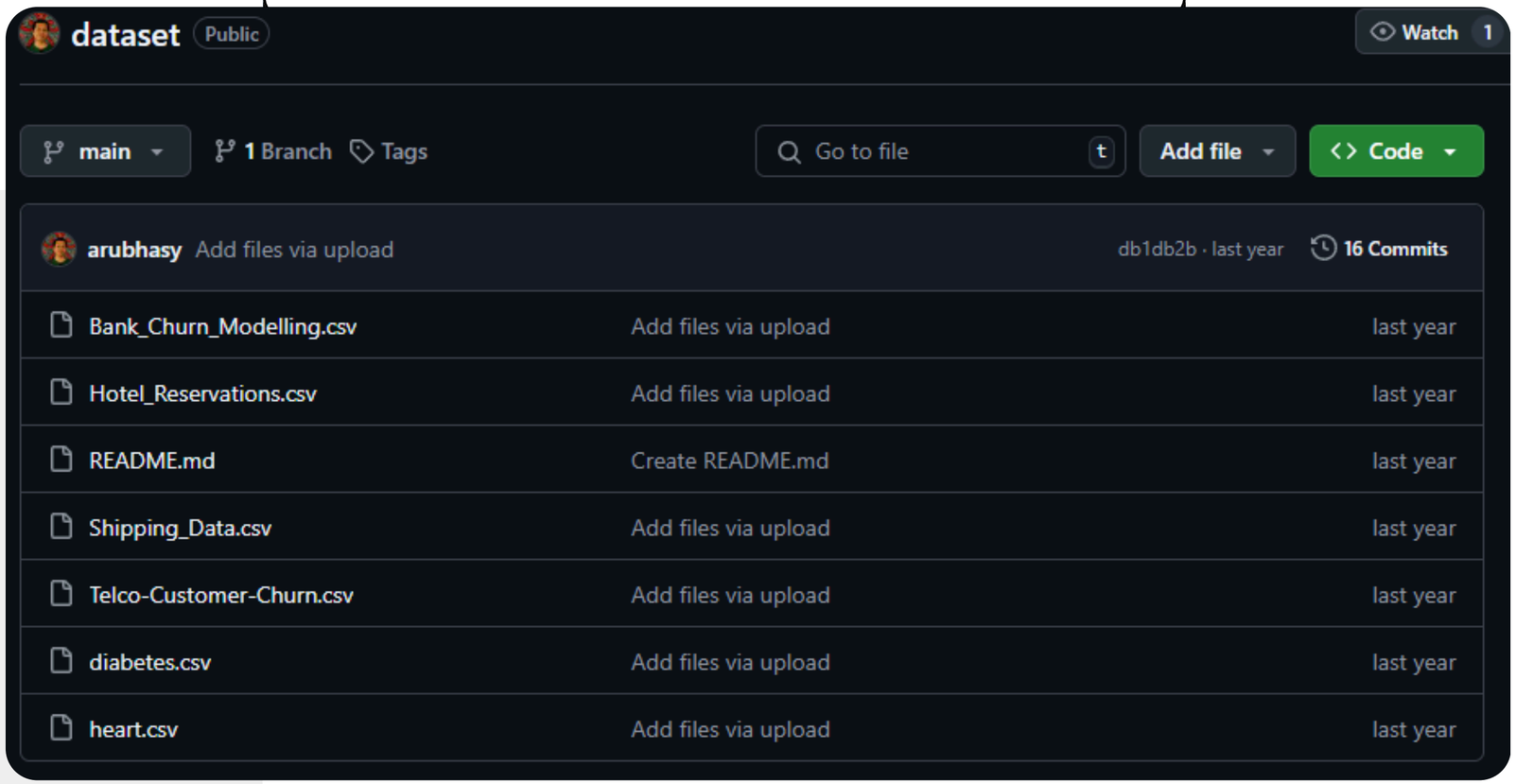
ASSESSMENT PRESENTATION

Presented by
Sunday B. Putera Mandiri

MENGUMPULKAN DATA

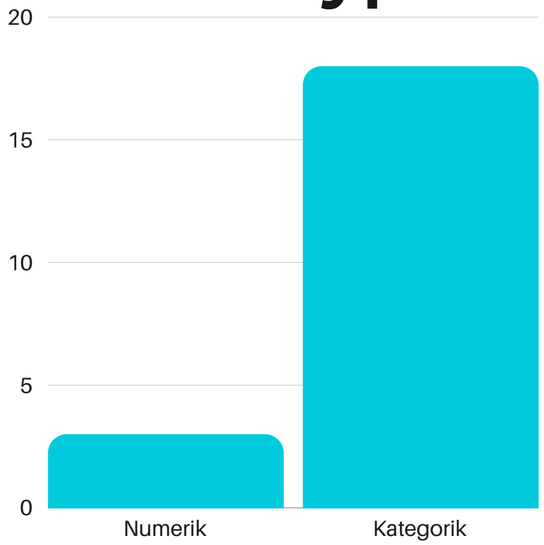
Sumber data



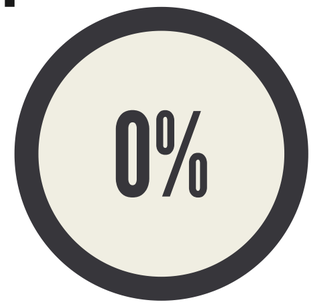
MENELAAH DATA

Sindhya Central Bank

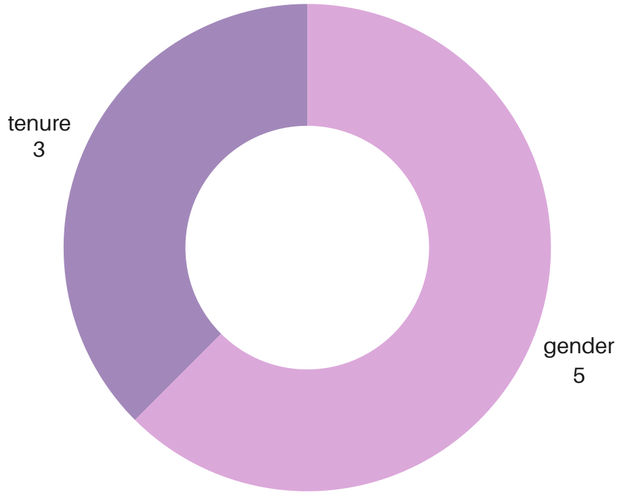
Data Type



Zero Duplicate Data

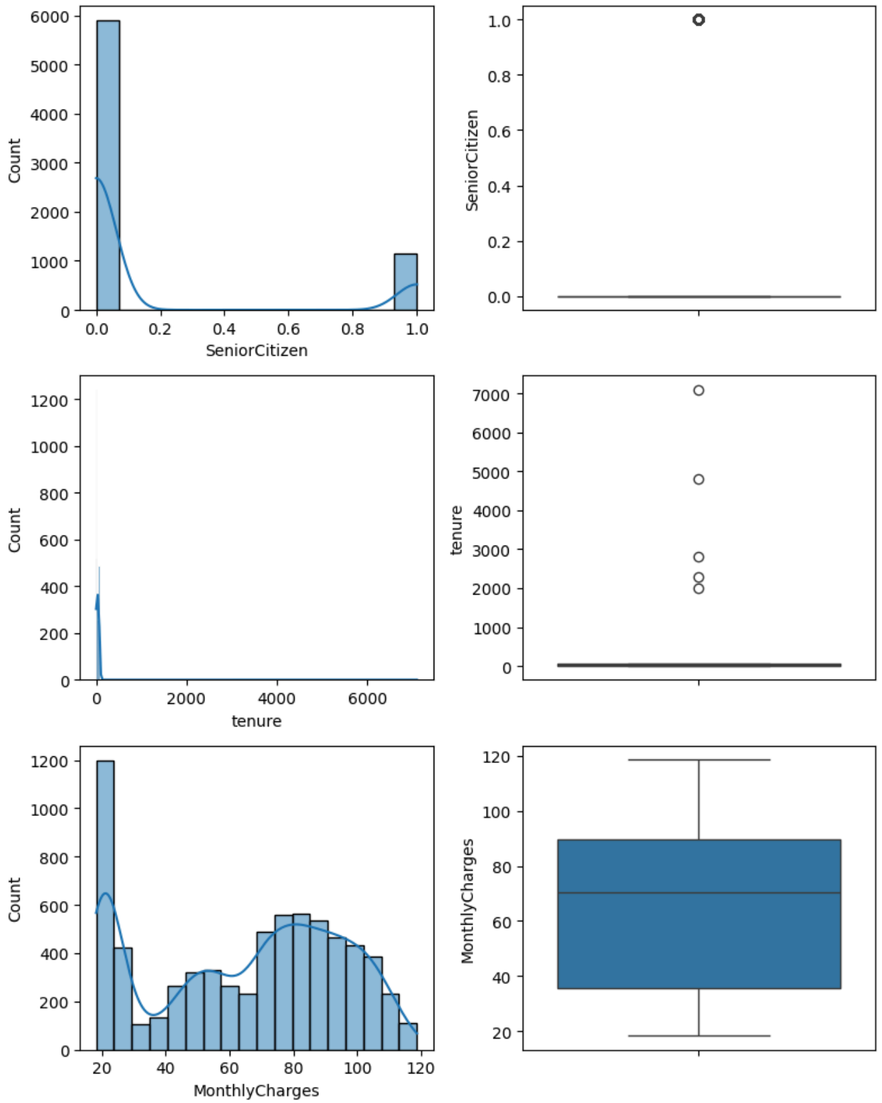


Missing Value



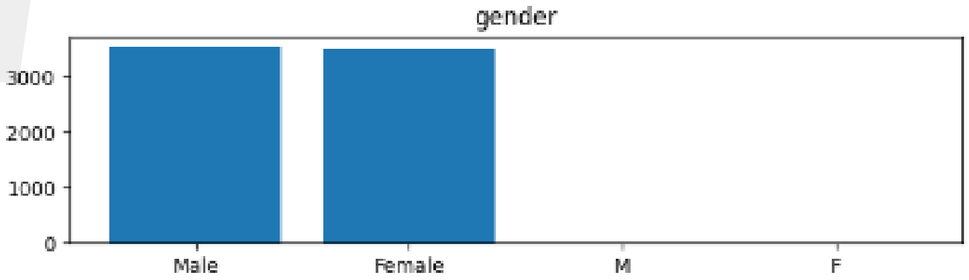
MENELAAH DATA

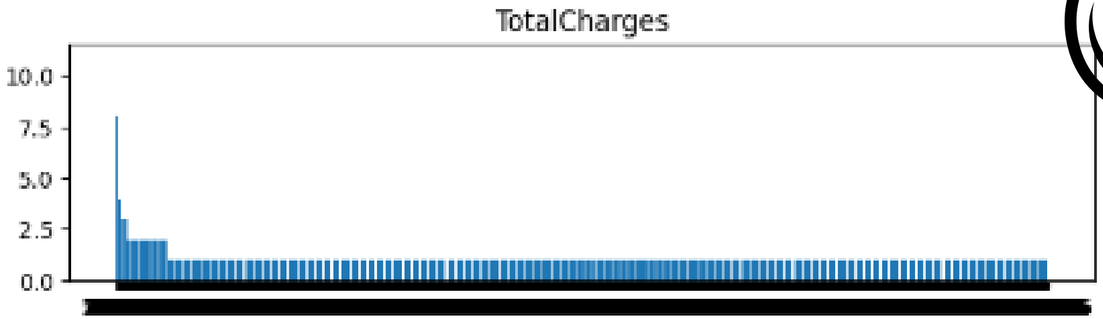
Numerik

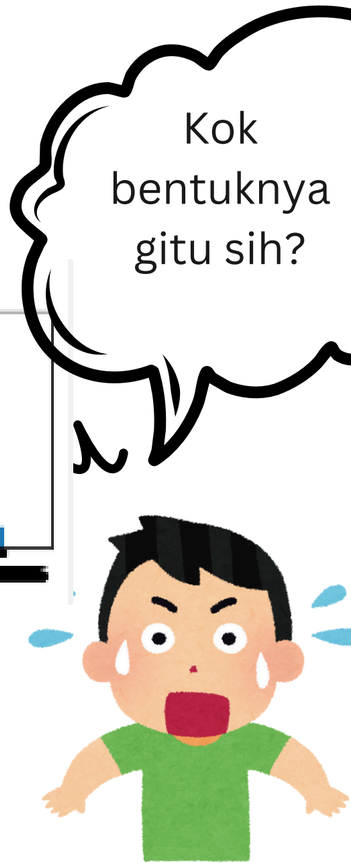


MENELAAH DATA

Kategorik







MENELAAH DATA

Kesimpulan

Terdapat Missing Value

Outliers

Terdapat variable dengan tipe data yang tidak sesuai

Inkonsisten Data

Agenda

Untuk menjadikan data lebih siap digunakan dalam analisis, perlu dilakukan beberapa langkah data cleaning

Stage 1

mengatasi missing value

Stage 2

mengatasi inkonsistensi data dengan menyamakan format pada gender

Stage 3

mengubah tipe data TotalCharges menjadi numerik

Stage 4

transformasi atau rekonstruksi data pada MonthlyCharges untuk distribusi yang lebih seimbang (Optional)

Stage 5

Mengatasih string kosong pada TotalCharges

DATA CLEANING

Gender

Menyamakan value

```
method replcae
df['gender'] = df['gender'].replace('F', 'Female')
df['gender'] = df['gender'].replace('M', 'Male')
```

Mengisih missing value berdasarkan proporsi

```
fill_values = random.choices(['Male', 'Female'], weights=[male_proportion, female_proportion], k=missing_count)
df.loc[df['gender'].isnull(), 'gender'] = fill_values
```

Output

gender	count
Male	3557
Female	3486

dtype: int64

DATA CLEANING

tenure

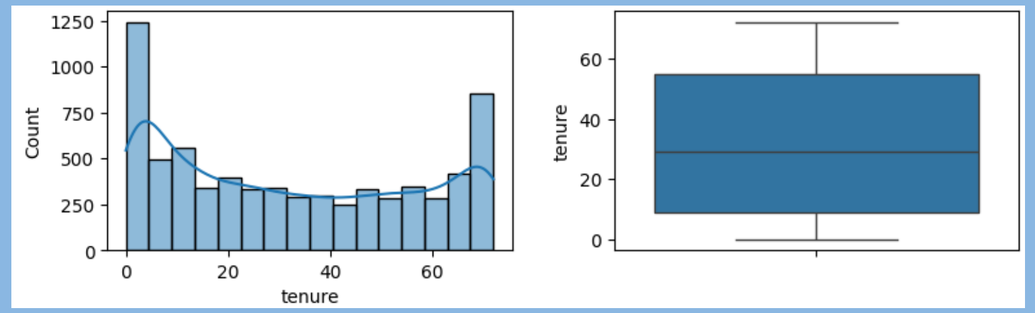
Mengisih missing value

```
df['tenure'] =
df['tenure'].fillna(df['tenure'].mean())
```

Mengatasih outliers

```
# Menghitung percentile ke-95
percintile_95 = df['tenure'].quantile(0.95)
# Mengganti outliers denga nilai percentile
ke-95
df['tenure'] = np.where(df['tenure'] >
percintile_95, percintile_95, df['tenure'])
```

Output



DATA CLEANING

TotalCharges

Mengubah tipe data

```
df['TotalCharges'] =
pd.to_numeric(df['TotalCharges'],error
s='coerce')
```

Mengisih missing value

```
df['TotalCharges'] =
df['TotalCharges'].fillna(df['TotalCharges'].me
dian())
```

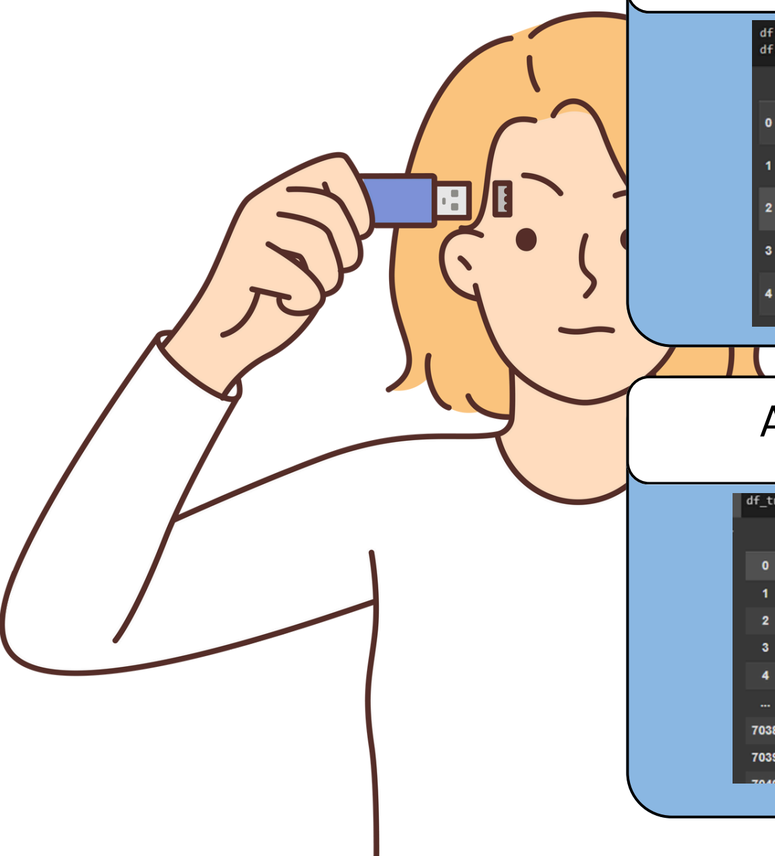
Output

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

```
0
```

TRANSFORMASI

Transformasi dilakukan menggunakan method LabelEncoder()



Before

```
df = pd.read_csv("/content/drive/MyDrive/2024_ITK_DS_Training/Telco-Customer-Churn.csv")
df.head()
```

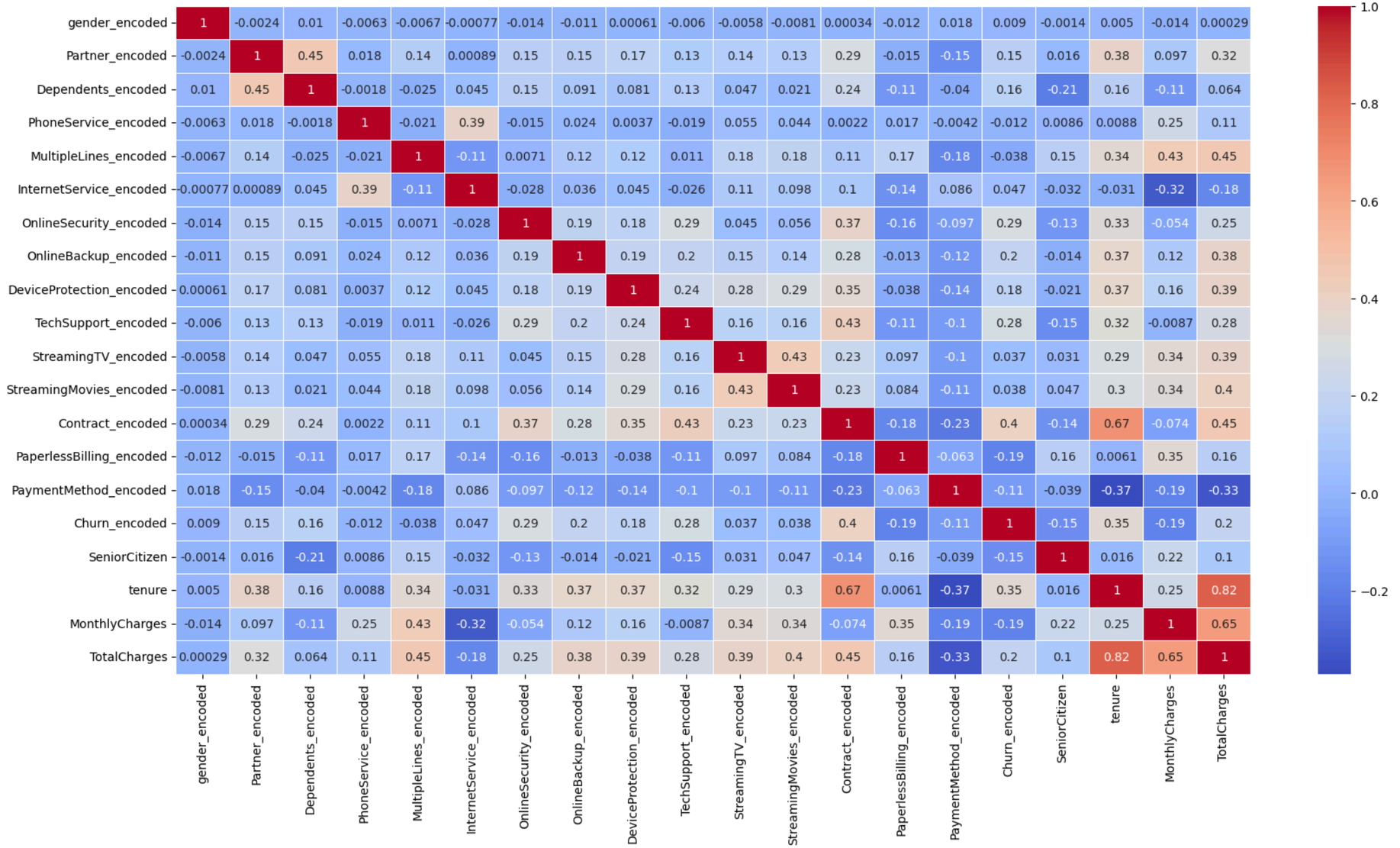
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	Paym
0	7590-VHVEG	F	0	Yes	No	1.00	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	
1	5575-GNVDE	Male	0	No	No	34.00	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	M
2	3668-QPYBK	Male	0	No	No	2.00	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	M
3	7795-CFOCW	Male	0	No	No	45.00	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	B
4	9237-HQITU	Female	0	No	No	2.00	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	

After

```
df_transformed
```

	gender_encoded	Partner_encoded	Dependents_encoded	PhoneService_encoded	MultipleLines_encoded	InternetService_encoded	OnlineSecurity_encoded	OnlineBackup_encoded	DeviceProtection_encoded	TechSupport_encoded	StreamingTV_
0	0	1	0	0	1	0	0	2	0	0	
1	1	0	0	1	0	0	2	0	2	0	
2	1	0	0	1	0	0	2	2	0	0	
3	1	0	0	0	1	0	2	0	2	2	
4	0	0	0	1	0	1	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	
7038	1	1	1	1	2	0	2	0	2	2	
7039	0	1	1	1	2	1	0	2	2	0	

KORELASI



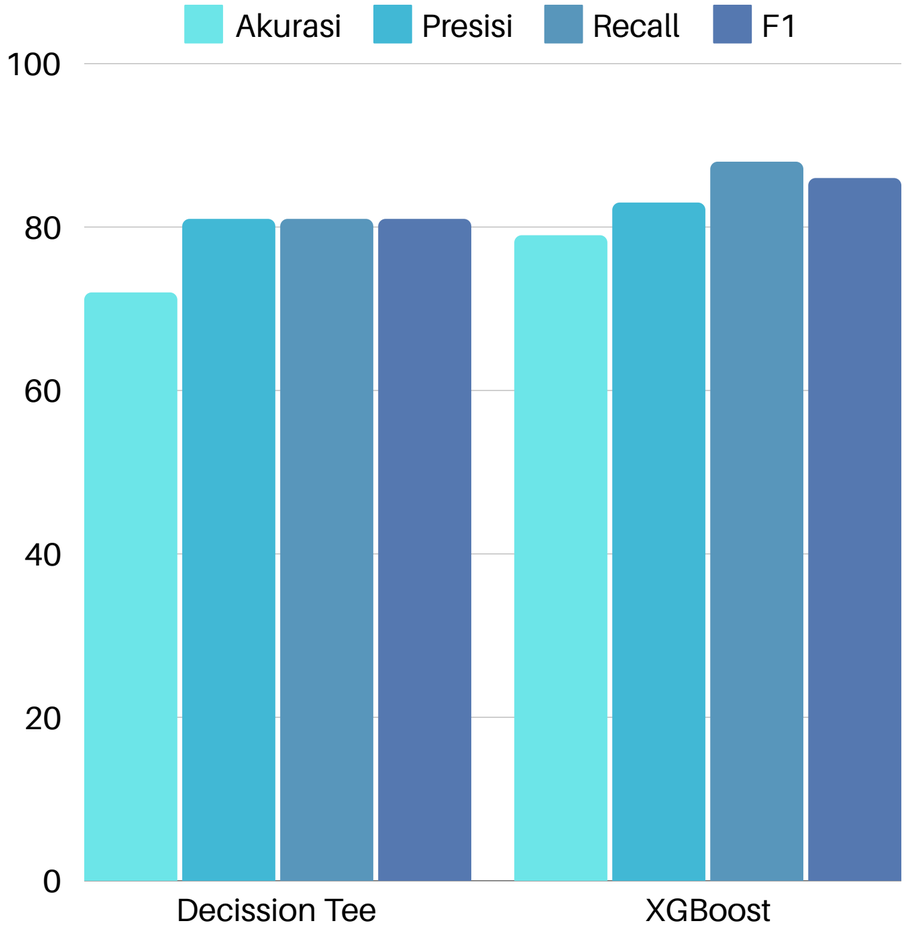
	Churn_encoded
Churn_encoded	1.00
Contract_encoded	0.40
tenure	0.35
OnlineSecurity_encoded	0.29
TechSupport_encoded	0.28
TotalCharges	0.20
OnlineBackup_encoded	0.20
MonthlyCharges	0.19
PaperlessBilling_encoded	0.19
DeviceProtection_encoded	0.18
Dependents_encoded	0.16
SeniorCitizen	0.15
Partner_encoded	0.15
PaymentMethod_encoded	0.11
InternetService_encoded	0.05
StreamingMovies_encoded	0.04
MultipleLines_encoded	0.04
StreamingTV_encoded	0.04
PhoneService_encoded	0.01
gender_encoded	0.01

dtype: float64

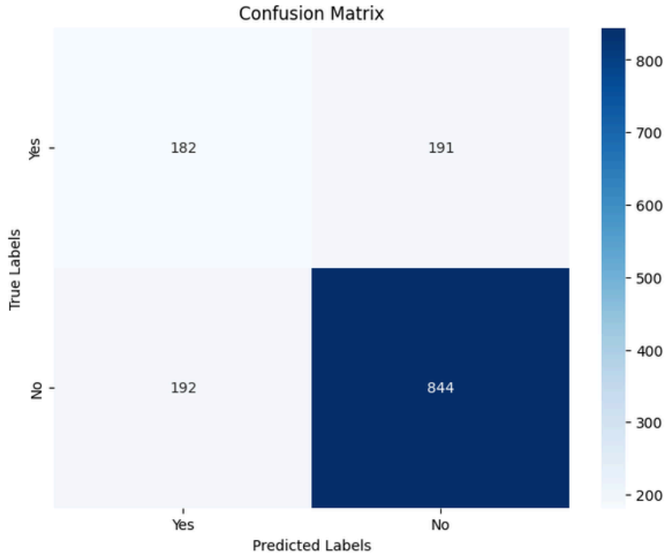
variable yang dipakai hanya variable yang memiliki nilai korelasi >0.15 dengan target variable

MODEL

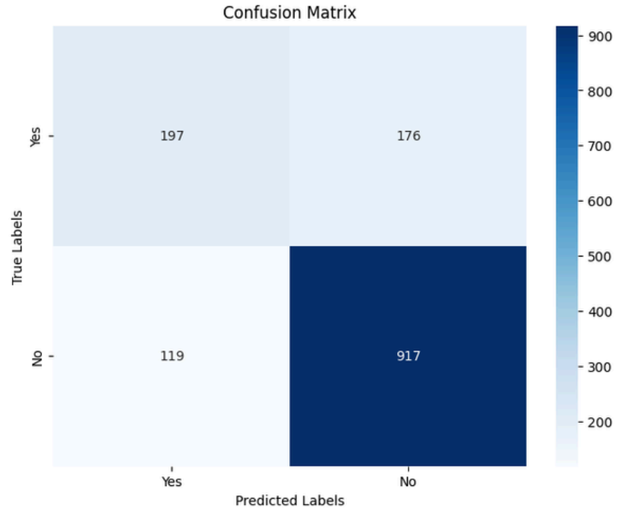
Metrik Evaluasi



Decission Tree



XGBoost



MODEL

Features Importance

Decission Tree

	Feature	Importance
6	MonthlyCharges	0.27
4	TotalCharges	0.23
0	Contract_encoded	0.17
1	tenure	0.12
2	OnlineSecurity_encoded	0.05
11	Partner_encoded	0.03
7	PaperlessBilling_encoded	0.03
9	Dependents_encoded	0.03
10	SeniorCitizen	0.02
3	TechSupport_encoded	0.02
5	OnlineBackup_encoded	0.02
8	DeviceProtection_encoded	0.02

XGBoost

	Feature	Importances
0	Contract_encoded	0.52
2	OnlineSecurity_encoded	0.12
3	TechSupport_encoded	0.07
1	tenure	0.04
6	MonthlyCharges	0.04
7	PaperlessBilling_encoded	0.04
10	SeniorCitizen	0.03
5	OnlineBackup_encoded	0.03
4	TotalCharges	0.03
8	DeviceProtection_encoded	0.03
9	Dependents_encoded	0.03
11	Partner_encoded	0.02

Dari kedua model tersebut, bisa disimpulkan bahwa variable Contract memiliki pengaruh paling besar dan disusul oleh variabel tenure yang meiliki pengaruh sedang.

TESTING

Diketahui data customer baru sebut saja Asep

Contract_encoded = 'Two year'
tenure = 10
OnlineSecurity_encoded = 'Yes'
TechSupport_encoded = 'Yes'
TotalCharges = 100
OnlineBackup_encoded = "Yes"
MonthlyCharges = 80
PaperlessBilling_encoded = 'Yes'
DeviceProtection_encoded = "Yes"
Dependents_encoded = "Yes"
SeniorCitizen = 1
Partner_encoded = "Yes"

Hasil test pada rill menggunakan model Decission Tree

```
# Lakukan prediksi pada data baru
new_predictions = model.predict(new_df)

# Jika Anda ingin melihat hasil prediksi
print("Hasil Prediksi untuk data baru:")
if new_predictions[0] == 1:
  print("Pelanggan akan churn")
else:
  print("Pelanggan tidak akan churn")
```

```
Hasil Prediksi untuk data baru:
Pelanggan akan churn
```

Hasil test pada rill menggunakan model XGBoost

```
# Lakukan prediksi pada data baru
new_predictions = model.predict(new_df)

# Jika Anda ingin melihat hasil prediksi
print("Hasil Prediksi untuk data baru:")
if new_predictions[0] == 1:
  print("Pelanggan akan churn")
else:
  print("Pelanggan tidak akan churn")
```

```
Hasil Prediksi untuk data baru:
Pelanggan akan churn
```

Thank you very much!

www.colab.research.google.com