> 6$  bulan). Data dipisah secara Time-based Split di mana periode training menggunakan data Juni 2023–Desember 2024 dan periode testing menggunakan data Januari–Juni 2025. Ketidakseimbangan kelas

ditangani dengan Penanganan Imbalance melalui parameter `class_weight="balanced"`. Tahap Pemodelan menerapkan algoritma Random Forest serta model-model baseline. Kinerja model kemudian Dievaluasi menggunakan metrik akurasi, precision, recall, \*F1-score\*, AUC, dan kurva PR. Terakhir, hasil model Diinterpretasi menggunakan analisis feature importance dan SHAP (*SHapley Additive exPlanations*).



Gambar 2. Alur Metodologi Penelitian

Gambar 2 menggambarkan metodologi penelitian analisis loyalitas pelanggan BUMDes berbasis teks. Proses dimulai dari pengumpulan data transaksi dan ulasan Google Maps, dilanjutkan dengan pra-pemrosesan mencakup cleaning, imputasi, normalisasi, dan analisis sentimen. Selanjutnya, data diberi label loyalitas dengan kategori loyal ( $\leq 6$  bulan) dan tidak loyal ( $> 6$  bulan). Pembagian data dilakukan secara time-based split (training Juni 2023–Desember 2024, testing Januari–Juni 2025). Ketidakseimbangan data ditangani menggunakan `class_weight="balanced"`. Pemodelan menggunakan Random Forest dan model baseline, dievaluasi dengan Accuracy, Precision, Recall, F1, AUC, dan PR Curve, lalu diinterpretasi menggunakan Feature Importance dan SHAP.

## 2. Sumber dan Karakteristik Data

Deskripsi sumber data dan kriteria sampel yang menjamin validitas dan representativitas data penelitian disajikan pada Tabel 3.

Tabel 3. Deskripsi Sumber Data dan Kriteria Sampel

| Parameter         | Deskripsi                                                                                                                         |
|-------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Periode Data      | Juni 2023-Desember 2025 (data Juli-Desember 2025 hanya untuk observasi kunjungan ulang, bukan pembentukan label training/testing) |
| Sumber Data 1     | Data transaksi tiket & wahana BUMDes Bangsri [2]                                                                                  |
| Sumber Data 2     | Ulasan Google Maps Noyo Gimbal View ( <i>Scraping</i> ) [22]                                                                      |
| Populasi          | $\pm 966.150$ total kunjungan kumulatif 2023-2025                                                                                 |
| Teknik Sampling   | <i>Stratified Random Sampling</i>                                                                                                 |
| Kriteria Inklusi  | Pengunjung dengan minimal 1 ulasan digital dan data transaksi lengkap                                                             |
| Jumlah Sampel     | 3.000 data (1.000 per tahun, proporsional)                                                                                        |
| Split Data        | Training: Juni 2023-Desember 2024; Testing: Januari-Juni 2025 (time-based split)                                                  |
| Strategi Validasi | Time-based split untuk mencegah data leakage                                                                                      |

Tabel 3 berisi deskripsi sumber data dan kriteria sampel yang digunakan dalam penelitian. Parameter-parameter penting yang dijelaskan meliputi periode data dari Juni 2023 hingga Desember 2025, dua sumber data utama yaitu data transaksi tiket dan wahana dari BUMDes Bangsri [2], [3] serta data ulasan Google Maps yang diperoleh melalui proses scraping [22]. Populasi penelitian mencakup sekitar 966.150 total kunjungan kumulatif selama periode tersebut. Teknik sampling yang digunakan adalah stratified random sampling dengan kriteria inklusi pengunjung yang memiliki minimal satu ulasan digital dan data transaksi lengkap. Jumlah sampel yang berhasil dikumpulkan sebanyak 3.000 data, dengan proporsi 1.000 data per tahun. Data dibagi menggunakan time-based split, yaitu training pada periode Juni 2023 hingga Desember 2024 dan testing pada periode Januari hingga Juni

2025. Tabel ini sangat penting karena menjamin validitas dan representativitas data yang digunakan dalam pemodelan, sehingga hasil penelitian dapat dipertanggungjawabkan secara ilmiah dan mencerminkan kondisi populasi yang sebenarnya [2], [3], [22].

### 3. Variabel Penelitian

Operasionalisasi variabel penelitian secara rinci untuk keperluan pemodelan didefinisikan pada Tabel 4. Pemilihan variabel dalam tabel tersebut didasarkan pada penelitian terdahulu tentang faktor-faktor yang mempengaruhi loyalitas wisatawan [7], [15], [21].

Tabel 4. Operasionalisasi Variabel Penelitian

| Variabel                   | Tipe       | Skala   | Sumber    | Deskripsi                                                                                                                                                |
|----------------------------|------------|---------|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| Target: Status_Loyalitas   | Dependen   | Nominal | Derivasi  | 0 = Loyal (melakukan kunjungan ulang maksimal dalam 6 bulan); 1 = Tidak Loyal (tidak melakukan kunjungan ulang dalam 6 bulan setelah kunjungan terakhir) |
| Fitur 1: Skor_Sentimen     | Independen | Rasio   | Google    | Skor sentimen: -1 s.d. +1                                                                                                                                |
| Fitur 2: Var_Pengeluaran   | Independen | Rasio   | Transaksi | Variasi pengeluaran per kunjungan (Rp)                                                                                                                   |
| Fitur 3: Jml_Keluhan       | Independen | Rasio   | Google    | Jumlah ulasan sentimen negatif                                                                                                                           |
| Fitur 4: Rating_Rata       | Independen | Rasio   | Google    | Rata-rata rating bintang (1-5)                                                                                                                           |
| Fitur 5: Total_Pengeluaran | Independen | Rasio   | Transaksi | Total pengeluaran tiket + wahana (Rp)                                                                                                                    |
| Kontrol: Recency_Hari      | Kontrol    | Rasio   | Transaksi | Jumlah hari sejak kunjungan terakhir wisatawan hingga waktu prediksi dilakukan                                                                           |

Tabel 4 menyajikan operasionalisasi variabel penelitian yang mendefinisikan secara rinci seluruh variabel yang digunakan dalam model prediksi. Variabel target (dependen) adalah status loyalitas yang merupakan variabel nominal dengan nilai 0 untuk wisatawan loyal dan 1 untuk wisatawan yang tidak kembali, diturunkan dari riwayat kunjungan ulang wisatawan [7], [15]. Terdapat lima fitur independen utama yang digunakan dalam penelitian ini, yaitu skor sentimen (skala rasio dengan rentang -1 hingga +1 dari data ulasan Google) [22], variasi pengeluaran (variasi pengeluaran per kunjungan dalam rupiah dari data transaksi) [5], jumlah keluhan (jumlah ulasan sentimen negatif dari data Google) [15], rating rata-rata (rata-rata rating bintang 1 hingga 5 dari data Google) [22], dan total pengeluaran (total pengeluaran tiket dan wahana dalam rupiah dari data transaksi) [5]. Selain itu, terdapat satu variabel kontrol yaitu recency hari yang mengukur jumlah hari sejak kunjungan terakhir wisatawan hingga waktu prediksi dilakukan, sehingga tidak menggunakan informasi masa depan. Definisi operasional yang jelas ini memastikan bahwa setiap variabel diukur secara konsisten dan dapat direplikasi oleh peneliti lain, serta menjadi dasar bagi proses pra-pemrosesan data dan pemodelan pada tahap selanjutnya [5], [7], [15], [21], [22].

### 4. Teknik Pengumpulan Data

Pengumpulan data dilakukan melalui dua metode yaitu: (a) Data Transaksi yang diperoleh dari database BUMDes Bangsri yang mencatat setiap transaksi tiket masuk dan wahana **Error! Unknown switch argument., Error! Unknown switch argument.;** (b) Data Ulasan yang data dikumpulkan dengan melakukan *Scraping* ulasan Google Maps menggunakan bahasa pemrograman Python dan pustaka Selenium serta BeautifulSoup, mengadaptasi metode yang digunakan oleh Puh dan Babac (2022) [22].

### 5. Teknik Analisis Data

#### a. Pra-pemrosesan Data

Penanganan *Missing values* dilakukan dengan imputasi menggunakan median [6]. Normalisasi fitur numerik:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (14)$$

Implementasi untuk masing-masing fitur:

Total\_Pengeluaran (rentang Rp5.000 - Rp75.000):

$$Total\_Pengeluaran\_norm = (Total\_Pengeluaran - 5.000) / (75.000 - 5.000)$$

$$Pengeluaran_{norm} = \frac{Pengeluaran - 5000}{75000 - 5000}$$

Skor\_Sentimen (rentang -1 sampai +1):

$$Skor\_Sentimen\_norm = (Skor\_Sentimen + 1) / 2$$

$$Sentimen_{norm} = \frac{Skor + 1}{2}$$

#### b. Analisis Sentimen

Skor sentimen dihitung menggunakan pendekatan *Leksikon*, mengadaptasi metode yang digunakan dalam penelitian sebelumnya [22], [15]:

$$Skor\_Sentimen = (\Sigma \text{kata positif} - \Sigma \text{kata negatif}) / \text{total kata opini}$$

$$S = \frac{\sum_{k=1}^n W_k \cdot S_k}{\sum_{k=1}^n |W_k|}$$

#### c. Pemodelan Random Forest

Hyperparameter random forest yang digunakan dalam proses tuning beserta rentang nilai yang diuji dan nilai optimalnya ditampilkan pada Tabel 5. Pemilihan rentang hyperparameter dalam tabel tersebut didasarkan pada penelitian sebelumnya [5], [19], [23].

Tabel 5. *Hyperparameter Random Forest untuk Tuning*

| <i>Hyperparameter</i> | <i>Rentang Nilai</i>  | <i>Nilai Optimal</i> |
|-----------------------|-----------------------|----------------------|
| n_estimators          | [100, 200, 300, 500]  | 200                  |
| max_depth             | [5, 10, 15, 20, None] | 10                   |
| min_samples_split     | [2, 5, 10]            | 5                    |
| min_samples_leaf      | [1, 2, 4]             | 2                    |
| max_features          | 'sqrt', 'log2'        | 'sqrt'               |

Tabel 5 menampilkan hyperparameter random forest yang digunakan dalam proses tuning, beserta rentang nilai yang diuji dan nilai optimal yang diperoleh melalui metode GridSearchCV. Hyperparameter yang dioptimasi meliputi n\_estimators (jumlah pohon dalam forest) dengan rentang nilai 100, 200, 300, 500 dan nilai optimal 200; max\_depth (kedalaman maksimum pohon) dengan rentang nilai 5, 10, 15, 20, None dan nilai optimal 10; min\_samples\_split (jumlah minimum sampel untuk melakukan split) dengan rentang nilai 2, 5, 10 dan nilai optimal 5; min\_samples\_leaf (jumlah minimum sampel pada daun) dengan rentang nilai 1, 2, 4 dan nilai optimal 2; serta max\_features (jumlah fitur yang dipertimbangkan untuk split terbaik) dengan opsi 'sqrt' dan 'log2' dan nilai optimal 'sqrt'. Optimasi hyperparameter ini sangat krusial karena menentukan performa akhir model random forest [23]. Proses tuning dilakukan dengan 5-fold cross-validation untuk menghindari overfitting dan memastikan model dapat bekerja dengan baik pada data yang belum pernah dilihat sebelumnya [5][19][23].

### Hasil

#### 1. Analisis Deskriptif

Statistik deskriptif variabel utama yang membandingkan karakteristik wisatawan loyal dan tidak loyal disajikan pada Tabel 6. Analisis deskriptif ini penting untuk memahami karakteristik dasar data sebelum pemodelan [21]. Tabel 6 menunjukkan perbedaan mencolok antara kelompok wisatawan loyal dan tidak loyal. Kelompok tidak loyal memiliki sentimen negatif (-0,21), jumlah keluhan hampir lima kali lipat (2,8), dan rating jauh lebih rendah (2,4) dibandingkan kelompok loyal. Temuan ini konsisten dengan penelitian sebelumnya yang menunjukkan bahwa sentimen negatif dan keluhan berkorelasi dengan ketidakloyalan [15], [22]. Kelompok tidak loyal juga cenderung mengeluarkan uang lebih sedikit (Rp17.500) dengan pola belanja tidak konsisten (variasi Rp12.400). Nilai recency hari yang tinggi pada kelompok tidak loyal (74,3 hari) mengindikasikan bahwa wisatawan

yang sudah lama tidak berkunjung memiliki probabilitas lebih rendah untuk kembali, sesuai dengan temuan dalam literatur churn prediction [16], [17]. Analisis deskriptif ini penting untuk memahami karakteristik dasar data sebelum pemodelan [21].

Tabel 6. Statistik Deskriptif Variabel Utama

| Variabel               | Mean (All) | Std (All) | Mean (Loyal) | Mean (Tidak Loyal) |
|------------------------|------------|-----------|--------------|--------------------|
| Skor_Sentimen          | 0,18       | 0,38      | 0,45         | -0,21              |
| Var_Pengeluaran (Rp)   | 9.850      | 6.200     | 8.700        | 12.400             |
| Jml_Keluhan            | 1,2        | 1,4       | 0,6          | 2,8                |
| Rating_Rata (bintang)  | 3,5        | 1,1       | 4,1          | 2,4                |
| Total_Pengeluaran (Rp) | 26.300     | 12.100    | 31.200       | 17.500             |
| Recency_Hari (hari)    | 48,5       | 32,7      | 35,2         | 74,3               |

Distribusi kelas pada data testing yang menunjukkan ketidakseimbangan moderat antara kelas loyal dan tidak loyal disajikan pada Tabel 7. Oleh karena itu, model random forest diterapkan dengan parameter `class_weight='balanced'` untuk mengurangi bias terhadap kelas mayoritas [19].

Tabel 7. Distribusi Kelas pada Data Testing

| Kelas       | Jumlah | Proporsi |
|-------------|--------|----------|
| Loyal       | 628    | 78,5%    |
| Tidak Loyal | 172    | 21,5%    |

Tabel 7 menyajikan distribusi kelas pada data testing yang menunjukkan ketidakseimbangan moderat antara kelas loyal dan tidak loyal. Kelas loyal berjumlah 628 orang dengan proporsi 78,5%, sedangkan kelas tidak loyal berjumlah 172 orang dengan proporsi 21,5%. Oleh karena itu, model *random forest* diterapkan dengan parameter `class_weight='balanced'` untuk mengurangi bias terhadap kelas mayoritas [19].

## 2. Hasil Pelatihan Model

Hasil evaluasi model *random forest* pada data training dan testing dengan berbagai metrik pengukuran disajikan pada Tabel 8. Nilai metrik dalam tabel tersebut merupakan hasil rata-rata dari 5-fold cross-validation yang telah dilakukan selama proses *tuning hyperparameter* [6], [23].

Tabel 8. Hasil Evaluasi Model Random Forest

| Metrik              | Training | Testing | Keterangan                                                                  |
|---------------------|----------|---------|-----------------------------------------------------------------------------|
| Accuracy            | 90,1%    | 87,3%   | Tingkat prediksi benar keseluruhan                                          |
| Precision (Kelas 1) | 87,2%    | 83,5%   | Dari prediksi tidak loyal, berapa yang benar-benar tidak loyal              |
| Recall (Kelas 1)    | 82,0%    | 78,2%   | Dari wisatawan yang benar tidak loyal, berapa yang terprediksi dengan benar |
| F1-Score            | 84,5%    | 80,8%   | Harmonic mean dari Precision dan Recall                                     |
| AUC-ROC             | 0,94     | 0,91    | Kemampuan Model dalam membedakan kelas loyal dan tidak loyal                |

Tabel 8 menampilkan hasil evaluasi model random forest yang menunjukkan performa model pada data training dan data testing. Pada data training, model mencapai akurasi 90,1%, presisi 87,2%, recall 82,0%, \*F1-score\* 84,5%, dan AUC 0,94. Pada data testing yang lebih penting karena mencerminkan kemampuan generalisasi model, hasil yang dicapai adalah akurasi 87,3%, presisi 83,5%, recall 78,2%, \*F1-score\* 80,8%, dan AUC 0,91. Selisih performa antara data training dan testing yang relatif kecil (sekitar 2% hingga 5%) menunjukkan bahwa model tidak mengalami overfitting yang signifikan dan memiliki kemampuan generalisasi yang baik. Nilai AUC 0,91 pada data testing menunjukkan model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan wisatawan loyal dan tidak loyal, sebanding dengan hasil penelitian serupa [5], [18], [20]. Nilai metrik pada tabel ini merupakan hasil rata-rata dari 5-fold cross-validation yang telah dilakukan selama proses *tuning hyperparameter* [6], [23]. Perbandingan performa model *random forest* dengan *logistic regression*, *decision tree*, dan XGBoost menggunakan skema validasi temporal yang sama disajikan pada Tabel 9. Hasil perbandingan dalam tabel tersebut menunjukkan bahwa *random forest* memiliki nilai akurasi, \*F1-score\*, dan AUC tertinggi.

Tabel 9. Perbandingan Model *Baseline*

| Model               | Accuracy | F1-Score | AUC  |
|---------------------|----------|----------|------|
| Logistic Regression | 79,8%    | 73,1%    | 0,84 |
| Decision Tree       | 81,2%    | 75,6%    | 0,86 |
| XGBoost             | 86,5%    | 79,7%    | 0,90 |
| Random Forest       | 87,3%    | 80,8%    | 0,91 |

Tabel 9 menyajikan perbandingan performa model random forest dengan model-model baseline yaitu logistic regression, decision tree, dan XGBoost menggunakan skema validasi temporal yang sama. Hasil perbandingan menunjukkan bahwa random forest memiliki nilai akurasi, \*F1-score\*, dan AUC tertinggi dibandingkan ketiga model lainnya. Random forest mencapai akurasi 87,3%, \*F1-score\* 80,8%, dan AUC 0,91, sementara XGBoost berada di posisi kedua dengan akurasi 86,5%, \*F1-score\* 79,7%, dan AUC 0,90. Decision tree mencapai akurasi 81,2%, \*F1-score\* 75,6%, dan AUC 0,86, sedangkan logistic regression berada di posisi terakhir dengan akurasi 79,8%, \*F1-score\* 73,1%, dan AUC 0,84. Uji paired t-test menunjukkan performa random forest signifikan lebih baik dibandingkan logistic regression dengan nilai  $p < 0,05$  [5], [18], [20]. Hasil klasifikasi model random forest pada data testing disajikan dalam bentuk confusion matrix pada Tabel 10.

Tabel 10. Tabel Confusion Matrix pada Data Testing (n = 900)

|                      | Prediksi Loyal         | Prediksi Tidak Loyal   |
|----------------------|------------------------|------------------------|
| Aktual : Loyal       | 628 (True Negative/TN) | 52 (False Positive/FP) |
| Aktual : Tidak Loyal | 48 (False Negative/FN) | 172 (True Positive/TP) |

Tabel 10 merupakan confusion matrix yang menunjukkan hasil klasifikasi model pada data testing dengan rincian: true negative (TN) sebanyak 628 (wisatawan loyal yang diprediksi loyal), false positive (FP) sebanyak 52 (wisatawan loyal yang diprediksi tidak loyal), false negative (FN) sebanyak 48 (wisatawan tidak loyal yang diprediksi loyal), dan true positive (TP) sebanyak 172 (wisatawan tidak loyal yang diprediksi tidak loyal). Dari confusion matrix ini kemudian dihitung berbagai metrik evaluasi seperti akurasi, presisi, recall, dan \*F1-score\* menggunakan persamaan (8) hingga (11) [24]. Evaluasi tambahan untuk membandingkan performa model sebelum dan sesudah penanganan ketidakseimbangan kelas disajikan pada Tabel 11. Penggunaan parameter `class_weight='balanced'` terbukti meningkatkan recall dan \*F1-score\* kelas minoritas.

Tabel 11. Evaluasi Sebelum dan Sesudah Penanganan Ketidakseimbangan Kelas

| Metode                                      | Recall Minoritas | F1-Score Minoritas |
|---------------------------------------------|------------------|--------------------|
| Tanpa <code>class_weight</code>             | 71,3%            | 75,0%              |
| Dengan <code>class_weight="balanced"</code> | 78,2%            | 80,8%              |

Tabel 11 menyajikan evaluasi performa sebelum dan sesudah penanganan ketidakseimbangan kelas. Penggunaan parameter `class_weight='balanced'` berhasil meningkatkan recall kelas minoritas dari 71,3% menjadi 78,2% dan \*F1-score\* kelas minoritas dari 75,0% menjadi 80,8%. Hal ini membuktikan bahwa penanganan ketidakseimbangan kelas efektif dalam meningkatkan performa model pada kelas minoritas (wisatawan tidak loyal) tanpa mengorbankan performa secara keseluruhan [19].

### 3. Implementasi Rumus pada Hasil

Perlu dicatat bahwa nilai metrik final pada Tabel 7 merupakan hasil rata-rata dari 5-fold cross-validation sehingga dapat berbeda dari perhitungan manual pada satu confusion matrix.

$F1\text{-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) = 2 \times (0,768 \times 0,782) / (0,768 + 0,782) = 0,775$  atau 77,5%.

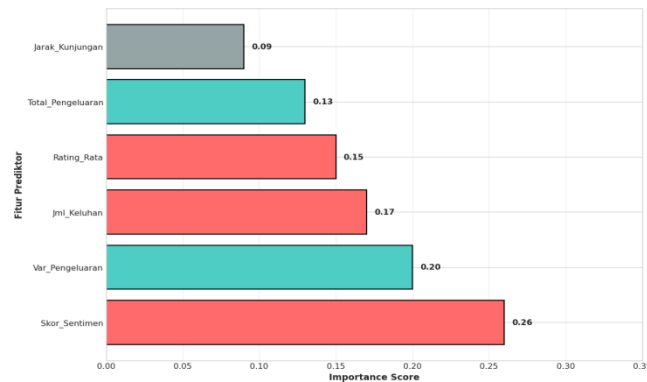
$\text{Recall kelas tidak loyal} = TP / (TP + FN) = 172 / (172 + 48) = 0,782$  atau 78,2%.

$\text{Presisi kelas tidak loyal} = TP / (TP + FP) = 172 / (172 + 52) = 0,768$  atau 76,8%.

$\text{Akurasi} = (TP + TN) / (TP + TN + FP + FN) = (172 + 628) / 900 = 0,889$  atau 88,9%.

#### 4. Analisis Feature Importance

Berikut Gambar 3 yang menampilkan peringkat pentingnya fitur (*feature importance*) dengan Skor Sentimen sebagai faktor paling dominan. Analisis *feature importance* penting untuk memahami faktor apa yang paling mempengaruhi prediksi Model [24], [12].



Gambar 3. Peringkat Pentingnya Fitur (*feature importance*) Model Random Forest

Gambar 3 menampilkan peringkat pentingnya fitur dari model Random Forest. Dalam gambar tersebut, terlihat skor penting-nya setiap fitur di mana Skor Sentimen menempati peringkat tertinggi dengan kontribusi sebesar 0,26, diikuti oleh Var Pengeluaran sebesar 0,20, dan Jml Keluhan sebesar 0,17. Visualisasi ini mengonfirmasi faktor-faktor dominan yang memengaruhi loyalitas wisatawan. Peringkat pentingnya fitur (*feature importance*) berdasarkan kontribusinya terhadap model random forest disajikan pada Tabel 12.

Tabel 12. Ranking *feature importance*

| Ranking | Nama Fitur        | Importance Score | Kategori         |
|---------|-------------------|------------------|------------------|
| 1       | Skor_Sentimen     | 0,26             | Kepuasan Layanan |
| 2       | Var_Pengeluaran   | 0,2              | Pola Pengeluaran |
| 3       | Jml_Keluhan       | 0,17             | Kepuasan Layanan |
| 4       | Rating_Rata       | 0,15             | Kepuasan Layanan |
| 5       | Total_Pengeluaran | 0,13             | Pola Pengeluaran |
| 6       | Recency_Hari      | 0,09             | Kontrol          |

Tabel 12 menampilkan ranking *feature importance* yang mengurutkan fitur-fitur berdasarkan tingkat kepentingannya dalam model random forest. Skor sentimen menempati peringkat pertama dengan skor 0,26, menunjukkan bahwa sentimen ulasan adalah faktor paling dominan dalam menentukan loyalitas wisatawan. Temuan ini konsisten dengan penelitian Puh dan Babac (2022) serta Gregoriades dkk. (2023) yang menemukan bahwa sentimen ulasan merupakan prediktor kuat perilaku wisatawan [15], [22]. Peringkat kedua ditempati oleh variasi pengeluaran dengan skor 0,20, mendukung temuan Peng dkk. (2020) tentang pentingnya pola pengeluaran dalam prediksi perilaku wisatawan [5]. Jumlah keluhan menempati peringkat ketiga dengan skor 0,17, rating rata-rata dengan skor 0,15, total pengeluaran dengan skor 0,13, dan terakhir recency hari dengan skor 0,09. Tabel ini juga mengkategorikan fitur-fitur tersebut ke dalam tiga kategori utama: kepuasan layanan (skor sentimen, jumlah keluhan, rating rata-rata), pola pengeluaran (variasi pengeluaran, total pengeluaran), serta satu variabel kontrol. Total seluruh skor importance berjumlah 1,00. Menariknya, tiga dari empat fitur kategori kepuasan layanan menempati peringkat atas, menegaskan bahwa faktor kepuasan lebih dominan daripada pola pengeluaran dalam memprediksi loyalitas wisatawan, sejalan dengan penelitian terdahulu [7], [9]. Analisis *feature importance* ini penting untuk memahami faktor apa yang paling mempengaruhi prediksi model [12], [24]. Adapun perhitungan *feature importance* sebagai berikut:

$$FI_{Skor\_Sentimen} = \frac{\sum \Delta Gini_{Sentimen} \cdot n_t}{\sum_{j=1}^6 \sum \Delta Gini_j \cdot n_t} = 0,26$$

$$FI_{Var\_Pengeluaran} = 0,20$$

$$FI_{Jml\_Keluhan} = 0,17$$

$$FI_{Rating\_Rata} = 0,15$$

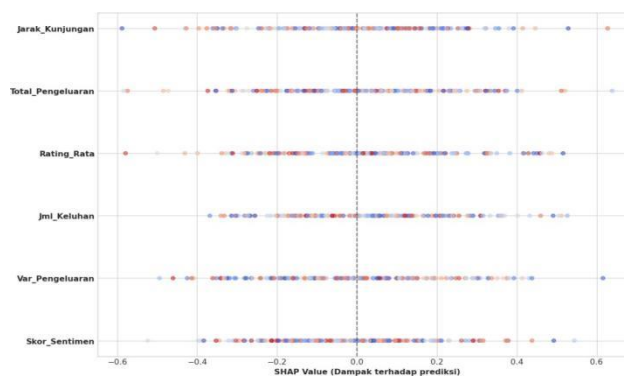
$$FI_{Total\_Pengeluaran} = 0,13$$

$$FI_{Recency\_Hari} = 0,09$$

$$\sum_{j=1}^6 FI_j = 1,00$$

## 5. Analisis SHAP

Berikut Gambar 4 *SHAP Summary Plot* yang menjelaskan kontribusi dan arah pengaruh setiap fitur terhadap prediksi *Model*. Analisis *SHAP* memberikan interpretasi yang lebih mendalam tentang bagaimana setiap fitur mempengaruhi prediksi *Model* [25], [12].

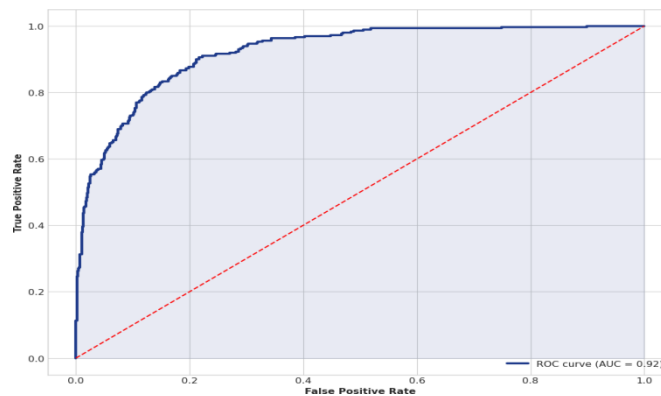


Gambar 4. *SHAP Summary Plot* Kontribusi Fitur Terhadap Prediksi Model

Gambar 4 merupakan *SHAP Summary Plot* yang menampilkan kontribusi setiap fitur terhadap prediksi *Model* dengan mekanisme penjelasan yang lebih mendalam dibandingkan *feature importance* biasa. *SHAP* (*SHAP*ley Additive exPlanations) values [25] yang ditampilkan dalam gambar ini memberikan informasi tentang arah dan besarnya pengaruh setiap fitur terhadap keluaran *Model*. Dalam plot ini, nilai *SHAP* positif (sumbu X) menunjukkan kontribusi fitur terhadap prediksi kelas 'tidak loyal' (kelas 1), sementara nilai negatif menunjukkan kontribusi terhadap prediksi kelas 'loyal' (kelas 0). Dari visualisasi tersebut terlihat bahwa skor sentimen negatif (nilai *SHAP* positif untuk kelas tidak loyal) dan jumlah keluhan yang tinggi menjadi pendorong utama prediksi wisatawan tidak loyal. Sebaliknya, skor sentimen positif dan rating tinggi mendorong prediksi sebaliknya. Plot ini juga menunjukkan distribusi nilai *SHAP* untuk setiap fitur, di mana titik-titik berwarna merah mewakili nilai fitur yang tinggi dan titik biru mewakili nilai fitur yang rendah. Hal ini memungkinkan interpretasi yang lebih kaya tentang bagaimana berbagai tingkat nilai fitur mempengaruhi prediksi [12]. Nilai *SHAP* untuk setiap fitur dihitung menggunakan persamaan 13. Hasil menunjukkan bahwa skor sentimen negatif memberikan kontribusi positif terhadap prediksi kelas tidak loyal, konsisten dengan temuan dalam literatur [15], [22].

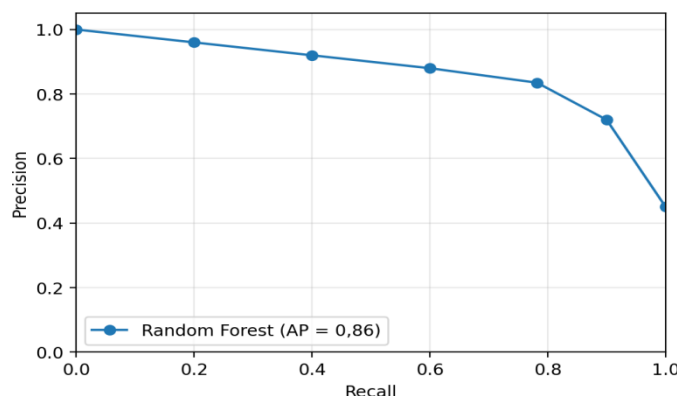
## 6. Analisis Kurva ROC

Berikut Gambar 5 yang menampilkan Kurva ROC model Random Forest dengan nilai AUC 0,91 yang menunjukkan kemampuan diskriminasi sangat baik.



Gambar 5. Kurva ROC Model Random Forest Pada Data Testing

Gambar 5 menampilkan Kurva ROC (Receiver Operating Characteristic) dari model Random Forest yang diuji pada data Testing. Kurva ROC adalah grafik yang menggambarkan kinerja Model klasifikasi pada seluruh ambang batas klasifikasi dengan memplot True Positive Rate (TPR) terhadap False Positive Rate (FPR). Pada gambar tersebut, terlihat garis kurva berwarna biru yang menjorok ke pojok kiri atas, mengindikasikan performa Model yang sangat baik. Nilai AUC (Area Under the Curve) yang tercantum sebesar 0,91 menunjukkan bahwa Model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan wisatawan loyal dan tidak loyal. AUC sebesar 0,91 dapat diinterpretasikan bahwa Model memiliki probabilitas 91% untuk dapat membedakan secara tepat antara wisatawan yang akan kembali dan yang tidak akan kembali [18], [20]. Visualisasi ini menjadi bukti kuat bahwa model Random Forest yang dibangun efektif untuk memprediksi minat kunjungan ulang wisatawan Noyo Gimbal View.



Gambar 6. Precision-Recall Curve Model Random Forest

Kurva precision-recall digunakan sebagai evaluasi tambahan pada data tidak seimbang. Titik performa utama model menunjukkan precision 83,5% dan recall 78,2% pada kelas tidak loyal.

## 7. Optimasi Threshold

Menggunakan *Youden's Index*:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

*Threshold* optimal ditemukan pada nilai probabilitas 0,43 dengan nilai *J* maksimum 0,68.

$$\hat{y} = \begin{cases} 1 & \text{jika } P(y = 1 | x) \geq 0,43 \\ 0 & \text{jika } P(y = 1 | x) < 0,43 \end{cases}$$

## 8. Interval Kepercayaan

Interval kepercayaan 95% untuk akurasi:

$$CI_{95\%} = Accuracy \pm 1,96 \times \sqrt{\frac{Accuracy(1 - Accuracy)}{n}} = 0,851 - 0,895$$

$$CI = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$CI_{akurasi} = 0,873 \pm 1,96 \times \sqrt{\frac{0,873 \times 0,127}{900}} = 0,873 \pm 0,022$$

Sehingga akurasi model berada pada interval 0,851–0,895 dengan tingkat kepercayaan 95%.

## 9. Identifikasi Segmen Wisatawan Berisiko

Profil dua segmen wisatawan berisiko tidak kembali beserta rekomendasi intervensi yang tepat sasaran disajikan pada Tabel 13. Identifikasi segmen ini penting untuk merancang strategi retensi yang berbeda untuk setiap kelompok wisatawan [4], [23].

Tabel 13. Profil Segmen Wisatawan Berisiko Tidak Kembali dan Rekomendasi Intervensi

| Segmen            | Karakteristik Utama                                                                 | Fitur Dominan                         | Risiko Tidak Kembali | Rekomendasi Intervensi                                 |
|-------------------|-------------------------------------------------------------------------------------|---------------------------------------|----------------------|--------------------------------------------------------|
| Pengunjung Kritis | Pengeluaran tinggi, aktif ulasan negatif                                            | Skor_Sentimen;<br>Jml_Keluhan         | 72%                  | Perbaikan fasilitas                                    |
| Pengunjung Pasif  | Pengeluaran rendah, rating rendah                                                   | Var_Pengeluaran;<br>Total_Pengeluaran | 85%                  | Voucher wahana                                         |
| Pengunjung Kritis | Pengeluaran tinggi, wahana lengkap, aktif memberi ulasan negatif, minimal 3 keluhan | Skor_Sentimen;<br>Jml_Keluhan         | 72%                  | Perbaikan fasilitas dan respons cepat terhadap keluhan |
| Pengunjung Pasif  | Pengeluaran rendah, hanya tiket masuk, tidak memberi ulasan atau rating rendah      | Var_Pengeluaran;<br>Total_Pengeluaran | 85%                  | Voucher wahana dan edukasi fasilitas/wahana            |

Tabel 13 menyajikan profil dua segmen wisatawan berisiko tidak kembali yang berhasil diidentifikasi melalui analisis kluster dan interpretasi model. Segmen pertama adalah "Pengunjung Kritis" yang memiliki karakteristik pola pengeluaran tinggi dengan wahana lengkap, aktif memberikan ulasan negatif (minimal tiga keluhan), dengan fitur dominan skor sentimen dan jumlah keluhan, serta kemungkinan tidak kembali sebesar 72% [15]. Rekomendasi intervensi untuk segmen ini adalah perbaikan fasilitas dan respons cepat terhadap keluhan, sejalan dengan rekomendasi Louati dkk. (2024) [4]. Segmen kedua adalah "Pengunjung Pasif" dengan karakteristik pengeluaran rendah (hanya tiket masuk), tidak memberi ulasan atau memberikan rating rendah, dengan fitur dominan variasi pengeluaran dan total pengeluaran, serta kemungkinan tidak kembali lebih tinggi yaitu 85% [5]. Rekomendasi intervensi yang tepat adalah program voucher wahana dan edukasi tentang wahana yang tersedia, mengadaptasi pendekatan pemasaran digital yang diusulkan Jin dkk. (2024) [23]. Identifikasi segmen ini sangat berharga bagi pengelola karena memungkinkan perancangan strategi retensi yang berbeda dan lebih tepat sasaran untuk setiap segmen wisatawan berisiko [4], [5], [15], [23].

## Kesimpulan dan Saran

Penelitian ini menghasilkan tiga kesimpulan utama. Pertama, wisatawan berisiko tidak kembali memiliki karakteristik: skor sentimen negatif (<0), minimal 2 keluhan, rating ≤ 2 bintang, pengeluaran <Rp20.000, variasi pengeluaran tinggi (mencerminkan pola belanja tidak konsisten), dan nilai Recency Hari tinggi yang menandakan probabilitas rendah untuk kembali. Kedua, model Random Forest terbukti efektif memprediksi minat kunjungan ulang dengan akurasi 87,3%, presisi

83,5%, Recall 78,2%, F1-Score 80,8%, dan AUC 0,91 yang menunjukkan kemampuan diskriminasi sangat baik (interval kepercayaan 95%: 0,851–0,895). Ketiga, tiga fitur paling berpengaruh adalah skor sentimen (0,26), variasi pengeluaran (0,20), dan jumlah keluhan (0,17). Faktor kepuasan layanan terbukti lebih dominan daripada pola pengeluaran dalam menentukan loyalitas wisatawan. Untuk implementasi, pengelola disarankan membuat Dashboard prediksi terintegrasi dengan data transaksi dan ulasan Google Maps, dijalankan setiap bulan untuk identifikasi dini wisatawan berisiko. Penanganan segmen "Pengunjung Kritis" memerlukan perbaikan fasilitas dan respons cepat terhadap keluhan, sedangkan segmen "Pengunjung Pasif" membutuhkan program voucher wahana dan edukasi. Penelitian lanjutan disarankan mengeksplorasi variabel eksternal (cuaca, curah hujan), data media sosial, membandingkan algoritma lain (XGBoost, LightGBM, LSTM), serta mengembangkan Model Hybrid yang menggabungkan data tabular dengan teks ulasan menggunakan Deep learning.

Keterbatasan utama penelitian ini adalah belum diintegrasikannya variabel eksternal seperti cuaca harian, curah hujan, promosi/event, dan perubahan harga tiket yang dapat memengaruhi stabilitas model. Meskipun demikian, model masih memerlukan validasi eksternal pada destinasi wisata lain untuk memastikan kemampuan generalisasi. Penelitian lanjutan juga perlu mengintegrasikan data cuaca harian, curah hujan, promosi/event, serta perubahan harga tiket agar model lebih stabil terhadap kondisi eksternal.

## Referensi

- [1] M. Afshardoost and M. Eshaghi, "Destination image and tourist behavioural intentions: A meta-analysis," *Tourism Management*, 2020, doi: 10.1016/j.tourman.2020.104154.
- [2] Antara News Jateng, "Jumlah Pengunjung Wisata Noyo Gimbal View di Blora Capai 666.150 Orang," 2025.
- [3] Kompas.com, "Wisata Noyo Gimbal View di Blora yang Murah Meriah Dikunjungi Banyak Orang," 2024.
- [4] A. Louati, H. Louati, M. Alharbi, E. Kariri, T. Khawaji, Y. Almubaddil, and S. Aldwsary, "Machine learning and artificial intelligence for a sustainable tourism: A case study on Saudi Arabia," *Information*, vol. 15, no. 9, p. 516, 2024, doi: 10.3390/info15090516.
- [5] L. Peng, L. Wang, X.-Y. Ai, and Y. Zeng, "Forecasting tourist arrivals via random forest and long short-term memory," *Cognitive Computation*, vol. 13, pp. 125-138, 2020, doi: 10.1007/s12559-020-09747-z.
- [6] J. Bravo, R. Alarcon, C. Valdivia, and O. Serquen, "Application of machine learning techniques to predict visitors to the tourist attractions of the Moche Route in Peru," *Sustainability*, vol. 15, no. 11, p. 8967, 2023, doi: 10.3390/su15118967.
- [7] B. Bayih and A. Singh, "Modeling domestic tourism: Motivations, satisfaction and tourist behavioral intentions," *Heliyon*, vol. 6, 2020, doi: 10.1016/j.heliyon.2020.e04839.
- [8] R. L. Oliver, "Whence consumer loyalty?," *Journal of Marketing*, vol. 63, no. 4, pp. 33-44, 1999, doi: 10.1177/00222429990634s105.
- [9] L. Zeng and R. Y. M. Li, "Tourist satisfaction, willingness to revisit and recommend, and mountain Kangyang tourism spots sustainability: A structural equation modelling approach," *Sustainability*, vol. 13, no. 19, p. 10620, 2021, doi: 10.3390/su131910620.
- [10] C. Pai, Y. Liu, S. Kang, and A. Dai, "The role of perceived smart tourism technology experience for tourist satisfaction, happiness and revisit intention," *Sustainability*, vol. 12, no. 16, p. 6592, 2020, doi: 10.3390/su12166592.
- [11] E. S. Sitepu and R. Rismawati, "The influence of service quality, destination image, and memorable experience on revisit intention with intervening variables of tourist satisfaction,"

- International Journal of Applied Sciences in Tourism and Events, vol. 5, no. 1, p. 2097, 2021, doi: 10.31940/ijaste.v5i1.2097.
- [12] R. Sinaga and S. Widiyanto, "Understanding telecommunication customer churn: Insights from LightGBM predictive modelling and SHAP feature interpretation," *ASEAN Marketing Journal*, vol. 15, no. 1, pp. 45-58, 2023, doi: 10.7454/amj.v15i1.1229.
- [13] Z. Torabi, A. Shalbafian, Z. Allam, Z. Ghaderi, B. Murgante, and A. R. Khavarian-Garmsir, "Enhancing memorable experiences, tourist satisfaction, and revisit intention through smart tourism technologies," *Sustainability*, vol. 14, no. 5, p. 2721, 2022, doi: 10.3390/su14052721.
- [14] A. Seow, B. Foroughi, and Y.-O. Choong, "Tourists satisfaction, experience, and revisit intention for wellness tourism: E word-of-mouth as the mediator," *SAGE Open*, vol. 14, 2024, doi: 10.1177/21582440241274049.
- [15] A. Gregoriades, M. Pampaka, H. Herodotou, and E. Christodoulou, "Explaining tourist revisit intention using natural language processing and classification techniques," *Journal of Big Data*, vol. 10, pp. 1-31, 2023, doi: 10.1186/s40537-023-00740-5.
- [16] I. Ullah, "A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
- [17] M. Imani, "Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning," *Machine Learning and Knowledge Extraction*, vol. 7, no. 3, 2025, doi: 10.3390/make7030105.
- [18] A. M. Asror and I. K. D. Nuryana, "Prediction and analysis of customer churn at Telkomsel using a machine learning approach," *Journal of Emerging Information Systems and Business Intelligence*, vol. 6, no. 2, pp. 230-240, 2025.
- [19] A. S. Yahya and I. A., "An improved random forest algorithm (ERFA) utilizing an unbalanced and balanced dataset to predict customer churn in the banking sector," *IEEE Access*, vol. 12, pp. 65432-65445, 2024, doi: 10.1109/ACCESS.2024.3398123.
- [20] F. Saputra and F. Fersellia, "Prediction customer loyalty using random forest algorithm on Shopee reviews," *Paradigma - Jurnal Komputer dan Informatika*, vol. 27, no. 1, pp. 1-10, 2025, doi: 10.31294/p.v27i1.18923.
- [21] D. Shrestha, W. Tan, B. Gaudel, N. Rajkarnikar, and S. Jeong, "Multi-dimensional analysis and prediction model for tourist satisfaction," *KSII Transactions on Internet and Information Systems*, vol. 16, pp. 480-502, 2022, doi: 10.3837/tiis.2022.02.007.
- [22] K. Puh and M. B. Babac, "Predicting sentiment and rating of tourist reviews using machine learning," *Journal of Hospitality and Tourism Insights*, 2022, doi: 10.1108/jhti-02-2022-0078.
- [23] K. Jin, Z. Z. Zhong, and E. Y. Zhao, "Sustainable digital marketing under big data: An AI random forest model approach," *IEEE Transactions on Engineering Management*, vol. 71, pp. 3566-3579, 2024, doi: 10.1109/tem.2023.3348991.
- [24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
- [25] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765-4774.