Rakamin Academy

Virtual Internship Experience

KALBE Nutritionals

Clustering Model

Disclaimer

Outline

1. *Clustering* Model
2. *K-Means Clustering*
3. Aplikasi Clustering
4. Studi Kasus

1 Clustering Model

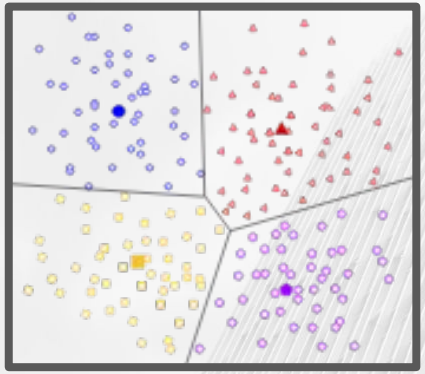
❑ Pengertian

Clustering adalah algoritma machine learning unsupervised yang mengelompokkan titik data bersama berdasarkan kesamaan mereka.

Tujuan dari clustering adalah untuk menemukan kelompok titik data yang mirip satu sama lain dan berbeda dari titik data dalam kelompok lain.

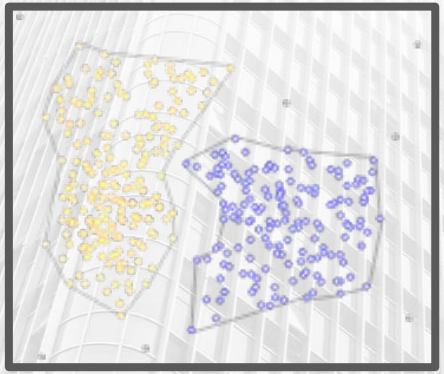
1 Clustering Model

❑ Jenis Clustering



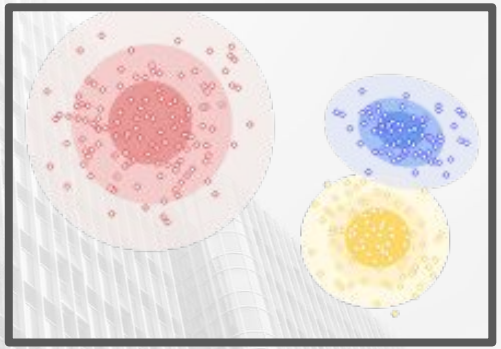
Gambar 1. Centroid-Based

Membagi data ke dalam sejumlah kluster tertentu, di mana setiap kluster memiliki sebuah sentroid yang mewakili rata-rata dari semua titik data dalam kluster tersebut.



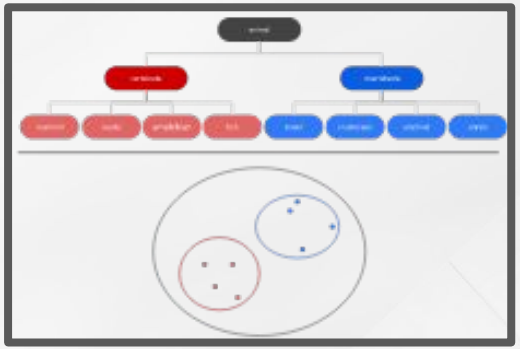
Gambar 2. Density-Based

Algoritma ini menemukan kluster yang padat dan terpisah dengan baik satu sama lain.



Gambar 3. Distribution-Based

Algoritma ini mengasumsikan bahwa titik data dalam setiap kelompok mengikuti distribusi tertentu, seperti distribusi Gaussian.



Gambar 4. Hierarchical

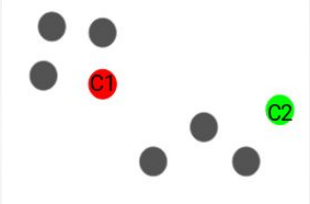
Algoritma ini membangun hierarki dari kelompok, dengan setiap kelompok menjadi anak dari kelompok yang lebih besar. Kelompok-kelompok tersebut digabungkan atau dibagi berdasarkan kesamaan mereka.

2 K-Means Clustering

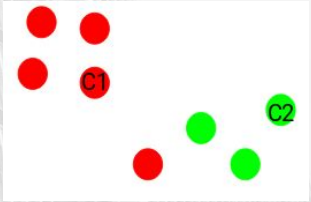
- Metode K-Means Clustering:

1. Pengguna menentukan jumlah kluster k.
2. Secara acak memilih k titik data sebagai pusat awal dari kluster.
3. mengelompokkan setiap titik data ke kluster dengan pusat terdekat.
4. Menghitung kembali nilai rata-rata centroid berdasarkan semua titik data yang telah dikelompokkan.
5. Mengulangi langkah 3 dan 4 sampai pusat tidak berubah lagi.

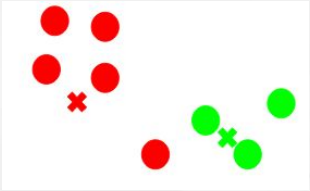
Gambar 1. Sebaran data



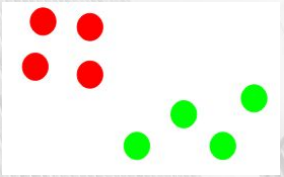
Gambar 2. pemilihan titik k secara random



Gambar 3. mengelompokkan data ke kluster terdekat



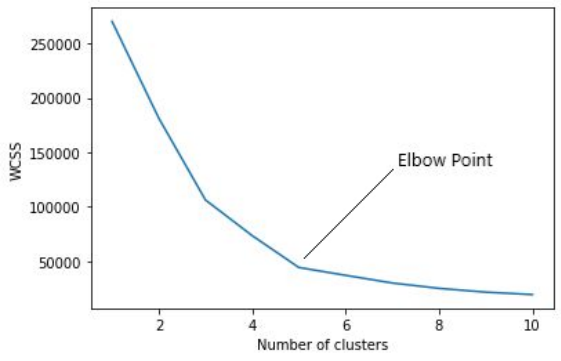
Gambar 4. Menghitung ulang pusat setiap kluster



Gambar 5. mengelompokkan data ke pusat baru

2 *K-Means Clustering*

- Menentukan banyaknya K (kluster / kelompok)

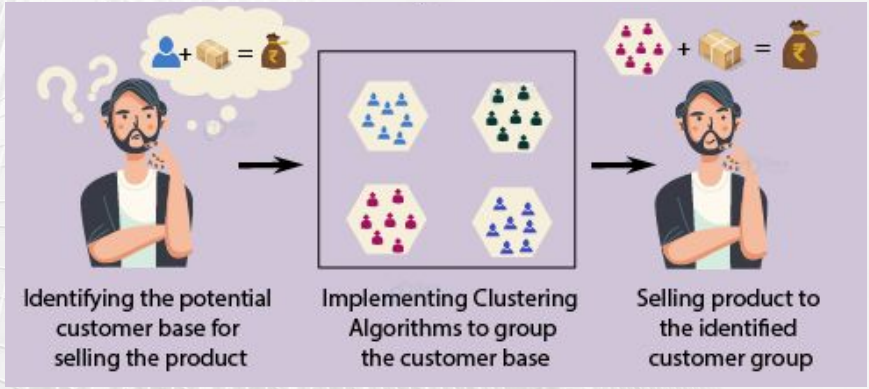


Gambar 6. Contoh Elbow Method

1. Mendefinisikan rentang nilai K untuk menjalankan *K-Means Clustering*
2. Mengevaluasi *Sum of Squared Errors* (SSE) untuk model menggunakan setiap jumlah klaster yang telah ditentukan.

3 Aplikasi Clustering

- **Segmentasi Pelanggan**



Gambar 1. Keuntungan Segmentasi pelanggan

- **Implementasi K-Means Clustering**

Gambar 2. Logo scikit-learn

4 Studi Kasus

Melihat penjualan produk yang menurun, tim marketing ingin membuat suatu program promo untuk meningkatkan penjualan produk. Sebelum memulai program, tim ingin membuat promo khusus kepada setiap pelanggan Kalbe Nutritionals. Sebagai seorang Data Scientist, anda memahami permasalahan tersebut dapat diselesaikan dengan clustering dan akan membuat model untuk permasalahan ini.

Petunjuk Pengerjaan:

1. Anda dapat mengerjakan menggunakan *jupyter notebook* atau *google colab*
2. Lakukan data cleaning dan eksplorasi data terdahulu.
3. Pastikan format data sudah sesuai, bila tidak anda terapkan *data transformation*.
4. selamat mencoba

Link Dataset

Solusi

- Memanggil Library yang diperlukan dan Load dataset
- Menampilkan informasi dataset
- Mengecek *missing values*
- Membuat histogram pada data pembelian dan data umur
- Membuat bar chart pada data jenis kelamin

Load Libraries dan Dataset

```
# Library untuk pengolahan data
import pandas as pd
import numpy as np

# Library untuk visualisasi data
import matplotlib.pyplot as plt
import seaborn as sns

# Library untuk machine learning model
from sklearn.cluster import KMeans
```

```
df = pd.read_csv('Dataset CCV 6 - Clustering Model.csv')
df.head()
```

	id_pelanggan	umur	jenis_kelamin	id_produk	nama_produk	kategori_produk	harga_pembelian
0	1	31	wanita	2	prenagen esensis	woman	54810
1	2	28	wanita	2	prenagen esensis	woman	55539
2	3	21	pria	3	hic1000 vitamin lemon	beverages	56225
3	4	38	pria	6	nutrive benecol	special needs	48295
4	5	42	pria	6	nutrive benecol	special needs	45963

Eksploratory Data Analysis (EDA)

```
[3] df.info()
```

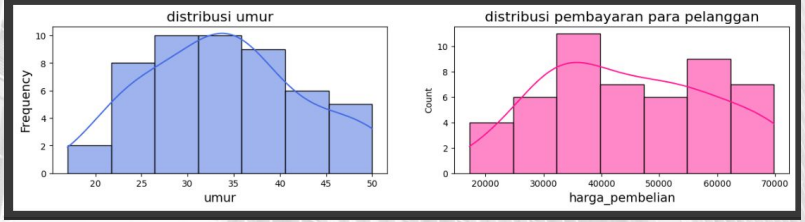
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id_pelanggan     50 non-null     int64
 1   umur             50 non-null     int64
 2   jenis_kelamin    50 non-null     object
 3   id_produk        50 non-null     int64
 4   nama_produk      50 non-null     object
 5   kategori_produk  50 non-null     object
 6   harga_pembelian  50 non-null     int64
dtypes: int64(4), object(3)
memory usage: 2.9+ KB
```

Gambar 1. Import libraries dan dataset

```
df.isnull().sum()
```

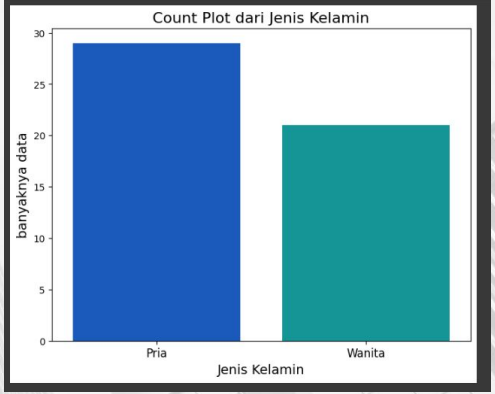
```
id_pelanggan     0
umur             0
jenis_kelamin    0
id_produk        0
nama_produk      0
kategori_produk  0
harga_pembelian  0
dtype: int64
```

```
plt.figure(figsize=(15,3))

plt.subplot(1,2,1)
sns.histplot(df['umur'], color='royalblue', kde = True)
plt.title('distribusi umur', fontsize=16)
plt.xlabel('umur', fontsize=14)
plt.ylabel('Frequency', fontsize=14)


plt.subplot(1,2,2)
sns.histplot(df['harga_pembelian'], color='deeppink', kde = True)
plt.title('distribusi pembayaran para pelanggan', fontsize=16)
plt.xlabel('harga_pembelian', fontsize=14)

plt.show()
```



Gambar 2. mengecek missing values dan visualisasi distribusi data

```
plt.figure(figsize=(8,6))
sns.countplot(x='jenis_kelamin', data=df, palette='winter')
plt.title('Count Plot dari Jenis Kelamin', fontsize=16)
plt.xlabel('Jenis Kelamin', fontsize=14)
plt.ylabel('banyaknya data', fontsize=14)
plt.xticks([0,1],['Pria','Wanita'],fontsize=12 )
plt.show()
```

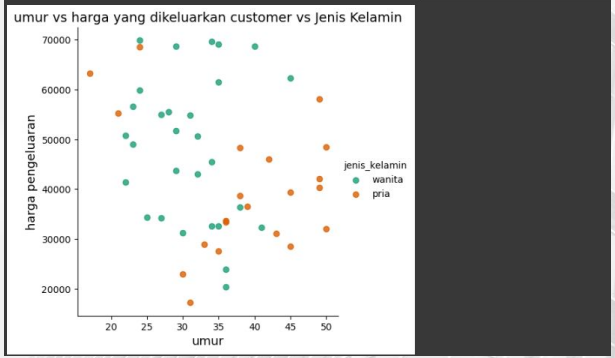


Gambar 3. membuat visualisasi jenis kelamin

Solusi

- Membuat scatterplot dengan tiga variabel:
  - Umur
  - Harga pembelian
  - Jenis Kelamin
- Membuat model clustering dengan K = 1 sampai K = 10 untuk menentukan cluster yang optimal

```
sns.lmplot(x='umur', y='harga_pembelian', data=df,fit_reg=False, hue='jenis_kelamin', height=5, palette='Dark2')
plt.title('umur vs harga yang dikeluarkan customer vs Jenis Kelamin', fontsize=14)
plt.xlabel('umur', fontsize=13)
plt.ylabel('harga pengeluaran', fontsize=13)
plt.show()
```



Gambar 4. Membuat visualisasi sebaran data antara 3 variabel

K-Means Clustering

Segmentasi berdadarkan Umur dan harga yang dikeluarkan

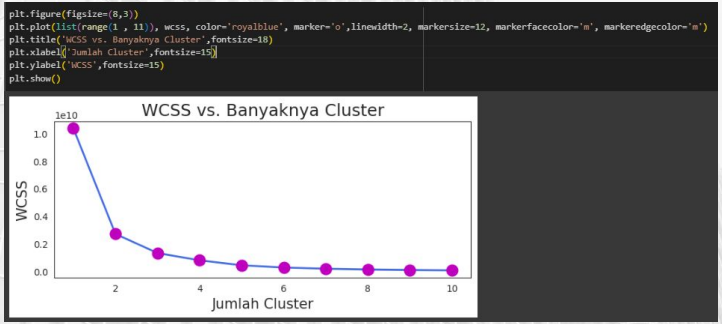
Menentukan K (banyaknya kelompok) yang optimal

```
[8] X1=df[['umur','harga_pembelian']]
    wcss=[]
    for n in range(1 , 11):
        model1=KMeans(n_clusters=n, init='k-means++', n_init=10, max_iter=300, tol=0.0001, random_state=100)
        model1.fit(X1)
        wcss.append(model1.inertia_)
    print(wcss)

[10367243957.220001, 2709910312.055195, 1326172182.178728, 797897839.7089127, 439957213.08888894, 271933067.5527777, 190306062.31944445, 132217680.18452379, 96276902.79999998, 72086028.63333333]
```

Gambar 5. Membangun model dengan K=1 sampai K=10

Solusi

- Membuat visualisasi antara jumlah cluster (K) dengan nilai WCSS (within cluster sum of squares) / elbow method
- Mendapati bahwa K optimal saat K= 3
- Membangun kembali model clustering dengan K = 3
- Melatih model dengan data harga pembelian dan umur pelanggan



Gambar 6. elbow method

Membuat model clustering dengan K yang optimal

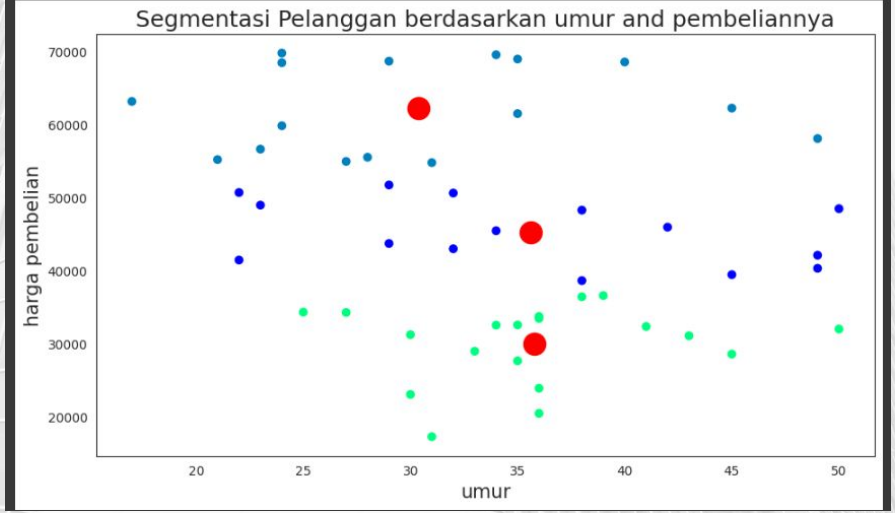
```
model1=KMeans(n_clusters=3, init='k-means++', n_init=10, max_iter=300, tol=0.0001, random_state=100)
model1.fit(X1)
labels1=model1.labels_
centroids1=model1.cluster_centers_
```

Gambar 7. Membuat model K-Means dengan K=3

Solusi

- Model telah membuat cluster pada sebaran data
- Membuat visualisasi persebaran data untuk menunjukkan segmentasi pelanggan dengan fitur umur dan harga pembelian produk

```
plt.figure(figsize=(11,6))
sns.set_style('white')
plt.scatter(x=df['umur'], y=df['harga_pembelian'],c=labels1,cmap='winter')
plt.scatter(x = centroids1[: , 0] , y =  centroids1[: , 1], s = 300 , c = 'red')
plt.xlabel('umur',fontsize=15)
plt.ylabel('harga pembelian',fontsize=15)
plt.title('Segmentasi Pelanggan berdasarkan umur and pembeliannya',fontsize=18)
plt.show()
```



Gambar 8. Hasil akhir segmentasi pelanggan berdasar umur dan harga beli


Thank You!
KALBE
Nutritionals