LAPORAN PROJECT UAS DATA MINING & DATA WAREHOUSE

"PERBANDINGAN METODE KLASIFIKASI PADA DATA KEPUASAN PENUMPANG PENERBANGAN"

Dosen Pengampu :

Ika Nurlaili Isnainiyah, S.Kom., M.Sc.

Disusun oleh : Kelompok 1

2210511046	Hanifah Az Zahra
2210511052	Widya Amellia Putri
2210511056	Adinda Rizki Sya'bana Diva
2210511068	Muhammad Nur Alam
2210511075	Kemas Alauddin Riayat Syah Indrakusuma
2210511084	Dzulfikri Adjmal

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAKARTA

2023

Laporan Proyek Ujian Akhir Semester
Praktikum Data Mining dan Data Warehouse

December 7, 2023

1 Pendahuluan

Data adalah sumber informasi yang bentuknya masih mentah dan menggambarkan suatu kejadian kejadian dan kesatuan nyata. Data dapat diperoleh dalam bentuk simbol-simbol karakter huruf, angka, gambar, suara, dan lain sebagainya. Agar data digunakan maka data harus diolah terlebih dahulu.

Data mining adalah proses pengumpulan dan pengolahan informasi data yang memiliki tujuan untuk mencari informasi penting pada data. Proses pengumpulan dan pencarian informasi tersebut dapat dilakukan dengan menggunakan perangkat lunak dengan bantuana perhitungan statistika, matematika atau teknologi *Machine Learning* (ML) maupun *Artificial Intelligence* (AI).

Proses Data Mining memiliki tujuan untuk sebagai sarana eksplorasi, konfirmasi dan eksplanasi. Terdapat beberapa metode untuk melakukan proses data mining, yaitu *Association* yang berbasis hubungan variabel dalam dataset, *Classification* yang digunakan untuk memprediksi suatu kelas, *Regression* yang menjelaskan variabel dependen melalui proses analisis variabel independen, dan *Clustering* yang digunakan untuk membagi kumpulan data menjadi beberapa kelompok atau disebut *cluster* berdasarkan kemiripan atribut yang dimiliki.

Proses tersebut dapat diterapkan di berbagai dataset, termasuk dataset yang kami pilih. Dataset yang kami pilih adalah dataset `Airline Passenger Satisfaction` yang berisi kumpulan data penilaian pelanggan terhadap layanan yang diberikan oleh maskapai. Metode yang kami gunakan untuk dataset ini adalah metode klasifikasi. Dengan penggunaan metode klasifikasi, kami ingin membuat sebuah model yang dapat digunakan untuk memprediksi status kepuasan pelanggan berdasarkan beberapa kolom penilaian yang diberikan, hasil prediksi tersebut dibagi menjadi dua label yaitu label neutral or dissatisfied yang artinya pelanggan tersebut netral atau tidak puas dengan pelayanan yang diberikan dan label satisfied yang artinya pelanggan tersebut puas dengan pelayanan yang diberikan.

Kami juga ingin melihat sebuah perbandingan dari tiga algoritma metode klasifikasi yang kami pilih yaitu Naive Bayes, K-Nearest Neighbor, dan Decision Tree. Naive Bayes merupakan algoritma klasifikasi yang berdasarkan probabilitas, K-Nearest Neighbor adalah algoritma yang berdasarkana dengan jarak antar data, dan Decision Tree yang membuat prediksi menggunakan struktur pohon. Dalam membandingkan ketiga algoritma tersebut, kami menggunakan acuan nilai confusion matrix untuk menghitung *Accuracy*, *Precision*, dan *Recall* setiap algoitma untuk dibandingkan.

2 Pembahasan

2.1 Preprocessing Data

2.1.1 Gathering

Dataset yang kami peroleh berasal dari situs kaggle. Dataset ini terdiri dari dua file yaitu `train.csv` yang terdiri dari 103904 baris dan `test.csv` terdiri dari 25976 baris. Kedua dataset ini memiliki 24 kolom dengan rincian sebagai berikut:

- id
- Gender
- Customer Type
- Age
- Type of Travel
- Class
- Flight Distance
- Inflight wifi service
- Departure/Arrival time convenient
- Ease of Online booking
- Gate location
- Food and drink
- Online boarding
- Seat comfort
- Inflight entertainment
- On-board service
- Leg room service
- Baggage handling
- Checkin service
- Inflight service

- Cleanliness
- Departure Delay in Minutes
- Arrival Delay in Minutes
- satisfaction

Bentuk dataset yang kami ambil termasuk data terstruktur yang memiliki kolom dan baris. Jenis pemodelan data kami yaitu supervised sehingga kami menggunakan metode klasifikasi dalam analisis data.

2.1.2 Assesing Data

[1]:
```
# Support Libraries
import pandas as pd
import math
from six import StringIO
import pydotplus
import os
import shutil

# Classification and Clustering Libraries
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.metrics import accuracy_score, confusion_matrix,␣
 ↪classification_report

# Visualization Libraries
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px

df = pd.read_csv('./Dataset/Airline-Passenger-Satisfaction/train.csv',␣
 ↪index_col=0)
df_test = pd.read_csv('./Dataset/Airline-Passenger-Satisfaction/test.csv',␣
 ↪index_col=0)
```

Menampilkan metadata dari dataset

[2]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 103904 entries, 0 to 103903
Data columns (total 24 columns):
```

```
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   id                                103904 non-null  int64
 1   Gender                            103904 non-null  object
 2   Customer Type                     103904 non-null  object
 3   Age                               103904 non-null  int64
 4   Type of Travel                    103904 non-null  object
 5   Class                             103904 non-null  object
 6   Flight Distance                   103904 non-null  int64
 7   Inflight wifi service             103904 non-null  int64
 8   Departure/Arrival time convenient 103904 non-null  int64
 9   Ease of Online booking            103904 non-null  int64
 10  Gate location                     103904 non-null  int64
 11  Food and drink                    103904 non-null  int64
 12  Online boarding                   103904 non-null  int64
 13  Seat comfort                      103904 non-null  int64
 14  Inflight entertainment            103904 non-null  int64
 15  On-board service                  103904 non-null  int64
 16  Leg room service                  103904 non-null  int64
 17  Baggage handling                  103904 non-null  int64
 18  Checkin service                   103904 non-null  int64
 19  Inflight service                  103904 non-null  int64
 20  Cleanliness                       103904 non-null  int64
 21  Departure Delay in Minutes        103904 non-null  int64
 22  Arrival Delay in Minutes          103594 non-null  float64
 23  satisfaction                      103904 non-null  object
dtypes: float64(1), int64(18), object(5)
memory usage: 19.8+ MB
```

[3]:
```
df_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 25976 entries, 0 to 25975
Data columns (total 24 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   id                                25976 non-null  int64
 1   Gender                            25976 non-null  object
 2   Customer Type                     25976 non-null  object
 3   Age                               25976 non-null  int64
 4   Type of Travel                    25976 non-null  object
 5   Class                             25976 non-null  object
 6   Flight Distance                   25976 non-null  int64
 7   Inflight wifi service             25976 non-null  int64
 8   Departure/Arrival time convenient 25976 non-null  int64
 9   Ease of Online booking            25976 non-null  int64
 10  Gate location                     25976 non-null  int64
 11  Food and drink                    25976 non-null  int64
```

```
 12  Online boarding                    25976 non-null  int64
 13  Seat comfort                       25976 non-null  int64
 14  Inflight entertainment             25976 non-null  int64
 15  On-board service                   25976 non-null  int64
 16  Leg room service                   25976 non-null  int64
 17  Baggage handling                   25976 non-null  int64
 18  Checkin service                    25976 non-null  int64
 19  Inflight service                   25976 non-null  int64
 20  Cleanliness                        25976 non-null  int64
 21  Departure Delay in Minutes         25976 non-null  int64
 22  Arrival Delay in Minutes           25893 non-null  float64
 23  satisfaction                       25976 non-null  object
dtypes: float64(1), int64(18), object(5)
memory usage: 5.0+ MB
```

Menampilkan 5 record data teratas

[4]:
```
df.head()
```

[4]:

	id	Gender	Customer Type	Age	Type of Travel	Class	\
0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	
1	5047	Male	disloyal Customer	25	Business travel	Business	
2	110028	Female	Loyal Customer	26	Business travel	Business	
3	24026	Female	Loyal Customer	25	Business travel	Business	
4	119299	Male	Loyal Customer	61	Business travel	Business	

	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	\
0	460	3	4	
1	235	3	2	
2	1142	2	2	
3	562	2	5	
4	214	3	3	

	Ease of Online booking	…	Inflight entertainment	On-board service	\
0	3	…	5	4	
1	3	…	1	1	
2	2	…	5	4	
3	5	…	2	2	
4	3	…	3	3	

	Leg room service	Baggage handling	Checkin service	Inflight service	\
0	3	4	4	5	
1	5	3	1	4	
2	3	4	4	4	
3	5	3	1	4	
4	4	4	3	3	

	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	\

```
0            5                        25                      18.0
1            1                         1                       6.0
2            5                         0                       0.0
3            2                        11                       9.0
4            3                         0                       0.0

              satisfaction
0  neutral or dissatisfied
1  neutral or dissatisfied
2                satisfied
3  neutral or dissatisfied
4                satisfied

[5 rows x 24 columns]
```

[5]:
```
df_test.head()
```

[5]:
```
      id  Gender      Customer Type  Age   Type of Travel     Class  \
0  19556  Female     Loyal Customer   52  Business travel       Eco
1  90035  Female     Loyal Customer   36  Business travel  Business
2  12360    Male  disloyal Customer   20  Business travel       Eco
3  77959    Male     Loyal Customer   44  Business travel  Business
4  36875  Female     Loyal Customer   49  Business travel       Eco

   Flight Distance  Inflight wifi service  Departure/Arrival time convenient  \
0              160                      5                                  4
1             2863                      1                                  1
2              192                      2                                  0
3             3377                      0                                  0
4             1182                      2                                  3

   Ease of Online booking  …  Inflight entertainment  On-board service  \
0                       3  …                       5                 5
1                       3  …                       4                 4
2                       2  …                       2                 4
3                       0  …                       1                 1
4                       4  …                       2                 2

   Leg room service  Baggage handling  Checkin service  Inflight service  \
0                 5                 5                2                 5
1                 4                 4                3                 4
2                 1                 3                2                 2
3                 1                 1                3                 1
4                 2                 2                4                 2

   Cleanliness  Departure Delay in Minutes  Arrival Delay in Minutes  \
0            5                         50                      44.0
```

```
1            5                         0                      0.0
2            2                         0                      0.0
3            4                         0                      6.0
4            4                         0                     20.0

             satisfaction
0               satisfied
1               satisfied
2  neutral or dissatisfied
3               satisfied
4               satisfied

[5 rows x 24 columns]
```

Menggunakan method `describe()` untuk melihat distribusi statistik

[6]:
```
df.describe()
```

[6]:
```
                 id            Age  Flight Distance  Inflight wifi service  \
count  103904.000000  103904.000000    103904.000000          103904.000000
mean    64924.210502      39.379706      1189.448375               2.729683
std     37463.812252      15.114964       997.147281               1.327829
min         1.000000       7.000000        31.000000               0.000000
25%     32533.750000      27.000000       414.000000               2.000000
50%     64856.500000      40.000000       843.000000               3.000000
75%     97368.250000      51.000000      1743.000000               4.000000
max    129880.000000      85.000000      4983.000000               5.000000

       Departure/Arrival time convenient  Ease of Online booking  \
count                      103904.000000           103904.000000
mean                            3.060296                2.756901
std                             1.525075                1.398929
min                             0.000000                0.000000
25%                             2.000000                2.000000
50%                             3.000000                3.000000
75%                             4.000000                4.000000
max                             5.000000                5.000000

       Gate location  Food and drink  Online boarding   Seat comfort  \
count  103904.000000   103904.000000    103904.000000  103904.000000
mean        2.976883        3.202129         3.250375       3.439396
std         1.277621        1.329533         1.349509       1.319088
min         0.000000        0.000000         0.000000       0.000000
25%         2.000000        2.000000         2.000000       2.000000
50%         3.000000        3.000000         3.000000       4.000000
75%         4.000000        4.000000         4.000000       5.000000
max         5.000000        5.000000         5.000000       5.000000
```

```
       Inflight entertainment  On-board service  Leg room service  \
count           103904.000000     103904.000000     103904.000000
mean                 3.358158          3.382363          3.351055
std                  1.332991          1.288354          1.315605
min                  0.000000          0.000000          0.000000
25%                  2.000000          2.000000          2.000000
50%                  4.000000          4.000000          4.000000
75%                  4.000000          4.000000          4.000000
max                  5.000000          5.000000          5.000000

       Baggage handling  Checkin service  Inflight service    Cleanliness  \
count     103904.000000    103904.000000     103904.000000  103904.000000
mean           3.631833         3.304290          3.640428       3.286351
std            1.180903         1.265396          1.175663       1.312273
min            1.000000         0.000000          0.000000       0.000000
25%            3.000000         3.000000          3.000000       2.000000
50%            4.000000         3.000000          4.000000       3.000000
75%            5.000000         4.000000          5.000000       4.000000
max            5.000000         5.000000          5.000000       5.000000

       Departure Delay in Minutes  Arrival Delay in Minutes
count                103904.000000             103594.000000
mean                     14.815618                 15.178678
std                      38.230901                 38.698682
min                       0.000000                  0.000000
25%                       0.000000                  0.000000
50%                       0.000000                  0.000000
75%                      12.000000                 13.000000
max                    1592.000000               1584.000000
```

[7]:
```
df_test.describe()
```

[7]:
```
                 id           Age  Flight Distance  Inflight wifi service  \
count   25976.000000  25976.000000     25976.000000           25976.000000
mean    65005.657992     39.620958      1193.788459               2.724746
std     37611.526647     15.135685       998.683999               1.335384
min        17.000000      7.000000        31.000000               0.000000
25%     32170.500000     27.000000       414.000000               2.000000
50%     65319.500000     40.000000       849.000000               3.000000
75%     97584.250000     51.000000      1744.000000               4.000000
max    129877.000000     85.000000      4983.000000               5.000000

       Departure/Arrival time convenient  Ease of Online booking  \
count                       25976.000000            25976.000000
mean                            3.046812                2.756775
std                             1.533371                1.412951
min                             0.000000                0.000000
```

25%	2.000000	2.000000
50%	3.000000	3.000000
75%	4.000000	4.000000
max	5.000000	5.000000

	Gate location	Food and drink	Online boarding	Seat comfort	\
count	25976.000000	25976.000000	25976.000000	25976.000000	
mean	2.977094	3.215353	3.261665	3.449222	
std	1.282133	1.331506	1.355536	1.320090	
min	1.000000	0.000000	0.000000	1.000000	
25%	2.000000	2.000000	2.000000	2.000000	
50%	3.000000	3.000000	4.000000	4.000000	
75%	4.000000	4.000000	4.000000	5.000000	
max	5.000000	5.000000	5.000000	5.000000	

	Inflight entertainment	On-board service	Leg room service	\
count	25976.000000	25976.000000	25976.000000	
mean	3.357753	3.385664	3.350169	
std	1.338299	1.282088	1.318862	
min	0.000000	0.000000	0.000000	
25%	2.000000	2.000000	2.000000	
50%	4.000000	4.000000	4.000000	
75%	4.000000	4.000000	4.000000	
max	5.000000	5.000000	5.000000	

	Baggage handling	Checkin service	Inflight service	Cleanliness	\
count	25976.000000	25976.000000	25976.000000	25976.000000	
mean	3.633238	3.314175	3.649253	3.286226	
std	1.176525	1.269332	1.180681	1.319330	
min	1.000000	1.000000	0.000000	0.000000	
25%	3.000000	3.000000	3.000000	2.000000	
50%	4.000000	3.000000	4.000000	3.000000	
75%	5.000000	4.000000	5.000000	4.000000	
max	5.000000	5.000000	5.000000	5.000000	

	Departure Delay in Minutes	Arrival Delay in Minutes
count	25976.00000	25893.000000
mean	14.30609	14.740857
std	37.42316	37.517539
min	0.00000	0.000000
25%	0.00000	0.000000
50%	0.00000	0.000000
75%	12.00000	13.000000
max	1128.00000	1115.000000

2.1.3 Data Cleaning

Membersihkan missing value dan menghapus kolom indeks pada data, kemudian disimpan ke file `.csv`

Melihat total jumlah missing value di setiap kolom menggunakan method `isnull()` dan `sum()`

[8]:
```
df.isnull().sum()
```

[8]:
```
id                                     0
Gender                                 0
Customer Type                          0
Age                                    0
Type of Travel                         0
Class                                  0
Flight Distance                        0
Inflight wifi service                  0
Departure/Arrival time convenient      0
Ease of Online booking                 0
Gate location                          0
Food and drink                         0
Online boarding                        0
Seat comfort                           0
Inflight entertainment                 0
On-board service                       0
Leg room service                       0
Baggage handling                       0
Checkin service                        0
Inflight service                       0
Cleanliness                            0
Departure Delay in Minutes             0
Arrival Delay in Minutes             310
satisfaction                           0
dtype: int64
```

[9]:
```
df_test.isnull().sum()
```

[9]:
```
id                                     0
Gender                                 0
Customer Type                          0
Age                                    0
Type of Travel                         0
Class                                  0
Flight Distance                        0
Inflight wifi service                  0
Departure/Arrival time convenient      0
Ease of Online booking                 0
Gate location                          0
Food and drink                         0
```

```
Online boarding                    0
Seat comfort                       0
Inflight entertainment             0
On-board service                   0
Leg room service                   0
Baggage handling                   0
Checkin service                    0
Inflight service                   0
Cleanliness                        0
Departure Delay in Minutes         0
Arrival Delay in Minutes          83
satisfaction                       0
dtype: int64
```

Membersihkan missing value menggunakan method `dropna` untuk menghapus record yang memiliki `NaN`

[10]:
```
df.dropna(inplace=True)
```

[11]:
```
df_test.dropna(inplace=True)
```

Melakukan pengecekan kembali jumlah missing value

[12]:
```
df.isnull().sum()
```

[12]:
```
id                                 0
Gender                             0
Customer Type                      0
Age                                0
Type of Travel                     0
Class                              0
Flight Distance                    0
Inflight wifi service              0
Departure/Arrival time convenient  0
Ease of Online booking             0
Gate location                      0
Food and drink                     0
Online boarding                    0
Seat comfort                       0
Inflight entertainment             0
On-board service                   0
Leg room service                   0
Baggage handling                   0
Checkin service                    0
Inflight service                   0
Cleanliness                        0
Departure Delay in Minutes         0
Arrival Delay in Minutes           0
```

```
satisfaction                         0
dtype: int64
```

[13]:
```
df_test.isnull().sum()
```

[13]:
```
id                                   0
Gender                               0
Customer Type                        0
Age                                  0
Type of Travel                       0
Class                                0
Flight Distance                      0
Inflight wifi service                0
Departure/Arrival time convenient    0
Ease of Online booking               0
Gate location                        0
Food and drink                       0
Online boarding                      0
Seat comfort                         0
Inflight entertainment               0
On-board service                     0
Leg room service                     0
Baggage handling                     0
Checkin service                      0
Inflight service                     0
Cleanliness                          0
Departure Delay in Minutes           0
Arrival Delay in Minutes             0
satisfaction                         0
dtype: int64
```

Data Yang Sudah Dibersihkan

[14]:
```
df.head()
```

[14]:

	id	Gender	Customer Type	Age	Type of Travel	Class	\
0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	
1	5047	Male	disloyal Customer	25	Business travel	Business	
2	110028	Female	Loyal Customer	26	Business travel	Business	
3	24026	Female	Loyal Customer	25	Business travel	Business	
4	119299	Male	Loyal Customer	61	Business travel	Business	

	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	\
0	460	3	4	
1	235	3	2	
2	1142	2	2	
3	562	2	5	
4	214	3	3	

```
   Ease of Online booking  …  Inflight entertainment  On-board service  \
0                       3  …                       5                 4
1                       3  …                       1                 1
2                       2  …                       5                 4
3                       5  …                       2                 2
4                       3  …                       3                 3

   Leg room service  Baggage handling  Checkin service  Inflight service  \
0                 3                 4                4                 5
1                 5                 3                1                 4
2                 3                 4                4                 4
3                 5                 3                1                 4
4                 4                 4                3                 3

   Cleanliness  Departure Delay in Minutes  Arrival Delay in Minutes  \
0            5                          25                      18.0
1            1                           1                       6.0
2            5                           0                       0.0
3            2                          11                       9.0
4            3                           0                       0.0

              satisfaction
0  neutral or dissatisfied
1  neutral or dissatisfied
2                satisfied
3  neutral or dissatisfied
4                satisfied

[5 rows x 24 columns]
```

[15]:
```
df_test.head()
```

[15]:
```
      id  Gender      Customer Type  Age   Type of Travel     Class  \
0  19556  Female     Loyal Customer   52  Business travel       Eco
1  90035  Female     Loyal Customer   36  Business travel  Business
2  12360    Male  disloyal Customer   20  Business travel       Eco
3  77959    Male     Loyal Customer   44  Business travel  Business
4  36875  Female     Loyal Customer   49  Business travel       Eco

   Flight Distance  Inflight wifi service  Departure/Arrival time convenient  \
0              160                      5                                  4
1             2863                      1                                  1
2              192                      2                                  0
3             3377                      0                                  0
4             1182                      2                                  3
```

```
   Ease of Online booking  …  Inflight entertainment  On-board service  \
0                       3  …                       5                 5
1                       3  …                       4                 4
2                       2  …                       2                 4
3                       0  …                       1                 1
4                       4  …                       2                 2

   Leg room service  Baggage handling  Checkin service  Inflight service  \
0                 5                 5                2                 5
1                 4                 4                3                 4
2                 1                 3                2                 2
3                 1                 1                3                 1
4                 2                 2                4                 2

   Cleanliness  Departure Delay in Minutes  Arrival Delay in Minutes  \
0            5                          50                      44.0
1            5                           0                       0.0
2            2                           0                       0.0
3            4                           0                       6.0
4            4                           0                      20.0

              satisfaction
0                satisfied
1                satisfied
2  neutral or dissatisfied
3                satisfied
4                satisfied

[5 rows x 24 columns]
```

[112]:
```
df.describe()
```

[112]:
```
                 id            Age  Flight Distance  Inflight wifi service  \
count  103594.000000  103594.000000    103594.000000          103594.000000
mean    64942.428625      39.380466      1189.325202               2.729753
std     37460.816597      15.113125       997.297235               1.327866
min         1.000000       7.000000        31.000000               0.000000
25%     32562.250000      27.000000       414.000000               2.000000
50%     64890.000000      40.000000       842.000000               3.000000
75%     97370.500000      51.000000      1743.000000               4.000000
max    129880.000000      85.000000      4983.000000               5.000000

       Departure/Arrival time convenient  Ease of Online booking  \
count                      103594.000000           103594.000000
mean                            3.060081                2.756984
std                             1.525233                1.398934
min                             0.000000                0.000000
```

```
25%                              2.000000                2.000000
50%                              3.000000                3.000000
75%                              4.000000                4.000000
max                              5.000000                5.000000

       Gate location  Food and drink  Online boarding   Seat comfort  \
count  103594.000000   103594.000000    103594.000000  103594.000000
mean        2.977026        3.202126         3.250497       3.439765
std         1.277723        1.329401         1.349433       1.318896
min         0.000000        0.000000         0.000000       0.000000
25%         2.000000        2.000000         2.000000       2.000000
50%         3.000000        3.000000         3.000000       4.000000
75%         4.000000        4.000000         4.000000       5.000000
max         5.000000        5.000000         5.000000       5.000000

       Inflight entertainment  On-board service  Leg room service  \
count           103594.000000     103594.000000     103594.000000
mean                 3.358341          3.382609          3.351401
std                  1.333030          1.288284          1.315409
min                  0.000000          0.000000          0.000000
25%                  2.000000          2.000000          2.000000
50%                  4.000000          4.000000          4.000000
75%                  4.000000          4.000000          4.000000
max                  5.000000          5.000000          5.000000

       Baggage handling  Checkin service  Inflight service    Cleanliness  \
count     103594.000000    103594.000000     103594.000000  103594.000000
mean           3.631687         3.304323          3.640761       3.286397
std            1.181051         1.265396          1.175603       1.312194
min            1.000000         0.000000          0.000000       0.000000
25%            3.000000         3.000000          3.000000       2.000000
50%            4.000000         3.000000          4.000000       3.000000
75%            5.000000         4.000000          5.000000       4.000000
max            5.000000         5.000000          5.000000       5.000000

       Departure Delay in Minutes  Arrival Delay in Minutes
count                103594.000000             103594.000000
mean                     14.747939                 15.178678
std                      38.116737                 38.698682
min                       0.000000                  0.000000
25%                       0.000000                  0.000000
50%                       0.000000                  0.000000
75%                      12.000000                 13.000000
max                    1592.000000               1584.000000
```

[113]:
```
df_test.describe()
```

[113]:

	id	Age	Flight Distance	Inflight wifi service
count	25893.000000	25893.000000	25893.000000	25893.000000
mean	65021.974858	39.621983	1193.753254	2.723709
std	37606.098635	15.134224	998.626779	1.334711
min	17.000000	7.000000	31.000000	0.000000
25%	32209.000000	27.000000	414.000000	2.000000
50%	65344.000000	40.000000	849.000000	3.000000
75%	97623.000000	51.000000	1744.000000	4.000000
max	129877.000000	85.000000	4983.000000	5.000000

	Departure/Arrival time convenient	Ease of Online booking
count	25893.000000	25893.000000
mean	3.046422	2.755996
std	1.532971	1.412552
min	0.000000	0.000000
25%	2.000000	2.000000
50%	3.000000	3.000000
75%	4.000000	4.000000
max	5.000000	5.000000

	Gate location	Food and drink	Online boarding	Seat comfort
count	25893.000000	25893.000000	25893.000000	25893.000000
mean	2.976442	3.214923	3.261615	3.448886
std	1.281661	1.331895	1.355505	1.320254
min	1.000000	0.000000	0.000000	1.000000
25%	2.000000	2.000000	2.000000	2.000000
50%	3.000000	3.000000	4.000000	4.000000
75%	4.000000	4.000000	4.000000	5.000000
max	5.000000	5.000000	5.000000	5.000000

	Inflight entertainment	On-board service	Leg room service
count	25893.000000	25893.000000	25893.000000
mean	3.356969	3.385587	3.349786
std	1.338643	1.282033	1.319045
min	0.000000	0.000000	0.000000
25%	2.000000	2.000000	2.000000
50%	4.000000	4.000000	4.000000
75%	4.000000	4.000000	4.000000
max	5.000000	5.000000	5.000000

	Baggage handling	Checkin service	Inflight service	Cleanliness
count	25893.000000	25893.000000	25893.000000	25893.000000
mean	3.632681	3.313907	3.648824	3.285521
std	1.176220	1.269138	1.180650	1.319355
min	1.000000	1.000000	0.000000	0.000000
25%	3.000000	3.000000	3.000000	2.000000
50%	4.000000	3.000000	4.000000	3.000000

```
75%            5.000000          4.000000          5.000000       4.000000
max            5.000000          5.000000          5.000000       5.000000

       Departure Delay in Minutes  Arrival Delay in Minutes
count                25893.000000              25893.000000
mean                    14.225080                 14.740857
std                     37.185919                 37.517539
min                      0.000000                  0.000000
25%                      0.000000                  0.000000
50%                      0.000000                  0.000000
75%                     12.000000                 13.000000
max                   1128.000000               1115.000000
```

Menggabungkan dataframe menggunakan method `concat()` dan parameter `ignore_index` untuk mengabaikan indeks pada dataframe.

[16]:
```
df_merge = pd.concat([df, df_test], ignore_index=True)
df_merge.shape
```

[16]:
```
(129487, 24)
```

Menyimpan dataframe yang sudah digabung ke dalam file `.csv`

[17]:
```
output_file_path = './Dataset/Airline-Passenger-Satisfaction/merge_cleaned_data.
↪csv'
df_merge.to_csv(output_file_path, index=False)
```

2.2 EDA (Exploratory Data Analysis)

2.2.1 Pertanyaan Analisis

1. Layanan apa yang memberikan kepuasan penumpang tertinggi?
2. Berapa rata-rata penilaian pelanggan untuk semua layanan?
3. Bagaimana perbandingan tingkat kepuasan penumpang antar kelas?
4. Bagaimana perbandingan hasil antara algoritma klasifikasi Naive Bayes, k-NN, dan Decision Tree?

2.2.2 Tujuan

1. Mengidentifikasi layanan spesifik yang paling berkontribusi terhadap kepuasan penumpang dengan menganalisis korelasi antara berbagai aspek layanan dan tingkat kepuasan.
2. Menghitung nilai rata-rata dari seluruh layanan yang tersedia dalam dataset untuk memperoleh gambaran umum tentang kepuasan penumpang.
3. Melakukan perbandingan secara langsung antara kepuasan penumpang dari kelas ekonomi, ekonomi plus dan business untuk mengetahui apakah terdapat perbedaan signifikan dalam pengalaman penumpang di kategori kelas tersebut.
4. Menggunakan 3 teknik klasifikasi untuk membedakan penumpang berdasarkan tipe penumpang dan tingkat kepuasan yang mereka berikan terhadap layanan, serta membandingkan performa dari algoritma tersebut.

2.2.3 Eksplorasi Data

Menemukan pola hubungan antar kolom dan mengubah kolom kategorikal menjadi numerik untuk digunakan pada proses selanjutnya.

Melihat data yang sudah dibersihkan

[116]:
```
df_merge
```

[116]:

	id	Gender	Customer Type	Age	Type of Travel	Class	\
0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	
1	5047	Male	disloyal Customer	25	Business travel	Business	
2	110028	Female	Loyal Customer	26	Business travel	Business	
3	24026	Female	Loyal Customer	25	Business travel	Business	
4	119299	Male	Loyal Customer	61	Business travel	Business	
…	…	…	…	…	…	…	
129482	78463	Male	disloyal Customer	34	Business travel	Business	
129483	71167	Male	Loyal Customer	23	Business travel	Business	
129484	37675	Female	Loyal Customer	17	Personal Travel	Eco	
129485	90086	Male	Loyal Customer	14	Business travel	Business	
129486	34799	Female	Loyal Customer	42	Personal Travel	Eco	

	Flight Distance	Inflight wifi service	\
0	460	3	
1	235	3	
2	1142	2	
3	562	2	
4	214	3	
…	…	…	
129482	526	3	
129483	646	4	
129484	828	2	
129485	1127	3	
129486	264	2	

	Departure/Arrival time convenient	Ease of Online booking	…	\
0	4	3	…	
1	2	3	…	
2	2	2	…	
3	5	5	…	
4	3	3	…	
…	…	…	…	
129482	3	3	…	
129483	4	4	…	
129484	5	1	…	
129485	3	3	…	
129486	5	2	…	

```
        Inflight entertainment  On-board service  Leg room service  \
0                            5                 4                 3
1                            1                 1                 5
2                            5                 4                 3
3                            2                 2                 5
4                            3                 3                 4
…                          …                 …                 …
129482                       4                 3                 2
129483                       4                 4                 5
129484                       2                 4                 3
129485                       4                 3                 2
129486                       1                 1                 2

        Baggage handling  Checkin service  Inflight service  Cleanliness  \
0                      4                4                 5            5
1                      3                1                 4            1
2                      4                4                 4            5
3                      3                1                 4            2
4                      4                3                 3            3
…                    …                …                 …          …
129482                 4                4                 5            4
129483                 5                5                 5            4
129484                 4                5                 4            2
129485                 5                4                 5            4
129486                 1                1                 1            1

        Departure Delay in Minutes  Arrival Delay in Minutes  \
0                               25                      18.0
1                                1                       6.0
2                                0                       0.0
3                               11                       9.0
4                                0                       0.0
…                              …                       …
129482                           0                       0.0
129483                           0                       0.0
129484                           0                       0.0
129485                           0                       0.0
129486                           0                       0.0

                satisfaction
0       neutral or dissatisfied
1       neutral or dissatisfied
2                     satisfied
3       neutral or dissatisfied
4                     satisfied
…                           …
129482  neutral or dissatisfied
```

```
129483               satisfied
129484  neutral or dissatisfied
129485               satisfied
129486  neutral or dissatisfied

[129487 rows x 24 columns]
```

Menampilkan beberapa baris pertama dari dataset

[117]:
```
df_merge.head()
```

[117]:

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	…	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3	…	5	4	3	4	4	5	5	25	18.0	neutral or dissatisfied
1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	3	…	1	1	5	3	1	4	1	1	6.0	
2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2	…	5	4	3	4	4	4	5	0	0.0	
3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	5	…	2	2	5	3	1	4	2	11	9.0	
4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	3	…	3	3	4	4	3	3	3	0	0.0	

```
1  neutral or dissatisfied
2                satisfied
3  neutral or dissatisfied
4                satisfied

[5 rows x 24 columns]
```

Statistik deskriptif untuk data numerik

[118]:
```
df_merge.describe()
```

[118]:
```
                 id            Age  Flight Distance  Inflight wifi service  \
count  129487.000000  129487.000000    129487.000000          129487.000000
mean    64958.335169      39.428761      1190.210662               2.728544
std     37489.781165      15.117597       997.560954               1.329235
min         1.000000       7.000000        31.000000               0.000000
25%     32494.500000      27.000000       414.000000               2.000000
50%     64972.000000      40.000000       844.000000               3.000000
75%     97415.500000      51.000000      1744.000000               4.000000
max    129880.000000      85.000000      4983.000000               5.000000

       Departure/Arrival time convenient  Ease of Online booking  \
count                      129487.000000           129487.000000
mean                            3.057349                2.756786
std                             1.526787                1.401662
min                             0.000000                0.000000
25%                             2.000000                2.000000
50%                             3.000000                3.000000
75%                             4.000000                4.000000
max                             5.000000                5.000000

       Gate location  Food and drink  Online boarding   Seat comfort  \
count  129487.000000   129487.000000    129487.000000  129487.000000
mean        2.976909        3.204685         3.252720       3.441589
std         1.278506        1.329905         1.350651       1.319168
min         0.000000        0.000000         0.000000       0.000000
25%         2.000000        2.000000         2.000000       2.000000
50%         3.000000        3.000000         3.000000       4.000000
75%         4.000000        4.000000         4.000000       5.000000
max         5.000000        5.000000         5.000000       5.000000

       Inflight entertainment  On-board service  Leg room service  \
count           129487.000000     129487.000000     129487.000000
mean                3.358067          3.383204          3.351078
std                 1.334149          1.287032          1.316132
min                 0.000000          0.000000          0.000000
25%                 2.000000          2.000000          2.000000
50%                 4.000000          4.000000          4.000000
```

```
75%                 4.000000          4.000000          4.000000
max                 5.000000          5.000000          5.000000

       Baggage handling  Checkin service  Inflight service    Cleanliness  \
count     129487.000000    129487.000000     129487.000000  129487.000000
mean           3.631886         3.306239          3.642373       3.286222
std            1.180082         1.266146          1.176614       1.313624
min            1.000000         0.000000          0.000000       0.000000
25%            3.000000         3.000000          3.000000       2.000000
50%            4.000000         3.000000          4.000000       3.000000
75%            5.000000         4.000000          5.000000       4.000000
max            5.000000         5.000000          5.000000       5.000000

       Departure Delay in Minutes  Arrival Delay in Minutes
count                129487.000000             129487.000000
mean                     14.643385                 15.091129
std                      37.932867                 38.465650
min                       0.000000                  0.000000
25%                       0.000000                  0.000000
50%                       0.000000                  0.000000
75%                      12.000000                 13.000000
max                    1592.000000               1584.000000
```

1. Layanan yang memberikan kepuasan paling tinggi

[119]:
```
df_service = df_merge.iloc[:, 7:21]
df_service.mean().round(2)
```

[119]:
```
Inflight wifi service               2.73
Departure/Arrival time convenient   3.06
Ease of Online booking              2.76
Gate location                       2.98
Food and drink                      3.20
Online boarding                     3.25
Seat comfort                        3.44
Inflight entertainment              3.36
On-board service                    3.38
Leg room service                    3.35
Baggage handling                    3.63
Checkin service                     3.31
Inflight service                    3.64
Cleanliness                         3.29
dtype: float64
```

[120]:
```
# print service and the highest mean
print('Service with the highest mean:')
print(df_service.mean().round(2).idxmax())
print('Mean: ', df_service.mean().round(2).max())
```

```
# print service and the lowest mean
print('\nService with the lowest mean:')
print(df_service.mean().round(2).idxmin())
print('Mean: ', df_service.mean().round(2).min())
```

```
Service with the highest mean:
Inflight service
Mean:  3.64

Service with the lowest mean:
Inflight wifi service
Mean:  2.73
```

Dari hasil perhitungan didapatkan pelayanan dengan rata-rata paling tinggi yaitu `Inflight Service` dengan nilai 3.64 dan pelayanan dengan rata-rata paling rendah yaitu `Inflight wifi service` dengan nilai `2.73`.

2. Rata-rata penilaian seluruh pelanggan untuk semua layanan

[121]:
```
df_AllAverage = df_merge.iloc[:, 7:21]
df_AllAverage['Average'] = df_AllAverage.mean(axis=1)
df_AllAverage
```

[121]:

	Inflight wifi service	Departure/Arrival time convenient	\
0	3	4	
1	3	2	
2	2	2	
3	2	5	
4	3	3	
…	…	…	
129482	3	3	
129483	4	4	
129484	2	5	
129485	3	3	
129486	2	5	

	Ease of Online booking	Gate location	Food and drink	\
0	3	1	5	
1	3	3	1	
2	2	2	5	
3	5	5	2	
4	3	3	4	
…	…	…	…	
129482	3	1	4	
129483	4	4	4	
129484	1	5	2	
129485	3	3	4	
129486	2	5	4	

```
        Online boarding  Seat comfort  Inflight entertainment  \
0                     3             5                       5
1                     3             1                       1
2                     5             5                       5
3                     2             2                       2
4                     5             5                       3
…                   …             …                       …
129482                3             4                       4
129483                4             4                       4
129484                1             2                       2
129485                4             4                       4
129486                2             2                       1

        On-board service  Leg room service  Baggage handling  Checkin service  \
0                      4                 3                 4                4
1                      1                 5                 3                1
2                      4                 3                 4                4
3                      2                 5                 3                1
4                      3                 4                 4                3
…                    …                 …                 …              …
129482                 3                 2                 4                4
129483                 4                 5                 5                5
129484                 4                 3                 4                5
129485                 3                 2                 5                4
129486                 1                 2                 1                1

        Inflight service  Cleanliness   Average
0                      5            5  3.857143
1                      4            1  2.285714
2                      4            5  3.714286
3                      4            2  3.000000
4                      3            3  3.500000
…                    …            …       …
129482                 5            4  3.357143
129483                 5            4  4.285714
129484                 4            2  3.000000
129485                 5            4  3.642857
129486                 1            1  2.142857

[129487 rows x 15 columns]
```

[122]:
```
print("Rata-rata keseluruhan:", df_AllAverage['Average'].mean())
```

```
Rata-rata keseluruhan: 3.2412608436147474
```

Nilai rata-rata yang didapatkan untuk keseluruhan layanan sebesar 3.24. Dapat disimpulkan setiap pelanggan merasa cukup puas dengan pelayaan yang diberikan oleh maskapai penerbangan.

3. Perbandingan tingkat kepuasan penumpang antar kelas

[123]:
```
df_avg_class_satisfaction =df_merge.iloc[:,[5, 7, 8, 9, 10, 11, 12, 13, 14, 15,
↪16, 17, 18, 19, 20]].groupby('Class').mean()
df_avg_class_satisfaction
```

[123]:

Class	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness
Business	2.775657	2.907582	2.915373	2.984981	3.329795	3.719035	3.763704	3.639313	3.682529	3.646169	3.844539	3.520745	3.846007	3.481933
Eco	2.673882	3.192560	2.602801	2.969699	3.086429	2.814478	3.142041	3.096426	3.120171	3.083848	3.450264	3.124507	3.467144	3.104617
Eco Plus	2.755864	3.209382	2.662793	2.968230	3.110554	2.886247	3.168763	3.120469	3.034755	3.056610	3.351812	3.014606	3.382303	3.118017

[124]:
```
df_avg_class_satisfaction.mean(axis=1)
```

[124]:
```
Class
Business    3.432669
Eco         3.066348
Eco Plus    3.060029
dtype: float64
```

Dari data di atas, kita dapat melihat rata-rata penilaian kelas business lebih besar dibandingkan dengan kedua kelas lainnya, yaitu dengan nilai rata-rata `3.432669`. Hal ini menandakan bahwa pelanggan yang memilih kelas business merasa puas dengan pelayanan yang diberikan. Sedangkan kelas ekonomi dan ekonomi plus mendapatkan rata-rata yang hampir sama, yaitu `3.066348` dan

3.060029. Hal ini berarti pelanggan kelas tersebut merasa cukup puas dengan pelayanan yang diberikan.

Mengubah kolom kategorikal menjadi numerik menggunakan LabelEncoder dari Scikit-learn

[19]:
```
label_encoder = LabelEncoder()
df_merge_encoded = df_merge.copy()
categorical_columns = df_merge_encoded.select_dtypes(include=['object']).
 ↪columns.tolist()

for col in categorical_columns:
    df_merge_encoded[col] = label_encoder.fit_transform(df_merge_encoded[col])
```

Menampilkan hasil transformasi kolom kategorikal menjadi numerik

[126]:
```
df_merge_encoded.head()
```

[126]:

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	\
0	70172	1	0	13	1	2	460	
1	5047	1	1	25	0	0	235	
2	110028	0	0	26	0	0	1142	
3	24026	0	0	25	0	0	562	
4	119299	1	0	61	0	0	214	

	Inflight wifi service	Departure/Arrival time convenient	\
0	3	4	
1	3	2	
2	2	2	
3	2	5	
4	3	3	

	Ease of Online booking	…	Inflight entertainment	On-board service	\
0	3	…	5	4	
1	3	…	1	1	
2	2	…	5	4	
3	5	…	2	2	
4	3	…	3	3	

	Leg room service	Baggage handling	Checkin service	Inflight service	\
0	3	4	4	5	
1	5	3	1	4	
2	3	4	4	4	
3	5	3	1	4	
4	4	4	3	3	

	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	\
0	5	25	18.0	
1	1	1	6.0	

```
2            5                          0                       0.0
3            2                         11                       9.0
4            3                          0                       0.0

   satisfaction
0             0
1             0
2             1
3             0
4             1

[5 rows x 24 columns]
```

Mencari korelasi antar kolom numerik

[127]:
```
correlation_matrix = df_merge_encoded.corr()
print("Matriks Korelasi:")
correlation_matrix
```

```
Matriks Korelasi:
```

[127]:

	id	Gender	Customer Type	\
id	1.000000	-0.001027	0.001359	
Gender	-0.001027	1.000000	-0.030803	
Customer Type	0.001359	-0.030803	1.000000	
Age	0.020443	0.008984	-0.284275	
Type of Travel	-0.000734	0.009215	-0.308210	
Class	-0.104469	-0.011655	0.042959	
Flight Distance	0.095027	0.003836	-0.226134	
Inflight wifi service	-0.023242	0.005968	-0.005884	
Departure/Arrival time convenient	-0.002056	0.008772	-0.206916	
Ease of Online booking	0.013247	0.006129	-0.018183	
Gate location	-0.000019	-0.000860	0.004647	
Food and drink	-0.000183	0.001631	-0.057126	
Online boarding	0.055428	-0.044850	-0.189329	
Seat comfort	0.052352	-0.030847	-0.156383	
Inflight entertainment	0.001944	0.003798	-0.106157	
On-board service	0.055454	0.006441	-0.054040	
Leg room service	0.043914	0.031031	-0.046885	
Baggage handling	0.074618	0.036414	0.025015	
Checkin service	0.079154	0.008392	-0.031258	
Inflight service	0.078838	0.038176	0.023567	
Cleanliness	0.024425	0.002818	-0.081433	
Departure Delay in Minutes	-0.017506	0.003111	0.004131	
Arrival Delay in Minutes	-0.035657	0.001309	0.004730	
satisfaction	0.012990	0.011496	-0.185925	

	Age	Type of Travel	Class	\

id	0.020443	-0.000734	-0.104469
Gender	0.008984	0.009215	-0.011655
Customer Type	-0.284275	-0.308210	0.042959
Age	1.000000	-0.044910	-0.116967
Type of Travel	-0.044910	1.000000	0.486598
Class	-0.116967	0.486598	1.000000
Flight Distance	0.099863	-0.267064	-0.427144
Inflight wifi service	0.015779	-0.105574	-0.024912
Departure/Arrival time convenient	0.036780	0.257208	0.087185
Ease of Online booking	0.022294	-0.133891	-0.094989
Gate location	-0.000709	-0.029882	-0.005656
Food and drink	0.023283	-0.068728	-0.080732
Online boarding	0.207485	-0.223781	-0.297634
Seat comfort	0.159229	-0.127404	-0.212241
Inflight entertainment	0.074990	-0.152708	-0.183178
On-board service	0.056743	-0.059700	-0.210748
Leg room service	0.038992	-0.139540	-0.198828
Baggage handling	-0.048192	-0.032921	-0.166507
Checkin service	0.033182	0.016530	-0.157380
Inflight service	-0.051778	-0.023417	-0.159110
Cleanliness	0.052575	-0.084257	-0.129715
Departure Delay in Minutes	-0.009263	-0.006336	0.009553
Arrival Delay in Minutes	-0.011248	-0.005830	0.014162
satisfaction	0.134001	-0.449794	-0.448338

	Flight Distance	Inflight wifi service	\
id	0.095027	-0.023242	
Gender	0.003836	0.005968	
Customer Type	-0.226134	-0.005884	
Age	0.099863	0.015779	
Type of Travel	-0.267064	-0.105574	
Class	-0.427144	-0.024912	
Flight Distance	1.000000	0.006554	
Inflight wifi service	0.006554	1.000000	
Departure/Arrival time convenient	-0.018901	0.344846	
Ease of Online booking	0.064959	0.714888	
Gate location	0.005378	0.338547	
Food and drink	0.057136	0.132109	
Online boarding	0.215082	0.457422	
Seat comfort	0.157825	0.121373	
Inflight entertainment	0.130518	0.207887	
On-board service	0.111224	0.120028	
Leg room service	0.134548	0.160414	
Baggage handling	0.064810	0.120548	
Checkin service	0.073635	0.043847	
Inflight service	0.059182	0.110300	
Cleanliness	0.095658	0.131163	

Departure Delay in Minutes	0.001992	-0.016046
Arrival Delay in Minutes	-0.001935	-0.017749
satisfaction	0.298206	0.283291

	Departure/Arrival time convenient	\
id	-0.002056	
Gender	0.008772	
Customer Type	-0.206916	
Age	0.036780	
Type of Travel	0.257208	
Class	0.087185	
Flight Distance	-0.018901	
Inflight wifi service	0.344846	
Departure/Arrival time convenient	1.000000	
Ease of Online booking	0.437697	
Gate location	0.447411	
Food and drink	0.001057	
Online boarding	0.072175	
Seat comfort	0.008707	
Inflight entertainment	-0.008189	
On-board service	0.067046	
Leg room service	0.010634	
Baggage handling	0.070646	
Checkin service	0.091217	
Inflight service	0.072166	
Cleanliness	0.010021	
Departure Delay in Minutes	0.000610	
Arrival Delay in Minutes	-0.000942	
satisfaction	-0.054457	

	Ease of Online booking	…	\
id	0.013247	…	
Gender	0.006129	…	
Customer Type	-0.018183	…	
Age	0.022294	…	
Type of Travel	-0.133891	…	
Class	-0.094989	…	
Flight Distance	0.064959	…	
Inflight wifi service	0.714888	…	
Departure/Arrival time convenient	0.437697	…	
Ease of Online booking	1.000000	…	
Gate location	0.460155	…	
Food and drink	0.030638	…	
Online boarding	0.404944	…	
Seat comfort	0.028602	…	
Inflight entertainment	0.046669	…	
On-board service	0.039039	…	

Leg room service	0.109341	…
Baggage handling	0.039215	…
Checkin service	0.008835	…
Inflight service	0.035356	…
Cleanliness	0.015150	…
Departure Delay in Minutes	-0.005330	…
Arrival Delay in Minutes	-0.007033	…
satisfaction	0.168704	…

	Inflight entertainment	On-board service	\
id	0.001944	0.055454	
Gender	0.003798	0.006441	
Customer Type	-0.106157	-0.054040	
Age	0.074990	0.056743	
Type of Travel	-0.152708	-0.059700	
Class	-0.183178	-0.210748	
Flight Distance	0.130518	0.111224	
Inflight wifi service	0.207887	0.120028	
Departure/Arrival time convenient	-0.008189	0.067046	
Ease of Online booking	0.046669	0.039039	
Gate location	0.002751	-0.029109	
Food and drink	0.623366	0.057476	
Online boarding	0.284008	0.154272	
Seat comfort	0.611949	0.130654	
Inflight entertainment	1.000000	0.418863	
On-board service	0.418863	1.000000	
Leg room service	0.300573	0.357877	
Baggage handling	0.379291	0.520400	
Checkin service	0.119664	0.244620	
Inflight service	0.406561	0.551460	
Cleanliness	0.692491	0.122208	
Departure Delay in Minutes	-0.027166	-0.030471	
Arrival Delay in Minutes	-0.030230	-0.034789	
satisfaction	0.398334	0.322329	

	Leg room service	Baggage handling	\
id	0.043914	0.074618	
Gender	0.031031	0.036414	
Customer Type	-0.046885	0.025015	
Age	0.038992	-0.048192	
Type of Travel	-0.139540	-0.032921	
Class	-0.198828	-0.166507	
Flight Distance	0.134548	0.064810	
Inflight wifi service	0.160414	0.120548	
Departure/Arrival time convenient	0.010634	0.070646	
Ease of Online booking	0.109341	0.039215	
Gate location	-0.005146	0.001097	

Food and drink	0.033215	0.035413
Online boarding	0.123149	0.083563
Seat comfort	0.104244	0.074617
Inflight entertainment	0.300573	0.379291
On-board service	0.357877	0.520400
Leg room service	1.000000	0.371599
Baggage handling	0.371599	1.000000
Checkin service	0.152715	0.234732
Inflight service	0.369833	0.629492
Cleanliness	0.096777	0.097107
Departure Delay in Minutes	0.014339	-0.004425
Arrival Delay in Minutes	0.011346	-0.007935
satisfaction	0.312557	0.248651

	Checkin service	Inflight service	\
id	0.079154	0.078838	
Gender	0.008392	0.038176	
Customer Type	-0.031258	0.023567	
Age	0.033182	-0.051778	
Type of Travel	0.016530	-0.023417	
Class	-0.157380	-0.159110	
Flight Distance	0.073635	0.059182	
Inflight wifi service	0.043847	0.110300	
Departure/Arrival time convenient	0.091217	0.072166	
Ease of Online booking	0.008835	0.035356	
Gate location	-0.039294	0.000337	
Food and drink	0.085103	0.035424	
Online boarding	0.204215	0.073973	
Seat comfort	0.189838	0.068912	
Inflight entertainment	0.119664	0.406561	
On-board service	0.244620	0.551460	
Leg room service	0.152715	0.369833	
Baggage handling	0.234732	0.629492	
Checkin service	1.000000	0.237737	
Inflight service	0.237737	1.000000	
Cleanliness	0.176611	0.090565	
Departure Delay in Minutes	-0.018632	-0.054329	
Arrival Delay in Minutes	-0.021705	-0.059853	
satisfaction	0.237146	0.245027	

	Cleanliness	Departure Delay in Minutes	\
id	0.024425	-0.017506	
Gender	0.002818	0.003111	
Customer Type	-0.081433	0.004131	
Age	0.052575	-0.009263	
Type of Travel	-0.084257	-0.006336	
Class	-0.129715	0.009553	

Flight Distance	0.095658	0.001992
Inflight wifi service	0.131163	-0.016046
Departure/Arrival time convenient	0.010021	0.000610
Ease of Online booking	0.015150	-0.005330
Gate location	-0.006066	0.005943
Food and drink	0.658026	-0.029351
Online boarding	0.329331	-0.019319
Seat comfort	0.679657	-0.027711
Inflight entertainment	0.692491	-0.027166
On-board service	0.122208	-0.030471
Leg room service	0.096777	0.014339
Baggage handling	0.097107	-0.004425
Checkin service	0.176611	-0.018632
Inflight service	0.090565	-0.054329
Cleanliness	1.000000	-0.014553
Departure Delay in Minutes	-0.014553	1.000000
Arrival Delay in Minutes	-0.016546	0.965291
satisfaction	0.306891	-0.051032

	Arrival Delay in Minutes	satisfaction
id	-0.035657	0.012990
Gender	0.001309	0.011496
Customer Type	0.004730	-0.185925
Age	-0.011248	0.134001
Type of Travel	-0.005830	-0.449794
Class	0.014162	-0.448338
Flight Distance	-0.001935	0.298206
Inflight wifi service	-0.017749	0.283291
Departure/Arrival time convenient	-0.000942	-0.054457
Ease of Online booking	-0.007033	0.168704
Gate location	0.005658	-0.002923
Food and drink	-0.031715	0.211164
Online boarding	-0.022730	0.501620
Seat comfort	-0.030521	0.348576
Inflight entertainment	-0.030230	0.398334
On-board service	-0.034789	0.322329
Leg room service	0.011346	0.312557
Baggage handling	-0.007935	0.248651
Checkin service	-0.021705	0.237146
Inflight service	-0.059853	0.245027
Cleanliness	-0.016546	0.306891
Departure Delay in Minutes	0.965291	-0.051032
Arrival Delay in Minutes	1.000000	-0.058275
satisfaction	-0.058275	1.000000

[24 rows x 24 columns]

Menyimpan data yang telah diubah

[128]:
```
output_file_path = './Dataset/Airline-Passenger-Satisfaction/
↪hasil_eksplorasi_data.csv'
df_merge_encoded.to_csv(output_file_path, index=False)
```

Melihat data yang telah diubah

[129]:
```
df_merge_encoded
```

[129]:

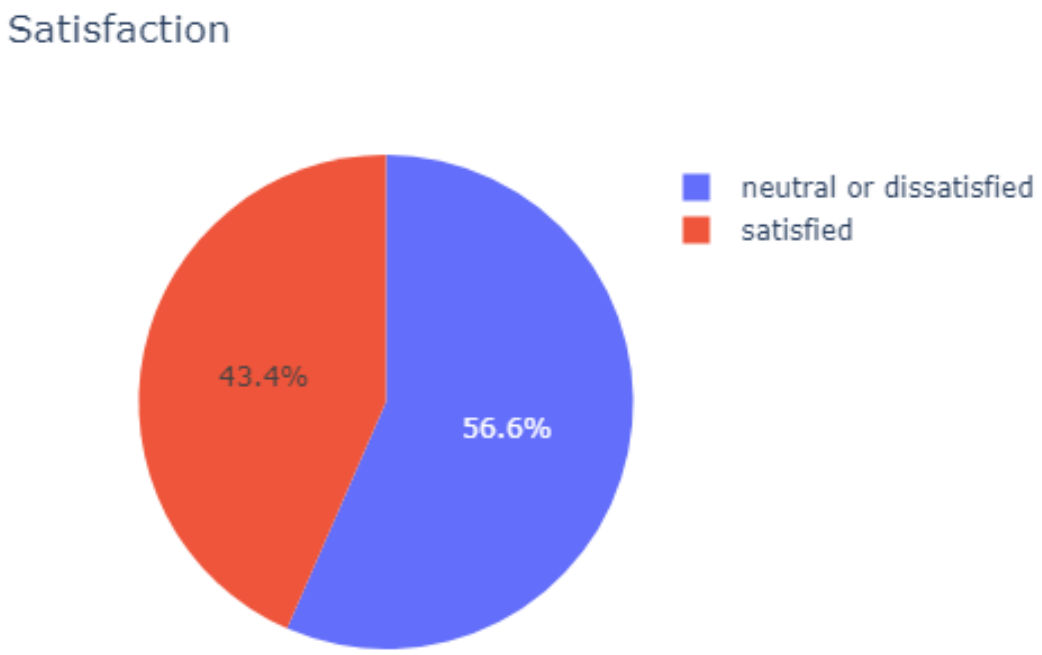
```
             id  Gender  Customer Type  Age  Type of Travel  Class  \
0         70172       1              0   13               1      2
1          5047       1              1   25               0      0
2        110028       0              0   26               0      0
3         24026       0              0   25               0      0
4        119299       1              0   61               0      0
…           …     …              …   …              …    …
129482    78463       1              1   34               0      0
129483    71167       1              0   23               0      0
129484    37675       0              0   17               1      1
129485    90086       1              0   14               0      0
129486    34799       0              0   42               1      1

         Flight Distance  Inflight wifi service  \
0                    460                      3
1                    235                      3
2                   1142                      2
3                    562                      2
4                    214                      3
…                   …                      …
129482               526                      3
129483               646                      4
129484               828                      2
129485              1127                      3
129486               264                      2

         Departure/Arrival time convenient  Ease of Online booking  …  \
0                                        4                       3  …
1                                        2                       3  …
2                                        2                       2  …
3                                        5                       5  …
4                                        3                       3  …
…                                      …                     …  …
129482                                   3                       3  …
129483                                   4                       4  …
129484                                   5                       1  …
129485                                   3                       3  …
129486                                   5                       2  …

         Inflight entertainment  On-board service  Leg room service  \
```

```
0                          5                 4                 3
1                          1                 1                 5
2                          5                 4                 3
3                          2                 2                 5
4                          3                 3                 4
…                        …                 …                 …
129482                     4                 3                 2
129483                     4                 4                 5
129484                     2                 4                 3
129485                     4                 3                 2
129486                     1                 1                 2

        Baggage handling  Checkin service  Inflight service  Cleanliness  \
0                      4                4                 5            5
1                      3                1                 4            1
2                      4                4                 4            5
3                      3                1                 4            2
4                      4                3                 3            3
…                    …                …                 …          …
129482                 4                4                 5            4
129483                 5                5                 5            4
129484                 4                5                 4            2
129485                 5                4                 5            4
129486                 1                1                 1            1

        Departure Delay in Minutes  Arrival Delay in Minutes  satisfaction
0                               25                      18.0             0
1                                1                       6.0             0
2                                0                       0.0             1
3                               11                       9.0             0
4                                0                       0.0             1
…                              …                       …             …
129482                           0                       0.0             0
129483                           0                       0.0             1
129484                           0                       0.0             0
129485                           0                       0.0             1
129486                           0                       0.0             0

[129487 rows x 24 columns]
```

[130]:
```
df_merge_encoded['satisfaction'].value_counts()
```

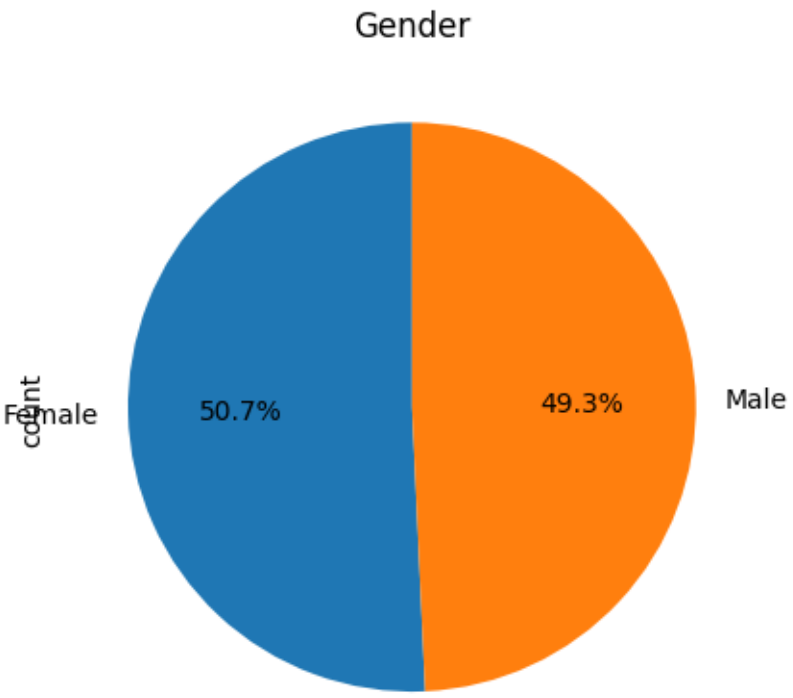
[130]:
```
satisfaction
0    73225
1    56262
Name: count, dtype: int64
```

[131]:
```
df_satisfaction = df_merge['satisfaction'].value_counts().reset_index()

plot = px.pie(df_satisfaction, values='count',␣
 ↪names=df_satisfaction['satisfaction'].unique(), title='Satisfaction',␣
 ↪height=400, width=500)
plot.show()
```



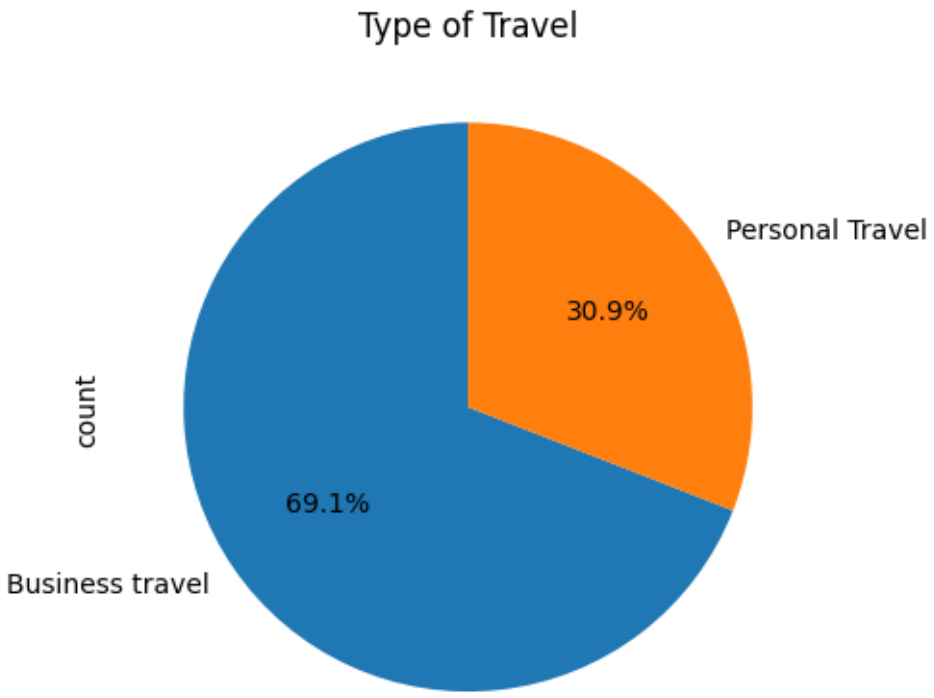
Berdasarkan hasil data tersebut, label kepuasan pelanggan terhadap layanan penerbangan, terbagi menjadi 2 kategori. Kategori `satisfied` memiliki persentase lebih sedikit dari kategori `neutral or dissatisfied`, tetapi persentase 56.6% tersebut terbagi menjadi dua kategori netral dan tidak puas yang hitungannya lebih kecil dari persentase puas secara penuh sehingga pelanggan cukup puas dan menikmati layanan penerbangan yang disediakan.

[132]:
```
df_merge['Gender'].value_counts().plot.pie(autopct='%1.1f%%', startangle=90)
plt.title('Gender')
plt.show()
```



Kumpulan data tersebut menunjukkan distribusi kepuasan penumpang yang secara umum seimbang antar gender, dengan jumlah penumpang perempuan sedikit lebih banyak dibandingkan laki-laki. Kepuasan penumpang cukup merata antar gender, dengan proporsi individu yang puas sedikit lebih tinggi baik pada kelompok pria maupun wanita. Temuan ini menyoroti keterwakilan penumpang laki-laki dan perempuan yang relatif setara, dengan jumlah penumpang perempuan yang sedikit lebih tinggi.

[133]:
```
df_merge['Type of Travel'].value_counts().plot.pie(autopct='%1.1f%%',␣
↪startangle=90)
plt.title('Type of Travel')
plt.show()
```



[134]:
```
df_travel = df_merge.groupby(['Type of Travel', 'satisfaction']).
 ↪agg({'satisfaction': 'count'}).rename(columns={'satisfaction': 'Jumlah'})
df_travel.reset_index(inplace=True)
df_travel
```

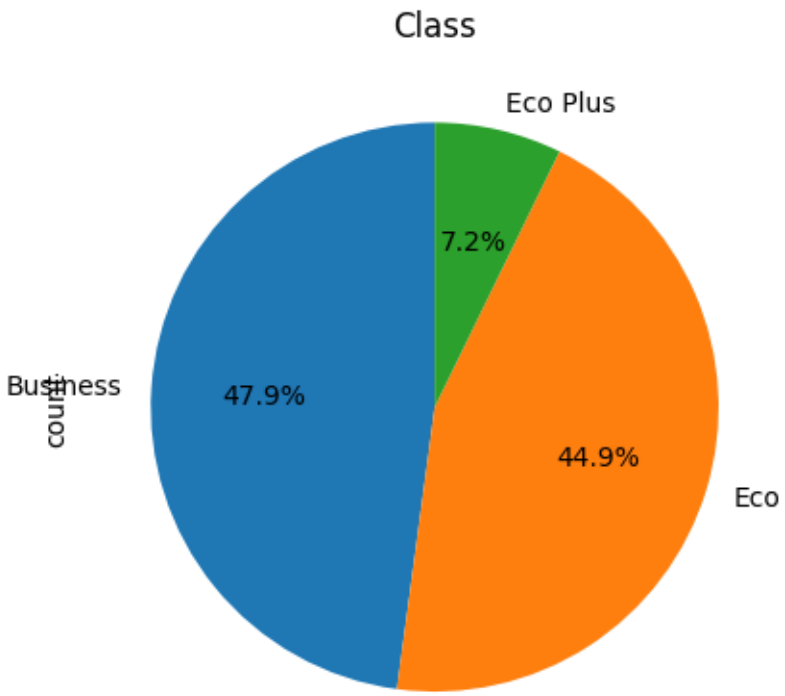
[134]:

	Type of Travel	satisfaction	Jumlah
0	Business travel	neutral or dissatisfied	37238
1	Business travel	satisfied	52207
2	Personal Travel	neutral or dissatisfied	35987
3	Personal Travel	satisfied	4055

[135]:
```
plot = px.bar(df_travel, x='Type of Travel', y='Jumlah', color='satisfaction',␣
 ↪barmode='group', title='Type of Travel vs Satisfaction', height=400,␣
 ↪width=700, text='Jumlah')
plot.show()
```

Jika dilihat dari tipe perjalanan yang dipilih oleh penumpang, persentase banyaknya penumpang yang paling besar yaitu tipe perjalanan bisnis. Jika dikaitkan dengan tingkat kepuasan, dapat diperoleh penumpang dengan perjalanan bisnis lebih dominan menyatakan label satisfication dan penumpang dengan perjalanan personal dominan menyatakan kepuasan label neutral or dissatisfaction. Dapat disimpulkan bahwa tipe perjalanan business dengan fasilitas yang lebih dapat mempengaruhi kepuasan pelanggan.

[136]:
```
df_merge['Class'].value_counts().plot.pie(autopct='%1.1f%%', startangle=90)
plt.title('Class')
plt.show()
```



[137]:
```
df_class = df_merge.groupby(['Class', 'satisfaction']).agg({'satisfaction':␣
↪'count'}).rename(columns={'satisfaction': 'Jumlah'})
df_class.reset_index(inplace=True)
df_class
```

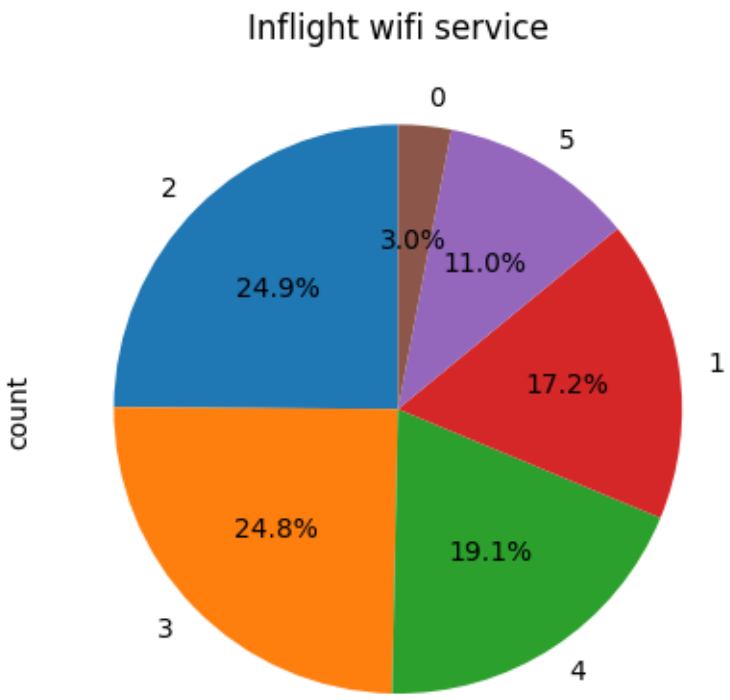
[137]:

	Class	satisfaction	Jumlah
0	Business	neutral or dissatisfied	18940
1	Business	satisfied	43050
2	Eco	neutral or dissatisfied	47215
3	Eco	satisfied	10902
4	Eco Plus	neutral or dissatisfied	7070
5	Eco Plus	satisfied	2310

[138]:
```
plot = px.bar(df_class, x='Class', y='Jumlah', color='satisfaction',␣
↪barmode='group', title='Class vs Satisfaction', height=400, width=700,␣
↪text='Jumlah')
plot.show()
```

Berdasarkan data di atas, dihasilkan bahwa jumlah penumpang pada class kategori business menggambarkan jumlah terbesar dengan label kepuasan satisfied, sedangkan jumlah penumpang pada urutan kedua masuk ke dalam class kategori eco dengan dominan label kepuasan neutral or dissatisfied, dan pada class kategori eco plus memiliki jumlah penumpang terkecil dengan tingkat kepuasan neutral or dissatisfied lebih besar. Dapat disimpulkan class penerbangan dapat mempengaruhi tingkat kepuasan.

[139]:
```
df_merge_encoded['Inflight wifi service'].value_counts().plot.pie(autopct='%1.
  ↪1f%%', startangle=90)
plt.title('Inflight wifi service')
plt.show()
```



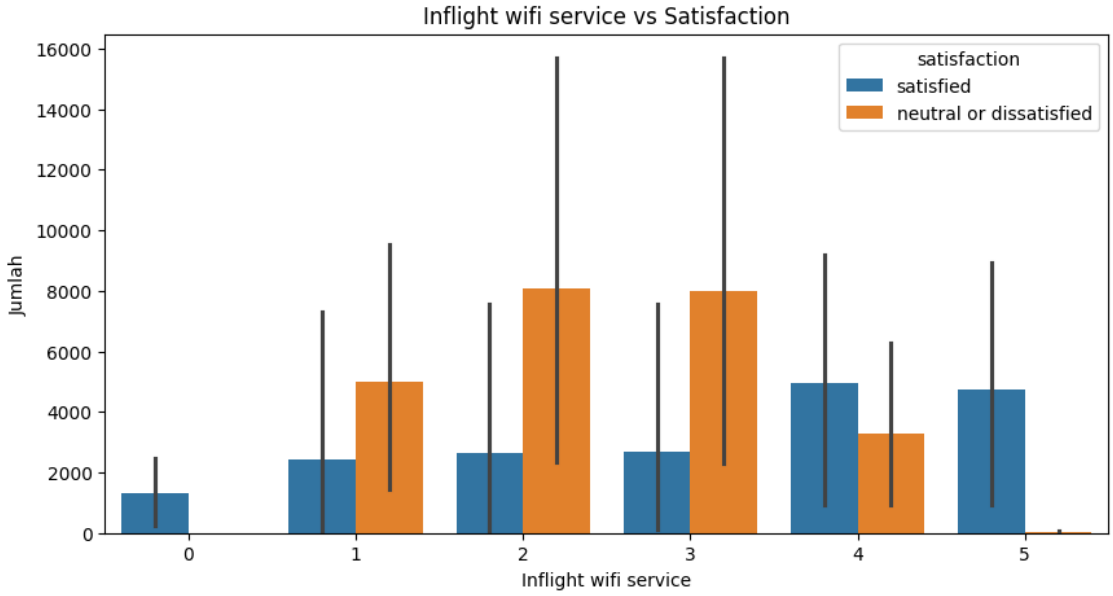
[140]:
```
df_inflight = df_merge.groupby(['Class', 'Inflight wifi service',␣
↪'satisfaction']).agg({'satisfaction': 'count'}).
↪rename(columns={'satisfaction': 'Jumlah'})
df_inflight.reset_index(inplace=True)
df_inflight
```

[140]:

	Class	Inflight wifi service	satisfaction	Jumlah
0	Business	0	satisfied	2487
1	Business	1	neutral or dissatisfied	3964
2	Business	1	satisfied	7299
3	Business	2	neutral or dissatisfied	6230
4	Business	2	satisfied	7569
5	Business	3	neutral or dissatisfied	6014
6	Business	3	satisfied	7565
7	Business	4	neutral or dissatisfied	2637
8	Business	4	satisfied	9208
9	Business	5	neutral or dissatisfied	95
10	Business	5	satisfied	8922
11	Eco	0	neutral or dissatisfied	4
12	Eco	0	satisfied	1160
13	Eco	1	neutral or dissatisfied	9534
14	Eco	1	satisfied	2

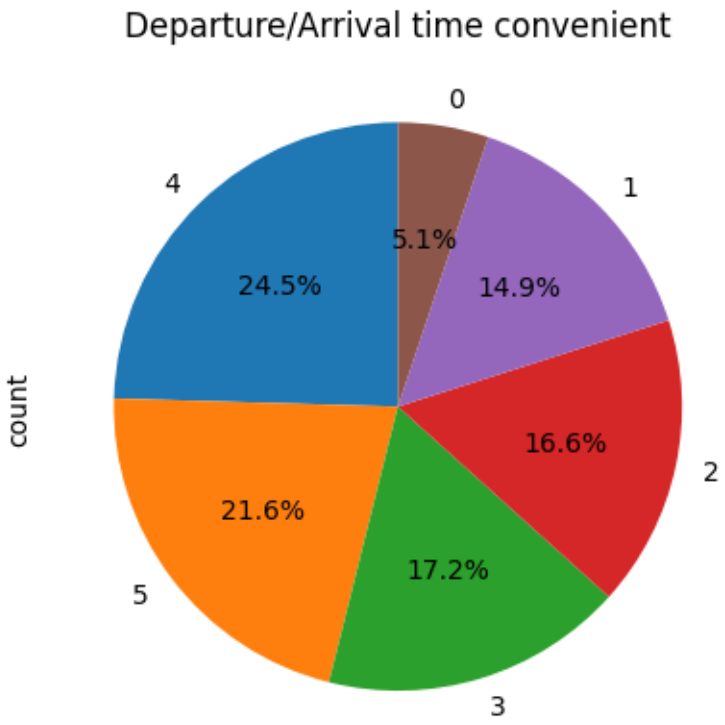
15	Eco	2	neutral or dissatisfied	15694
16	Eco	2	satisfied	331
17	Eco	3	neutral or dissatisfied	15678
18	Eco	3	satisfied	403
19	Eco	4	neutral or dissatisfied	6275
20	Eco	4	satisfied	4711
21	Eco	5	neutral or dissatisfied	30
22	Eco	5	satisfied	4295
23	Eco Plus	0	neutral or dissatisfied	6
24	Eco Plus	0	satisfied	251
25	Eco Plus	1	neutral or dissatisfied	1450
26	Eco Plus	1	satisfied	1
27	Eco Plus	2	neutral or dissatisfied	2337
28	Eco Plus	2	satisfied	75
29	Eco Plus	3	neutral or dissatisfied	2315
30	Eco Plus	3	satisfied	112
31	Eco Plus	4	neutral or dissatisfied	946
32	Eco Plus	4	satisfied	925
33	Eco Plus	5	neutral or dissatisfied	16
34	Eco Plus	5	satisfied	946

[141]:
```
plt.figure(figsize=(10, 5))
sns.barplot(x='Inflight wifi service', y='Jumlah', hue='satisfaction',␣
 ↪data=df_inflight)
plt.title('Inflight wifi service vs Satisfaction')
plt.show()
```



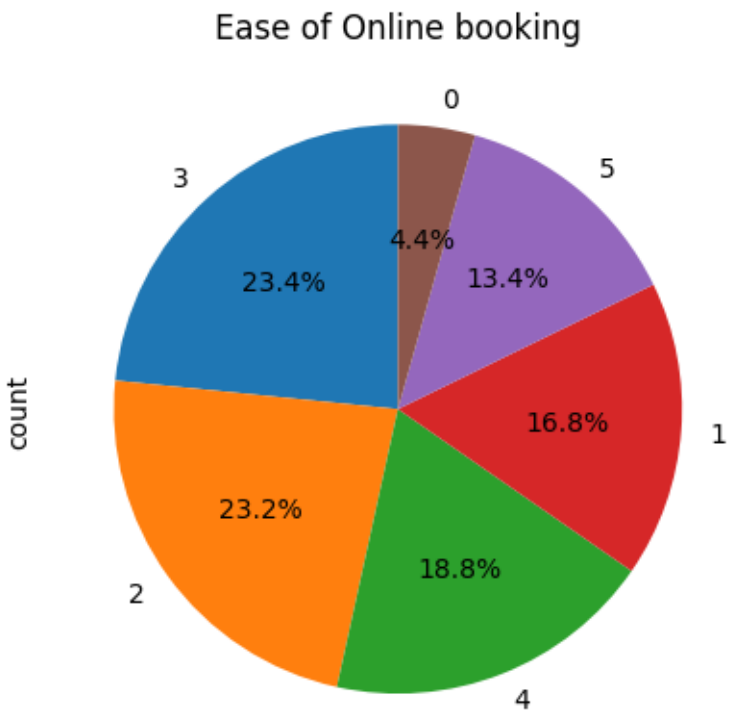
Berdasarkan data di atas, layanan inflight wifi mendapat penilaian kepuasan pelanggan dengan skala 2 dari 5 paling banyak dari keseluruhan data penumpang sebesar 24.9%. Secara rincian penyumbang penilaian `satisfied` didominasi oleh pelanggan class business menunjukkan bahwa class penumpang mempengaruhi kualitas penyediaan layanan.

[142]:
```
df_merge_encoded['Departure/Arrival time convenient'].value_counts().plot.
  ↪pie(autopct='%1.1f%%', startangle=90)
plt.title('Departure/Arrival time convenient')
plt.show()
```



Berdasarkan data di atas, layanan departure/arrival time convenient mendapat penilaian kepuasan pelanggan dengan skala 4 dari 5 paling banyak dari keseluruhan data penumpang sebesar 24.5%. Dapat disimpulkan bahwa layanan ini cukup memberikan pelayanan yang baik.

[143]:
```
df_merge_encoded['Ease of Online booking'].value_counts().plot.pie(autopct='%1.
  ↪1f%%', startangle=90)
plt.title('Ease of Online booking')
plt.show()
```



Berdasarkan data di atas, layanan ease of online booking mendapat penilaian kepuasan pelanggan dengan skala 3 dari 5 paling banyak dari keseluruhan data penumpang sebesar 23.4%. Dapat disimpulkan bahwa layanan ini cukup memberikan pelayanan yang baik dengan memberikan kemudahan dalam pemesanan tiket via online.

[144]:
```
df_merge_encoded['Food and drink'].value_counts().plot.pie(autopct='%1.1f%%',␣
↪startangle=90)
plt.title('Food and drink')
plt.show()
```

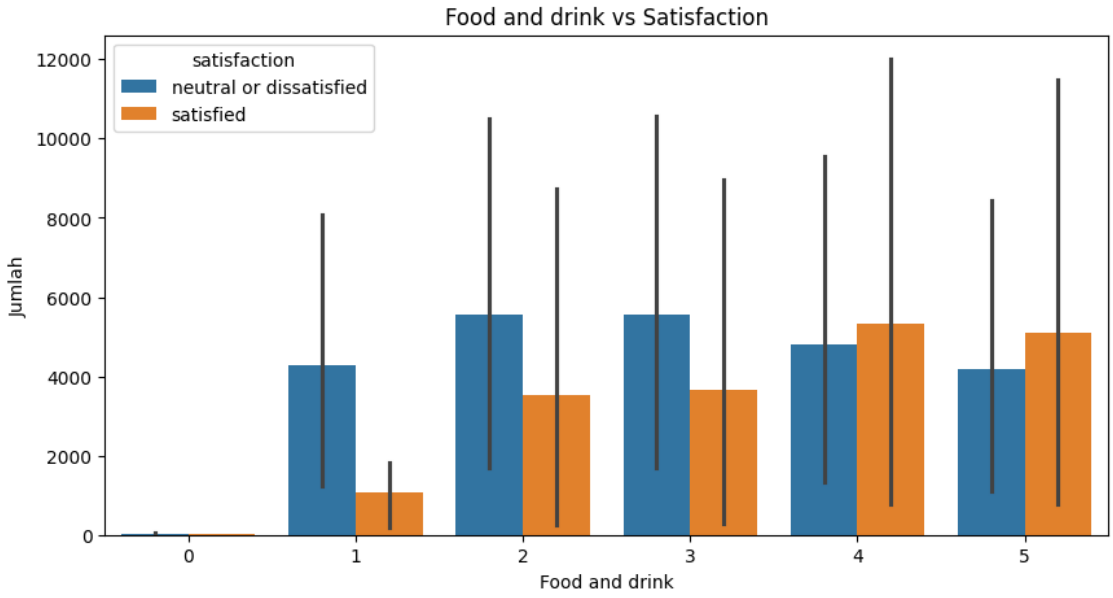
[145]:
```
df_food = df_merge.groupby(['Class', 'Food and drink', 'satisfaction']).
 ↪agg({'satisfaction': 'count'}).rename(columns={'satisfaction': 'Jumlah'})
df_food.reset_index(inplace=True)
df_food
```

[145]:

	Class	Food and drink	satisfaction	Jumlah
0	Business	0	neutral or dissatisfied	13
1	Business	0	satisfied	24
2	Business	1	neutral or dissatisfied	3521
3	Business	1	satisfied	1839
4	Business	2	neutral or dissatisfied	4450
5	Business	2	satisfied	8736
6	Business	3	neutral or dissatisfied	4437
7	Business	3	satisfied	8970
8	Business	4	neutral or dissatisfied	3531
9	Business	4	satisfied	12008
10	Business	5	neutral or dissatisfied	2988
11	Business	5	satisfied	11473
12	Eco	0	neutral or dissatisfied	52
13	Eco	0	satisfied	20
14	Eco	1	neutral or dissatisfied	8067
15	Eco	1	satisfied	1163

16	Eco	2	neutral or dissatisfied	10513
17	Eco	2	satisfied	1638
18	Eco	3	neutral or dissatisfied	10583
19	Eco	3	satisfied	1745
20	Eco	4	neutral or dissatisfied	9566
21	Eco	4	satisfied	3256
22	Eco	5	neutral or dissatisfied	8434
23	Eco	5	satisfied	3080
24	Eco Plus	0	neutral or dissatisfied	11
25	Eco Plus	0	satisfied	10
26	Eco Plus	1	neutral or dissatisfied	1225
27	Eco Plus	1	satisfied	195
28	Eco Plus	2	neutral or dissatisfied	1707
29	Eco Plus	2	satisfied	249
30	Eco Plus	3	neutral or dissatisfied	1681
31	Eco Plus	3	satisfied	296
32	Eco Plus	4	neutral or dissatisfied	1346
33	Eco Plus	4	satisfied	770
34	Eco Plus	5	neutral or dissatisfied	1100
35	Eco Plus	5	satisfied	790

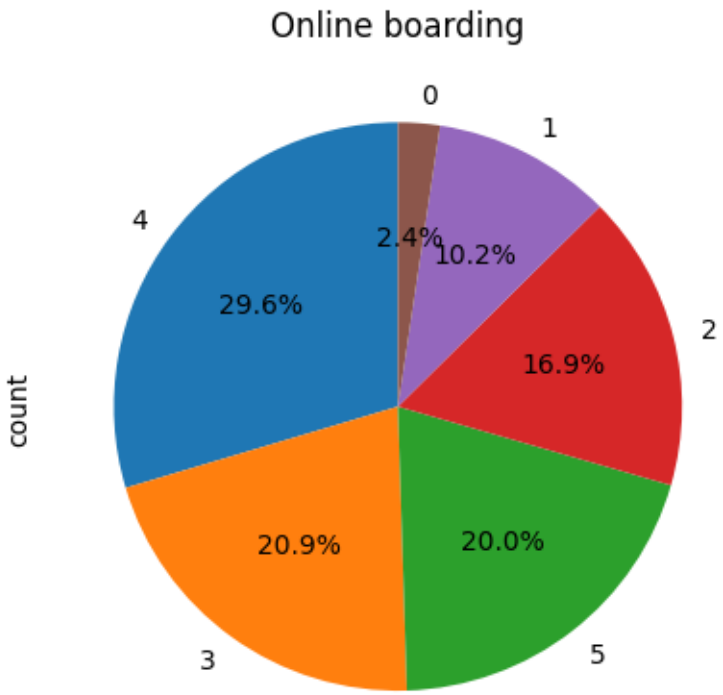
[146]:
```
plt.figure(figsize=(10, 5))
sns.barplot(x='Food and drink', y='Jumlah', hue='satisfaction', data=df_food)
plt.title('Food and drink vs Satisfaction')
plt.show()
```



Berdasarkan data di atas, layanan food and drink mendapat penilaian kepuasan pelanggan dengan skala 4 dari 5 paling banyak dari keseluruhan data penumpang sebesar 23.5%. Dari rincian data di

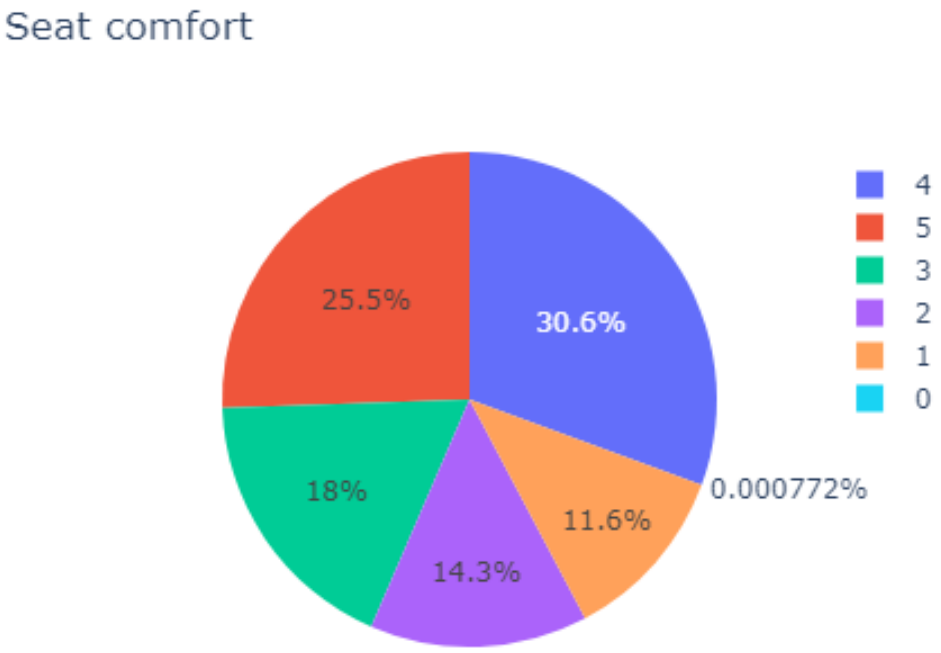
atas, berdasarkan kategori class dapat mempengaruhi kepuasan terhadap makanan dan minuman yang sediakan.

[147]:
```
df_merge_encoded['Online boarding'].value_counts().plot.pie(autopct='%1.1f%%',␣
↪startangle=90)
plt.title('Online boarding')
plt.show()
```

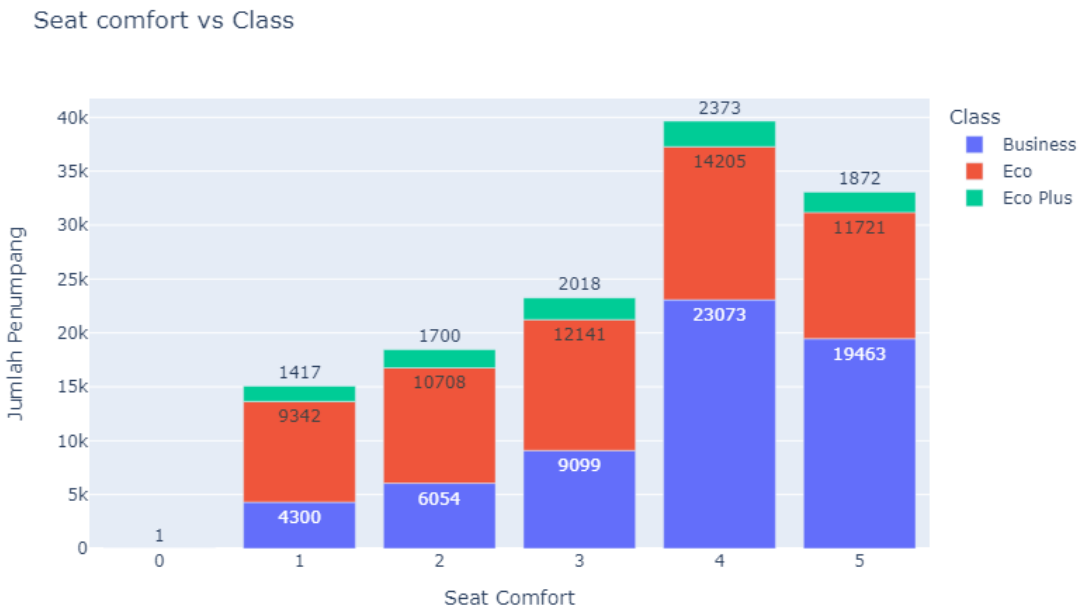


Berdasarkan data di atas, layanan online boarding mendapat penilaian kepuasan pelanggan dengan skala 4 dari 5 paling banyak dari keseluruhan data penumpang sebesar 29.6%. Dapat disimpulkan bahwa layanan ini cukup memberikan pelayanan yang baik.

[148]:
```
df_seat_comfot = df_merge_encoded['Seat comfort'].value_counts().reset_index().
↪rename(columns={'count':'Jumlah'})
plot = px.pie(df_seat_comfot, values='Jumlah', names=df_seat_comfot['Seat␣
↪comfort'], title='Seat comfort', height=400, width=500)
plot.show()
```

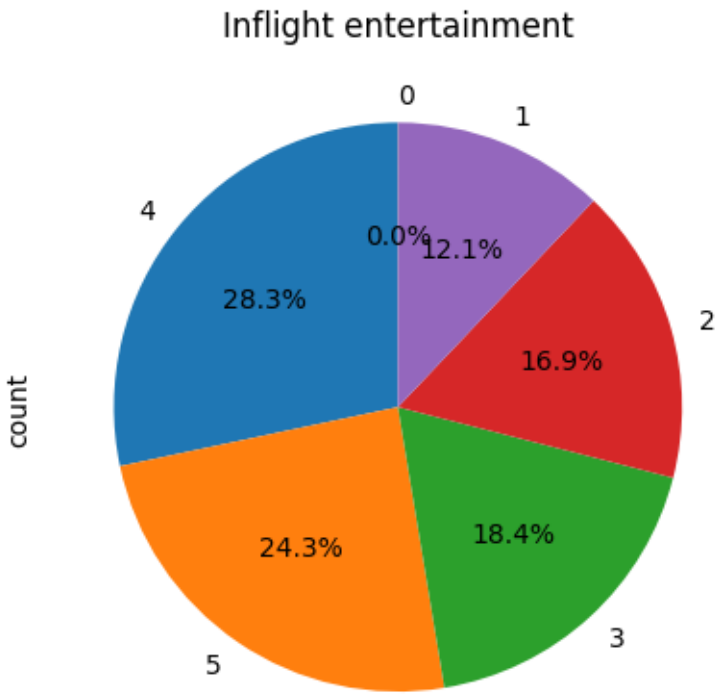


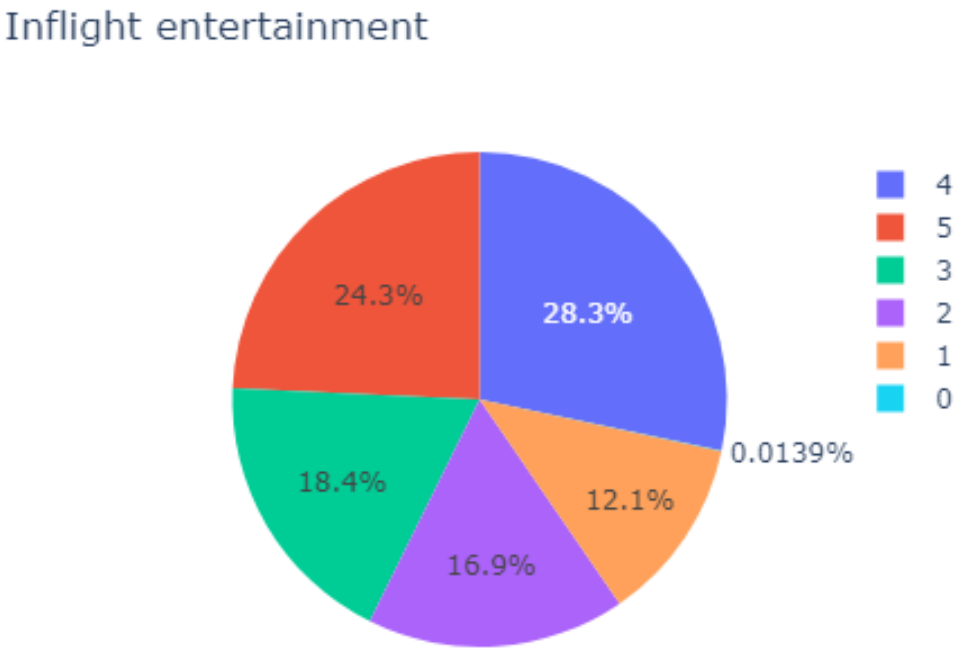
[149]:
```
df_seat = df_merge.groupby(['Seat comfort', 'Class']).size().
 ↪reset_index(name='Jumlah')

fig_bar_seat = px.bar(df_seat, x='Seat comfort', y='Jumlah', color='Class',␣
 ↪title='Seat comfort vs Class',
                      labels={'Seat comfort': 'Seat Comfort', 'Jumlah': 'Jumlah␣
 ↪Penumpang'}, text='Jumlah')
fig_bar_seat.update_layout(width=800, height=500)
fig_bar_seat.show()
```



Berdasarkan data di atas, layanan seat comfort mendapat penilaian kepuasan pelanggan dengan skala 4 dari 5 paling banyak dari keseluruhan data penumpang sebesar 30.6%. Dengan detail, pada rating 4 dan 5 ditempati oleh class business dan pada rating 4 cukup banyak dibandingkan seluruh rate di semua class. Dapat disimpulkan tempat duduk di semua class cukup nyaman.

[150]:
```
df_inflight_entertainment = df_merge_encoded['Inflight entertainment'].
  ↪value_counts().reset_index().rename(columns={'count':'Jumlah'})
plot = px.pie(df_inflight_entertainment, values='Jumlah',␣
  ↪names=df_inflight_entertainment['Inflight entertainment'], title='Inflight␣
  ↪entertainment', height=400, width=500)
plot.show()
```

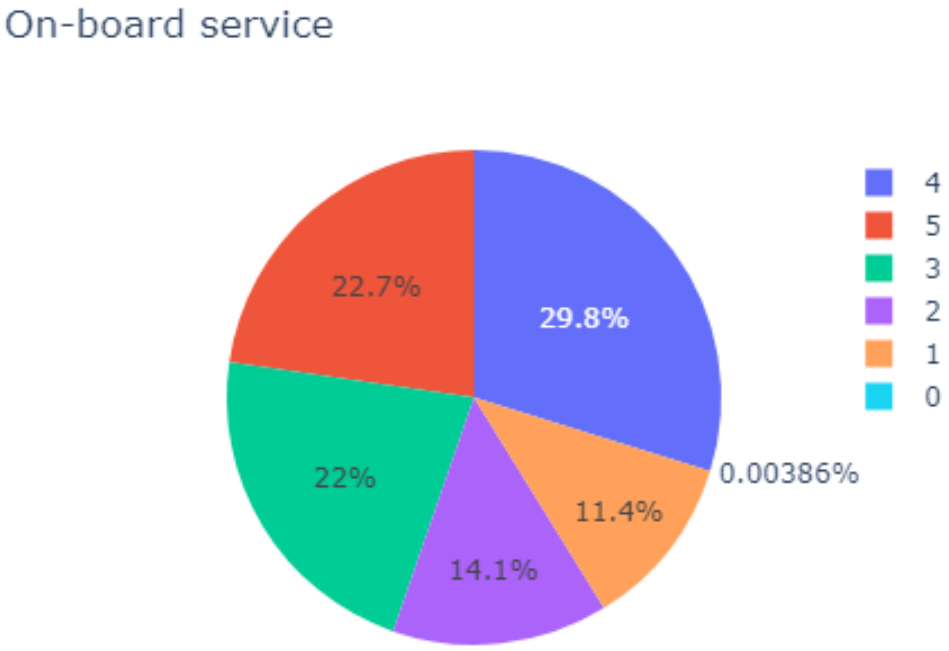

Inflight entertainment
0
1
2
3
4
5
0.0%
12.1%
16.9%
18.4%
24.3%
28.3%
count



[151]:
```
df_inflight = df_merge.groupby(['Inflight entertainment', 'Class']).size().
 ↪reset_index(name='Jumlah')

fig_bar_inflight = px.bar(df_inflight, x='Inflight entertainment', y='Jumlah',␣
 ↪color='Class', title='Inflight entertainment vs Class',
                      labels={'Inflight entertainment': 'Inflight␣
 ↪entertainment', 'Jumlah': 'Jumlah Penumpang'}, text='Jumlah')
fig_bar_inflight.update_layout(width=800, height=500)
fig_bar_inflight.show()
```

Berdasarkan data di atas, layanan inflight entertainment mendapat penilaian kepuasan pelanggan dengan skala 4 dari 5 paling banyak dari keseluruhan data penumpang sebesar 28.3%. Dari visualisasi jumlah rating untuk masing-masing class yang tertinggi terdapat di rating 4 yang dapat dikategorikan cukup baik untuk pelayanan ini untuk semua class.

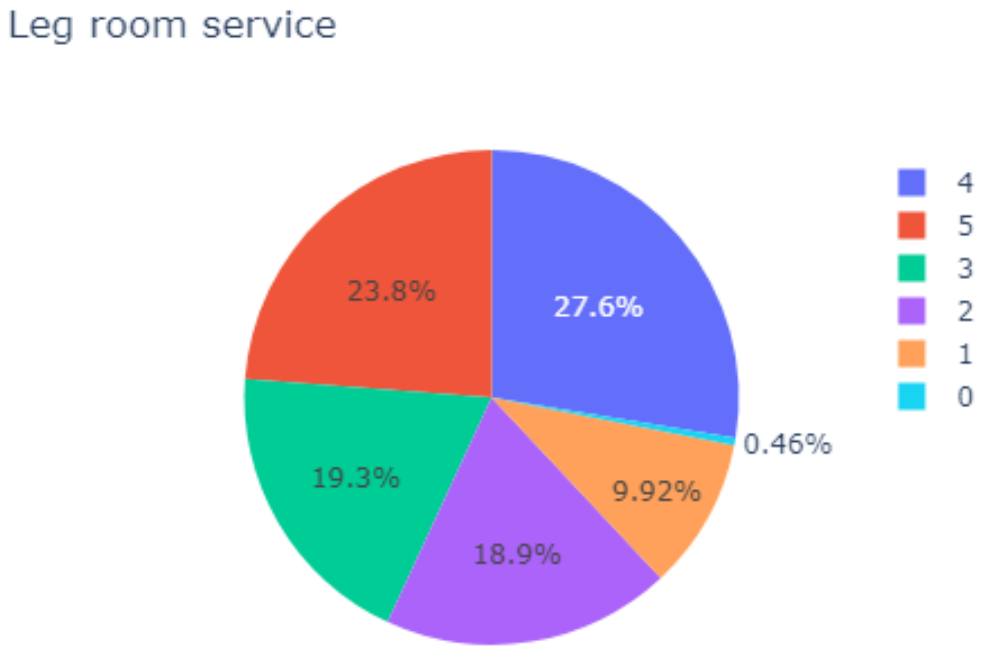
[152]:
```
df_onboard = df_merge_encoded['On-board service'].value_counts().reset_index().
  ↪rename(columns={'count': 'Jumlah'})
plot = px.pie(df_onboard, values='Jumlah', names=df_onboard['On-board␣
  ↪service'], title='On-board service', height=400, width=500)
plot.show()
```



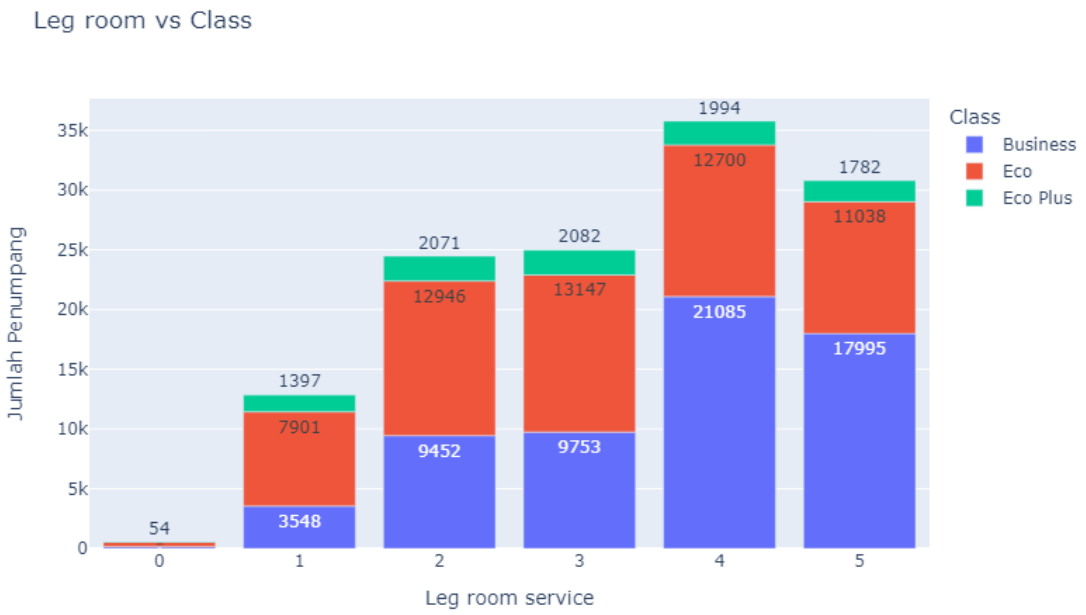
[153]:
```
df_board = df_merge.groupby(['On-board service', 'Class']).size().
 ↪reset_index(name='Jumlah')

fig_bar_board = px.bar(df_board, x='On-board service', y='Jumlah',␣
 ↪color='Class', title='On-board service vs Class',
                       labels={'On-board service': 'On-board service', 'Jumlah':␣
 ↪'Jumlah Penumpang'}, text='Jumlah')
fig_bar_board.update_layout(width=800, height=500)
fig_bar_board.show()
```

Berdasarkan data di atas, layanan on-board service terdapat tren signifikan di mana sebagian besar penumpang menilai layanan dalam kisaran 1 hingga 3 merupakan kelompok class ekonomi yang melaporkan tingkat kepuasan yang lebih rendah. Sedangkan, penumpang dengan rating 4 atau 5 menunjukkan tingkat kepuasan yang lebih tinggi berasal dari class business. Oleh karena itu, dapat disimpulkan untuk pelayanan di dalam pesawat harus ditingkatkan pada class ekonomi.

[154]:
```
df_legroom = df_merge_encoded['Leg room service'].value_counts().reset_index().
  ↪rename(columns={'count': 'Jumlah'})
plot = px.pie(df_legroom, values='Jumlah', names=df_legroom['Leg room␣
  ↪service'], title='Leg room service', height=400, width=500)
plot.show()
```

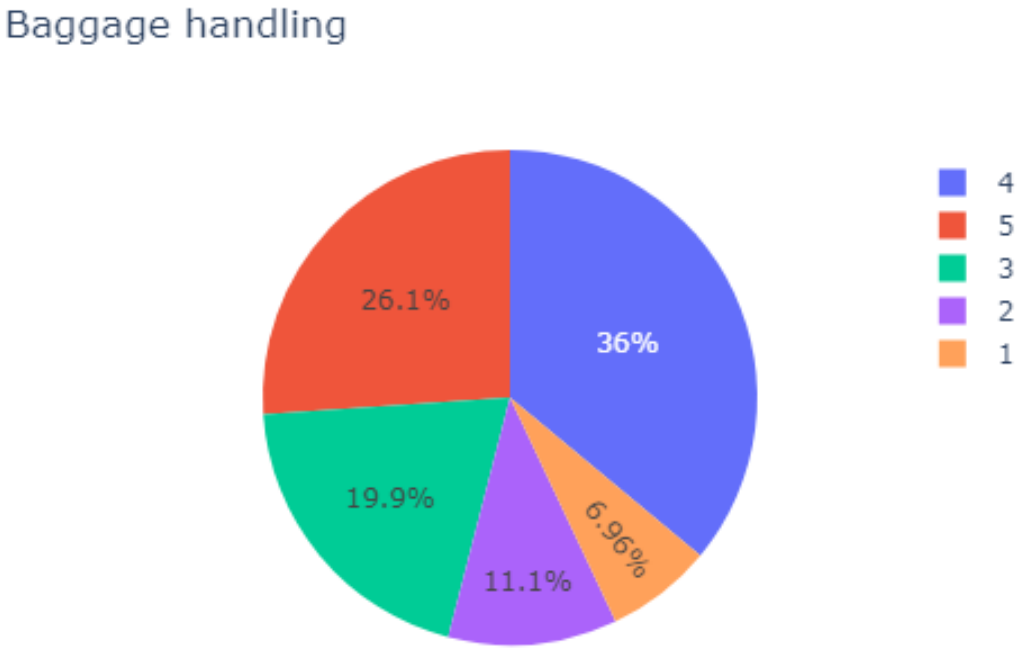


[155]:
```
df_legroom = df_merge.groupby(['Leg room service', 'Class']).size().
 ↪reset_index(name='Jumlah')

fig_bar_legroom = px.bar(df_legroom, x='Leg room service', y='Jumlah',␣
 ↪color='Class', title='Leg room vs Class',
                      labels={'Leg room': 'Leg room', 'Jumlah': 'Jumlah␣
 ↪Penumpang'}, text='Jumlah')
fig_bar_legroom.update_layout(width=800, height=500)
fig_bar_legroom.show()
```

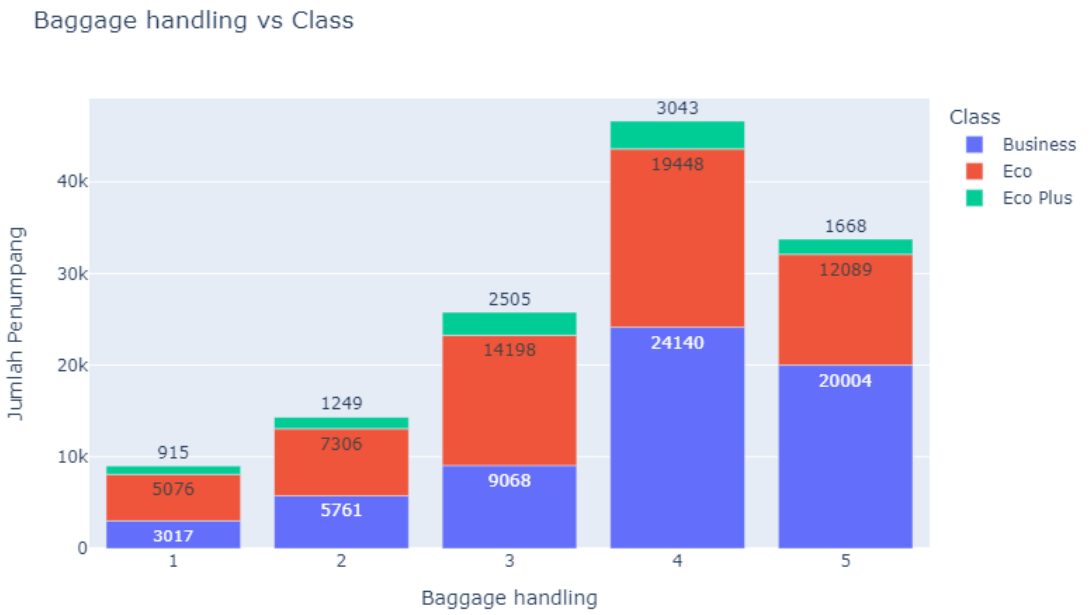


Berdasarkan data di atas, layanan leg room terdapat tren signifikan di mana sebagian besar penumpang menilai layanan dalam kisaran 1 hingga 3 merupakan kelompok class ekonomi yang melaporkan tingkat kepuasan yang lebih tinggi. Sedangkan, penumpang dengan rating 4 atau 5 menunjukkan tingkat kepuasan yang lebih tinggi berasal dari class business. Dapat disimpulkan untuk pelayanan leg room pada class ekonomi masih kurang.

[156]:
```
df_baggage_handling = df_merge_encoded['Baggage handling'].value_counts().
  ↪reset_index().rename(columns={'count': 'Jumlah'})
plot = px.pie(df_baggage_handling, values='Jumlah',␣
  ↪names=df_baggage_handling['Baggage handling'], title='Baggage handling',␣
  ↪height=400, width=500)
plot.show()
```

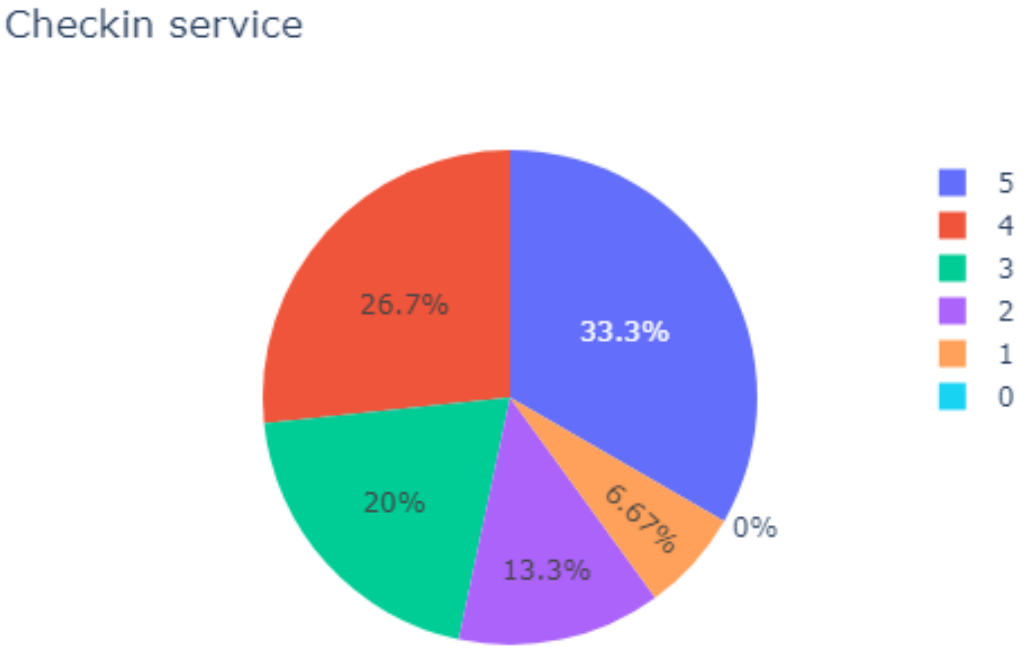


[157]:
```
df_baggage = df_merge.groupby(['Baggage handling', 'Class']).size().
 ↪reset_index(name='Jumlah')

fig_bar_baggage = px.bar(df_baggage, x='Baggage handling', y='Jumlah',␣
 ↪color='Class', title='Baggage handling vs Class',
                      labels={'Baggage handling': 'Baggage handling', 'Jumlah':␣
 ↪'Jumlah Penumpang'}, text='Jumlah')
fig_bar_baggage.update_layout(width=800, height=500)
fig_bar_baggage.show()
```



Berdasarkan data di atas, layanan baggage handling terdapat tren signifikan di mana sebagian besar penumpang menilai layanan dalam kisaran 1 hingga 3 merupakan kelompok class ekonomi yang melaporkan tingkat kepuasan yang lebih tinggi. Sedangkan, penumpang dengan rating 4 atau 5 menunjukkan tingkat kepuasan yang lebih tinggi berasal dari class business. Dapat disimpulkan untuk pelayanan bagasi pada class business cukup baik dan class ekonomi masih kurang maka perlu ditingkatkan.

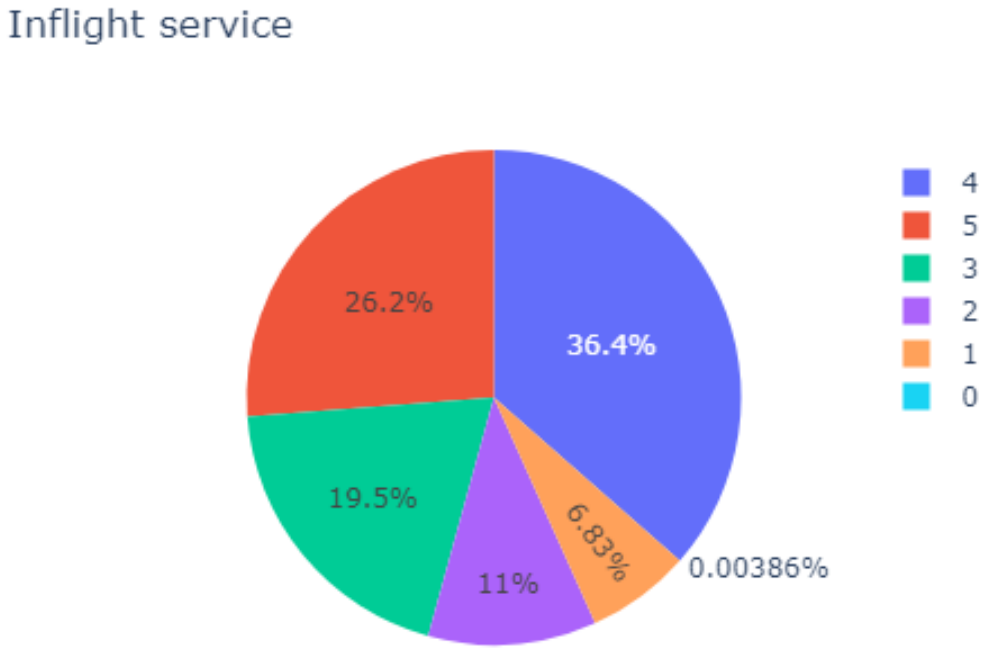
[158]:
```
df_checkin = df_merge_encoded['Checkin service'].value_counts().reset_index().
  ↪rename(columns={'count': 'Jumlah'})
plot = px.pie(df_checkin, values='Checkin service', names=df_checkin['Checkin␣
  ↪service'], title='Checkin service', height=400, width=500)
plot.show()
```



[159]:
```
df_checkin = df_merge.groupby(['Checkin service', 'Class']).size().
 ↪reset_index(name='Jumlah')

fig_bar_checkin = px.bar(df_checkin, x='Checkin service', y='Jumlah',␣
 ↪color='Class', title='Checkin service vs Class',
                      labels={'Checkin service': 'Checkin service', 'Jumlah':␣
 ↪'Jumlah Penumpang'}, text='Jumlah')
fig_bar_checkin.update_layout(width=800, height=500)
fig_bar_checkin.show()
```

Berdasarkan data di atas, layanan checkin service terdapat tren signifikan di mana sebagian besar penumpang menilai layanan dalam kisaran 1 dan 2 merupakan kelompok class ekonomi yang melaporkan tingkat kepuasan yang lebih tinggi. Sedangkan, penumpang dengan rating 3, 4, dan 5 menunjukkan tingkat kepuasan yang lebih tinggi berasal dari class business. Dapat disimpulkan untuk pelayanan checkin pada penumpang business lebih baik.

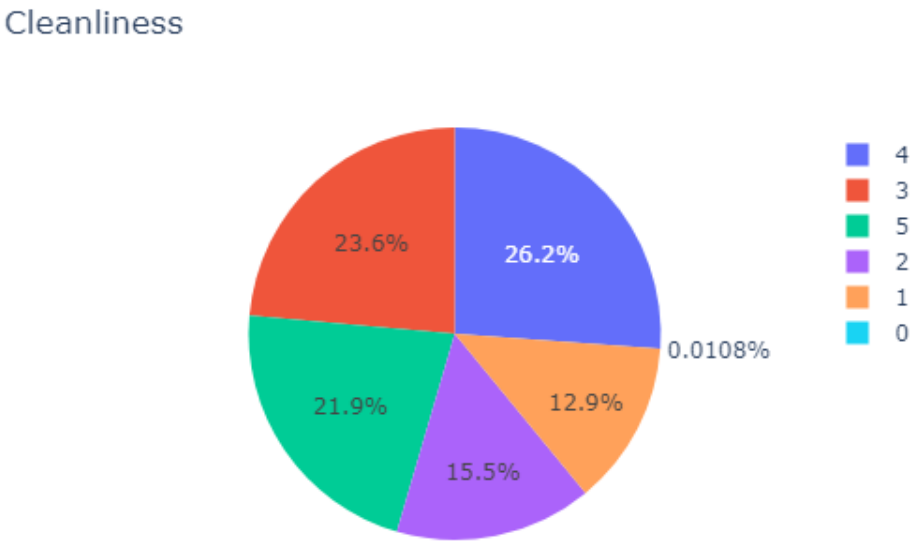
[160]:
```
df_inflight_service = df_merge_encoded['Inflight service'].value_counts().
  ↪reset_index().rename(columns={'count': 'Jumlah'})
plot = px.pie(df_inflight_service, values='Jumlah',␣
  ↪names=df_inflight_service['Inflight service'], title='Inflight service',␣
  ↪height=400, width=500)
plot.show()
```



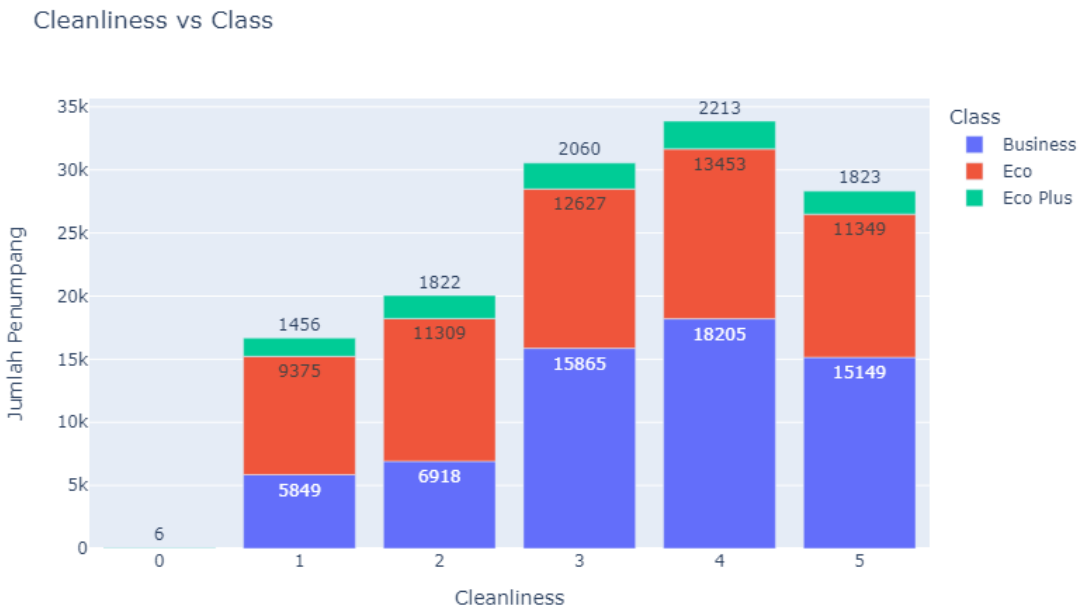
[161]:
```
df_inflight_service = df_merge.groupby(['Inflight service', 'Class']).size().
↪reset_index(name='Jumlah')

fig_bar_inflight_service = px.bar(df_inflight_service, x='Inflight service',␣
↪y='Jumlah', color='Class', title='Inflight service vs Class',
                    labels={'Inflight service': 'Inflight service', 'Jumlah':␣
↪'Jumlah Penumpang'}, text='Jumlah')
fig_bar_inflight_service.update_layout(width=800, height=500)
fig_bar_inflight_service.show()
```

Berdasarkan data di atas, layanan inflight service terdapat tren signifikan di mana sebagian besar penumpang menilai layanan dalam kisaran 1 hingga 3 merupakan kelompok class ekonomi yang melaporkan tingkat kepuasan yang lebih tinggi. Sedangkan, penumpang dengan rating 4 atau 5 menunjukkan tingkat kepuasan yang lebih tinggi berasal dari class business. Dapat disimpulkan untuk pelayanan penerbangan harus ditingkatkan pada class ekonomi.

[162]:
```
df_cleanliness = df_merge_encoded['Cleanliness'].value_counts().reset_index().
  ↪rename(columns={'count': 'Jumlah'})
plot = px.pie(df_cleanliness, values='Jumlah',␣
  ↪names=df_cleanliness['Cleanliness'], title='Cleanliness', height=400,␣
  ↪width=600)
plot.show()
```



[163]:
```
df_cleanliness = df_merge.groupby(['Cleanliness', 'Class']).size().
 ↪reset_index(name='Jumlah')

fig_bar_cleanliness = px.bar(df_cleanliness, x='Cleanliness', y='Jumlah',␣
 ↪color='Class', title='Cleanliness vs Class',
                      labels={'Cleanliness': 'Cleanliness', 'Jumlah': 'Jumlah␣
 ↪Penumpang'}, text='Jumlah')
fig_bar_cleanliness.update_layout(width=800, height=500)
fig_bar_cleanliness.show()
```



Untuk pelayanan kebersihan, penilaian terbesar terdapat pada rating 4 dengan setiap classnya paling banyak menilai dengan rating 4, tetapi untuk rating 1 sampai 3 masih terbilang besar, sehingga diperlukan peningkatan kebersihan pada layanan maskapai guna meningkatkan kenyamanan penumpang.

2.3 Pengukuran Kinerja Algoritma

Untuk mengukur kinerja dari model klasifikasi terdapat beberapa metrik yang dapat dijadikan acuan. Dalam penelitian ini kami menggunakan metrik *Accuracy*, *Precision*, dan *Recall*.

- Accuracy adalah rasio prediksi yang benar (baik positif maupun negatif) dengan keseluruhan data. Rumusnya adalah:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- Precision adalah rasio prediksi yang benar positif dengan keseluruhan hasil yang diprediksi positif. Rumusnya adalah:

$$Precision = \frac{TP}{TP + FP}$$

- Recall adalah rasio prediksi yang benar positif dengan keseluruhan data ynag sebenarnya positif. Rumusnya adalah:

$$Recall = \frac{TP}{TP + FN}$$

2.3.1 Klasifikasi Naive Bayes

Algoritma Naive Bayes adalah algoritma klasifikasi yang berbasis probabilitas menggunakan teorema Bayes untuk menghitung kemungkinan suatu data masuk ke dalam kelas tertentu. Rumus teorema Bayes adalah sebagai berikut:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

Berikut adalah implementasi metode klasifikasi menggunakan algoritma `Naive Bayes` menggunakan Python. Jumlah data train sebesar 75% dan data test sebesar 25% dari total keseluruhan.

[20]:
```
df_encode = df_merge_encoded.iloc[:, 1:23]
df_encode.head()
```

[20]:

	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	\
0	1	0	13	1	2	460	
1	1	1	25	0	0	235	
2	0	0	26	0	0	1142	
3	0	0	25	0	0	562	
4	1	0	61	0	0	214	

	Inflight wifi service	Departure/Arrival time convenient	\
0	3	4	
1	3	2	
2	2	2	
3	2	5	
4	3	3	

	Ease of Online booking	Gate location	…	Seat comfort	\
0	3	1	…	5	
1	3	3	…	1	
2	2	2	…	5	
3	5	5	…	2	
4	3	3	…	5	

	Inflight entertainment	On-board service	Leg room service	\
0	5	4	3	
1	1	1	5	
2	5	4	3	
3	2	2	5	
4	3	3	4	

	Baggage handling	Checkin service	Inflight service	Cleanliness	\
0	4	4	5	5	
1	3	1	4	1	
2	4	4	4	5	
3	3	1	4	2	

```
4                 4                 3                 3             3

   Departure Delay in Minutes  Arrival Delay in Minutes
0                          25                      18.0
1                           1                       6.0
2                           0                       0.0
3                          11                       9.0
4                           0                       0.0

[5 rows x 22 columns]
```

[165]:
```
X = df_encode
y = df_merge['satisfaction']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

[166]:
```
nbModel = GaussianNB()
nbModel = nbModel.fit(X_train, y_train)
y_pred = nbModel.predict(X_test)
y_pred
```

[166]:
```
array(['neutral or dissatisfied', 'satisfied', 'neutral or dissatisfied',
       …, 'neutral or dissatisfied', 'neutral or dissatisfied',
       'neutral or dissatisfied'], dtype='<U23')
```

Menampilkan confusion matrix untuk merepresentasikan prediksi dengan kondisi sebenarnya (aktual) pada data yang dianalisis.

[167]:
```
confusion_matrix(y_test, y_pred)
```

[167]:
```
array([[16457,  1793],
       [ 2605, 11517]], dtype=int64)
```

[168]:
```
y_pred = nbModel.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred).round(2))
report = classification_report(y_test, y_pred)
print(report)
```

```
Accuracy: 0.86
                          precision    recall  f1-score   support

neutral or dissatisfied       0.86      0.90      0.88     18250
              satisfied       0.87      0.82      0.84     14122

               accuracy                           0.86     32372
```

```
           macro avg       0.86      0.86      0.86     32372
        weighted avg       0.86      0.86      0.86     32372
```

Klasifikasi menggunakan metode Naive Bayes menghasilkan nilai sebagai berikut:

- Accuracy dengan nilai `0.86` yang berarti model klasifikasi menggunakan algorimt Naive Bayes dapat mengklasifikasikan data dengan benar sebanyak `86%`.
- Precision dengan nilai `0.87` yang berarti semua data yang diprediksi sebagai label `1` atau `satisfied` dengan benar sebesar `87%`.
- Recall dengan nilai `0.82` yang berarti data yang sebenarnya diberi label `1` sebanyak `82%` dari keseluruhan data.

2.3.2 Klasifikasi K-NN

Algoritma K-NN adalah algoritma klasifikasi dengan prinsip mencari jarak terdekat antara data yang dievaluasi dengan K tetangga (neighbor). Dalam penentuan jarak antar data terdapat beberapa rumus, misalnya `Euclidean`, `Hamming`, `Manhattan`, dan `Minkowski`. Pada penelitin ini kami menggunakan rumus `Euclidean`, dengan rumusnya sebagai berikut:

$$d_i = \sqrt{\sum_{i=1}^{p}(X_{2i} - X_{1i})^2}$$

di mana:

X1 = Sampel Data

X2 = Data Uji

i = variabel data

d = jarak

p = dimensi data

Berikut adalah implementasi K-NN menggunakan Python. Jumlah data train sebesar 75% dan data test sebesar 25% dari total keseluruhan.

[169]:
```
X = df_encode
y = df_merge['satisfaction']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

[170]:
```
knn_models = KNeighborsClassifier(n_neighbors=int(math.sqrt(X_train.shape[0])),␣
 ↪metric='euclidean')
knn_models.fit(X_train_scaled, y_train)
```

[170]:
```
KNeighborsClassifier(metric='euclidean', n_neighbors=311)
```

[171]:
```
y_pred = knn_models.predict(X_test_scaled)
y_pred
```

[171]:
```
array(['satisfied', 'neutral or dissatisfied', 'satisfied', …,
       'satisfied', 'neutral or dissatisfied', 'neutral or dissatisfied'],
      dtype=object)
```

Menampilkan confusion matrix untuk merepresentasikan prediksi dengan kondisi sebenarnya (aktual) pada data yang dianalisis.

[172]:
```
confusion_matrix(y_test, y_pred)
```

[172]:
```
array([[17671,   726],
       [ 2244, 11731]], dtype=int64)
```

[173]:
```
accuracy = accuracy_score(y_test, y_pred).round(2)
report = classification_report(y_test, y_pred)
print("Accuracy:", accuracy)
print(report)
```

```
Accuracy: 0.91
                         precision    recall  f1-score   support

neutral or dissatisfied       0.89      0.96      0.92     18397
              satisfied       0.94      0.84      0.89     13975

               accuracy                           0.91     32372
              macro avg       0.91      0.90      0.91     32372
           weighted avg       0.91      0.91      0.91     32372
```

Klasifikasi menggunakan metode K-Nearest Neighbor (K-NN) menghasilkan peningkatan nilai di beberapa metrik dengan rincian sebagai berikut:

- Accuracy dengan nilai `0.91` yang berarti model klasifikasi menggunakan algorimta K-NN dapat mengklasifikasikan data dengan benar sebanyak `91%`.
- Precision dengan nilai `0.94` yang berarti semua data yang diprediksi sebagai label `1` atau `satisfied` dengan benar sebesar `94%`.
- Recall dengan nilai `0.84` yang berarti data yang sebenarnya diberi label `1` sebanyak `84%` dari keseluruhan data.

Dari hasil di atas menunjukkan algoritma K-NN lebih optimal daripada algoritma Naive Bayes

2.3.3 Decision Tree

Decision Tree adalah metode pemodelan prediktif dalam analisis data yang menggunakan struktur pohon. Tujuannya untuk menggambarkan serta membuat keputusan berdasarkan serangkain aturan dan kondisi.

Gini adalah suatu index untuk membagi data menjadi kelas-kelas yang berbeda. Nilai Gini bervariasi antara 0 dan 1, di mana 0 berarti data murni (semua data termasuk dalam satu kelas) dan 1 berarti data tidak murni (data tersebar secara acak di berbagai kelas). Rumus untuk menghitung nilai Gini untuk suatu fitur adalah:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

di mana n adalah jumlah kelas yang ada, dan pi adalah proporsi data yang termasuk dalam kelas ke-1.

Berikut adalah implementasi Decisiton Tree menggunakan Python. Jumlah data train sebesar 75% dan data test sebesar 25% dari total keseluruhan.

[174]:
```
treeModel = DecisionTreeClassifier(max_depth=10)
X = df_encode
y = df_merge['satisfaction']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)


treeModel.fit(X_train, y_train)

y_pred = treeModel.predict(X_test)
y_pred
```

[174]:
```
array(['neutral or dissatisfied', 'satisfied', 'satisfied', …,
       'satisfied', 'neutral or dissatisfied', 'satisfied'], dtype=object)
```

Menampilkan confusion matrix untuk merepresentasikan prediksi dengan kondisi sebenarnya (aktual) pada data yang dianalisis.

[175]:
```
confusion_matrix(y_test, y_pred)
```

[175]:
```
array([[17760,   611],
       [ 1129, 12872]], dtype=int64)
```

[176]:
```
print("Accuracy:", accuracy_score(y_test, y_pred).round(2))
report = classification_report(y_test, y_pred)
print(report)
```

```
Accuracy: 0.95
                          precision    recall  f1-score   support

neutral or dissatisfied       0.94      0.97      0.95     18371
              satisfied       0.95      0.92      0.94     14001

               accuracy                           0.95     32372
              macro avg       0.95      0.94      0.94     32372
```

```
weighted avg       0.95      0.95      0.95     32372
```

Klasifikasi menggunakan metode Decision Tree menghasilkan peningkatan nilai di beberapa metrik dengan rincian sebagai berikut:

- Accuracy dengan nilai `0.95` yang berarti model klasifikasi menggunakan algoritma Decision Tree dapat mengklasifikasikan data dengan benar sebanyak `95%`.
- Precision dengan nilai `0.95` yang berarti semua data yang diprediksi sebagai label `1` atau `satisfied` dengan benar sebesar `95%`.
- Recall dengan nilai `0.92` yang berarti data yang sebenarnya diberi label `1` sebanyak `92%` dari keseluruhan data.

Dari hasil di atas menunjukkan algoritma Decision Tree lebih optimal daripada algoritma K-NN

[177]:
```
tree_dot = StringIO()
export_graphviz(treeModel, out_file=tree_dot, feature_names=X.columns,␣
 ↪class_names=df_merge['satisfaction'].unique().tolist(), rounded=True,␣
 ↪filled=True, special_characters=True)
graph = pydotplus.graph_from_dot_data(tree_dot.getvalue())

name_file = './Assets/DecisionTree.dot'

# check if directory exist
if not os.path.exists('./Assets'):
    os.makedirs('./Assets')

with open(name_file, 'w') as png:
    png.write(graph.to_string())
```

[178]:
```
if shutil.which('dot') is None:
    print("Install Graphviz for Windows: https://graphviz.gitlab.io/_pages/
 ↪Download/Download_windows.html")
else:
    os.system('dot -Tsvg ./Assets/DecisionTree.dot -o ./Assets/DecisionTree.svg␣
 ↪&& start \"\" ./Assets/DecisionTree.svg')
```

3 Kesimpulan

Berdasarkan hasil analisis data di atas label kepuasan pelanggan dibagi menjadi 2 kategori yaitu: kategori `satisfied` dengan persentase `43.4%`, lebih kecil dari kategori `neutral or dissatisfied` yang memiliki persentase `56.6%`. Tingkat kepuasan ini ditentukan dari hasil penilaian kepuasan pelanggan terhadap layanan yang disediakan maskapai penerbangan. Dalam penilaian pelayanan dikaitkan dengan kategori kelas penumpang, berdasarkan data kelas `business` memiliki persentase kelas terbesar dari jumlah seluruh penumpang.

Dari hasil penilaian pelayanan yang telah dilakukan dapat disimpulkan dengan rincian layanan sebagai berikut:

- Inflight Wifi Service

- Departure/Arrival Time Convenient
- Ease of Online Booking
- Food and Drink
- Online Boarding
- Seat Comfort
- Inflight Entertainment
- Onboard Service
- Leg Room Service
- Baggage Handling
- Check-in Service
- Inflight Service
- Cleanliness

Hasil penilaian pelanggan menunjukkan sebagian besar layanan mendapatkan tingkat kepuasan di skala 4 dari 5 yang di dominasi oleh kelas `business`. Untuk tingkat kepuasan pelanggan skala 1 sampai 3 di semua layanan didominasi oleh kelas `ekonomi`. Dapat disimpulkan kelas penerbangan mempengaruhi penyediaan layanan, sehingga perlu adanya peningkatan layanan pada kelas `ekonomi` agar meningkatkan kepuasan pelanggan.

Untuk hasil perbandingan kinerja algoritma klasifikasi yang kami lakukan, didapatkan hasil yang tercantum di tabel berikut:

Algoritma	Label	Accuracy	Precision	Recall
Naive Bayes	satisfied	0.86	0.87	0.82
K-NN	satisfied	0.91	0.94	0.84
Decision Tree	satisfied	0.95	0.95	0.92

Berdasarkan analisis confusion matrix dari ketiga algoritma dihasilkan False Negative pada algoritma Naive dan K-NN memiliki nilai yang lebih besar dibandingkan False Positive sehingga meningkatkan peluang type error lebih besar dalam pengambilan keputusan. Sedangkan, menggunakan algoritma Decision Tree nilai False Negative lebih kecil dari False Positive sehingga prediksi lebih akurat untuk mendapatkan hasil prediksi kepuasan.

4 Daftar Pustaka

Arthana, R. (2019, April 5). *Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning.* Medium. Retrieved December 6, 2023, from https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8

Dewi, A. C. (2021, February 6). *Klasifikasi Menggunakan Algoritma Decision Tree.* Medium. Retrieved December 6, 2023, from https://agneschintiadewi.medium.com/klasifikasi-menggunakan-algoritma-decision-tree-446d500ba73c

Nugroho, K. S. (2019, November 13). *Confusion Matrix untuk Evaluasi Model pada Supervised Learning.* Medium. Retrieved December 6, 2023, from https://ksnugroho.medium.com/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f

Rizky, M. S. (2019, February 17). *How to: Mengubah Data Kategori Menjadi Numerik dengan Python*. Medium. Retrieved December 6, 2023, from https://medium.com/@msifaulkiki/how-to-mengubah-data-kategori-menjadi-numerik-dengan-python-73f7b6e1639f

Srivastava, T. (2023, October 20). *A Complete Guide to K-Nearest Neighbors (Updated 2023)*. Analytics Vidhya. Retrieved December 6, 2023, from https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

W., D. D. (2021, July 9). *Data Preparation*. Medium. Retrieved December 6, 2023, from https://medium.com/sysinfo/data-preparation-78ef3da24347