

Analisis Prediktif Pembatalan Reservasi Hotel Menggunakan Random Forest Berbasis PySpark untuk Mendukung Keputusan Industri Perhotelan

Predictive Analysis of Hotel Reservation Cancellation using Random Forest Based on PySpark to Support Decision-Making in the Hospitality Industry

Vellisya Afifa Qonita, Siti Laila Nurjannah, Indira Sistamarien
Silvia Ariani Daulay, Gema Parasti Mindara*, Aditya Wicaksono
IPB University, Bogor, Indonesia

Informasi Artikel:

Diterima: 5 Juni 2026, Direvisi: 19 Juni 2026, Disetujui: 28 Juni 2026

Abstrak-

Latar Belakang: Pembatalan *booking* hotel (*hotel booking cancellation*) dapat menyebabkan kerugian finansial dan menurunkan efektivitas operasional hotel.

Tujuan: Penelitian ini bertujuan untuk menganalisis faktor-faktor yang memengaruhi pembatalan reservasi hotel serta membangun model prediksi *cancellation booking* sebagai pendukung pengambilan keputusan pada industri perhotelan.

Metode: Metode penelitian yang digunakan adalah CRISP-DM (Cross Industry Standard Process for Data Mining) dengan pendekatan Big Data dan Machine Learning berbasis PySpark. Dataset yang digunakan adalah Hotel Booking Demand Dataset dengan jumlah 119.390 data reservasi. Tahapan pengolahan data meliputi data *cleaning*, penghapusan data duplikat, *encoding* data kategorikal, *feature assembling*, serta pembangunan model menggunakan Spark MLlib. Algoritma yang digunakan dalam penelitian ini adalah Logistic Regression sebagai *baseline* model dan Random Forest Classifier sebagai model prediksi utama.

Hasil: Hasil penelitian menunjukkan bahwa Random Forest memiliki performa terbaik dengan nilai Accuracy sebesar 78,42%, F1 Score sebesar 76,83%, dan ROC AUC sebesar 80,04%. Berdasarkan analisis feature importance, faktor yang paling berpengaruh terhadap pembatalan reservasi adalah *lead time*, *market segment*, dan *total special requests*.

Kesimpulan: Model yang dihasilkan dapat digunakan sebagai dasar penerapan *risk scoring* reservasi sehingga pihak hotel dapat mengidentifikasi *booking* berisiko tinggi dan menyusun strategi mitigasi pembatalan secara lebih efektif.

Kata Kunci: *Big Data; Business Intelligence; Hotel Booking Cancellation; PySpark; Random Forest.*

Abstract-

Background: Hotel booking cancellation may cause financial losses and reduce hotel operational effectiveness.

Objective: This study aims to analyze the factors influencing hotel reservation cancellations and develop a cancellation prediction model to support decision-making in the hospitality industry.

Methods: The research method used is CRISP-DM (Cross Industry Standard Process for Data Mining) with a Big Data and Machine Learning approach based on PySpark. The dataset used is the Hotel Booking Demand Dataset consisting of 119,390 reservation records. Data processing stages include data cleaning, duplicate data removal, categorical data encoding, feature assembling, and model development using Spark MLlib. The algorithms used in this study are Logistic Regression as a baseline model and Random Forest Classifier as the main prediction model.

Result: The results show that Random Forest achieved the best performance with an Accuracy of 78.42%, an F1 Score of 76.83%, and a ROC AUC of 80.04%. Based on feature importance analysis, the most influential factors affecting reservation cancellation are lead time, market segment, and total special requests.

Conclusion: The developed model can be used as a basis for implementing reservation risk scoring, enabling hotels to identify high-risk bookings and formulate more effective cancellation mitigation strategies.

Keywords: *Big Data; Business Intelligence; Hotel Booking Cancellation; PySpark; Random Forest.*

How to Cite: V. A. Qonita, S. L. Nurjannah, I. Sistamarien, S. A. Daulay, G. P. Mindara, & A. Wicaksono, "Analisis Prediktif Pembatalan Reservasi Hotel Menggunakan Random Forest Berbasis PySpark untuk Mendukung Keputusan Industri Perhotelan," *Jurnal Bumigora Information Technology (BITe)*, vol. 8, no. 1, pp. 14-28, Jun 2026, doi: 10.30812/bite.v8i1.6492.

This is an open access article under the CC BY-SA license (<https://creativecommons.org/licenses/by-sa/4.0/>)

Penulis Korespondensi:

Gema Parasti Mindara,
IPB University,
Email: gemaparasti@apps.ipb.ac.id

1. PENDAHULUAN

Industri perhotelan merupakan salah satu sektor usaha yang sangat bergantung pada pengelolaan reservasi untuk menjaga tingkat hunian kamar serta kestabilan pendapatan perusahaan. Reservasi adalah pengaturan untuk seseorang pada waktu tertentu [1]. Dalam kegiatan operasionalnya, hotel sering menghadapi permasalahan berupa pembatalan reservasi (*booking cancellation*) yang dilakukan oleh pelanggan [2]. Tingginya angka pembatalan reservasi dapat menimbulkan berbagai dampak negatif, karena kamar yang telah dipesan berpotensi tidak terjual kembali pada waktu yang tersedia [3]. Selain menyebabkan kerugian finansial, kondisi ini juga dapat mengganggu proses perencanaan operasional hotel, seperti pengaturan kapasitas kamar, penjadwalan karyawan, pengelolaan layanan, hingga penyusunan strategi pemasaran [4]. Dalam industri perhotelan, *travel*, dan pariwisata, pembatalan pesanan juga sangat memengaruhi proses pengambilan keputusan manajerial karena pihak pengelola perlu menyesuaikan strategi operasional berdasarkan perubahan tingkat kepastian reservasi pelanggan [5]. Permasalahan ini semakin kompleks dengan meningkatnya penggunaan platform reservasi daring dan layanan *Online Travel Agency* (OTA), yang memberikan kemudahan serta fleksibilitas bagi pelanggan dalam melakukan perubahan maupun pembatalan reservasi. Selain itu, fenomena *multiple booking* untuk mencari penawaran terbaik juga turut meningkatkan tingkat pembatalan reservasi [6]. Transformasi digital ini telah menjadi kebutuhan utama dalam industri perhotelan modern, di mana kecepatan, kenyamanan, dan personalisasi layanan semakin dihargai oleh pelanggan [7]. Digitalisasi layanan reservasi menjadi kebutuhan strategis untuk meningkatkan efisiensi operasional dan pengalaman pelanggan, sehingga hotel perlu memanfaatkan teknologi berbasis data untuk mengoptimalkan proses pengelolaan reservasi [8]. Perubahan tersebut juga mendorong hotel untuk memanfaatkan data pelanggan secara lebih optimal guna memahami preferensi serta perilaku reservasi pelanggan. Di sisi lain, perkembangan teknologi digital menyebabkan jumlah data reservasi hotel terus meningkat dalam skala besar dan semakin kompleks [9]. Data tersebut mencakup informasi penting seperti perilaku pelanggan, pola pemesanan, metode pembayaran, waktu reservasi, hingga riwayat pembatalan sebelumnya. Kondisi ini mendorong perlunya pemanfaatan teknologi *Big Data* dan *Business Intelligence* untuk membantu pihak hotel dalam mengelola serta menganalisis data secara lebih efektif. Melalui pendekatan *Machine Learning*, data historis reservasi dapat dianalisis untuk menemukan pola yang berkaitan dengan kemungkinan terjadinya pembatalan reservasi, sehingga hasilnya dapat digunakan sebagai dasar pengambilan keputusan bisnis yang lebih cepat, akurat, dan berbasis data [10].

Berbagai penelitian sebelumnya telah mengkaji penerapan *Machine Learning* dalam prediksi pembatalan reservasi hotel menggunakan algoritma seperti *Logistic Regression*, *Decision Tree*, *Random Forest*, *XGBoost*, dan *LightGBM*. Penelitian oleh Hermawan [11] menerapkan algoritma *Logistic Regression* dan *Decision Tree* untuk memprediksi kemungkinan terjadinya pembatalan reservasi hotel, dengan hasil menunjukkan bahwa variabel seperti *lead time*, *deposit type*, dan *market segment* memiliki pengaruh signifikan terhadap keputusan pembatalan dengan nilai F_2 -score sebesar ,7897 setelah dilakukan penyesuaian *decision threshold* menjadi ,35. Selanjutnya, penelitian Putro [12] membandingkan algoritma *Deep Neural Network* dan *Logistic Regression* untuk memprediksi pembatalan reservasi hotel. Hasil penelitian menunjukkan bahwa *Deep Neural Network* memberikan performa yang lebih baik, dengan arsitektur *Encoder–Decoder* menggunakan *optimizer Adamax* mencapai akurasi 86,57%, sedangkan *Logistic Regression* memperoleh akurasi 79,66%. Selain itu, penelitian tersebut mengidentifikasi bahwa atribut *country* dan *total_of_special_requests* merupakan fitur yang paling berpengaruh terhadap proses klasifikasi. Penelitian lain oleh Prasetya [13] mengembangkan model *Stacking*

Machine Learning untuk memprediksi pembatalan reservasi hotel dengan memanfaatkan *Naive Bayes*, *Logistic Regression*, dan *Gradient Boosting Machine* sebagai *base learner*, serta *Random Forest* sebagai *meta learner*. Hasil penelitian menunjukkan bahwa model *Stacking* dan *Gradient Boosting Machine* memperoleh akurasi tertinggi sebesar ,87. Selain itu, model *Stacking* menghasilkan sensitivitas sebesar ,86, sedangkan *Gradient Boosting Machine* memperoleh spesifisitas tertinggi sebesar ,88, yang menunjukkan kemampuan model dalam membedakan pelanggan yang melakukan dan tidak melakukan pembatalan reservasi. Penelitian oleh Sun [14] mengevaluasi berbagai algoritma *Machine Learning* dan *Deep Learning*, yaitu *Logistic Regression*, *Decision Tree*, *Random Forest*, *XGBoost*, MLP, CNN, DNN, dan LSTM untuk memprediksi pembatalan reservasi hotel. Hasil penelitian menunjukkan bahwa seluruh model memperoleh akurasi di atas 80%, sedangkan model berbasis *neural network* mencapai akurasi yang mendekati 100%. Selain itu, penelitian tersebut menunjukkan bahwa *feature selection* menggunakan PCA dan *Lasso* berkontribusi besar terhadap peningkatan performa model, sementara *Random Forest* menjadi algoritma klasik dengan performa terbaik. Penulis juga menyoroti bahwa kompleksitas model *Deep Learning* dan keterbatasan interpretasi masih menjadi tantangan dalam implementasi pada lingkungan bisnis nyata. Namun demikian, sebagian besar penelitian sebelumnya masih berfokus pada pengembangan model prediksi tanpa mengaitkan secara mendalam dengan kebutuhan pengambilan keputusan bisnis di industri perhotelan, serta masih terbatas dalam pemanfaatan teknologi *Big Data* pada *dataset* berskala besar [15]. Penelitian oleh Li [16] melalui *systematic review* terhadap 406 artikel menyimpulkan bahwa *machine learning* telah menjadi pendekatan utama dalam penelitian pariwisata dan perhotelan, terutama pada prediksi permintaan, *revenue management*, dan pembatalan reservasi hotel. *Random Forest* termasuk algoritma yang paling banyak digunakan karena mampu menangani data numerik yang kompleks. Namun demikian, penelitian-penelitian sebelumnya masih menghadapi tantangan terkait validasi model, *hyperparameter tuning*, *explainable AI*, kualitas data, serta generalisasi model sehingga aspek-aspek tersebut direkomendasikan sebagai arah pengembangan penelitian selanjutnya.

Penelitian ini menawarkan kebaruan melalui integrasi pemrosesan *Big Data* berbasis *PySpark* dengan pembangunan model *Machine Learning* untuk prediksi pembatalan reservasi hotel yang tidak hanya berfokus pada evaluasi performa model, tetapi juga menginterpretasikan hasil prediksi melalui analisis *feature importance*, simulasi prediksi pada profil pelanggan, serta penyusunan rekomendasi bisnis berupa *risk scoring* reservasi. Pendekatan tersebut diharapkan dapat menjembatani kesenjangan antara pengembangan model prediksi dan implementasinya sebagai sistem pendukung pengambilan keputusan pada industri perhotelan. Penelitian ini bertujuan untuk menganalisis faktor yang memengaruhi pembatalan reservasi, memahami pola perilaku pelanggan, membangun model prediksi *booking cancellation* menggunakan *Logistic Regression* dan *Random Forest*, serta memberikan rekomendasi untuk mendukung pengambilan keputusan. Sistem prediksi yang dikembangkan diharapkan mampu membantu hotel mengidentifikasi reservasi berisiko tinggi sejak awal sehingga strategi pencegahan pembatalan dan optimalisasi okupansi dapat dilakukan secara lebih efektif. Pemanfaatan *Machine Learning* pada data reservasi hotel dapat membantu pihak hotel dalam memahami pola perilaku pelanggan dan meningkatkan efektivitas strategi bisnis hotel [17]. Selain itu, penggunaan algoritma klasifikasi pada data reservasi hotel juga terbukti mampu membantu proses identifikasi pelanggan dengan risiko *cancellation* tinggi sehingga hotel dapat melakukan strategi pencegahan secara lebih efektif [18]. Penerapan *explainable artificial intelligence* (XAI) pada sistem prediksi hotel juga dinilai mampu meningkatkan transparansi model serta membantu proses *revenue management* dan efisiensi operasional hotel [19].

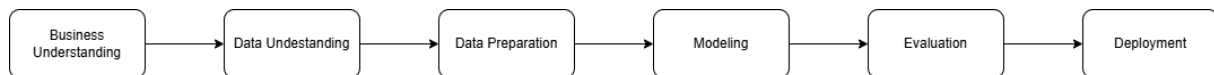
Sistematika penulisan artikel ini disusun ke dalam beberapa bagian sebagai berikut. Bagian pendahuluan menjelaskan latar belakang masalah, penelitian terdahulu, serta tujuan penelitian. Bagian metode penelitian menguraikan tahapan penelitian yang menggunakan metodologi CRISP-DM berbasis *Big Data* dan *PySpark*. Bagian hasil dan pembahasan menyajikan hasil analisis data, pembangunan model *Machine Learning*, evaluasi performa model, serta *insight* bisnis yang diperoleh. Selanjutnya, bagian kesimpulan berisi ringkasan hasil penelitian dan rekomendasi untuk pengembangan penelitian selanjutnya.

2. METODE PENELITIAN

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan *predictive analysis* berbasis *Machine Learning*. Studi kasus yang digunakan adalah permasalahan pembatalan reservasi hotel (*hotel booking cancellation*). Data yang digunakan berasal dari *Hotel Booking Demand Dataset* dengan jumlah 119.390 data reservasi yang berisi informasi karakteristik pelanggan, detail reservasi, serta status pembatalan (*is_canceled*) sebagai variabel target.

Pengolahan data dilakukan menggunakan *framework PySpark* pada lingkungan *Google Colab*. Pemilihan *PySpark* dilakukan karena mampu mendukung pemrosesan data berskala besar melalui konsep *distributed data processing* serta menyediakan *library Spark MLlib* untuk implementasi algoritma *Machine Learning* [20]. Algoritma yang digunakan dalam penelitian ini adalah *Logistic Regression* dan *Random Forest Classifier* untuk melakukan klasifikasi kemungkinan terjadinya pembatalan reservasi.

Metode penelitian yang digunakan adalah CRISP-DM (*Cross Industry Standard Process for Data Mining*) karena memiliki tahapan yang sistematis dalam proses analisis data, pembangunan model *Machine Learning*, hingga menghasilkan *insight* dan rekomendasi bisnis [21]. Tujuan dari metode ini adalah untuk melakukan proses analisis strategi yang digunakan untuk memecahkan masalah penelitian ataupun permasalahan dari sebuah bisnis atau perusahaan [22]. Berkat strukturnya yang sederhana namun ampuh, CRISP-DM dengan cepat menjadi populer dalam proyek analisis data [23]. Metodologi ini menyediakan kerangka kerja komprehensif yang mencakup enam fase berurutan [24] yang dapat dilihat dengan jelas pada Gambar 1 berikut.



Gambar 1. Alur Penelitian

2.1. Business Understanding

Tahap pertama adalah memahami permasalahan bisnis yang terjadi pada industri hotel terkait tingginya tingkat pembatalan reservasi. Pada tahap ini dilakukan identifikasi masalah, tujuan penelitian, serta kebutuhan bisnis yang ingin dicapai melalui penerapan analisis data dan *Machine Learning*. Berdasarkan permasalahan tersebut, tujuan penelitian ini adalah untuk menganalisis faktor-faktor yang memengaruhi *cancellation booking* hotel, memahami karakteristik pelanggan yang memiliki kecenderungan melakukan pembatalan reservasi, membangun model prediksi *cancellation booking*, serta menghasilkan *insight* dan rekomendasi bisnis berbasis data untuk membantu hotel mengurangi risiko *cancellation*.

2.2. Data Understanding

Pada tahap *data understanding* dilakukan proses pemahaman terhadap *dataset Hotel Booking Demand* yang digunakan dalam penelitian. *Dataset* terdiri dari data historis reservasi hotel yang memuat informasi terkait karakteristik pelanggan, proses pemesanan, tipe reservasi, serta status pembatalan reservasi. Variabel target yang digunakan adalah *is_canceled*, sedangkan atribut lainnya digunakan sebagai variabel prediktor dalam proses analisis dan pemodelan.

2.3. Data Preparation

Tahap *data preparation* dilakukan untuk membersihkan dan mempersiapkan *dataset* sebelum digunakan pada proses *Machine Learning*. Proses yang dilakukan meliputi pengecekan struktur *dataset*, pengecekan tipe data, pengecekan *missing value*, pengecekan *duplicate data*, pembersihan data, *feature selection*, dan transformasi data. Selanjutnya dilakukan proses *encoding* menggunakan *StringIndexer* pada fitur kategorikal seperti *hotel*, *market_segment*, *deposit_type*, *customer_type*, dan *arrival_date_month*. Setelah proses *encoding* selesai, seluruh fitur digabungkan menggunakan *VectorAssembler* menjadi *feature vector* untuk kebutuhan *Machine Learning*.

2.4. Modeling

Pada tahap *modeling* dilakukan pembangunan model *Machine Learning* untuk memprediksi kemungkinan *customer* melakukan *cancellation booking* hotel. Model yang digunakan pada penelitian ini yaitu *Logistic Regression* dan *Random Forest Classifier*. *Dataset* dibagi menjadi 80% data *training* dan 20% data *testing*. Model *Logistic Regression* digunakan sebagai model *baseline*, yang cukup banyak diimplementasikan dan efektif pada berbagai bidang penelitian [25]. Sedangkan *Random Forest* digunakan karena mampu menangani hubungan non-linear dan pola data yang kompleks, serta memiliki keandalan dalam menangani data yang tidak seimbang (*data imbalance*) [26]. Proses *modeling* dilakukan menggunakan *Spark MLlib* berbasis *PySpark* agar pengolahan data dapat dilakukan secara lebih efisien dalam lingkungan *Big Data*.

2.5. Evaluation

Tahap *evaluation* dilakukan untuk mengevaluasi performa model *Machine Learning* dalam memprediksi *cancellation booking* hotel. Metrik evaluasi yang digunakan meliputi *Accuracy*, *F1-Score*, dan *ROC-AUC*. Hasil evaluasi kemudian digunakan untuk membandingkan performa masing-masing model dan menentukan model terbaik yang akan digunakan pada tahap *deployment*.

2.6. Deployment/Business Insight

Tahap *deployment* dilakukan dengan menginterpretasikan hasil analisis dan model *Machine Learning* untuk menghasilkan *insight* dan rekomendasi bisnis yang dapat diterapkan pada industri perhotelan. Berdasarkan hasil analisis data dan evaluasi model, ditemukan bahwa *customer* dengan *lead_time* tinggi cenderung

3. HASIL DAN PEMBAHASAN

Pada bab ini dibahas hasil analisis data, evaluasi model *Machine Learning*, serta *insight* bisnis yang diperoleh dari penelitian prediksi pembatalan *booking* hotel menggunakan pendekatan *Big Data* dan *Business Intelligence* berbasis *PySpark*.

3.1. Logistic Regression

Setelah *training* pada 80% data, model *Logistic Regression* digunakan untuk mempelajari hubungan antara fitur-fitur reservasi dengan kemungkinan terjadinya pembatalan *booking*. Pengujian model selanjutnya dilakukan pada 20% data *testing* untuk mengukur kemampuan model dalam memprediksi data baru. Performa model yang diperoleh ditunjukkan pada Tabel 1 berikut.

Tabel 1. Metrik Evaluasi Logistic Regression

Metrik	Nilai
Accuracy	75.09%
F1 Score	70.80%
ROC AUC	74.17%

Model *Logistic Regression* sebagai model *baseline* sudah memberikan hasil yang cukup baik dalam melakukan prediksi *cancellation booking* hotel. Model ini mampu mengenali sebagian besar data dengan benar, terutama pada *customer* yang tidak melakukan *cancellation*. Model menghasilkan prediksi dalam bentuk probabilitas antara 0 hingga 1, di mana nilai di atas ,5 dikategorikan sebagai *cancel* dan di bawah ,5 sebagai tidak *cancel*. Dengan *accuracy* 75,09%, model sudah mampu memisahkan kedua kelas dengan cukup baik sebagai model *baseline*. Model ini cenderung lebih kuat dalam mengenali *customer* yang tidak *cancel* (kelas mayoritas) dibandingkan yang *cancel*, yang merupakan karakteristik umum dari *Logistic Regression* pada data yang tidak sepenuhnya seimbang.

3.2. Random Forest Classifier

Model *Random Forest* dilatih menggunakan kombinasi parameter berupa 100 *decision tree* dengan kedalaman maksimum 10 pada data *testing* yang sama dengan model sebelumnya. Kombinasi parameter tersebut dipilih untuk mendapatkan keseimbangan antara kemampuan model dalam mempelajari pola data dan kemampuannya dalam melakukan prediksi pada data yang belum pernah dilihat sebelumnya. Dengan konfigurasi tersebut, *Random Forest* menghasilkan performa yang lebih unggul dibandingkan *Logistic Regression*, sebagaimana terlihat pada Tabel 2 berikut.

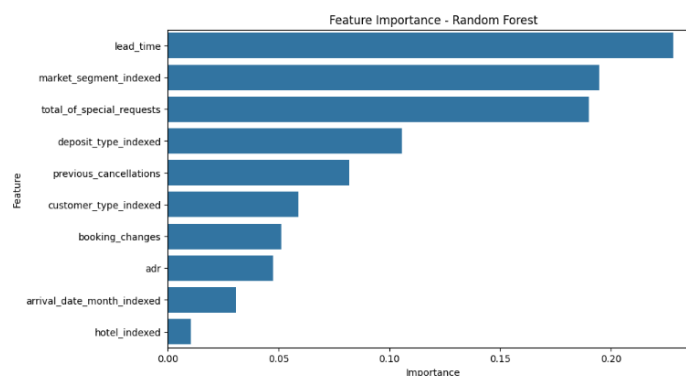
Tabel 2. Metrik Evaluasi Random Forest Classifier

Metrik	Nilai
Accuracy	78.42%
F1 Score	76.83%
ROC AUC	80.04%

Model *Random Forest* berhasil memberikan hasil prediksi yang lebih baik dibandingkan *Logistic Regression*, terutama dalam mendeteksi *customer* yang benar-benar melakukan *cancellation*. Hal ini terlihat dari nilai *recall* yang lebih tinggi, sehingga model lebih mampu mengenali kasus pembatalan *booking* dengan tepat. Model menghasilkan prediksi berdasarkan *voting* mayoritas dari 100 *decision tree* yang dibangun secara paralel. Dengan *accuracy* 78,42%, model berhasil memberikan performa yang lebih unggul dibandingkan *Logistic Regression*. Dibandingkan dengan *Logistic Regression*, *Random Forest* menghasilkan nilai *recall* yang lebih tinggi pada kelas *cancel*, artinya model lebih mampu menangkap *customer* yang benar-benar berpotensi membatalkan reservasi. Selain prediksi, model juga menghasilkan skor *feature importance* yang menunjukkan kontribusi masing-masing fitur terhadap keputusan prediksi [27]. Hasil ini memperkuat temuan penelitian sebelumnya yang menyatakan bahwa model *ensemble* seperti *Random Forest* lebih unggul dalam menangani data yang kompleks dan nonlinier dibandingkan model linear.

3.3. Feature Importance

Hasil analisis *feature importance* pada model *Random Forest* ditampilkan pada Gambar 2 berikut. Skor *feature importance* dihitung berdasarkan kontribusi masing-masing fitur dalam proses pembentukan *decision tree* pada model *Random Forest*. Semakin tinggi skor suatu fitur, semakin besar pengaruh fitur tersebut terhadap keputusan model dalam memprediksi *cancellation booking*.



Gambar 2. Feature Importance Random Forest

Berdasarkan hasil *feature importance* pada model *Random Forest*, fitur yang paling berpengaruh terhadap prediksi *cancellation* adalah *lead_time*, *market_segment_indexed*, dan *total_of_special_requests*. Ketiga fitur tersebut memiliki kontribusi terbesar dalam membantu model menentukan kemungkinan *customer* melakukan pembatalan *booking*. Selain itu, fitur *deposit_type_indexed* dan *previous_cancellations* juga menunjukkan pengaruh yang cukup signifikan terhadap proses prediksi. Hasil *feature importance* menunjukkan bahwa model *Random Forest* lebih

banyak memanfaatkan informasi yang berkaitan dengan karakteristik reservasi dan perilaku pelanggan selama proses pemesanan dibandingkan atribut lainnya dalam menentukan kemungkinan terjadinya *cancellation booking*. Secara keseluruhan, hasil *feature importance* memberikan gambaran mengenai variabel-variabel yang paling berkontribusi dalam proses prediksi sehingga dapat menjadi dasar dalam interpretasi model serta penyusunan strategi bisnis untuk mengurangi risiko *cancellation booking*.

3.4. Simulasi Prediksi

Untuk memvalidasi kegunaan model secara praktis, dilakukan simulasi prediksi pada dua profil tamu fiktif. Profil pertama merepresentasikan tamu dengan karakteristik reservasi berisiko tinggi sebagaimana ditampilkan pada Tabel 3, sedangkan profil kedua merepresentasikan tamu dengan karakteristik reservasi yang cenderung berkomitmen sebagaimana ditampilkan pada Tabel 4. Kedua profil tersebut disusun berdasarkan kombinasi fitur yang memiliki pengaruh besar pada hasil *feature importance*, yaitu *lead time*, *market segment*, dan *total special requests*.

Customer 1 (Profil Berisiko Tinggi)

Tabel 3. Profil *Customer* Berisiko Tinggi.

Fitur	Nilai
Lead Time	120 Hari
ADR	Rp. 150
Deposit Type	No Deposit
Market Segment	Online TA
Hotel	City Hotel
Special Request	0

Prediksi: *CANCEL* dengan probabilitas 82,08%. Profil tamu dengan *lead time* 120 hari, *Online TA*, tanpa *special request*, dan *No Deposit* cenderung memiliki risiko *cancellation* yang lebih tinggi. Faktor-faktor tersebut menunjukkan karakteristik reservasi yang pada data historis lebih sering dikaitkan dengan kejadian *cancellation*, sehingga kemungkinan untuk membatalkan *booking* menjadi lebih besar. Model berhasil mengenali pola tersebut dan memprediksi *customer* sebagai berpotensi melakukan *cancellation*.

Customer 2 (Profil Berkomitmen)

Tabel 4. Profil *Customer* Berkomitmen

Metrik	Nilai
Lead Time	2 Hari
ADR	Rp. 50
Deposit Type	Non Refund
Market Segment	Direct
Hotel	Resort Hotel
Special Request	5

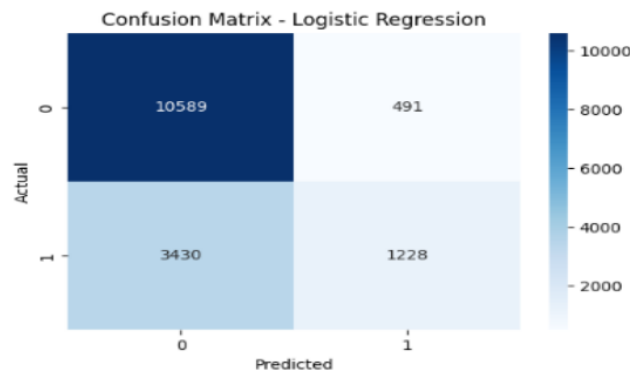
Prediksi: *TIDAK CANCEL* dengan probabilitas 87,13%. Sebaliknya, tamu dengan *lead time* hanya 2 hari, memesan langsung (*Direct*), mengajukan 5 *special request*, dan deposit *Non Refund* cenderung menunjukkan karakteristik reservasi yang pada data historis lebih sering dikaitkan dengan reservasi yang tidak dibatalkan. Berdasarkan kombinasi fitur tersebut, model memprediksi bahwa *customer* memiliki kemungkinan kecil untuk melakukan *cancellation*.

3.5. Evaluasi Model

3.5.1. Evaluasi Logistic Regression

Hasil evaluasi model *Logistic Regression* divisualisasikan melalui *confusion matrix* pada Gambar 3 berikut. *Confusion matrix* menampilkan jumlah prediksi benar dan salah untuk masing-masing kelas, yaitu *cancel* dan

tidak *cancel*. Melalui matriks ini, dapat diketahui sebaran *true positive*, *true negative*, *false positive*, dan *false negative* dari hasil prediksi model.

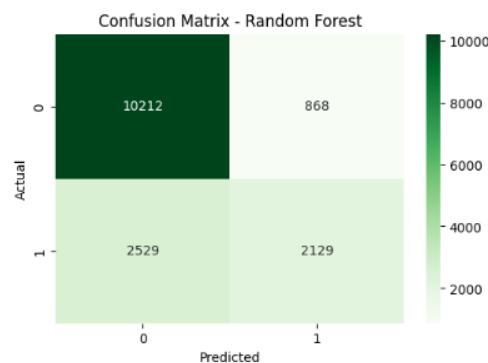


Gambar 3. *Confusion Matrix Logistic Regression*

Berdasarkan *confusion matrix*, model masih memiliki cukup banyak *false negative*, yaitu kasus ketika *customer* sebenarnya melakukan *cancellation* tetapi diprediksi tidak *cancel* oleh model. Hal ini menunjukkan bahwa *Logistic Regression* belum mampu mendeteksi seluruh pola *cancellation* secara optimal. Kondisi tersebut terjadi karena *Logistic Regression* bekerja dengan hubungan linear antarfitur, sedangkan pola pembatalan *booking* hotel pada *dataset* ini cenderung lebih kompleks dan dipengaruhi oleh banyak faktor yang saling berhubungan.

3.5.2. Evaluasi Random Forest

Hasil evaluasi model *Random Forest* divisualisasikan melalui *confusion matrix* pada Gambar 4 berikut. *Confusion matrix* ini menampilkan jumlah prediksi benar dan salah model *Random Forest* untuk masing-masing kelas. Hasil ini digunakan untuk menilai sejauh mana model mampu mendeteksi *customer* yang benar-benar melakukan *cancellation*.

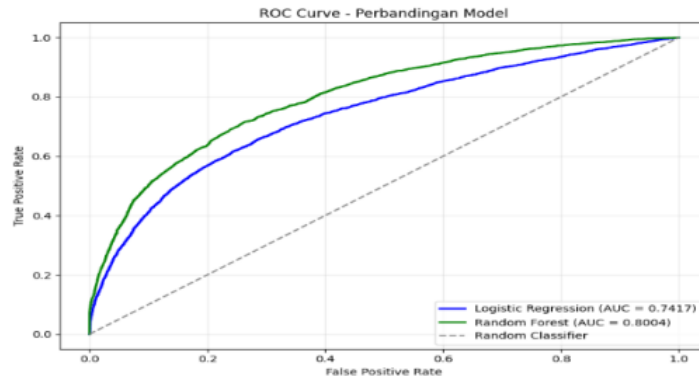


Gambar 4. *Confusion matrix Random Forest*

Berdasarkan *confusion matrix*, jumlah *false negative* pada *Random Forest* lebih sedikit dibandingkan *Logistic Regression*. Artinya, jumlah kasus *customer* yang sebenarnya *cancel* tetapi diprediksi tidak *cancel* menjadi lebih rendah. Dengan demikian, model lebih sensitif dalam mendeteksi *customer* yang berpotensi membatalkan *booking*. Selain itu, nilai ROC-AUC sebesar 80,04% menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan *customer* yang akan *cancel* dan tidak *cancel*. Nilai tersebut jauh di atas nilai acak (0,5), sehingga model dapat dikatakan cukup andal untuk membantu pengambilan keputusan bisnis hotel.

3.5.3. ROC-AUC

ROC-AUC merupakan metode evaluasi pada model klasifikasi yang digunakan untuk mengukur kemampuan model dalam membedakan kelas positif dan negatif [28]. *ROC Curve* menampilkan visualisasi performa model yang memetakan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai *threshold* [29], sedangkan AUC menunjukkan tingkat performa model dalam bentuk nilai numerik [30].



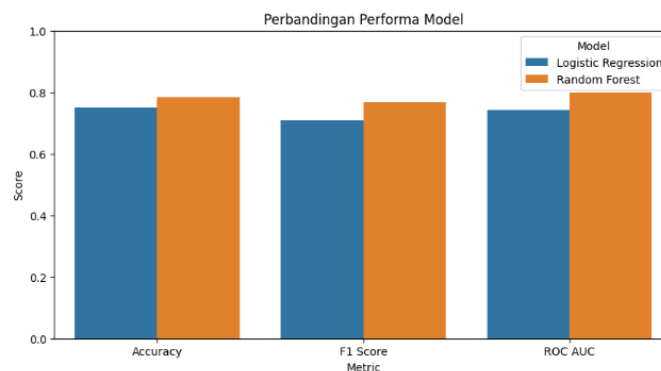
Gambar 5. Kurva ROC-AUC

Berdasarkan hasil *ROC Curve*, model *Random Forest* menunjukkan performa yang lebih baik dibandingkan *Logistic Regression* dengan nilai AUC sebesar ,8004 (80,04%), sedangkan *Logistic Regression* memperoleh nilai AUC sebesar ,7417 (74,17%). Nilai tersebut menunjukkan bahwa *Random Forest* memiliki kemampuan yang lebih baik dalam membedakan pelanggan yang melakukan pembatalan reservasi dan yang tidak melakukan pembatalan.

Kedua kurva ROC berada jauh di atas garis diagonal yang merepresentasikan prediksi acak ($AUC = ,5$), sehingga menunjukkan bahwa kedua model mampu mempelajari pola data dengan baik. Namun, *Random Forest* dinilai lebih efektif dalam menangkap hubungan kompleks antarfitur yang bersifat non-linear. Oleh karena itu, berdasarkan hasil evaluasi *Accuracy*, *F1-Score*, dan ROC-AUC, *Random Forest* dapat disimpulkan sebagai model terbaik dalam penelitian ini.

3.6. Perbandingan Kedua Model

Setelah kedua model dilatih dan dievaluasi pada data *testing* yang sama, diperoleh perbandingan performa kedua model sebagaimana ditunjukkan pada Gambar 6. Perbandingan tersebut mencakup tiga metrik evaluasi, yaitu *Accuracy*, *F1-Score*, dan ROC-AUC. Rincian nilai dari masing-masing metrik untuk kedua model disajikan secara lengkap pada Tabel 5 berikut:



Gambar 6. Grafik perbandingan kedua model

Selain itu, dapat dilihat juga perbandingan kedua model yang dikemas dalam bentuk Tabel 5. Tabel tersebut menyajikan hasil evaluasi model *Decision Tree* dan *Random Forest* berdasarkan metrik *Accuracy*, *Precision*, *Recall*, *F1-Score*, dan AUC. Melalui perbandingan tersebut, dapat diketahui model yang memiliki kinerja lebih baik dalam memprediksi pembatalan reservasi hotel.

Tabel 5. Perbandingan Kedua Model

Fitur	Logistic Regression	Random Forest	Selisih
Accuracy	75.09%	78.42%	+3.33%
F1 Score	70.80%	76.83%	+6.03%
ROC AUC	74.17%	80.04%	+5.87%

Dari tabel di atas, *Random Forest* secara konsisten memberikan performa lebih baik dibandingkan *Logistic Regression* pada seluruh metrik pengujian. Peningkatan *Accuracy* sebesar 3,33% menunjukkan kemampuan prediksi yang lebih tinggi, sedangkan peningkatan *F1-Score* menunjukkan bahwa *Random Forest* lebih seimbang dalam menangani kesalahan prediksi *false positive* dan *false negative*. Nilai ROC-AUC sebesar 80,04% juga menunjukkan kemampuan model yang baik dalam membedakan reservasi yang berpotensi *cancel* dan tidak *cancel*.

Random Forest lebih unggul karena mampu menangkap hubungan kompleks antarfitur yang bersifat non-linear, seperti interaksi antara *lead time*, *market segment*, *special request*, dan *deposit type*, yang sulit dipelajari oleh *Logistic Regression*. Hasil ini sejalan dengan penelitian sebelumnya oleh Hermawan et al. yang menunjukkan bahwa karakteristik reservasi seperti *lead time*, *market segment*, dan *deposit type* merupakan faktor penting yang memengaruhi pembatalan reservasi hotel. Temuan tersebut memperkuat bahwa informasi yang tersedia pada saat proses pemesanan dapat dimanfaatkan untuk mengidentifikasi risiko pembatalan sejak dini. Meskipun demikian, kedua model memiliki keterbatasan, yaitu adanya ketidakseimbangan kelas (*class imbalance*) pada *dataset* serta ketergantungan model terhadap data historis sehingga perubahan perilaku pelanggan dapat memengaruhi akurasi prediksi. Secara keseluruhan, *Random Forest* dipilih sebagai model terbaik karena lebih efektif dalam menangani kompleksitas data reservasi dan berpotensi digunakan sebagai sistem pendukung keputusan untuk mengidentifikasi *booking* berisiko tinggi serta membantu strategi pencegahan pembatalan.

3.7. Insight Bisnis

3.7.1. Lead Time dan Market Segment Sebagai Indikator Risiko Utama

Berdasarkan hasil *feature importance* *Random Forest*, *lead time* dan *market segment* merupakan faktor yang paling berpengaruh terhadap kemungkinan *cancellation booking*. *Customer* yang melakukan reservasi jauh hari sebelum tanggal *check-in* memiliki kecenderungan lebih tinggi untuk membatalkan *booking*. Pada *dataset* yang digunakan, *lead time* yang tinggi lebih sering ditemukan pada reservasi yang berakhir dengan pembatalan. Selain itu, beberapa *market segment* seperti *Online TA* juga menunjukkan tingkat *cancellation* yang lebih tinggi dibandingkan segmen lainnya.

3.7.2. Special Request Mencerminkan Komitmen Pelanggan

Hasil analisis menunjukkan bahwa *total-of-special-requests* merupakan salah satu fitur terpenting dalam prediksi *cancellation*. *Customer* yang memberikan lebih banyak permintaan khusus cenderung memiliki tingkat *cancellation* yang lebih rendah. Hasil tersebut menunjukkan adanya hubungan antara jumlah *special request* dan tingkat *cancellation*, di mana pelanggan dengan jumlah *special request* yang lebih tinggi cenderung memiliki tingkat pembatalan yang lebih rendah.

3.7.3. Kombinasi Lead Time Tinggi, Market Segment Online TA, dan Special Request Rendah Cenderung Berisiko Cancel

Berdasarkan hasil simulasi prediksi, *customer* dengan *lead time* tinggi, berasal dari *market segment* *Online TA*, dan memiliki jumlah *special request* yang rendah menunjukkan risiko *cancellation* yang lebih tinggi

dibandingkan profil *customer* lainnya. Pada simulasi *Customer* 1, tamu yang melakukan reservasi 120 hari sebelum *check-in*, berasal dari *market segment Online TA*, dan tidak memberikan *special request* memiliki probabilitas *cancellation* sebesar 82,08%. Kondisi tersebut menunjukkan bahwa kombinasi *lead time* yang tinggi, *market segment* tertentu, dan rendahnya jumlah *special request* dapat menjadi indikator risiko *cancellation* yang kuat.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa *lead time*, *market segment*, dan *total special requests* menjadi faktor utama yang memengaruhi pembatalan reservasi hotel. Reservasi dengan *lead time* tinggi, segmen pelanggan tertentu, serta jumlah *special request* yang rendah memiliki risiko pembatalan lebih besar. Model *Random Forest* terbukti lebih efektif dibandingkan *Logistic Regression* dalam menangkap hubungan kompleks antarvariabel, sehingga mampu meningkatkan prediksi pembatalan reservasi.

Kontribusi penelitian ini adalah mengintegrasikan *Big Data* berbasis *PySpark*, analisis *feature importance*, dan penerapan hasil prediksi sebagai rekomendasi bisnis bagi industri perhotelan. Penelitian ini tidak hanya mengevaluasi performa model, tetapi juga mengidentifikasi faktor penyebab pembatalan sebagai dasar strategi bisnis dan *risk scoring* reservasi. Penelitian selanjutnya dapat berfokus pada penanganan ketidakseimbangan data (*class imbalance*), pengujian algoritma lain, serta penerapan model secara *real-time* untuk menyesuaikan perubahan perilaku pelanggan.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan, bantuan, dan arahan selama proses penelitian ini berlangsung. Ucapan terima kasih disampaikan kepada dosen pembimbing yang telah memberikan bimbingan dan masukan dalam penyusunan penelitian, serta kepada Program Studi Rekayasa Perangkat Lunak IPB University yang telah mendukung pelaksanaan penelitian ini. Penulis juga mengucapkan terima kasih kepada seluruh anggota tim yang telah bekerja sama dalam proses pengumpulan data, pengolahan data, analisis, hingga penyusunan jurnal ini sehingga penelitian dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

- [1] R. Kivania, A. Novianti, dan R. Firmansyah, "Analisis Implementasi Peranan Sistem Reservasi Pada Bisnis Di Sektor Industri," *Student Scientific Creativity Journal*, vol. 1, no. 2, pp. 286–299, Mar. 30, 2023. DOI: [10.55606/sscj-amik.v1i2.1277](https://doi.org/10.55606/sscj-amik.v1i2.1277)
- [2] Z. Zafitri dan M. I. Jambak, "Karakteristik Pembatalan Reservasi Kamar Hotel Pada Online Travel Agent Menggunakan Algoritma C4.5," *The Indonesian Journal of Computer Science*, vol. 12, no. 4, Aug. 30, 2023. DOI: [10.33022/ijcs.v12i4.3268](https://doi.org/10.33022/ijcs.v12i4.3268)
- [3] E. W. Solang dan F. X. Adu, "Machine learning evaluation for hotel cancellation prediction with threshold adjustment and cost-based evaluation evaluasi machine learning untuk prediksi pembatalan hotel dengan threshold adjustment dan cost-based evaluation," *MALCOM: Journal of Machine Learning and Computer Science*, vol. 6, no. 1, pp. 193–204, 2026. DOI: [10.57152/malcom.v6i1.2466](https://doi.org/10.57152/malcom.v6i1.2466)
- [4] S. Ohoira dan S. Monica, "Prediksi Pembatalan Pemesanan Hotel menggunakan Algoritma Machine Learning," *Prosiding Sains dan Teknologi*, vol. 4, no. 1, pp. 495–503, 2025.
- [5] Y. Azhar, G. A. Mahesa, dan M. C. Mustaqim, "Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 1, pp. 15–21, Jan. 31, 2021. DOI: [10.14710/jtsiskom.2020.13790](https://doi.org/10.14710/jtsiskom.2020.13790)
- [6] J. Christian, I. Ernawati, dan N. Chamidah, "Implementasi penggunaan algoritma categorical boosting (catboost) dengan optimisasi hiperparameter dalam memprediksi pembatalan pesanan kamar hotel,"

- Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya*, vol. 3, no. 2, pp. 835–844, Aug. 20, 2022.
- [7] H. Herryani dan A. Y. Nugroho, *Buku Ajar Pengantar Industri Perhotelan*. Pekanbaru: CV. Bravo Press Indonesia, Aug. 2025.
- [8] F. A. Kurniawan dan A. R. A'yun, "Penerapan Sistem Reservasi Digital Untuk Meningkatkan Efisiensi Layanan Restoran: Studi Kasus Restoran Rasa Nusantara," *PROSENAMA*, vol. 5, pp. 209–225, 2025.
- [9] K. Z. Ramadhan dan B. Suharto, "Transformasi Digital Di Bidang Hospitalitas: Menjelajahi Peran Teknologi dalam Memperlancar Operasional Hotel," *Jurnal Ilmiah Hospitality Management*, vol. 14, no. 2, pp. 181–192, Jun. 30, 2024. DOI: [10.22334/jihm.v14i2.281](https://doi.org/10.22334/jihm.v14i2.281)
- [10] J. D. Devi dan K. Rambabu, "An Intelligent Hotel Booking Cancellation Prediction System Using Machine Learning Techniques," *International Journal of Engineering Research and Science & Technology*, vol. 22, no. 2, pp. 388–400, Apr. 3, 2026.
- [11] A. Hermawan et al., "Predicting Hotel Booking Cancellations Using Machine Learning for Revenue Optimization," *Router : Jurnal Teknik Informatika dan Terapan*, vol. 3, no. 1, pp. 37–48, Mar. 12, 2025. DOI: [10.62951/router.v3i1.400](https://doi.org/10.62951/router.v3i1.400)
- [12] N. A. Putro et al., "Prediction of Hotel Booking Cancellation Using Deep Neural Network and Logistic Regression Algorithm," *Jurnal Techno Nusa Mandiri*, vol. 18, no. 1, pp. 1–8, Mar. 15, 2021. DOI: [10.33480/techno.v18i1.2056](https://doi.org/10.33480/techno.v18i1.2056)
- [13] J. Prasetya, S. I. Fallo, dan M. A. Aprihartha, "Stacking Machine Learning Model for Predict Hotel Booking Cancellations," *Jurnal Matematika, Statistika dan Komputasi*, vol. 20, no. 3, pp. 525–537, May 15, 2024. DOI: [10.20956/j.v20i3.32619](https://doi.org/10.20956/j.v20i3.32619)
- [14] J. Sun, "Hotel booking cancellation and machine learning," *Advances in Engineering Innovation*, vol. 15, no. 1, pp. 45–62, Feb. 7, 2025. DOI: [10.54254/2977-3903/2025.20826](https://doi.org/10.54254/2977-3903/2025.20826)
- [15] O. Sanda, M. Pavlidis, dan N. Polatidis, "A Regulatory Readiness Assessment Framework for Blockchain Adoption in Healthcare," *Digital*, vol. 2, no. 1, pp. 65–87, Mar. 11, 2022. DOI: [10.3390/digital2010005](https://doi.org/10.3390/digital2010005)
- [16] N. Li, X. Liu, dan F. Meng, "Machine Learning Techniques in Tourism and Hospitality Research: A Critical Assessment," *Journal of Travel Research*, p. 00472875251401912, Dec. 31, 2025. DOI: [10.1177/00472875251401912](https://doi.org/10.1177/00472875251401912)
- [17] R. Cahyadi dan B. Setiawan, "Analisis Pemasaran Digitalisasi Berbasis AI pada Industri Perhotelan di Indonesia," *Jurnal Sains Terapan Pariwisata*, vol. 10, no. 3, pp. 279–287, Oct. 31, 2025. DOI: [10.56743/jstp.v10i3.690](https://doi.org/10.56743/jstp.v10i3.690)
- [18] C. Agustina dan E. Rahmawati, "Perbandingan Teknik Resample pada Algoritma K-NN dan SVM untuk Prediksi Pembatalan Pemesanan Kamar Hotel," *Jurnal Teknologi Informasi dan Terapan*, vol. 10, no. 2, pp. 102–107, Dec. 28, 2023. DOI: [10.25047/jtit.v10i2.333](https://doi.org/10.25047/jtit.v10i2.333)
- [19] I. Mohammed dan B. Denizci Guillet, "Stakeholder perspectives on integrating explainable artificial intelligence into hotel revenue management system," *Tourism Economics*, vol. 32, no. 4, pp. 804–824, Jun. 2026. DOI: [10.1177/13548166251382289](https://doi.org/10.1177/13548166251382289)
- [20] A. K. C. S. Eeti dan T. Singh, "Real-Time Data Processing: An Analysis of PySpark's Capabilities," *IJRAR-International Journal of Research and Analytical Reviews (IJRAR)*, E-ISSN, vol. 8, no. 3, pp. 2348–1269, 2021.
- [21] A. Rianti, N. W. A. Majid, dan A. Fauzi, "CRISP-DM: Metodologi Proyek Data Science," *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, pp. 106–114, Jul. 25, 2023.

- [22] M. Fitriani, G. F. Nama, dan M. Mardiana, “Implementasi Association Rule Dengan Algoritma Apriori Pada Data Peminjaman Buku UPT Perpustakaan Universitas Lampung Menggunakan Metodologi CRISP-DM,” *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 10, no. 1, pp. 41–49, Jan. 7, 2022. DOI: [10.23960/jitet.v10i1.2263](https://doi.org/10.23960/jitet.v10i1.2263)
- [23] M. Mendes dan T. Farinha, “MIDA—Method for Industrial Data Analysis Based on CRISP-DM,” *Machine Learning and Knowledge Extraction*, vol. 8, no. 4, p. 108, Apr. 18, 2026. DOI: [10.3390/make8040108](https://doi.org/10.3390/make8040108)
- [24] R. Hidayat, “Product segmentation for apparel msme using k-means and crisp-dm approach,” vol. 2, no. 2, pp. 93–104, 2026.
- [25] F. H. Wardhani dan K. M. Lhaksana, “Predicting Employee Attrition Using Logistic Regression With Feature Selection,” *Sinkron*, vol. 7, no. 4, pp. 2214–2222, Oct. 3, 2022. DOI: [10.33395/sinkron.v7i4.11783](https://doi.org/10.33395/sinkron.v7i4.11783)
- [26] S. Kurniawan, A. Nugroho, dan S. Suherman, “Analisis Faktor yang Mempengaruhi Promosi Karyawan Menggunakan Random Forest pada Dataset Employee Promotion,” *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 5, no. 2, pp. 177–187, Aug. 2, 2025. DOI: [10.55382/jurnalpustakaai.v5i2.1115](https://doi.org/10.55382/jurnalpustakaai.v5i2.1115)
- [27] V. M. A. Simalango dan W. F. Senjaya, “Evaluasi Algoritma Decision Tree dan Random Forest serta Efektivitas Feature Selection dalam Memprediksi Kesehatan Mental,” *Jurnal STRATEGI-Jurnal Maranatha*, vol. 7, no. 1, pp. 201–215, 2025.
- [28] J. R. Sari et al., “Evaluasi Model Klasifikasi dalam Deteksi Penipuan Transaksi: Studi Kasus pada Data Tidak Seimbang,” *Jurnal Gaussian*, vol. 14, no. 2, pp. 565–576, Dec. 10, 2025. DOI: [10.14710/j.gauss.14.2.565-576](https://doi.org/10.14710/j.gauss.14.2.565-576)
- [29] O. Y. Inonu, K. Magda, dan A. Amarudin, “Analisis Kinerja Algoritma Random Forest Dengan Model Machine Learning Pada Dataset Penyakit Diabetes,” *EXPERT: Jurnal Manajemen Sistem Informasi dan Teknologi*, vol. 15, no. 1, p. 1, Jun. 4, 2025. DOI: [10.36448/expert.v15i1.4312](https://doi.org/10.36448/expert.v15i1.4312)
- [30] A. Adrian dan I. Vewawati, “Analisis Performa Logistic Regression dan Random Forest dalam Klasifikasi Kelayakan Penerimaan Kredit,” *The Indonesian Journal of Computer Science Research*, vol. 4, no. 2, pp. 148–158, Jul. 23, 2025. DOI: [10.59095/ijcsr.v4i2.205](https://doi.org/10.59095/ijcsr.v4i2.205)

[Halaman ini sengaja dikosongkan.]