

# Telco Project

Using Decision Tree Machine Learning Models

# About Telco Dataset

## Context

"Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs." [IBM Sample Data Sets]

## Content

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

### The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Source : ([https://www.kaggle.com/datasets/blastchar/telco-customer-churn?select=WA\\_Fn-UseC\\_-Telco-Customer-Churn.csv](https://www.kaggle.com/datasets/blastchar/telco-customer-churn?select=WA_Fn-UseC_-Telco-Customer-Churn.csv))

# Telco Project

1. Exploratory Data Analysis
2. Feature Engineering
3. Data Modelling
4. Confusion Matrix
5. Conclusion



# 1. Exploratory Data Analysis

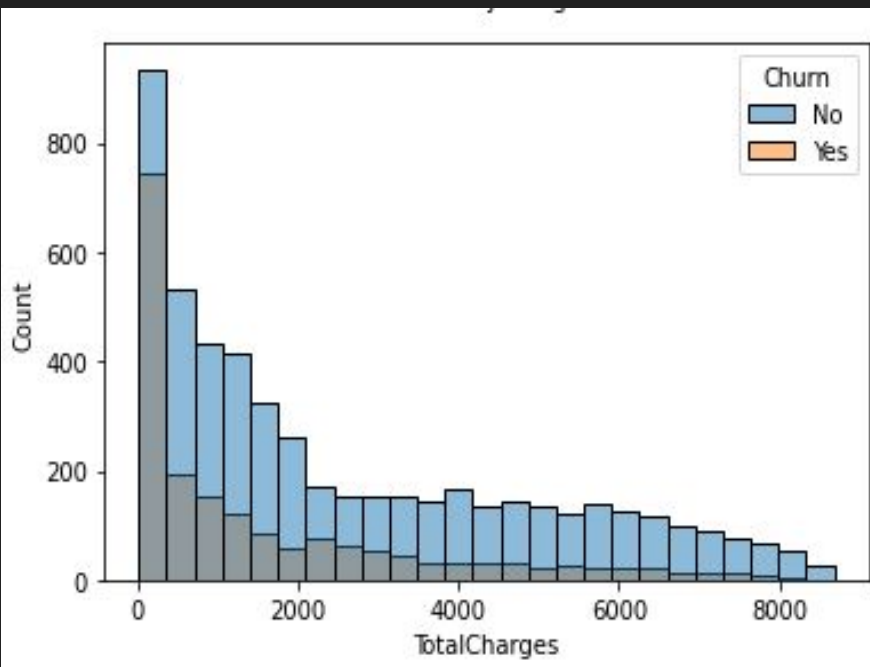
Dataset Telco berisi 7043 baris dan 21 kolom, yang berisi informasi tentang data pelanggan telco. Referensi dari dataset ini bersumber dari kaggle. Selanjutnya dataset ini akan dilakukan analisa machine learning dengan menggunakan metode regresi decision tree. Tujuan dari analisa ini adalah membuat prediksi pelanggan telco akan churn atau tidak. Bagaimana keputusan bisnis untuk mengatasi pelanggan yang akan churn.

```
df.shape
```

```
(7043, 21)
```

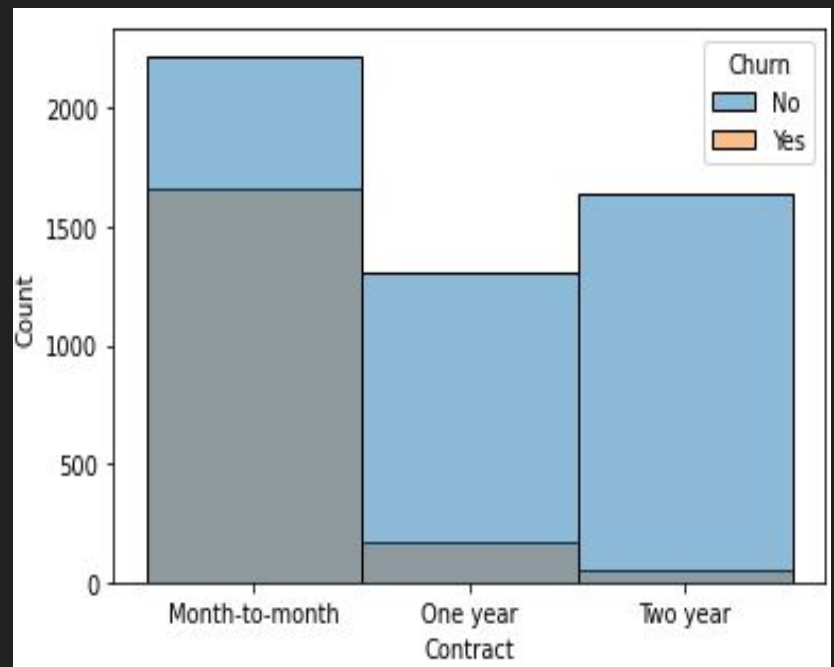
```
df.describe()
```

	SeniorCitizen	tenure	MonthlyCharges
count	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692
std	0.368612	24.559481	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	72.000000	118.750000



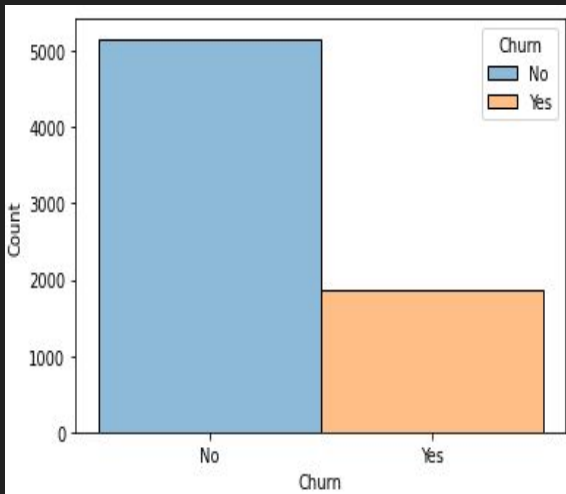
## Total Charges

Pelanggan dengan Total Charges rendah cenderung memiliki churn yang tinggi, sedangkan pelanggan dengan Total Charges tinggi cenderung memiliki churn yang rendah



## Contract

Pelanggan dengan kontrak yang pendek cenderung memiliki churn yang tinggi, sedangkan pelanggan dengan kontrak yang panjang cenderung memiliki churn yang rendah.

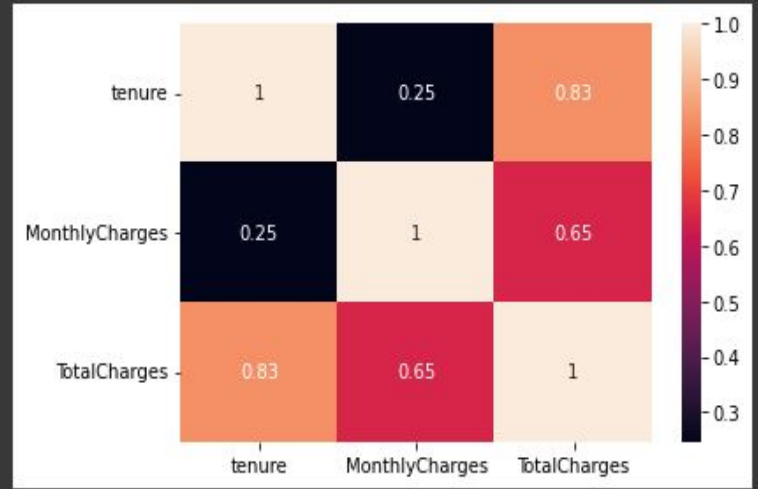


```
count of Churn
No    5163
Yes   1869
Name: Churn, dtype: int64

% of Churn
No    73.421502
Yes   26.578498
Name: Churn, dtype: float64
```

## Churn VS Not Churn

Berdasarkan grafik dan total pelanggan yang churn dan tidak churn, maka dapat diambil kesimpulan pelanggan yang tidak churn lebih banyak dibanding pelanggan yang churn, artinya bisnis tersebut masih berjalan dengan baik.

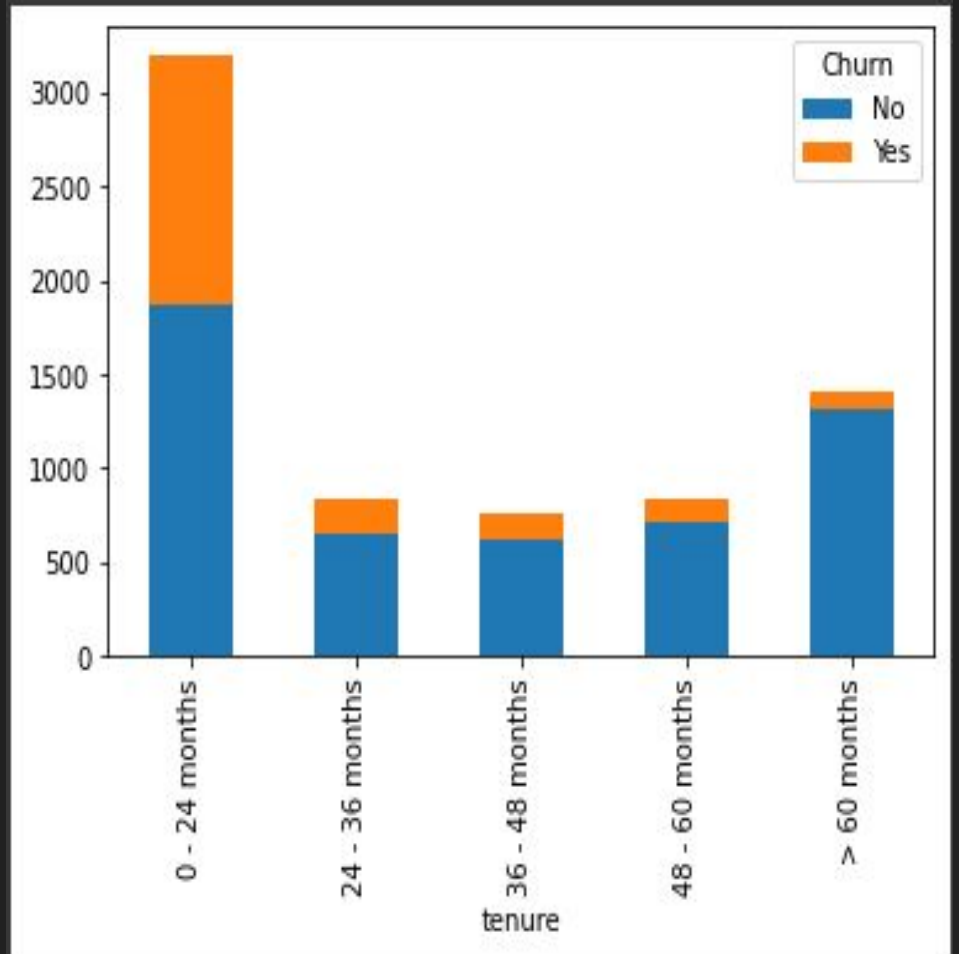


## Correlation

Berdasarkan grafik heatmap maka dapat diambil kesimpulan bahwa terdapat korelasi yang cukup kuat diantara Total Charges dan Tenure

## 2. Feature Engineering

Pada tahap ini tenure dikelompokkan menjadi beberapa kategori berdasarkan periodenya. Tenure dengan periode awal memiliki churn yang lebih tinggi dibandingkan tenure dengan periode yang lama.



### 3. Data Modelling

Data dimodelkan dengan model machine learning regresi decision tree, karena hasil yang diharapkan untuk memprediksi apakah pelanggan akan churn atau tidak, regresi decision tree adalah model yang cocok untuk memodelkan data tersebut.

Model regresi decision tree menghasilkan accuracy score 79 %. Kesimpulannya model dari regresi decision tree ini cukup bagus untuk memodelkan dan memprediksi data.

```
from sklearn.tree import DecisionTreeClassifier
# instantiate the DecisionTreeClassifier model with criterion gini index

clf_entropy = DecisionTreeClassifier(criterion='entropy', max_depth=3, random_state=0)

# fit the model
clf_entropy.fit(X_train, y_train)
```

	precision	recall	f1-score	support
0	0.91	0.82	0.86	1139
1	0.46	0.65	0.54	268
accuracy			0.79	1407
macro avg	0.69	0.73	0.70	1407
weighted avg	0.82	0.79	0.80	1407

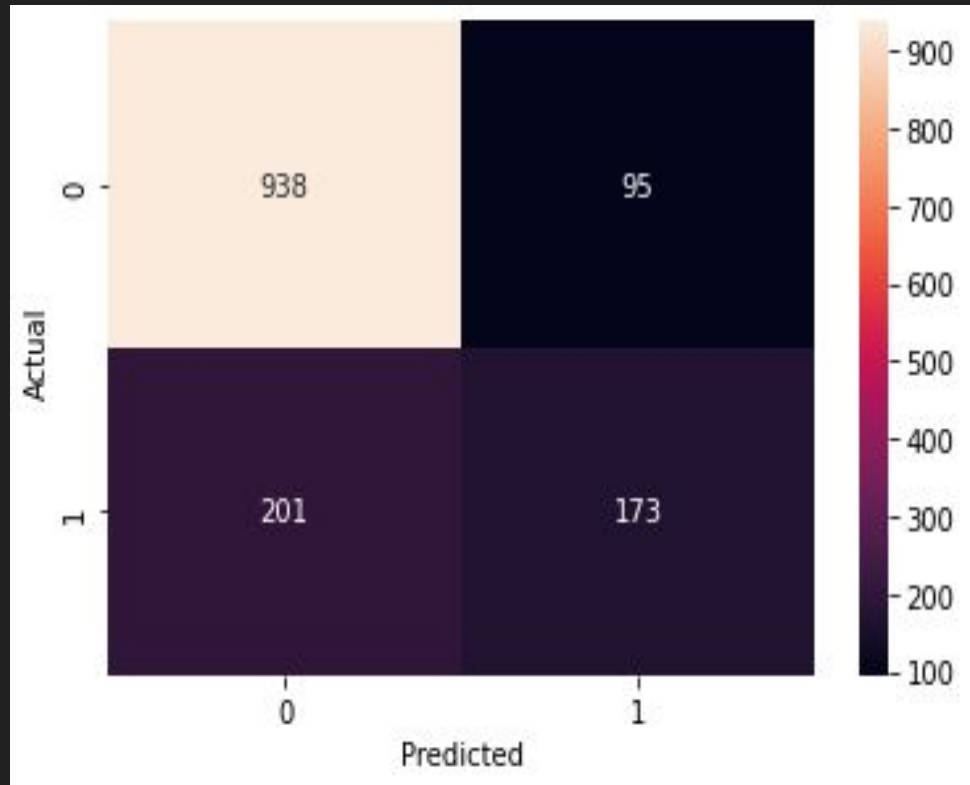


## 4. Confusion Matrix

Berdasarkan hasil dari confusion matrix ketika model regresi decision tree memprediksi churn dan nilai aktual churn (True Positif) terdapat 173 pelanggan. Ketika model regresi decision tree memprediksi tidak churn dan nilai aktual tidak churn (True Negatif) terdapat 938 pelanggan.

Namun jika model memprediksi churn tetapi aktualnya tidak churn terdapat 95 pelanggan, dan ketika model memprediksi tidak churn tetapi aktualnya churn terdapat 201 pelanggan.

Jumlah model prediksi yang sesuai dengan aktualnya jauh lebih banyak dibandingkan dengan model prediksi yang tidak sesuai dengan aktualnya, maka dapat disimpulkan bahwa model cukup baik.



## 5. Conclusion

Ketika model regresi decision tree memprediksi churn dan nilai aktual churn terdapat 173 pelanggan, untuk pelanggan yang akan churn dapat ditawarkan promosi paket yang bisa membuat pelanggan tetap bertahan untuk menggunakan produk ini sebagai bentuk strategi marketing untuk mempertahankan pelanggan.

Ketika model memprediksi churn tetapi aktualnya tidak churn terdapat 95 pelanggan, artinya cost marketing untuk menarik pelanggan ini untuk tetap bertahan tidak tepat sasaran.



# Aldilah Ariwibowo

**Linkedin**

**[www.linkedin.com/in/aldilah-ariwibowo](https://www.linkedin.com/in/aldilah-ariwibowo)**

**Github**

**<https://github.com/aldilahariwibowo>**