Predict Customer Personality to boost marketing campaign by using Machine Learning

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

Created by:
Andi Eka Nugraha
an.ekanugraha@gmail.com
linkedin.com/in/andi-eka-nugraha

Bachelor in Physics with major expertise in Instrumentation & Robotics and has attended a Datascience bootcamp for 4 months. Experienced in programming microcontrollers and machine learning to process data or images, as well as creating robotic systems that can support human work. Able to understand business, especially for data analysis, studying statistics and machine learning, as well as the ability to create regression models, classification, and clustering. Skills in identifying and analyzing patterns in data and presenting analytical results well.

Overview

Rakamin Academy

“Sebuah perusahaan dapat berkembang dengan pesat saat mengetahui perilaku customer personality nya, sehingga dapat memberikan layanan serta manfaat lebih baik kepada customers yang berpotensi menjadi loyal customers. Dengan mengolah data historical marketing campaign guna menaikkan performa dan menyasar customers yang tepat agar dapat bertransaksi di platform perusahaan, dari insight data tersebut fokus kita adalah membuat sebuah model prediksi kluster sehingga memudahkan perusahaan dalam membuat keputusan ”

Data Overview

#	Column	Non-Null Count	Dtype
0	ID	2240 non-null	int64
1	Year_Birth	2240 non-null	int64
2	Education	2240 non-null	object
3	Marital_Status	2240 non-null	object
4	Income	2216 non-null	float64
5	Kidhome	2240 non-null	int64
6	Teenhome	2240 non-null	int64
7	Dt_Customer	2240 non-null	object
8	Recency	2240 non-null	int64
9	MntCoke	2240 non-null	int64
10	MntFruits	2240 non-null	int64
11	MntMeatProducts	2240 non-null	int64
12	MntFishProducts	2240 non-null	int64
13	MntSweetProducts	2240 non-null	int64
14	MntGoldProds	2240 non-null	int64
15	NumDealsPurchases	2240 non-null	int64
16	NumWebPurchases	2240 non-null	int64
17	NumCatalogPurchases	2240 non-null	int64
18	NumStorePurchases	2240 non-null	int64
19	NumWebVisitsMonth	2240 non-null	int64
20	AcceptedCmp3	2240 non-null	int64
21	AcceptedCmp4	2240 non-null	int64
22	AcceptedCmp5	2240 non-null	int64
23	AcceptedCmp1	2240 non-null	int64
24	AcceptedCmp2	2240 non-null	int64
25	Complain	2240 non-null	int64
26	Z_CostContact	2240 non-null	int64
27	Z_Revenue	2240 non-null	int64
28	Response	2240 non-null	int64

- Dataset berisi data **Marketing Campaign** pada perusahaan marketing, data terdiri dari informasi pribadi pelanggan, pengeluaran belanja pada kategori tertentu dalam dua tahun terakhir, dan respon pelanggan terhadap campaign yang dilakukan
- Dalam Dataset terdiri dari 29 kolom dan terdapat data null pada salah satu kolom
- Diharapkan hasil dari model **Respons** dapat memberikan dorongan yang signifikan terhadap efisiensi kampanye pemasaran dengan meningkatkan respons atau mengurangi biaya. Tujuannya adalah untuk memprediksi siapa yang akan menanggapi penawaran untuk suatu produk atau layanan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Engineering

```
df['Age'] = df['Dt_Customer'] - df['Year_Birth']
df['Children'] = df['Kidhome'] + df['Teenhome']
df['Total_Spents'] = df['MntCoke'] + df['MntFruits'] + df['MntMeatProducts'] + df['MntFishProducts'] + df['MntSweetProducts'] + df['MntGoldProds']
df['Total_Transactions'] = df['NumWebPurchases'] + df['NumCatalogPurchases'] + df['NumStorePurchases'] + df['NumDealsPurchases']
df['Total_AcceptedCmp'] = df['AcceptedCmp3'] + df['AcceptedCmp4'] + df['AcceptedCmp5'] + df['AcceptedCmp1'] + df['AcceptedCmp2']
```

Karena terdapat beberapa kolom data yang sejenis maka dilakukan Feature Engineering

- **Age** → Dt_Customer - Year_Birth
- **Children** → Kidhome + Teenhome
- **Total_Spents** → MntCoke + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds
- **Total_Transactions** → NumWebPurchases + NumCatalogPurchases + NumStorePurchases + NumDealsPurchases
- **Total_AcceptedCmp** → AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp + AcceptedCmp1 + AcceptedCmp2

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Engineering

```
df['Age_Segment'] = df['Age'].apply(lambda x: '16-30' if x <= 30 else '30-40' if x <= 40 else '40-50' if x <= 50 else '50-60' if x <= 60 else 'Lansia')
df['Age_Segment'] = df['Age_Segment'].astype('category')
```

Melakukan **segmentasi pada kelompok umur** untuk membantu menarik interpretasi terhada data khususnya pada kolom umur, umur dikelompokan menjadi 16-30, 30-40, 40-50, 50-60, dan Lansia untuk usia diatas 60 tahun.

```
x = df[df['Response']==0].groupby(['Age_Segment'])['Response'].count().reset_index()
y = df[df['Response']==1].groupby(['Age_Segment'])['Response'].count().reset_index()
x['Respond'] = y['Response']
x = x.rename(columns={'Response':'Not_Respond'})
x['conversion_rate'] = x['Respond']/(x['Not_Respond']+x['Respond'])*100
```

Membuat dataset baru yang berisi informasi jumlah pelanggan yang tidak merespon ('Not_Respond'), jumlah pelanggan yang merespon ('Respond'), dan tingkat konversi ('conversion_rate') untuk setiap segmen usia ('Age_Segment').
Tingkat konversi (CVR) dihasilkan dari

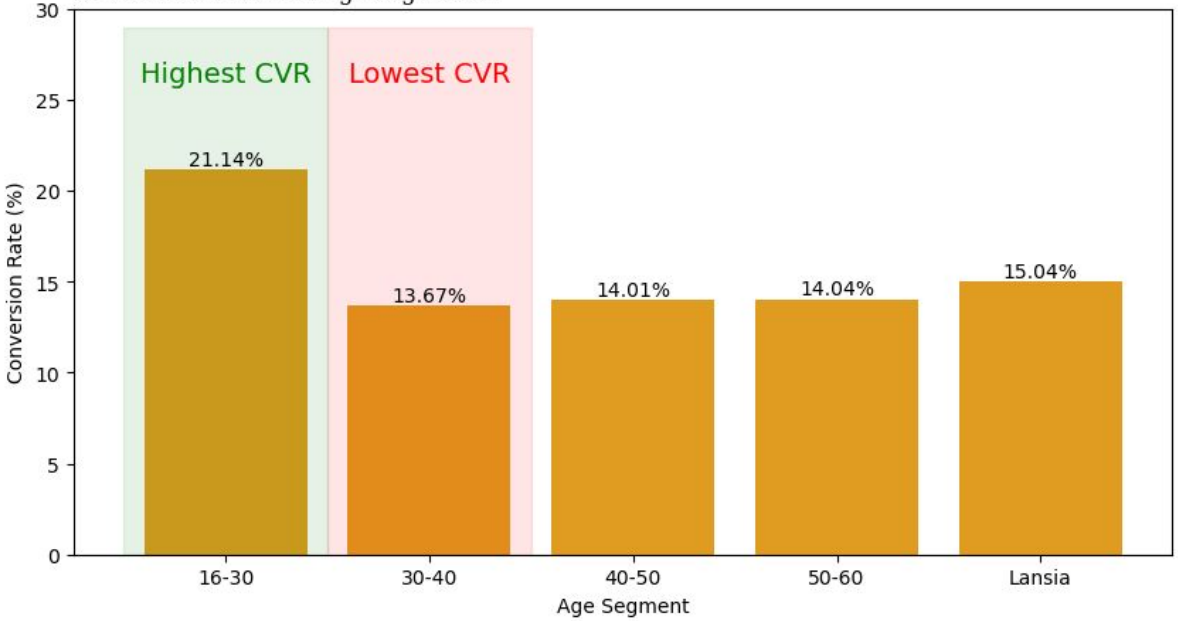
conversion_rate = Respond / (Not_Respond + Respond) x 100%

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Age

The largest CVR based on the largest age segmentation is at the age of 16-30 years

Based on these data, it can be seen that the 16-30 age segment has a CVR of 21.14%. This shows that people in the age range of 16-30 tend to be more responsive to certain actions or events than other age segments.

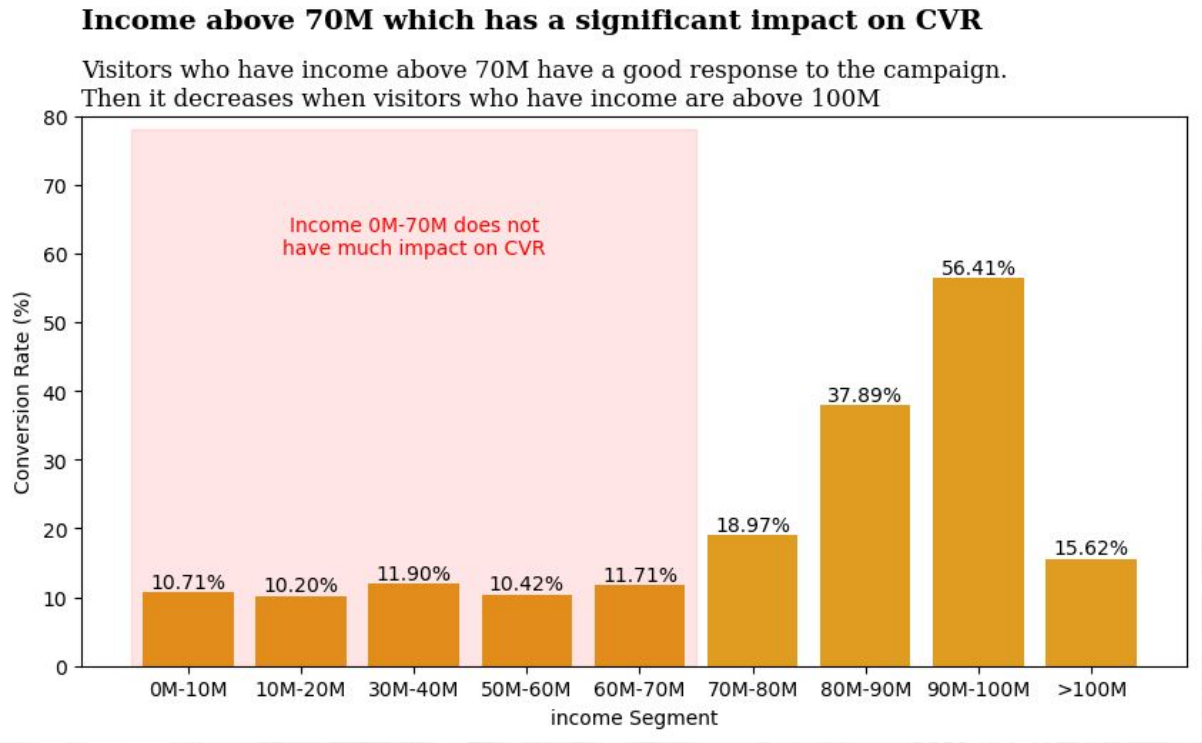


Insight

Segmen usia dengan tingkat konversi terendah adalah segmen usia "30-40" dengan tingkat konversi sebesar 13.67%. Hal ini menunjukkan bahwa pelanggan dalam rentang usia tersebut memiliki tingkat respons yang lebih rendah terhadap kampanye pemasaran. Segmen usia "50-60", "40-50", lanjut usia, memiliki tingkat konversi yang hampir serupa, meskipun tidak menjadi segmen dengan tingkat konversi tertinggi, kedua segmen ini masih memberikan respons yang cukup baik terhadap kampanye pemasaran.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income



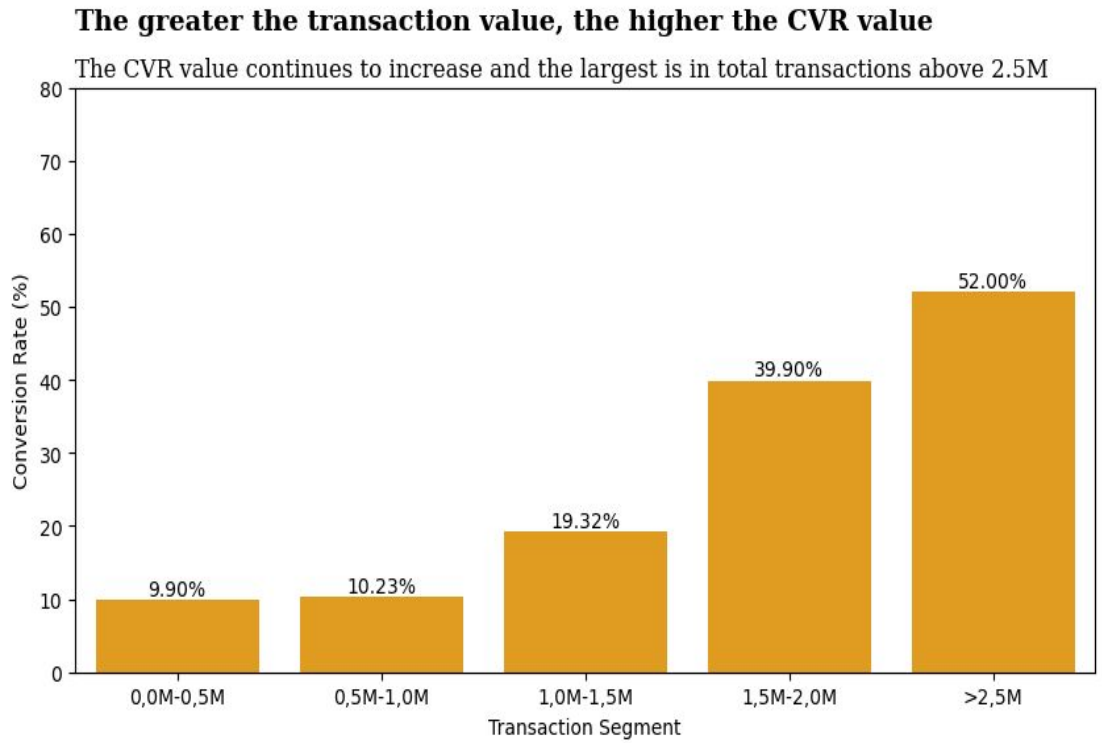
Insight

Segmen usia dengan tingkat CVR terendah berapa pada nilai pendapatan dibawah 70M. Sebaliknya pada rentan diatas 70M tingkat CVR yang lebih besar dan terus meningkat hingga pendapatan 100M, kemudian nilai mengecil untuk pendapatan diatas 100M.

CVR terbesar berada di rentan 91M sampai 100M, hal ini perlu diperhatikan dalam strategi campaign untuk meningkatkan hasil CVR

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Spending



Insight

Terlihat adanya korelasi antara tingkat pengeluaran dan tingkat konversi. Segmentasi dengan **pengeluaran yang lebih tinggi cenderung memiliki tingkat konversi yang lebih tinggi**. Ini menunjukkan bahwa pengeluaran pelanggan dapat menjadi faktor penting dalam menentukan tingkat respons dan konversi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Handling Missing Values

```python
for edu in edu_values:
    for age in age_values:
        idx = (df['Education'] == edu) & (df['Age_Segment'] == age) & (df['Income'].isnull())
        if idx.any():
            mean_income = df.loc[(df['Education']== edu) & (df['Age_Segment'] == age), 'Income'].mean()
            df.loc[idx, 'Income'] = mean_income
            print(f"{edu} - {age} di isi: {mean_income}")
        else:
            print (f"{edu} - {age} tidak dkosong")
```

Column	Non-Null Count
ID	2229 non-null
Age	2229 non-null
Children	2229 non-null
Education	2229 non-null
Marital_Status	2229 non-null
Income	2206 non-null
Dt_Customer	2229 non-null
Recency	2229 non-null
Total_Spents	2229 non-null
Total_Transactions	2229 non-null
NumWebVisitsMonth	2229 non-null
Total_AcceptedCmp	2229 non-null
Complain	2229 non-null
Z_CostContact	2229 non-null
Z_Revenue	2229 non-null
Response	2229 non-null
Age_Segment	2229 non-null

```
S1 - 50-60 di isi: 57504832.61802575
S1 - 40-50 di isi: 51483233.22683706
S1 - 16-30 di isi: 47514496.93251534
S1 - 30-40 di isi: 51258618.58974359
S1 - Lansia di isi: 57204304.347826086
S3 - 50-60 di isi: 56860960.78431372
S3 - 40-50 di isi: 53225453.333333336
S3 - 16-30 tidak dkosong
S3 - 30-40 di isi: 53702130.841121495
S3 - Lansia tidak dkosong
S2 - 50-60 di isi: 57506602.27272727
S2 - 40-50 di isi: 51242562.5
S2 - 16-30 tidak dkosong
S2 - 30-40 di isi: 48446651.1627907
S2 - Lansia di isi: 58882361.70212766
SMA - 50-60 tidak dkosong
SMA - 40-50 tidak dkosong
SMA - 16-30 tidak dkosong
SMA - 30-40 tidak dkosong
SMA - Lansia tidak dkosong
D3 - 50-60 tidak dkosong
D3 - 40-50 di isi: 52019076.92307692
D3 - 16-30 tidak dkosong
D3 - 30-40 di isi: 41282532.46753247
D3 - Lansia tidak dkosong
```

Gambar pertama sebelah kiri merupakan kolom yang digunakan setelah melewati Feature Engineering, terdapat 17 kolom yang akan digunakan dalam Data Preprocessing ini.

Terdapat **data kosong** pada Dataset ‘Income’, untuk menjaga interpretasi data maka nilai kosong tersebut diisi berdasarkan segmentasi ‘Education’ dan ‘Age_Segment’. Data kosong diisi berdasarkan **nilai rata-rata ‘Income’ pada Setiap segmentasi ‘Education’ dan pengelompokan ‘Age_Segement’**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Handling Duplicate Data

```
df = df.drop('ID', axis=1)
df.duplicated().sum()
```

✓ 0.0s

```
183
```

```
df.drop_duplicates(inplace=True, keep='first')
```

✓ 0.0s

Drop **kolom ID** untuk menghindari kondisi yang sama pada dataset meskipun usernya berbeda, terdapat **183 duplikat** yang di drop untuk menghindari ganguan pada kualitas modeling.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Encoding

- **Label Encoding** → ‘Education’, ‘Age_Segment’
- **One Hot Encoding** → ‘Marital_Status’

* Ket:

- Label Encoding dilakukan ketika terdapat urutan yang jelas,
- One Hot Encoding dilakukan ketika data tidak memiliki hirarki jelas

Before

	Age_Segment	Education	Marital_Status
0	50-60	S1	Lajang
1	50-60	S1	Lajang
2	40-50	S1	Bertunangan
3	16-30	S1	Bertunangan
4	30-40	S3	Menikah

After

	Age_Segment	Education	status_Bertunangan	status_Cerai	status_Duda	status_Janda	status_Lajang	status_Menikah
0	0	2	0	0	0	0	1	0
1	0	2	0	0	0	0	1	0
2	3	2	1	0	0	0	0	0
3	1	2	1	0	0	0	0	0
4	2	4	0	0	0	0	0	1

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Transformation

Rakamin Academy

Before

	Income	Total_Spents	Dt_Customer
0	0.241864	1.682545	-1.505593
1	-0.230842	-0.958489	1.417029
2	0.781944	0.285620	-0.044282
3	-1.020341	-0.915302	1.417029
4	0.248076	-0.302383	1.417029

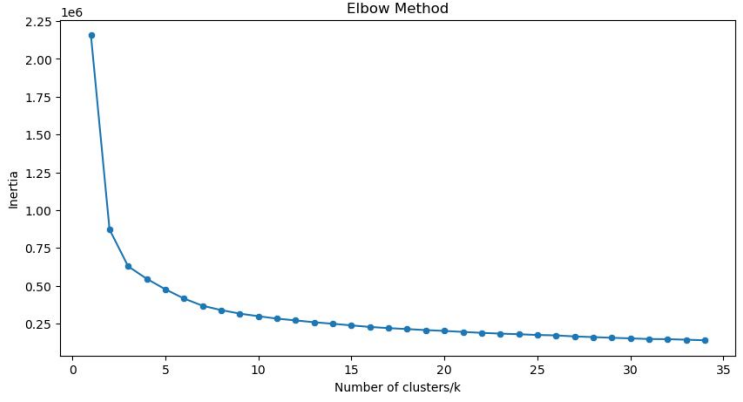
After

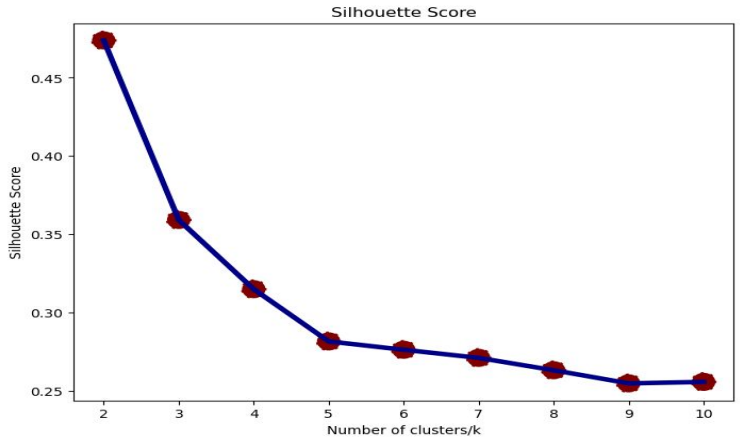
	Income	Total_Spents	Dt_Customer
0	58138000.0	1617000	2012
1	46344000.0	27000	2014
2	71613000.0	776000	2013
3	26646000.0	53000	2014
4	58293000.0	422000	2014

Kolom **'Income'** , **'TotalSpents'**, dan **'Dt_Customer'** dilakukan **Standarisasi** untuk menghindari **over weight** hanya pada beberapa kolom saja yang disebabkan ketimpangan ukuran nilai data yang ada.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling





Elbow merupakan titik di mana penurunan inersia mulai melambat secara signifikan. Pada titik ini, penambahan clster tidak lagi memberikan keuntungan signifikan dalam mengurangi inersia. Pada grafil Elbow Method disamping, **titik Elbow berada pada cluster 2**, nilai cluster selanjutnya kurang efektif dalam mengurangi inersia.

Silhouette Score mengukur sejauh mana setiap sampel data berada di dalam klasternya sendiri dan sejauh mana sampel tersebut terpisah dari klaster lain. konfigurasi klaster dengan nilai Silhouette **Score tertinggi adalah 2**, Ini menandakan bahwa klastering tersebut memberikan pemisahan yang optimal antara klaster dengan menggunakan parameter atau jumlah klaster yang sesuai.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Grid Search

```
from sklearn.model_selection import GridSearchCV
from sklearn.cluster import KMeans
# Tentukan parameter grid
param_grid = {'n_clusters': range(2, 11), 'init': ['k-means++', 'random']}
# Inisialisasi model
kmeans = KMeans(random_state=0)
# Inisialisasi GridSearchCV
grid_search = GridSearchCV(kmeans, param_grid, cv=5,
scoring='adjusted_rand_score')
# Fit model dan cari hyperparameter optimal
grid_search.fit(df)
# Print hyperparameter terbaik
print('Best parameters: ', grid_search.best_params_)
```

```
Best parameters:  {'init': 'k-means++', 'n_clusters': 2}
```

GridSearchCV akan mencoba semua kombinasi hyperparameter yang mungkin, melatih model KMeans dengan setiap kombinasi, dan menghitung skor adjusted_rand_score untuk setiap model.

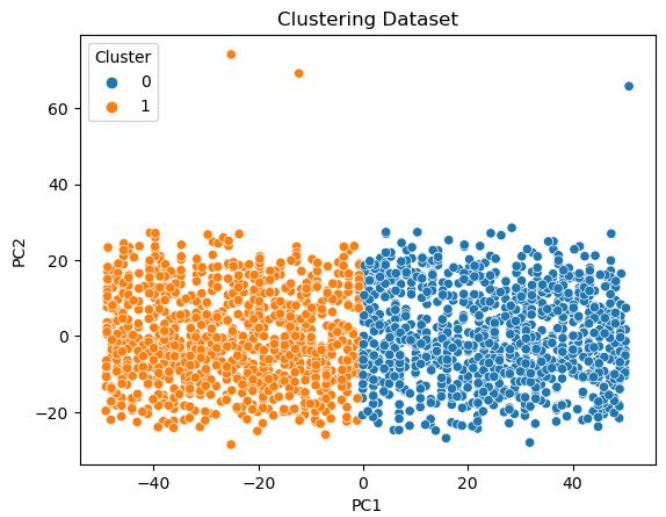
Setelah proses fitting selesai, best_params_ dari objek GridSearchCV mencetak hyperparameter terbaik yang ditemukan. Jika hasilnya adalah {'n_clusters': 2, 'init': 'k-means++'}, maka itu berarti hyperparameter terbaik yang ditemukan adalah menggunakan dua kluster dengan metode inisialisasi 'k-means++'.

Dengan demikian, hasil clustering terbaik yang ditemukan menggunakan metode di atas adalah dengan menggunakan dua kluster (n_clusters = 2) dan metode inisialisasi 'k-means++'.

Untuk selengkapnya, dapat melihat jupyter notebook disini

PCA & Scatter Plot

	PC1	PC2	Cluster
0	12.238079	8.134271	1
1	14.041108	-11.423570	1
2	4.822195	5.558390	1
3	-15.226275	-4.502967	0
4	-10.330942	5.932080	0



Setelah mendapatkan hasil clustering yang optimal, kemudian memasukkan hasil klastering ini ke dalam Dataset. Selanjutnya, melakukan analisis PCA (Principal Component Analysis) untuk mereduksi dimensi dari Dataset.

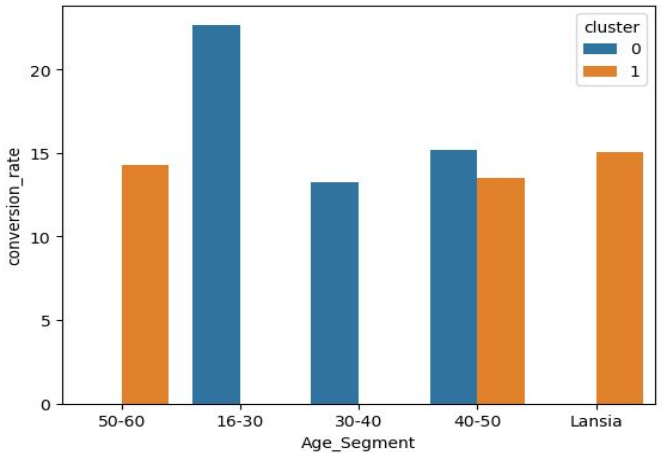
PCA membantu dalam mengurangi dimensi dataset dengan memproyeksikannya ke ruang fitur yang lebih rendah namun masih mempertahankan sebagian besar informasi variabilitas dalam data. Kemudian, lalu membuat grafik scatterplot dengan menggunakan hasil PCA untuk memvisualisasikan hasil clustering.

Grafik scatterplot menampilkan clustering sebagai titik-titik dalam ruang feature baru yang dihasilkan oleh PCA. Hasil dari grafik menunjukan bahwa dua clusteryang dihasilkan oleh K-means **terpisah dengan jelas satu sama lain**. Meskipun pemisahan antara dua klaster terlihat jelas, **jarak antara klaster tersebut sangat tipis**. Ini mengindikasikan bahwa terdapat **sejumlah overlap atau tumpang tindih antara klaster** tersebut. Beberapa titik data dalam satu klaster mungkin terletak sangat dekat dengan batas klaster lainnya. Hal ini menandakan bahwa **ada kemiripan dalam fitur atau atribut tertentu antara dua kelompok data tersebut**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR Based on Age



	Age_Segment	cluster	Not_Respond	Respond	conversion_rate
0	50-60	1	373	62	14.252874
1	16-30	0	208	61	22.676580
2	30-40	0	490	75	13.274336
3	40-50	0	302	54	15.168539
4	40-50	1	186	29	13.488372
5	Lansia	1	175	31	15.048544

Segmentasi usia 16-30 memiliki tingkat konversi yang paling tinggi, sehingga dapat menjadi target utama untuk meningkatkan respons terhadap suatu tindakan atau kampanye.

Segmentasi usia 30-40 memiliki jumlah Responden yang Tidak Merespon yang paling tinggi, sehingga perlu dilakukan analisis lebih lanjut untuk memahami faktor-faktor yang menyebabkan rendahnya respons pada segmentasi ini.

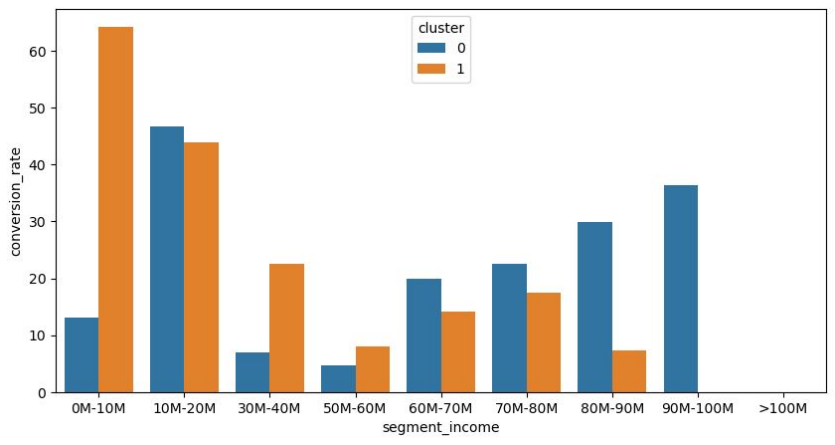
Kluster 1 (segmentasi usia 50-60 dan Lansia) memiliki jumlah Responden yang lebih sedikit, namun tetap memiliki potensi untuk meningkatkan tingkat konversi dengan strategi yang tepat.

Pada segmentasi usia 40-50 di kluster 0, terdapat potensi untuk meningkatkan tingkat konversi, mengingat jumlah Responden yang Merespon relatif lebih rendah dibandingkan dengan segmentasi usia lainnya di kluster yang sama.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR based on Income



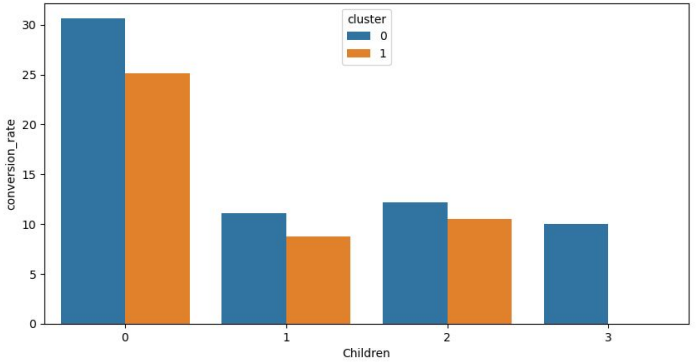
Pada rentan pendapatan diatas 90M, pada cluster 1 tidak ada yang merespon campaign. Sedangkan pada cluster 0, pendapatan diatas 100M tidak ada yang merespon campaign.

Segmen dengan pendapatan yang lebih tinggi cenderung memiliki tingkat respons yang lebih rendah dalam kluster 1. Selain itu, beberapa segmen dengan pendapatan rendah juga menunjukkan tingkat konversi yang baik di kluster 1. Hal ini bisa menjadi pertimbangan dalam mengembangkan strategi pemasaran yang lebih efektif untuk segmen-segmen ini, dengan **fokus pada kluster 1 untuk meningkatkan tingkat respons dan konversi pada pendapatan rendah dan fokus pada kulster 0 pada pendapatan tinggi.**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR based on number of children



	Children	cluster	Not_Respond	Respond	conversion_rate
0	0	0	206	91.0	30.639731
1	0	1	203	68.0	25.092251
2	1	0	603	75.0	11.061947
3	1	1	333	32.0	8.767123
4	2	0	173	24.0	12.182741
5	2	1	170	20.0	10.526316
6	3	0	18	2.0	10.000000
7	3	1	28	NaN	NaN

Jika kita melihat tingkat respons pada setiap kategori jumlah anak, terlihat bahwa **respons pada kluster 0 lebih tinggi daripada kluster 1 di semua kategori jumlah anak**. Namun, dalam kluster 1, terjadi penurunan jumlah respons seiring dengan peningkatan jumlah anak hingga pada jumlah anak 3 tidak ada yang memberikan respon. Hal ini menunjukkan bahwa dalam kluster 1, jumlah anak berpotensi memiliki pengaruh negatif terhadap tingkat respons.

kita dapat menyimpulkan bahwa dalam kluster 0, jumlah anak tidak memiliki pengaruh yang signifikan terhadap tingkat respons, sementara dalam kluster 1, peningkatan jumlah anak dapat berhubungan dengan penurunan tingkat respons.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conclusion

- **Secara umum**, karakteristik kluster 0 adalah respons dan konversi yang relatif stabil, tanpa pengaruh yang signifikan dari jumlah anak, pendapatan, atau usia. Di sisi lain, kluster 1 memiliki respons yang menurun seiring dengan peningkatan jumlah anak dan memiliki tingkat konversi yang bervariasi tergantung pada pendapatan dan usia.
- **Cluster 0:** Fokuskan upaya pemasaran pada segmen anak tanpa memperhatikan jumlah anak, pendapatan tinggi dalam segmen 10M-20M, serta segmen usia 16-30, 30-40, dan 40-50. Peningkatan respons dan konversi dapat dicapai dengan strategi pemasaran yang tepat sesuai dengan karakteristik demografis ini.
- **Cluster 1:** Teliti lebih lanjut mengapa tingkat respons dan konversi rendah dalam segmen anak dengan jumlah yang lebih tinggi, pendapatan rendah dalam segmen 0M-10M, dan segmen usia 50-60 dan Lansia. Identifikasi faktor-faktor yang mempengaruhi respons dan konversi rendah dalam kluster ini, seperti penyebab potensial ketidakrelevanan produk atau kurangnya kesadaran merek. Kemudian, lakukan penyesuaian strategi pemasaran dan komunikasi yang lebih efektif untuk meningkatkan respons dan konversi.

Untuk selengkapnya, dapat melihat jupyter notebook disini