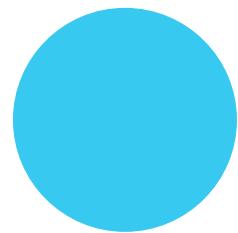


Dataset and Modeling: Telco Customer Churn

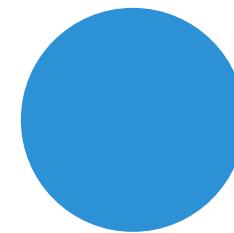
Final Project - Data Science Batch 11

Anggota Kelompok 6

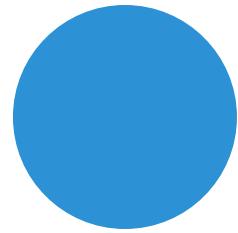
(Coldplay)



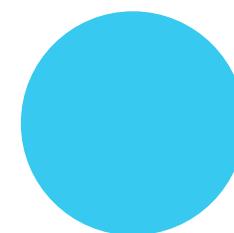
Adel Nor Muhammad



Muhammad Ali Akbar



Annisa Rahma Melyanta



Zulfi Jauharul Ikhsan

Agenda

Stage 1 - Understanding The Dataset

Stage 2 - Identify Which Activities Should be Done

Stage 3 - Analyze The Data Using Exploratory and Visualization

Stage 4 - Data Preprocessing

Stage 5 - Develop Model & Evaluation



Stage 1: Understanding The Dataset

Dataset



BLASTCHAR · UPDATED 4 YEARS AGO

▲ 2036

New Notebook

Download (176 kB)



Telco Customer Churn

Focused customer retention programs

[Data](#) [Code \(832\)](#) [Discussion \(15\)](#) [Metadata](#)

About Dataset

Context

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

Usability ⓘ

8.82

License

Data files © Original Authors

Expected update frequency

Not specified

<https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

Latar Belakang Dataset

- Telco adalah perusahaan yang menyediakan layanan telekomunikasi.
- Dataset berisi mengenai perilaku customer Telco baik yang masih berlangganan maupun yang sudah berhenti.
- Dataset terdiri dari 7043 baris dan 21 kolom yang mencakup:
 1. Pelanggan yang Churn dalam sebulan terakhir
 2. Layanan Telco yang dipilih
 3. Akun pelanggan
 4. Demografi pelanggan



Tujuan Dataset

- Dengan menganalisis dataset ini diharapkan dapat memprediksi perilaku pelanggan baru berdasarkan data yang sudah ada.
- Tujuan dari mengetahui perilaku ini adalah untuk dapat mengembangkan program retensi (program mempertahankan pelanggan).



Deskripsi Fitur Data

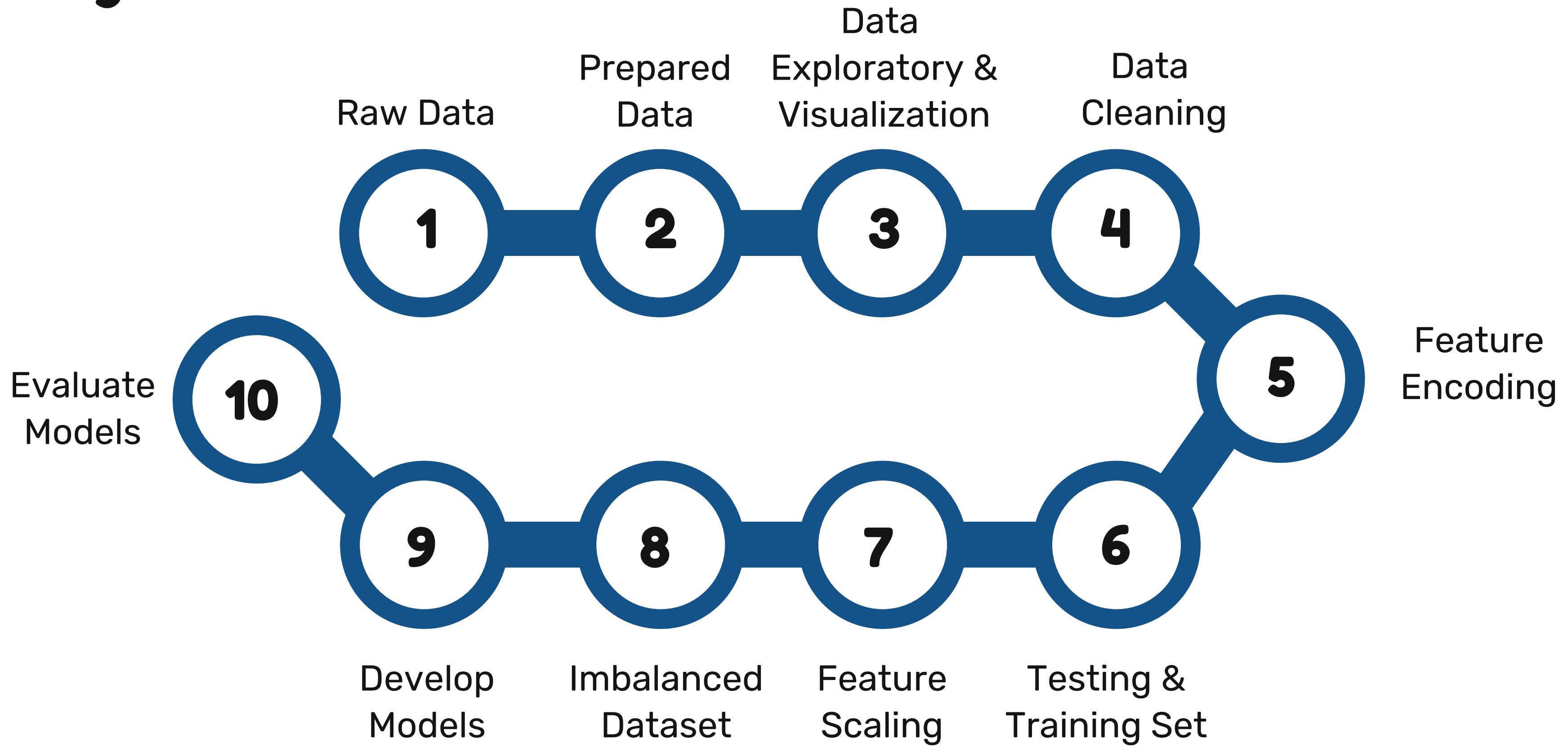
customerID	object	Menunjukkan ID yang membedakan setiap pelanggan
Gender	object	Membedakan pelanggan berdasarkan jenis kelamin
SeniorCitizen	int	Membedakan pelanggan berdasarkan kelompok usia
Partner	object	Menunjukkan pelanggan memiliki pasangan atau tidak
Dependents	object	Menunjukkan pelanggan memiliki tanggungan (anak) atau tidak
Tenure	int	Menunjukkan masa berlangganan yang telah dilewati dalam hitungan bulan
phone service	object	Menunjukkan apakah pelanggan memiliki layanan phone service atau tidak
Multiple Line	object	Menunjukkan apakah pelanggan menggunakan layanan multiple line atau tidak
Internet Services	object	Menunjukkan jenis layanan internet yang dimiliki customer
Online Security	object	Menunjukkan apakah pelanggan menggunakan layanan Online Security atau tidak

DeviceProtection	object	Menunjukan apakah pelanggan menggunakan layanan Device Protection
OnlineBackup	object	Menunjukan apakah pelanggan menggunakan layanan Online Backup atau tidak
TechSupport	object	Menunjukan apakah pelanggan menggunakan layanan Technical Support atau tidak
Streamin TV	object	Menunjukan apakah pelanggan menggunakan layanan Streaming TV atau tidak
Streaming Movie	object	Menunjukan apakah pelanggan menggunakan layanan Streaming Movie atau tidak
Contract	object	Menunjukan jenis jangka waktu kontrak yang dipilih oleh pelanggan
Paperless Billing	object	Menunjukan jenis tagihan yang digunakan oleh pelanggan
Payment Method	object	Menunjukkan metode pembayaran yang digunakan oleh pelanggan
Monthly Charges	float	Menunjukkan tagihan bulanan yang perlu dibayar oleh setiap pelanggan
Total Charges	object	Menunjukkan jumlah total tagihan yang sudah dibebankan kepada pelanggan
Churn	object	Untuk mengetahui customer tersebut churn atau tidak



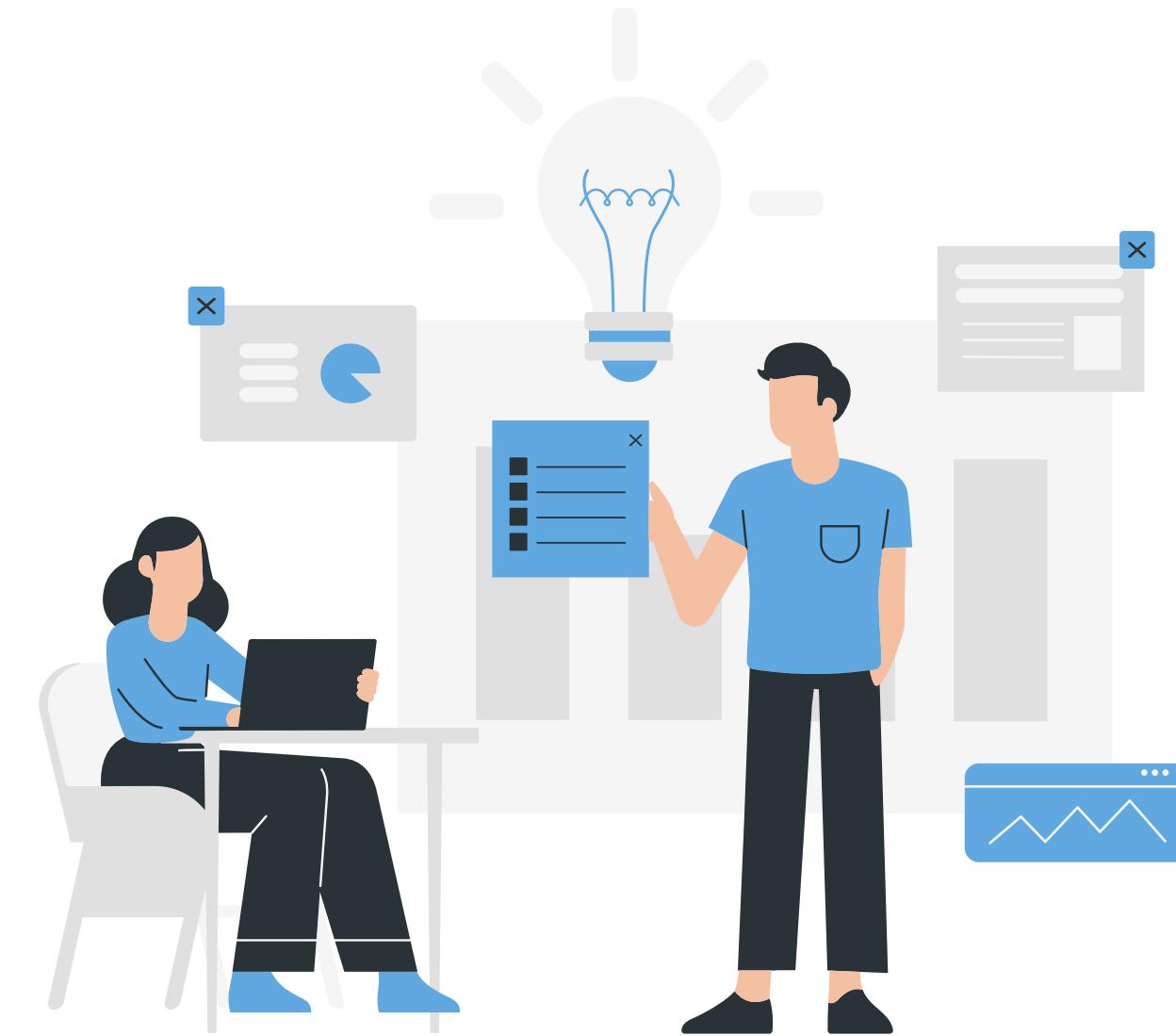
Stage 2 : Identify Which Activities Should be Done

Kegiatan



Objective

- Dari dataset yang dimiliki, akan dibuat model **Churn Prediction** dan melakukan **Customer Segmentation** untuk mencari tahu alasan kenapa pelanggan tersebut churn.
- Kegiatan tersebut dilakukan untuk membantu perusahaan untuk membuat tindakan churn prevention.





Stage 3 : Analyze The Data Using Exploratory and Visualization

Data Exploratory

Dataset

7043

18

0

Baris

21

3

0

Data
Kategorikal

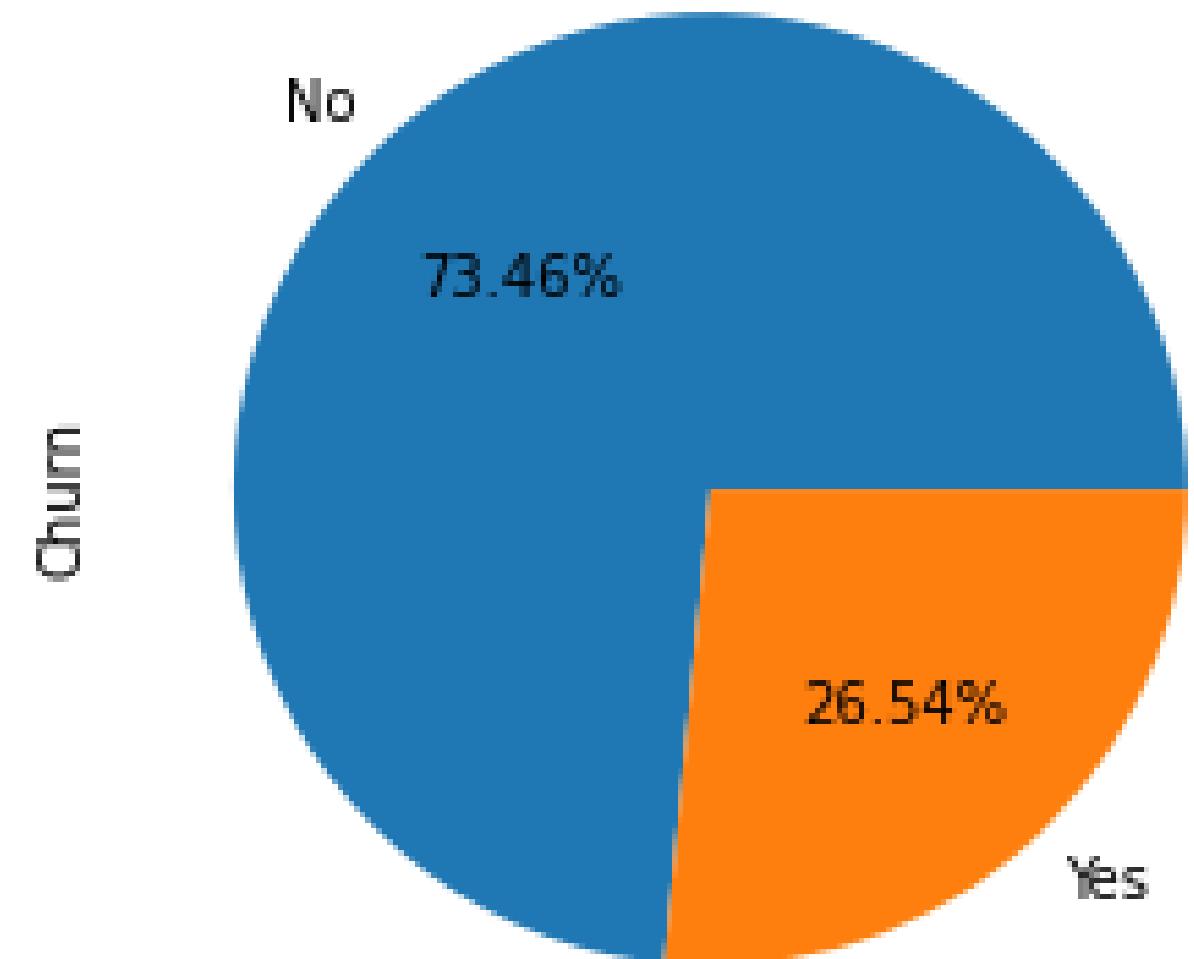
Missing
Values

Kolom

Data
Numerik

Data
Duplicates

Data Visualization



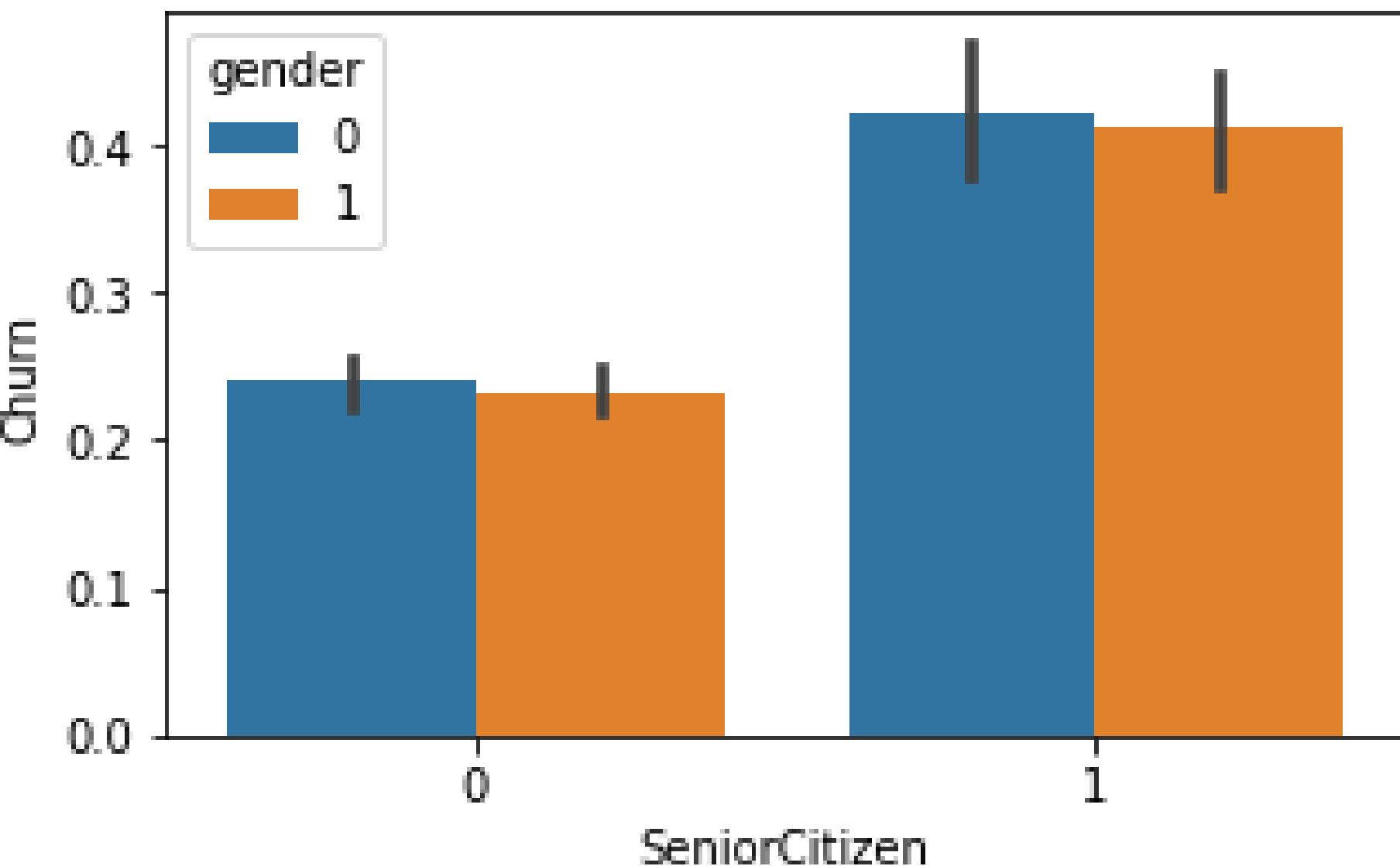
Frekuensi Data Churn

Sebanyak **73.5%** pelanggan masih berlangganan, sedangkan **26.5%** pelanggan sudah tidak berlangganan. Dimana jumlah pelanggan yang churn adalah **1869** data point dan pelanggan yang tidak churn adalah **5174** data point.

Data Kategorikal

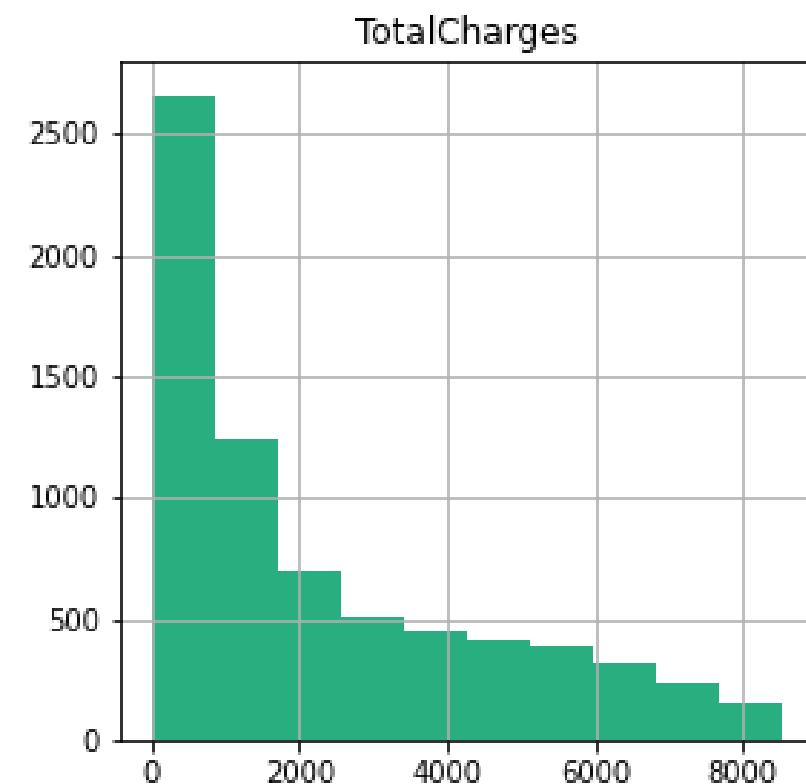
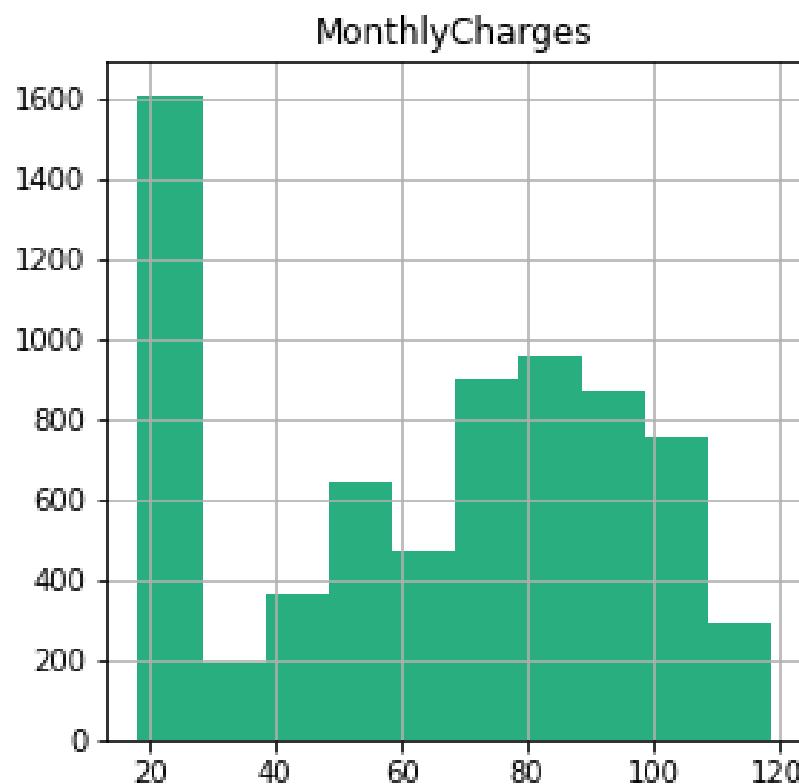
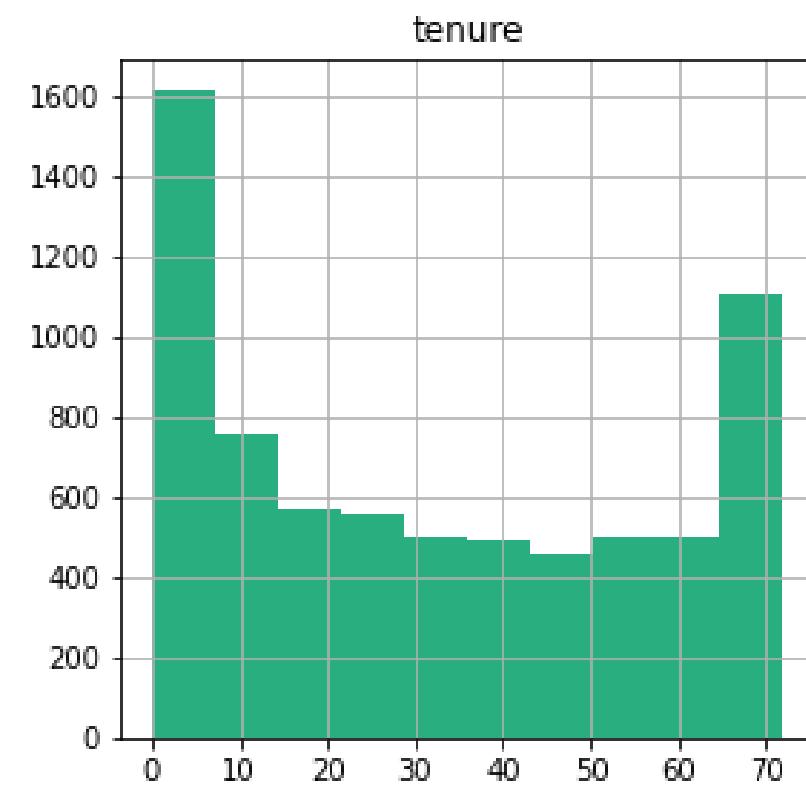
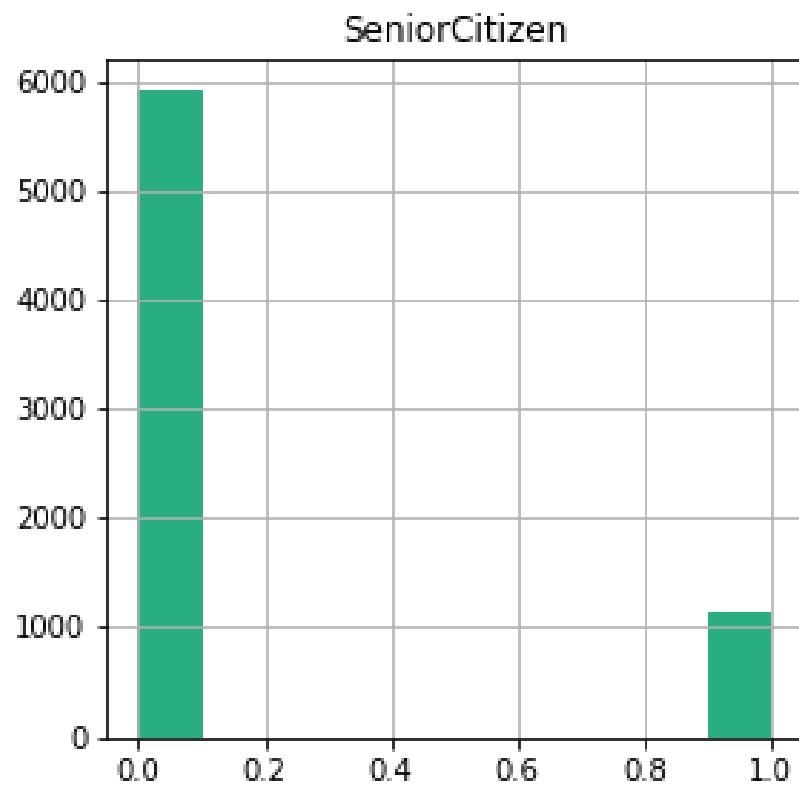
- Jika melihat customer behavior dari jenis kelamin dan usia berdasarkan data yang dimiliki adalah:

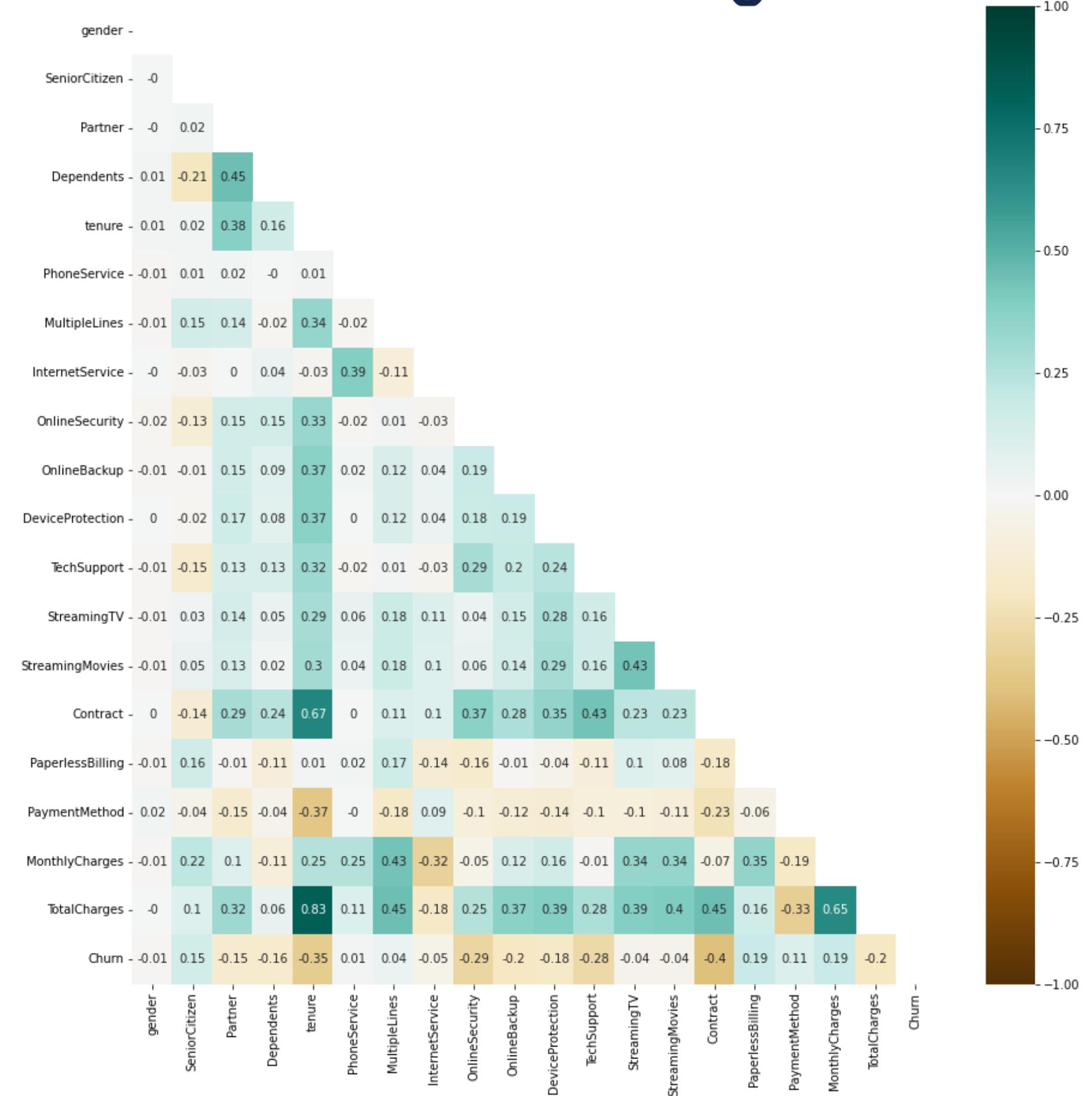
Pelanggan yang lanjut usia (lansia) dan berjenis kelamin perempuan memiliki potensi churn yang lebih tinggi.



Distribusi Data

Dapat dilihat bahwa data terdistribusi secara tidak normal baik untuk Tenure, MonthlyCharges, maupun TotalCharges.





Heatmap

- Tenure dan TotalCharges** memiliki korelasi yang kuat (positif).
- Tenure dan Contract** memiliki korelasi yang kuat (positif).
- TotalCharges dan MonthlyCharges** memiliki korelasi yang kuat (positif).



Stage 4 : Data Preprocessing

Data Cleaning

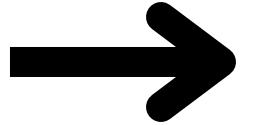
- Untuk membuat model yang baik, dataset harus sudah bersih dari anomali.
Untuk mengidentifikasi anomali tersebut akan dilakukan data exploratory dan visualisasi, sehingga lebih mudah untuk melihat sebaran data.
- Adapun beberapa langkah pengecekan dalam pengolahan data yang akan dilakukan adalah:
 1. Melakukan Identifikasi pada fitur data yang missing values
 2. Melakukan identifikasi pada fitur data yang duplicate
 3. Melakukan identifikasi pada fitur data yang inconsistency
 4. Melakukan identifikasi pada fitur data yang outliers

Setelah dilakukan pengecekan, terdapat beberapa anomali yang terdeteksi, yaitu:

- Terdapat blank string (**missing values**) pada kolom **TotalCharges**, sebab **tenure** belum lebih dari 1 bulan.

tenure	MonthlyCharges	TotalCharges
488	0	52.55
753	0	20.25
936	0	80.85
1082	0	25.75
1340	0	56.05
3331	0	19.85
3826	0	25.35
4380	0	20.00
5218	0	19.70
6670	0	73.35
6754	0	61.90

Blank String



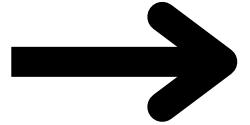
tenure	MonthlyCharges	TotalCharges
488	0	52.55
753	0	20.25
936	0	80.85
1082	0	25.75
1340	0	56.05
3331	0	19.85
3826	0	25.35
4380	0	20.00
5218	0	19.70
6670	0	73.35
6754	0	61.90

Sehingga kolom **TotalCharges** perlu diubah dengan mengkalikan **tenure** dan **MonthlyCharges**.

- Nilai pada kolom TotalCharges tidak sesuai (**inconsistent values**), yang mana seharusnya merupakan hasil dari **tenure x MonthlyCharges**.

	tenure	MonthlyCharges	TotalCharges
0	1	29.85	29.85
1	34	56.95	1889.5
2	2	53.85	108.15
3	45	42.30	1840.75
4	2	70.70	151.65
5	8	99.65	820.5
6	22	89.10	1949.4
7	10	29.75	301.9
8	28	104.80	3046.05
9	62	56.15	3487.95
10	13	49.95	587.45

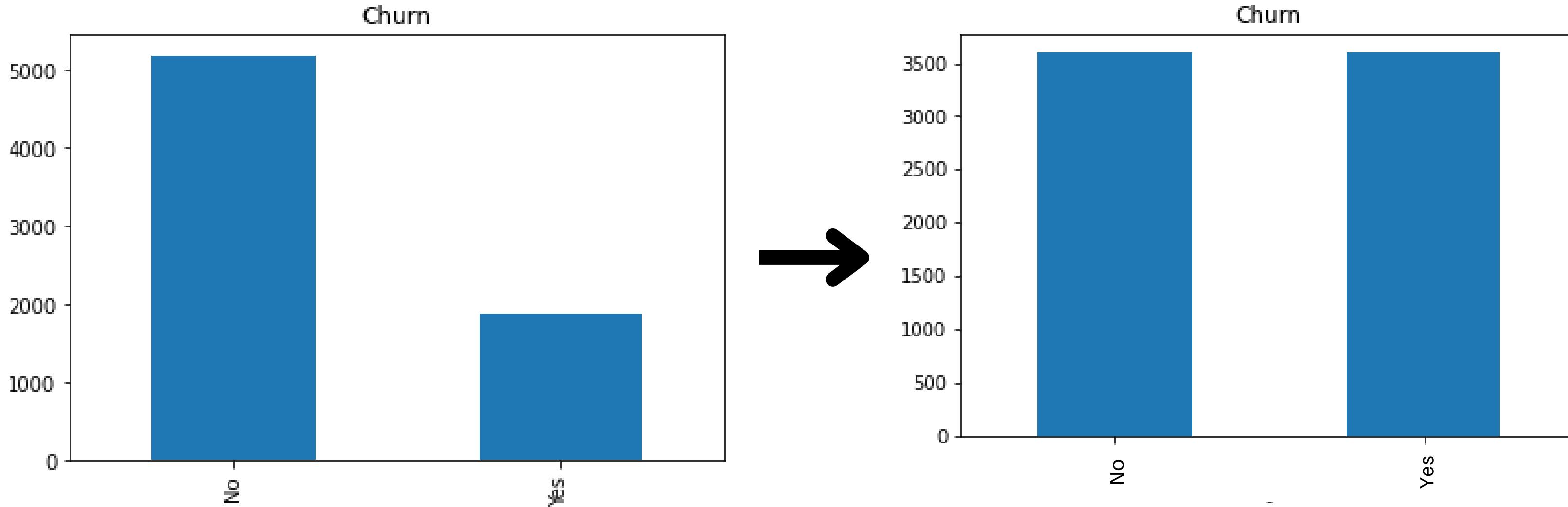
False Values



	tenure	MonthlyCharges	TotalCharges
0	1	29.85	29.85
1	34	56.95	1936.30
2	2	53.85	107.70
3	45	42.30	1903.50
4	2	70.70	141.40
5	8	99.65	797.20
6	22	89.10	1960.20
7	10	29.75	297.50
8	28	104.80	2934.40
9	62	56.15	3481.30
10	13	49.95	649.35

Sehingga kolom **TotalCharges** perlu diubah dengan mengkalikan **tenure** dan **MonthlyCharges**.

- Distribusi data **Churn** tidak seimbang (**Imbalanced Dataset**).



Dataset ini memiliki jumlah data class yang tidak seimbang, dengan jumlah **Churn 1869 data point**, sedangkan **No Churn 5174 data point**. Dalam mengatasi imbalanced dataset, dapat dilakukan resampling data pada data training.

Feature Encoding

Feature encoding dapat digunakan untuk melakukan transformasi dari fitur kategorikal menjadi variabel numerik. Terdapat 2 jenis feature encoding yang digunakan, yaitu:

- Label Encoding

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	
0	0	0	1	0	1	0	1	0	0
1	1	0	0	0	34	1	0	0	0
2	1	0	0	0	2	1	0	0	0
3	1	0	0	0	45	0	1	0	0
4	0	0	0	0	2	1	0	1	0

- One Hot Encoding

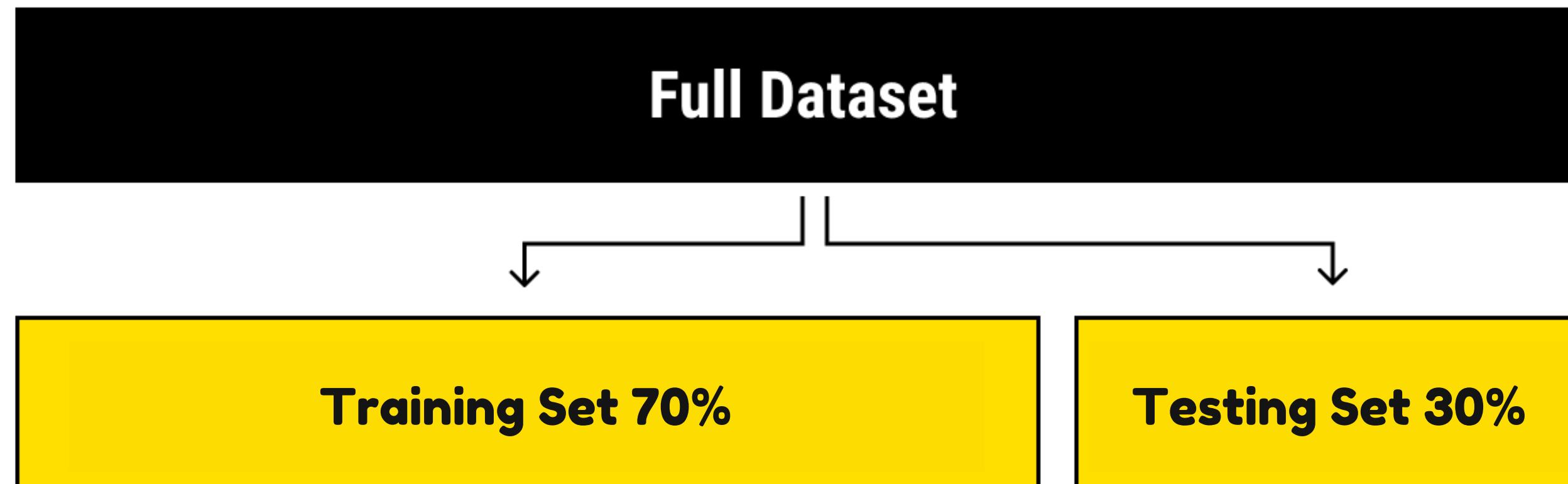
	gender_Male	Partner_Yes	Dependents_Yes	PhoneService_Yes	MultipleLines_No phone service	MultipleLines_Yes	
0	1	0	0	0	1	0	0
1	0	1	0	1	0	1	0
1	0	0	1	1	0	0	1
1	0	0	0	0	1	0	0
0	0	0	0	1	0	0	0

Testing & Training Set

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

Komposisi:

Training set 70% dan Testing set 30% dari keseluruhan data.



Feature Scaling

Feature scaling merupakan suatu cara untuk fitur numerikal pada dataset agar memiliki rentang nilai yang sama.

- Standardization

Membuat data menjadi mengikuti standar normal distribution dengan mean = 0.

- Normalization

Menskalakan data yang nilai minimum dan maksimum masing-masing menjadi 0 dan 1.

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

```
scaler = MinMaxScaler().fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

Imbalanced Dataset

Imbalanced dataset merupakan keadaan dimana dataset tidak dalam keadaan data yang seimbang, jumlah kelas data yang satu lebih sedikit dibandingkan dengan jumlah kelas data lainnya.

- Under Sampling

```
sebelum undersampling
Churn_Yes
0.0      3589
1.0      1341
dtype: int64
setelah undersampling
Churn_Yes
0.0      1341
1.0      1341
dtype: int64
```

- Over Sampling

```
sebelum oversampling
Churn_Yes
0.0      3589
1.0      1341
dtype: int64
setelah oversampling
Churn_Yes
0.0      3589
1.0      3589
dtype: int64
```

- SMOTE

```
sebelum smote
Churn_Yes
0.0      3589
1.0      1341
dtype: int64
setelah smote
Churn_Yes
0.0      3589
1.0      3589
dtype: int64
```

- ADASYN

```
sebelum adasyn
0.0      3589
1.0      1341
dtype: int64
setelah adasyn
1.0      3617
0.0      3589
dtype: int64
```



Stage 5 : Develop Model & Evaluation

Modelling I: Churn Prediction

KNN

Decision
Tree

Random
Forest

AdaBoost

MLP

Training Set

70% dari keseluruhan data

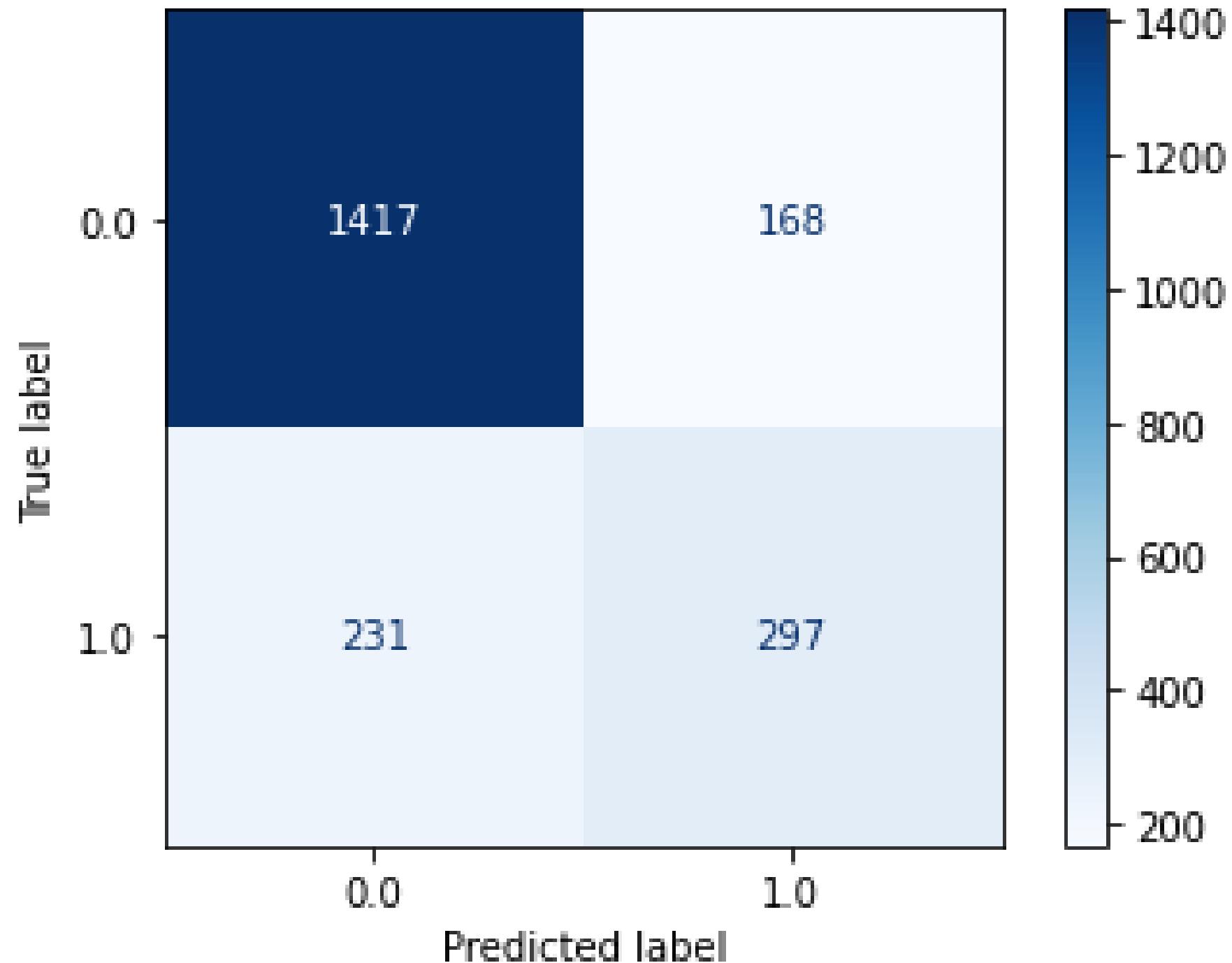
Testing Set

30% dari keseluruhan data

Performance Comparison (%)

	Model	Label Encoder			One Hot		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall
Random Under Sampling	KNN	79.18	72.13	69.07	78.47	71.09	69.67
	DT	73.83	65.97	67.27	73.78	65.95	67.30
	RF	80.41	73.91	71.72	79.98	73.27	71.31
	AdaBoost	80.64	74.20	72.64	81.12	74.93	72.83
	MLP	78.28	71.11	71.50	77.24	69.77	70.24
Random Over Sampling	KNN	72.31	66.68	70.55	71.88	66.50	70.52
	DT	73.02	64.85	65.91	73.49	65.68	67.11
	RF	78.65	71.83	73.26	78.41	71.58	73.17
	AdaBoost	74.68	71	77.18	74.82	71.20	77.47
	MLP	78.08	71.07	72.32	71.74	69.52	75.79
SMOTE	KNN	73.73	67.85	71.69	73.78	67.78	71.72
	DT	72.26	64.99	67.17	74.11	66.52	68.15
	RF	78.93	72.24	73.96	79.17	72.18	71.84
	AdaBoost	76.99	71.68	76.64	79.93	72.66	75.41
	MLP	77.66	71.67	75.57	80.40	74.23	69.76
ADASYN	KNN	71.42	66.72	71.22	70.75	66.20	70.65
	DT	72.08	65.14	67.68	74.78	66.95	68.03
	RF	78.09	71.38	73.59	78.80	71.71	71.60
	AdaBoost	76.01	70.76	75.67	79.13	72.47	74.15
	MLP	74.87	70.46	76.05	77.61	71.69	75.73

Kesimpulan



Setelah dilakukan perbandingan antar model prediksi, diketahui model paling baik adalah Ada Boost Classifier dengan menggunakan One Hot Encoder.

Hasil prediksi adalah:

Akurasi: 81.12%

Presisi: 74.93%

Recall: 72.83%

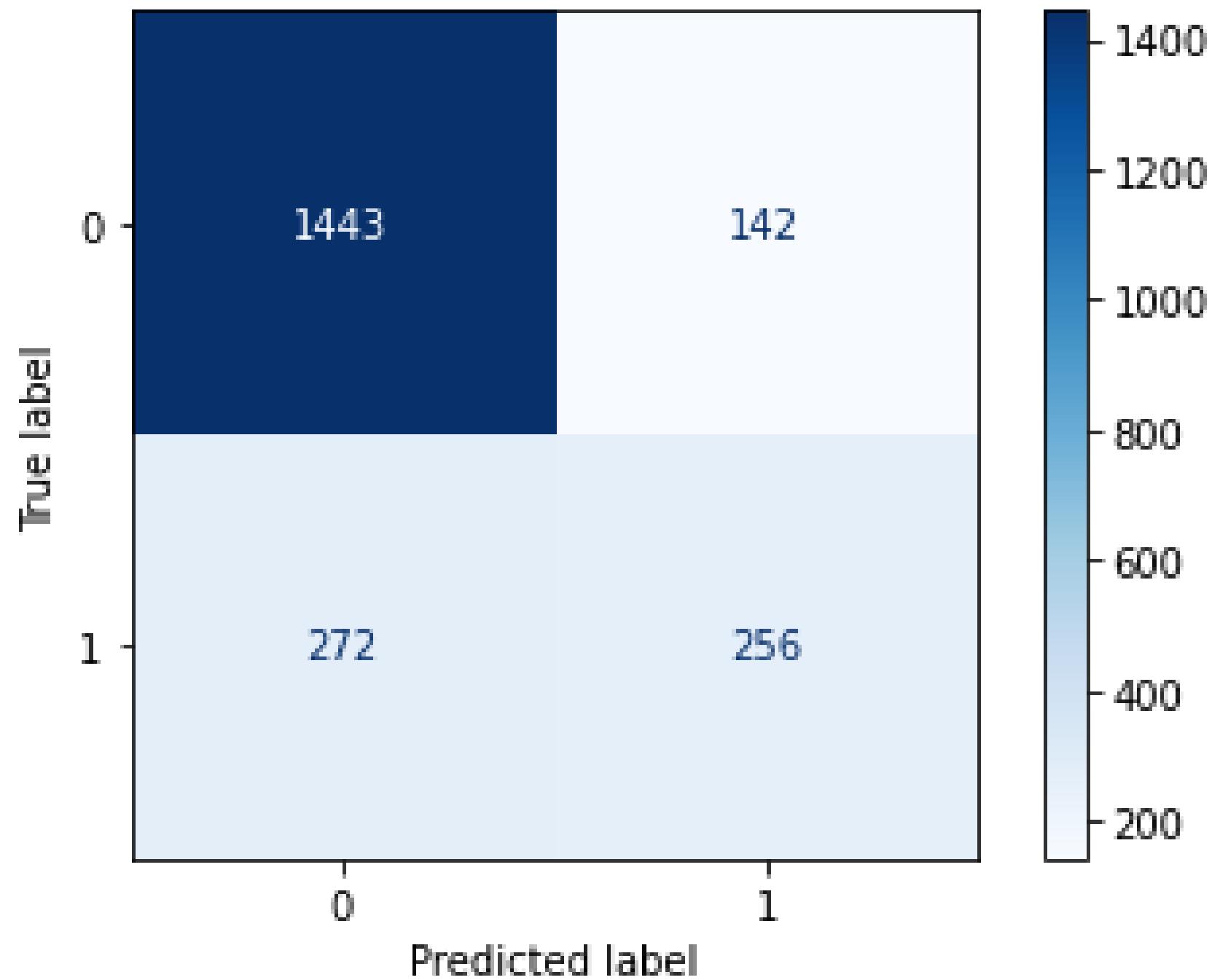
Dengan Hasil

1417 True No Churn / True Positive

168 False No Churn / False Positive

231 False Churn / False Negative

297 True Churn / True Negative



Model prediksi terbaik ke 2 adalah Multi Layer Perceptron dengan menggunakan One Hot Encoder.

Hasil prediksi adalah:

Akurasi: 80.40%

Presisi: 74.23%

Recall: 69.76%

Dengan Hasil

1443 True No Churn / True Positive

142 False No Churn / False Positive

272 False Churn / False Negative

256 True Churn / True Negative

Modeling II: Customer Segmentation

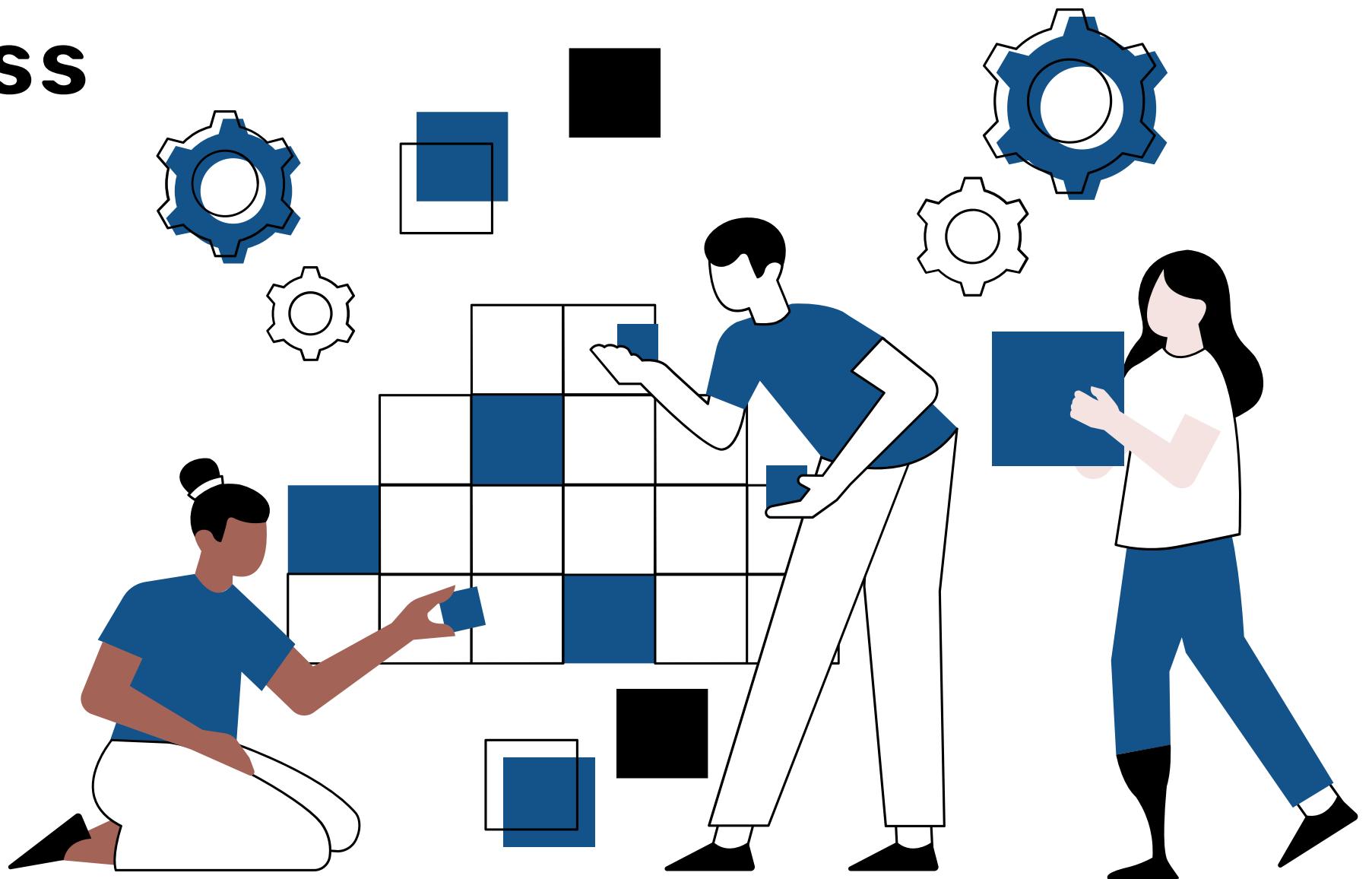
Menentukan Jumlah Class

Metode Segmentasi Data

- K-Means
- K-Medoid
- DBSCAN

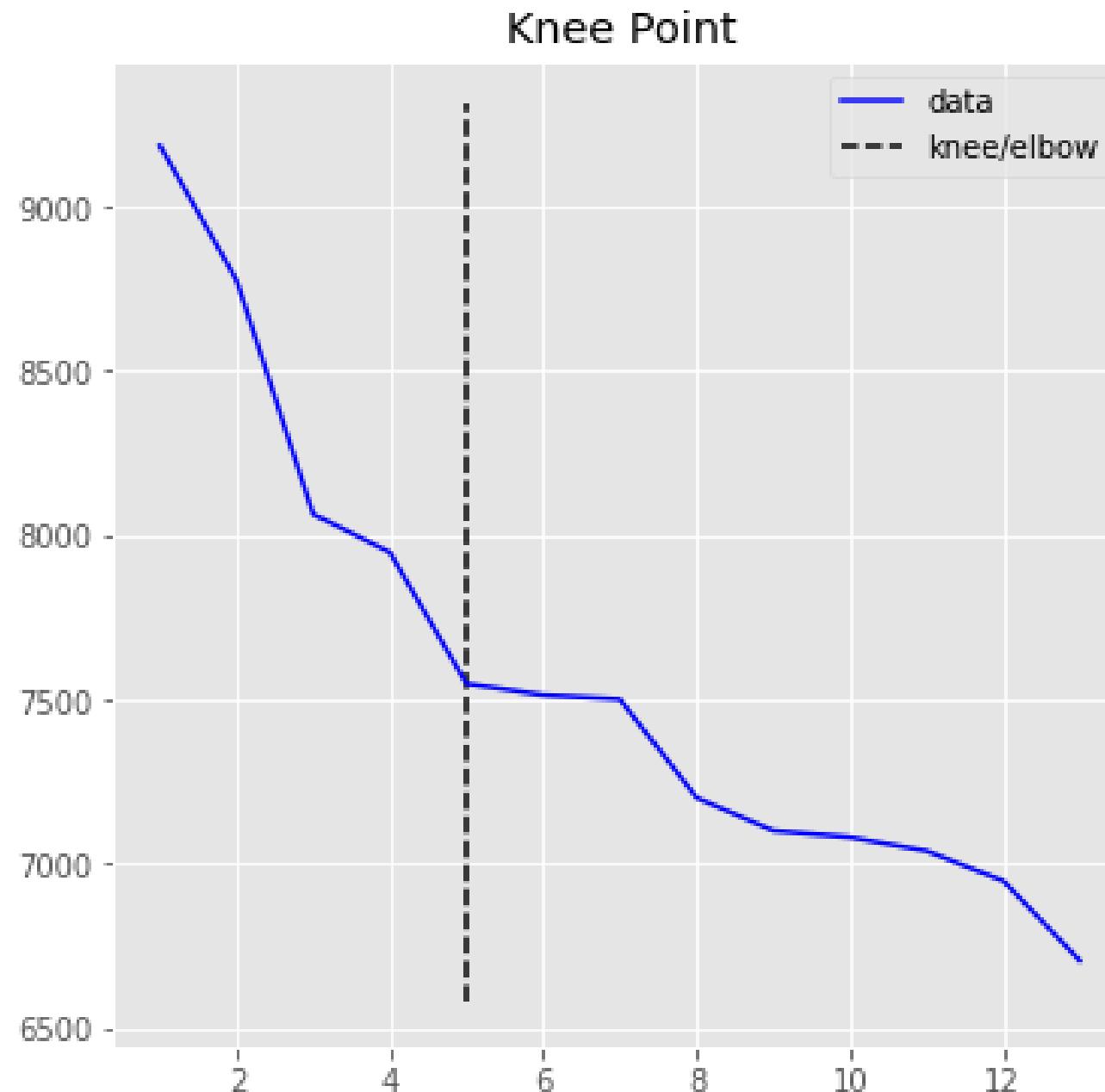
Menentukan Jumlah Class

- Elbow Method
- Silhouette Method

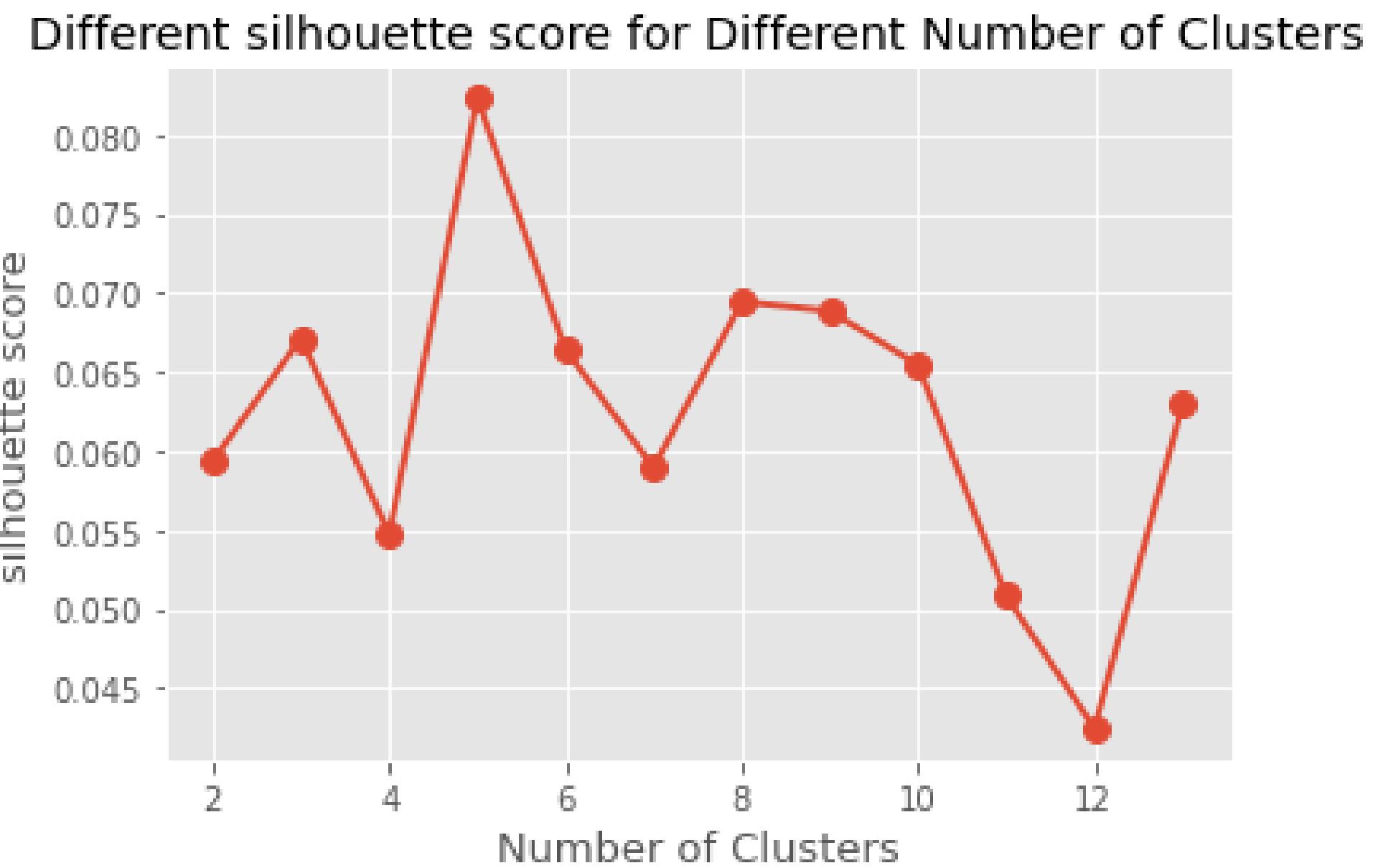


K-Medoids

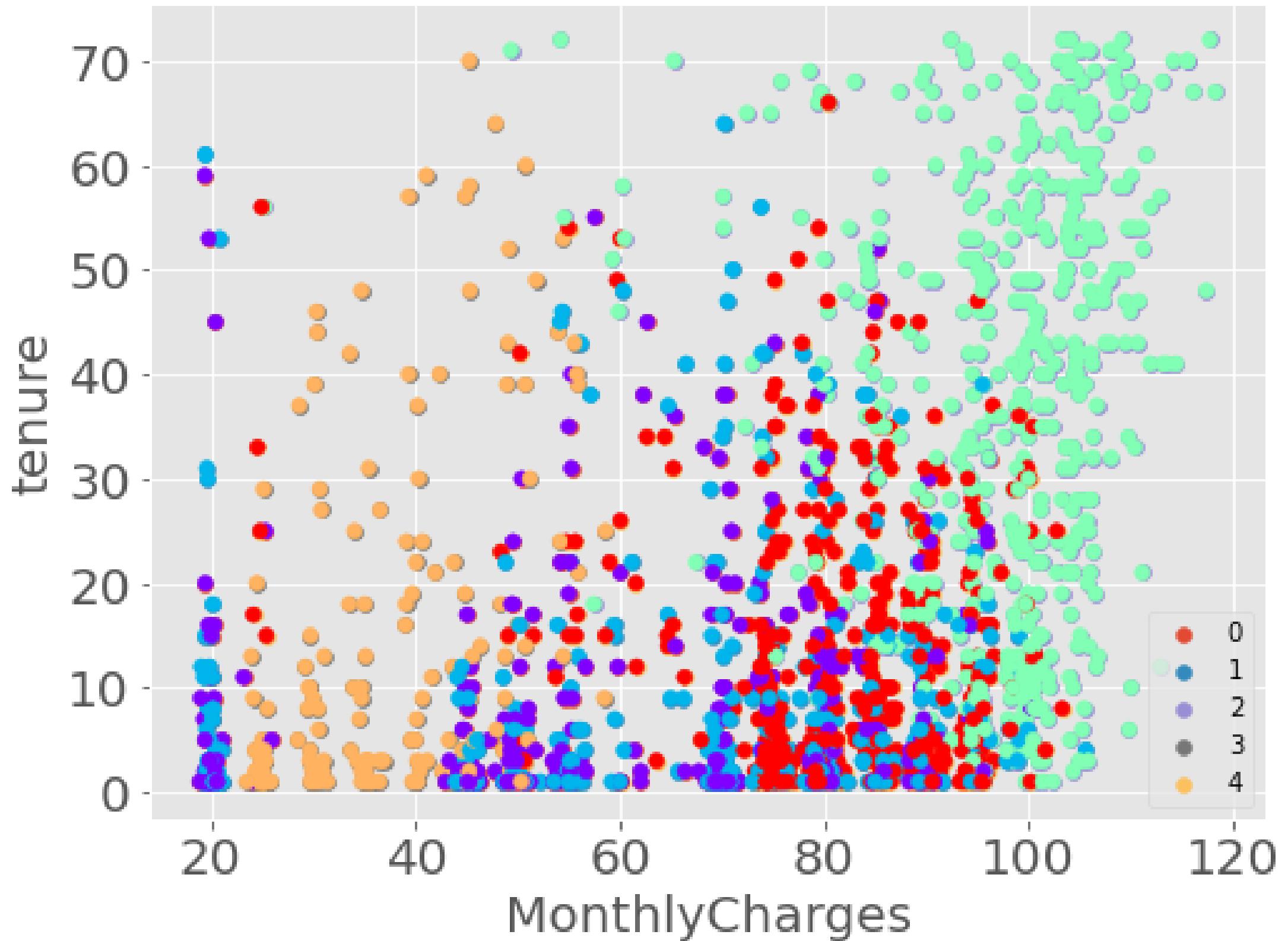
Elbow Method



Silhouette Method



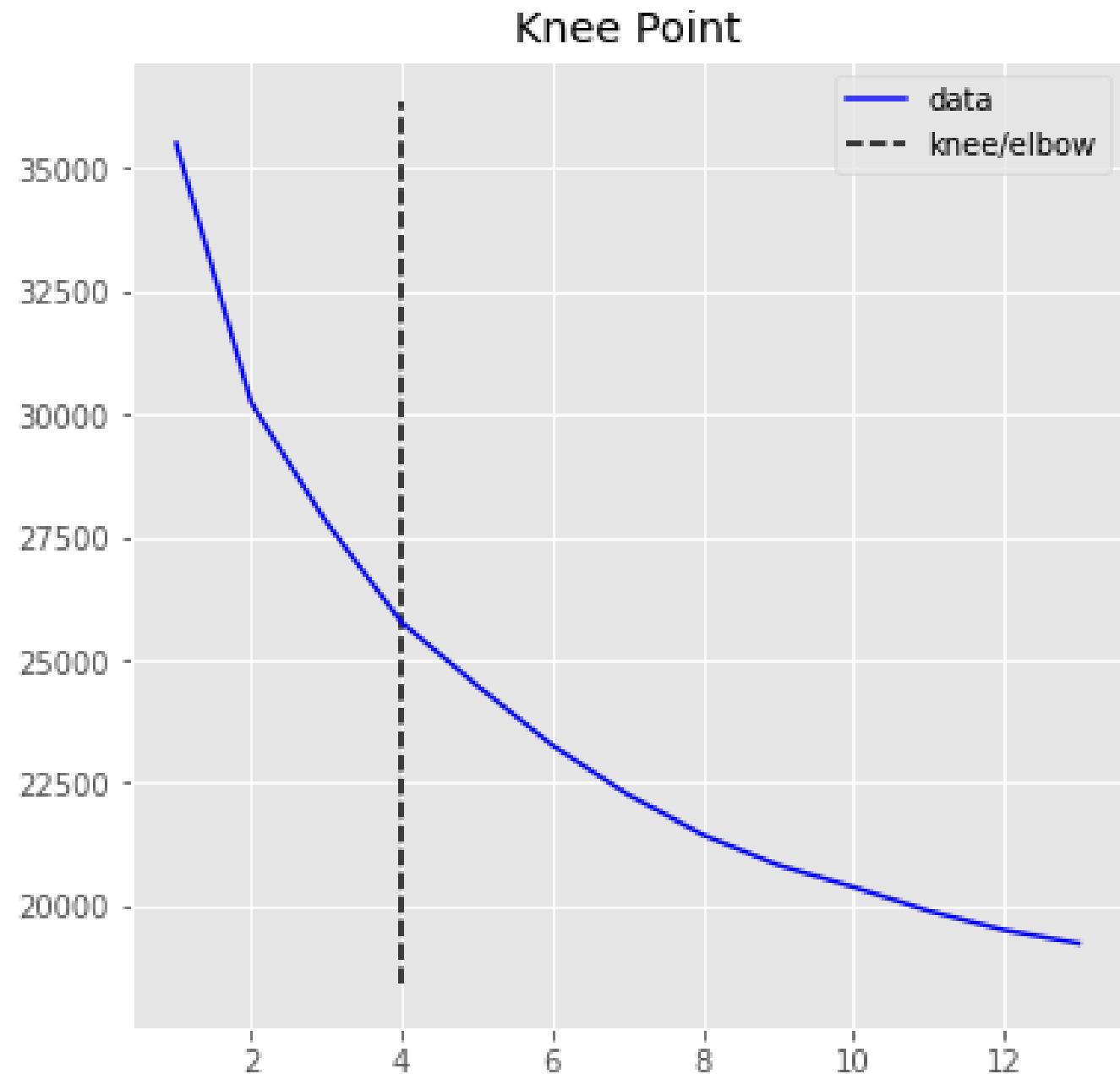
K-medoids clustering



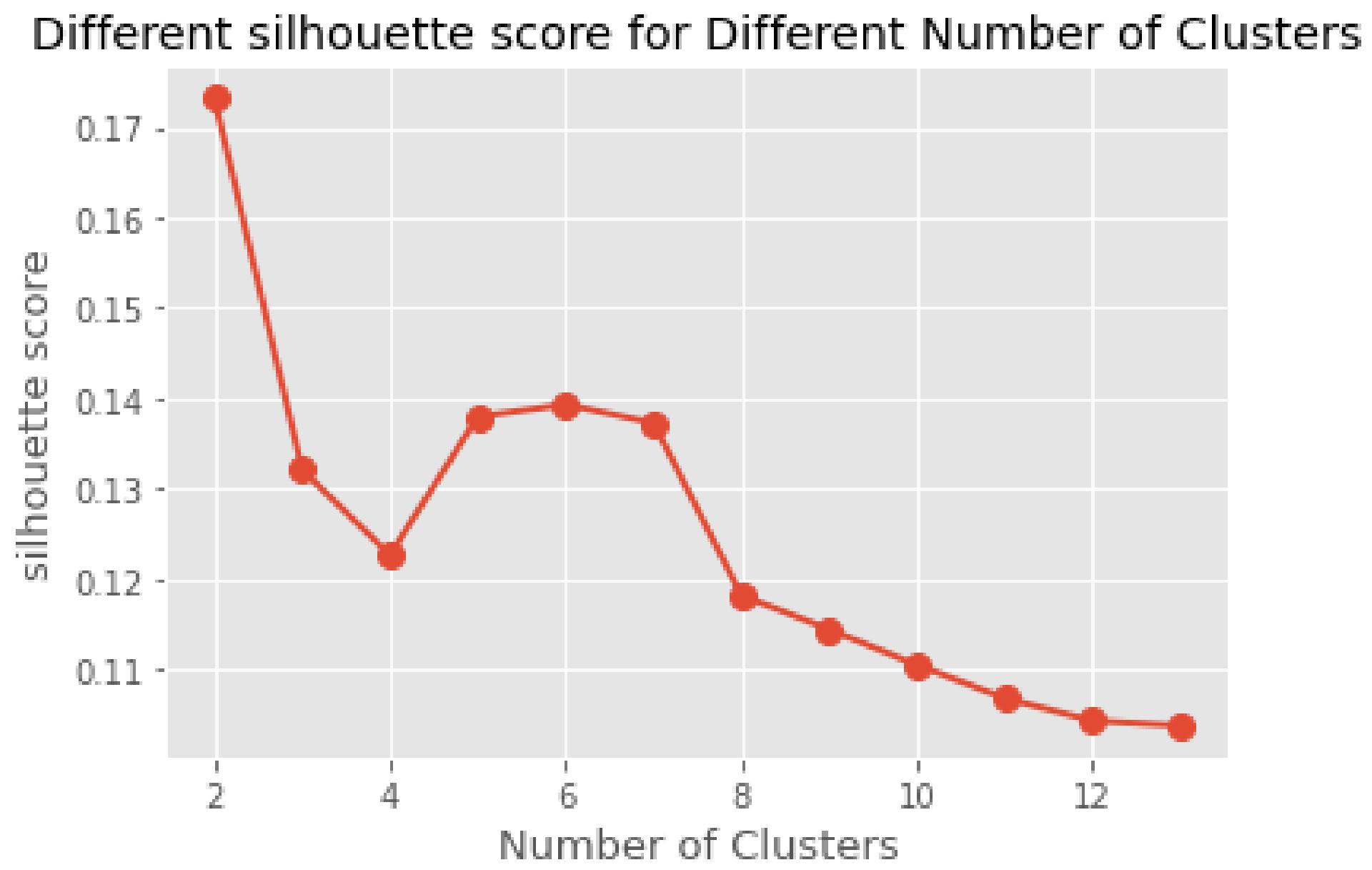
K-Medoids menunjukkan jumlah kelas optimal adalah 5. Setelah dilakukan visualisasi, ternyata pengelompokan terlihat kurang baik.

K-Means

Elbow Method



Silhouette Method

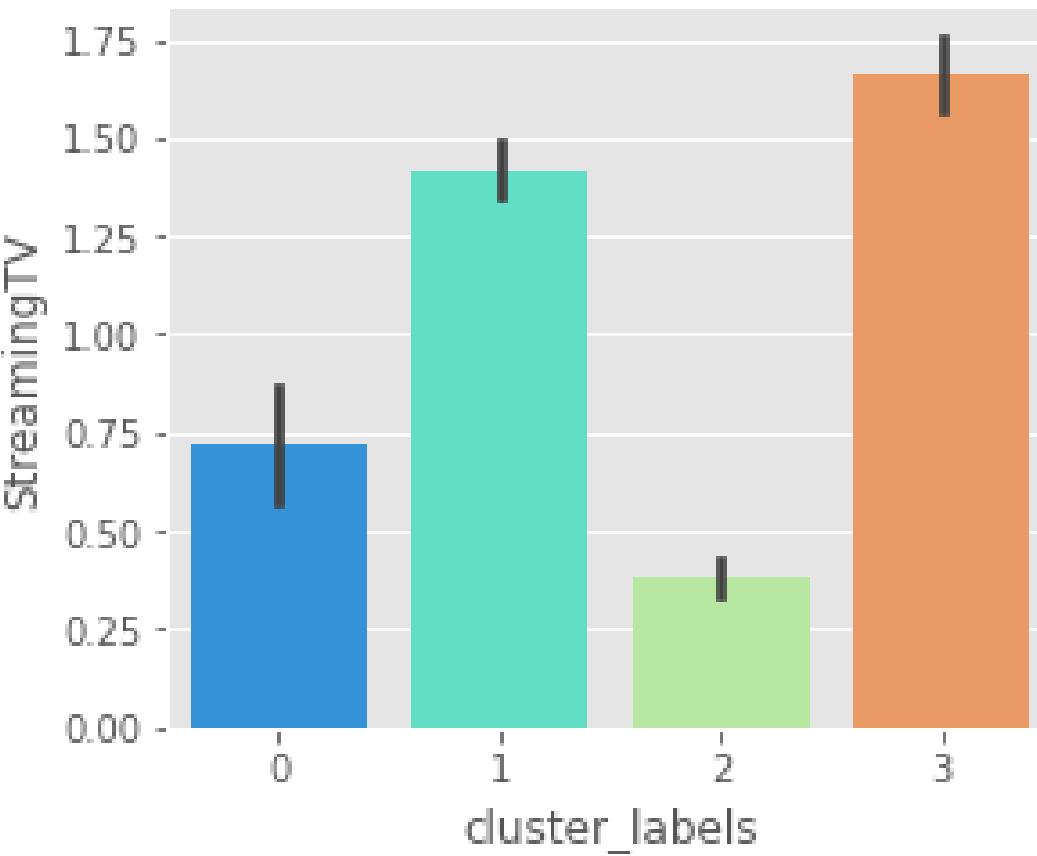
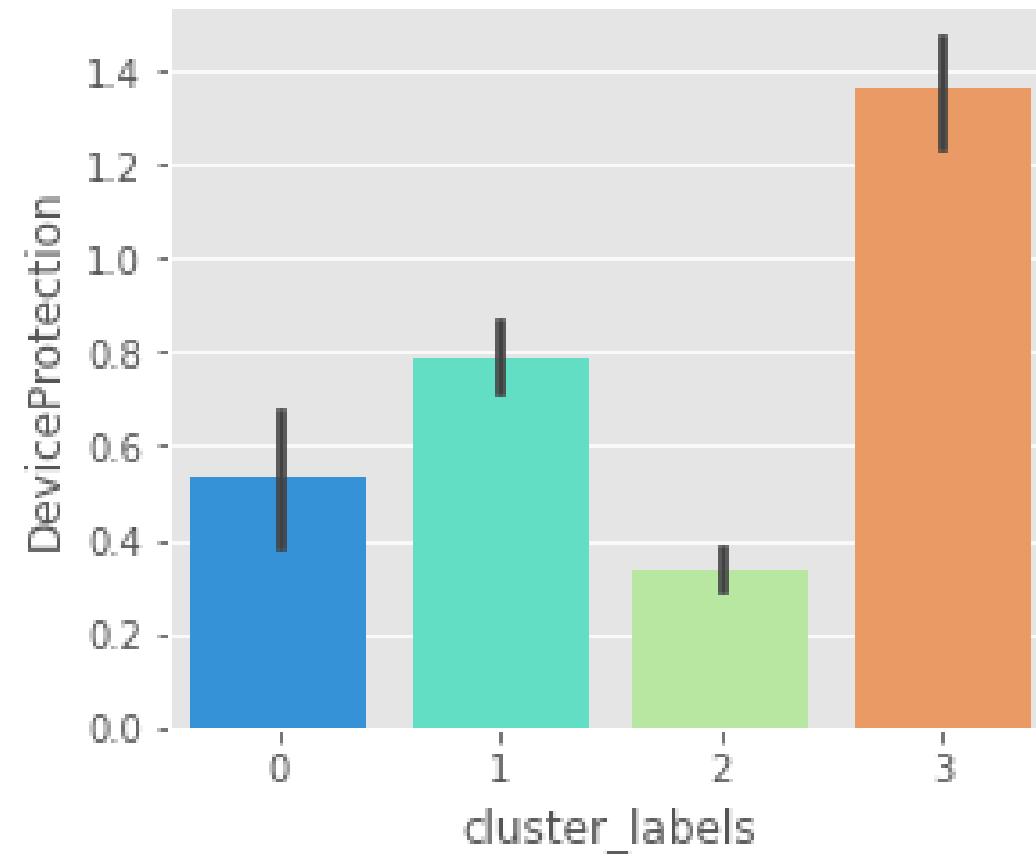
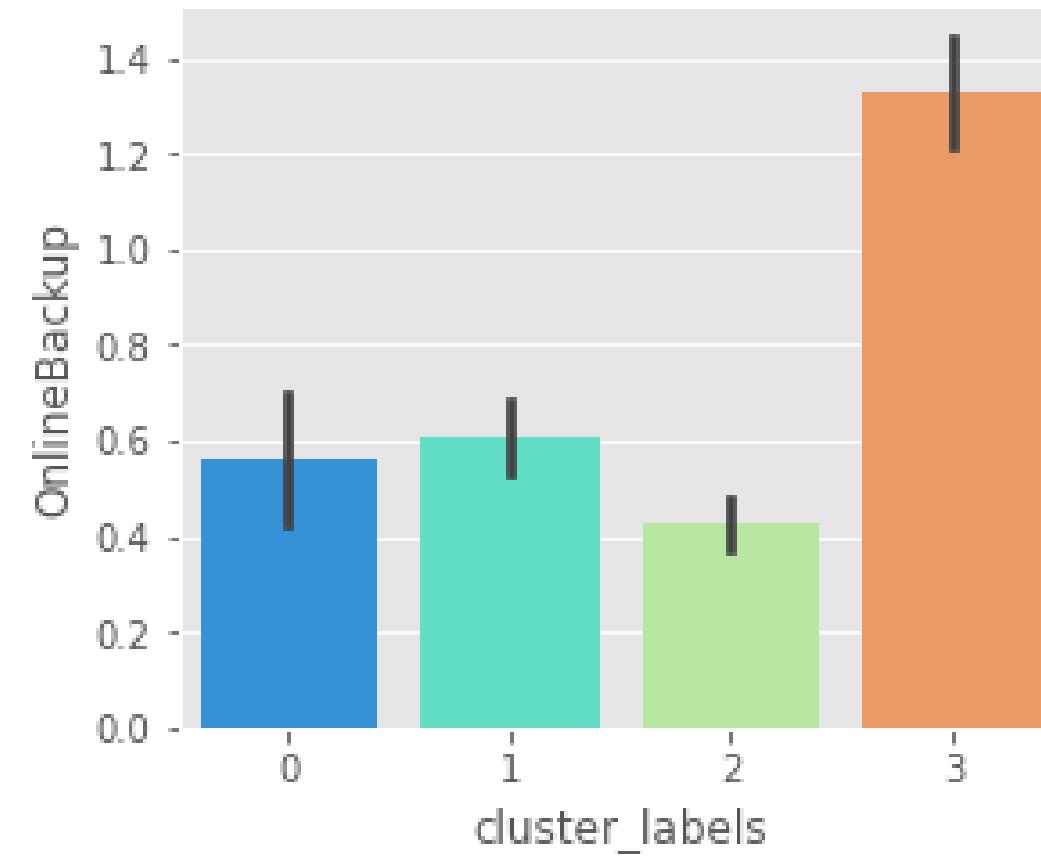
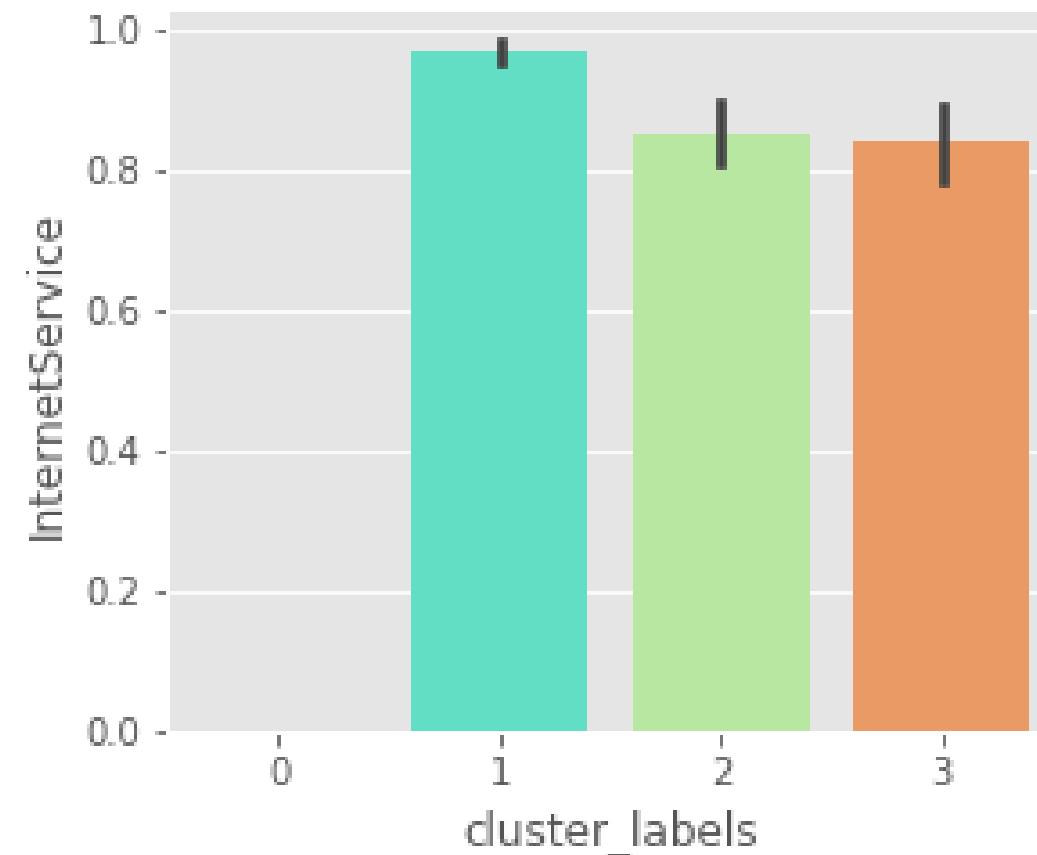
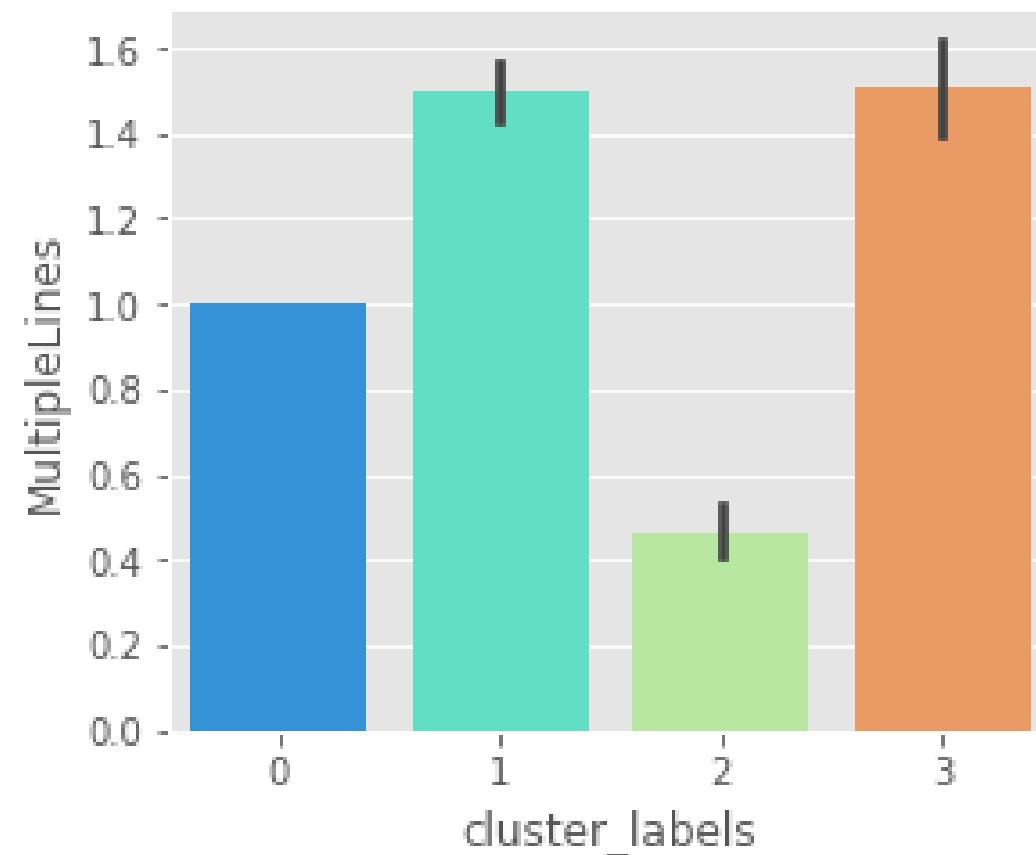
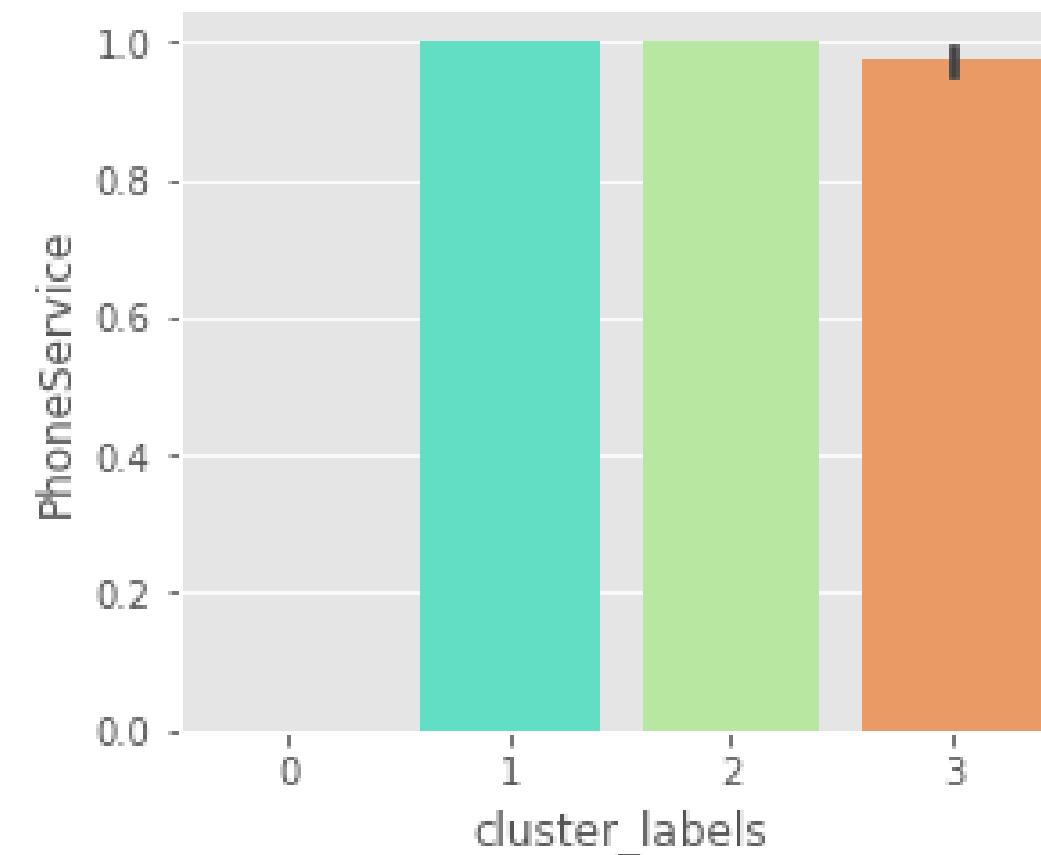


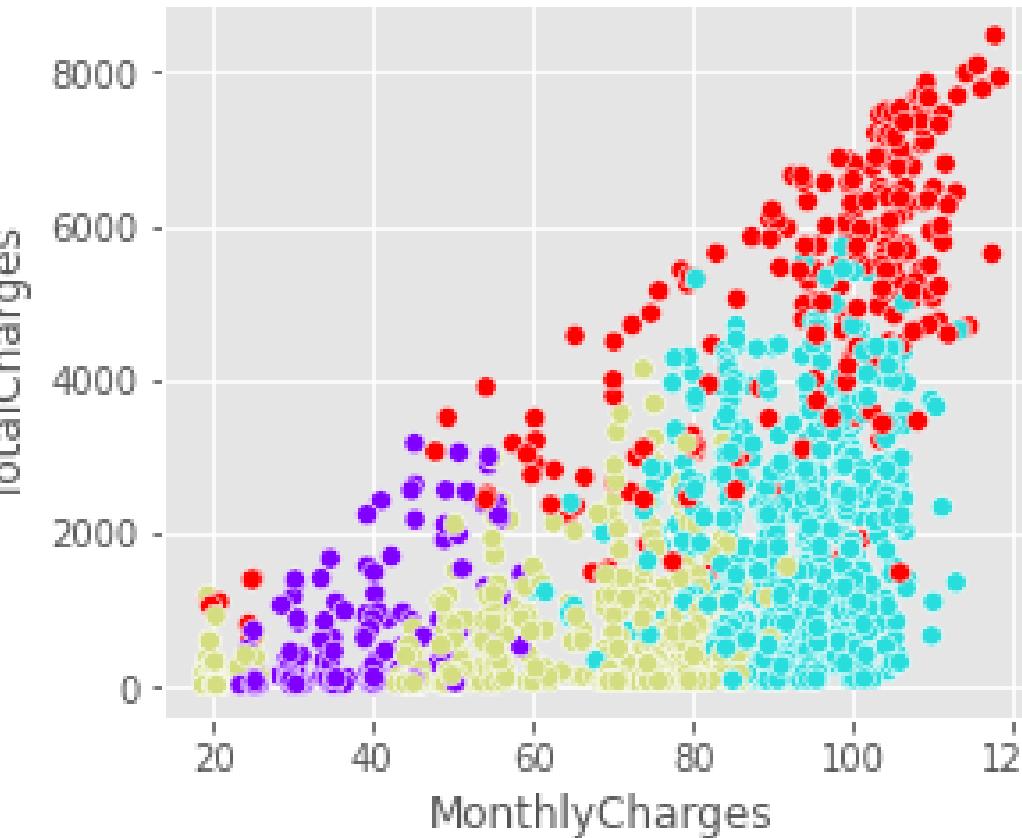
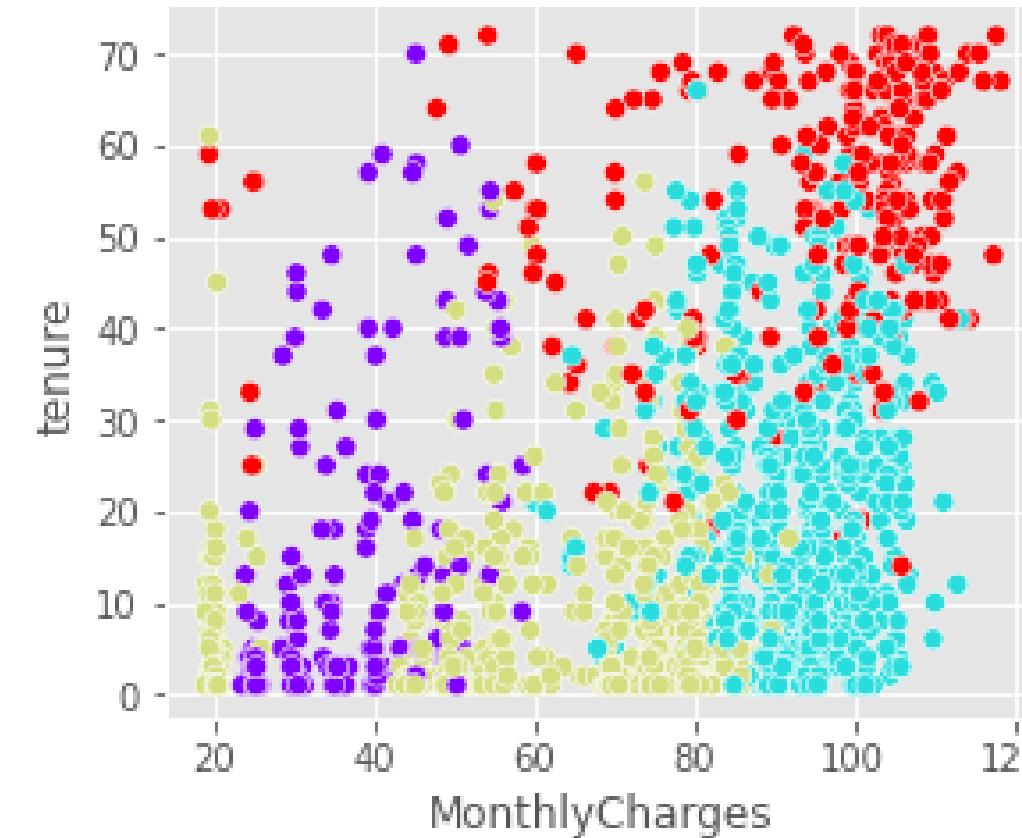
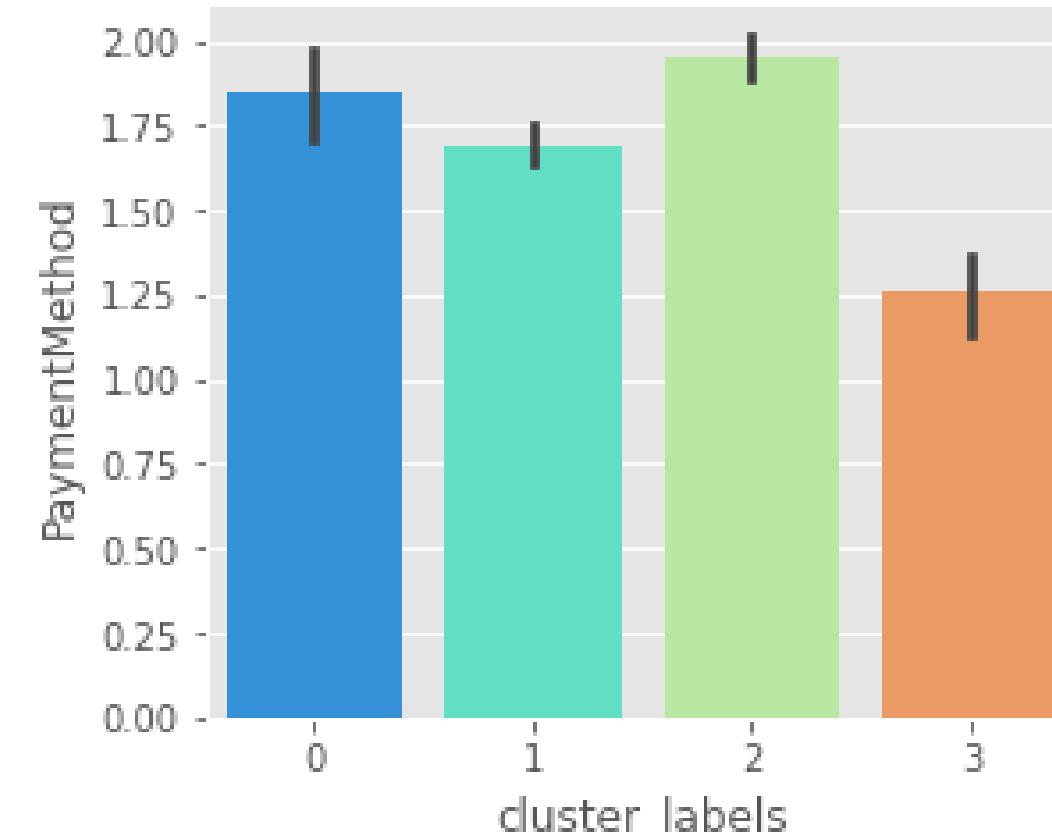
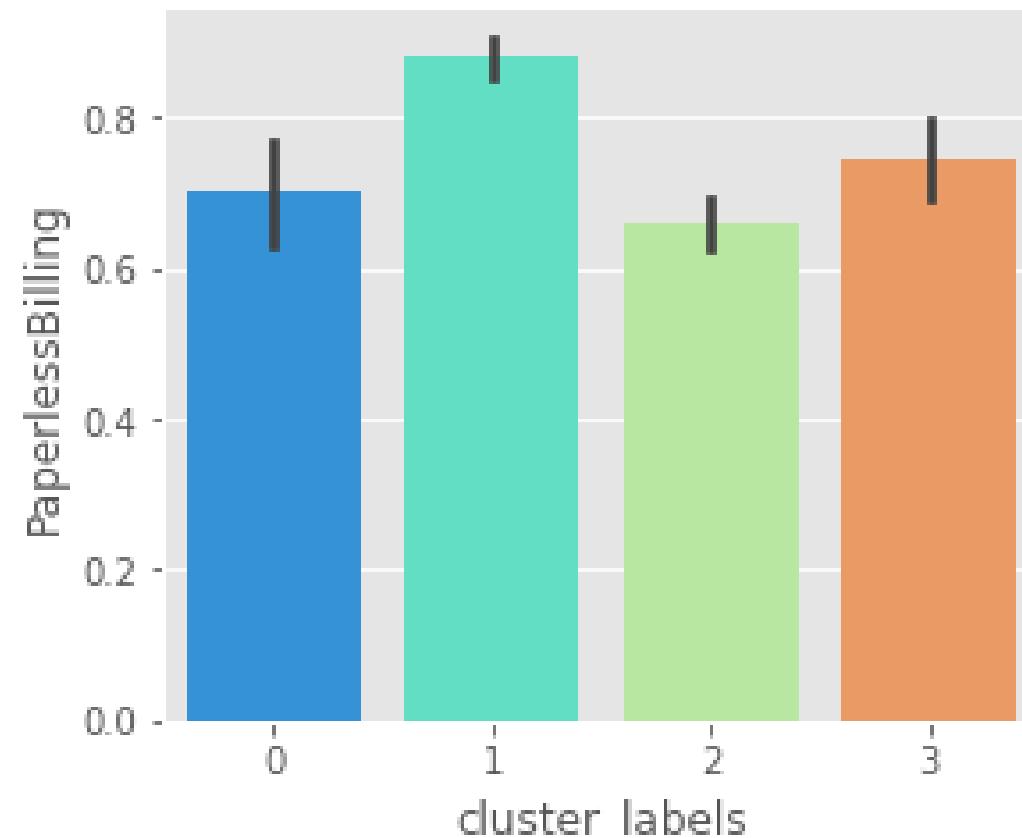
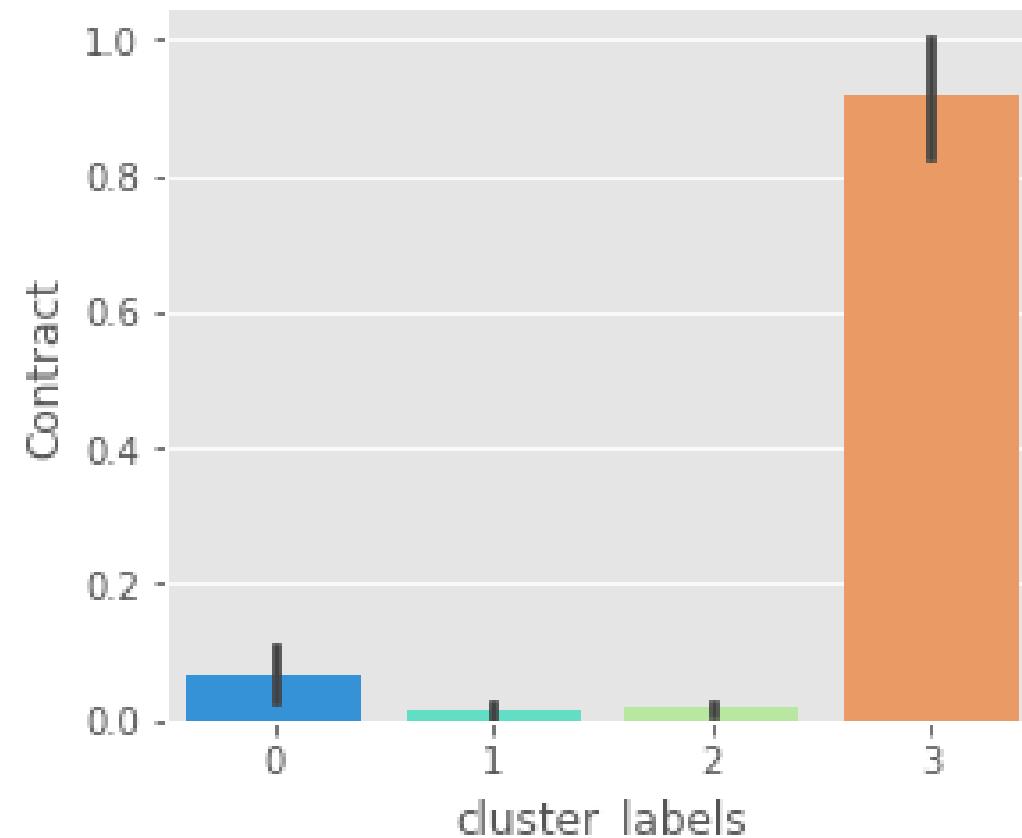
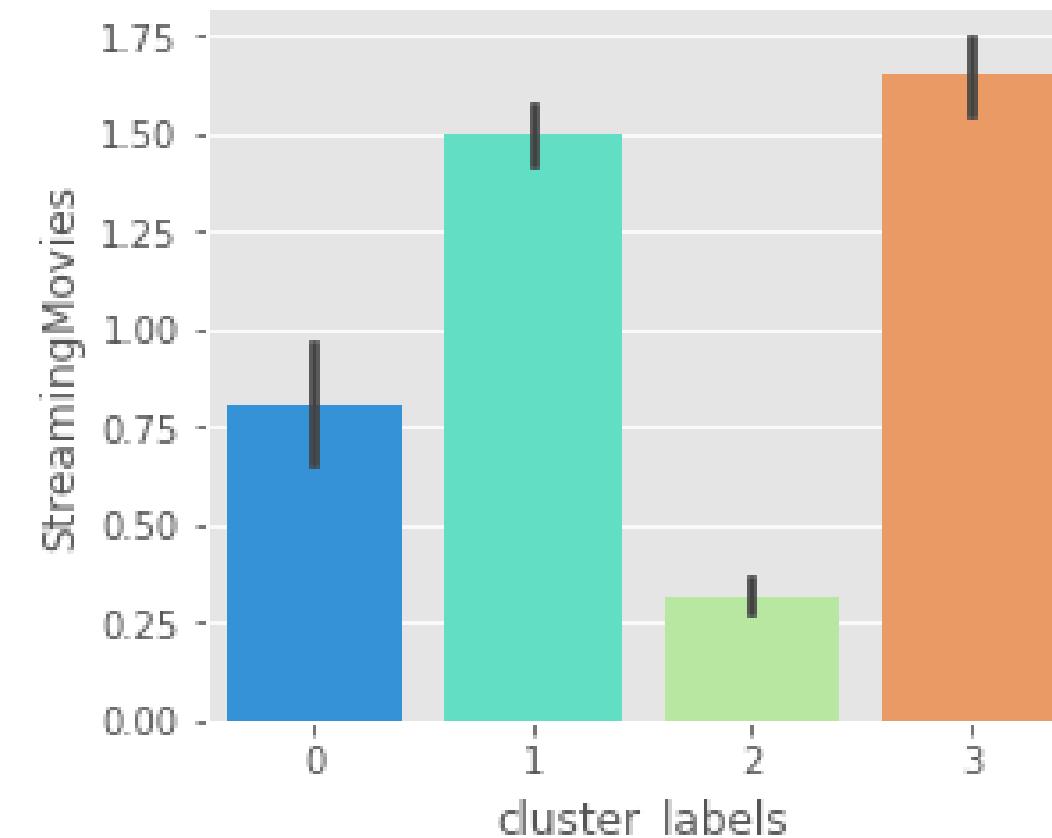
K-means clustering



K-Means dengan elbow method menunjukan jumlah kelas optimal adalah 4 dengan sedangkan silhouette method adalah 2. Setelah dilakukan visualisasi, kelas berjumlah 4 kelas terlihat cukup baik.

Selanjutnya dilakukan analisis terhadap masing-masing untuk mengetahui karakteristik / pola yang dimiliki setiap kelas





Customer Behaviour (%)

Phone Service	Multiple Lines	Internet Service	Online Backup	Device Protection	TV	Movies	Contract	Paperless Billing	Payment Method
N: 100 Y: 0	N: 0 NP: 100 Y: 0	D: 100 F: 0 N: 0	N: 72 NI: 0 Y: 28	N: 73 NI: 0 Y: 27	N: 64 NI: 0 Y: 36	N: 60 NI: 0 Y: 40	M: 93 1Y: 7 2Y: 0	N: 30 Y: 70	B: 12 C: 12 E: 55 M: 21
N: 0 Y: 100	N: 25 NP: 0 Y: 75	D: 3 F: 97 N: 0	N: 70 NI: 0 Y: 30	N: 60 NI: 0 Y: 40	N: 29 NI: 0 Y: 71	N: 25 NI: 0 Y: 75	M: 99 1Y: 1 2Y: 0	N: 12 Y: 88	B: 13 C: 11 E: 69 M: 7
N: 0 Y: 100	N: 77 NP: 0 Y: 23	D: 27 F: 60 N: 13	N: 72 NI: 13 Y: 15	N: 77 NI: 13 Y: 10	N: 75 NI: 13 Y: 12	N: 77 NI: 13 Y: 1	M: 98 1Y: 2 2Y: 0	N: 34 Y: 66	B: 11 C: 10 E: 53 M: 26
N: 2 Y: 98	N: 23 NP: 2 Y: 75	D: 19 F: 78 N: 3	N: 32 NI: 3 Y: 65	N: 30 NI: 3 Y: 67	N: 15 NI: 3 Y: 82	N: 16 NI: 3 Y: 81	M: 28 1Y: 53 2Y: 19	N: 26 Y: 74	B: 28 C: 25 E: 6 M: 41

N: No | Y: Yes | NP: No Phone | D: DSL | F: Fiber Optic | NI: No Internet | M: Month to month | 1Y: 1 Year | 2Y: 2 Years | B: Bank Transfer | C: Credit Card | E: Electronic Check | M: Mailed Check

Kesimpulan

Class	Karakteristik
0	Merupakan pelanggan yang berfokus pada penggunaan internet DSL tanpa menggunakan layanan telefon. Sebagian besar tidak menggunakan layanan keamanan dan entertainment
1	Rata-rata berlangganan telefon dengan multiple lines, tidak menggunakan layanan keamanan, dan berlangganan layanan entertainment. Rata-rata masa berlangganan adalah 19 bulan.
2	Menggunakan layanan telefon, sebagian besar tanpa multiple lines dan tanpa layanan keamanan, hampir seluruhnya tidak menggunakan layanan entertainment. Rata-rata masa berlangganan 7 bulan
3	Rata-rata menggunakan seluruh layanan yang desediakan oleh telco (Full services) dan berlangganan sudah cukup lama sebelum berhenti dengan rata-rata 59 bulan. Kemungkinan adalah pelaku bisnis

Terima Kasih

