Analisis Eksploratif Terhadap Faktor Kunci Pada Dataset Pemesanan Hotel

Thoriq Hidayansyah[#1], Setia Budi[*2]

[#] *Program Studi S1 Teknik Informatika, Fakultas Teknologi dan Rekayasa Cerdas, Universitas Kristen Maranatha*
Jalan Prof. Drg. Surya Sumantri No. 65, Bandung 40164, Indonesia

[1]2172019@maranatha.ac.id

[2]setia.budi@it.maranatha.edu

Abstract — **This research titled "Exploratory Analysis of Key Factors in Hotel Booking Dataset" aims to apply various data analysis and visualization methods to enhance skills in data exploration, preprocessing, and deriving insights from real-world datasets. The dataset used includes comprehensive information regarding hotel types, booking time, cancellations, and other relevant variables.**

The study employs data exploration techniques such as data cleaning, Exploratory Data Analysis (EDA), and in-depth analysis of correlation matrices, cancellation rates, stay duration, Average Daily Rate (ADR), seasonal patterns, room distribution, customer characteristics, guest numbers, and food package preferences. Additionally, machine learning models including Logistic Regression, Random Forest, and Multilayer Perceptron are utilized to predict booking cancellations.

The research reveals several key findings: lead time positively correlates with cancellation rates (+0.18), indicating that bookings with longer lead times are more likely to be canceled; Online Travel Agent customers show the highest cancellation rates; summer in Portugal sees the highest peak in bookings and cancellations; and Bed & Breakfast (BB) is the most preferred food package across all market segments. Logistic Regression outperforms Random Forest in recall performance, while Multilayer Perceptron achieves the highest overall recall.

This study provides valuable insights for the hospitality industry in understanding customer behavior, reducing cancellation risks, and optimizing operational strategies. The findings also serve as a means to enhance technical competency in applying data science methods to real-world cases.

Keywords— **Cancellation Analysis, Stay Duration, Data Exploration, Lead Time, Machine Learning Models, Seasonal Patterns**

I. PENDAHULUAN

Industri hotel merupakan salah satu sektor yang sangat beragam, mencakup jaringan multinasional besar hingga operator kecil independen. Meskipun terdapat banyak variasi dalam skala operasional, industri ini juga menunjukkan konsentrasi yang signifikan, dengan sejumlah perusahaan besar mendominasi pasar. Brand-brand besar seperti Marriott International, Hilton Worldwide, dan Accor bukanlah operator langsung dari hotel mereka, melainkan bekerja dengan sistem waralaba dan memberikan lisensi kepada pemilik hotel. Namun, mereka tetap bertanggung jawab dalam memastikan standar kualitas, konsistensi pengalaman tamu, dan penyediaan platform pemesanan yang efisien. Di samping itu, para pemilik hotel sering kali merupakan investor atau lembaga keuangan yang menggunakan dana investor untuk memiliki properti tersebut, sementara operator hotel mengelola operasional harian.

Seiring dengan kemajuan teknologi, khususnya di era digital, industri perhotelan mengalami transformasi digital yang mendalam. Digitalisasi telah memperkenalkan inovasi baru yang berdampak besar pada cara hotel beroperasi, dengan semakin banyaknya penggunaan teknologi digital yang memungkinkan pengolahan dan analisis data dalam jumlah besar. Hotel-hotel mulai mengadopsi PMS berbasis cloud untuk mengelola operasional, memantau pemesanan, dan meningkatkan efisiensi pelayanan. Selain itu, munculnya OTAs dan situs perbandingan harga memperkuat persaingan, yang berujung pada peningkatan tingkat hunian serta perubahan strategi penetapan harga [1].

Namun, dengan pesatnya perkembangan teknologi, tantangan baru muncul, terutama dalam menentukan strategi teknologi informasi yang tepat. Banyak operator hotel berusaha mencari keunggulan kompetitif melalui penerapan teknologi baru, termasuk penggunaan metode data science untuk menganalisis pola pemesanan dan pembatalan. Proses ini disebut sebagai transformasi digital, yang menurut Vial (2019) adalah proses untuk meningkatkan entitas dengan memicu perubahan signifikan melalui kombinasi teknologi informasi, komputasi, komunikasi, dan konektivitas [1].

Penelitian ini bertujuan untuk menganalisis berbagai faktor yang memengaruhi kinerja hotel menggunakan dataset pemesanan hotel. Melalui analisis ini, diharapkan dapat diidentifikasi faktor-faktor signifikan yang memengaruhi keputusan pelanggan dan dapat digunakan untuk memprediksi pembatalan pemesanan secara lebih akurat.

Data yang digunakan dalam penelitian ini merupakan dataset pemesanan hotel yang terdiri dari berbagai informasi terkait pemesanan hotel, seperti waktu pemesanan, segmen pasar, kanal distribusi, jenis hotel, durasi menginap, dan berbagai fitur lainnya. Dengan 32 kolom informasi yang mencakup data historis pemesanan, dataset ini menyediakan gambaran yang lengkap tentang operasional hotel dan perilaku pelanggan.

Penelitian ini berfokus pada beberapa aspek penting dalam operasional hotel, termasuk analisis pembatalan, durasi menginap, harga harian rata-rata, dan yang lainnya. Dengan menggunakan metode Analisis Deskriptif, penelitian ini akan mengeksplorasi berbagai hubungan yang mungkin tersembunyi dalam data.

Selain itu, penelitian ini juga akan membangun model prediksi pembatalan menggunakan teknik pembelajaran mesin Logistic Regression dan Random Forest. Model ini diharapkan mampu membantu hotel memprediksi kemungkinan pembatalan di masa depan, sehingga mereka dapat merespon lebih proaktif dengan strategi yang tepat.

II. KAJIAN TEORI

A. Data Science

Data science merupakan bidang yang melibatkan pengolahan data untuk menemukan jawaban atas pertanyaan yang relevan. Sebagai disiplin ilmu yang mulai dikenal sejak tahun 80-an dan 90-an, data science secara resmi dipublikasikan sekitar tahun 2009 hingga 2011. Para pionir dalam bidang ini antara lain Andrew Gelman dan DJ Patil.

Secara sederhana, data science adalah proses untuk mengeksplorasi, menganalisis, dan memanipulasi data guna memperoleh insights atau wawasan yang bermanfaat. Fokus utama data science adalah pada data itu sendiri dan pemanfaatan berbagai ilmu dan teknologi untuk menggali informasi penting. Insights yang dihasilkan dari proses ini, meskipun mungkin kecil, sangat berharga dalam mendukung pengambilan keputusan yang lebih baik.

Data science digunakan untuk menjawab pertanyaan melalui proses pengolahan data yang dapat berukuran sedang hingga besar. Seorang data scientist bertugas untuk mengolah data tersebut dengan rasa ingin tahu yang kuat atau berdasarkan kebutuhan dari organisasi tempat mereka bekerja, dan menggunakan berbagai alat serta metode yang tepat untuk mengungkap wawasan yang tersembunyi dalam data.

Tujuan akhir dari data science adalah menemukan insights dari data, baik berupa informasi penting atau model yang berguna dalam penyelesaian masalah dan pengambilan keputusan di berbagai sektor, seperti industri, perdagangan, transportasi, layanan, kesehatan, pendidikan, dan lain-lain [3].

B. Python

Python merupakan salah satu bahasa pemrograman yang sangat populer dan kuat, terutama sejak diperkenalkan pada tahun 1991. Sebagai bahasa yang bersifat interpreted dan sering digunakan dalam scripting, Python memiliki fleksibilitas tinggi untuk membangun berbagai aplikasi, termasuk untuk otomatisasi tugas dan pengembangan perangkat lunak. Walaupun pada awalnya Python dipandang sebagai bahasa pemrograman eksperimental, dalam 20 tahun terakhir Python telah menjadi bahasa yang penting dalam komputasi ilmiah dan analisis data.

Popularitas Python dalam analisis data dan visualisasi dipengaruhi oleh dukungan komunitas ilmiah yang besar, serta pengembangan berbagai open-source libraries seperti pandas dan scikit-learn, yang sangat memudahkan pengolahan data. Python juga bersaing dengan bahasa lain seperti R, MATLAB, dan SAS dalam ekosistem analisis data. Fleksibilitas Python dalam mengintegrasikan kode dari bahasa seperti C, C++, dan FORTRAN membuatnya semakin banyak digunakan untuk scientific computing, terutama dalam memanfaatkan pustaka legacy yang digunakan untuk komputasi numerik.

Dengan kekuatan Python dalam rekayasa perangkat lunak secara umum, serta kemampuannya dalam analisis data, Python telah berkembang menjadi pilihan utama untuk berbagai aplikasi analisis data di akademisi dan industri [4].

C. NumPy

NumPy adalah pustaka Python yang muncul sebagai solusi untuk dua paket array sebelumnya, yaitu Numeric dan Numarray. Numeric, yang dikembangkan pada pertengahan 1990-an, menyediakan objek array dan fungsi yang efisien dengan kecepatan tinggi, ditulis dalam C dan terhubung dengan implementasi aljabar linier yang cepat. Sementara itu, Numarray dirancang untuk menangani citra besar dari Teleskop Luar Angkasa Hubble, menambahkan dukungan untuk array terstruktur dan pengindeksan fleksibel. Meskipun keduanya kompatibel, perbedaan di antara mereka memecah komunitas pengguna.

Pada tahun 2005, NumPy muncul sebagai unifikasi dari kedua paket tersebut, menggabungkan fitur-fitur terbaik dari Numeric dan Numarray. Sejak saat itu, NumPy telah menjadi dasar bagi hampir semua pustaka Python yang berfokus pada komputasi ilmiah dan numerik, seperti SciPy, Matplotlib, pandas, scikit-learn, dan scikit-image. NumPy menyediakan objek array multidimensional dan fungsi yang beroperasi pada array tersebut, menjadikannya format pertukaran data yang umum digunakan.

NumPy beroperasi pada array yang disimpan dalam memori menggunakan central processing unit (CPU). Saat ini, dengan munculnya banyak paket array baru, NumPy berfungsi sebagai mekanisme koordinasi yang menetapkan API pemrograman array terdefinisi dengan baik, mendistribusikannya ke implementasi yang lebih khusus untuk akses teknologi baru [5].

D. Pandas

Pandas merupakan pustaka Python yang berperan penting dalam pengolahan dan analisis data. Sebagai pustaka open-source, Pandas sangat fleksibel dan dapat digunakan secara luas, baik untuk data numerik maupun teks. Pandas adalah pengembangan dari pustaka NumPy, yang memiliki keterbatasan dalam penanganan data dalam bentuk larik berdimensi, khususnya pada relasi data yang masih memerlukan indeks penanda. Pandas mengadopsi proses operasi numerik dari NumPy, namun memperluasnya dengan struktur data yang lebih kompleks dan beragam.

Pandas menyediakan tiga jenis struktur data utama: Series, DataFrame, dan Panel. Series adalah struktur satu dimensi yang berisi kumpulan data dalam satu atribut, sedangkan DataFrame adalah gabungan dari beberapa Series yang terhubung dan membentuk tabel dua dimensi yang serupa dengan matriks, dengan minimal dua atribut dan satu perekaman. Panel adalah representasi data tiga dimensi yang terdiri dari beberapa DataFrame yang saling berhubungan.

Dengan berbagai tipe struktur data ini, Pandas memungkinkan manipulasi data yang lebih efisien dan mudah, serta dapat menangani berbagai jenis data dengan lebih tepat guna dibandingkan dengan NumPy. Pandas menjadi alat yang sangat esensial dalam data analysis dan data science karena kemampuannya yang fleksibel dan efisien dalam menangani dan memproses data [6].

E. Descriptive Statistics

Descriptive statistics adalah metode statistik yang digunakan untuk memberikan ringkasan yang terperinci tentang pengamatan atau sampel data. Statistik deskriptif dapat bersifat kuantitatif, seperti statistik ringkasan (mean, median, modus, persentil, nilai maksimal dan minimal), atau bersifat visual, seperti grafik dan plot. Teknik ini membentuk dasar dari analisis data yang lebih kompleks dan memberikan wawasan awal mengenai kumpulan data yang sedang dianalisis.

Dalam dunia bisnis, statistik deskriptif membantu dalam merangkum berbagai jenis data yang besar dan kompleks. Sebagai contoh, pemasar dan tenaga penjualan dapat menggunakan pola pembelian dan pengeluaran historis pelanggan untuk membuat keputusan produk yang lebih baik. Dengan menerapkan analisis deskriptif sederhana pada data ini, mereka dapat mengidentifikasi tren dan pola perilaku yang relevan.

Statistik deskriptif dapat dibagi menjadi tiga jenis analisis utama, yaitu univariat, bivariat, dan multivariat. Analisis univariat melibatkan satu variabel dan bertujuan untuk menganalisis distribusi serta karakteristik dasar dari variabel tersebut. Analisis bivariat melibatkan dua variabel dan bertujuan untuk mempelajari hubungan atau korelasi antara kedua variabel tersebut. Sedangkan analisis multivariat melibatkan lebih dari dua variabel, dan digunakan untuk memahami interaksi yang lebih kompleks antar variabel.

Penggunaan statistik deskriptif sangat penting dalam membantu analis atau pengambil keputusan untuk mendapatkan pemahaman yang lebih baik tentang data yang mereka miliki, sehingga mereka dapat mengidentifikasi pola dan tren yang mungkin tersembunyi, serta mendukung pembuatan keputusan yang lebih baik di berbagai bidang, termasuk bisnis, pemasaran, dan penelitian ilmiah [7].

F. Data Cleaning

Data Cleaning adalah proses untuk memperbaiki masalah sistematis atau kesalahan pada data yang kacau atau "messy data". Proses ini melibatkan identifikasi dan penanganan pengamatan yang salah, yang membutuhkan pemahaman mendalam dari domain terkait. Banyak alasan yang menyebabkan data memiliki nilai yang tidak benar, seperti salah ketik, data yang rusak, duplikasi, dan sebagainya. Pengalaman di bidang terkait memungkinkan pengamatan yang salah teridentifikasi karena berbeda dari apa yang diharapkan, contohnya tinggi seseorang yang tercatat setinggi 200 kaki.

Setelah data yang kacau, bising, korup, atau keliru teridentifikasi, data tersebut dapat diperbaiki. Perbaikan ini bisa dilakukan dengan menghapus baris atau kolom yang tidak sesuai, atau mengganti pengamatan yang salah dengan nilai yang benar. Operasi pembersihan data yang umum meliputi mengidentifikasi outliers dengan menggunakan statistik untuk mendefinisikan data normal, menghapus kolom yang tidak bervariasi atau memiliki nilai yang sama tanpa perbedaan yang berarti, menghapus baris data yang duplikat, menandai nilai yang kosong sebagai nilai hilang (missing values), dan mengisi nilai hilang dengan menggunakan statistik atau model yang telah dipelajari.

Proses data cleaning merupakan langkah pertama yang harus dilakukan sebelum mempersiapkan data untuk operasi analisis data yang lebih kompleks. Dengan membersihkan data, analisis yang dilakukan akan lebih akurat dan dapat diandalkan [8].

G. Exploratory Data Analysis

Exploratory Data Analysis (EDA) adalah langkah awal yang sangat penting dalam proses penemuan pengetahuan pada data, di mana seorang data scientist secara interaktif mengeksplorasi dataset yang belum dikenal melalui serangkaian operasi analisis, seperti filter, agregasi, dan visualisasi. EDA memainkan peran kunci dalam memahami struktur data, mengidentifikasi pola, dan menemukan anomali atau wawasan awal yang relevan dengan analisis lebih lanjut. EDA sering dianggap sebagai tugas yang menantang karena memerlukan keterampilan analitis yang mendalam, pengalaman dalam pengolahan data, serta pengetahuan mendalam di bidang terkait.

Selama dekade terakhir, berbagai sistem telah dikembangkan untuk mempermudah pelaksanaan EDA, terutama seiring dengan kemajuan dalam penelitian machine learning. Teknologi ini tidak hanya membantu mempermudah EDA tetapi juga menciptakan peluang baru untuk mengotomatisasi proses tersebut. Sistem rekomendasi telah muncul sebagai salah satu pendekatan yang membantu data scientist dalam menentukan langkah eksplorasi berikutnya, seperti tindakan filter atau agregasi yang tepat. Selain itu, metode seperti k-Nearest Neighbors (kNN) dan active learning digunakan untuk memprediksi preferensi pengguna dalam hal ketertarikan terhadap pola atau fitur tertentu dalam data.

Penelitian terbaru bahkan telah mendorong pengembangan metode yang sepenuhnya mengotomatisasi EDA, misalnya dengan menggunakan deep reinforcement learning dan model sequence-to-sequence. Ini membuka jalan bagi upaya pengurangan upaya manual yang signifikan dalam melakukan EDA, terutama dalam proses eksplorasi data yang kompleks. Meski begitu, masih terdapat berbagai tantangan dan pertanyaan yang harus dijawab untuk dapat sepenuhnya menggantikan peran manusia dalam proses EDA, seperti cara menangani kompleksitas data dan preferensi pengguna yang berubah-ubah [9].

H. Correlation Matrix

Correlation Matrix adalah representasi matematis hubungan antara variabel dalam bentuk matriks, yang digunakan untuk mengevaluasi hubungan linear antara pasangan variabel. Elemen-elemen dalam matriks ini menunjukkan koefisien korelasi, yang berkisar antara -1 dan 1, dengan nilai 1 menunjukkan hubungan linear positif sempurna, nilai -1 menunjukkan hubungan linear negatif sempurna, dan nilai 0 menunjukkan tidak adanya hubungan linear [10].

I. Machine Learning and Deep Learning

Machine Learning adalah cabang kecerdasan buatan (AI) yang memungkinkan sistem untuk belajar dari data tanpa pemrograman eksplisit. Proses ini melibatkan pembuatan model analitik yang dapat memprediksi, mengklasifikasi, atau memberikan rekomendasi berdasarkan data. ML terdiri dari beberapa jenis utama, yaitu Supervised Learning, di mana sistem dilatih menggunakan data yang telah diberi label; Unsupervised Learning, di mana sistem mencari pola dari data tanpa label; dan Reinforcement Learning, di mana sistem belajar melalui umpan balik berdasarkan tindakan yang diambil untuk memaksimalkan reward.

Salah satu subbidang ML yang lebih kompleks adalah Deep Learning (DL), yang memanfaatkan artificial neural networks (ANNs) dengan banyak lapisan tersembunyi (hidden layers), sering disebut sebagai deep neural networks (DNNs). DL secara otomatis melakukan proses feature learning, sehingga lebih efektif dalam menangani data besar dan tidak terstruktur, seperti gambar dan teks. Beberapa arsitektur populer dalam DL meliputi Convolutional Neural Networks (CNNs), yang digunakan untuk analisis gambar dan visi komputer; Recurrent Neural Networks (RNNs), yang dirancang untuk data sekuensial seperti deret waktu dan pemrosesan bahasa alami; serta Generative Adversarial Networks (GANs), yang digunakan untuk menghasilkan data baru yang menyerupai data asli [11].

III. Analisis Dan Rancangan Sistem

A. Analisis Matriks Korelasi (Correlation Matrix Analysis)

Analisis ini bertujuan untuk mengeksplorasi hubungan antara variabel-variabel dalam dataset pemesanan hotel, khususnya untuk memahami faktor-faktor yang memengaruhi pembatalan pemesanan. Dengan memanfaatkan matriks korelasi, hubungan antara variabel independen dengan variabel dependen dapat diidentifikasi dan diinterpretasikan. Nilai korelasi digunakan untuk menentukan kekuatan dan arah hubungan antara independen dengan variabel dependen, dengan nilai positif menunjukkan hubungan langsung, dan nilai negatif menunjukkan hubungan berlawanan. Proses ini memberikan wawasan tentang bagaimana setiap variabel independen berkontribusi terhadap peningkatan atau penurunan kemungkinan pembatalan pemesanan.

B. Analisis Prediksi Pembatalan Menggunakan Pembelajaran Mesin

Untuk memberikan prediksi yang lebih akurat mengenai pembatalan pemesanan, analisis ini menggunakan model pembelajaran mesin untuk mengidentifikasi pola dalam data yang dapat digunakan oleh hotel untuk memprediksi kemungkinan pembatalan di masa depan.

IV. IMPLEMENTASI

A. Penjelasan Dataset

Dataset yang digunakan dalam penelitian ini merupakan dataset pemesanan hotel dengan 32 kolom dan 119.390 entri. Dataset ini mencakup berbagai informasi terkait pemesanan, seperti jenis hotel (City Hotel atau Resort Hotel), status pembatalan, waktu tunggu sebelum kedatangan, jumlah tamu, lama menginap, serta asal negara pelanggan berdasarkan kode ISO 3166-1 alpha-3. Terdapat juga informasi tentang jenis paket makanan yang dipesan, termasuk Bed & Breakfast (BB), Half Board (HB), Full Board (FB), dan Self Catering (SC). Selain itu, dataset mencatat segmen pasar pemesanan seperti Direct, Online Travel Agency (OTA), Corporate, dan Groups, serta kanal distribusi seperti Corporate, Direct, dan Global Distribution System (GDS).

Faktor lain yang dicatat meliputi jenis deposit (No Deposit, Non-Refundable, dan Refundable), status tamu apakah merupakan pelanggan berulang, serta jumlah perubahan dalam pemesanan. Dataset juga mencantumkan informasi tentang tipe pelanggan, yang diklasifikasikan sebagai Contract (berdasarkan perjanjian jangka panjang), Group (pemesanan kelompok), Transient (tamu individu tanpa keterikatan), dan Transient-Party (kelompok kecil dalam satu reservasi). Selain itu, terdapat data tentang permintaan khusus, jumlah tempat parkir yang diminta, serta status akhir reservasi seperti Canceled, Check-Out, atau No-Show. Dengan berbagai informasi ini, dataset dapat digunakan untuk menganalisis tren pemesanan, pola pembatalan, serta preferensi pelanggan dalam industri perhotelan.

B. Menilai Dataset

Pada tahap ini, dilakukan persiapan data dengan mengimpor dataset menggunakan Pandas serta memuat pustaka pendukung seperti NumPy, Matplotlib, dan Seaborn untuk analisis data. Dataset disimpan dalam sebuah dataframe bernama df.

1. Menampilkan Dataset
   Pada tahap ini, dilakukan persiapan data dengan mengimpor dataset menggunakan Pandas serta memuat pustaka pendukung seperti NumPy, Matplotlib, dan Seaborn untuk analisis data. Dataset disimpan dalam sebuah dataframe bernama df.
2. Memeriksa Tipe Data dan Missing Value
   Pemeriksaan tipe data dan jumlah nilai kosong dilakukan dengan df.info(). Hasilnya menunjukkan bahwa tiga kolom mengandung missing value, yaitu Company, Country, Children, dan Agent. Kolom Company memiliki jumlah missing value yang paling tinggi, sementara tiga lainnya memiliki jumlah yang lebih sedikit.
3. Mengidentifikasi Missing Value
   Untuk mengetahui jumlah missing value pada setiap kolom secara lebih rinci, digunakan df.isnull().sum(). Hasil analisis menunjukkan bahwa kolom Company memiliki jumlah nilai kosong paling banyak, yang memerlukan penanganan lebih lanjut dalam tahap pembersihan data.
4. Memeriksa Data Duplikat
   Pemeriksaan data duplikat dilakukan menggunakan df.duplicated().sum(), yang mengidentifikasi sebanyak 31.994 data duplikat. Data duplikat ini perlu diatasi agar tidak menyebabkan bias dalam analisis.
5. Menampilkan Ringkasan Statistik
   Ringkasan statistik dataset dianalisis menggunakan df.describe(include="all"), yang memberikan informasi terkait distribusi data. Pada fitur adr, ditemukan nilai maksimum sebesar 5400, yang jauh lebih tinggi dibandingkan kuartil ketiga (126) dan rata-rata (101,83). Nilai ini dianggap anomali dan akan ditangani dalam tahap pembersihan data. Dengan analisis ini, pola dalam dataset dapat diidentifikasi lebih akurat sebelum dilakukan tahap pemrosesan lebih lanjut.

C. Pembersihan Dataset

Proses pembersihan data dilakukan untuk memastikan kualitas data yang digunakan dalam analisis. Tahapan ini mencakup penghapusan fitur tidak relevan, penanganan missing value, penghapusan data duplikat, dan penanganan nilai tidak valid.

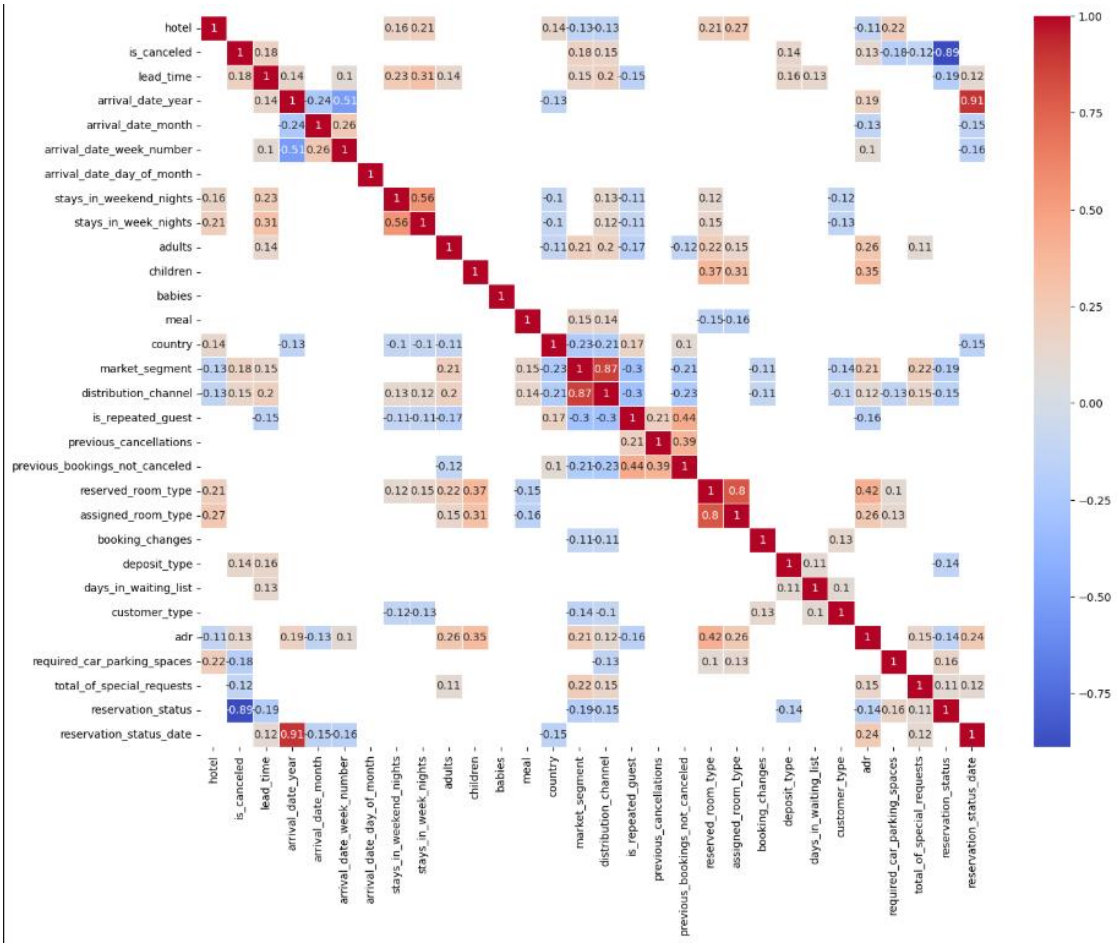
1. Penghapusan Fitur Tidak Relevan
   Fitur Company dan Agent dihapus karena hanya berisi ID yang tidak memberikan informasi signifikan serta memiliki banyak missing value.
2. Penanganan Missing Value pada Fitur Country dan Children
   Missing value pada Country diisi dengan nilai yang paling sering muncul (PRT), mengingat dataset berasal dari pemesanan hotel di Portugal. Sementara itu, hanya terdapat empat missing value pada fitur Children, sehingga data tersebut dihapus untuk menjaga integritas dataset.
3. Penghapusan Data Duplikasi
   Sebanyak 31.994 data duplikat diidentifikasi dan dihapus agar tidak menyebabkan bias dalam analisis.

4. Menghapus Invalid Value pada Fitur adr
   Nilai adr maksimum sebesar 5400 dihapus karena jauh melebihi kuartil ketiga dan dianggap sebagai anomali yang dapat memengaruhi hasil analisis.

Dengan selesainya proses pembersihan data, dataset menjadi lebih siap untuk tahap berikutnya, yaitu Exploratory Data Analysis. Langkah ini memastikan bahwa data bebas dari nilai kosong, fitur tidak berguna, dan data duplikat sehingga hasil analisis lebih akurat dan dapat dipercaya.

D. Correlation Matrix

Hasil perhitungan menunjukkan tujuh variabel prediktor dengan korelasi tertinggi terhadap pembatalan reservasi (is_canceled). Nilai korelasi berkisar antara -1 hingga +1, dengan nilai positif menunjukkan hubungan langsung dan nilai negatif menunjukkan hubungan berlawanan.



Gambar 1. *Correlation Matrix*

1. Reservation Status (-0.89)
   Korelasi negatif yang sangat kuat menunjukkan bahwa status reservasi secara langsung menentukan pembatalan, dengan "Canceled" mengindikasikan pembatalan, sedangkan "Checked-Out" atau "No-Show" tidak.

reservation_status	Canceled	Check-Out	No-Show
is_canceled			
0	0	63331	0
1	23006	0	1014

Gambar 2. Tabel *Crosstab* untuk *is_canceled* dan *reservation_status*

2. Lead Time (+0.18)
   Semakin lama jarak antara pemesanan dan kedatangan, semakin besar kemungkinan pembatalan, karena pelanggan lebih rentan terhadap perubahan rencana.

	is_canceled	lead_time				
		mean	min	max	std	nunique
0	0	70.127394	0	737	81.637058	422
1	1	105.721274	0	629	91.883773	465

Gambar 3. Tabel Statistik Deskriptif *lead_time* Berdasarkan *is_canceled*

3. Market Segment (+0.18)
   Segmen pasar tertentu, seperti Online TA, memiliki tingkat pembatalan lebih tinggi karena kemudahan pembatalan yang ditawarkan.

market_segment	Aviation	Complementary	Corporate	Direct	Groups	Offline TA/TO	Online TA	Undefined
is_canceled								
0	182	614	3690	10059	3604	11819	33363	0
1	45	88	510	1737	1335	2059	18244	2

Gambar 4. Tabel *Crosstab* untuk i*s_canceled* dan *maket_segment*

4. Distribution Channel (+0.15)
   Kanal TA/TO berkontribusi pada korelasi positif, karena kebijakan pembatalan yang fleksibel membuat pelanggan lebih mungkin membatalkan pemesanan.

distribution_channel	Corporate	Direct	GDS	TA/TO	Undefined
is_canceled					
0	4421	11053	145	47711	1
1	648	1925	36	21407	4

Gambar 5. Tabel *Crosstab* untuk *is_canceled* dan *distribution_channel*

5. Deposit Type (+0.14)
   Tipe deposit no refund memiliki tingkat pembatalan tertinggi, karena pelanggan lebih cenderung membatalkan jika menemukan alternatif yang lebih murah.

deposit_type	No Deposit	Non Refund	Refundable
is_canceled			
0	63195	55	81
1	23011	983	26

Gambar 6. Tabel *Crosstab* untuk *is_canceled* dan *deposit_type*

6. Average Daily Rate (+0.13)
   Reservasi dengan tarif rata-rata lebih tinggi cenderung dibatalkan lebih sering, kemungkinan karena fleksibilitas pelanggan dalam mengubah rencana mereka.

	is_canceled	adr				
		mean	min	max	std	nunique
0	0	102.016114	-6.38	510.0	51.381641	7610
1	1	117.578095	0.00	540.0	52.043350	3941

Gambar 7. Tabel Statistik Deskriptif *adr* berdasarkan *is_canceled*

7. Required Car Parking Spaces (-0.18)
   Pemesanan dengan permintaan tempat parkir lebih kecil kemungkinannya untuk dibatalkan, menunjukkan bahwa pelanggan dengan kebutuhan parkir lebih berkomitmen terhadap reservasi mereka.

required_car_parking_spaces	0	1	2	3	8
is_canceled					
0	56018	7280	28	3	2
1	24020	0	0	0	0

Gambar 8. Tabel *Crosstab* untuk *is_canceled* dan *required_car_parking_spaces*

8. Total of Special Requests (-0.12)
   Pemesanan dengan lebih banyak permintaan khusus, seperti kamar tertentu atau fasilitas tambahan, cenderung tidak dibatalkan karena pelanggan lebih terikat dengan detail pemesanan.

	is_canceled	total_of_special_requests				
		mean	min	max	std	nunique
0	0	0.760496	0	5	0.849938	6
1	1	0.535762	0	5	0.758928	6

Gambar 9. Tabel *Crosstab* untuk *is_canceled* dan *total_of_special_requests*

Studi ini mengeksplorasi hubungan antarvariabel prediktor, terutama saat variabel tertentu dijadikan dependent. Berikut adalah variabel dengan korelasi tertinggi, baik positif maupun negatif:

1. Lead Time (+0.31 Stays In Week Nights)
   Semakin lama tamu menginap pada malam hari kerja, semakin jauh hari mereka melakukan reservasi, terutama untuk perjalanan bisnis atau liburan panjang.

	stays_in_week_nights	lead_time				
		mean	min	max	std	nunique
0	0	36.654345	0	737	61.504109	285
1	1	49.072340	0	629	74.671811	429
2	2	74.855242	0	629	81.012151	422
3	3	89.530494	0	479	83.492385	374
4	4	100.391006	0	504	84.593125	362
5	5	127.629334	0	542	89.154136	378
6	6	132.591008	0	435	88.130343	329
7	7	145.811789	0	542	93.673057	317
8	8	145.691680	0	454	95.106861	271
9	9	125.292237	0	406	95.768195	126
10	10	163.116255	0	406	99.900413	331

Gambar 10. Tabel Statistik Deskriptif *lead_time* berdasarkan *stays_in_week_nights*

2. Market Segment (+0.87 Distribution Channel, +0.22 Total of Special Requests)
   Segmen pasar memiliki hubungan erat dengan saluran distribusi, misalnya Online TA dominan di saluran Online TA sendiri. Segmen tertentu juga lebih sering mengajukan permintaan khusus, seperti tamu Direct dan Online TA.

distribution_channel	Corporate	Direct	GDS	TA/TO	Undefined
market_segment					
Aviation	217	0	0	10	0
Complementary	81	546	0	75	0
Corporate	3891	154	0	155	0
Direct	82	11482	1	229	2
Groups	669	649	0	3621	0
Offline TA/TO	95	16	44	13723	0
Online TA	34	131	136	51305	1
Undefined	0	0	0	0	2

Gambar 11. Tabel *Crosstab* untuk *market_segment* dan *distribution_channel*

total_of_special_requests	0	1	2	3	4	5
market_segment						
Aviation	209	9	9	0	0	0
Complementary	289	228	125	45	15	0
Corporate	3293	705	165	33	4	0
Direct	7180	2915	1259	365	67	10
Groups	4246	590	94	9	0	0
Offline TA/TO	9973	3081	688	119	15	2
Online TA	18674	21477	9468	1745	219	24
Undefined	0	1	1	0	0	0

Gambar 12. Tabel *Crosstab* antara *total_of_special_requests* dan *market_segment*

3. Distribution Channel (+0.87 Market Segment, +0.20 Lead Time, +0.20 Adults)
   Saluran distribusi memengaruhi lead time dan jumlah tamu dewasa. TA/TO memiliki lead time panjang dan sering melayani kelompok, sedangkan Corporate memiliki lead time lebih pendek dengan mayoritas pemesanan untuk individu atau pasangan.

	distribution_channel	lead_time				
		mean	min	max	std	nunique
0	Corporate	33.486092	0	390	73.182762	182
1	Direct	52.395901	0	737	74.910635	345
2	GDS	20.121547	0	220	27.091955	54
3	TA/TO	88.647979	0	629	86.748145	476
4	Undefined	23.000000	1	103	44.816292	4

Gambar 13. Tabel Statistik Deskriptif *lead_time* berdasarkan *distribution_channel*

adults	0	1	2	3	4	5	6	10	20	26	27	40	50	55
distribution_channel														
Corporate	19	3779	1207	64	0	0	0	0	0	0	0	0	0	0
Direct	67	2774	9389	712	25	2	1	1	2	0	2	1	1	1
GDS	0	166	14	1	0	0	0	0	0	0	0	0	0	0
TA/TO	299	9762	53860	5157	35	0	0	0	0	5	0	0	0	0
Undefined	0	0	4	1	0	0	0	0	0	0	0	0	0	0

Gambar 14. Tabel *Crosstab* antara *distribution_channel* dan *adults*

4. Deposit Type (-0.14 Reservation Status)
   Jenis deposit berpengaruh pada status reservasi. No refund lebih sering berakhir dengan pembatalan, sedangkan no deposit lebih fleksibel dan meningkatkan peluang tamu menyelesaikan masa inap.

reservation_status	Canceled	Check-Out	No-Show
deposit_type			
No Deposit	22004	63195	1007
Non Refund	977	55	6
Refundable	25	81	1

Gambar 15. Tabel *Crosstab* antara *deposit_type* dan *reservation_status*

5. Average Daily Rate (+0.42 Reserved Room Type)
   Harga rata-rata harian (ADR) berkorelasi dengan tipe kamar. Kamar premium memiliki ADR lebih tinggi, sementara kamar ekonomis memiliki harga lebih rendah tetapi jumlah pemesanan lebih banyak.

	reserved_room_type	adr				
		mean	min	max	std	nunique
0	A	92.226247	-6.38	540.00	40.823209	5372
1	B	90.377878	0.00	284.10	35.693315	425
2	C	160.561770	0.00	367.00	71.719435	555
3	D	122.080654	0.00	375.50	48.383344	3686
4	E	125.946427	0.00	451.50	60.308968	1981
5	F	168.229848	0.00	392.00	63.321948	1170
6	G	176.727904	0.00	426.25	79.361501	909
7	H	188.763993	0.00	437.00	75.893216	304
8	L	124.666667	8.00	200.00	69.451182	6
9	P	0.000000	0.00	0.00	0.000000	1

Gambar 16. Tabel Statistik Deskrptif *adr* berdasarkan *reserved_room_type*

6. Required Car Parking Spaces (+0.22 Hotel)
   Tamu Resort Hotel lebih sering membutuhkan tempat parkir dibandingkan City Hotel, yang lebih banyak diakses dengan transportasi umum.

required_car_parking_spaces	0	1	2	3	8
hotel					
City Hotel	51518	1891	3	2	0
Resort Hotel	28520	5389	25	1	2

Gambar 17. Tabel *Crosstab* berdasarkan *hotel* dan *required_car_parking_spaces*

E. Machine Learning

Pada bagian ini, digunakan dua algoritma pembelajaran mesin, yaitu Logistic Regression dan Random Forest, untuk memprediksi pembatalan pada dataset. Pada tahap preprocessing, data dipersiapkan menggunakan pipeline untuk memastikan setiap langkah preprocessing dilakukan secara konsisten dan efisien. Pipeline ini mencakup beberapa komponen penting, yang pertama adalah penggunaan StandardScaler untuk menstandarkan data numerik. Standarisasi ini dilakukan dengan menghilangkan rata-rata dan menskalakan data ke standar deviasi, sehingga algoritma pembelajaran mesin dapat bekerja secara optimal tanpa terpengaruh oleh skala fitur yang berbeda.

Selain itu, karena dataset memiliki distribusi kelas yang tidak seimbang, digunakan teknik SMOTE untuk melakukan oversampling pada kelas minoritas. Teknik ini menghasilkan data sintetis untuk kelas yang lebih kecil sehingga distribusi kelas menjadi lebih seimbang, yang pada akhirnya meningkatkan performa model dalam mendeteksi kelas pembatalan. Kombinasi pipeline, standardisasi, dan oversampling ini bertujuan untuk memaksimalkan keakuratan prediksi dan memastikan model dapat generalisasi dengan baik.

Evaluasi dilakukan menggunakan teknik cross-validation untuk memastikan konsistensi performa model, terutama mengingat dataset yang tidak seimbang. Skema StratifiedShuffleSplit digunakan dalam proses ini untuk menjaga distribusi kelas yang seimbang pada setiap fold, sehingga hasil evaluasi lebih representatif terhadap karakteristik data. Selain itu, beberapa percobaan dilakukan dengan variasi ukuran data uji sebesar 30%, 20%, dan 10%. Variasi ini bertujuan untuk menganalisis pengaruh proporsi data uji terhadap performa model, termasuk bagaimana perubahan ukuran data uji memengaruhi kemampuan model dalam generalisasi.

Metode evaluasi utama yang digunakan adalah recall karena fokus penelitian ini adalah mendeteksi sebanyak mungkin kasus pembatalan secara akurat. Recall lebih relevan dibandingkan accuracy, yang kurang informatif pada dataset tidak seimbang, karena model dapat mencapai accuracy tinggi hanya dengan memprediksi kelas mayoritas. Sementara itu, precision, yang mengukur ketepatan prediksi positif, tidak diprioritaskan karena kesalahan berupa prediksi positif palsu dianggap kurang berdampak dibandingkan gagal mendeteksi pembatalan. Meskipun F1-score menggabungkan recall dan precision, fokus utama pada deteksi penuh kasus pembatalan menjadikan recall sebagai metrik utama yang paling sesuai.

1. Logistic Regression

   Logistic Regression menunjukkan hasil performa yang cukup stabil dengan rata-rata recall yang berkisar antara 0.762 hingga 0.766 pada berbagai ukuran data uji. Berikut adalah detail hasilnya:

   a. Data uji: 30%
      - Cross-validation scores: [0.772, 0.758, 0.761, 0.758, 0.760]
      - Mean Recall: 0.762

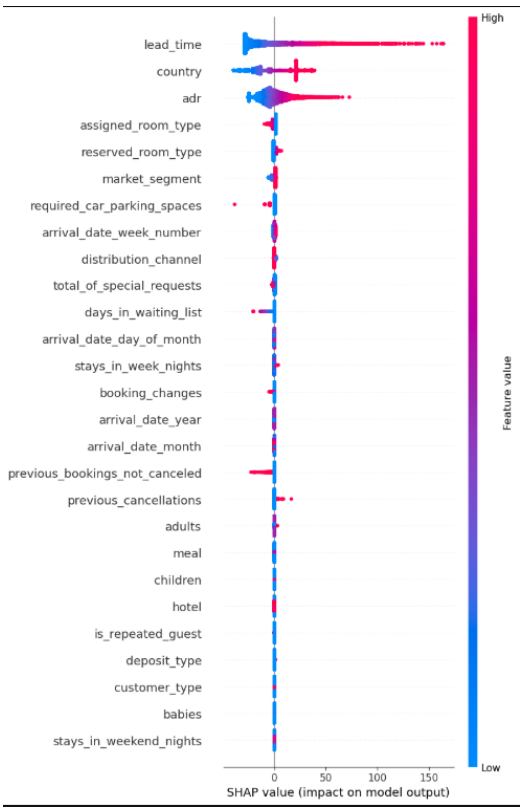
   b. Data uji: 20%
      - Cross-validation scores: [0.768, 0.761, 0.754, 0.761, 0.771]
      - Mean Recall: 0.763

   c. Data uji: 10%
      - Cross-validation scores: [0.761, 0.759, 0.782, 0.763, 0.763]
      - Mean Recall: 0.766

   Logistic Regression menunjukkan performa yang konsisten pada berbagai ukuran data uji, meskipun terdapat sedikit variasi dalam nilai rata-rata recall. Rata-rata recall tertinggi diperoleh pada ukuran data uji 10% dengan nilai 0.766, sedangkan ukuran data uji 30% menghasilkan rata-rata recall terendah sebesar 0.762.

   Meskipun selisih antara rata-rata recall tidak terlalu besar di semua ukuran data uji, tren ini menunjukkan bahwa performa model Logistic Regression sedikit lebih baik saat data pelatihan lebih besar. Hal ini mencerminkan bahwa model Logistic Regression mampu memanfaatkan lebih banyak data pelatihan untuk meningkatkan kemampuannya dalam menangkap pola dalam data.

   Namun, secara keseluruhan, perbedaan ini tergolong kecil, yang menandakan bahwa Logistic Regression cukup robust terhadap variasi ukuran data uji.



Gambar 18. *SHAP Summary Plot*

Visualisasi *SHAP* digunakan untuk mengevaluasi pentingnya fitur terhadap keluaran model *Logistic Regression*. Grafik ini menunjukkan seberapa besar kontribusi masing-masing fitur terhadap prediksi pembatalan.

1. *lead_time*

Fitur ini merupakan faktor yang paling signifikan dalam memprediksi pembatalan. *SHAP* value yang tinggi untuk nilai besar pada *lead_time* menunjukkan bahwa semakin lama waktu antara pemesanan dan kedatangan, semakin besar kemungkinan terjadi pembatalan. Ini masuk akal karena pemesanan dengan waktu lebih awal lebih berisiko untuk berubah.

2. *country*

Fitur ini memiliki dampak yang bervariasi, tergantung pada nilai spesifik dari negara asal pelanggan. Warna merah (nilai tinggi) menunjukkan bahwa asal negara tertentu mungkin meningkatkan kemungkinan pembatalan.

3. *adr* (*Average Daily Rate*)

Fitur ini juga berpengaruh signifikan. Nilai *adr* yang tinggi (ditunjukkan dengan warna merah) cenderung meningkatkan probabilitas pembatalan, sedangkan nilai rendah memiliki dampak sebaliknya.

4. Random Forest

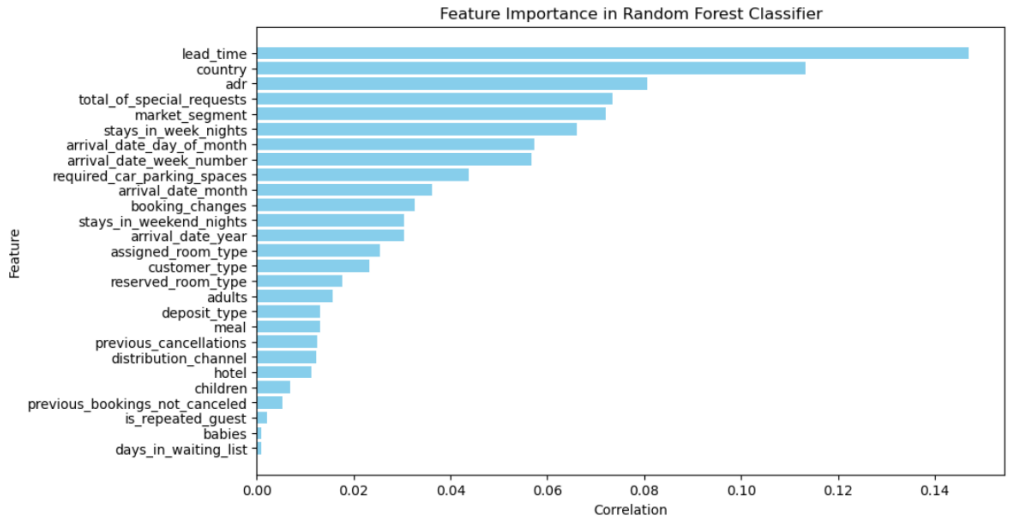
Random Forest menunjukkan performa yang sedikit lebih rendah dibandingkan Logistic Regression dengan rata-rata recall berkisar antara 0.675 hingga 0.681. Berikut adalah detail hasilnya:

a. Data uji: 30%
   - Cross-validation scores: [0.664, 0.672, 0.671, 0.676, 0.670]
   - Mean Recall: 0.670

b. Data uji: 20%
   - Cross-validation scores: [0.663, 0.667, 0.666, 0.682, 0.679]
   - Mean Recall: 0.674

c. Data uji: 10%
   - Cross-validation scores: [0.660, 0.666, 0.87, 0.681, 0.681]
   - Mean Recall: 0.675

Random Forest menghasilkan rata-rata recall tertinggi pada ukuran data uji 10% dengan nilai 0.675. Sementara itu, pada ukuran data uji 30% dan 20%, nilai rata-rata recall masing-masing adalah 0.670 dan 0.674. Secara keseluruhan, performa Random Forest menunjukkan konsistensi yang baik, tetapi tidak setinggi yang diharapkan jika dibandingkan dengan Logistic Regression, yang cenderung menunjukkan performa yang lebih stabil pada berbagai ukuran data uji.



Gambar 19. *Feature Importance Random Forest Classifier*

Fitur penting yang dihasilkan oleh *Random Forest* memiliki kesamaan dengan analisis dari *Logistic Regression*, di mana fitur *lead_time*, *adr*, dan *country* muncul sebagai variabel paling signifikan dalam memengaruhi pembatalan. Hal ini menunjukkan bahwa kedua model sepakat mengenai faktor utama yang berkontribusi terhadap prediksi pembatalan. Kesamaan ini menegaskan bahwa pemilihan fitur yang relevan dalam dataset sudah cukup baik dan memberikan kontribusi signifikan terhadap performa model pembelajaran mesin.

5. Multilayer Perceptron

Model Multilayer Perceptron dibuat untuk memprediksi pembatalan dengan konfigurasi arsitektur sederhana, terdiri dari 3 lapisan penuh dengan fungsi aktivasi ReLU untuk lapisan tersembunyi dan Sigmoid untuk keluaran karena ini adalah masalah klasifikasi biner. Evaluasi dilakukan menggunakan teknik Stratified Shuffle Split Cross-Validation dengan 5 fold dan variasi ukuran data uji (30%, 20%, dan 10%).

a. Data uji: 30%
   - Fold Recalls: [0.806, 0.80, 0.805, 0.813, 0.802]
   - Mean Recall: 0.805

b. Data uji: 20%
   - Fold Recalls: [0.805, 0.794, 0.797, 0.799, 0.801]
   - Mean Recall: 0.799

c. Data uji: 10%
   - Fold Recalls: [0.795, 0.794, 0.790, 0.792, 0.810]
   - Mean Recall: 0.796

Model Multilayer Perceptron menunjukkan performa yang cukup konsisten dalam memprediksi pembatalan dengan nilai recall rata-rata di atas 0.79 pada berbagai ukuran data uji. Hasil ini mencerminkan kemampuan Multilayer Perceptron untuk menangkap pola non-linear dari fitur yang digunakan. Performa terbaik dicapai pada ukuran data uji sebesar 30% dengan nilai recall rata-rata 0.805, sementara performa menurun sedikit pada ukuran data uji 20% dan 10%. Hal ini menunjukkan bahwa ukuran data uji yang lebih besar dapat membantu model lebih memahami pola dalam data validasi, tetapi penurunan performa pada ukuran data uji yang lebih kecil tetap dalam rentang yang wajar.

Secara keseluruhan, urutan performa model dari yang terbaik hingga terendah adalah sebagai berikut:

- Multilayer Perceptron: Kemampuan menangkap pola kompleks dan non-linear membuatnya unggul pada dataset ini.
- Logistic Regression: Model sederhana namun efektif untuk hubungan linier, menghasilkan performa yang stabil.
- Random Forest: Meskipun cocok untuk data yang tidak terstruktur, model ini kurang optimal dalam menangkap pola sederhana atau linier pada dataset ini.

V. KESIMPULAN

Penelitian ini menganalisis faktor-faktor utama yang memengaruhi pembatalan pemesanan hotel dengan menggunakan teknik eksplorasi data dan pembelajaran mesin. Hasil analisis menunjukkan bahwa lead time, market segment, distribution channel, deposit type, dan average daily rate memiliki korelasi signifikan terhadap pembatalan.

Untuk memprediksi pembatalan, penelitian ini membandingkan tiga model pembelajaran mesin: Logistic Regression, Random Forest, dan Multilayer Perceptron (MLP). MLP menunjukkan performa terbaik dalam mendeteksi pembatalan, diikuti oleh Logistic Regression dan Random Forest. Teknik SMOTE digunakan untuk mengatasi ketidakseimbangan kelas, sementara recall dipilih sebagai metrik utama karena fokus penelitian ini adalah mendeteksi sebanyak mungkin pembatalan.

DAFTAR PUSTAKA

[1] M. Wynn and P. Jones, "IT Strategy in the Hotel Industry in the Digital Era," *Sustainability*, p. 14, 28 Agustus 2022.

[2] G. Vial, "Understanding digital transformation: A review and a research agenda," *The Journal of Strategic Information Systems*, vol. 28, no. 2, pp. 118-144, 2019.

[3] V. S. Moertini and M. T. Adithia, Pengantar Data Science dan Aplikasinya bagi Pemula, Bandung: Unpar Press, 2020.

[4] W. McKinney, Pyton for Data Analysis, United States: 0'Reilly Media, 2022.

[5] C. R. Harris and K. J. Millman, "Array programming with NumPy," vol. 585, p. 358, 2020.

[6] I. Albanna and R. T. h. Laksono, "Implementasi Pandas Data framesebagai Agregasi dan Tabulasi Penyajian Data Luaran Survei Kepuasan Pengguna Proses Pembelajaran dalam Pendidikan Tinggi," *Institut TeknologiAdhi Tama Surabaya*, pp. 2-3, 2022.

[7] R. Odegua, "DataSist: A Python-based library for easy data analysis, visualization and modeling," *Department of Computer Science*, 2017.

[8] J. Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, 2020.

[9] T. Milo and A. Somech, "Automating Exploratory Data Analysis via Machine Learning: An Overview," *ACM SIGMOD International Conference on Management of Data*, pp. 2617-2622, 2020.

[10] D. J. Irvine, L. J. S. Halloran and P. Brunner, "Opportunities and limitations of the ChatGPT Advanced Data Analysis plugin for hydrological analyses," *Hydrological Processes*, vol. 37, no. 10, p. e15015, 2023.

[11] . C. Janiesch, P. Zschech and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 4, p. 685–695, 2021.