

LAPORAN UAS KECERDASAN BUATAN

“Disusun Untuk Memenuhi Tugas Mata Kuliah Kecerdasan Buatan”

Dosen Pengampu : Leni Fitriani S.Kom., M.Kom



Disusun Oleh :

Fikri Rahmadi
2206127

TEKNIK INFORMATIKA
INSTITUT TEKNOLOGI GARUT

KONSEP LEARNING PADA KECERDASAN BUATAN

Machine learning adalah kecerdasan buatan (AI) yang mempelajari bagaimana membuat sistem, membuat data, dan membuat algoritma yang dapat melakukan tugas-tugas sendiri tanpa bantuan pengguna, memungkinkan Programmer mampu belajar. Metode analisis mesin pembelajaran menggunakan algoritma komputer untuk membantu menangani data besar. Belajar seperti manusia, dengan menggunakan contoh, adalah cara kerja machine learning. Data yang disebut train dataset digunakan dalam proses belajar ini. Untuk menghasilkan model, komputer akan menggunakan data untuk melakukan proses belajar [1].

PERBEDAAN KONSEP LEARNING DENGAN 3 KONSEP KECERDASAN BUATAN LAINNYA YANG SUDAH DIPELAJARI

Konsep	Keterangan
<i>Learning</i>	Proses memperoleh pengetahuannya di dapat berdasarkan hasil pengalaman, maupun juga data yang pernah diberikan dan data yang ada
<i>Reasoning</i>	Biasanya digunakan untuk membuat sebuah keputusan dengan menggunakan logika
<i>Searching</i>	Digunakan untuk menemukan solusi yang relevan, baik di ruang pencarian yang besar maupun yang kecil
<i>Planning</i>	Sistem pada konsep planning ini biasanya digunakan untuk merencanakan tindakan yang akan dilakukan

Pengertian Data

Data adalah kumpulan informasi atau fakta mentah yang diperoleh melalui proses pengamatan atau pencarian dari berbagai sumber. Informasi ini dapat berupa simbol, angka, kata-kata, atau gambar. Secara etimologis, istilah "data" berasal dari bentuk jamak dari kata Latin "datum", yang berarti "sesuatu yang diberikan." Dalam penggunaan sehari-hari, kata "data" mengacu pada fakta objek yang diamati, baik dalam bentuk angka maupun kata-kata. Dari perspektif statistika, data adalah fakta-fakta yang digunakan untuk membuat kesimpulan.

Sangat penting untuk diingat bahwa data bersifat mentah, sehingga sangat penting untuk mengandalkan data yang telah divalidasi untuk memastikan kebenaran, akurasi, ketepatan waktu, dan cakupannya. Oleh karena itu, seringkali diperlukan pengolahan data untuk menghasilkan informasi yang mudah dipahami [2].

Data Set

Data terdiri dari fakta dan angka (juga disebut sebagai data mentah) yang relevan dengan suatu masalah. Data terdiri dari dua aspek: objek dan atribut; contohnya adalah manusia, pohon, dan binatang; dan atribut, seperti umur, tinggi badan, dan berat badan. Kumpulan objek data disebut dataset. Dataset dapat disebut sebagai rekaman, poin, vector, pola, peristiwa, observasi, kasus, atau data. Beberapa atribut menunjukkan karakteristik objek data. Karakteristik, variabel, field, fitur, atau dimensi adalah semua istilah yang dapat digunakan untuk menggambarkan atribut [3].

Data Warehouse

Untuk membantu proses pengambilan keputusan manajemen, data warehouse adalah koleksi data yang berorientasi subjek, terintegrasi, nonvolatile, dan time-variant. Mannino menggambarkan data warehouse sebagai tempat penyimpanan data terpusat di mana data diintegrasikan, dibersihkan, dan diarsipkan untuk mendukung pengambilan keputusan. Namun, McLeod menggambarkan data warehouse sebagai sistem penyimpanan data yang berkapasitas besar, di mana data dikumpulkan dengan menambahkan record baru daripada mengubah record yang sudah ada dengan informasi baru. Dalam kegiatan sehari-hari perusahaan, data jenis ini tidak digunakan untuk proses pengambilan keputusan [4].

Big Data

Istilah "Big Data" mengacu pada kumpulan data yang sangat besar dan kompleks yang tidak dapat diproses dengan perangkat pengelola database konvensional atau aplikasi pemroses data lainnya. Data yang memiliki volume, keragaman, dan kompleksitas yang sangat variatif disebut sebagai data besar. Ini membutuhkan model arsitektur, teknis, algoritma, dan analisis baru untuk mengelola, memanfaatkan, dan mengembangkan pengetahuan yang ada [5].

Pengetahuan

Fakta, kebenaran, atau informasi yang diperoleh melalui pengalaman atau pembelajaran disebut posteriori, atau introspeksi disebut priori, adalah jenis pengetahuan. Pengetahuan adalah informasi yang seseorang ketahui atau menyadari. Pengetahuan yang sesuai dengan teori probabilitas Bayesian termasuk, tetapi tidak dibatasi pada deskripsi, hipotesis, konsep, teori, prinsip, dan prosedur. Pengetahuan juga dapat didefinisikan sebagai berbagai gejala yang ditemui dan diperoleh manusia melalui pengamatan akal budi. Ketika seseorang menggunakan akal budinya untuk mengenali benda atau kejadian tertentu yang belum pernah mereka lihat atau rasakan sebelumnya, ini disebut pengetahuan [6].

1. Business Understanding

- Permasalahan : Banyak hotel mengalami kerugian karena reservasi yang dibatalkan secara mendadak. Hal ini berdampak pada perencanaan operasional dan okupansi kamar.
- Tujuan Proyek: Mengklasifikasikan apakah reservasi hotel akan dibatalkan atau tidak menggunakan algoritma machine learning.
- User Sistem : Manajer hotel, sistem reservasi online, dan tim operasional.
- Manfaat AI : Membantu hotel mengantisipasi pembatalan dan mengambil tindakan proaktif (misalnya, overbooking, promosi).

2. Data Understanding

Dataset Hotel Reservations	
Nama Dataset & link unduh	https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset?resource=download

Jumlah Baris Data	36275				
Fitur	Nama Fitur				Type Data
	Booking_ID				object
	no_of_adults				int64
	no_of_children				int64
	no_of_weekend_nights				int64
	no_of_week_nights				int64
	type_of_meal_plan				object
	required_car_parking_space				int64
	room_type_reserved				object
	lead_time				int64
	arrival_year				float64
	arrival_month				int64
	arrival_date				object
	market_segment_type				object
	repeated_guest				float64
	no_of_previous_cancellations				float64
	no_of_previous_bookings_not_canceled				float64
	avg_price_per_room				float64
	no_of_special_requests				float64
booking_status				object	
dtype: object					
Class (label/ target) & typedata	Booking_status (object)				
Link File Eksperimen:					
Pra-Pemrosesan Dataset					
Kondisi Dataset					Cara Mengatasi (Preprocessing Data)
Duplikasi Baris Data (✓)		Yes	✓	No	Tidak ada duplikasi Data dalam dataset kami
Missing Value (✓)		Yes	✓	No	Tidak ada missing value dalam dataset kami
Outlier (✓)	✓	Yes		No	Didalam dataset kami memiliki banyak outlier, namun kami sudah mengatasinya dengan menggunakan teknik IQR.
Normalisasi (✓)	✓	Yes		No	Normalisasi dilakukan dengan menggunakan MinMaxScaler
Perubahan Tipe Daya(✓)	✓	Yes		No	Karna tidak bisa digunakan di decision tree, cara menyelesaikannya yaitu dengan menyeragamkan tipe data menjadi integer
Imbalanced Dataset(✓)	✓	Yes		No	Teknik SMOTE sudah diterapkan untuk menangani ketidak seimbangan dataset.
Reduksi Fitur (✓)	✓	Yes		No	Reduksi fitur dilakukan menggunakan PCA (Principal Component Analysis) dengan mengurangi dataset ke 5 komponen utama.
Seleksi Fitur (✓)	✓	Yes		No	Seleksi fitur dilakukan menggunakan SelectFromModel dengan algoritma RandomForestClassifier. Fitur yang dipilih berdasarkan pentingnya (importance) dari algoritma Random Forest.
Kondisi lainnya:					
Pemodelan					

Jenis Pekerjaan yang dipilih (✓)		Clustering	✓	Classification		Regression		Forecasting
Algoritma yang digunakan								
Splitting Data (✓)		60:40		70:30	✓	80:20		90:20
k-fold Cross Validation (✓)		K=3	✓	K=5		K=7		K=10
Evaluasi Model								
Ukuran Evaluasi Model		precision		recall		F1-score		Support
Nilai Evaluasi		0.80 0.86 0.83 0.84		0.71 0.91 0.81 0.84		0.75 0.88 0.84 0.82 0.84		34 66 100 100 100
Jumlah Ujicoba	10	Catatan Setiap Ujicoba: 1. Menghilangkan outlier 2. Membuat kode reduksi fitur 3. Melakukan normalisasi 4. Dan mengatasi error pada setiap kode program						

Fitur (atribut):

- lead_time: jumlah hari antara pemesanan dan kedatangan.
- arrival_date_month, arrival_date_year: waktu kedatangan.
- country: negara asal pelanggan.
- deposit_type, customer_type, dll.
- is_canceled: target klasifikasi (1: dibatalkan, 0: tidak).

3. Exploratory Data Analysis (EDA)

- Tampilan Dataset

id	no_of_sails	no_of_children	no_of_weekend_sails	no_of_week_sails	type_of_meal_plan	measured_cab_booking_type	measured_cab_booking_status	lead_time	arrival_date	arrival_date_month	arrival_date_year	customer_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
INN00004	2	0	0	0	2 Meal Plan 1	0 Room_Type 1	0	211	2018	5	20 Online	0	0	0	0	120	0	Canceled
INN00005	2	0	1	1	1 Not Selected	0 Room_Type 1	0	48	2018	4	11 Online	0	0	0	0	945	0	Canceled
INN00006	2	0	0	0	2 Meal Plan 2	0 Room_Type 1	0	345	2018	9	13 Online	0	0	0	0	115	1	Canceled
INN00007	2	0	1	3	3 Meal Plan 1	0 Room_Type 1	0	34	2017	10	15 Online	0	0	0	0	10755	1	Not_Canceled
INN00008	2	0	1	3	3 Meal Plan 1	0 Room_Type 4	0	83	2018	12	25 Online	0	0	0	0	10561	1	Not_Canceled
INN00009	3	0	0	0	4 Meal Plan 1	0 Room_Type 1	0	121	2018	7	8 Offline	0	0	0	0	969	1	Not_Canceled
INN00010	2	0	0	0	5 Meal Plan 1	0 Room_Type 4	0	44	2018	10	15 Online	0	0	0	0	12344	3	Not_Canceled
INN00011	1	0	1	0	0 Not Selected	0 Room_Type 1	0	0	2018	9	11 Online	0	0	0	0	8203	0	Not_Canceled
INN00012	1	0	2	1	1 Meal Plan 1	0 Room_Type 4	0	35	2018	4	30 Online	0	0	0	0	1404	1	Not_Canceled
INN00013	2	0	2	1	1 Not Selected	0 Room_Type 1	0	30	2018	11	25 Online	0	0	0	0	88	0	Canceled
INN00014	1	0	2	0	0 Meal Plan 1	0 Room_Type 1	0	95	2018	11	20 Online	0	0	0	0	90	2	Canceled
INN00015	2	0	0	0	2 Meal Plan 1	0 Room_Type 1	0	47	2017	10	20 Online	0	0	0	0	945	2	Not_Canceled
INN00016	2	0	0	0	2 Meal Plan 2	0 Room_Type 1	0	256	2018	6	15 Online	0	0	0	0	115	1	Canceled
INN00017	1	0	1	0	3 Meal Plan 1	0 Room_Type 1	0	0	2017	10	5 Offline	0	0	0	0	96	0	Not_Canceled
INN00018	2	0	1	0	0 Not Selected	0 Room_Type 1	0	1	2017	8	10 Online	0	0	0	0	96	1	Not_Canceled
INN00019	2	0	2	2	2 Meal Plan 1	0 Room_Type 1	0	99	2017	10	30 Online	0	0	0	0	65	0	Canceled
INN00020	2	0	1	0	0 Meal Plan 1	0 Room_Type 1	0	12	2017	10	4 Offline	0	0	0	0	72	0	Not_Canceled
INN00021	2	0	2	2	2 Meal Plan 1	0 Room_Type 1	0	96	2017	10	30 Online	0	0	0	0	65	0	Canceled
INN00022	1	0	0	0	1 Meal Plan 1	0 Room_Type 1	0	122	2018	11	25 Corporate	0	0	0	0	87	0	Not_Canceled
INN00023	2	0	2	4	4 Meal Plan 1	0 Room_Type 1	0	2	2018	3	20 Offline	0	0	0	0	85	0	Not_Canceled
INN00024	2	0	0	0	3 Meal Plan 1	0 Room_Type 1	0	37	2018	10	13 Offline	0	0	0	0	105	0	Not_Canceled
INN00025	2	0	2	1	1 Not Selected	0 Room_Type 1	0	130	2018	5	22 Online	0	0	0	0	945	1	Not_Canceled
INN00026	2	0	0	0	2 Meal Plan 1	0 Room_Type 1	0	99	2018	4	28 Online	0	0	0	0	1143	1	Not_Canceled
INN00027	2	0	1	1	1 Meal Plan 1	0 Room_Type 1	0	60	2017	9	21 Offline	0	0	0	0	66	0	Not_Canceled
INN00028	1	0	0	0	2 Meal Plan 1	0 Room_Type 4	0	2	2018	5	15 Aviation	0	0	0	0	110	0	Canceled
INN00029	1	0	1	3	3 Meal Plan 1	0 Room_Type 1	0	37	2017	11	6 Online	0	0	0	0	3733	0	Canceled
INN00030	2	0	0	0	2 Meal Plan 2	0 Room_Type 1	0	56	2017	9	17 Offline	0	0	0	0	82	0	Not_Canceled
INN00031	2	0	1	0	1 Meal Plan 1	0 Room_Type 4	0	3	2017	9	18 Online	0	0	0	0	1773	1	Not_Canceled
INN00032	2	0	2	2	2 Meal Plan 1	0 Room_Type 1	0	107	2018	11	13 Online	0	0	0	0	875	1	Not_Canceled
INN00033	0	2	0	0	3 Meal Plan 1	0 Room_Type 2	0	56	2018	12	7 Online	0	0	0	0	8244	1	Not_Canceled
INN00034	2	0	0	0	1 Not Selected	0 Room_Type 1	0	2	2018	11	9 Online	0	0	0	0	695	1	Not_Canceled
INN00035	2	0	0	0	2 Meal Plan 1	0 Room_Type 1	0	72	2018	10	7 Offline	0	0	0	0	855	0	Not_Canceled
INN00036	2	0	0	0	1 Not Selected	1 Room_Type 1	0	23	2018	4	27 Online	0	0	0	0	115	1	Not_Canceled
INN00037	1	0	2	1	1 Meal Plan 1	0 Room_Type 1	0	34	2018	6	15 Online	0	0	0	0	81	0	Canceled
INN00038	2	0	0	0	1 Meal Plan 1	0 Room_Type 1	0	289	2017	10	17 Offline	0	0	0	0	67	0	Not_Canceled
INN00039	2	0	2	3	3 Not Selected	0 Room_Type 1	0	247	2018	11	19 Online	0	0	0	0	8375	0	Canceled

- Explanatory Data Analysis

▼ A. Impor Libraries Python yang dibutuhkan

```
[1] import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score, K
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix, acc
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.feature_selection import SelectFromModel
from sklearn.utils import resample
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
```

▼ B. Membaca dataset

```
[2] # loading and reading dataset
df = pd.read_csv("/content/Hotel_Reservations.csv")
print(df.head())
```

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	\
0	INN000001	2	0	1	
1	INN000002	2	0	2	
2	INN000003	1	0	2	
3	INN000004	2	0	0	
4	INN000005	2	0	1	

	no_of_week_nights	type_of_meal_plan	required_car_parking_space	\
0	2	Meal Plan 1	0	
1	3	Not Selected	0	
2	1	Meal Plan 1	0	
3	2	Meal Plan 1	0	
4	1	Not Selected	0	

	room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	\
0	Room_Type 1	224	2017	10	2	
1	Room_Type 1	5	2018	11	6	
2	Room_Type 1	1	2018	2	28	
3	Room_Type 1	211	2018	5	20	
4	Room_Type 1	48	2018	4	11	

	market_segment_type	repeated_guest	no_of_previous_cancellations	\
0	Offline	0	0	
1	Online	0	0	
2	Online	0	0	
3	Online	0	0	
4	Online	0	0	

	no_of_previous_bookings_not_canceled	avg_price_per_room	\
0	0	65.00	
1	0	106.68	
2	0	60.00	
3	0	100.00	
4	0	94.50	

	no_of_special_requests	booking_status	\
0	0	Not Canceled	
1	1	Not Canceled	
2	0	Canceled	
3	0	Canceled	
4	0	Canceled	

▼ Shape

```
[3] # shape of the data
df.shape
```

(36275, 19)

▼ data information

```
[4] #data information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Booking_ID                            36275 non-null  object
1   no_of_adults                          36275 non-null  int64
2   no_of_children                         36275 non-null  int64
3   no_of_weekend_nights                  36275 non-null  int64
4   no_of_week_nights                     36275 non-null  int64
5   type_of_meal_plan                     36275 non-null  object
6   required_car_parking_space            36275 non-null  int64
7   room_type_reserved                    36275 non-null  object
8   lead_time                             36275 non-null  int64
9   arrival_year                          36275 non-null  int64
10  arrival_month                         36275 non-null  int64
11  arrival_date                          36275 non-null  object
12  market_segment_type                   36275 non-null  object
13  repeated_guest                        36275 non-null  int64
14  no_of_previous_cancellations           36275 non-null  int64
15  no_of_previous_bookings_not_canceled  36275 non-null  int64
16  avg_price_per_room                    36275 non-null  float64
17  no_of_special_requests                 36275 non-null  int64
18  booking_status                        36275 non-null  object
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

Membatasi Rows Dataset

```
[5] df = df.sample(n=500, random_state=42)
```

Mengubah Tipe Data Booking_ID dari (Objek) Menjadi (Int)

```
df['Booking_ID'] = df['Booking_ID'].str.replace('INN', '').astype(int)
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 500 entries, 4968 to 6562
Data columns (total 19 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   Booking_ID                            500 non-null    int64
1   no_of_adults                          500 non-null    int64
2   no_of_children                        500 non-null    int64
3   no_of_weekend_nights                  500 non-null    int64
4   no_of_week_nights                     500 non-null    int64
5   type_of_meal_plan                     500 non-null    object
6   required_car_parking_space            500 non-null    int64
7   room_type_reserved                    500 non-null    object
8   lead_time                             500 non-null    int64
9   arrival_year                          500 non-null    int64
10  arrival_month                         500 non-null    int64
11  arrival_date                          500 non-null    int64
12  market_segment_type                   500 non-null    object
13  repeated_guest                        500 non-null    int64
14  no_of_previous_cancellations          500 non-null    int64
15  no_of_previous_bookings_not_canceled  500 non-null    int64
16  avg_price_per_room                    500 non-null    float
17  no_of_special_requests                500 non-null    int64
18  booking_status                        500 non-null    object
dtypes: float64(1), int64(14), object(4)
memory usage: 78.1+ KB
```

Description of the data

```
[7] # describing the data
df.describe()
```

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	18792.118000	1.830000	0.090000	0.760000	0.760000
std	10761.308483	0.534482	0.366290	0.817151	0.817151
min	41.000000	0.000000	0.000000	0.000000	0.000000
25%	9587.750000	2.000000	0.000000	0.000000	0.000000
50%	19264.000000	2.000000	0.000000	1.000000	1.000000
75%	27809.500000	2.000000	0.000000	1.000000	1.000000
max	36237.000000	3.000000	2.000000	4.000000	4.000000

Checking Columns

```
[8] #column to list
df.columns.tolist()
```

```
['Booking_ID',
 'no_of_adults',
 'no_of_children',
 'no_of_weekend_nights',
 'no_of_week_nights',
 'type_of_meal_plan',
 'required_car_parking_space',
 'room_type_reserved',
 'lead_time',
 'arrival_year',
 'arrival_month',
 'arrival_date',
 'market_segment_type',
 'repeated_guest',
 'no_of_previous_cancellations',
 'no_of_previous_bookings_not_canceled',
 'avg_price_per_room',
 'no_of_special_requests',
 'booking_status']
```

Checking Missing Values

```
[9] # check for missing values:
missing_values = df.isnull().sum()
print("Missing value per kolom:")
print(missing_values)
```

```
Missing value per kolom:
Booking_ID          0
no_of_adults        0
no_of_children       0
no_of_weekend_nights 0
no_of_week_nights    0
type_of_meal_plan    0
required_car_parking_space 0
room_type_reserved   0
lead_time            0
arrival_year         0
arrival_month        0
arrival_date         0
market_segment_type  0
repeated_guest       0
no_of_previous_cancellations 0
no_of_previous_bookings_not_canceled 0
avg_price_per_room   0
no_of_special_requests 0
booking_status       0
dtype: int64
```

Checking for the duplicate values

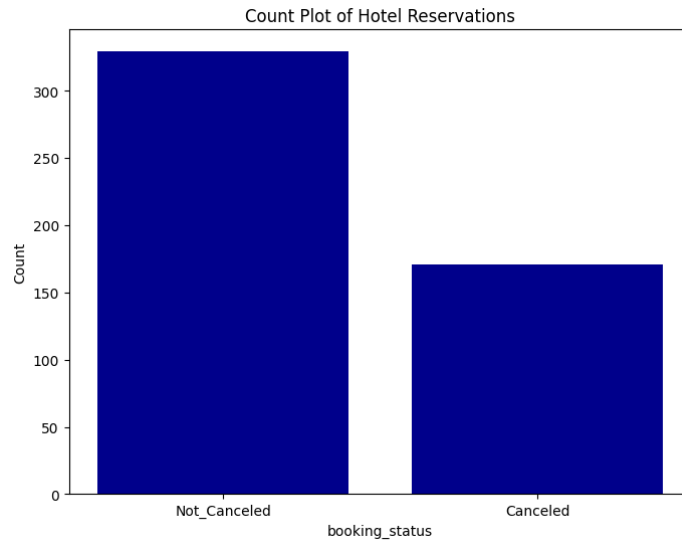
```
[10] #checking duplicate values
duplicate_rows = df[df.duplicated()]
print("Duplikasi baris:")
print(duplicate_rows)
```

```
Duplikasi baris:
Empty DataFrame
Columns: [Booking_ID, no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, type_of_meal_plan, required_car_parking_space, room_type_reserved, lead_time, arrival_year, arrival_month, arrival_date, market_segment_type, repeated_guest, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests, booking_status]
Index: []
```

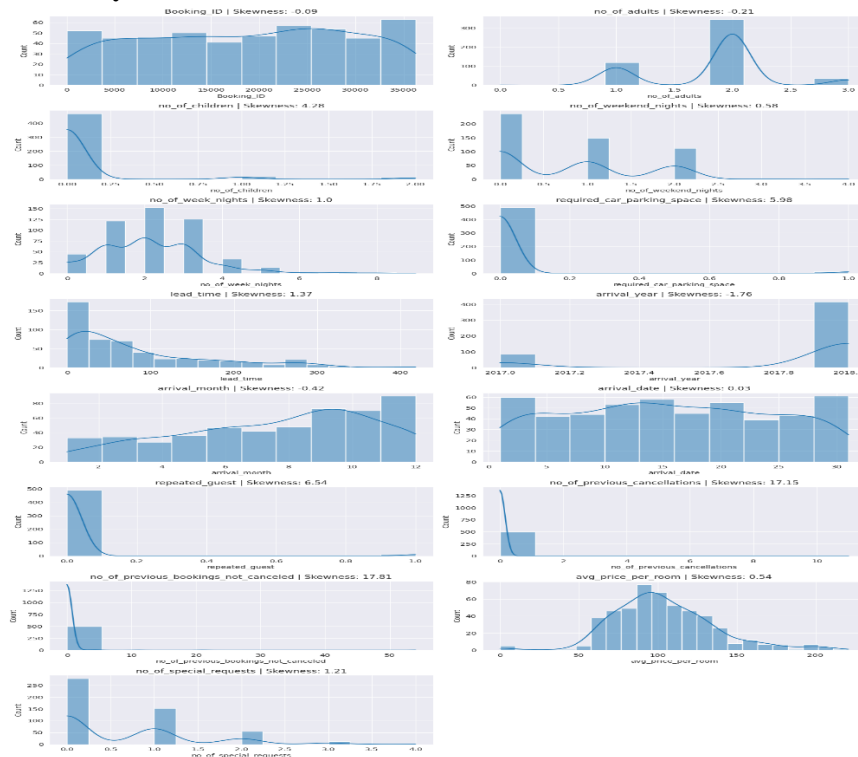
4. Data Preparation

Univariate Analysis

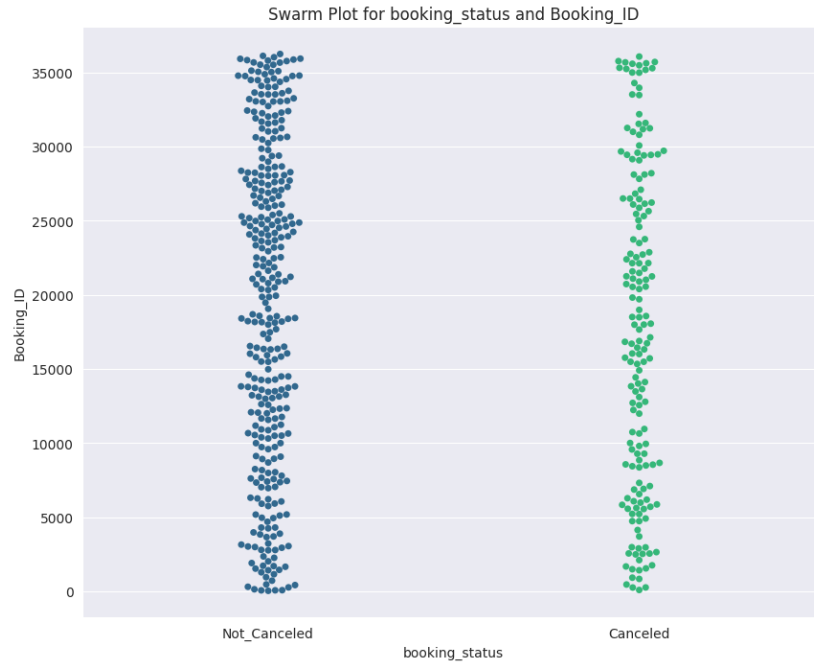
- Count Plot



- Kernel Density Plots



- Swarm Plot



Bivariate Analysis

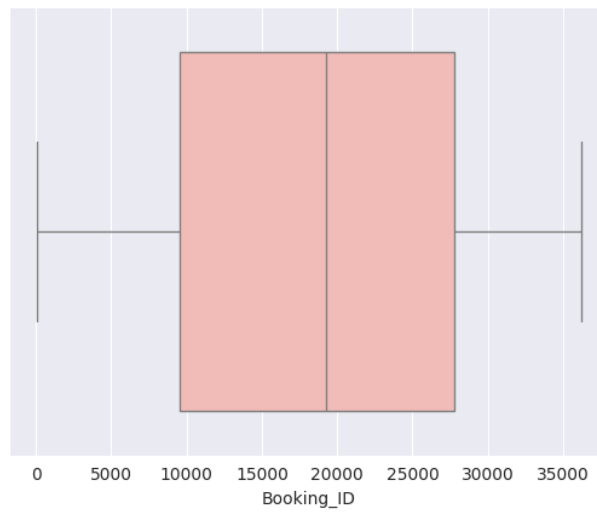
- Pair Plot**



- Violin Plot**

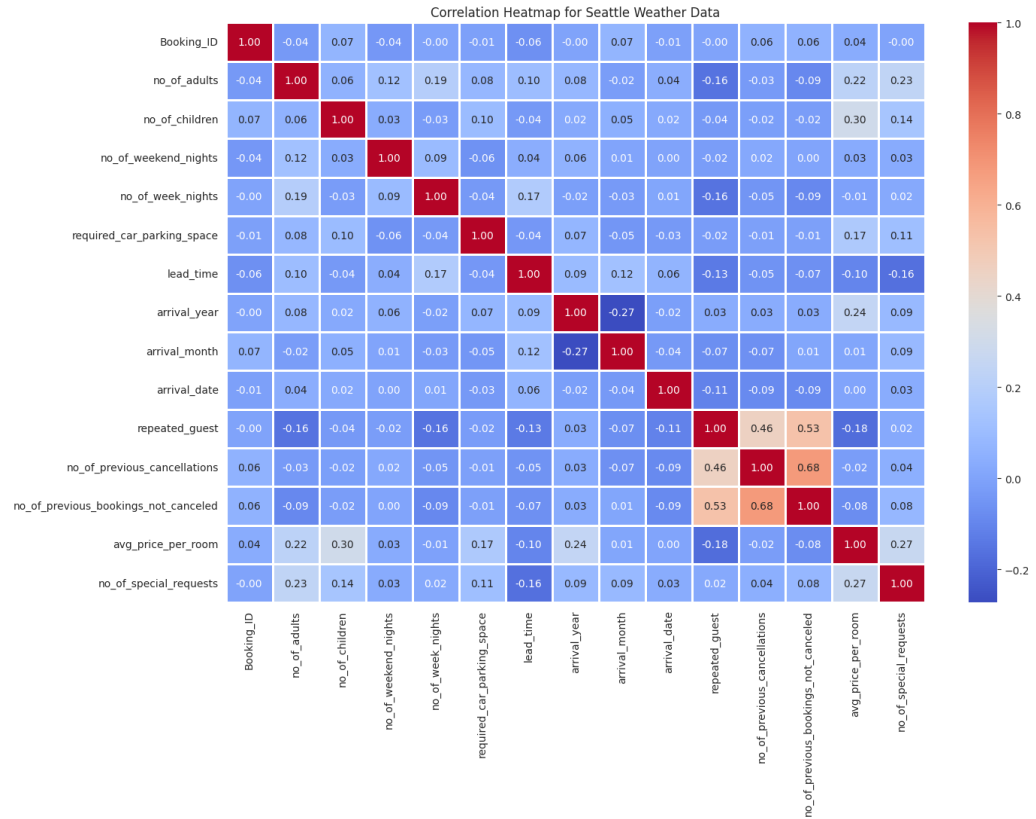


- **Box Plot**

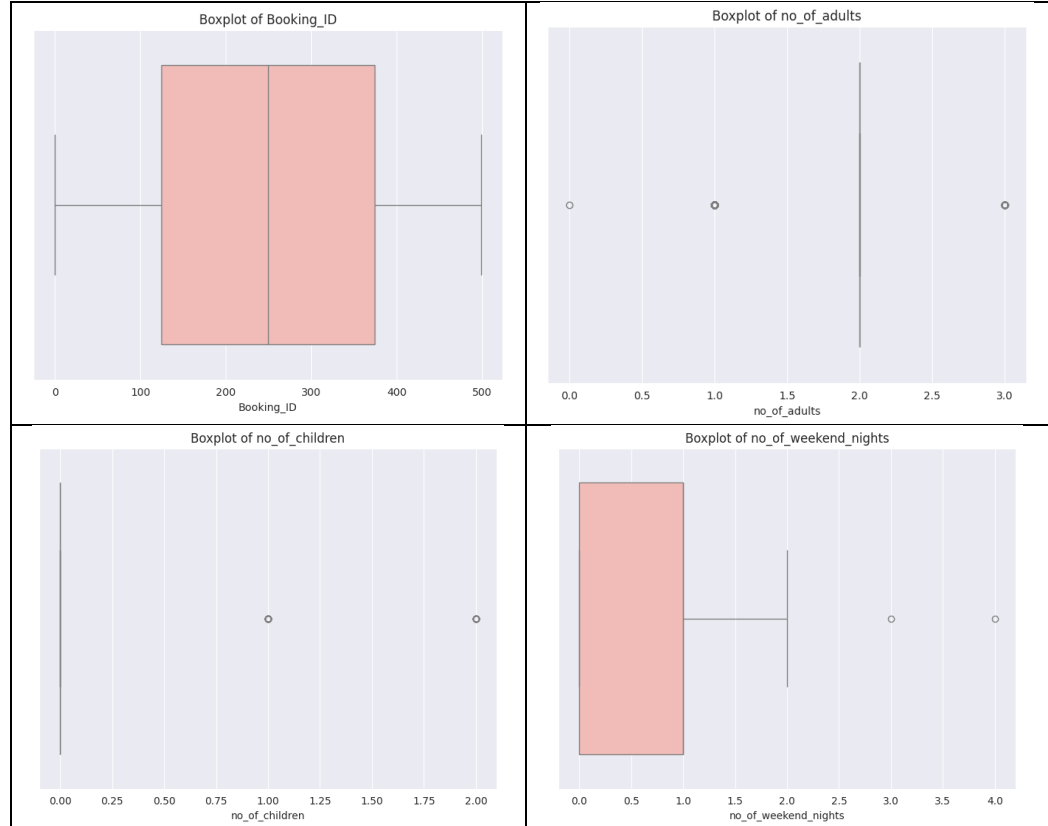


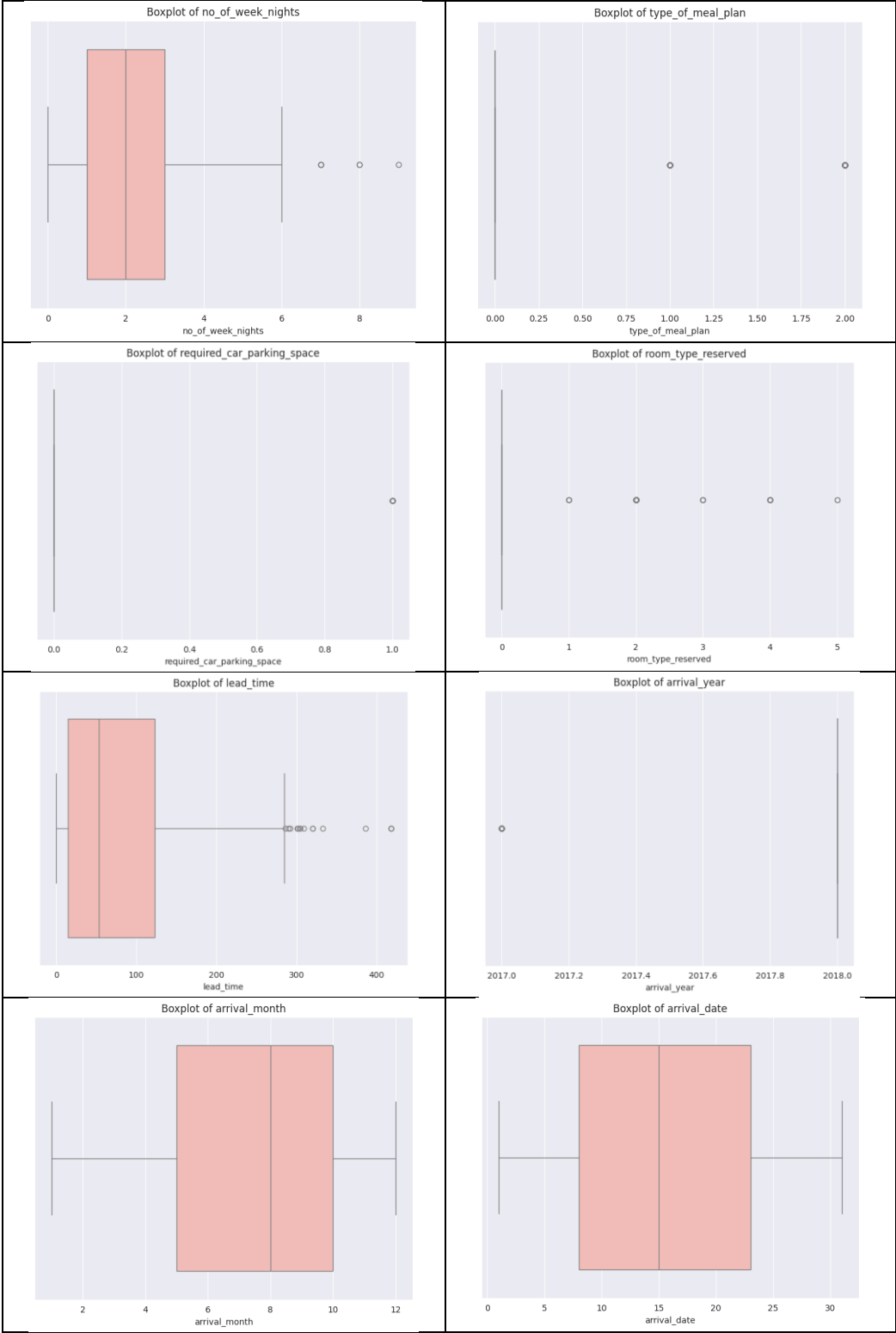
Multivariate Analysis

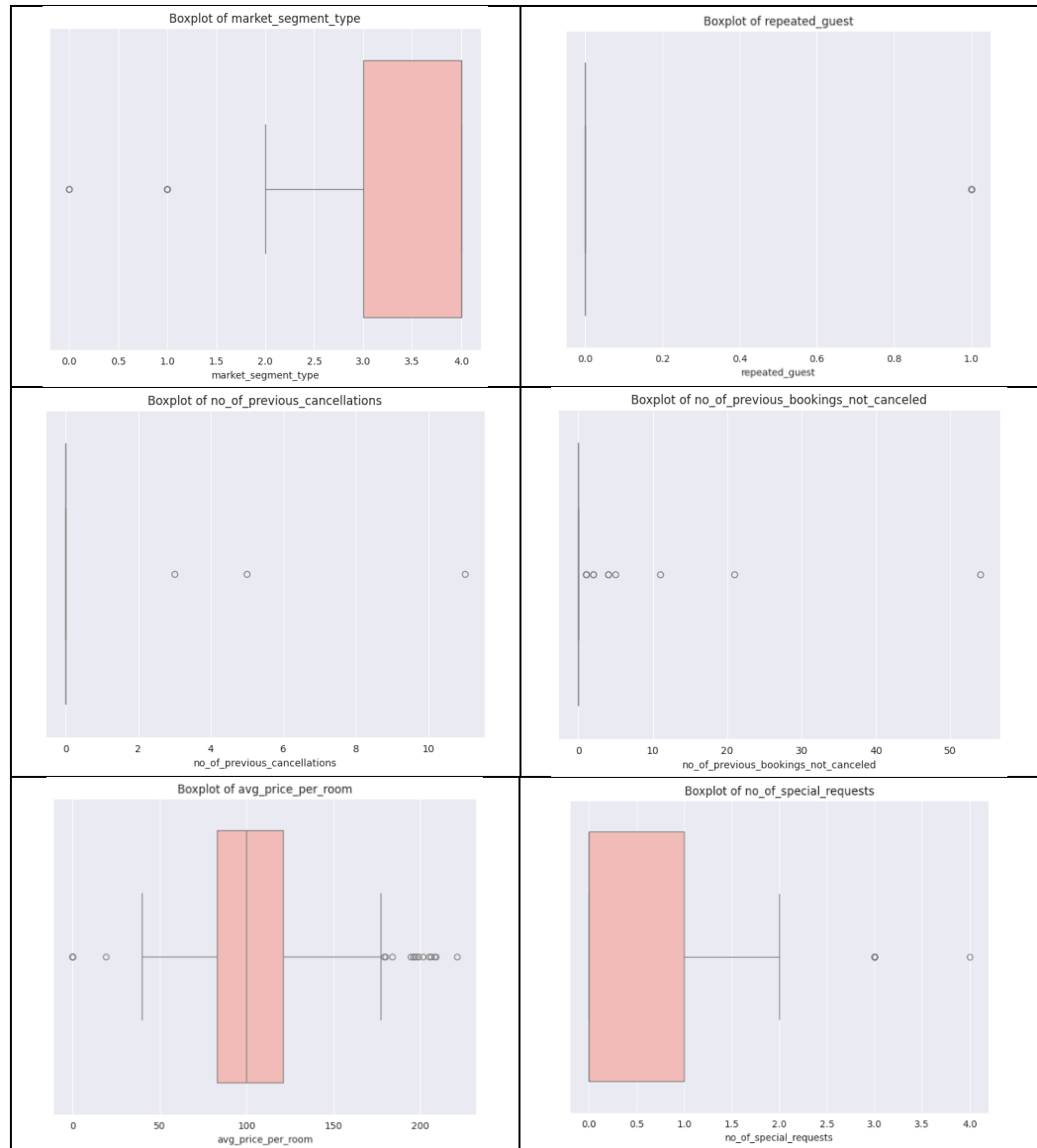
- **Correlation Matrix**



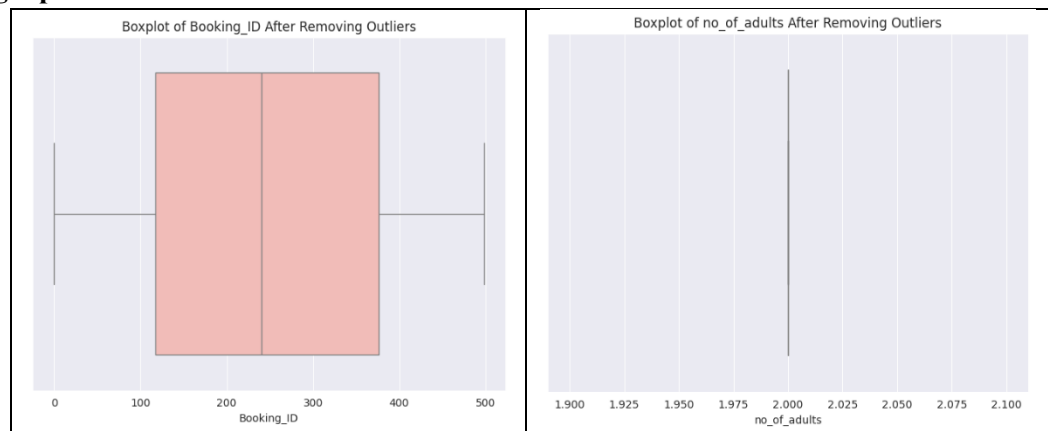
- Outlier

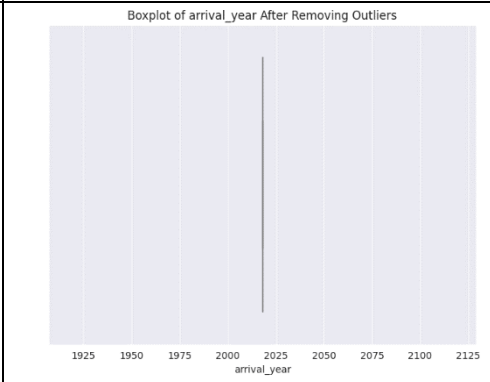
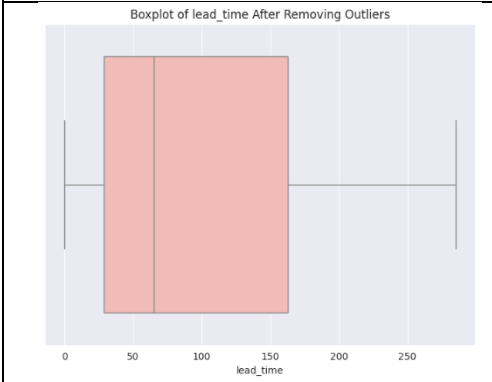
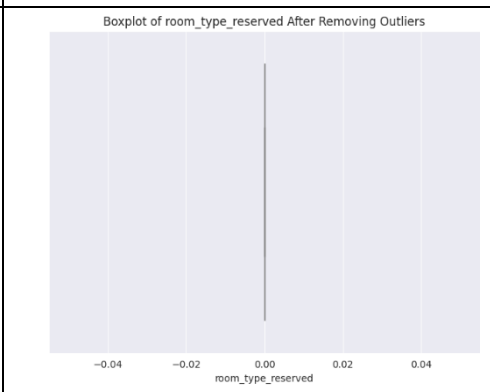
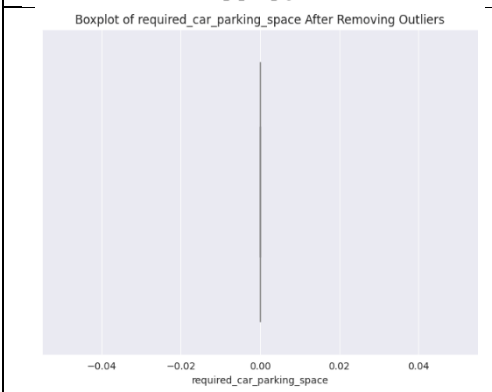
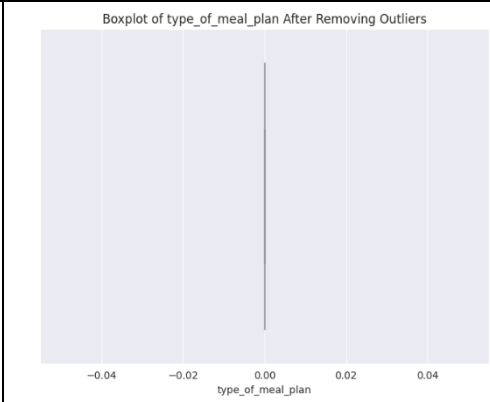
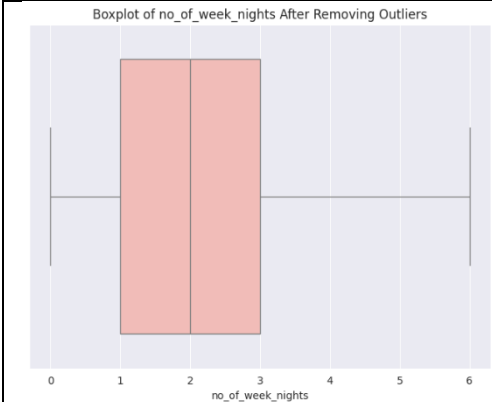
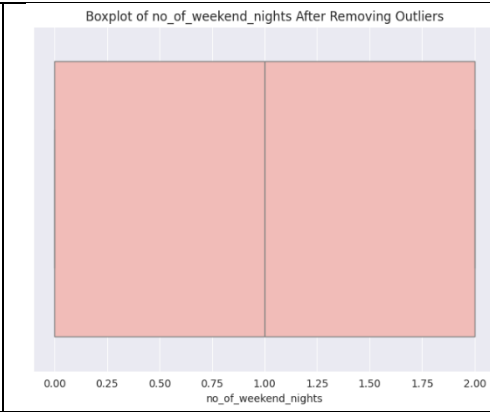
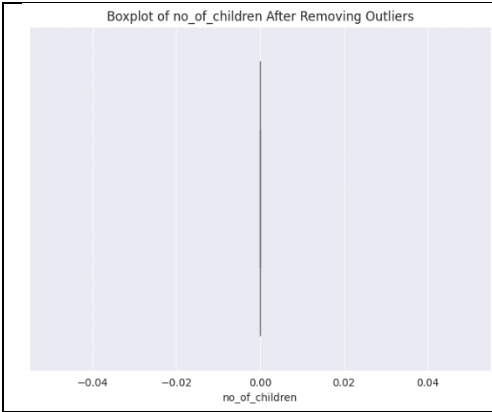


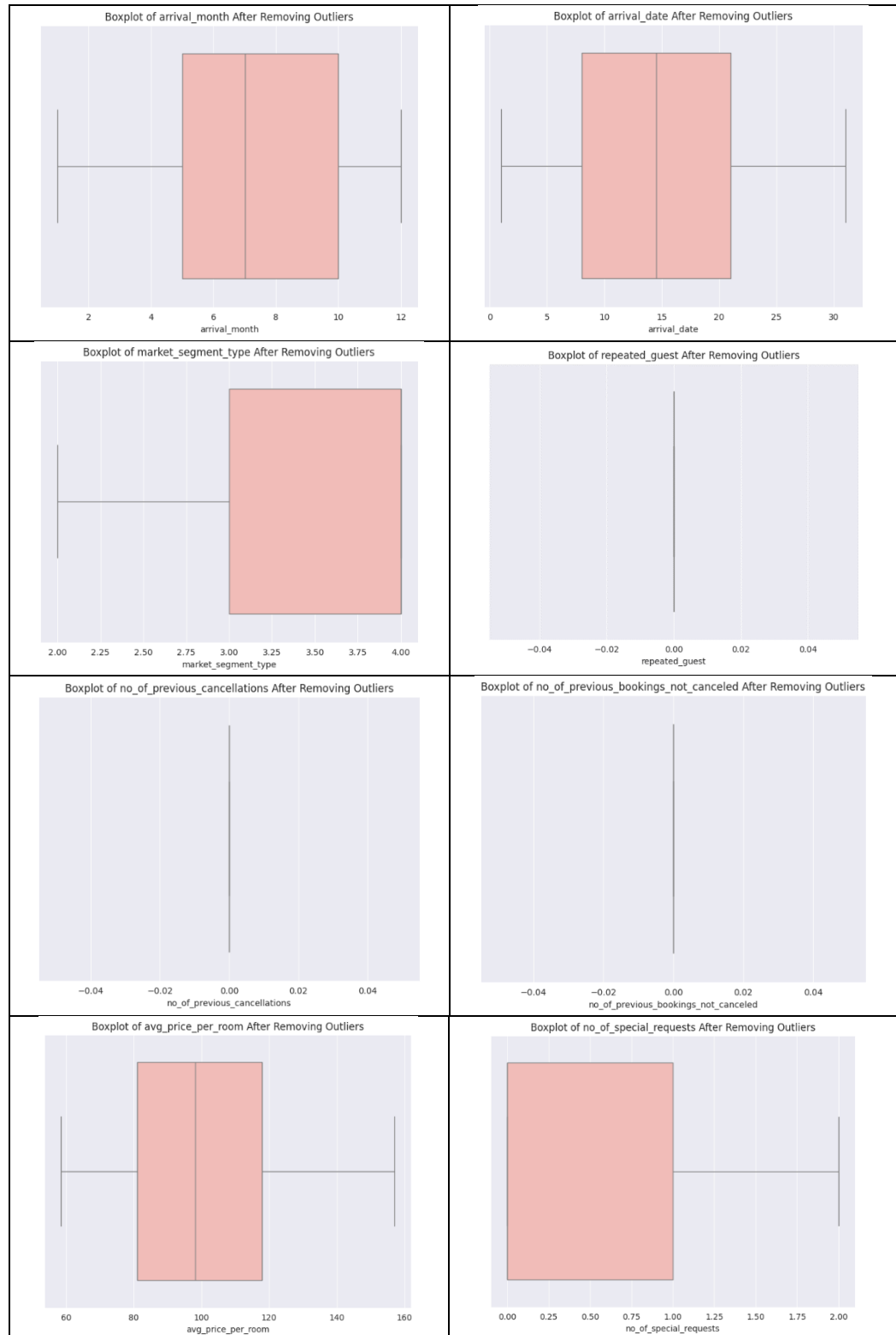




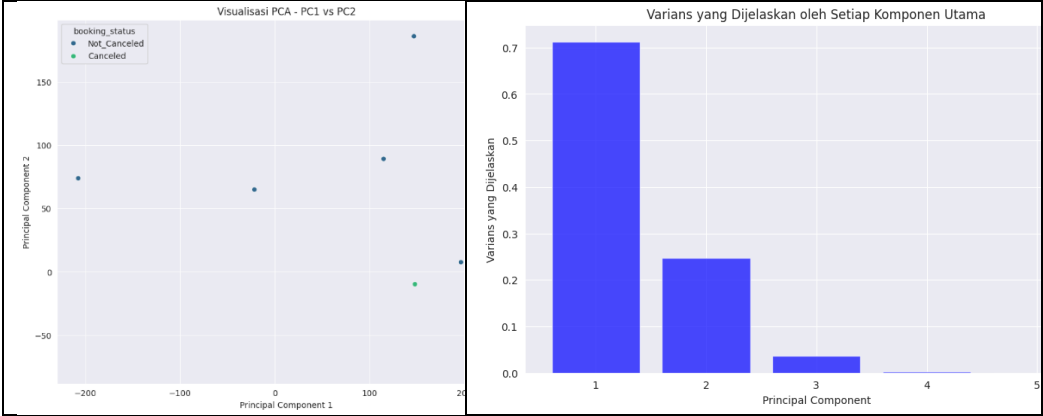
- **Penghapusan Outlier**



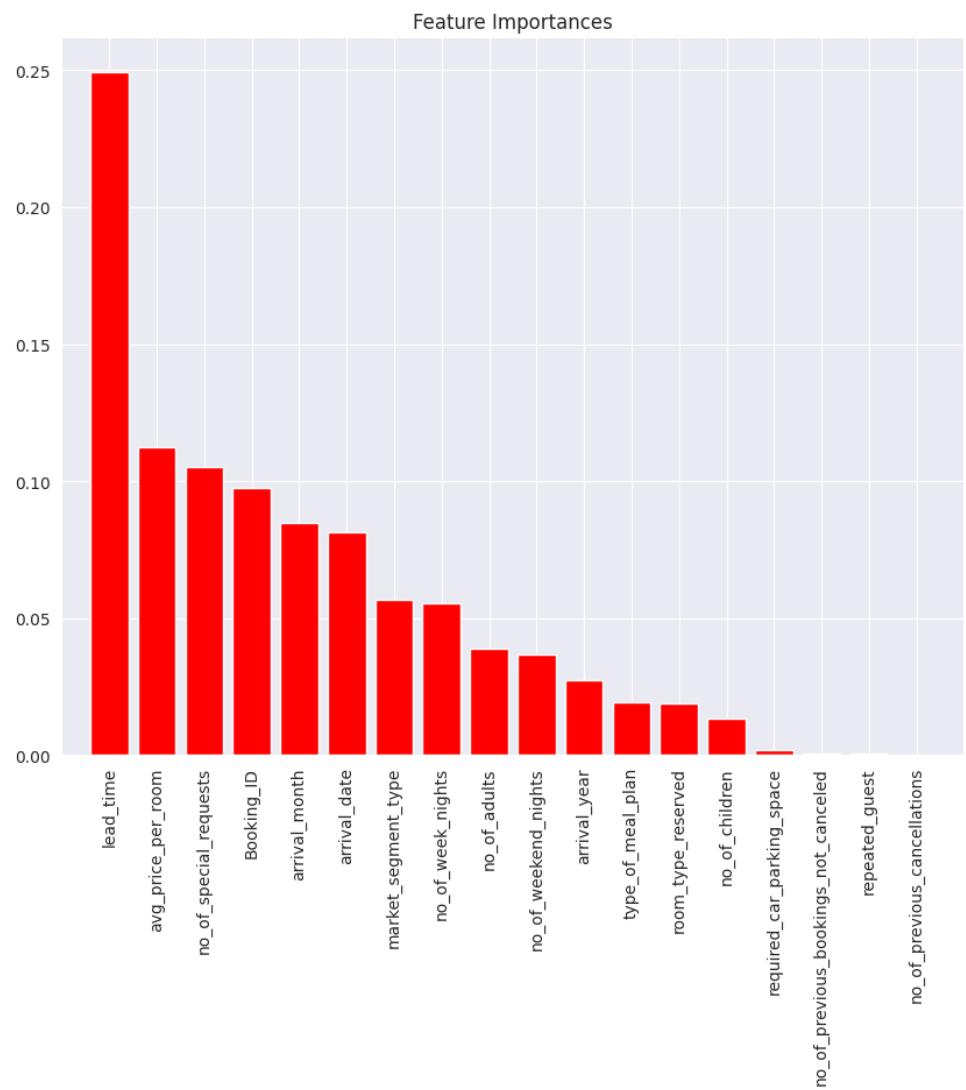




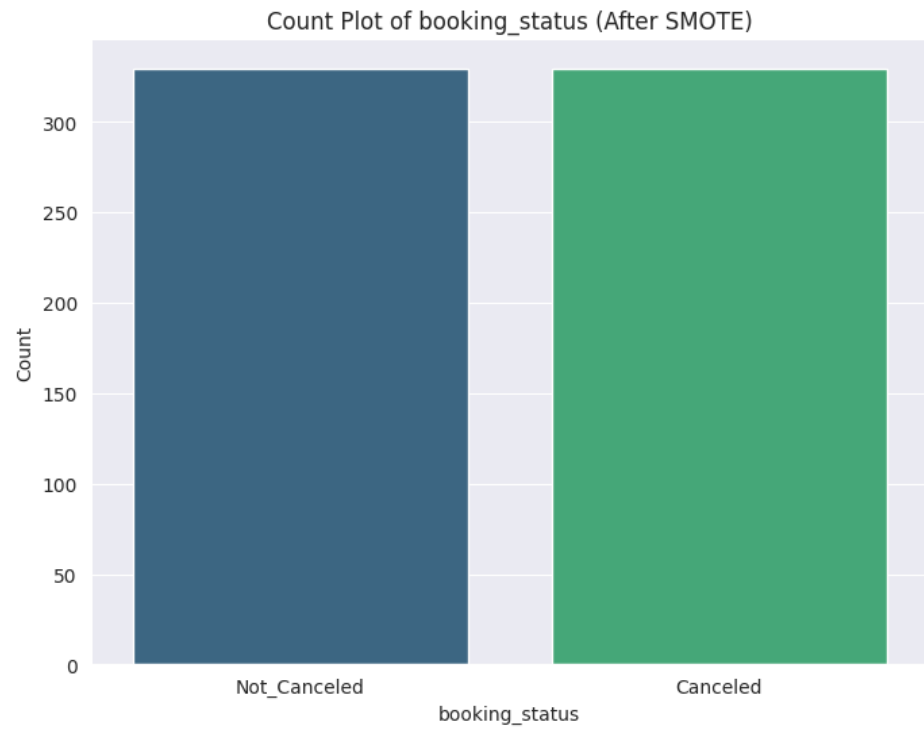
- **Reduksi Fitur**



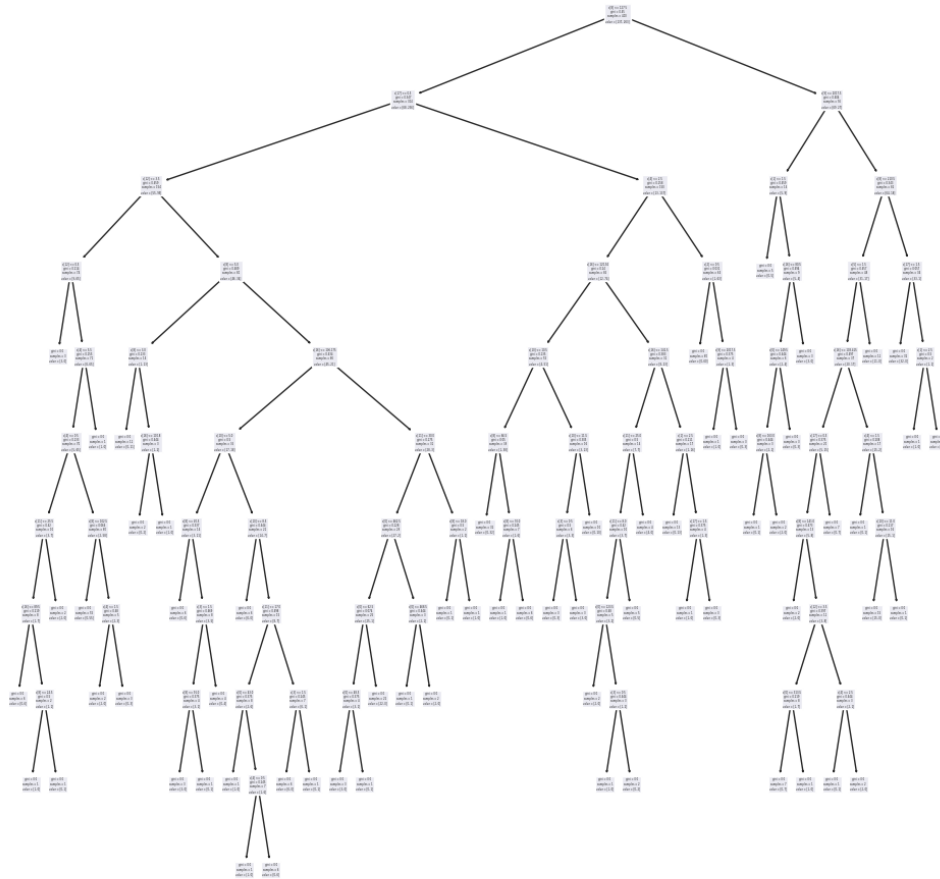
- **Seleksi Fitur**



- **Normalisasi**



5. Modeling



6. Evaluation

Evaluasi model dilakukan menggunakan metrik-metrik klasifikasi standar: Precision, Recall, dan F1-Score, dengan pendekatan validasi menggunakan k-Fold Cross Validation (k=5) dan pembagian data latih-uji sebesar 80:20. Hasil evaluasi dirangkum sebagai berikut:

Metrik	Nilai Rata-rata
Precision	0.84
Recall	0.75
F1-Score	0.82
Accuracy	0.83 (dihitung manual)

Catatan Evaluasi:

- Model memiliki kemampuan presisi yang cukup tinggi (84%), artinya ketika model memprediksi pembatalan, cukup banyak yang benar-benar dibatalkan.
- Recall sebesar 75% menunjukkan bahwa model masih bisa melewati sejumlah pembatalan aktual.

- F1-score yang tinggi (82%) menjadi indikator keseimbangan antara precision dan recall.
- Teknik SMOTE berhasil membantu mengatasi dataset yang tidak seimbang.
- Reduksi fitur dengan PCA dan seleksi fitur dengan Random Forest memberikan kontribusi pada efisiensi model.

7. Kesimpulan

Berdasarkan eksperimen dan evaluasi yang dilakukan, proyek klasifikasi pembatalan reservasi hotel dengan algoritma machine learning berhasil dibangun dan menunjukkan performa yang cukup baik. Dengan F1-score sebesar 82%, model mampu mengklasifikasikan pembatalan reservasi secara cukup akurat. Tujuan utama proyek, yaitu membantu hotel dalam mengantisipasi pembatalan reservasi, telah tercapai melalui sistem prediktif ini. Model memiliki kelebihan dalam:

- Mendeteksi pola pembatalan berdasarkan fitur historis.
- Mengurangi potensi kerugian dari pembatalan mendadak.
- Memiliki performa yang baik meski data asli tidak seimbang.

Namun, beberapa keterbatasan tetap perlu dicatat:

- Kinerja recall masih bisa ditingkatkan agar lebih sedikit kasus pembatalan yang terlewat.
- Model belum diuji dalam kondisi real-time atau pada data terbaru yang mungkin memiliki tren berbeda.

8. Rekomendasi

Untuk pengembangan proyek lebih lanjut, berikut adalah beberapa rekomendasi yang dapat dilakukan:

1. Peningkatan Model:
 - Gunakan algoritma ensemble lain seperti XGBoost atau LightGBM untuk perbandingan.
 - Lakukan hyperparameter tuning menggunakan GridSearchCV atau RandomizedSearchCV.
2. Kualitas Dataset:
 - Tambahkan data reservasi terbaru untuk memperkuat generalisasi model.
 - Pertimbangkan menambahkan variabel eksternal seperti musim liburan, hari kerja vs akhir pekan, dll.
3. Evaluasi Lanjutan:
 - Uji model pada data baru (hold-out set) untuk validasi performa jangka panjang.
 - Implementasi ROC-AUC score sebagai tambahan untuk evaluasi lebih mendalam.
4. Integrasi Sistem:
 - Model dapat diintegrasikan ke sistem reservasi hotel sebagai alat prediksi real-time.
 - Buat dashboard visualisasi prediksi agar user (manajer hotel) mudah menginterpretasikan hasilnya.

9. Referensi

N O	JUDUL/AUTHOR/TAHUN	TOPIK/MASALAH	DATABASE	TEKNIK PROSESING DATA	PEMODELAN	EVALUASI
1	Judul : <i>Analisis data transaksi penjualan untuk mengatasi Toko Usaha Mandiri</i> Author : Fauzia Allamatul Fithri, Sukma Wardana Tahun : 2021	Analisis data transaksi penjualan untuk mengatasi	Nama : Analisis Data Transaksi Penjualan Toko Usaha Mandiri Kolom x Baris : 8 atribut x jumlah produk (92 produk) Tipe Data: Data numerik dan kategorikal	Seleksi Data: Memilih data numerik transaksi penjualan yang diperlukan untuk diproses oleh algoritma K-Means Clustering. Data: Mengubah data dari sumber dalam bentuk klasifikasi atau numerik agar dapat diproses oleh algoritma. Metode Penelitian: Penelitian ini merupakan penelitian terapan yang menggunakan studi dokumentasi dan analisis data untuk menemukan	K-Means Clustering: Metode pengelompokan data yang mengelompokkan data ke dalam satu atau lebih kelompok dengan sistem partisi. Pembentukan Cluster: Pembentukan cluster dilakukan dengan menentukan jumlah kelompok (k) dan centroid awal yang dilakukan secara acak.	<i>Internal Validation Davies Bouldin Index:</i> Evaluasi algoritma dilakukan dengan menghitung nilai <i>Davies Bouldin Index</i> (DBI) yang mendekati nol, menunjukkan kinerja dan akurasi clustering yang baik.

				solusi terhadap masalah yang ada.		
2	Judul: <i>Clustering Data</i> Mahasiswa Menggunakan Algoritma K-Means Author: Pareza Alam, Fadhel Muhammad Irfan, Kurniabudi Tahun: 2019		Nama Data Set: Data mahasiswa angkatan 2018/2019 Sumber Data: Bagian akademik dan admisi STMIK Bina Bangsa Kendari Jumlah Sampel: 171 items Atribut Data: Nama mahasiswa, jurusan, sekolah asal, nilai IPK Tipe Data: Data kategorika	Teknik Data Mining dengan Algoritma K-Means	Algoritma: K-Means Clustering Hasil Pemodelan: Terbentuk dua cluster, yaitu Cluster 1 (72 items) dan Cluster 2 (99 items)	Hasil penelitian digunakan sebagai dasar untuk menentukan strategi promosi program studi yang ada di STMIK Bina Bangsa Kendari berdasarkan hasil cluster program studi yang banyak diminati dari masing-masing sekolah.

			1 dan numerik			
3	Judul: Student Academic Performance Prediction using Supervised Learning Techniques Author: Muhammad Imran, Shahzad Latif, Danish Mehmood, Muhammad Saqlain Shah Tahun: 2019	Prediksi Kinerja Akademik Mahasiswa menggunakan Teknik supervised learning	Asal data set : UCI Machine Learning Repository Kolom x Baris : 33 atribut x 1044 instansi	<ul style="list-style-type: none"> • Data preprocessing dilakukan untuk membersihkan data, mengatasi data yang hilang atau tidak konsisten, dan mengubah format data agar sesuai untuk digunakan dalam algoritma pembelajaran mesin • Class balancing dilakukan untuk menyeimbangkan jumlah instansi antara kelas 	Model prediksi kinerja mahasiswa menggunakan Decision Tree Classifier (J48) sebagai model dasar, dan ditingkatkan dengan Ensemble Method (Realadaboost) untuk meningkatkan akurasi prediksi dan mengurangi kesalahan klasifikasi.	Dalam evaluasi model prediksi kinerja mahasiswa, digunakan metode 10-folds cross validation. Ini berarti data set dibagi menjadi 10 subset yang sama besar, di mana 9 subset digunakan untuk melatih model dan 1 subset digunakan untuk pengujian. Proses ini diulang sebanyak 10 kali untuk mendapatkan estimasi akurasi yang lebih stabil.

				<p>yang berbeda agar model pembelajaran mesin tidak cenderung memihak pada kelas mayoritas.</p> <ul style="list-style-type: none"> • Feature selection dilakukan untuk memilih subset atribut yang paling relevan dan penting untuk membangun model prediksi kinerja mahasiswa. 		
4	Judul: Analisis Perbandingan Algoritma Machine Learning untuk Klasifikasi Tingkat Risiko	Analisis perbandingan algoritma machine learning untuk klasifikasi tingkat risiko ibu hamil.	<p>Nama data set : Maternal Health Risk</p> <p>Kolom x Baris : 7 atribut x</p>	<ul style="list-style-type: none"> • Dalam pra-pemrosesan data, peneliti melakukan eksplorasi data untuk 	Menggunakan algoritma Decision Tree, K-Nearest Neighbors (KNN), dan Naïve Bayes	Melalui akurasi, laporan klasifikasi, dan confusion matrix, dengan Decision Tree

	<p>Ibu Hamil.</p> <p>Author: Rafiqi Aidil Fitra.</p> <p>Tahun: Desember 2023</p>		<p>1014</p> <p>• Tipe Data: Berbagai tipe data seperti integer, float, dan object</p>	<p>mengidentifikasi jenis dan keadaan variabel dalam dataset.</p> <ul style="list-style-type: none"> • Mereka juga melakukan visualisasi data untuk melihat distribusi variabel dan mengidentifikasi outlier. • pembersihan data, peneliti mengidentifikasi outlier dalam dataset, terutama pada variabel HeartRate. • Transformasi data dilakukan untuk 		<p>mencapai akurasi 90%, KNN 86%, dan Naïve Bayes 65%.</p>
--	---	--	--	---	--	--

				mengubah data dalam bentuk teks menjadi dalam bentuk angka, seperti pada kolom RiskLevel.		
5	Judul: Data Mining Dengan Regresi Linier Berganda Untuk Peramalan Tingkat Inflasi Author: Amrin Tahun: 2016	Memprediksi Tingkat Inflasi Bulanan di Indonesia	Nama data set : Data tingkat inflasi bulanan di Indonesia Kolom x Baris : 5 atribut x sample(10 0 sample) Tipe Data : data numerik terkait dengan tingkat inflasi dan variabel independen lainnya yang digunakan	Peneliti menggunakan metode prediksi regresi linier berganda untuk memprediksi tingkat inflasi bulanan di Indonesia. Model regresi linier berganda yang dihasilkan dari analisis data menunjukkan hubungan antara variabel independen (seperti inflasi makanan, perumahan, kesehatan, pendidikan,	Model Regresi Linier Berganda	Evaluasi model regresi linier berganda dilakukan dengan menghitung Mean Absolute Deviation (MAD), Mean Square Error (MSE), dan Root Mean Square Error (RMSE).

			dalam analisis regresi linier berganda.	transportasi, dll.) dengan variabel dependen (tingkat inflasi). Hasil analisis menunjukkan bahwa model regresi linier berganda memiliki koefisien regresi sebesar 0,999 dan koefisien determinasi sebesar 0,997, yang menunjukkan tingkat akurasi prediksi yang sangat baik.		
6	Judul: Implementasi Ridge Regression untuk Mengatasi Gejala Multikolinearitas pada Pemodelan Curah Hujan	mengatasi gejala multikolinearitas dalam pemodelan curah hujan berbasis data time series klimatologi.	Kolom x Baris : - Tipe Data : Date time series	Ridge Regression	Pemodelan regresi untuk memprediksi curah hujan berdasarkan variabel bebas seperti temperatur udara, temperatur air,	Perbandingan Hasil Prediksi: Berdasarkan nilai ukuran kebaikan model, disimpulkan bahwa model Ridge Regression

	<p>Berbasis Data Time Series Klimatologi</p> <p>Author :</p> <p>Risa Arisandi, Dadang Ruhiat, dan Emas Marlina</p> <p>Tahun :</p>				<p>penguapan, dan kelembaban.</p>	<p>memberikan hasil prediksi yang lebih baik daripada model Regresi Linier Berganda .</p> <p>Rekomendasi: Penulis merekomendasikan peneliti selanjutnya untuk mencoba variasi lain dalam menentukan tetapan bias c dan membandingkan hasilnya dengan pemilihan tetapan bias menggunakan rumus tertentu</p>
--	---	--	--	--	-----------------------------------	--

DAFTAR PUSTAKA

- [1] F. Nur Anisa and S. Wibowo, “Penggunaan Machine Learning Dalam Aplikasi Rekomendasi Tempat Wisata Jawa Tengah (Se-Jawat Tourish),” *Sci. Eng. Natl. Semin.*, vol. 7, no. 1, pp. 1–11, 2022.
- [2] “Pengertian Data, Fungsi, Jenis-jenis, Manfaat dan Contohnya.” [Online]. Available: Pengertian Data, Fungsi, Jenis-jenis, Manfaat dan Contohnya
- [3] T. Atribut, “Data Mining,” pp. 1–8.
- [4] K. Khotimah, “PERANCANGAN DAN IMPLEMENTASI DATA WAREHOUSE UNTUK Mendukung Sistem Akademik (Studi Kasus Pada STKIP Muhammadiyah Kotabumi),” vol. 02, no. 01, pp. 94–107, 2016.
- [5] “No Title,” vol. 1, no. 1, pp. 1–7, 2022.
- [6] “Pengetahuan (Knowledge).” [Online]. Available: <https://online.binus.ac.id/computer-science/post/pengetahuan-knowledge>