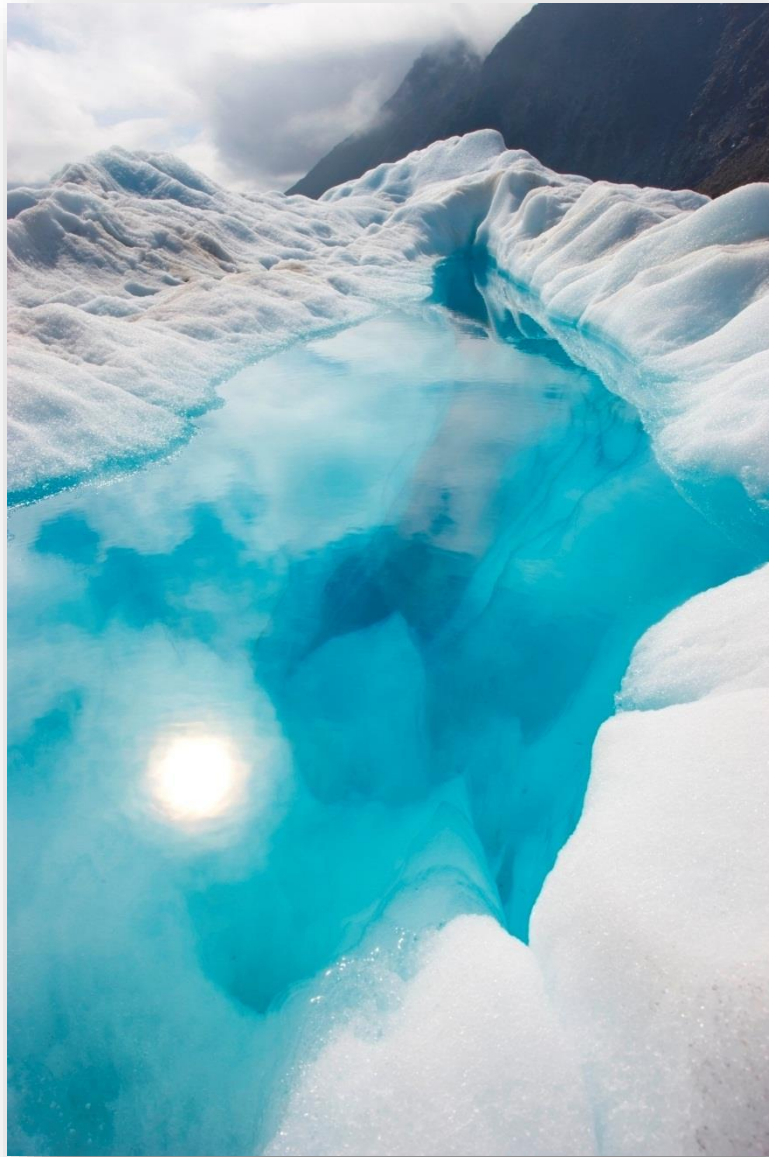


2023



BUKU KERJA/JOB SHEET

ASSOCIATE DATA SCIENTIST

Nama Peserta	:	Heldha Ayu Setia
Nomor Urut	:	

DAFTAR ISI

DAFTAR ISI	1
BUKTI 1-ADS	3
1. Kebutuhan Data	3
2. Pengambilan Data.....	4
3. Pengintegrasian Data	6
BUKTI 2-ADS	7
1. Analisis Tipe dan Relasi Data	7
2. Analisis Karakteristik Data.....	8
3. Laporan Telaah Data	10
BUKTI 3-ADS	11
1. Pengecekan Kelengkapan Data	11
2. Rekomendasi Kelengkapan Data.....	13
BUKTI 4-ADS	14
1. Kriteria dan Teknik Pemilihan Data	14
2. Attributes (Columns) dan Records (Row) Data.....	15
BUKTI 5-ADS	17
1. Pembersihan Data Kotor	17
2. Laporan dan Rekomendasi Hasil Pembersihan Data Kotor.....	18
BUKTI 6-ADS	20
1. Analisis Teknik Transformasi Data	20
2. Transformasi Data	21
3. Dokumentasi Konstruksi Data.....	21
BUKTI 7-ADS	23
1. Pelabelan Data.....	23
2. Laporan Hasil Pelabelan Data	23
BUKTI 8-ADS	25
1. Parameter Model	25
2. Tools Pemodelan	26
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

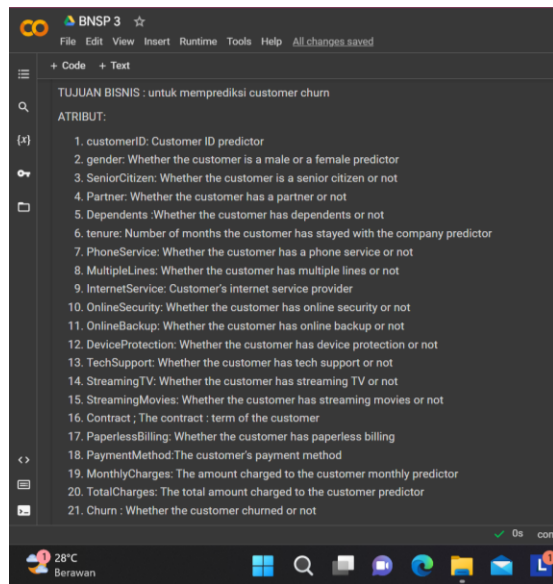
1. KEBUTUHAN DATA

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

JAWABAN

- Tujuan dari kebutuhan data pada project ini adalah memprediksi customer churn yang setia dengan Perusahaan dan yang pergi dengan Perusahaan, untuk kepentingan keberlanjutan Perusahaan tersebut, serta variable yang menyebabkan mereka keluar.
- Atribut yang dibutuhkan untuk memprediksi studi kasus ini sudah tersedia dalam sebuah data set dalam bentuk csv. Yang memiliki 21 kolom dan 7043 baris (ada yang berisi nilai kosong didalamnya)



- Volume data

```
[173] # untuk mengetahui informasi rinci pada dataset atau DataFrame
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
#   column                Non-Null Count  Dtype
---  ---
0   customerID            7043 non-null   object
1   gender                 7038 non-null   object
2   SeniorCitizen          7043 non-null   int64
3   Partner                7043 non-null   object
4   Dependents             7043 non-null   object
5   tenure                 7040 non-null   float64
6   PhoneService           7043 non-null   object
7   MultipleLines          7043 non-null   object
8   InternetService         7043 non-null   object
9   OnlineSecurity         7043 non-null   object
10  OnlineBackup           7043 non-null   object
11  DeviceProtection       7043 non-null   object
12  TechSupport            7043 non-null   object
13  StreamingTV            7043 non-null   object
14  StreamingMovies        7043 non-null   object
15  Contract               7043 non-null   object
16  PaperlessBilling       7043 non-null   object
17  PaymentMethod          7043 non-null   object
18  MonthlyCharges         7043 non-null   float64
19  TotalCharges           7043 non-null   object
20  Churn                  7043 non-null   object
dtypes: float64(2), int64(1), object(18)
memory usage: 1.1+ MB
```

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

JAWABAN

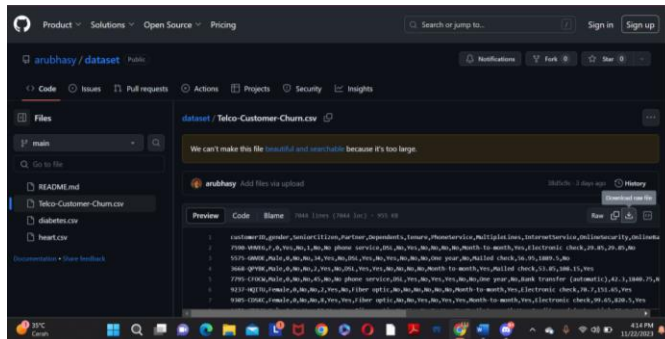
- Dataset yang saya ambil sudah tersedia dan di donwold dalam bentuk csv yang ditaruh di link github <https://github.com/arubhasy/dataset/blob/main/Telco-Customer-Churn.csv> dan merupakan metode web scraming. Tools yang saya gunakan untuk mendowold ialah dengan google choreme dan pemrosesan data melalui google collab

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	predictor (independent) variables
2.	gender	Whether the customer is a male or a female	predictor (independent) variables
3.	SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)	predictor (independent) variables
4.	Partner	Whether the customer has a partner or not (Yes, No)	predictor (independent) variables
5.	Dependents	Whether the customer has dependents or not (Yes, No)	predictor (independent) variables
6.	tenure	Number of months the customer has stayed with the company	predictor (independent) variables
7.	PhoneService	Whether the customer has a phone service or not (Yes, No)	predictor (independent) variables
8.	MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)	predictor (independent) variables
9.	InternetService	Customer's internet service provider (DSL, Fiber optic, No)	predictor (independent) variables
10.	OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)	predictor (independent) variables
11.	OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)	predictor (independent) variables



- Tools yang saya gunakan untuk menganalisa ialah menggunakan Google Collab notebook. Dengan menggunakan library pandas untuk menampilkan load data, seperti dibawah ini.

#untuk membangun model klasifikasi berbeda.

import numpy as np #untuk numpy

import pandas as pd #untuk dataframe

from sklearn.model_selection import train_test_split #digunakan untuk membagi dataset menjadi dua subset

from sklearn.preprocessing import LabelEncoder #mengubah label atau kategori dalam data menjadi angka

from sklearn.neighbors import KNeighborsClassifier #model klasifikasi

from sklearn.tree import DecisionTreeClassifier #model klasifikasi

from sklearn.ensemble import RandomForestClassifier #model klasifikasi

import xgboost as xgb #model klasifikasi

from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score

mengimpor berbagai metrik evaluasi yang digunakan untuk mengevaluasi kinerja model

from xgboost import XGBClassifier

```

1 import library yang dibutuhkan
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn as sk
import sklearn.datasets as datasets
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import auc
from sklearn.metrics import roc_auc_score
from sklearn.metrics import variance_inflation_factor
from sklearn.metrics import mean_squared_error
import pickle
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif
import statistics

2 setelah itu selanjutnya adalah panggil data pada folder laptop
from google.colab import files
uploaded = files.upload()

3 kemudian data yang sudah diupload tadi
df = pd.read_csv('Telo-Customer-Churn.csv')
df

```

	customerID	gender	seniorcitizen	partner	dependents	tenure	phonoservice	multiplelines	internetservice	onlinesecurity	...	deviceprotection	techsupport	streamingtv	streamingmusic	contract	paperlessbilling	paymentmethod	monthlycharges	totalcharges	churn
0	7596-VIVE6	F	0	Yes	No	1.0	No	No phone service	DSL	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
1	5575-GAY6E	Male	0	No	No	34.0	Yes	No	DSL	Yes	...	Yes	No	No	No	One year	No	Mailed check	56.95	1089.5	No
2	3866-GFV6K	Male	0	No	No	2.0	Yes	No	DSL	Yes	...	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
3	7795-GFV6K	Male	0	No	No	45.0	No	No phone service	DSL	Yes	...	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.36	1548.75	No
4	9237-HZTU7	Female	0	No	No	2.0	Yes	No	Fiber optic	No	...	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes
...
7038	6840-REJVS	Male	0	Yes	Yes	24.0	Yes	Yes	DSL	Yes	...	Yes	Yes	Yes	Yes	One year	Yes	Mailed check	84.85	1090.5	No
...
7043	7795-GFV6K	Male	0	Yes	No	75.0	Yes	Yes	Fiber optic	No	...	No	No	No	No	Month-to-month	No	Bank transfer (automatic)	439.10	7387.10	No

3. PENGINTEGRASIAN DATA

Instruksi Kerja:

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

JAWABAN

- Data tidak dintegrasikan hanya karena ada satu file csv, dan atribut yang dibutuhkan untuk tujuan analys sudah tersedia dalam file csv tersebut. Dengan terdapat 7043 raw dan 21 kolom. Setelah menampilkan data tersebut dapat dianalisis apay g dibutuhkan. Tetapi di dalam data tersebut ada nilai perlu perbaikan yang terdapat pada seniorcitizen, total charge.

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

JAWABAN

- Tipe data yang terkumpul adalah 21 kolom. Dalam dataset tersebut terdapat beberapa tipe data yaitu object dan 2 tipe data numerik (float, integer)

```
[173] # untuk mengetahui informasi rinci pada dataset atau DataFrame
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype  
---  --
 0   customerID            7043 non-null   object  
 1   gender                7043 non-null   object  
 2   SeniorCitizen         7043 non-null   int64   
 3   Partner               7043 non-null   object  
 4   Dependents            7043 non-null   object  
 5   tenure                7040 non-null   float64  
 6   PhoneService          7043 non-null   object  
 7   MultipleLines         7043 non-null   object  
 8   InternetService       7043 non-null   object  
 9   OnlineSecurity        7043 non-null   object  
10  OnlineBackup          7043 non-null   object  
11  DeviceProtection      7043 non-null   object  
12  TechSupport           7043 non-null   object  
13  StreamingTV           7043 non-null   object  
14  StreamingMovies       7043 non-null   object  
15  Contract              7043 non-null   object  
16  PaperlessBilling       7043 non-null   object  
17  PaymentMethod         7043 non-null   object  
18  MonthlyCharges        7043 non-null   float64  
19  TotalCharges          7043 non-null   object  
20  Churn                 7043 non-null   object  
dtypes: float64(2), int64(1), object(18)
memory usage: 1.1+ MB
```

- Kita dapat menganalisa suatu data dengan menampilkan visualisasi korelasi antar atribut dengan menggunakan heatmap seperti gambar dibawah ini

- **Keterangan:** jika warna yang dihasilkan berwarna biru dan angka bernilai negative, maka korelasi antar variable tersebut lemah, tetapi jika warna yang dihasilkan lebih condong ke merah dan nilainya positive maka hubungan antar variable tersebut kuat.
- **Contoh korelasi kuat:** hubungan antara variable Internet Service dan variable Streaming movie yang lebih condong ke warna merah dengan nilai variable positif 0,71, yang berarti jika internetnya buruk atau baik dapat mempengaruhi streaming movie tersebut. Sehingga berpengaruh terhadap pengguna yang melakukan streaming movie
- **Contoh korelasi lemah:** hubungan antara variable Internet Service dan variable Monthly charge yang lebih condong ke biru merah dengan nilai variable negative -0,29, yang berarti jika internetnya buruk atau tidak terlalu berpengaruh pada biaya bulanan. Sehingga menurut prediksi jika perbaikan internet bisa jadi tidak adanya penambahan biaya.



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

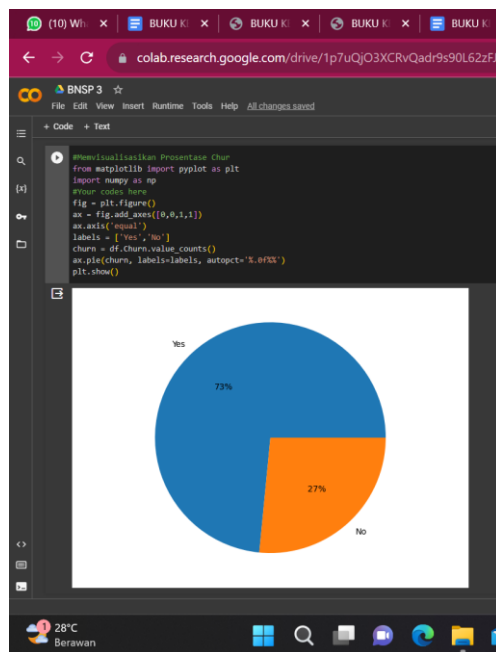
JAWABAN

- Untuk menampilkan sebuah deskriptif statistic dapat menggunakan code describe(). Yang dimana nanti akan muncul tampilan seperti count, mean, standar deviasi, nilai minimal dan maksimal, Q1-Q3, 25%,50%,75%.

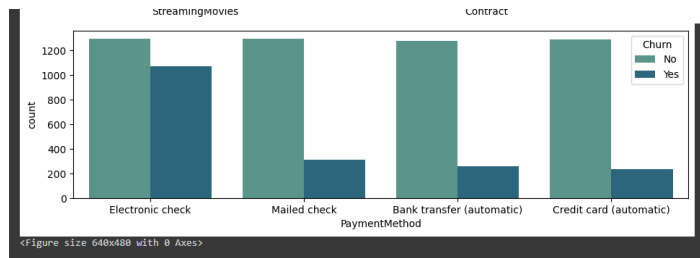
df.describe()

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

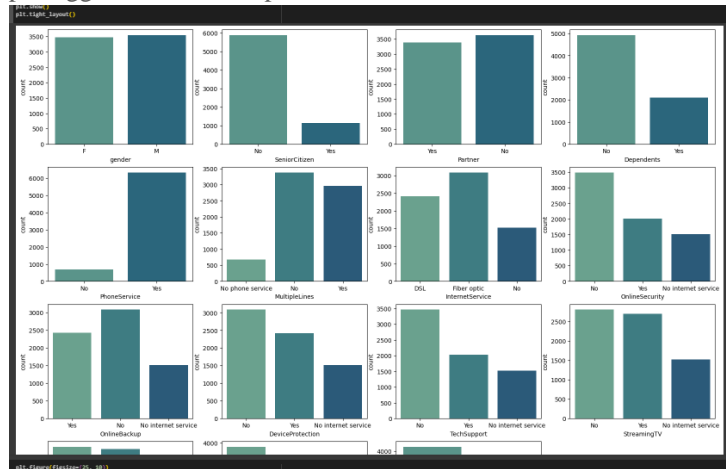
- Dalam EDA saya mencoba menampilkan data grafik pada data churn untuk melihat bahwa seberapa banyak orang yang churn dan tidak churn. Dalam table tersebut terlihat terdapat 75% yang tidak churn dan ada 27% yang churn. Sehingga kita dapat mengambil kesimpulan bahwa pelanggan yang tidak churn lebih banyak



- Selain itu, saya juga menampilkan visualisasi grafik dari masing masing variable, seperti gambar dibawah. Yang dimana saya mengambil contoh dari variable paymentMethod. Dalam grafik tersebut terlihat bahwa pelanggan yang banyak bertahan pada Perusahaan atau yg tidak churn dengan yang melakukan pembayaran melalui credit card dan bank tranfer.
- Dalam haal ini dapat menjadi evaluasi bagi pihan Perusahaan untuk lebih meningkatkan pembayaran dengan metode lainnya. Dan diberi kemudahan dalam melakukan transaksi dengan metode lainnya. Sehingga pelanggan tetep setia pada Perusahaan.



- Grafik dibawah ini adalah grafik dari variable lainnya yang dapat memprediknya churnnya pelanggan atau tidak, seperti contoh diatas



3. LAPORAN TELAHAH 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.

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

JAWABAN

- Untuk kualitas yang ada dalam data set yang telah tersedia sudah cukup lengkap, maka tidak perlu mencari data yg lain, tetapi jika di cek terdapat nilai kosong atau missing value. Missing value dapat mempengaruhi pada kualitas data, maka kita dapat mengeceknya code dibawah. Dapat terlihat bahwa terdapat 5 missing value pada Senior Citizen dan 3 missing value yang terdapat pada tenure.

```
[ ] df.isnull().sum()
customerID      0
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    0
Churn           0
dtype: int64
```

- Setelah itu saya pada atribut TotalCharge saya mengubah tipedatanya ke float atau numerik

```
[ ] # mengisi nilai yang kosong dengan NaN
df['TotalCharges'] = df['TotalCharges'].replace(' ', np.nan)
# konversi object ke float
df['TotalCharges'] = df['TotalCharges'].astype(float)
```

- Ketika adanya perubahan dari tipe data TotalCharge yang semula object menjadi float terdapat missing value pada Total Charge

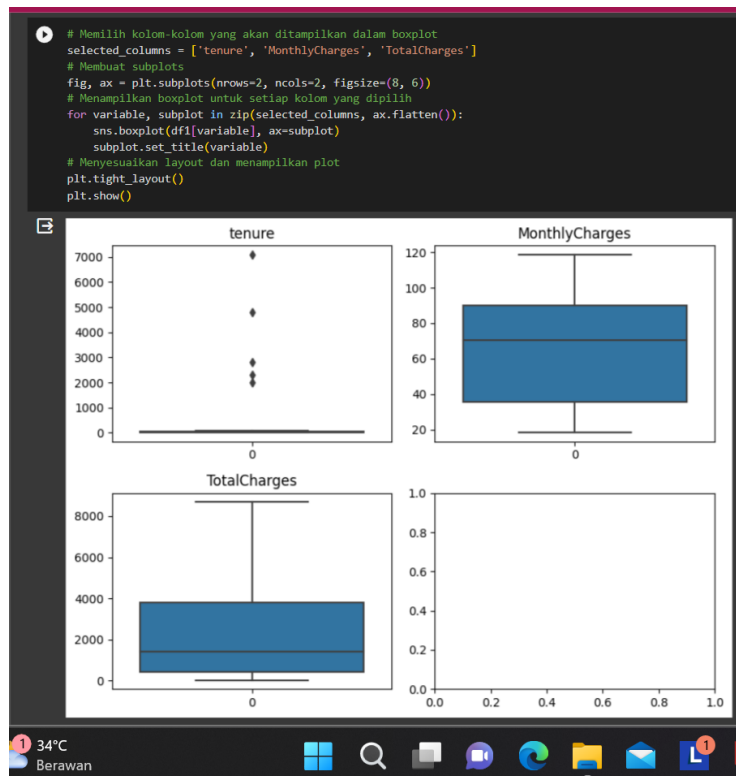
```
[ ] df.isnull().sum()
```

```
customerID      0
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
```

- Berikut adalah pengecekan Kembali terhadap kelengkapan data. Dalam hal ini data sudah cukup lengkap jdengan dilihat dari kualitas data, volume data dan data yang relevan.

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  -
 0   customerID            7043 non-null   object
 1   gender                7043 non-null   object
 2   SeniorCitizen         7043 non-null   object
 3   Partner               7043 non-null   object
 4   Dependents            7043 non-null   object
 5   tenure                7040 non-null   float64
 6   PhoneService          7043 non-null   object
 7   MultipleLines         7043 non-null   object
 8   InternetService       7043 non-null   object
 9   OnlineSecurity        7043 non-null   object
10  OnlineBackup          7043 non-null   object
11  DeviceProtection      7043 non-null   object
12  TechSupport           7043 non-null   object
13  StreamingTV           7043 non-null   object
14  StreamingMovies       7043 non-null   object
15  Contract              7043 non-null   object
16  PaperlessBilling      7043 non-null   object
17  PaymentMethod         7043 non-null   object
18  MonthlyCharges        7043 non-null   float64
19  TotalCharges          7042 non-null   float64
20  Churn                 7043 non-null   object
dtypes: float64(3), object(18)
memory usage: 1.1+ MB
```

Selain itu juga dilakukan untuk pengecekan outliers, terdapat outlier pada data, outliers dapat mempengaruhi hasil analisis. Dibawah ini terlihat bahwa dalam variable tenure terdapat 5 outliers, meskipun 5 outliers dapat mempengaruhi analisis nantinya.



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

JAWABAN

- Rekomendasi untuk penilaian kelengkapan data atau kualitas data adalah dengan melakukan handling terhadap missing value yang ada yaitu pada atribut Gender, Tenur dan Total Charge. Handling ini dilakukan dengan menghapus missing value menggunakan dropna() yang ada, karena jika missing value dibiarkan akan bedapampak pada analys nanti.
- Selanjutnya rekomendasi terhadap outliers, adanya 5 outliers yang terlihat perlu ditangani dengan menggunakan handling IQR pada outliers. Outlier dalam metode IQR sering diidentifikasi sebagai nilai yang berada di luar batas tertentu. Perlunya dilakukan handling outliers karena outliers sangat berpengaruh terhadap hasil analys nanti, dengan ini dilakukan handling menggunakan metode IQR yang terbilang cukup sederhana dan efektif dalam menangani outliers.

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)

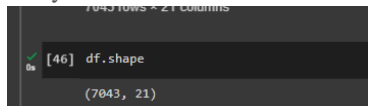
1. KRITERIA DAN TEKNIK PEMILIHAN DATA

Instruksi Kerja:

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

JAWABAN:

- Kriteria Teknik pemilihan data yaitu data yang lengkap, dengan tidak memiliki banyak missing value, outliers dan duplicate data. Jika terdapat maka perlu ditangani agar tidak mempengaruhi analys.



- Dalam gambar dibawah ini terlihat adanya missing value yang harus ditangani

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

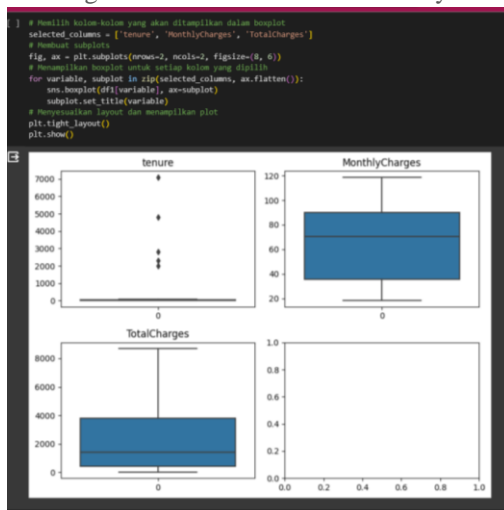
customerID      0
gender          5
SeniorCitizen   0
Partner         0
Dependents      0
tenure          3
PhoneService    0
MultipleLines   0
InternetService 0
OnlineSecurity  0
OnlineBackup    0
DeviceProtection 0
TechSupport     0
StreamingTV     0
StreamingMovies 0
Contract        0
PaperlessBilling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

- Dalam gambar dibawah ini terlihat tidak adanya duplicate data

```
[ ] df1.duplicated().sum()

0
```

- Dalam gambar dibawah ini terlihat adanya outliers yang perlu ditangani



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

JAWABAN

- Perlu dilakukannya penghapusan pada kolom customersID karena tidak berpengaruh


```
✓ 0s #menghapus kolom customerID karena tidak perlu digunakan
df1.drop(columns=['customerID'], inplace=True)
```

- Dengan adanya penghapusan pada kolom customersID, serta handling missing value dan outliers terjadi adanya perubahan yang menjadi 20 kolom dan 7020 row

```
✓ 0s [28] df1.shape
(7020, 20)
```

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

JAWABAN

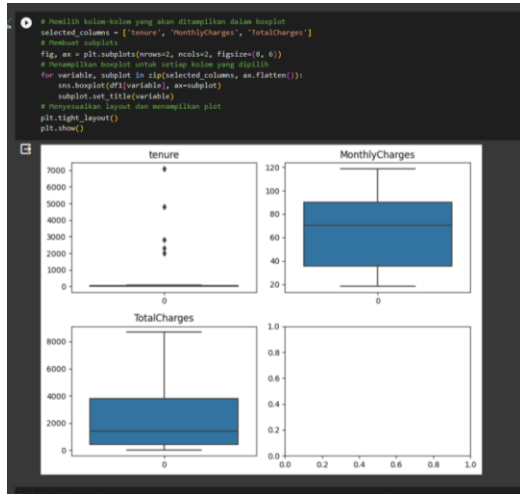
- Pada gambar dibawah ini terlihat bahwa adanya missing value pada attribute gender, teanure, dan TotalCharge

```
[68] df.isnull().sum()
customerID    0
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
```

Pada gambar dibawah ini terlihat bahwa tidak adanya duplicate

```
[71] df1.duplicated().sum()
0
tidak ada duplicate
```

- Pada gambar dibawah ini terlihat adanya 5 outliers yang ada pada atribut teanuer dan perlu ditangani menggunakan metode IQR. Karena jika tidak ditangani akan berdampak pada hasil analys



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

JAWABAN

- Pembersihan data yang dilakukan ialah menghapus nilai yang ada missing value tadi menggunakan dropna() . kenapa dihapus? Karena data kosong terbilang sedikit. Jika dibiarkan saja maka akan berpengaruh terhadap hasil data analys nantinya. Dapat dilihat seperti gambar dibawah ini bahwa tidak adanya missing value lagi

```
[ ] df1 = df.dropna()

df1.isnull().sum()

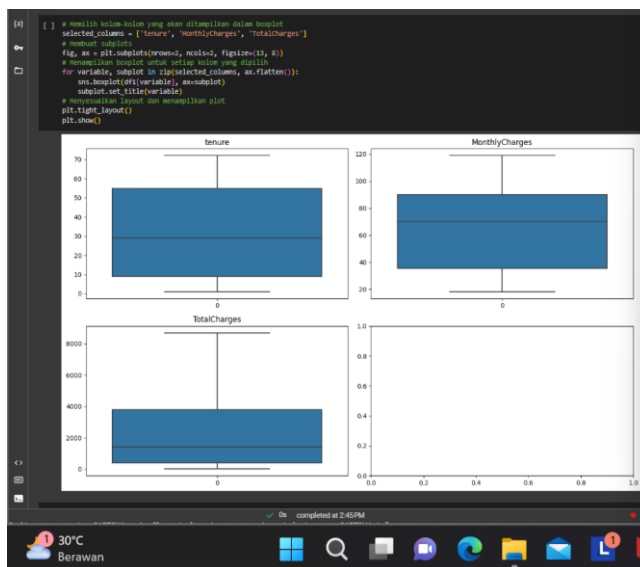
customerID    0
gender         0
SeniorCitizen 0
Partner       0
Dependents    0
tenure        0
PhoneService  0
MultipleLines 0
InternetService 0
onlinesecurity 0
onlinesbackup 0
deviceprotection 0
TechSupport   0
StreamingTV   0
StreamingMovies 0
Contract      0
PaperlessBilling 0
PaymentMethod 0
MonthlyCharges 0
TotalCharges  0
Churn         0
dtype: int64

[ ] df1.duplicated().sum()

0
```

- Dan untuk pembersihan data outlier dengan handling menggunakan metode IQR yang nantinya agar tidak ada outlier, karena outliers sangat berpengaruh terhadap analysis. Seperti gambar dibawah ini setelah adanya penanganan terlihat bahwa tidak ada outliers lagi

```
#handling outlier
Q1 = df1.quantile(0.25)
Q3 = df1.quantile(0.75)
IQR = Q3 - Q1
df1 = df1[~((df1 < (Q1 - 1.5 * IQR)) | (df1 > (Q3 + 1.5 * IQR))).any(axis=1)]
```



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

JAWABAN

- Pada Teknik transformasi data ini yang perlu dilakukan adalah dengan menyaring atribut yang tidak digunakan seperti atribut customerID, atribut ini tidak digunakan karena tidak berpengaruh pada churn atau tidaknya
- Selanjutnya penanganan terkait tipe data gender yang tidak konsisten yaitu pada atribut gender. Terdapat nilai 'Female', 'F', 'Male' dan 'M'. Oleh karena itu, perlu dilakukan standardisasi untuk memudahkan pemahaman data, nilai 'Female' akan diubah menjadi 'F' dan nilai 'Male' akan diubah menjadi 'M'.
- Selanjutnya Mengubah tipe data Senior Citizen menjadi 'Yes:1' dan 'No:0' yang nantinya agar lebih mudah untuk dipahami
- Selanjutnya adanya perubahan terkait dengan tipe data, yang dimana akan dirubah ke dalam tipe data numerik semua, karena untuk melakukan modeling tipe data haruslah numerik.

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

JAWAABAN:

- Dilakukannya penghapusan pada kolom customersID, seperti yang sudah dijelaskan sebelumnya

```
[72] #menghapus kolom customerID karena tidak perlu digunakan
df1.drop(columns=['customerID'], inplace=True)
```

- Dilakukannya Standarisasi atribut gender dengan mengubah nilainya menjadi 'F dan M' seperti yang sudah dijelaskan sebelumnya

```
[51] df['gender'].value_counts()
Male    3550
Female  3483
M         3
F         2
Name: gender, dtype: int64

[52] # standarisasi nilai data 'gender'
df['gender'] = df['gender'].str.replace('Female', 'F')
df['gender'] = df['gender'].str.replace('Male', 'M')

[53] df['gender'].value_counts()
M     3553
F     3485
Name: gender, dtype: int64
```

- Dilakukannya mengubah nilai pada atribut SeniorCitizen menjadi 'Yes;1 dan No;0' seperti yang sudah dijelaskan sebelumnya

```
#konversi int ke object
df['SeniorCitizen'] = df['SeniorCitizen'].replace({1: 'Yes', 0: 'No'})
```

- Dilakukannya Label encoding dengan mengubah semua tipe data menjadi numerik karena akan dilakukannya modeling, seperti yang sudah dijelaskan sebelumnya. Pada gambar terlihat hasilnya numerik semua

```
[32] from sklearn.preprocessing import LabelEncoder
def object_to_int(dataframe_series):
    if dataframe_series.dtype == object:
        dataframe_series = LabelEncoder().fit_transform(dataframe_series)
    return dataframe_series

df1 = df1.apply(lambda x: object_to_int(x))
df1.head()
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	Multiplatforms	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	Total
0	0	0	1	0	1.0	0	1	0	0	2	0	0	0	0	0	1	2	29.85	
1	1	0	0	0	34.0	1	0	0	2	0	2	0	0	0	1	0	3	56.95	
2	1	0	0	0	2.0	1	0	0	2	2	0	0	0	0	0	1	3	53.85	
3	1	0	0	0	45.0	0	1	0	2	0	2	2	0	0	1	0	0	42.30	
4	0	0	0	0	2.0	1	0	1	0	0	0	0	0	0	0	1	2	70.70	

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.

JAWABAN

- Sudah dijelaskan sebelumnya pada poin 1 dan 2

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

JAWABAN

- Disini dibagi menjadi 2 label, yaitu label X yang berisi semua kolom atribut, kecuali atribut churn. Dan label y yang menjadi target, yang berisi kolom atribut Churn. Kenapa dilakukan pelabelan? Agar memudahkan machine learning dalam pemodelannya nanti

```
[34] X = df1.drop(columns=['Churn'])  
      y = df1['Churn']
```

2. LAPORAN HASIL PELABELAN DATA

Instruksi Kerja:

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

JAWABAN

- Gambar pada label dibawah ini merupakan isi dari label X yang yang berisi semua kolom atribut, kecuali atribut Churn

```

[14]: X = df.drop(columns=['Churn'])
      y = df['Churn']

```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultiLine	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMusic	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges
0	0	0	1	0	18.0	0	1	0	0	0	0	0	0	0	0	1	2	29.85	29.85
1	1	0	0	0	34.0	1	0	0	2	0	0	0	0	0	0	0	3	59.05	1099.95
2	1	0	0	0	2.0	1	0	0	2	2	0	0	0	0	0	1	3	53.05	105.15
3	1	0	0	0	48.0	0	1	0	2	0	2	2	0	0	1	0	0	42.30	1840.15
4	0	0	0	0	2.0	1	0	1	0	0	0	0	0	0	0	1	2	70.70	151.05
7038	1	0	1	1	24.0	1	2	0	2	0	2	2	2	2	1	1	3	84.00	1990.00
7039	0	0	1	1	72.0	1	2	1	0	2	2	0	2	2	1	1	1	403.20	7392.00
7040	0	0	1	1	35.0	0	1	0	2	0	0	0	0	0	0	1	3	39.00	146.45
7041	1	1	1	0	4.0	1	2	1	0	0	0	0	0	0	0	1	3	74.40	305.00
7042	1	0	0	0	60.0	1	0	1	2	0	2	2	2	2	2	1	0	105.05	6044.00

7020 rows x 20 columns

- Gambar pada label dibawah ini merupakan isi dari label y yang berisi Churn

```

y
0      0
1      0
2      1
3      0
4      1
...
7038   0
7039   0
7040   0
7041   1
7042   0
Name: Churn, Length: 7020, dtype: int64

```

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

JAWABAN

- Disini saya membuat split data train dan test dengan rasio 80:40. Random state ditentukan berdasarkan nilai umum yaitu 42. Random state merupakan parameter yang bersungsi untuk mengontrol randomization yang terjadi selama proses pemisahan data.
- Dalam dataset tersebut, algoritma model algoritma machine learning yang saya gunakan adalah KNN, Decision Tree, Random Forest, XGBoost. Dengan masing masing parameternya pada gambar dibawah ini, terdapat parameter n_neighbors, random state, max_depth, n_estimator untuk mengoptimalisasikan nilai hasil evaluasi

```
Random Forest, dan XGBoost. Hasil dr disimpan dalam knn_pred, dt_pred, rf_pred, dan xgb_pred.

# K-Nearest Neighbors (KNN)
knn_model = KNeighborsClassifier(n_neighbors=5)
knn_model.fit(X_train, y_train)
knn_pred = knn_model.predict(X_test)

# Decision Tree
dt_model = DecisionTreeClassifier(max_depth=5)
dt_model.fit(X_train, y_train)
dt_pred = dt_model.predict(X_test)

# Random Forest
rf_model = RandomForestClassifier(n_estimators=100, max_depth=5, random_state=42)
rf_model.fit(X_train, y_train)
rf_pred = rf_model.predict(X_test)

# XGBoost
xgb_model = xgb.XGBClassifier(n_estimators=100, max_depth=3)
xgb_model.fit(X_train, y_train)
xgb_pred = xgb_model.predict(X_test)
```

- Dan untuk mengukur sejauh mana model dapat memprediksi nilai terbaik maka digunakan confusion matrik dengan parameter evaluasi yang diminta oleh soal yaitu Accuracy : mengukur sejauh mana model klasifikasi memberikan prediksi yang benar secara keseluruhan.
Precision : mengukur sejauh mana prediksi positif dari model benar, menghitung rasio True Positives terhadap total prediksi positif
Recall : mengukur sejauh mana model dapat mendeteksi semua kasus positif yang seharusnya, menghitung rasio True Positives terhadap total kasus positif yang seharusnya
F1-Score : keseimbangan antara kemampuan model mengidentifikasi kasus positif dan menghindari kesalahan positif.

```
#K-Nearest Neighbors (KNN)
knn_accuracy = accuracy_score(y_test, knn_pred)
knn_f1 = f1_score(y_test, knn_pred)
knn_recall = recall_score(y_test, knn_pred)
knn_precision = precision_score(y_test, knn_pred)

#Decision Tree
dt_accuracy = accuracy_score(y_test, dt_pred)
dt_f1 = f1_score(y_test, dt_pred)
dt_recall = recall_score(y_test, dt_pred)
dt_precision = precision_score(y_test, dt_pred)

#Random Forest
rf_accuracy = accuracy_score(y_test, rf_pred)
rf_f1 = f1_score(y_test, rf_pred)
rf_recall = recall_score(y_test, rf_pred)
rf_precision = precision_score(y_test, rf_pred)

#XGBoost
xgb_accuracy = accuracy_score(y_test, xgb_pred)
xgb_f1 = f1_score(y_test, xgb_pred)
xgb_recall = recall_score(y_test, xgb_pred)
xgb_precision = precision_score(y_test, xgb_pred)
```

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

JAWABAN

- Pada analys kali ini saya menggunakan tools google collab dengan pemrograman phyton
- Gambar dibawah ini merupakan code untuk membangun model algoritma KNN, Decision Tree, Random Forest, XGBoost dengan menggunakan parameter parameter n_neighbors, random state, max_depth, n_estimator yang berfungsi untuk mengoptimisasikan ke nilai evaluasi parameter yang digunakan

Random Forest, dan XGBoost. Hasil dr disimpan dalam knn_pred, dt_pred, rf_pred, dan xgb_pred.

```
# K-Nearest Neighbors (KNN)
knn_model = KNeighborsClassifier(n_neighbors=5)
knn_model.fit(X_train, y_train)
knn_pred = knn_model.predict(X_test)

# Decision Tree
dt_model = DecisionTreeClassifier(max_depth=5)
dt_model.fit(X_train, y_train)
dt_pred = dt_model.predict(X_test)

# Random Forest
rf_model = RandomForestClassifier(n_estimators=100, max_depth=5, random_state=42)
rf_model.fit(X_train, y_train)
rf_pred = rf_model.predict(X_test)

# XGBoost
xgb_model = xgb.XGBClassifier(n_estimators=100, max_depth=3)
xgb_model.fit(X_train, y_train)
xgb_pred = xgb_model.predict(X_test)
```

- Confusion matrix evaluasi menggunakan Accuracy, Precision, Recall, dan F1-Score, seperti yang diminta pada soal dan sudah dijelaskan diatas. Pada gambar dibawah ini merupakan code mndefiniskan confusion matrix

```
#K-Nearest Neighbors (KNN)
knn_accuracy = accuracy_score(y_test, knn_pred)
knn_f1 = f1_score(y_test, knn_pred)
knn_recall = recall_score(y_test, knn_pred)
knn_precision = precision_score(y_test, knn_pred)

#Decision Tree
dt_accuracy = accuracy_score(y_test, dt_pred)
dt_f1 = f1_score(y_test, dt_pred)
dt_recall = recall_score(y_test, dt_pred)
dt_precision = precision_score(y_test, dt_pred)

#Random Forest
rf_accuracy = accuracy_score(y_test, rf_pred)
rf_f1 = f1_score(y_test, rf_pred)
rf_recall = recall_score(y_test, rf_pred)
rf_precision = precision_score(y_test, rf_pred)

#XGBoost
xgb_accuracy = accuracy_score(y_test, xgb_pred)
xgb_f1 = f1_score(y_test, xgb_pred)
xgb_recall = recall_score(y_test, xgb_pred)
xgb_precision = precision_score(y_test, xgb_pred)
```

- Setelah itu memanggil hasil evaluasi

```

#memanggil hasil evaluasi
print('KNN Accuracy: {knn_accuracy}')
print('KNN precision: {knn_precision}')
print('KNN recall: {knn_recall}')
print('KNN f1 score: {knn_f1}\n')

print('Decision Tree Accuracy: {dt_accuracy}')
print('Decision Tree precision: {dt_precision}')
print('Decision Tree recall: {dt_recall}')
print('Decision Tree f1 score: {dt_f1}\n')

print('Random Forest Accuracy: {rf_accuracy}')
print('Random Forest precision: {rf_precision}')
print('Random Forest recall: {rf_recall}')
print('Random Forest f1 score: {rf_f1}\n')

print('XGB Accuracy: {xgb_accuracy}')
print('XGB precision: {xgb_precision}')
print('XGB recall: {xgb_recall}')
print('XGB f1 score: {xgb_f1}\n')

```

KNN Accuracy: 0.7592592592592593
 KNN precision: 0.5660377358498566
 KNN recall: 0.48214477211796246
 KNN f1 score: 0.4782194357366771

 Decision Tree Accuracy: 0.7777777777777778
 Decision Tree precision: 0.6320346320346321
 Decision Tree recall: 0.3914289115281501
 Decision Tree f1 score: 0.48344370860927155

 Random Forest Accuracy: 0.7849902849002849
 Random Forest precision: 0.6577777777777778
 Random Forest recall: 0.3967828418230563
 Random Forest f1 score: 0.4949832775919732

 XGB Accuracy: 0.7813390313390314
 XGB precision: 0.613013698630137
 XGB recall: 0.47989276139410186
 XGB f1 score: 0.5383458646616541

- Dan memanggil hasil model evaluasi terbaik

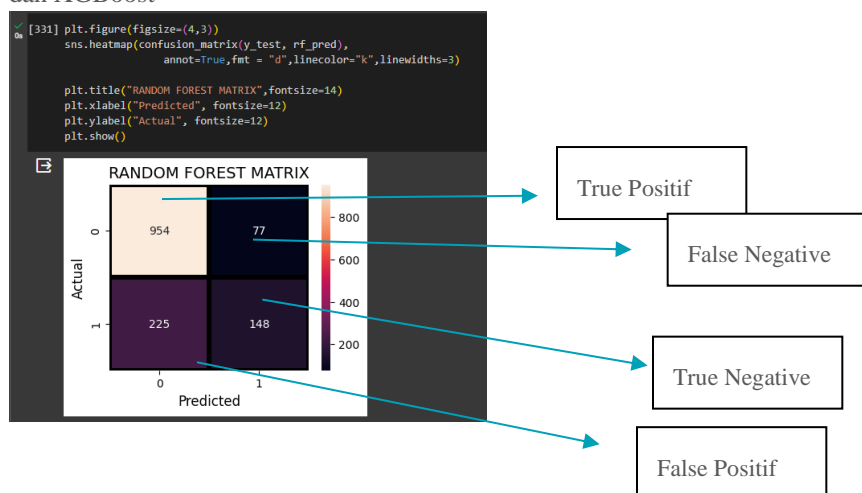
```

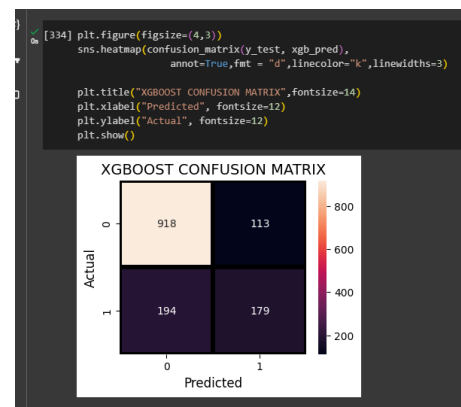
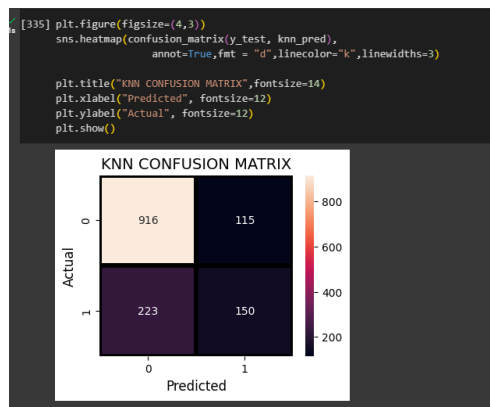
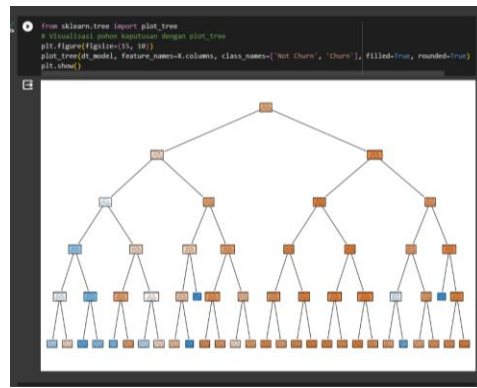
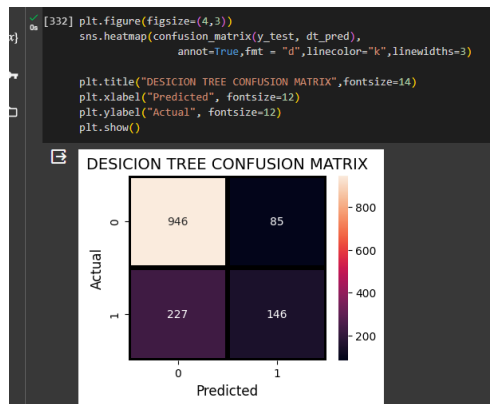
[329] # Memilih model terbaik berdasarkan confusion matrix
# Model dengan nilai akurasi tertinggi adalah model terba
models = {
    'Decision Tree': dt_accuracy,
    'Random Forest': rf_accuracy,
    'XGBoost': xgb_accuracy
}
best_model = max(models, key=models.get)
print("Best Model:", best_model)

Best Model: Random Forest

```

- Gambar dibawah merupakan visualisasi confusion matrix K-NN, Decision Tree, Random Forest, dan XGBoost





HASIL

- Dari hasil diatas Tujuannya dari analysis ini adalah untuk memprediksi chustomer churn. dari 4 model klasifikasi yang telah dievaluasi, yaitu K-NN, Decision Tree, Random Forest, dan XGBoost. Masing-masing model memiliki sejumlah metrik evaluasi kinerja yang berbeda, termasuk nilai akurasi, presisi, recall, dan skor F1.
- Sesuai yang diminta pada soal keberhasilan suatu akurasi ialah $>70\%$, makan dari evaluasi yang telas saya lakukan ialah **semua model memenuhi yaitu $>70\%$** , yang berarti berhasil. Dengan hasil evaluasinya **Random forest accuracy yang memiliki nilai paling tinggi yaitu 0.784900284**
- Dari **hasil confucion matrix** pada Random Forest :
 - True Positive : Hasil prediksi customer yang churn, bernar sebagai customer yang churn ada 954
 - True Negative : Hasil prediksi customer yang customer yang tidak churn dan model memprediksi benar bahwa customer tidak churn ada 148
 - False Positive : hasil prediksi customers yang tidak churn tetapi model memprediksi bahwa customer churn ada 225
 - False Negative : Hasil prediksi customer yang churn tetapi model memprediksi bahwa customer churn ada 77

```

#selanjutnya ialah melakukan feature importance untuk mengetahui variable mana yang penting dan berpengaruh
rf.feature_importance = rf.model.feature_importances_
feature_names = X.columns
rf.feature_importance_df = pd.DataFrame({'feature': feature_names, 'importance': rf.feature_importance})
rf.feature_importance_df = rf.feature_importance_df.sort_values(by='importance', ascending=False)
print(rf.feature_importance_df)

```

	Feature	Importance
14	Contract	0.221944
4	tenure	0.152630
8	OnlineSecurity	0.144403
11	TechSupport	0.118553
18	TotalCharges	0.093521
17	MonthlyCharges	0.076623
7	InternetService	0.065803
9	OnlineBackup	0.032487
16	PaymentMethod	0.030323
10	DeviceProtection	0.015072
15	PaperlessBilling	0.013003
1	SeniorCitizen	0.009244
3	Dependents	0.006701
12	StreamingTV	0.005204
13	StreamingMovies	0.004757
6	Multiplanelines	0.004193
2	Partner	0.002231
5	PhoneService	0.001719
0	gender	0.001499

- Dari hasil diatas Feature Importance dengan hasil paling atas adalah Contract
- Contract : Pelanggan yang berlangganan dengan jenis kontrak memiliki dampak yang signifikan terhadap churn nya customer, hal ini bisa dipengaruhi karena kontrak yang kurang menguntungkan untuk pelanggan
- Tenure : Pelanggan yang telah berlangganan lama juga mempengaruhi churn atau tidaknya. Mungkin saja hal ini terjadi karena pelanggan merasa pelayanan untuk mereka sama saja dengan pelanggan biasa, tidak ada reward atau apapun karena berlangganan lama, dan kualitas Perusahaan semakin menurun
- Online Security : hal ini bisa terjadi karena mungkin keamanan yang kurang sehingga pelanggan beralih
- TechSupport : techsupport juga memainkan peran penting dalam mempengaruhi keputusan pelanggan untuk melakukan churn. Dimana jika pelanggan mendapatkan pelayanan dukungan teknis customer lebih puas sehingga tidak terjadi churn

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

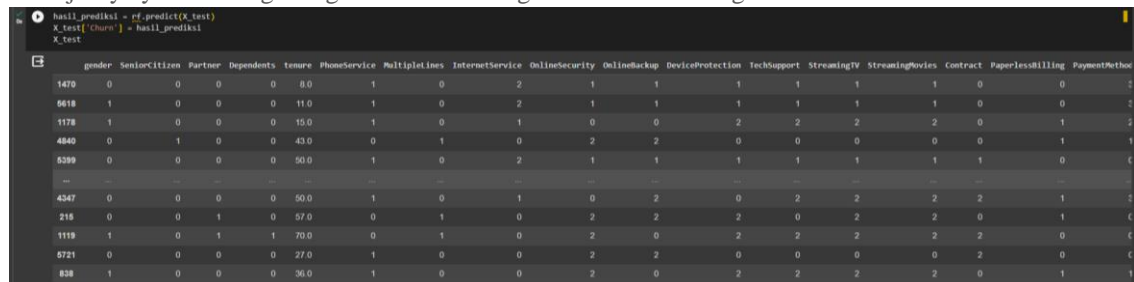
1. PENGGUNAAN MODEL DENGAN DATA RIIL

Instruksi Kerja:

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

JAWABAN

- Selanjutnya yaitu menngabungkan data test dengan data actual sebagai berikut.



```
hasil_prediksi = rf_predict(X_test)
X_test['Churn'] = hasil_prediksi
X_test
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	Multiplatforms	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod
1479	0	0	0	0	8.0	1	0	2	1	1	1	1	1	1	0	0	1
5618	1	0	0	0	11.0	1	0	2	1	1	1	1	1	1	0	0	1
1178	1	0	0	0	15.0	1	0	1	0	0	2	2	2	2	0	1	1
4840	0	1	0	0	43.0	0	1	0	2	2	0	0	0	0	0	1	1
5399	0	0	0	0	50.0	1	0	2	1	1	1	1	1	1	1	0	1
...
4347	0	0	0	0	50.0	1	0	1	0	2	0	2	2	2	2	1	1
216	0	0	1	0	57.0	0	1	0	2	2	2	0	2	2	0	1	1
1119	1	0	1	1	70.0	0	1	0	2	0	2	2	2	2	2	0	1
5721	0	0	0	0	27.0	1	0	0	2	2	0	0	0	0	2	0	1
838	1	0	0	0	36.0	1	0	0	2	0	2	2	2	2	0	1	1

2. PENILAIAN HASIL PEMODELAN

Instruksi Kerja:

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

JAWABAN

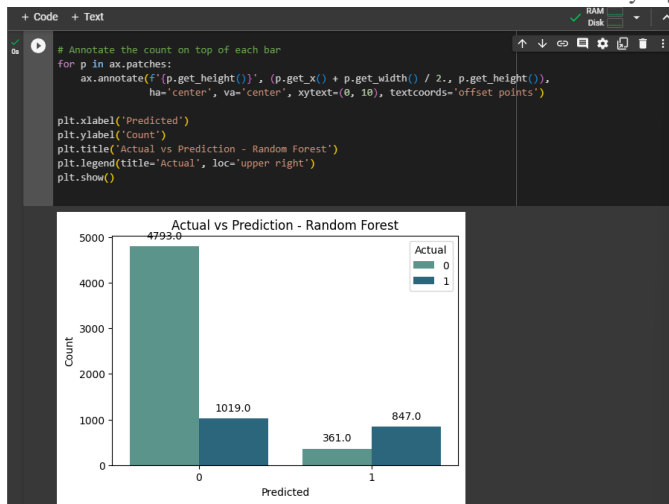
- Dibawah ini merupakan hasil data data actual dan predict

```
rf_predictions = rf_model.predict(X)
results_df_rf = pd.DataFrame({'Actual': y, 'RF_Predictions': rf_predictions})
print(results_df_rf)
```

	Actual	RF_Predictions
0	0	1
1	0	0
2	1	0
3	0	0
4	1	1
...
7038	0	0
7039	0	0
7040	0	0
7041	1	1
7042	0	0

[7020 rows x 2 columns]

- Pada hasil visualisasi dibawah ini bahwa Customer actual yang



<https://colab.research.google.com/drive/1p7uQjO3XCRvQadr9s90L62zFJPxSyfBv#scrollTo=nI5JXB83R5MY>