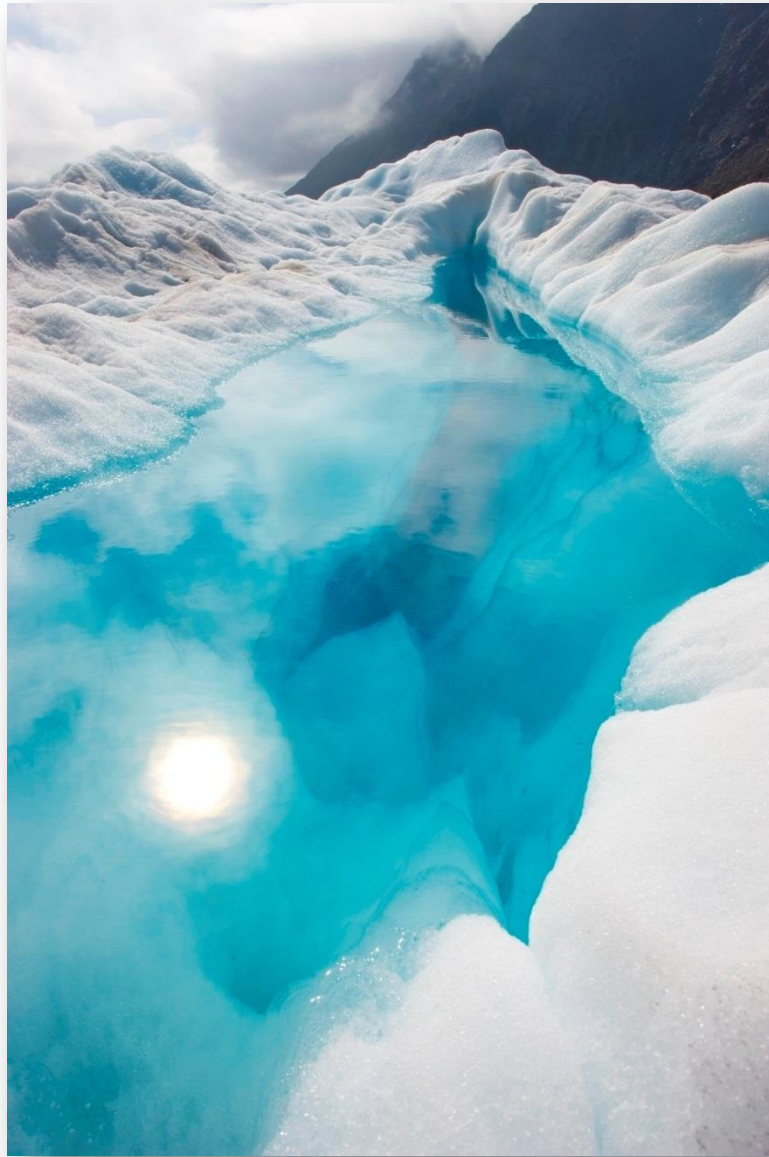


2023



BUKU KERJA/JOB SHEET

ASSOCIATE DATA SCIENTIST

Nama Peserta	:	Mohammad Hasan Tajuk Rizal
Nomor Urut	:	

DAFTAR ISI

DAFTAR ISI	1
BUKTI 1-ADS	2
1. Kebutuhan Data	2
2. Pengambilan Data.....	3
3. Pengintegrasian Data	5
BUKTI 2-ADS	7
1. Analisis Tipe dan Relasi Data	7
2. Analisis Karakteristik Data.....	8
3. Laporan Telaah Data	12
BUKTI 3-ADS	13
1. Pengecekan Kelengkapan Data	13
2. Rekomendasi Kelengkapan Data.....	15
BUKTI 4-ADS	16
1. Kriteria dan Teknik Pemilihan Data	16
2. Attributes (Columns) dan Records (Row) Data.....	16
BUKTI 5-ADS	18
1. Pembersihan Data Kotor	18
2. Laporan dan Rekomendasi Hasil Pembersihan Data Kotor.....	20
BUKTI 6-ADS	22
1. Analisis Teknik Transformasi Data	22
2. Transformasi Data	24
3. Dokumentasi Konstruksi Data.....	24
BUKTI 7-ADS	26
1. Pelabelan Data.....	26
2. Laporan Hasil Pelabelan Data	26
BUKTI 8-ADS	28
1. Parameter Model	28
2. Tools Pemodelan	29
BUKTI 9-ADS	31
1. Penggunaan Model dengan Data Riil	31
2. Penilaian Hasil Pemodelan	31

BUKTI 1-ADS

Kode Unit	:	J.62DMI00.004.1
Judul Unit	:	Mengumpulkan Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengumpulkan data untuk data science.

Langkah Kerja:

- 1) Menentukan kebutuhan data
- 2) Mengambil data
- 3) Mengintegrasikan data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengubah teks
 - Aplikasi basis data
 - Tools pengambilan data

1. KEBUTUHAN DATA

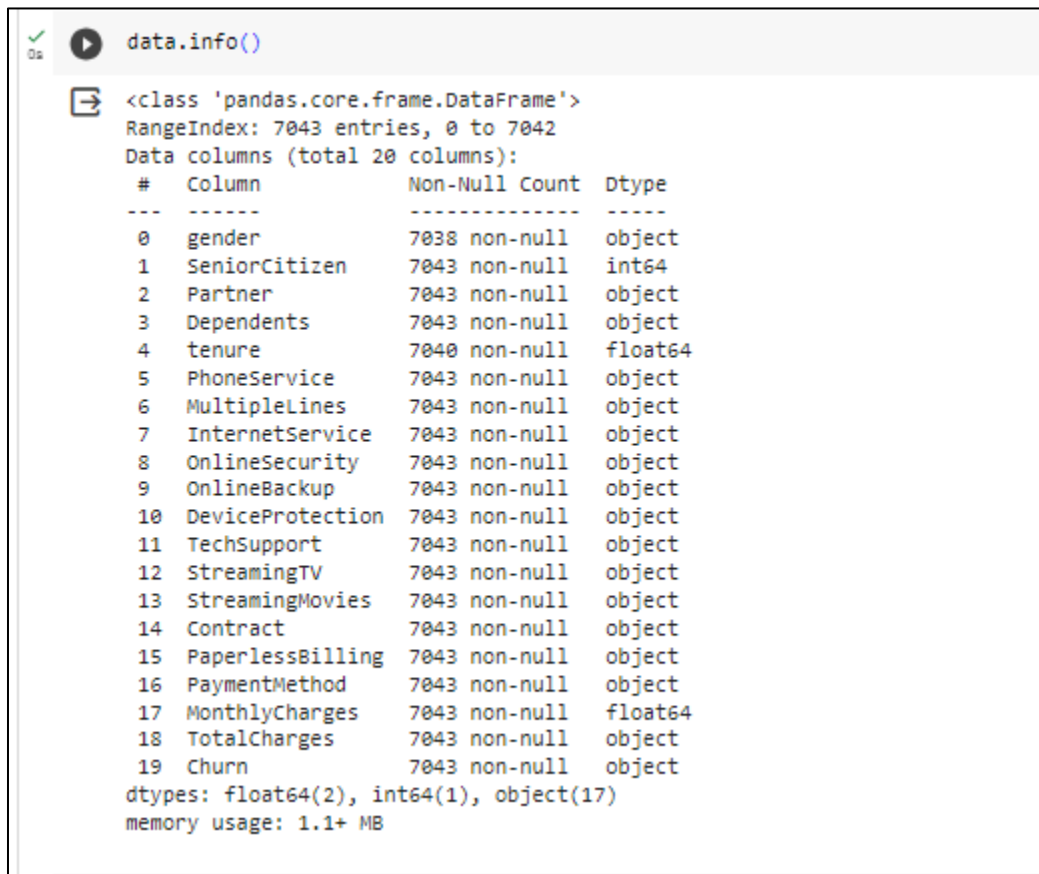
Instruksi Kerja:

- Identifikasi kebutuhan data sesuai tujuan teknis data science
- Periksa ketersediaan data berdasarkan kebutuhan data sesuai aturan yang berlaku
- Tentukan volume data berdasarkan kebutuhan data sesuai tujuan teknis data science

Kebutuhan data pada proyek ini untuk melakukan prediksi terhadap kemungkinan terjadinya customer churn. Diharapkan pemodelan yang dihasilkan dapat membantu perusahaan dalam memahami berapa banyak pelanggan yang meninggalkan bisnis, dan mengapa mereka keluar. Variable yang akan digunakan pada proyek ini adalah semua variable selain customerID, dan variable targetnya adalah kolom churn

```
0s [196] data = pd.read_csv('/content/Telco-Customer-Churn.csv')
0s [198] data.drop(['customerID'], axis=1, inplace=True)
0s data
```

Gambar 1. Screenshot code untuk menghapus customerID dari dataset



```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 20 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   gender                 7038 non-null  object  
1   SeniorCitizen          7043 non-null  int64   
2   Partner                7043 non-null  object  
3   Dependents             7043 non-null  object  
4   tenure                 7040 non-null  float64  
5   PhoneService           7043 non-null  object  
6   MultipleLines           7043 non-null  object  
7   InternetService        7043 non-null  object  
8   OnlineSecurity          7043 non-null  object  
9   OnlineBackup            7043 non-null  object  
10  DeviceProtection       7043 non-null  object  
11  TechSupport            7043 non-null  object  
12  StreamingTV            7043 non-null  object  
13  StreamingMovies        7043 non-null  object  
14  Contract               7043 non-null  object  
15  PaperlessBilling       7043 non-null  object  
16  PaymentMethod          7043 non-null  object  
17  MonthlyCharges         7043 non-null  float64  
18  TotalCharges           7043 non-null  object  
19  Churn                  7043 non-null  object  
dtypes: float64(2), int64(1), object(17)
memory usage: 1.1+ MB
```

Gambar 2. Screenshot code untuk memeriksa ketersediaan data



```
[202] data.shape

(7043, 20)
```

Gambar 3. Screenshot code untuk memeriksa volume data

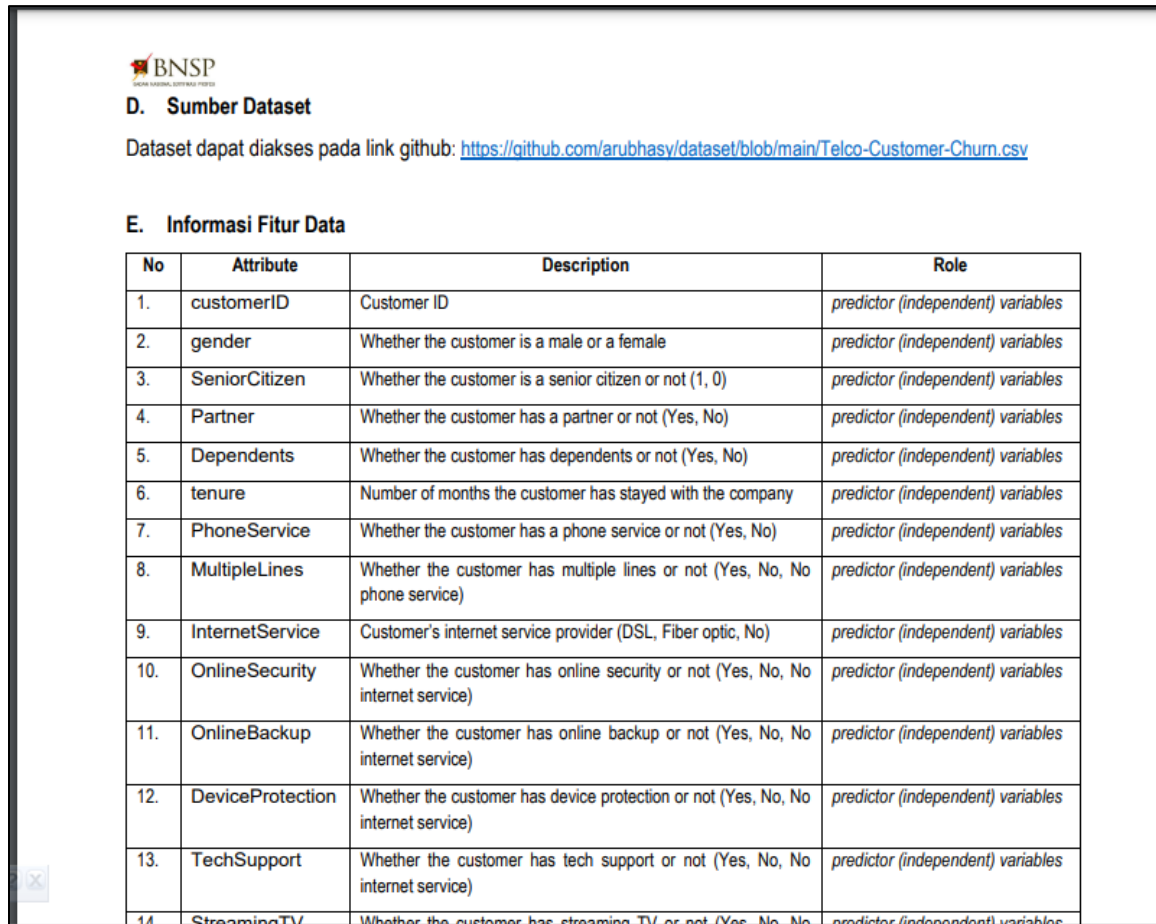
2. PENGAMBILAN DATA

Instruksi Kerja:

- Identifikasi metode dan tools pengambilan data sesuai tujuan teknis data science

- Tentukan tools pengambilan data sesuai tujuan teknis data science
- Siapkan tools pengambilan data sesuai tujuan teknis data science
- Jalankan proses pengambilan data sesuai dengan tools yang telah disiapkan

Metode pengambilan data yang digunakan pada proyek ini adalah metode web scrapping yaitu mendownload pada github melalui link yang telah disiapkan



D. Sumber Dataset

Dataset dapat diakses pada link github: <https://github.com/arubhasy/dataset/blob/main/Telco-Customer-Churn.csv>

E. Informasi Fitur Data

No	Attribute	Description	Role
1.	customerID	Customer ID	<i>predictor (independent) variables</i>
2.	gender	Whether the customer is a male or a female	<i>predictor (independent) variables</i>
3.	SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)	<i>predictor (independent) variables</i>
4.	Partner	Whether the customer has a partner or not (Yes, No)	<i>predictor (independent) variables</i>
5.	Dependents	Whether the customer has dependents or not (Yes, No)	<i>predictor (independent) variables</i>
6.	tenure	Number of months the customer has stayed with the company	<i>predictor (independent) variables</i>
7.	PhoneService	Whether the customer has a phone service or not (Yes, No)	<i>predictor (independent) variables</i>
8.	MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)	<i>predictor (independent) variables</i>
9.	InternetService	Customer's internet service provider (DSL, Fiber optic, No)	<i>predictor (independent) variables</i>
10.	OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)	<i>predictor (independent) variables</i>
11.	OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)	<i>predictor (independent) variables</i>
12.	DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)	<i>predictor (independent) variables</i>
13.	TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)	<i>predictor (independent) variables</i>
14.	StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)	<i>predictor (independent) variables</i>

Gambar 4. Screenshot sumber dataset

Tools pengambilan data yang digunakan adalah dengan mendownload menggunakan browser chrome dan tools pemrosesan yang digunakan adalah google colab dan library pandas



```
+ Code + Text

[238] import pandas as pd
      from sklearn.model_selection import train_test_split
      from sklearn.preprocessing import LabelEncoder
      from sklearn.tree import DecisionTreeClassifier
      from sklearn.linear_model import LogisticRegression
      from sklearn.pipeline import make_pipeline
      from sklearn.preprocessing import StandardScaler
      from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score
      import seaborn as sns
      import matplotlib.pyplot as plt

[196] data = pd.read_csv('/content/Telco-Customer-Churn.csv')
```

Gambar 5. Screenshot code pengambilan data dengan google colab dan pandas

3. PENGINTEGRASIAN DATA

Instruksi Kerja:

- Periksa integritas data sesuai tujuan teknis data science
- Integrasikan data sesuai tujuan teknis data science

Pada dataset ditemukan adanya inkonsistensi data pada beberapa kolom, sehingga perlu dilakukannya perbaikan pada type data tersebut. Kolom seniorcitizen adalah variable yang bersifat kategorikal dan variable total charges memiliki nilai numerik tetapi bertipe data objek sehingga harus diperbaiki.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 20 columns):
#   Column                Non-Null Count  Dtype
---  -
0   gender                 7038 non-null   object
1   SeniorCitizen          7043 non-null   int64
2   Partner                7043 non-null   object
3   Dependents             7043 non-null   object
4   tenure                 7040 non-null   float64
5   PhoneService           7043 non-null   object
6   MultipleLines          7043 non-null   object
7   InternetService        7043 non-null   object
8   OnlineSecurity         7043 non-null   object
9   OnlineBackup           7043 non-null   object
10  DeviceProtection       7043 non-null   object
11  TechSupport            7043 non-null   object
12  StreamingTV            7043 non-null   object
13  StreamingMovies        7043 non-null   object
14  Contract               7043 non-null   object
15  PaperlessBilling       7043 non-null   object
16  PaymentMethod          7043 non-null   object
17  MonthlyCharges         7043 non-null   float64
18  TotalCharges           7043 non-null   object
19  Churn                  7043 non-null   object
dtypes: float64(2), int64(1), object(17)
memory usage: 1.1+ MB

[381] data['SeniorCitizen'] = data['SeniorCitizen'].astype('object')
      data['TotalCharges'] = pd.to_numeric(data['TotalCharges'], errors='coerce')
      data['TotalCharges'] = data['TotalCharges'].astype('float64')
```

Gambar 6. Screenshot code perbaikan type data pada dataset

BUKTI 2-ADS

Kode Unit	:	J.62DMI00.005.1
Judul Unit	:	Menelaah Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam menelaah data untuk data science.

Langkah Kerja:

- 1) Menganalisis tipe dan relasi data
- 2) Menganalisis karakteristik data
- 3) Membuat laporan telaah data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Tools pengolahan data
 - Tools pembuat grafik

1. ANALISIS TIPE DAN RELASI DATA

Instruksi Kerja:

- Identifikasi tipe data yang terkumpul sesuai tujuan teknis
- Uraikan nilai atribut data yang terkumpul sesuai dengan batasan konteks bisnisnya
- Identifikasi relasi antar data yang terkumpul sesuai dengan tujuan teknis

Pada proses sebelumnya telah memperbaiki tipe data yang tidak konsisten. Untuk melakukan analisis korelasi data perlu dilakukannya pengelompokan terhadap variable yang bersifat numerik dan kategori. Seperti gambar berikut

```
kolom_kategori = [i for i in data.columns if data[i].dtype == 'object']
print(f'Data kategori: {kolom_kategori}\n')

kolom_numerik = [i for i in data.columns if data[i].dtype != 'object']
print(f'Data numerik: {kolom_numerik}')

Data kategori: ['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceF
Data numerik: ['tenure', 'MonthlyCharges', 'TotalCharges']

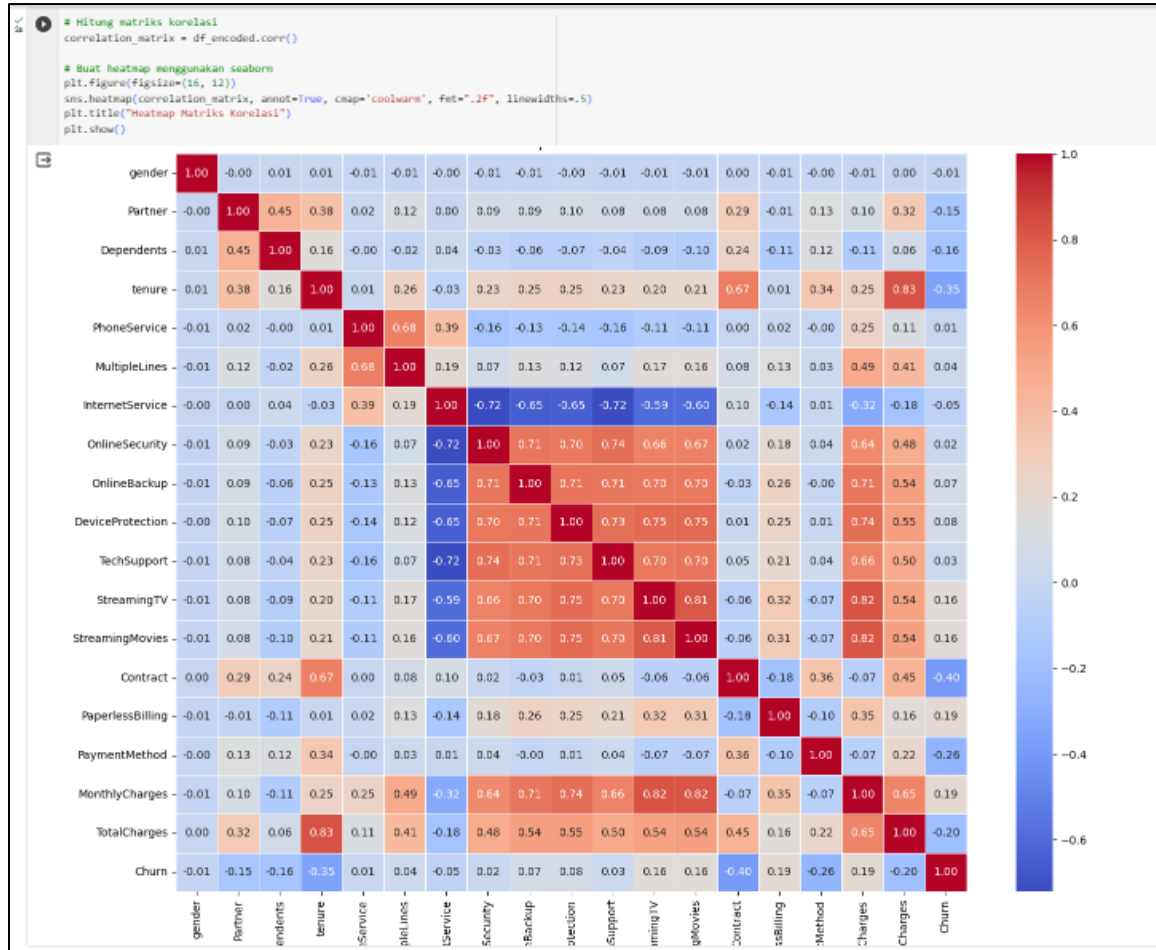
[211] kolom_numerik.append('Churn')

[212] data_numerik = data[kolom_numerik]
      data_kategori = data[kolom_kategori]

[213] data_numerik
```


Gambar 7. Screenshot code penguraian atribut kategorikal dan numerik

Setelah itu membuat sebuah heatmap dari matriks korelasi antar variabel dalam suatu dataset seperti gambar dibawah. Fungsi `pd.factorize()` digunakan untuk mengubah nilai-nilai kategorikal dalam setiap kolom menjadi bilangan bulat yang unik, dan kemudian dihitung matriks korelasi menggunakan `corr()`.



Gambar 8. Screenshot relasi antar data pada data numerik

2. ANALISIS KARAKTERISTIK DATA

Instruksi Kerja:

- Sajikan karakteristik data yang terkumpul dengan deskripsi statistik dasar
- Sajikan karakteristik data yang terkumpul dengan visualisasi grafik
- Analisis karakteristik data dari hasil penyajian data untuk telaah data

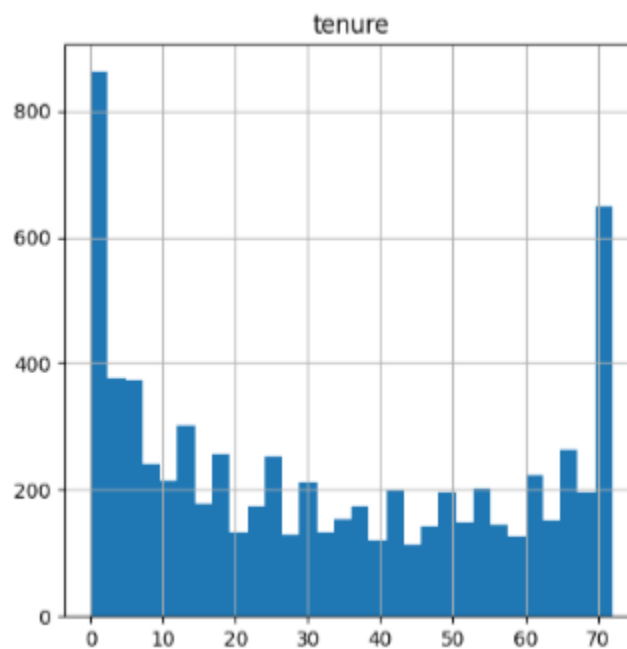
Deskripsi statistic pada dataset telco dapat dilihat pada gambar dibawah ini

```
data.describe()
```

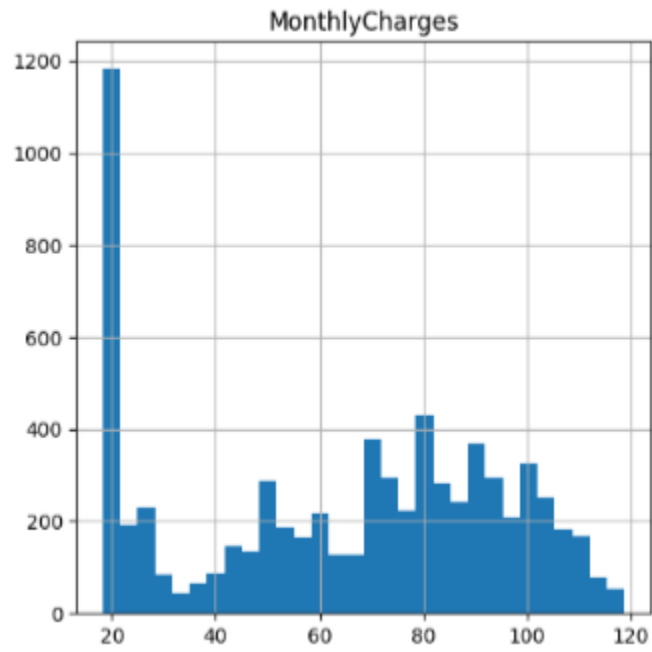
	tenure	MonthlyCharges	TotalCharges
count	7040.000000	7043.000000	7043.000000
mean	35.043892	64.761692	0.162147
std	115.282871	30.090047	0.368612
min	0.000000	18.250000	0.000000
25%	9.000000	35.500000	0.000000
50%	29.000000	70.350000	0.000000
75%	55.000000	89.850000	0.000000
max	7100.000000	118.750000	1.000000

Gambar 9. Screenshot deskripsi statistic dasar

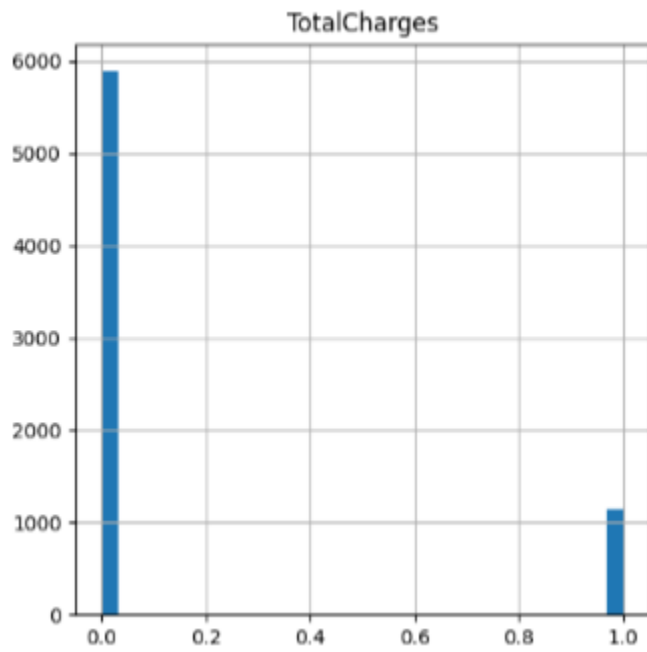
Untuk menentukan karakteristik dengan visualisasi grafik dilakukan dengan metode EDA



Gambar 10. Screenshot karakteristik tenure

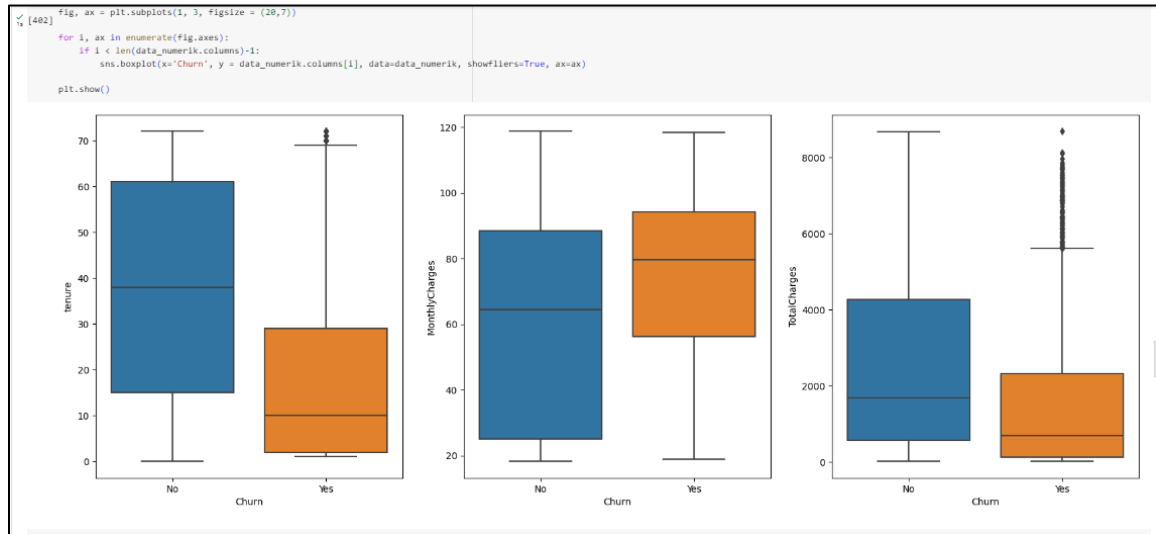


Gambar 11. Screenshot karakteristik MonthlyCharges



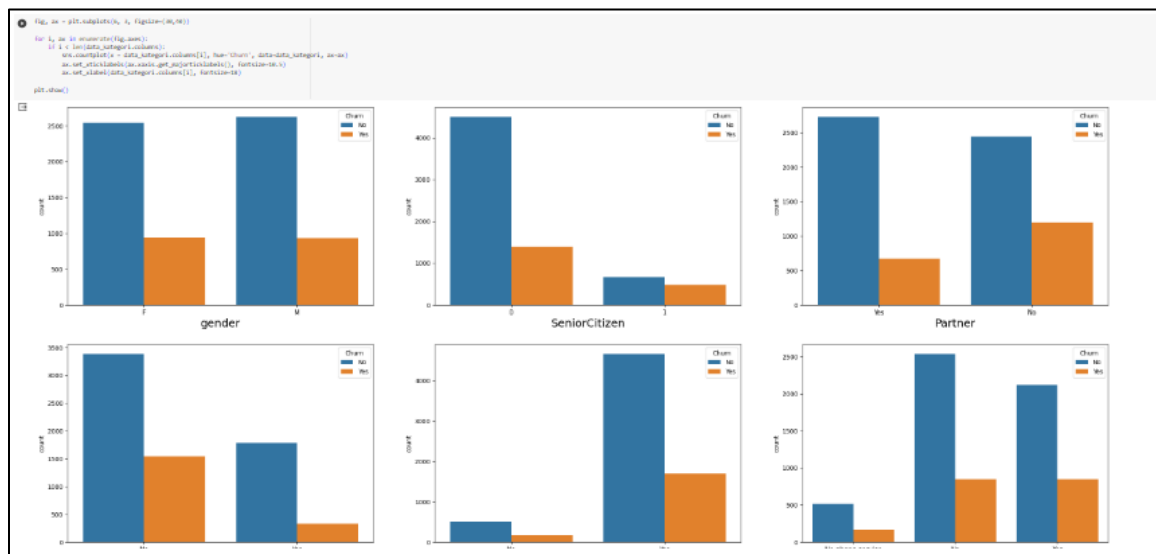
Gambar 12. Screenshot karakteristik TotalCharges

Dapat disimpulkan dari visualisasi diatas bahwa customer yang bersifat churn memiliki jangka tenor yang pendek terhadap telco, akan tetapi monthly charge nya lebih tinggi daripada customer yang tidak bersifat churn. Disini saya menggunakan box plot dan histogram untuk data numerik. Box plot berfungsi untuk memberikan ringkasan statistik distribusi, meliputi kuartil, rentang interkuartil (IQR), median, dan deteksi outliers.

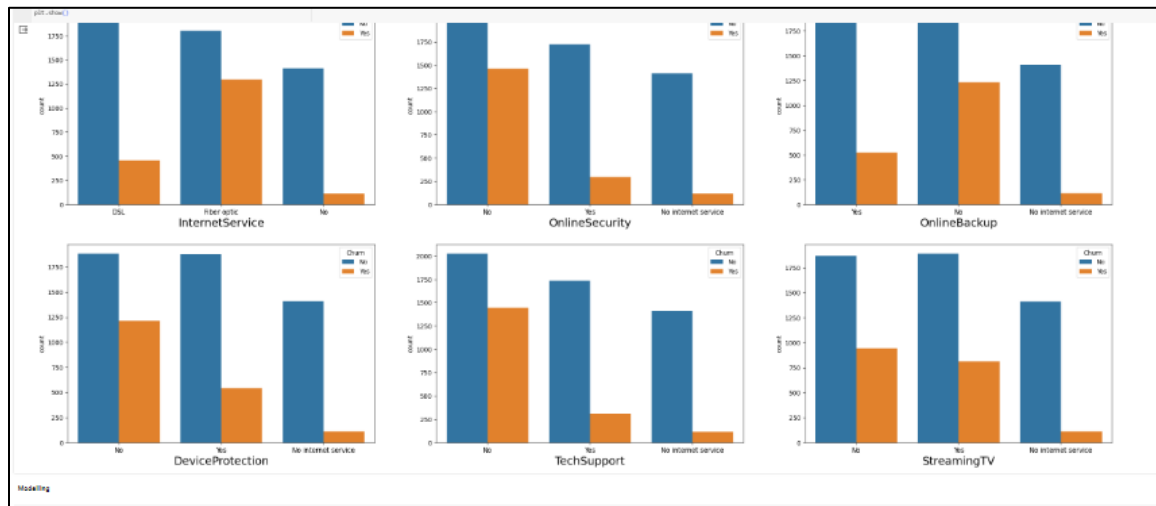


Gambar 13. Screenshot karakteristik data numerik terhadap churn

Pada kolom kategorikal visualisasi menggunakan histogram dikarenakan nilai bukan numerik untuk memberikan gambaran visual tentang bentuk keseluruhan distribusi data, termasuk bentuk distribusi, pusat, dan penyebaran nilai.



Gambar 14. Screenshot karakteristik data kategorikal terhadap churn



Gambar 15. Screenshot karakteristik data kategorikal terhadap churn

3. LAPORAN TELAAH DATA

Instruksi Kerja:

- Dokumentasikan hasil analisis dalam bentuk laporan sesuai dengan tujuan teknis
- Susun hipotesis berdasar hasil analisis sesuai tujuan teknis data science

Catatan:

- Langkah kerja ini dapat diintegrasikan dengan langkah-langkah kerja sebelumnya
- Bila pada langkah kerja (1) menganalisis tipe dan relasi data; dan (2) menganalisis karakteristik data; telah didokumentasikan dalam bentuk laporan yang memadai, maka langkah kerja (3) membuat laporan telaah data; dapat diabaikan.

Dapat disimpulkan dari langkah kerja sebelumnya customer yang bersifat churn memiliki jangka tenor yang pendek terhadap telco, akan tetapi monthly charge nya lebih tinggi daripada customer yang tidak bersifat churn.

BUKTI 3-ADS

Kode Unit	:	J.62DMI00.006.1
Judul Unit	:	Memvalidasi Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memvalidasi data untuk data science.

Langkah Kerja:

- 1) Melakukan pengecekan kelengkapan data
- 2) Membuat rekomendasi kelengkapan data

Peralatan dan Perlengkapan:

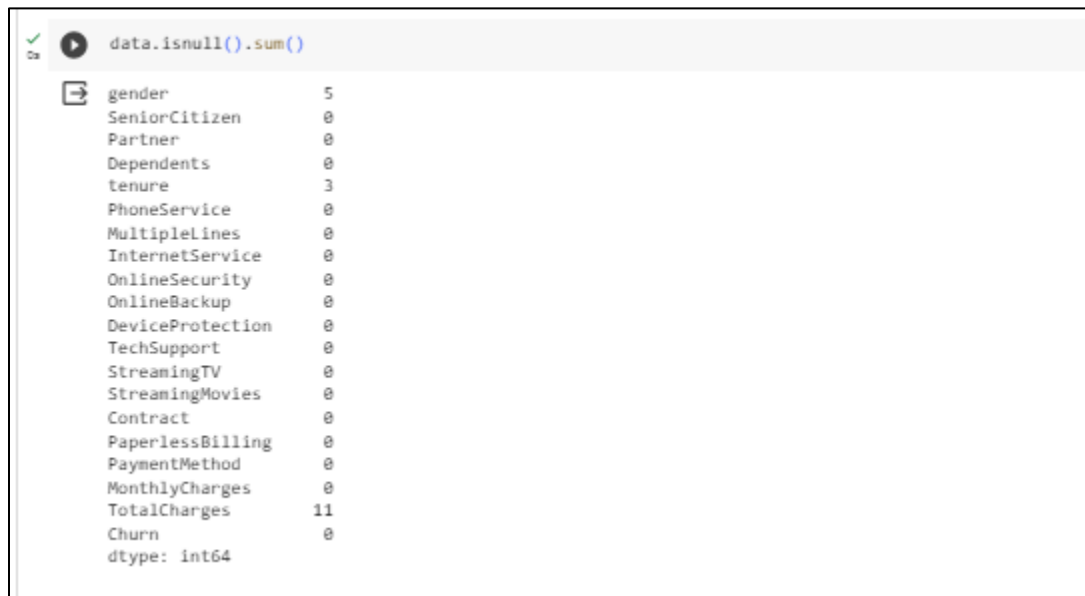
- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengubah teks

1. PENGECEKAN KELENGKAPAN DATA

Instruksi Kerja:

- Sajikan penilaian kualitas data dari hasil telaah sesuai tujuan teknis data science
- Sajikan penilaian tingkat kecukupan data dari hasil telaah sesuai tujuan teknis data science

Untuk melakukan pengecekan kelengkapan data dapat menggunakan kode seperti gambar dibawah. Terlihat pada dataset memiliki missing value yang mempengaruhi kualitas dari data tersebut.



```
data.isnull().sum()
```

gender	5
SeniorCitizen	0
Partner	0
Dependents	0
tenure	3
PhoneService	0
MultipleLines	0
InternetService	0
OnlineSecurity	0
OnlineBackup	0
DeviceProtection	0
TechSupport	0
StreamingTV	0
StreamingMovies	0
Contract	0
PaperlessBilling	0
PaymentMethod	0
MonthlyCharges	0
TotalCharges	11
Churn	0
dtype: int64	

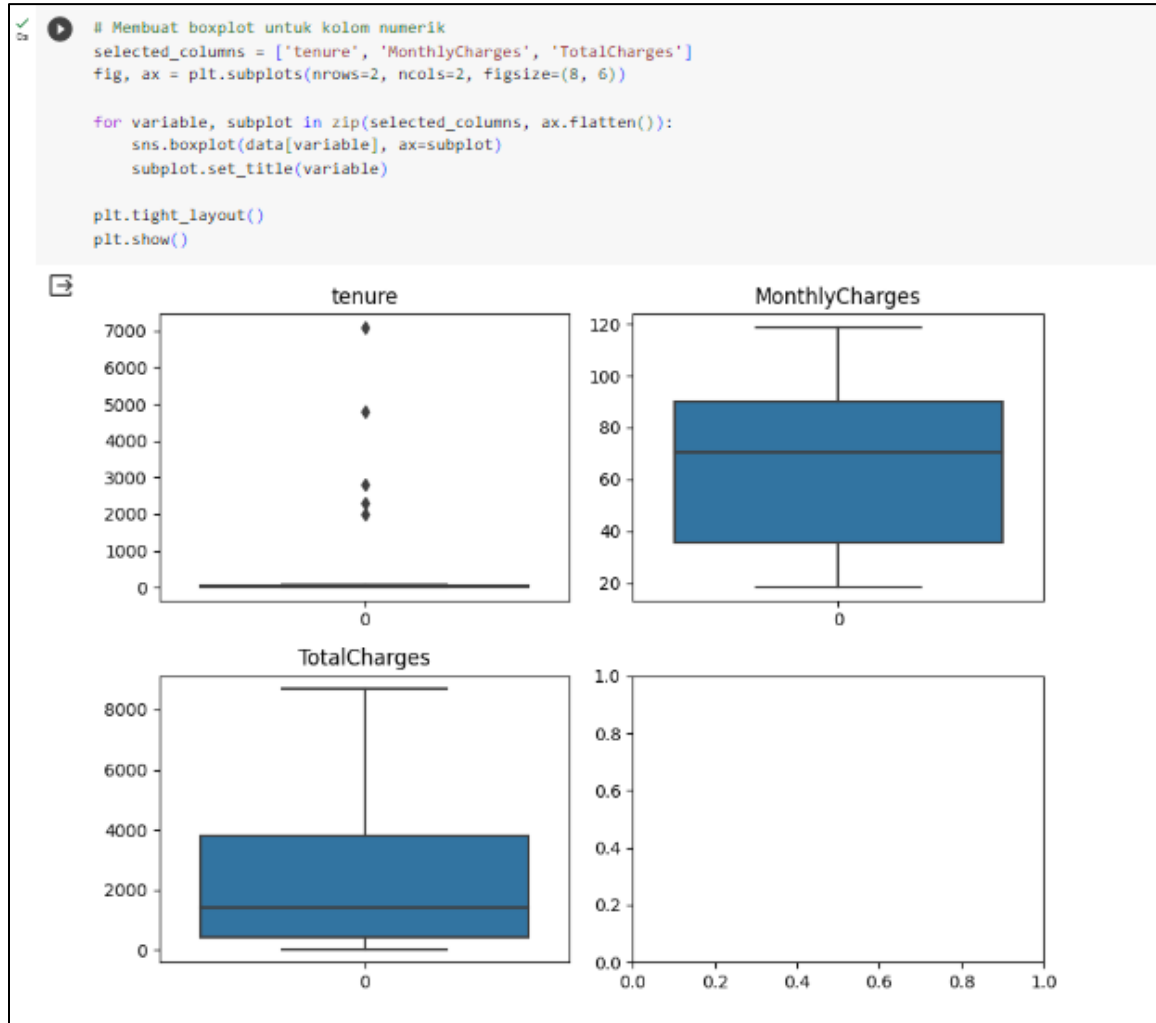
Gambar 16. Screenshot code pengecekan kelengkapan data

Dengan hanya 5 nilai yang hilang dari 7000 lebih records yang ada pada kolom gender, karena bersifat kategorikal, dapat menggantikan nilai yang hilang dengan nilai yang paling sering muncul (mode) dalam kolom tersebut. Ini bisa menjadi pendekatan yang cukup sederhana dan efektif untuk dataset.

Untuk rekomendasi pada kolom tenure dan totalcharges, dapat mengisi nilai yang kosong tersebut dengan median dikarenakan median bersifat lebih tahan pada outliers.

Setelah itu saya melakukan pengecekan outlier pada dataset menggunakan box plot. Hasil pengecekan terdapat beberapa outliers pada kolom tersebut. Dikarenakan jumlah outliers hanya 5, saya memutuskan

untuk menghapus 5 record tersebut.



Gambar 17. Screenshot code pengecekan outliers pada kolom tenure

Sejauh ini kualitas dapat terbilang cukup bagus karena missing value dan outliers sangat sedikit jumlahnya tetapi masih bisa mempengaruhi kualitas perhitungan statistik dan prediksi machine learning

2. REKOMENDASI KELENGKAPAN DATA

Instruksi Kerja:

- Susun rekomendasi hasil penilaian kualitas sesuai tujuan teknis data science
- Susun rekomendasi hasil penilaian kecukupan data sesuai tujuan teknis data science

Kualitas data sudah cukup bagus akan tetapi kita perlu mengevaluasi teknik pengumpulan data agar data yang diterima tidak terdapat kesalahan untuk mempertahankan kualitas data. Saya merekomendasikan agar pengumpulan data terus dilakukan karena semakin banyak data pelatihan maka algoritma machine learning semakin baik.

BUKTI 4-ADS

Kode Unit	:	J.62DMI00.007.1
Judul Unit	:	Menentukan Objek Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memilah dan memilih data yang sesuai permintaan atau kebutuhan.

Langkah Kerja:

- 1) Memutuskan kriteria dan teknik pemilihan data
- 2) Menentukan attributes (columns) dan records (row) data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Aplikasi spreadsheet
 - Aplikasi notepad plus
 - Aplikasi SQL (Structured Query Language)

1. KRITERIA DAN TEKNIK PEMILIHAN DATA

Instruksi Kerja:

- Identifikasi kriteria pemilihan data sesuai dengan tujuan teknis dan aturan yang berlaku
- Tetapkan teknik pemilihan data sesuai dengan kriteria pemilihan data

Teknik pemilihan data yang digunakan pada dataset ini adalah memasukkan semua kolom yang tidak bersifat kategorikal dan memasukkan semua data yang bersifat kategorikal. Setelah modelling selesai kita dapat me ngedrop salah satu variable yang tidak memiliki relasi kuat terhadap kolom churn melalui feature importance

2. ATTRIBUTES (COLUMNS) DAN RECORDS (ROW) DATA

Instruksi Kerja:

- Identifikasi attributes (columns) data sesuai dengan kriteria pemilihan data
- Identifikasi records (row) data sesuai dengan kriteria pemilihan data

Atribut yang akan dipakai adalah semuanya kecuali kolom customerID, dan semua record termasuk missing value juga digunakan karena telah diganti dengan nilai statistic. Saya juga memutuskan untuk tidak menggunakan outliers. Outliers adalah nilai yang secara signifikan berbeda dari sebagian besar data dan dapat mempengaruhi hasil analisis statistic sehingga tidak saya gunakan.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 20 columns):
#   Column                Non-Null Count  Dtype
---  -
0   gender                 7038 non-null   object
1   SeniorCitizen          7043 non-null   int64
2   Partner                7043 non-null   object
3   Dependents             7043 non-null   object
4   tenure                 7040 non-null   float64
5   PhoneService           7043 non-null   object
6   MultipleLines          7043 non-null   object
7   InternetService        7043 non-null   object
8   OnlineSecurity         7043 non-null   object
9   OnlineBackup           7043 non-null   object
10  DeviceProtection       7043 non-null   object
11  TechSupport            7043 non-null   object
12  StreamingTV            7043 non-null   object
13  StreamingMovies        7043 non-null   object
14  Contract               7043 non-null   object
15  PaperlessBilling       7043 non-null   object
16  PaymentMethod          7043 non-null   object
17  MonthlyCharges         7043 non-null   float64
18  TotalCharges           7043 non-null   object
19  Churn                  7043 non-null   object
dtypes: float64(2), int64(1), object(17)
memory usage: 1.1+ MB

[256] data['SeniorCitizen'] = data['SeniorCitizen'].astype('object')
      data['TotalCharges'] = data['SeniorCitizen'].astype('float64')
```

Gambar 18. Screenshot identifikasi kolom yang akan di encode

BUKTI 5-ADS

Kode Unit	:	J.62DMI00.008.1
Judul Unit	:	Membersihkan Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membersihkan data yang sesuai permintaan atau kebutuhan.

Langkah Kerja:

- 1) Melakukan pembersihan data yang kotor
- 2) Membuat laporan dan rekomendasi hasil membersihkan data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Aplikasi spreadsheet
 - Aplikasi text editor
 - Aplikasi SQL (Structured Query Language)

1. PEMBERSIHAN DATA KOTOR

Instruksi Kerja:

- Tentukan strategi pembersihan data berdasarkan hasil telaah data
- Koreksi data yang kotor berdasarkan strategi pembersihan data

Strategi pembersihan data dilakukan dengan mengisi modus pada kolom yang bersifat kategorikal atau non numerik dan mengisi dengan median untuk data numerik karena median bersifat tahan pada outliers

```
data.isnull().sum()

gender          5
SeniorCitizen   0
Partner         0
Dependents      0
tenure          3
PhoneService    0
MultipleLines   0
InternetService 0
OnlineSecurity  0
OnlineBackup    0
DeviceProtection 0
TechSupport     0
StreamingTV     0
StreamingMovies 0
Contract        0
PaperlessBilling 0
PaymentMethod   0
MonthlyCharges  0
TotalCharges    11
Churn           0
dtype: int64

data['gender'] = data['gender'].str.replace('Female', 'F')
data['gender'] = data['gender'].str.replace('Male', 'M')

[286] data['gender'].fillna(data['gender'].mode()[0], inplace=True)
```

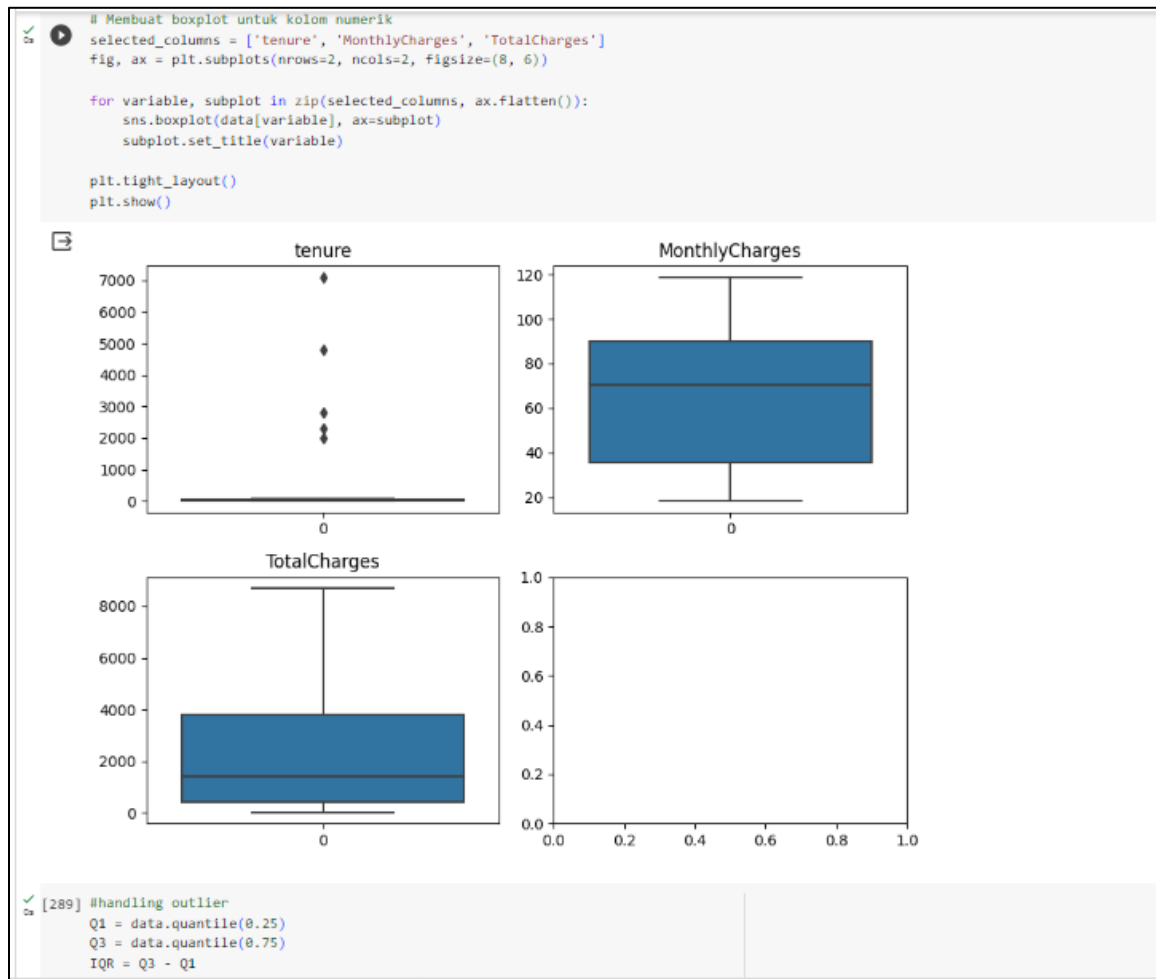
Gambar 19. Screenshot melakukan cleaning pada gender

```
data['tenure'].fillna(data['tenure'].median(), inplace=True)

<ipython-input-264-4cfa3bdb40f5>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy
data['tenure'].fillna(data['tenure'].median(), inplace=True)
```

Gambar 20. Screenshot code mengatasi missing value pada tenure



Gambar 21. Screenshot code mengatasi outlier pada dataset

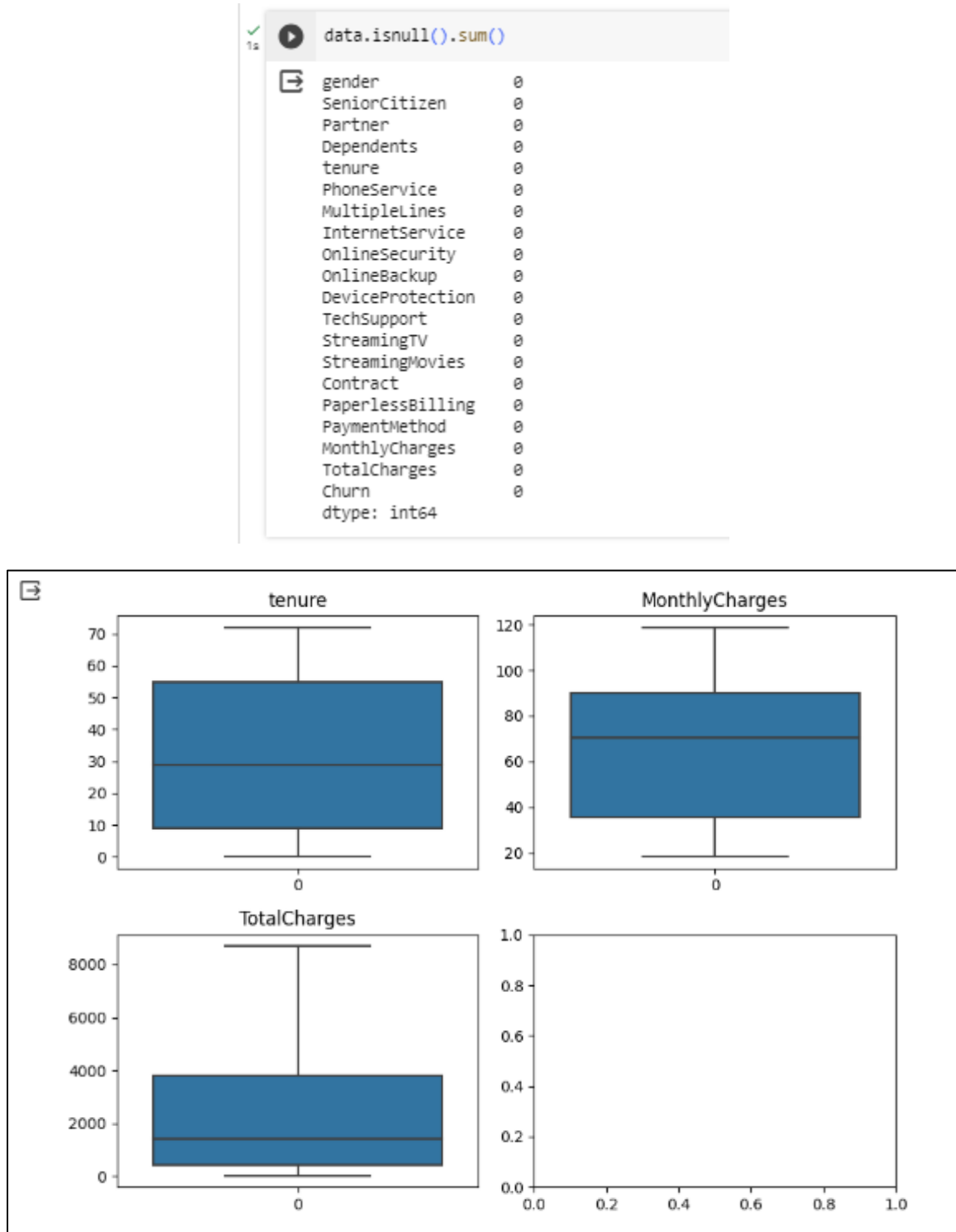
Karena outliers hanya berjumlah sedikit maka dilakukan penghapusan record pada dataset yang mengalami outliers pada dataset

2. LAPORAN DAN REKOMENDASI HASIL PEMBERSIHAN DATA KOTOR

Instruksi Kerja:

- Deskripsikan masalah dan teknis koreksi data sesuai dengan kondisi data dan strategi pembersihan data
- Lakukan evaluasi berdasarkan analisis koreksi yang telah dilakukan
- Dokumentasikan evaluasi proses dan hasil pembersihan data kotor

Hasil dokumentasi setelah melakukan perbaikan pada missing value dan outlier



Gambar 22. Screenshot setelah mengatasi missing value dan outliers

BUKTI 6-ADS

Kode Unit	:	J.62DMI00.009.1
Judul Unit	:	Mengkonstruksi Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengkonstruksi data untuk proyek data science.

Langkah Kerja:

- 1) Menganalisis teknik transformasi data
- 2) Melakukan transformasi data
- 3) Membuat dokumentasi konstruksi data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Tools pengolah kata

1. ANALISIS TEKNIK TRANSFORMASI DATA

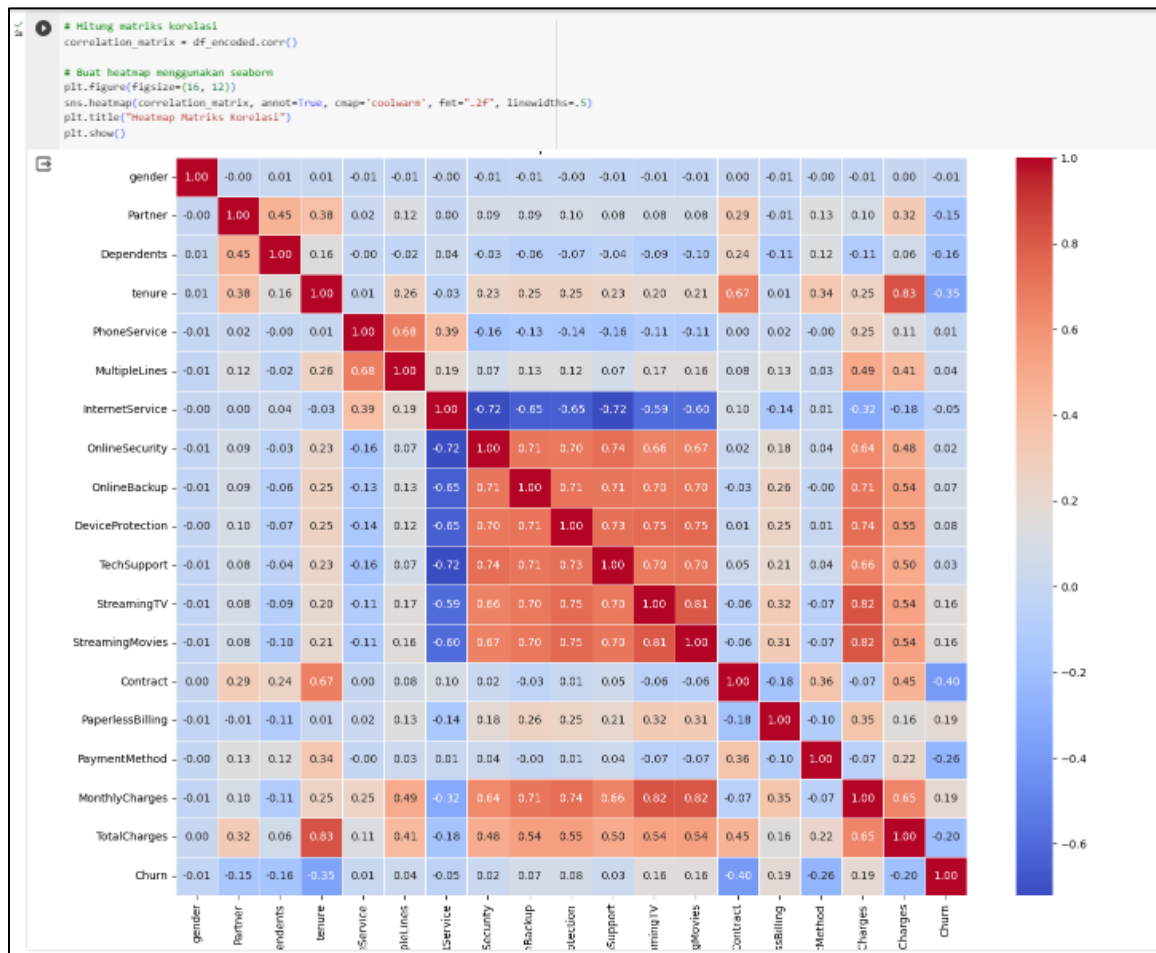
Instruksi Kerja:

- Lakukan analisis data untuk menentukan representasi fitur data awal
- Lakukan analisis representasi fitur data awal untuk menentukan teknik rekayasa fitur yang diperlukan untuk pembangunan model data science

Dikarenakan data memiliki atribut kategorikal dan numerik, jadi diperlukan adanya transformasi data untuk pengecekan relasi antar atribut.

```
[372] df_encoded = data.copy()

df_encoded['gender'] = data['gender'].map({'M': 1, 'F': 0})
df_encoded['SeniorCitizen'] = data['SeniorCitizen']
df_encoded['Partner'] = data['Partner'].map({'No': 0, 'Yes': 1})
df_encoded['Dependents'] = data['Dependents'].map({'No': 0, 'Yes': 1})
df_encoded['PhoneService'] = data['PhoneService'].map({'No': 0, 'Yes': 1})
df_encoded['MultipleLines'] = data['MultipleLines'].map({'No phone service': 0, 'No': 1, 'Yes': 2})
df_encoded['InternetService'] = data['InternetService'].map({'DSL': 0, 'Fiber optic': 1, 'No': 2})
df_encoded['OnlineSecurity'] = data['OnlineSecurity'].map({'No internet service': 0, 'No': 1, 'Yes': 2})
df_encoded['OnlineBackup'] = data['OnlineBackup'].map({'No internet service': 0, 'No': 1, 'Yes': 2})
df_encoded['DeviceProtection'] = data['DeviceProtection'].map({'No internet service': 0, 'No': 1, 'Yes': 2})
df_encoded['TechSupport'] = data['TechSupport'].map({'No internet service': 0, 'No': 1, 'Yes': 2})
df_encoded['StreamingTV'] = data['StreamingTV'].map({'No internet service': 0, 'No': 1, 'Yes': 2})
df_encoded['StreamingMovies'] = data['StreamingMovies'].map({'No internet service': 0, 'No': 1, 'Yes': 2})
df_encoded['Contract'] = data['Contract'].map({'Month-to-month': 0, 'One year': 1, 'Two year': 2})
df_encoded['PaperlessBilling'] = data['PaperlessBilling'].map({'No': 0, 'Yes': 1})
df_encoded['PaymentMethod'] = data['PaymentMethod'].map({'Electronic check': 0, 'Mailed check': 1, 'Bank transfer (automatic)': 2, 'Credit card (automatic)': 3})
df_encoded['Churn'] = data['Churn'].map({'No': 0, 'Yes': 1})
```



Gambar 23. Screenshot code transformasi data

Untuk persiapan modelling saya memutuskan untuk menggunakan teknik one hot encoding pada setiap atribut yang bersifat kategorikal. Atribut tersebut diantaranya :

- **gender**: Male, Female
- **InternetService**: DSL, Fiber optic, No
- **OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies**: Yes, No, No internet service
- **Contract**: Month-to-month, One year, Two year
- **PaperlessBilling**: Yes, No
- **PaymentMethod**: Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)
- **SeniorCitizen**: 1, 0

- Partner, Dependents, PhoneService: Yes, No
- **Churn:** Yes, No (Target variable)

```
data_kategori_encoded = data_kategori_encoded.rename(columns={'Churn_Yes': 'Churn'})
data_kategori_encoded.head()
```

	gender_M	SeniorCitizen_1	Partner_Yes	Dependents_Yes	PhoneService_Yes	MultipleLines_No phone service	MultipleLines_Yes	InternetService_Fiber optic	InternetService_No	OnlineSecurity_No Internet service	...	StreamingTV_Yes	StreamingMovies Internet serv
0	0	0	1	0	0	1	0	0	0	0	...	0	0
1	1	0	0	0	1	0	0	0	0	0	...	0	0
2	1	0	0	0	1	0	0	0	0	0	...	0	0
3	1	0	0	0	0	1	0	0	0	0	...	0	0
4	0	0	0	0	1	0	0	1	0	0	...	0	0

5 rows x 28 columns

Gambar 24. Screenshot code melakukan one hot encoding

2. TRANSFORMASI DATA

Instruksi Kerja:

- Lakukan transformasi untuk mendapatkan fitur data awal
- Lakukan rekayasa fitur data untuk mendapatkan fitur baru yang diperlukan untuk pembangunan model data science

Disini saya membuat dataframe baru yang menggabungkan data kategorikal yang sudah di encode dengan data yang bersifat numerik. Langkah ini merupakan tahap awal untuk melakukan modelling

```
#gabungkan data
new_data = pd.concat([data_numerik, data_kategori_encoded], axis=1)
```

Gambar 25. Screenshot code melakukan penggabungan data

3. DOKUMENTASI KONSTRUKSI DATA

Instruksi Kerja:

- Jabarkan teknis transformasi data dalam bentuk tertulis
- Tuangkan hasil transformasi data dan rekomendasi hasil transformasi dalam bentuk tertulis

Catatan:

- Langkah kerja ini dapat diintegrasikan dengan langkah-langkah kerja sebelumnya

- Bila pada langkah kerja (1) menganalisis teknik transformasi data; dan (2) melakukan transformasi data; telah didokumentasikan dalam bentuk laporan yang memadai, maka langkah kerja (3) membuat dokumentasi konstruksi data; dapat diabaikan.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

BUKTI 7-ADS

Kode Unit	:	J.62DMI00.010.1
Judul Unit	:	Menentukan Label Data

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam menentukan label data untuk pembangunan model data science.

Langkah Kerja:

- 1) Melakukan pelabelan data
- 2) Membuat laporan hasil pelabelan data

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Aplikasi pengolah kata
 - Aplikasi pelabelan data

1. PELABELAN DATA

Instruksi Kerja:

- Uraikan kesesuaian antara analisis hasil pelabelan data sejenis yang sudah ada dengan Standard Operating Procedure (SOP) pelabelan
- Lakukan pelabelan data sesuai dengan SOP pelabelan

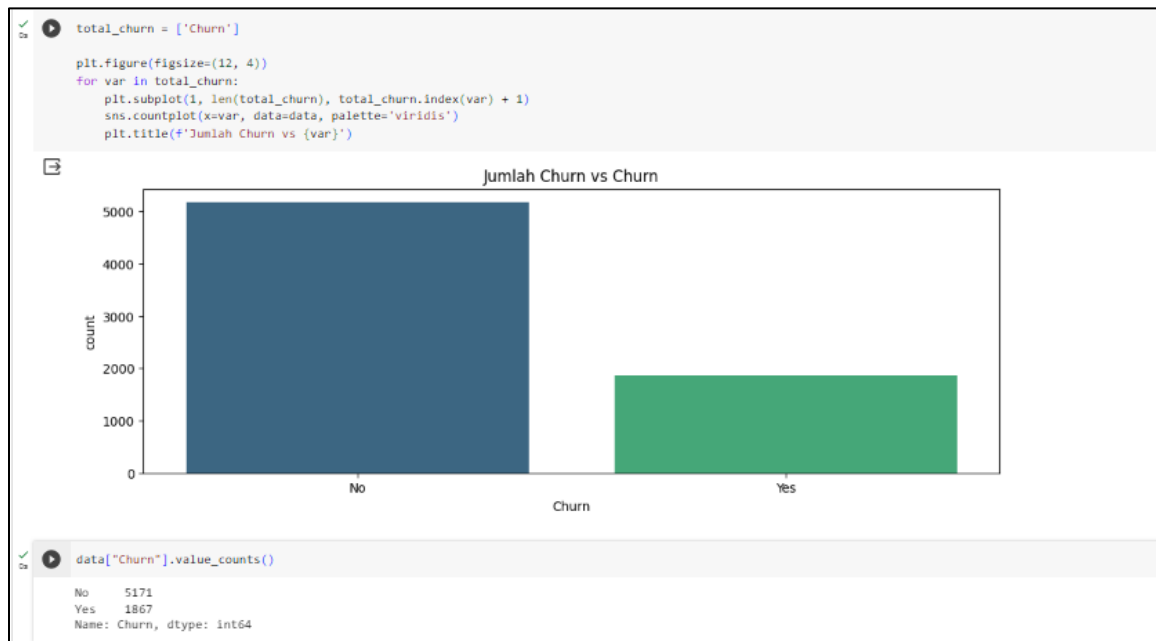
Pada kasus dataset ini variable target (churn) telah dilabeli. Hasil perbandingan yang telah dilabeli churn atau tidak dapat dilihat pada gambar berikut ini :

2. LAPORAN HASIL PELABELAN DATA

Instruksi Kerja:

- Uraikan statistik hasil pelabelan pada laporan
- Uraikan evaluasi proses pelabelan pada laporan

. Hasil perbandingan yang telah dilabeli churn atau tidak dapat dilihat pada gambar berikut ini, customer yang dilabeli churn berjumlah 1867 sedangkan yang tidak churn berjumlah 5171



Gambar 26. Screenshot hasil pelabelan churn dan tidak

BUKTI 8-ADS

Kode Unit	:	J.62DMI00.013.1
Judul Unit	:	Membangun Model

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam membangun model.

Langkah Kerja:

- 1) Menyiapkan parameter model
- 2) Menggunakan tools pemodelan

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer dan peralatannya
 - Perangkat lunak data science di antaranya: rapid miner, weka, atau development untuk bahasa pemrograman tertentu seperti Python atau R.
- Perlengkapan
 - Dokumen best practices kriteria dan evaluasi penilaian

1. PARAMETER MODEL

Instruksi Kerja:

- Identifikasi parameter-parameter yang sesuai dengan model
- Tetapkan nilai toleransi parameter evaluasi pengujian sesuai dengan tujuan teknis

Pada kasus ini saya menggunakan decisiontree tanpa menyediakan nilai spesifik untuk setiap parameter, maka otomatis akan menggunakan nilai default dari setiap parameter.

Beberapa parameter default utama yang penting pada Decision Tree adalah:

1. **criterion{"gini", "entropy", "log_loss"}, default="gini"**
Menentukan fungsi untuk mengukur kualitas split. 'Gini' mengukur sejauh mana pembersihan (impurity) terjadi di dalam node.
2. **splitter: 'best'**
Parameter ini menentukan strategi untuk memilih pembagi di setiap node. 'Best' berarti pemilihan pembagi didasarkan pada kriteria terbaik.
3. **max_depth: None**
Kedalaman maksimum dari pohon. None berarti tidak ada batasan kedalaman, dan pohon akan tumbuh hingga tidak ada lagi sampel yang dapat dibagi atau sampel mencapai jumlah minimum untuk daun.
4. **min_samples_split: 2**
Jumlah minimum sampel yang diperlukan untuk membagi node internal.
5. **min_samples_leaf :1**

- Menentukan jumlah sampel minimum yang dibutuhkan dalam leaf node.
6. **max_features :None**
Menentukan jumlah maksimum fitur yang akan dipertimbangkan untuk splitting di setiap node.
 7. **random_state :None**
Menentukan seed untuk mengontrol randomness dan membuat hasil reproducible.

Dan juga menggunakan LogisticRegression tanpa menyediakan nilai untuk parameter-parameter tertentu, maka scikit-learn akan menggunakan nilai default yang telah ditentukan. Parameter yang umum digunakan pada logistic regression sebagai berikut :

1. **penalty (default='l2'):**
Menentukan jenis regulasi yang akan diterapkan. 'l2' menggunakan regulasi L2 (Ridge), 'l1' menggunakan regulasi L1 (Lasso), dan 'none' tidak menggunakan regulasi
2. **Cfloat, default=1.0**
Parameter ini mengontrol kekuatan regulasi. Semakin kecil nilai C, semakin kuat regulasinya.
3. **solver (default='lbfgs'):**
Menentukan algoritma solver yang digunakan untuk menyelesaikan masalah optimasi. 'lbfgs' adalah solver yang umum digunakan untuk masalah multikelas. Dalam contoh Anda, nilai default 'lbfgs' digunakan.
4. **max_iter (default=100):**
Menentukan jumlah maksimum iterasi untuk solver konvergensi. Nilai ini dapat diatur sesuai kebutuhan, terutama jika konvergensi belum tercapai dalam jumlah iterasi default
5. **class_weight (default=None):**
Memberikan bobot kelas yang berbeda. Nilai None berarti semua kelas memiliki bobot yang sama.
6. **random_state (default=None):**
Menentukan seed untuk mengontrol randomness dan membuat hasil reproducible

2. TOOLS PEMODELAN

Instruksi Kerja:

- Identifikasi tools untuk membuat model sesuai dengan tujuan teknis data science
- Bangun algoritma untuk teknik pemodelan yang ditentukan menggunakan tools yang dipilih
- Eksekusi algoritma pemodelan sesuai dengan skenario pengujian dan tools untuk membuat model yang telah ditetapkan
- Optimasi parameter model algoritma untuk menghasilkan nilai parameter evaluasi yang sesuai dengan skenario pengujian

Tools pemodelan yang dipakai pada dataset ini adalah decision tree dan logistic regression. Berikut gambar code untuk melakukan modelling

```
✓ [225] X = new_data.drop(columns=['Churn'], axis=1)
0s      y = new_data['Churn']

✓ [226] #split into train and test set
0s      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)

✓ [228] # Decision Tree
0s      dt = DecisionTreeClassifier()
      dt.fit(X_train, y_train)
      dt_pred = dt.predict(X_test)
```

Gambar 27. Screenshot code decision tree

```
✓ [245] #Logistic Regression
0s      log_reg = LogisticRegression()
      pipe = make_pipeline(StandardScaler(), LogisticRegression())
      pipe.fit(X_train, y_train)
      y_pred_log = pipe.predict(X_test)
      acc_log_reg = accuracy_score(y_test, y_pred_log)
```

Gambar 28. Screenshot code logistic regression

BUKTI 9-ADS

Kode Unit	:	J.62DMI00.014.1
Judul Unit	:	Mengevaluasi Hasil Pemodelan

Deskripsi:

Bukti ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam mengevaluasi hasil pemodelan.

Langkah Kerja:

- 1) Menggunakan model dengan data riil
- 2) Menilai hasil pemodelan

Peralatan dan Perlengkapan:

- Peralatan
 - Komputer
- Perlengkapan
 - Tools untuk mengeksekusi model
 - Tools untuk pengumpulan data riil

1. PENGGUNAAN MODEL DENGAN DATA RIIL

Instruksi Kerja:

- Kumpulkan data baru untuk evaluasi pemodelan sesuai kebutuhan yang mengacu kepada parameter evaluasi
- Uji model dengan menggunakan data riil yang telah dikumpulkan

Pada model ini kita tidak bisa menggunakan data baru kepada model, kita hanya menggunakan data lama yang sudah ada

2. PENILAIAN HASIL PEMODELAN

Instruksi Kerja:

- Nilai keluaran pengujian model berdasarkan metrik kesuksesan
- Dokumentasikan hasil penilaian sesuai standar yang berlaku

Berikut adalah hasil evaluasi dari model decision tree. berdasarkan metrik kesuksesan

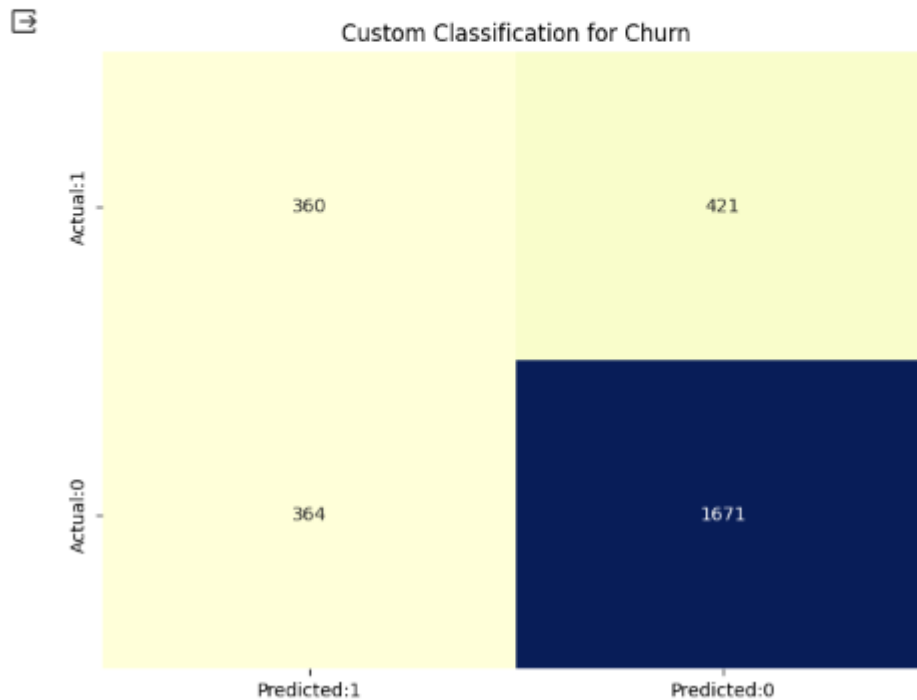

```

cm_dt = confusion_matrix(y_test, dt_pred)

conf_matrix_dt_custom = pd.DataFrame(data=[[cm_dt[1, 1], cm_dt[0, 1]],
                                           [cm_dt[1, 0], cm_dt[0, 0]]],
                                     columns=['Predicted:1', 'Predicted:0'],
                                     index=['Actual:1', 'Actual:0'])

plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix_dt_custom, annot=True, fmt='d', cmap="YlGnBu", cbar=False)
plt.title('Custom Classification for Churn')
plt.show()

```



```

[229] # Evaluate decision tree
accuracy = accuracy_score(y_test, dt_pred)
precision = precision_score(y_test, dt_pred)
recall = recall_score(y_test, dt_pred)
f1 = f1_score(y_test, dt_pred)

print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1}")

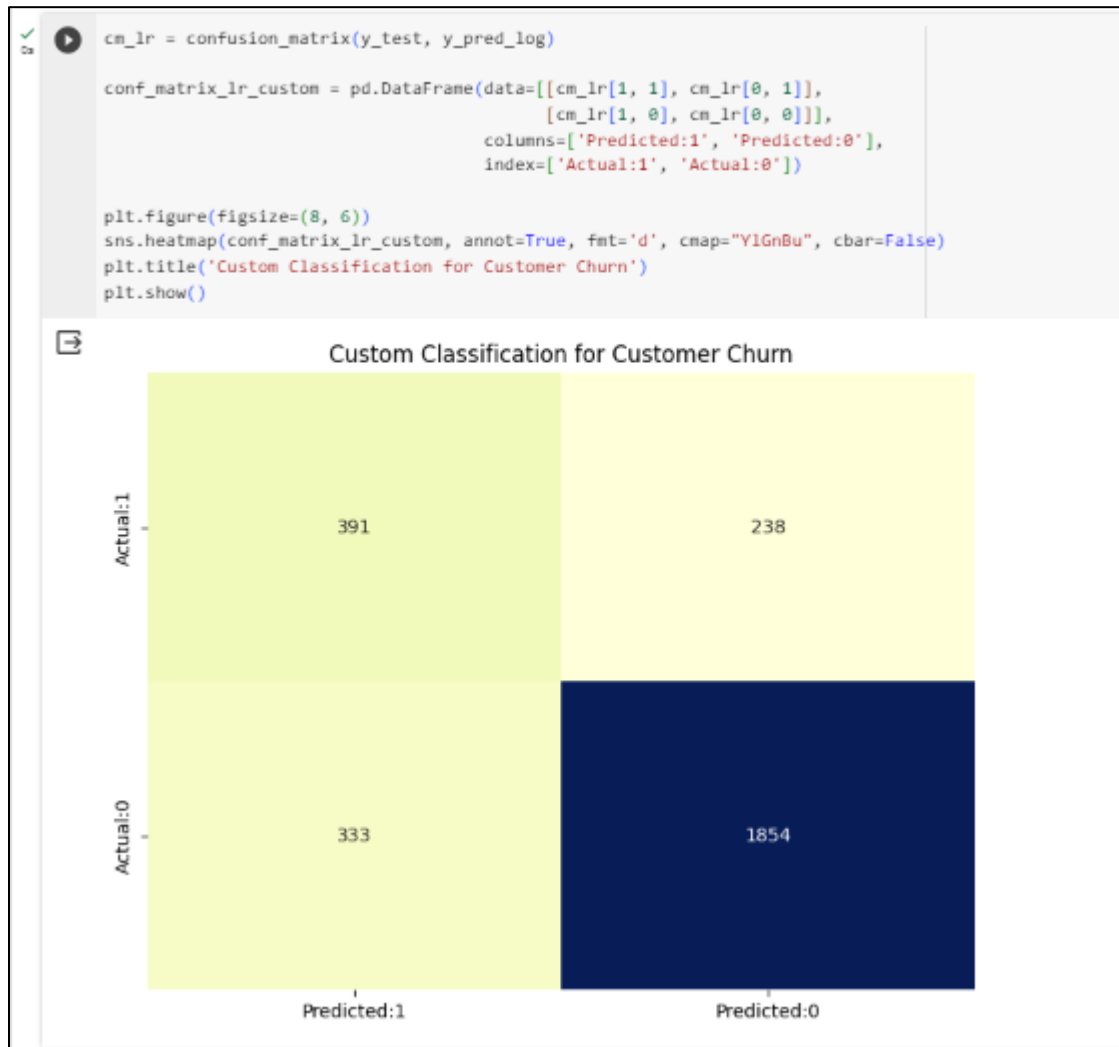
```

```

Accuracy: 0.7123579545454546
Precision: 0.4457070707070707
Recall: 0.48756906077348067
F1 Score: 0.46569920844327184

```

Dan berikut ini adalah hasil evaluasi dari logistic regression



```
[246] # Evaluate the model
accuracy = accuracy_score(y_test, y_pred_log)
precision = precision_score(y_test, y_pred_log)
recall = recall_score(y_test, y_pred_log)
f1 = f1_score(y_test, y_pred_log)

print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1}")

Accuracy: 0.796875
Precision: 0.6206349206349207
Recall: 0.5400552486187845
F1 Score: 0.5775480059084196
```

Model logistic regression memiliki kinerja lebih bagus dengan accuracy 0.79%, karena pada proyek ini bertujuan untuk memprediksi customer churn maka nilai False Positive perlu diperhatikan. Dilihat dari perbandingan presisi dari kedua model, logistic regression lebih unggul dari segi presisi.

```

> coefs = pipe.named_steps['logisticregression'].coef_[0]

# Buat DataFrame untuk menampilkan hasil
feature_importance_df = pd.DataFrame({
    'Feature': X.columns,
    'Coefficient': coefs
})

feature_importance_df = feature_importance_df.reindex(feature_importance_df['Coefficient'].abs().sort_values(ascending=False).index)

feature_importance_df

```

```
coefs = pipe.named_steps['logisticregression'].coef_[0]

# Buat DataFrame untuk menampilkan hasil
feature_importance_df = pd.DataFrame({
    'Feature': X.columns,
    'Coefficient': coefs
})

feature_importance_df = feature_importance_df.reindex(feature_importance_df['Coefficient'].abs().sort_values(ascending=False).index)
feature_importance_df
```

	Feature	Coefficient
0	tenure	-1.343814
25	Contract_Two year	-0.677094
2	TotalCharges	0.573809
10	InternetService_Fiber optic	0.423772
24	Contract_One year	-0.333803
13	OnlineSecurity_Yes	-0.173202
28	PaymentMethod_Electronic check	0.168033
21	StreamingTV_Yes	0.152106
9	MultipleLines_Yes	0.130567
1	MonthlyCharges	-0.126732
23	StreamingMovies_Yes	0.122945
19	TechSupport_Yes	-0.113836
26	PaperlessBilling_Yes	0.098425
6	Dependents_Yes	-0.090023
15	OnlineBackup_Yes	-0.077729
4	SeniorCitizen_1	0.076894
7	PhoneService_Yes	-0.069115
8	MultipleLines_No phone service	0.069115
3	gender_M	-0.063728
16	DeviceProtection_No internet service	-0.063046

Terlihat pada table feature importance pada logistic regression diatas bahwa atribut yang paling berpengaruh adalah tenure, semakin pendek jangka tenur maka semakin tinggi customer churn, ini juga didukung dengan attrribut contract_twoyear dimana pelanggan yang churn memiliki karakteristik contract dibawah 2 tahun dengan totalcharges tinggi dan internet service bertipe fiber optic