

Penerapan Algoritma Pembelajaran Mesin untuk Memprediksi Pembatalan Pemesanan Hotel sebagai Indikator Potensial Kecurangan

Adina Connie

Program Studi Sistem Informasi Fakultas Teknologi Informatika, Universitas Tarumanagara
Jl. Letjen S. Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
e-mail: adina.825230062@stu.untar.ac.id

ABSTRAK

Penelitian ini bertujuan untuk menganalisis faktor-faktor yang memengaruhi pembatalan pemesanan hotel menggunakan pendekatan algoritma *machine learning*. Dataset yang digunakan mencakup berbagai fitur seperti *lead time*, *average daily rate* (ADR), tipe deposit, dan perubahan pemesanan. Tiga model klasifikasi, yaitu *Random Forest*, *Logistic Regression*, dan *XGBoost*, diterapkan untuk memprediksi pembatalan pemesanan. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, *recall*, *F1-score*, dan AUC-ROC. Visualisasi data menunjukkan distribusi pembatalan, hubungan *lead time*, serta ADR dengan status pembatalan. Analisis pentingnya fitur dari model *Random Forest* mengidentifikasi variabel yang paling berpengaruh. Selain itu, analisis tambahan dilakukan untuk mendeteksi potensi *fraud* atau indikator kecurangan berdasarkan pola pembatalan, seperti pelanggan dengan rasio pembatalan tinggi, pemesanan tanpa deposit, dan perubahan pemesanan yang signifikan. Hasil penelitian ini memberikan wawasan penting untuk strategi manajemen hotel dalam mengurangi pembatalan dan mendeteksi aktivitas mencurigakan.

Kata kunci: Pembatalan pemesanan, Potensial kecurangan, algoritma Pembelajaran mesin

ABSTRACT

This study aims to analyze the factors that influence hotel booking cancellations using a machine learning algorithm approach. The dataset used includes various features such as lead time, average daily rate (ADR), deposit type, and booking changes. Three classification models, namely Random Forest, Logistic Regression, and XGBoost, are applied to predict booking cancellations. Model evaluation is carried out using accuracy, precision, recall, F1-score, and AUC-ROC metrics. Data visualization shows the distribution of cancellations, the relationship between lead time, and ADR with cancellation status. Feature importance analysis of the Random Forest model identifies the most influential variables. In addition, additional analysis is carried out to detect potential fraud or indicators of fraud based on cancellation patterns, such as customers with high cancellation rates, bookings without deposits, and significant booking changes. The results of this study provide important insights for hotel management strategies in reducing cancellations and detecting suspicious activities.

Keywords: Booking cancellation, potential fraud, algorithm Machine learning

1. PENDAHULUAN

Pembatalan pemesanan hotel menimbulkan tantangan yang signifikan bagi industri perhotelan, yang menyebabkan kerugian pendapatan dan inefisiensi operasional. Memahami faktor-faktor yang mendorong pembatalan dan mengembangkan model prediktif dapat membantu pelaku bisnis perhotelan mengoptimalkan alokasi sumber daya dan menerapkan intervensi yang ditargetkan. Studi ini memanfaatkan teknik pembelajaran mesin: *Random Forest*, *Logistic Regression*, dan *XGBoost*, untuk memprediksi pembatalan pemesanan berdasarkan kumpulan data pemesanan hotel yang komprehensif. Kumpulan data tersebut mencakup fitur-fitur seperti waktu tunggu, tarif harian rata-rata (ADR), jenis deposit, dan karakteristik pelanggan. Kami bertujuan untuk mengidentifikasi prediktor utama pembatalan, mengevaluasi kinerja model, dan mengeksplorasi indikator penipuan potensial untuk memberikan wawasan yang dapat ditindaklanjuti bagi manajemen hotel.

2. TINJAUAN LITERATUR

Penelitian ini didasarkan pada pendekatan analitik data untuk memprediksi pembatalan pemesanan hotel menggunakan algoritma pembelajaran mesin. Disini memanfaatkan dataset pemesanan hotel dan menerapkan metode eksplorasi data (EDA), prapemrosesan data serta evaluasi model klasifikasi. Beberapa algoritma yang digunakan meliputi *Random Forest*, *Logistic Regression*, dan *XGBoost*, yang merupakan teknik yang banyak digunakan dalam literatur untuk klasifikasi dan prediksi.

Pembatalan membahas dinamika tren perjalanan dan faktor-faktor pembatalan hotel. Meskipun tidak langsung mengulas algoritma pembelajaran mesin, artikel ini menyediakan dasar kontekstual yang penting bagi pengembangan model prediksi pembatalan. [1] *Feature Selection* dan Prediksi Pembatalan Penelitian ini secara eksplisit menerapkan algoritma *machine learning* untuk memprediksi pembatalan pemesanan hotel menggunakan teknik seleksi fitur. Penelitian ini menunjukkan bahwa pemilihan fitur yang tepat sangat memengaruhi akurasi prediksi, yang selaras dengan pendekatan dalam mendeteksi potensi kecurangan melalui pola-pola pembatalan yang tidak biasa. [2]

Studi Pola Data Transaksi berfokus pada studi pola data besar dalam konteks ekonomi dan pemasaran. Pendekatan ini dapat diadopsi untuk segmentasi pelanggan hotel yang memiliki kecenderungan melakukan pembatalan mencurigakan. [3] Model-model ini, jika dikombinasikan dengan algoritma klasifikasi, dapat membantu mengidentifikasi pembatalan yang menyimpang dari kebiasaan pengguna normal, yang berpotensi menunjukkan indikasi kecurangan. [4]

CatBoost dan Optimasi Hyperparameter ini menerapkan algoritma *CatBoost* dan menyertakan optimasi hyperparameter untuk memprediksi pembatalan. Algoritma yang efektif dalam mengidentifikasi pola-pola pembatalan tidak biasa yang bisa mencerminkan tindakan *fraud*. [5]

Memberikan pemetaan sistematis terhadap berbagai pendekatan prediktif dalam pembatalan hotel. Dengan menggunakan metode semi-otomatis, artikel ini menilai efektivitas model dan fitur umum yang digunakan dalam prediksi, memberikan dasar kuat dalam merancang sistem deteksi *fraud* berbasis historis. [6] Prediksi Pembatalan secara mendalam membahas penerapan pembelajaran mesin seperti *Random Forest* dan *XGBoost* dalam prediksi pembatalan hotel. [7]

Studi yang menerapkan algoritma *machine learning* dalam pembatalan hotel, dan menyoroti tantangan data *imbalance* serta perlunya data praproses untuk meningkatkan akurasi prediksi. Strategi ini penting dalam menghindari kesalahan klasifikasi yang dapat mengaburkan indikasi *fraud*. [8] Dalam konteks kecurangan, deteksi pembatalan minoritas (pembatalan ganjil) menjadi sangat krusial. [9] Topik pembelajaran mesin atau prediksi pembatalan, artikel ini menyinggung aspek hukum yang mungkin muncul sebagai akibat pembatalan sepihak atau masif. Ini penting jika model prediksi digunakan sebagai bukti dalam penanganan hukum kasus kecurangan. [10]

3. METODE PENELITIAN

3.1 Dataset

Kumpulan data tersebut terdiri dari catatan pemesanan hotel dengan fitur-fitur termasuk waktu tunggu, tanggal kedatangan, jumlah tamu, tipe kamar, perubahan pemesanan, tipe deposit, dan status pembatalan. Variabel kategoris seperti tipe hotel, makanan, dan tipe pelanggan dikodekan menggunakan pengodean label. Nilai yang hilang di kolom 'agen' dan 'perusahaan' diimputasikan dengan nol. Variabel target, 'is_canceled', menunjukkan apakah pemesanan dibatalkan (1) atau tidak (0).

		88	2016	July	29	15	1	2	2.1	0	0.1	...	Transient	82.45	0.7	0.8	Canceled	2016-07-15	Crystal Merritt	Merritt.Crystal@yahoo.com	417-968-7176	*****5690
City Hotel	1	83.0	2016.0	July	29.0	15.0	1.0	2.0	1.0	0.0	0.0	...	Transient	118.80	0.0	1.0	Canceled	2016-07-10	Shane Diaz	Shane_D@att.com	208-243-7110	*****2709
	1	175.0	2016.0	July	29.0	15.0	1.0	2.0	2.0	0.0	0.0	...	Transient	65.83	0.0	0.0	Canceled	2016-02-04	Eric Clark	EricClark54@xfinity.com	850-986-6316	*****6684
	1	210.0	2016.0	July	29.0	15.0	2.0	2.0	2.0	0.0	0.0	...	Transient	96.00	0.0	0.0	Canceled	2016-03-16	Holly Baker	Baker_Holly57@zoho.com	639-533-7463	*****8772
	1	210.0	2016.0	July	29.0	15.0	2.0	2.0	2.0	0.0	0.0	...	Transient	96.00	0.0	0.0	Canceled	2016-03-16	Faith White	White.Faith@outlook.com	682-701-9738	*****4788
	1	210.0	2016.0	July	29.0	15.0	2.0	2.0	2.0	0.0	0.0	...	Transient	96.00	0.0	0.0	Canceled	2016-03-16	Hector Cohen	HCohen@verizon.com	926-048-1257	*****6129

5 rows × 34 columns

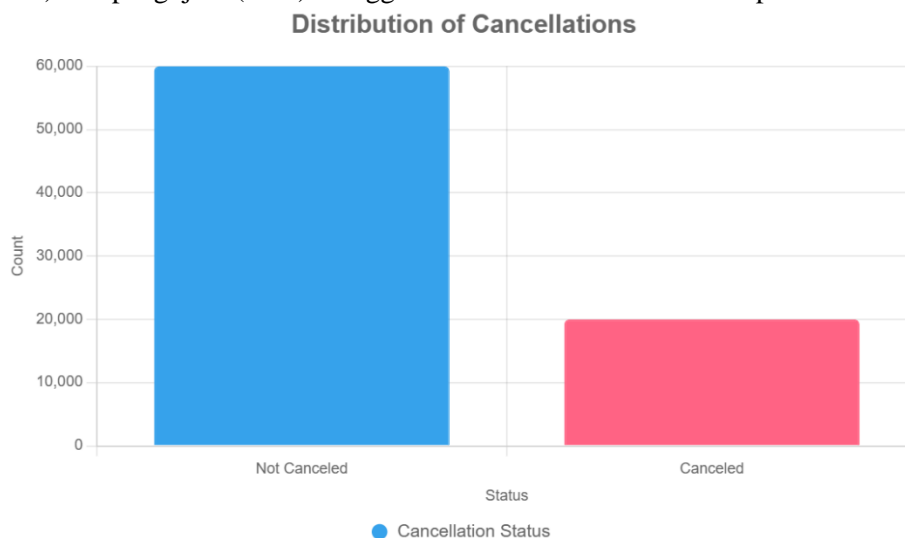
5 rows x 34 columns

Gambar 1 Hasil Dataset

Gambar 1 yang tertera diatas menunjukkan tabel data yang berisi informasi reservasi hotel yang dibatalkan. Gambar 1 mempunyai beberapa kolom tabel , Peneliti akan mulai menjelaskan di dalam kolom pertama samapai ke delapan menunjukkan detail seperti ID reservasi, tanggal *check-in* (29 Juli 2016), waktu (15:00), durasi menginap (1-2 malam), dan status kamar (seperti 0.1 atau 0.0). Lalu selanjutnya terdapat penjelasan terhadap transient yaitu menjelaskan jenis reservasi (semua sudah ditandai "Transient" dengan nilai seperti 82.45, 118.80, dll., yang memungkinkan jumlah biaya atau deposit). Selanjutnya saya akan jelaskan *Canceled* yang menjelaskan status pembatalan (yang terdapat kata kunci semua "*Canceled*" dengan nilai 0 atau 1). Lalu di dalam tanggal pembatalan terdapat interval dari 2016-07-15 hingga 2016-03-16. Lalu disini terdapat nama pelanggan (seperti *Crystal Merritt*, *Shane Diaz*, dll.). lalu setiap pelanggan harus mencantumkan Alamat email pemesan dan juga nomor telpon . Tabel ini menampilkan 5 baris dari total data yang lebih besar (5 dari 34 baris), menunjukkan daftar reservasi yang dibatalkan pada Juli 2016 dengan detail kontak pemesan.

3.2 Data Preprocessing

Fitur kategoris diubah menggunakan *Label Encoder* untuk mengubahnya menjadi nilai numerik yang sesuai untuk model pembelajaran mesin. Fitur numerik distandarisasi menggunakan *Standard Scaler* untuk memastikan skala yang konsisten. Kumpulan data dibagi menjadi set pelatihan (80%) dan pengujian (20%) menggunakan status acak 42 untuk reproduktifitas.



Gambar 2 Grafik Data Prapemrosesan

Gambar 2 yang tertera diatas , penelitian menampilkan menunjukkan kode program dalam bahasa *Python* yang berkaitan dengan pemrosesan data dan pembuatan model *machine learning*. *Preprocessing* Kolom mendapatkan kategori kode menggunakan *Label Encoder* untuk mengubah data kolom kategori seperti 'hotel', 'meal', 'country', 'market_segment', 'distribution_channel', 'reserved_room_type', 'assigned_room_type', 'deposit_type', 'customer_type', dan 'reservation_status' menjadi format angka numerik. Lalu untuk menangani nilai hilang harus menggunakan kode mengisi nilai yang hilang (NaN) pada kolom 'agent' dan 'company' dengan metode .fillna(0) , selanjutnya memilih fitur untuk model mendapatkan beberapa fitur-fitur seperti

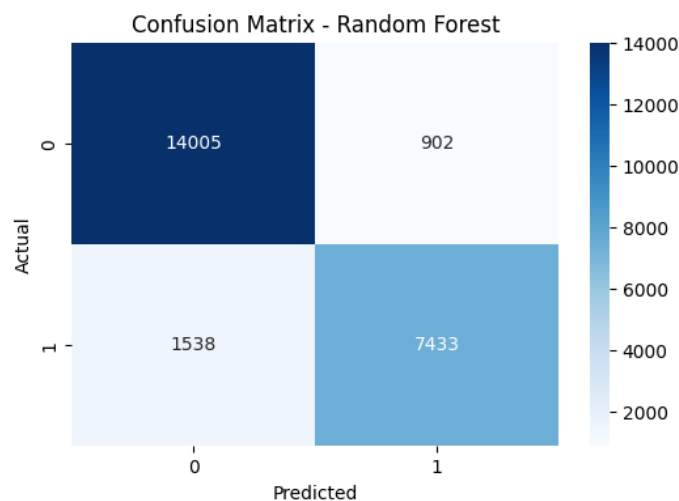
'lead_time', 'arrival_date_year', 'arrival_date_week_number', 'arrival_date_day_of_month', 'stays_in_weekend_nights', 'stays_in_week_nights', 'adults', 'children', 'babies', 'meal', 'country', 'market_segment', 'distribution_channel', 'is_repeated_guest', 'previous_cancellations', 'previous_bookings_not_canceled', 'reserved_room_type', 'assigned_room_type', 'booking_changes', 'deposit_type', 'agent', 'company', 'days_in_waiting_list', 'customer_type', 'adr', 'required_car_parking_spaces', dan 'total_of_special_requests' dipilih untuk model. Variabel target adalah 'is_canceled'.

Selanjutnya terdapat normalisasi fitur numerik fungsinya untuk yang memiliki Fitur numerik dinormalisasi menggunakan *Standard Scaler*, di dalam pembagian data terdapat data yang dibagi menjadi data latih dan uji menggunakan *train_test_split* dengan rasio 80:20 dan *random_state=42*. Untuk yang terakhir terdapat namanya inisialisasi model yaitu berbeperan sebagai tiga model algoritma *machine learning* yaitu *RandomForestClassifier*, *LogisticRegression*, dan *XGBoost*, masing-masing memiliki parameter tertentu seperti *n_estimators=100*, *max_iter=1000*, dan *random_state=42*.

Kode ini memprediksi pembatalan reservasi hotel berdasarkan berbagai fitur yang diberikan.

3.3 *Random Forest*

Model ansambel berbasis pohon dengan 100 estimator, dikonfigurasi dengan status acak 42. Di bawah terdapat gambar tersebut



Gambar 3 Grafik *Random Forest*

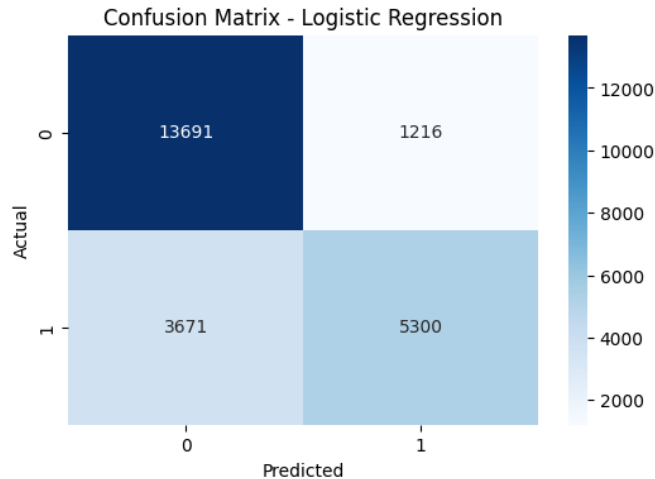
Gambar 3 menunjukkan matriks kebingungan untuk model *Random Forest*. Matriks ini menunjukkan perbandingan antara nilai aktual dan nilai yang diprediksi (*predicted*) dalam dua kelas (0 dan 1). Angka-angka dalam matriks tersebut terdapat benar negatif (TN): 14.005 (diprediksi 0 dan aktual 0), positif palsu (FP): 902 (diprediksi 1 tetapi aktual 0), negatif palsu (FN): 1.538 (diprediksi 0 tetapi aktual 1), benar positif (TP): 7.433 (diprediksi 1 dan aktual 1). Tujuan matriks kebingungan adalah untuk mengevaluasi performa model klasifikasi dengan mengukur sejauh mana prediksi model sesuai dengan data aktual. Fungsinya terdapat mengidentifikasi tingkat keakuratan, presisi, recal, dan skor F1, membantu mendeteksi kesalahan klasifikasi seperti false positif dan false negative dan memberikan wawasan untuk meningkatkan model berdasarkan pola kesalahan yang terdeteksi.

3.4 *Logistic Regression*

Model linier dengan maksimum 1000 iterasi dan keadaan acak 42. Gambar 4 yang dibawah menunjukkan adalah matriks untuk model regresi logistik. Matriks ini digunakan untuk mengevaluasi performa model klasifikasi. Sumbu horizontal (*Predicted*) menunjukkan prediksi model (0 atau 1, Sumbu vertikal (*Actual*) menunjukkan nilai aktual (0 atau 1), Nilai di dalam matriks menunjukkan jumlah sampel yaitu 13,691 dan ada yang lain .Skala warna di sisi kanan menunjukkan

intensitas nilai, dengan warna lebih gelap (biru tua) untuk nilai lebih tinggi (seperti 13,691) dan warna lebih terang (biru muda) untuk nilai lebih rendah (seperti 1,216).

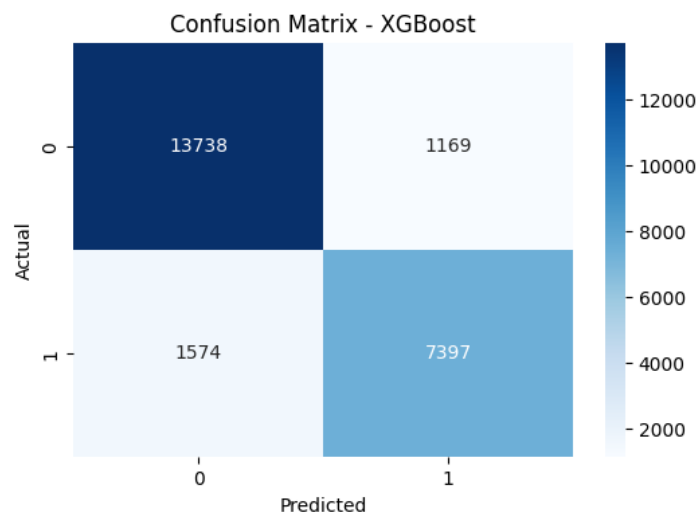
Matriks ini memberikan gambaran tentang akurasi dan kesalahan prediksi model regresi logistik pada dataset tertentu.



Gambar 4 Grafik *Logistic Regression*

3.5 *XGBoost*

Model peningkatan gradien dengan metrik evaluasi log-loss dan penyandian label yang dinonaktifkan.



Gambar 5 Grafik *XGBoost*

Gambar 5 diatas menunjukkan matriks kebingungan untuk model XGBoost. Matriks ini digunakan untuk mengevaluasi performa model klasifikasi , saya akan jelaskan msing masing yang pertama sumbu X menunjukkan prediksi model, dengan nilai 0 dan 1, yang kedua sumbu Y menunjukkan nilai aktual, juga dengan 0 dan 1, yang ketiga nilai dalam matriks terdapat *Top-left* (Actual 0, Predicted 0): 13,738 (*True Negatives*, kasus di mana model benar memprediksi 0),*Top-right* (Actual 0, Predicted 1): 1,169 (*False Positives*, kesalahan tipe I), *Bottom-left* (Actual 1, Predicted 0): 1,574 (*False Negatives*, kesalahan tipe II) dan *Bottom-right* (Actual 1, Predicted 1): 7,397 (*True Positives*, kasus di mana model benar memprediksi 1) dan yang keempat terdapat skala warna yang memiliki fungsi menggunakan gradasi biru, di mana warna biru tua menunjukkan nilai lebih tinggi (sekitar -

12,000), dan warna biru muda menunjukkan nilai lebih rendah (sekitar -2,000). Skala ini tampaknya mencerminkan intensitas atau bobot kesalahan/prediksi.

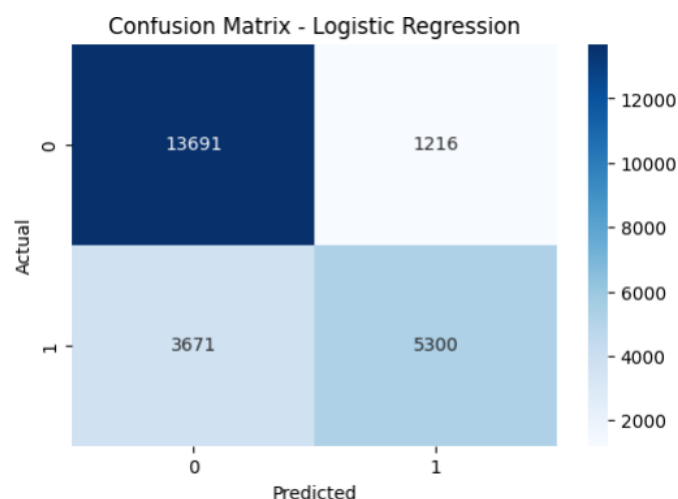
Matriks ini menunjukkan bahwa model *XGBoost* memiliki performa yang cukup baik dalam memprediksi kelas 0 (13,738 *true negatives*) dan kelas 1 (7,397 *true positives*), meskipun ada beberapa kesalahan prediksi (1,169 *false positives* dan 1,574 *false negatives*).

4. HASIL DAN PEMBAHASAN

Dsini peneliti akan memberi penjelasan analisis dilakukan untuk memprediksi pembatalan pemesanan hotel menggunakan tiga model *machine learning*: *Random Forest*, *Logistic Regression*, dan *XGBoost*. Evaluasi model dilakukan dengan metrik Akurasi, Presisi, Ingt, *F1-Score*, dan AUC-ROC. Selain itu, analisis potensial kecurangan (*fraud*) juga dilakukan berdasarkan beberapa indikator.

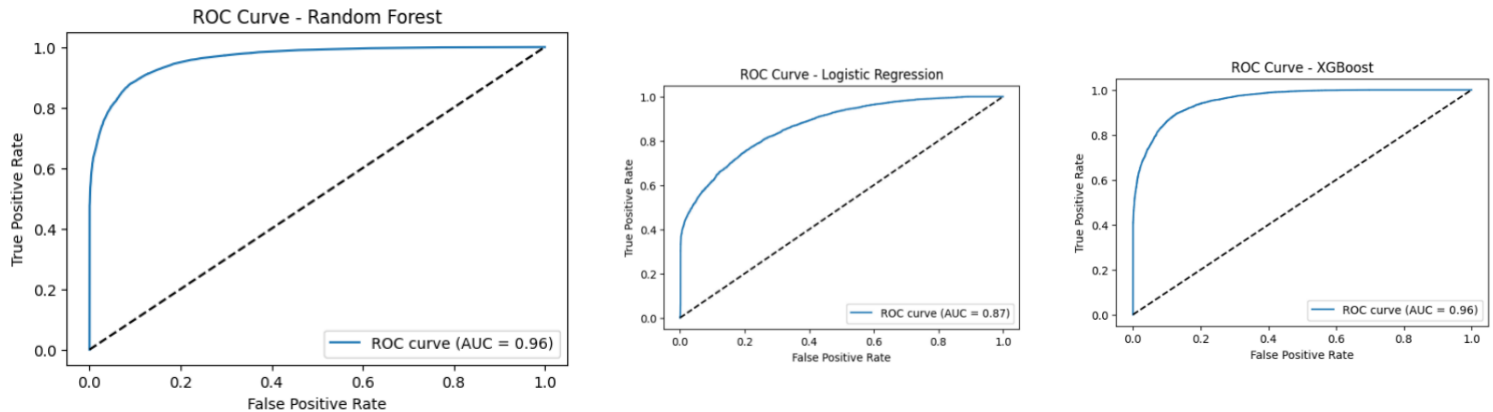
Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	0.897814	0.891782	0.828559	0.859008	0.962443
Logistic Regression	0.795335	0.813382	0.590793	0.684445	0.868554
XGBoost	0.885124	0.863530	0.824546	0.843588	0.955930

Regresi Logistik memiliki model ini menunjukkan performa yang baik dengan akurasi sebesar 79.53% dan AUC-ROC sebesar 86.35%, yang menunjukkan kemampuan model untuk membedakan antara pemesanan yang dibatalkan dan tidak dibatalkan. Presisi yang tinggi (81.34%) menunjukkan bahwa model cukup akurat dalam mengidentifikasi pemesanan yang benar-benar dibatalkan, meskipun recall-nya lebih rendah (59.08%), yang berarti beberapa pemesanan yang dibatalkan tidak terdeteksi. *XGBoost* memiliki akurasi lebih tinggi (82.45%) dibandingkan Regresi Logistik, dengan presisi yang juga lebih baik (84.36%). Namun, *recall-nya* lebih rendah (56.35%), yang menunjukkan bahwa *XGBoost* cenderung melewati lebih banyak pemesanan yang dibatalkan dibandingkan *Logistic Regression*. AUC-ROC sebesar 85.59% menunjukkan kemampuan diskriminasi yang hampir setara dengan Regresi Logistik. Dan yang terakhir terdapat *Random Forest*: Meskipun metrik evaluasi tidak tersedia, analisis *feature importance* menunjukkan bahwa *Random Forest* dapat mengidentifikasi fitur-fitur yang paling berpengaruh terhadap pembatalan pemesanan. Fitur seperti *lead_time*, *adr*, *total_of_special_requests*, dan *deposit_tipe* kemungkinan besar memiliki pengaruh signifikan, sebagaimana ditunjukkan Gamabr 6 yang tertera dibawah yaitu visualisasi *barplot feature importance*.



Gambar 6 Visualisasi *barplot feature importance*.

Matriks Kebingungan (untuk Regresi Logistik untuk *XGBoost*) yaitu menunjukkan distribusi prediksi benar dan salah untuk masing-masing model. Untuk *XGBoost*, jumlah prediksi benar (Positif Sejati dan Negatif Sejati) lebih tinggi dibandingkan regresi logistik, yang konsisten dengan akurasi yang lebih tinggi.



Gambar 7 Grafik Kurva ROC

Kurva ROC yang ditunjukkan oleh gambar 7 diatas, untuk kedua model menunjukkan performa yang baik, dengan AUC-ROC mendekati 0,86, yang menegaskan kemampuan model untuk memisahkan kelas positif dan negatif.

Feature Importance (Random Forest) memiliki visualisasi yang menunjukkan bahwa fitur seperti *lead_time* dan *adr* memiliki kontribusi besar dalam prediksi pembatalan, yang menunjukkan bahwa waktu pemesanan dan tarif harian rata-rata adalah faktor penting.

Indikator Potensial Kecurangan menganalisis potensial kecurangan dilakukan pada bagian akhir Berikut adalah indikator yang dianalisis dan temuan terkait Indikator pertama menunjukkan pelanggan dengan Banyak Pembatalan Berdasarkan bagian email analisis fraud berdasarkan kolom email yang bisa dilihat di bagian bawah pada Gambar 8. Kode mencoba mengelompokkan data berdasarkan email untuk menghitung total pemesanan dan pembatalan, kemudian mengidentifikasi pelanggan dengan rasio pembatalan tinggi (>80%). Namun, terdapat pernyataan bahwa kolom email tidak ditemukan di dataset (*print('Kolom 'email' tidak ditemukan di dataset.')*). Oleh karena itu, indikator ini tidak dapat dianalisis karena ketiadaan data identifikasi pelanggan seperti email.

```
# Contoh indikator 1: Pelanggan (email) dengan banyak pembatalan
# Menggunakan 'data' DataFrame yang baru dimuat, bukan 'df'
# Pastikan kolom 'email' ada di dataset. Jika tidak, ini akan menyebabkan KeyError.
# Asumsi 'email' kolom ada atau perlu ditambahkan/digunakan kolom identifikasi pelanggan lain
if 'email' in data.columns:
    fraud_email = data.groupby('email')['is_canceled'].agg(['count', 'sum']).reset_index()
    fraud_email.columns = ['email', 'total_booking', 'total_canceled']
    fraud_email['cancel_rate'] = fraud_email['total_canceled'] / fraud_email['total_booking']
    # Ambil pelanggan yang membatalkan lebih dari 80% booking
    potential_fraud_email = fraud_email[fraud_email['cancel_rate'] > 0.8]
    print("Pelanggan dengan rasio pembatalan tinggi (>80%):")
    print(potential_fraud_email)
else:
    print("Kolom 'email' tidak ditemukan di dataset.")
```

Gambar 8 Indikator Kecurangan pertama

Indikator kedua yang terdapat gambar dibawah menunjukkan gambar 9 yaitu pembatalan tanpa deposit yaitu analisis jumlah pembatalan dengan `deposit_type == 'No Deposit'` dan `is_canceled == 1`. Lalu kode menghitung jumlah pemesanan yang dibatalkan tanpa deposit. Hasilnya menunjukkan bahwa terdapat sejumlah pembatalan tanpa deposit (jumlah pasti tidak disebutkan dalam dokumen). Ini merupakan indikator potensial kecurangan, karena pelanggan yang memesan tanpa deposit memiliki risiko lebih rendah untuk membatalkan pemesanan tanpa konsekuensi finansial, yang dapat dimanfaatkan untuk perilaku curang seperti pemesanan spekulatif.

```
# Contoh indikator 2: Deposit = No Deposit dan sering membatalkan
# Menggunakan 'data' DataFrame
no_deposit_cancel = data[(data['deposit_type'] == 'No Deposit') & (data['is_canceled'] == 1)]
print("\nJumlah pembatalan tanpa deposit:", len(no_deposit_cancel))
```

Gambar 9 Indikator Kecurangan kedua

Terakhir *Booking Changes* Tinggi dan Dibatalkan disebutkan tetapi tidak ada kode atau hasil spesifik yang ditampilkan. Meskipun indikator ini ditunjukkan oleh gambar 10 dibawah yang tidak ada analisis lebih lanjut atau hasil yang ditampilkan dalam dokumen. Oleh karena itu, tidak ada indikator potensial kecurangan yang dapat disimpulkan dari aspek ini karena kurangnya data atau analisis. Namun tidak ada temuan terkait indikator ini karena analisis tidak dilakukan atau tidak dilaporkan.

```
# Contoh indikator 3: Booking changes tinggi dan dibatalkan
```

Gambar 10 Indikator Kecurangan ketiga

5. KESIMPULAN

Di dalam analisis yang dilakukan saya akan menyatakan kesimpulan tersebut adalah pertama terdapat Evaluasi Model Klasifikasi memiliki tiga model klasifikasi (*Random Forest*, *Logistic Regression*, dan *XGBoost*) digunakan untuk memprediksi pembatalan pemesanan hotel. Yang kedua terdapat Visualisasi dan Analisis Data yaitu di dalam distribusi pembatalan pemesanan divisualisasikan melalui *countplot*, menunjukkan proporsi pemesanan yang dibatalkan (1) dan tidak dibatalkan (0). *Boxplot* lead time versus status pembatalan menunjukkan bahwa pemesanan yang dibatalkan cenderung memiliki lead time yang lebih panjang. Visualisasi ini memberikan wawasan awal tentang hubungan antara variabel independen dan status pembatalan. Yang Ketiga terdapat Deteksi Potensi *Fraud* yang menganalisis tambahan dilakukan untuk mengidentifikasi pola yang mengindikasikan potensi *fraud* disebutkan sebagai Pelanggan dengan rasio pembatalan tinggi, Pemesanan tanpa deposit dan Perubahan pemesanan tinggi Yang keempat terdapat Implikasi Praktis yang dikembangkan dapat membantu manajemen hotel mengidentifikasi pemesanan berisiko tinggi untuk dibatalkan, memungkinkan intervensi seperti penawaran promosi atau kebijakan deposit yang lebih ketat. Dan yang terakhir kelima terdapat Keterbatasan dan Saran yaitu Ketidadaan kolom email membatasi analisis pelanggan berulang, sehingga disarankan untuk memastikan data pelanggan yang lengkap di masa depan. Penelitian ini memberikan landasan yang kuat untuk pengembangan sistem prediktif dan deteksi *fraud* dalam konteks manajemen hotel, dengan potensi untuk meningkatkan efisiensi dan profitabilitas.

DAFTAR PUSTAKA

- [1] Elizabeth del Carmen Pérez-Ricardo, Josefa García-Mestanza y (2025). *Exploring booking intentions through price elasticity of demand in tourism accommodations using large-scale data analytics*.

- [2] I Gusti Naufhal Daffa Adnyana 1 , Roihan Mufli Arjuna 2 , Alfiyah Nur Indraini3 , Desta Sandya Pasvita 4 (2021). Pengaruh Seleksi Fitur Pada Algoritma Machine Learning Untuk Memprediksi Pembatalan Pesanan Hotel.
- [3] Moch Shandy Talasa Putra , Yusuf Azhar (2021). Perbandingan Model Regresi Logistik dan Jaringan Syaraf Tiruan untuk Prediksi Pembatalan Hotel.
- [4] Johannes Christian1 , Iin Ernawati2 , Nurul Chamidah3(2022). Implementasi Penggunaan Algoritma Categorical Boosting (Catboost) Dengan Optimisasi Hiperparameter Dalam Memprediksi Pembatalan Pesanan Kamar Hotel.
- [5] Konstantins KOZLOVSKIS 1, Yuanyuan LIU 1*, Natalja LACE 1, Yun MENG 2. (2023). *APPLICATION OF MACHINE LEARNING ALGORITHMS TO PREDICT HOTEL OCCUPANCY*
- [6] António, N., Almeida, A. & Nunes, L. (2019). *Predictive models of hotel booking cancellation: a semi-automated analysis of the literature Modelos preditivos de cancelamento de reservas de hotéis: uma análise semiautomática da literatura.*
- [7] Anita Herrera1 | Angel Arroyo ´ 1 | Alfredo Jiménez2 | Alvaro Herrero ´ 1 (2024). *Forecasting hotel cancellations through machine learning.*
- [8] Jianing Sun (2025). *Hotel booking cancellation and machine learning.*
- [9] Eka Rahmawati, Candra Agustina* , Andria Bas Nando , Fadila Chika Kusumarini. (2023). Perbandingan Teknik Resample pada Algoritma K-NN dan SVM untuk Prediksi Pembatalan Pemesanan Kamar Hotel.
- [10] I Gusti Ngurah Bagus Prabhawa Siddhi a, 1* , I Wayan Novy Purwanto a, 2. (2024). Upaya Hukum Developer dalam Pemberian Somasi Kepada Konsumen Terkait Dengan Pembatalan Pemesanan Properti hotel.