

# CLUSTERING NASABAH MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING PADA BANKALTIMTARA

Habibi <sup>1</sup>, Wahyuni <sup>2</sup>, Ita Arfyanti <sup>3</sup>

<sup>123</sup>Program Studi Teknik Informatika, STMIK Widya Cipta Dharma

<sup>123</sup>Jl. M. Yamin No.25, Gn. Kelua Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75123

E-mail: [2143903@wicida.ac.id](mailto:2143903@wicida.ac.id)<sup>1</sup>, [wahyuni@wicida.ac.id](mailto:wahyuni@wicida.ac.id)<sup>2</sup>, [ita@wicida.ac.id](mailto:ita@wicida.ac.id)<sup>3</sup>

## ABSTRAK

Penelitian ini bertujuan untuk mengelompokkan nasabah Bankaltimtara berdasarkan karakteristik tertentu menggunakan algoritma *K-Means Clustering*. Dalam proses pengembangan penelitian, metode *CRISP-DM (Cross-Industry Standard Process for Data Mining)* digunakan sebagai kerangka kerja yang sistematis. Metode ini terdiri dari enam fase utama: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan.

Fase pemahaman bisnis dilakukan untuk mengidentifikasi tujuan dan kebutuhan dari Bankaltimtara terkait segmentasi nasabah. Fase pemahaman data melibatkan pengumpulan data nasabah yang relevan, seperti data demografis dan transaksi. Selanjutnya, pada fase persiapan data, dilakukan proses pembersihan dan transformasi data agar siap digunakan dalam analisis.

Pada fase pemodelan, algoritma *K-Means Clustering* diterapkan untuk mengelompokkan nasabah ke dalam beberapa segmen berdasarkan kesamaan karakteristik. Proses evaluasi dilakukan untuk menilai keakuratan dan relevansi model yang dihasilkan. Terakhir, fase penerapan mencakup implementasi model dalam lingkungan operasional Bankaltimtara serta analisis hasil untuk memberikan rekomendasi strategi pemasaran yang lebih tepat sasaran.

Hasil dari penelitian ini menunjukkan bahwa algoritma *K-Means Clustering* mampu mengidentifikasi beberapa segmen nasabah yang berbeda, yang dapat digunakan oleh Bankaltimtara untuk meningkatkan pelayanan dan strategi pemasaran. Dengan segmentasi nasabah yang lebih terstruktur, Bankaltimtara dapat menawarkan produk dan layanan yang lebih sesuai dengan kebutuhan masing-masing segmen nasabah.

**Kata Kunci :** *K-Means, Clustering, Crisp-DM, Data Mining, Perbankan.*

## 1. PENDAHULUAN

Pertumbuhan pesat sektor perbankan di Kalimantan Timur dan Kalimantan Utara menunjukkan persaingan yang tinggi. Bank harus mengadopsi pendekatan cerdas dalam analisis data nasabah untuk memahami perilaku nasabah dan meningkatkan pangsa pasar.

Pentingnya pengelolaan data dan analisis untuk mengidentifikasi pola-pola tersembunyi dalam perilaku nasabah guna membuat strategi yang tepat. Bank perlu mengelompokkan nasabah berdasarkan karakteristik tertentu untuk memberikan layanan yang personal dan relevan.

Bankaltimtara saat ini menggunakan data nasabah, media sosial, dan kuesioner untuk memahami kebutuhan nasabah, namun masih memiliki kekurangan dalam segmentasi. Penggunaan metode *K-Means Clustering* dapat meningkatkan efektivitas pengelolaan nasabah.

Metode *K-Means Clustering* dapat otomatis mengelompokkan data tanpa label kelas sebelumnya, membantu Bank mengidentifikasi segmen nasabah yang berbeda, dan merancang strategi yang sesuai. Bank dapat meningkatkan targeting iklan, mengoptimalkan alokasi sumber daya, dan meningkatkan kepuasan nasabah dengan teknologi analisis data.

Penelitian "*Clustering Nasabah Menggunakan Algoritma K-Means Clustering pada Bankaltimtara*" bertujuan mengelompokkan nasabah guna meningkatkan layanan, memperluas pangsa pasar, serta memperkuat posisi Bankaltimtara di industri perbankan yang kompetitif.

## 2. RUANG LINGKUP PENELITIAN

### 2.1 RUMUSAN MASALAH

Berdasarkan latar belakang tersebut, maka rumusan masalah dalam penelitian ini adalah "*Bagaimana melakukan Clustering Nasabah Menggunakan Algoritma K-Means Clustering pada Bankaltimtara ?*".

### 2.2 BATASAN MASALAH

Untuk mengatasi masalah yang ada, maka penulis membatasi permasalahan sebagai berikut:

1. Penelitian ini berfokus pada pembuatan Clustering nasabah pada Bankaltimtara.
2. Penelitian ini menggunakan data nasabah periode Januari 2014 hingga April 2024 yang dimiliki oleh Bankaltimtara khusus pada nasabah Konvensional.
3. Data transaksi yang digunakan pada penelitian ini adalah data periode April 2024 - Mei 2024.

- Penelitian ini menggunakan Metode K-Means Clustering untuk melakukan Clustering pada Nasabah Bankaltimtara
- Untuk proses deployment pada website menggunakan tools streamlit.

### 2.3 TUJUAN PENELITIAN

Tujuan yang akan dicapai dalam penelitian ini adalah untuk membantu Bankaltimtara mengelompokkan nasabah menggunakan Algoritma *K-Means Clustering* dengan lebih tepat dan efektif. Penelitian ini bertujuan untuk mengidentifikasi segmen-segmen nasabah berdasarkan karakteristik yang relevan, meningkatkan pemahaman Bankaltimtara terhadap perilaku dan kebutuhan nasabah konvensional, serta mengoptimalkan layanan dan produk perbankan sesuai dengan segmen nasabah yang telah diidentifikasi.

## 3. BAHAN DAN METODE

Adapun bahan dan metode yang digunakan dalam membangun penelitian ini yaitu:

### 3.1 Clustering

Clustering adalah metode pengelompokan data yang digunakan untuk menghasilkan pengetahuan baru tentang data yang diolah (Xie, Lee, Wang, Chen, & Zhou, 2020). Dalam satu kelompok data memiliki kesamaan karakteristik yang berbeda dengan kelompok data lainnya. (Oyewole & Thopil, 2022). Metode Klustering telah banyak digunakan diberbagai sektor seperti, manufaktur, transportasi, ilmu kesehatan, pendidikan dan ritel

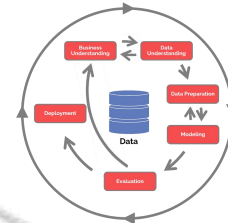
### 3.2 Algoritma K-Means

K-Means adalah algoritma clustering yang digunakan pada machine learning dan data mining. K-Means salah satu bentuk algoritma unsupervised learning yang bertujuan mengelompokkan objek yang hamper sama pada suatu wilayah. Algoritma ini melakukan perulangan untuk mengelompokkan dataset yang belum memiliki label ke dalam kluster yang memiliki properti data yang serupa (Ikotun, Ezugwu, Abualigah, Abuhajja, & Heming, 2023). Tahapan pada algoritma K-Means adalah sebagai berikut :

- Menentukan jumlah K kluster yang ingin dibentuk. Jumlah K Kluster yang dibentuk tidak boleh lebih besar atau sama dengan dengan jumlah data yang tersedia.
- Menentukan titik K yang menjadi pusat data/centroid dan dapat ditentukan secara acak/random untuk setiap atribut.
- Menghitung jarak data ke setiap centroid kluster yang dibentuk dan menempatkan data pada centroid kluster terdekat.
- Hitung ulang pusat kelompok dengan mengambil rata-rata dari seluruh data dalam kelompok.
- Ulangi tahap 3 dan 4 sampai tidak ada data yang berubah posisi kelompoknya atau sampai iterasi/perulangan yang ditentukan selesai.

### 3.3 CRISP-DM

*CRISP-DM* (Cross Industry Standard Process for Data Mining) suatu standarisasi pemrosesan data mining yang telah dikembangkan dimana data yang ada akan melewati setiap fase terstruktur dan terdefinisi dengan jelas dan efisien (Hasanah, Soim, & Handayani, 2021).



Gambar 1. *Crisp-DM*

Proses data mining berdasarkan CRISP-DM (Cross Industry Standard Process for Data Mining) terdiri dari enam fase yaitu :

#### 1. Business Understanding

Proses pengumpulan kebutuhan berdasarkan kasus yang akan dilakukan. Tahapan ini memerlukan pemahaman mendalam tentang tujuan bisnis, kebutuhan pengguna dan keberhasilan terkait.

#### 2. Data Understanding

Proses ini meliputi pengumpulan data yang digunakan pada data mining. Data ini akan di eksplorasi terkait sumber data, karakteristik, kualitas dan kecocokan atribut data dengan tujuan data mining.

#### 3. Data Preparation

Data yang terkumpul akan diolah untuk dianalisis lebih lanjut. Hal ini mencakup pembersihan data, menggabungkan data dari berbagai sumber, memilih atribut yang relevan, dan mengubah data jika diperlukan.

#### 4. Modeling

Pada proses ini melakukan Teknik data mining untuk pemodelan data sesuai dengan tujuan penelitian. Metode ini dilakukan untuk analisis dan ekstraksi informasi.

#### 5. Evaluation

Tahapan ini merupakan tahapan evaluasi hasil analisis pemodelan data. Evaluasi ini mencakup penggunaan metrik dan teknik validasi yang tepat untuk mengukur kinerja model.

#### 6. Deployment

Tahapan terakhir adalah mengimplementasikan solusi yang dihasilkan dari analisis data ke dalam lingkungan produksi.

### 3.4 Metode Slovin

Rumus slovin adalah salah satu rumus yang dipelajari dalam statistika. Rumus slovin ini diterapkan dalam penentuan banyaknya sampel minimum yang diperlukan dalam suatu penelitian. (Erno Sumantri, 2024).

$$n = \frac{N}{1 + N \cdot e^2}$$

Keterangan :

n = ukuran sample

N = Ukuran Populasi

e = margin of error

#### 4. RANCANGAN SISTEM ATAU APLIKASI

Pada tahapan ini memberikan suatu gambaran tentang hasil penelitian yang telah dilaksanakan.

##### 4.1 Business Understanding

Peneliti melakukan pemahaman bisnis untuk analisis pengelompokan nasabah Bankaltimara menggunakan algoritma K-Means Clustering. Bankaltimara adalah lembaga keuangan dengan layanan perbankan di Kalimantan Timur dan Kalimantan Utara. Bankaltimara membutuhkan cara efektif mengelompokkan nasabah untuk meningkatkan pelayanan dan strategi pemasaran. Penelitian ini bertujuan mengidentifikasi segmen nasabah berdasarkan data demografis dan perilaku transaksi.

##### 4.2 Data Understanding

Pada tahap ini peneliti melakukan proses ekstraksi data melalui query pada sistem corebanking dan mengeksportnya dalam format file CSV. Selanjutnya, data tersebut diimpor ke Jupyter Notebook menggunakan Bahasa pemrograman Python.

```
0 | # proses import library yang dibutuhkan
1 | import re
2 | import numpy as np
3 | import pandas as pd
4 | import matplotlib.pyplot as plt
5 | %matplotlib inline
6 | plt.style.use("fivethirtyeight")
7 | import seaborn as sns
8 | try:
9 |     import plotly.express as px
10 |     import plotly.graph_objects as go
11 |
12 | except:
13 |     !pip install plotly
14 |     import plotly.express as px
15 |     import plotly.graph_objects as go
16 | import warnings
17 | warnings.simplefilter(action='ignore', category=FutureWarning)
18 | warnings.simplefilter(action='ignore',
19 |                       category=DeprecationWarning)
20 | warnings.simplefilter(action='ignore', category=RuntimeWarning)
21 |
22 |
23 | import scipy.cluster.hierarchy as sch
24 | from sklearn.preprocessing import StandardScaler
25 | from sklearn.cluster import KMeans
26 | from sklearn.decomposition import PCA
27 | from sklearn.metrics import silhouette_score
28 |
29 | try:
30 |     from kneed import KneeLocator
31 | except:
32 |     !pip install kneed
33 |     from kneed import KneeLocator
34 | #-----
35 | try:
36 |     from yellowbrick.cluster import KElbowVisualizer
37 | except:
38 |     !pip install -U yellowbrick
39 |     from yellowbrick.cluster import KElbowVisualizer
```

Gambar 2. Sintak import library pada python

Gambar 2 menunjukkan berbagai pustaka atau file pendukung yang diperlukan untuk proses pengelompokan data pada Jupyter Notebook. Pustaka-

pustaka ini akan digunakan oleh peneliti, seperti *numpy*, *pandas*, *matplotlib.pyplot*, *seaborn*, *plotly*, *scipy*, *KMeans*, *PCA*, *Silhouette score*, *KneeLocator*, dan *KElbowVisualizer*.

Proses impor data CSV ke dalam dataframe di mana diambil sampel sebanyak 10.000 data atau baris sesuai dengan ukuran sampel yang telah ditentukan dalam populasi. Data tersebut kemudian dibentuk ke dalam dataframe untuk pengolahan selanjutnya. Dengan menggunakan sintaks *df.info()*, kita dapat melihat isi dari data yang telah dibentuk seperti yang terlihat pada gambar 4.5. Dari situ, kita dapat mengetahui nama kolom, tipe data, dan jumlah total data dalam file tersebut, yaitu sebanyak 7.617.931 baris data.

```
#proses untuk mengimport data yang akan diproses
df = pd.read_csv("D:\Skripsi\jupyter\dataset.csv", delimiter=';', quotechar='\"', doublequote=True, on_bad_lines='skip')
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7617931 entries, 0 to 7617930
Data columns (total 14 columns):
 #   Column  Dtype
---  ---
 0   BMNICO  int64
 1   PORECO  int64
 2   CUUCOE  int64
 3   RESUME  object
 4   REATCO  int64
 5   CUOTLH  int64
 6   CUUEKL  object
 7   CUADR3  object
 8   NOMINAL int64
 9   POREFN  object
10  PODISP  object
11  PODISC  object
12  PODTPO  int64
13  SALDO   int64
dtypes: int64(8), object(6)
memory usage: 811.7+ MB
```

Gambar 3. data deskriptif dataset.csv

Manfaat dari pemahaman data deskriptif adalah membantu dalam memahami struktur data, mengidentifikasi kolom dengan nilai yang kosong, menganalisis tipe data setiap kolom, dan mengoptimalkan penggunaan memori untuk analisis data lebih lanjut. Dari data di atas, terungkap bahwa data tersebut mencakup informasi seperti pada gambar 3 diatas.

##### 4.3 Data Preparation

```
[1]: df.isnull().sum()
[1]:
CUUCOE 0
CUOTLH 0
CUUEKL 0
CUADR3 0
NOMINAL 0
POREFN 0
PODISP 0
PODISC 0
SALDO 0
dtypes: int64

[7]: df.duplicated().sum()
[7]: 0
```

Gambar 4. Cek data Null dan Duplikat

Pada tahapan ini peneliti melakukan proses pembersihan data, transformasi data dan pra pengolahan data agar data yang telah didapat dapat dilakukan proses selanjutnya. Pada proses pembersihan data kita melakukan pengecekan data null, data duplikat dan data unik yang seharusnya tidak ada dalam dataframe. Pada data yang didapat terdapat kolom jenis kelamin dimana ternyata ada isinya ada data "P", "L" dan " " sesuai data yang seharusnya hanya ada data "P" dan "L".

```
[5]: df["NOMINAL"] = df["NOMINAL"].abs() #mengubah nilai pada kolom nominal menjadi nilai absolut karena sebelumnya nilainya minus
df.head()

[5]:
   CUUCOE  CUOTLH  CUUEKL  CUADR3  NOMINAL  POREFN  PODTPO  SALDO
232703  12799121  19920429  P  SAMARINDA  3000000  00000024195  20240401  54500
5833844  850865  19900626  L  BONTANG  54000  240503279739  20240503  3008516
3581072  294039  19946719  P  PASER  1000000  3361KTM00209  20240418  37344300
5048812  368176  19860225  L  BALIKPAPAN  6500  00001114564FEE  20240428  64220
325700  13400683  19790101  L  PASER  3119200  000000002689  20240402  1586886
```

Gambar 5. Hasil Setelah Transformasi

Pada gambar 5 diatas adalah hasil setelah dilakukan transformasi data pada kolom Nominal dimana data sebelumnya bernilai negatif menjadi nilai positif.

Selanjutnya pada proses pra pengolahan data peneliti melakukan distribusi data berdasarkan umur dan lokasi serta melakukan penanganan *outliers* serta melakukan analisis terhadap perilaku nasabah dengan menggunakan metode RFM yang merupakan metode menghitung berapa lama nasabah melakukan transaksi kembali (*Recency*), berapa kali nasabah melakukan transaksi (*Frequency*) dan berapa jumlah nominal transaksi yang dilakukan nasabah (*Monetary*).

Setelah tahapan tersebut peneliti melakukan proses normalisasi terhadap data tersebut menggunakan *library StandarScaler* sehingga data yang tersedia dimungkinkan untuk dilakukan penghitungan oleh *library k-means* dengan lebih efisien.

#### 4.4 Modelling K-Means

Dalam proses modelling K-Means Clustering dilakukan beberapa langkah yaitu penentuan jumlah kluster dengan Elbow Method dan mengujinya dengan data yang telah disiapkan dan memvisualisasikan hasil klusteringsnya.

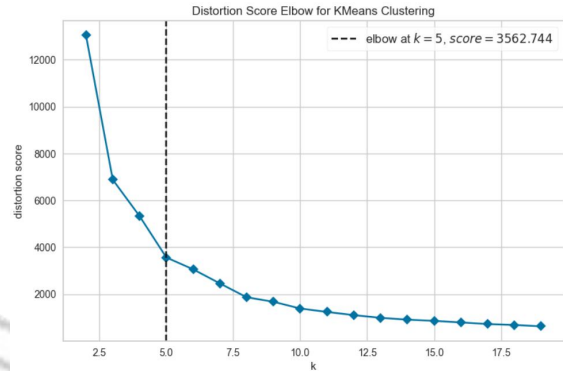
```

1 | import random
2 | from yellowbrick.cluster import KElbowVisualizer
3 | from sklearn.cluster import KMeans
4 |
5 | # mendefinisikan ilai X dan Y dari dataframe yang telah kita
   tentukan
6 | X = df2[['UMUR', 'NOMINAL']].values
7 | Y = df2[['CUADR3', 'NOMINAL']].values
8 |
9 | # menjalankan K-Means Clustering dan menghitung jumlah K
   optimal menggunakan Elbow Method
10| def perform_elbow_method(X, k_range, title):
11|     model = KMeans(init='k-means++', random_state=42)
12|     visualizer = KElbowVisualizer(model, k=k_range, timings=False)
13|     visualizer.fit(X)
14|     plt.title(title)
15|     visualizer.show()
16|
17| perform_elbow_method(X, k_range=(2, 20), title='Elbow Method
   untuk Clustering berdasarkan Umur Nasabah')
18| perform_elbow_method(Y, k_range=(2, 20), title='Elbow Method
   untuk Clustering berdasarkan Lokasi Nasabah')

```

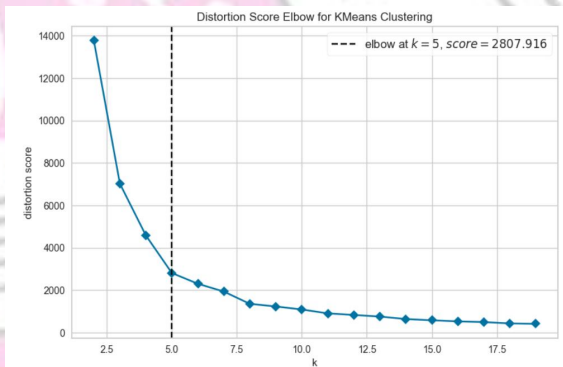
**Gambar 6. sintaks penggunaan metode Elbow**

Pada Gambar 6 menggunakan K-Means Clustering dan Elbow Method untuk menentukan jumlah cluster optimal bagi data nasabah Bankaltimtara berdasarkan dua set fitur yang berbeda: umur dan nominal transaksi, serta lokasi dan nominal transaksi. Dengan mengimpor pustaka yang diperlukan dan mendefinisikan data yang akan dianalisis, fungsi `perform\_elbow\_method` dijalankan untuk memvisualisasikan jumlah cluster optimal yang dapat digunakan oleh peneliti. Hasilnya adalah visualisasi yang membantu dalam pemilihan jumlah cluster yang tepat, memberikan wawasan yang lebih baik untuk segmentasi nasabah berdasarkan umur dan lokasi, yang dapat digunakan untuk analisis dan strategi pemasaran lebih lanjut.



**Gambar 7. Hasil metode elbow berdasarkan Umur**

Gambar 7 tersebut menampilkan grafik "Elbow Method" yang digunakan untuk menentukan jumlah cluster optimal dalam algoritma K-Means.. Titik elbow, yang ditandai dengan garis putus-putus, menunjukkan titik di mana penurunan nilai distorsi mulai melambat secara signifikan, yang dalam hal ini terjadi pada k = 5. Ini menunjukkan bahwa lima cluster adalah jumlah optimal, karena menambahkan lebih banyak cluster di luar titik ini memberikan pengurangan distorsi yang semakin kecil.



**Gambar 8. Hasil metode elbow berdasarkan Lokasi**

Pada gambar 8 menunjukkan hasil visualisasi dari jumlah nilai K yang optimal yang telah dibentuk melalui elbow method dan sama seperti pada penentuan cluster berdasarkan umur pada gambar 4.23 untuk penentuan jumlah cluster atau nilai K pada cluster berdasarkan lokasi juga didapatkan nilainya K= 5.

```

1 | # Function to perform KMeans clustering and return the cluster
   labels and centroids
2 | def perform_kmeans_clustering(X, num_clusters):
3 |     kmeans = KMeans(n_clusters=num_clusters, init='k-means++',
   random_state=42)
4 |     cluster_labels = kmeans.fit_predict(X)
5 |     cluster_centers = kmeans.cluster_centers_
6 |     return cluster_labels, cluster_centers
7 |
8 | x_cluster_labels, x_cluster_centers = perform_kmeans_clustering(X,
   num_clusters=5)
9 | y_cluster_labels, y_cluster_centers = perform_kmeans_clustering(Y,
   num_clusters=5)

```

**Gambar 9. sintaks penggunaan library k-means**

Kode pada gambar 9 mendefinisikan fungsi `perform\_kmeans\_clustering` yang melakukan clustering menggunakan algoritma K-Means pada data yang

diberikan dan mengembalikan label cluster serta centroid-nya. Fungsi ini menerima data `X` dan jumlah cluster (`num\_clusters`) sebagai parameter, kemudian menjalankan K-Means dengan inialisasi `k-means++` dan state acak yang tetap. Setelah itu, fungsi ini memprediksi label cluster untuk data dan menentukan pusat cluster, lalu mengembalikannya. Fungsi tersebut diterapkan pada dua set data: `X` dengan 5 cluster dan `Y` dengan 5 cluster, menghasilkan label cluster dan pusat cluster untuk masing-masing set data.

```
[36]: print("x_cluster_labels",x_cluster_labels)
      print("x_cluster_centers",x_cluster_centers)
      print("y_cluster_labels",y_cluster_labels)
      print("y_cluster_centers",y_cluster_centers)

x_cluster_labels [0 3 0 ... 4 1 0]
x_cluster_centers [[-1.01354113e+00 -1.04829358e-01]
 [ 1.43492169e+00  2.45209141e-02]
 [ 5.77245993e-01  2.46535559e+01]
 [ 1.62133048e-01 -6.26676920e-02]
 [ 4.79522487e-01  5.56067391e+00]]
y_cluster_labels [2 3 2 ... 4 2 0]
y_cluster_centers [[-0.1391267 -0.0584921 ]
 [ 0.20541661 24.65355586]
 [ 1.09191462 -0.05574106]
 [-1.42232623 -0.05592888]
 [-0.08238619  6.00199241]]
```

**Gambar 10. hasil pelabelan clustering**

Hasil clustering menunjukkan bahwa data telah dikelompokkan ke dalam cluster yang berbeda berdasarkan variabel yang dianalisis. Pusat cluster memberikan wawasan tentang nilai rata-rata dari setiap cluster di dalam masing-masing dimensi, yang dapat digunakan untuk memahami karakteristik utama dari setiap kelompok. Misalnya, dengan melihat `x\_cluster\_centers` dan `y\_cluster\_centers`, kita bisa melihat bahwa cluster dengan pusat terdekat ke asal (0,0) mungkin mewakili kelompok dengan nilai lebih rendah, sedangkan cluster dengan pusat lebih jauh mungkin mewakili kelompok dengan nilai lebih tinggi.

```
1 | # Add cluster labels to the DataFrame
2 | df2['cluster_umur'] = x_cluster_labels
3 | df2['cluster_lokasi'] = y_cluster_labels
4 |
5 | # Add centroids to the DataFrame
6 | df2['cen_xx'] = df2.cluster_umur.map({i: x_cluster_centers[i][0] for
i in range(5)})
7 | df2['cen_xy'] = df2.cluster_umur.map({i: x_cluster_centers[i][1] for
i in range(5)})
8 | df2['cen_yx'] = df2.cluster_lokasi.map({i: y_cluster_centers[i][0]
for i in range(5)})
9 | df2['cen_yy'] = df2.cluster_lokasi.map({i: y_cluster_centers[i][1]
for i in range(5)})
10|
11| df2.head()
```

**Gambar 11. sintak untuk menambahkan kolom baru cluster**

Pada gambar 11 tersebut menambahkan data pada DataFrame `df2` dengan hasil K-Means clustering yaitu label cluster dan koordinat centroid untuk setiap cluster yang dihasilkan. Kolom baru `cluster\_umur` dan `cluster\_lokasi` menunjukkan label cluster berdasarkan umur dan lokasi nasabah, sementara kolom `cen\_xx`, `cen\_xy`, `cen\_yx`, dan `cen\_yy` menyimpan koordinat centroid dari masing-masing cluster. Dengan ini, DataFrame `df2` setelah ini peneliti bisa untuk melakukan Analisa lebih lanjut dan visualisasi dari

distribusi dan karakteristik nasabah berdasarkan clustering yang telah terbentuk.

```
1 | # Define colors for each cluster
2 | colors_X = ['red', 'yellow', 'grey', 'green', 'blue', 'orange', 'pink']
3 | colors_Y = ['red', 'yellow', 'grey', 'green', 'blue', 'orange']
4 | # Add cluster colors to the DataFrame
5 | df2['color_umur km'] = df2.cluster_umur.map({i: colors_X[i] for i
in range(5)})
6 | df2['color_lokasi km'] = df2.cluster_lokasi.map({i: colors_Y[i] for i
in range(5)})
```

**Gambar 12. sintak untuk melabeli cluster dengan warna**

Kode program pada gambar 12 tersebut berfungsi untuk menambahkan warna pada setiap cluster ke dalam DataFrame `df2` untuk memudahkan visualisasi hasil clustering. Baris 2 dan 3 mendefinisikan daftar warna yang akan digunakan untuk setiap cluster pada dua set fitur yang berbeda: satu berdasarkan umur (`colors\_X`) dan satu lagi berdasarkan lokasi (`colors\_Y`). Kemudian, baris 5 dan 6 menambahkan kolom `color\_umur\_km` dan `color\_lokasi\_km` ke DataFrame `df2`, yang berisi warna yang sesuai dengan setiap label cluster. Warna-warna ini ditetapkan dengan memetakan label cluster (`cluster\_umur` dan `cluster\_lokasi`) ke daftar warna yang telah didefinisikan, sehingga setiap nasabah dapat diidentifikasi dengan warna yang mewakili cluster mereka. dan visualisasi dari hasil cluster tersebut bisa dilihat pada gambar 13 dibawah.

Secara manual kita akan mencoba menghitung pembentukan kluster dari contoh 3 data nasabah berdasarkan Umur dan Nominal transaksi dan diasumsikan pembentukan nilai K= 2 sebagai berikut:

**Tabel 1 Contoh Dataset Nasabah**

Data Ke	Umur	Nominal
1	32	300000
2	34	50000
3	28	100000

Lalu kita tentukan Centroid 1 yaitu [32,300000] dan Centroid 2 [34,50000] setelah kita tentukan Centroid 1 dan Centroid 2 kita melakukan perhitungan nilai jarak Euclidean dari dataset tersebut.

1. Jarak dari dataset ke C1

$$(Data\ 1\ ke\ C1) = \sqrt{(32 - 32)^2 + (300000 - 300000)^2} = 0$$

$$(Data\ 2\ ke\ C1) = \sqrt{(34 - 32)^2 + (50000 - 300000)^2} \approx 250000$$

$$(Data\ 3\ ke\ C1) = \sqrt{(28 - 32)^2 + (100000 - 300000)^2} \approx 200000.04$$

2. Jarak dari dataset ke C2

$$(Data\ 1\ ke\ C2) = \sqrt{(32 - 34)^2 + (300000 - 50000)^2} \approx 250000$$

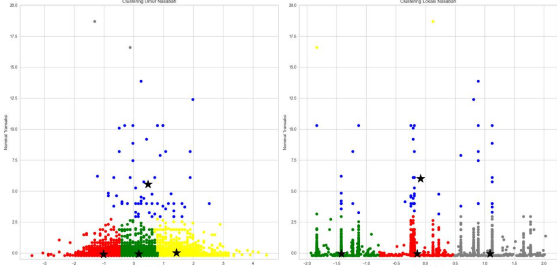
$$(Data\ 2\ ke\ C2) = \sqrt{(34 - 34)^2 + (50000 - 50000)^2} = 0$$

$$(Data\ 3\ ke\ C2) = \sqrt{(28 - 34)^2 + (100000 - 50000)^2} =$$

Dari perhitungan diatas kita bisa melihat bahwa data 1 lebih dekat ke C1 dan data 2 serta data 3 lebih dekat ke C2 dan setelah itu kita hitung kembali pusat dari centroid 2 atau C2 dengan nilai rata-rata dari data 2 dan data 3.

$$C2 = \frac{(34 + 28)}{2}, \frac{(50000 + 100000)}{5} = 31,75000$$

Dari perhitungan kembali centroid 2 maka diketahui bahwa nilai centroid 2 berubah menjadi [31,75000] dan perhitungan terus dilakukan hingga tidak ada perubahan dari nilai centroidnya.



Gambar 13. hasil visualisasi pembentukan cluster

Pada Gambar 13 menunjukkan hasil klustering nasabah berdasarkan dua dimensi utama: umur dan nominal transaksi (grafik kiri) serta lokasi dan nominal transaksi (grafik kanan). Pada grafik kiri, nasabah dikelompokkan menjadi beberapa cluster berdasarkan usia mereka dan jumlah nominal transaksi yang mereka lakukan, di mana setiap warna menunjukkan cluster yang berbeda. Bintang hitam menunjukkan posisi centroid dari setiap cluster. Grafik kanan menunjukkan klustering yang serupa tetapi berdasarkan lokasi nasabah dan nominal transaksi mereka. Distribusi nasabah dalam kedua grafik menunjukkan bahwa ada variasi yang signifikan dalam data, dan beberapa cluster memiliki centroid yang berdekatan, menandakan adanya kesamaan karakteristik antar nasabah dalam cluster tersebut. Clustering ini membantu dalam memahami segmen-segmen nasabah yang berbeda dan memungkinkan analisis yang lebih mendalam terhadap perilaku transaksi mereka.

#### 4.5 Model Evaluation

```
1 | from sklearn.metrics import silhouette_score
2 | # Silhouette Score for cluster umur
3 | silhouette_score_umur = silhouette_score(df2[['NOMINAL',
4 | 'UMUR']], df2['cluster umur'])
5 | print("Silhouette Score for cluster umur:", silhouette_score_umur)
6 | # Silhouette Score for cluster lokasi
7 | silhouette_score_lokasi =
silhouette_score(df2[['CUADR3', 'NOMINAL']], df2['cluster lokasi'])
8 | print("Silhouette Score for cluster_lokasi:", silhouette_score_lokasi)
```

```
from sklearn.metrics import silhouette_score
# Silhouette Score for cluster_umur
silhouette_score_umur = silhouette_score(df2[['NOMINAL', 'UMUR']], df2['cluster_umur'])
print("Silhouette Score for cluster_umur:", silhouette_score_umur)
# Silhouette Score for cluster_lokasi
silhouette_score_lokasi = silhouette_score(df2[['CUADR3', 'NOMINAL']], df2['cluster_lokasi'])
print("Silhouette Score for cluster_lokasi:", silhouette_score_lokasi)

Silhouette Score for cluster_umur: 0.487272593738574
Silhouette Score for cluster_lokasi: 0.6538875119797889
```

Gambar 13. hasil dari penghitungan Silhouette score

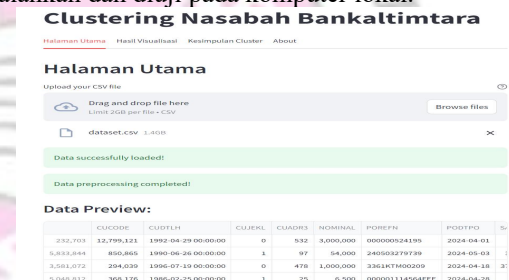
Dari gambar 13 bisa dilihat bahwa hasil dari perhitungan Silhouette Score untuk data sampling menunjukkan nilai untuk cluster\_umur 0.487272... dan cluster\_lokasi 0.653887..... nilai tersebut mendekati nilai 1 yang artinya bahwa cluster yang terbentuk adalah baik

dan menunjukkan bahwa titik data berada sangat dekat dengan titik-titik yang lain dalam cluster yang sama dan jauh dari titik-titik cluster yang lain.

## 5 IMPLEMENTASI

Setelah model sudah selesai dibangun dan telah dilakukan evaluasi atas model tersebut maka langkah selanjutnya peneliti akan mengimplementasikan model yang sudah dibangun dalam bentuk web. Dalam penelitian ini peneliti menggunakan web service dari Streamlit dengan menggunakan server local. langkah pertama yang perlu dilakukan adalah membuat folder proyek baru dan menyalin semua file yang diperlukan ke dalam folder tersebut. File-file yang diperlukan antara lain `functions.py` yang berisi fungsi-fungsi pendukung, `app.py` yang merupakan aplikasi utama Streamlit, dan dataset yang diperlukan.

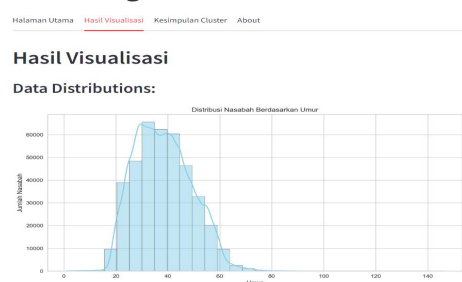
Setelah semua dependensi terinstal, jalankan aplikasi Streamlit secara lokal untuk memastikan bahwa semuanya berfungsi dengan baik. Gunakan perintah `streamlit run app.py` untuk menjalankan aplikasi. Buka browser dan arahkan ke `http://localhost:8501` untuk melihat tampilan aplikasi. Pastikan semua fungsi berjalan sesuai dengan yang diharapkan dan aplikasi dapat menampilkan hasil clustering dengan baik. Dengan mengikuti langkah-langkah ini, deployment model K-Means ke dalam aplikasi Streamlit secara lokal dapat dilakukan dengan sukses, sehingga aplikasi dapat dijalankan dan diuji pada komputer lokal.



Gambar 14. Tampilan halaman utama aplikasi clustering

Pada gambar diatas merupakan tampilan pada halaman utama aplikasi dimana terdapat tombol upload data yang berfungsi untuk mengupload data dalam bentuk CSV dan selanjutnya akan dilakukan preprocess data dan akan ditampilkan preview dari data tersebut.

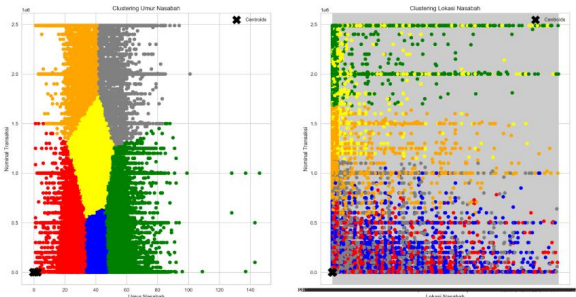
### Clustering Nasabah Bankaltimtara



Gambar 15. Tampilan halaman visualisasi aplikasi clustering

Selanjutnya pada halaman visualisasi kita diperlihatkan hasil visualisasi terhadap distribusi nasabah

berdasarkan umur dan bisa dilihat bahwa jumlah nasabah Bank paling tinggi ada pada kelompok umur 30-40 tahun.



**Gambar 16. Tampilan halaman visualisasi aplikasi clustering (2)**

Selanjutnya masih pada halaman visualisasi diperlihatkan juga hasil visualisasi clustering berdasarkan Umur dan Nominal transaksi serta clustering berdasarkan Lokasi dan Nominal Transaksi nasabah.

### Clustering Nasabah Bankaltimtara

Halaman Utama Hasil Visualisasi **Kesimpulan Cluster** About

#### Kesimpulan Cluster

Kesimpulan dan Ringkasan dari Cluster yang terbentuk:

#### Tabel Cluster berdasarkan Umur

cluster_umur	SALDO	NOMINAL	UMUR
0	3,459,573.6924	170,665,3778	26.91
1	6,224,518.6892	1,059,055.4882	38.454
2	9,311,212.5543	2,192,747.8114	51.0622
3	7,780,558.3892	333,828.9657	54.033
4	5,778,903.6799	141,572.3068	39.9185
5	6,038,135.9174	2,244,620.3108	32.3744

**Gambar 16. Tampilan halaman kesimpulan aplikasi clustering**

Pada halaman kesimpulan cluster kita diberikan informasi berapa banyak cluster yang telah terbentuk beserta nilai dari cluster yang terbentuk misalnya cluster berdasarkan umur pada gambar tersebut telah terbentuk sebanyak 6 cluster.

#### Kesimpulan Cluster berdasarkan Umur

Cluster 0 pada Umur: Rata-rata saldo nasabah adalah 3459573.69. Rata-rata nominal transaksi adalah 170665.38. Rata-rata umur nasabah adalah 26.91. Dari data ini, kita dapat menyimpulkan bahwa cluster ini terdiri dari nasabah yang memiliki karakteristik sebagai berikut: saldonya 3459573.69, dan melakukan transaksi dengan nominal sekitar 170665.38 dengan usia rata-rata 26.91.

Cluster 1 pada Umur: Rata-rata saldo nasabah adalah 6224518.69. Rata-rata nominal transaksi adalah 1059055.49. Rata-rata umur nasabah adalah 38.45. Dari data ini, kita dapat menyimpulkan bahwa cluster ini terdiri dari nasabah yang memiliki karakteristik sebagai berikut: saldonya 6224518.69, dan melakukan transaksi dengan nominal sekitar 1059055.49 dengan usia rata-rata 38.45.

Cluster 2 pada Umur: Rata-rata saldo nasabah adalah 9311212.55. Rata-rata nominal transaksi adalah 2192747.81. Rata-rata umur nasabah adalah 51.06. Dari data ini, kita dapat menyimpulkan bahwa cluster ini terdiri dari nasabah yang memiliki karakteristik sebagai berikut: saldonya 9311212.55, dan melakukan transaksi dengan nominal sekitar 2192747.81 dengan usia rata-rata 51.06.

Cluster 3 pada Umur: Rata-rata saldo nasabah adalah 7780558.39. Rata-rata nominal transaksi adalah 333828.97. Rata-rata umur nasabah adalah 54.03. Dari data ini, kita dapat menyimpulkan bahwa cluster ini terdiri dari nasabah yang memiliki karakteristik sebagai berikut: saldonya 7780558.39, dan melakukan transaksi dengan nominal sekitar 333828.97 dengan usia rata-rata 54.03.

**Gambar 16. Tampilan halaman kesimpulan aplikasi clustering (2)**

Pada halaman kesimpulan cluster ini juga kita diberikan informasi terkait kesimpulan dari cluster yang telah terbentuk sehingga kita dapat lebih memahami tentang cluster seperti apa yang telah terbentuk.

## 6 KESIMPULAN

Berdasarkan hasil pembahasan dari Penerapan Algoritma K-Means Clustering terhadap nasabah Bankaltimtara dapat ditarik Kesimpulan bahwa :

1. Pembuatan model algoritma K-Means Clustering menggunakan metode CRISP-DM memiliki 6 tahap dan ditambah tahapan pengujian untuk implementasinya pada webservice streamlit, tahapannya terdiri dari : Business Understanding, Data Understanding, Data Preparation, Modelling , Evaluation, Deployment dan Pengujian.
2. Proses Clustering menggunakan data sampling sebanyak 10.000 baris data dalam bentuk file CSV, penentuan jumlah kuster menggunakan elbow method dan dilakukan evaluasi terhadap pembentukan jumlah kuster menggunakan metode Slihouette score dimana mendapatkan nilai 0.487272 untuk cluster berdasarkan umur dan 0.653887 untuk cluster berdasarkan lokasi dan hal tersebut menunjukkan bahwa kuster yang terbentuk adalah kuster yang baik.
3. Untuk penerapan clustering Nasabah Bankaltimtara yang telah diimplementasikan pada streamlit data yang dimasukkan berupa file CSV.
4. Nasabah Muda (Cluster 0 Umur): Nasabah muda cenderung memiliki saldo dan nominal transaksi yang signifikan, menunjukkan aktivitas ekonomi yang tinggi di awal karier mereka.
5. Nasabah di Puncak Karier (Cluster 1 dan 5 Umur): Nasabah berusia sekitar 30-40 tahun memiliki saldo dan nominal transaksi yang sangat tinggi, menunjukkan puncak produktivitas ekonomi mereka.
6. Nasabah Lebih Tua (Cluster 2 dan 3 Umur): Nasabah yang lebih tua cenderung memiliki saldo yang sangat besar, dengan nominal transaksi yang beragam, menunjukkan akumulasi kekayaan tetapi dengan variasi dalam aktivitas ekonomi.
7. Lokasi Ekonomi Tinggi (Cluster 1 dan 3 Lokasi): Nasabah dari lokasi dengan aktivitas ekonomi tinggi memiliki saldo dan nominal transaksi yang sangat besar, menunjukkan kontribusi signifikan terhadap perekonomian lokal.
8. Lokasi Ekonomi Sedang (Cluster 0, 2, dan 4 Lokasi): Nasabah dari lokasi dengan aktivitas ekonomi sedang memiliki saldo dan nominal transaksi yang moderat, menunjukkan aktivitas ekonomi yang stabil tetapi tidak setinggi lokasi ekonomi tinggi.

## 7 SARAN

Terdapat beberapa saran untuk pengembangan penelitian selanjutnya dengan mempertimbangkan saran -saran dibawah :

1. Diharapkan bisa menggunakan data yang lebih banyak sehingga bisa mendapatkan wawasan yang lebih baik terhadap karakteristik nasabah.
2. Diharapkan untuk pengembangan aplikasinya bisa untuk ditambahkan fitur lainnya seperti fitur login sehingga hanya orang tertentu saja yang bisa menggunakannya.
3. Diharapkan kedepannya bisa menggunakan data selain tipe CSV dan memilih model algoritma yang akan digunakan.
4. Diharapkan juga kedepannya untuk bisa menentukan parameter apa yang akan digunakan dalam melakukan clustering misalnya berdasarkan pekerjaan nasabah, gaji nasabah, produk bank yang nasabah gunakan dan lain-lain.
5. Dan untuk selanjutnya juga tidak hanya menggunakan data nasabah Bankaltimara yang konvensional namun juga bisa mencakup nasabah Bankaltimara yang Syariah.

## 8 DAFTAR PUSTAKA

- Adams, K., Vilkomir, A., & Hills, M. (2023). A Comparison Of Machine Learning Code Quality In Python Script And Jupyter Notebooks. *The Journal Of Computing Science In Colleges*, 96.
- Alma. (2016).
- Awalina, E., & Rahayu, W. (2023). Optimalisasi Strategi Pemasaran Dengan Segmentasi Pelanggan Menggunakan . *Jurnal Teknologi Dan Informasi (Jati)* .
- Erno Sumantri, Y. M. (2024). Penerimaan Teknologi Kesehatan Masyarakat Alodokter Menggunakan Metode Technology Acceptance Model(Tam). 227-236.
- Gitosudarmo. (2014).
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts And Techniques*. Massachusetts : Morgan Kaufman.
- Harahap, M., Lubis, Y., & Situmorang, Z. (2022). Analisis Pemasaran Bisnis Dengan Data Science : Segmentasi Kepribadian. *Jurnal Data Science*. Doi:https://doi.org/10.47709/dsi.v1i2.1348
- Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi Crisp-Dm Model Menggunakan Metode Decision Tree Dengan Algoritma Cart Untuk Prediksi Curah Hujan Berpotensi Banjir, 103-108.
- Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhajja, B., & Heming, J. (2023). K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, And Advances In The Era Of Big Data. *Information Sciences*, 178-210 .
- Kotler, A. (2019).
- Muchtar, F. R., Muhyidin, Y., & Muttaqin, R. M. (2022). Penentuan Strategi Marketing Menggunakan Algoritma . *Jtsi*, 266-276.
- Mustafidah, H. (2010). Model Regresi Data Mining Motivasi Belajar Pengaruhnya Terhadap Tingkat Kedisiplinan Mahasiswa. *Juita*, 1(1), 1-4.
- Oyewole, G. J., & Thopil, G. A. (2022). Data Clustering: Application And Trends. *Artificial Intelligence Review*.
- Praniffa, A. C., Syahri, A., Sandes, F., Umi, F., Giansyah, Q. A., & Hamzah, M. L. (2023). Pengujian Black Box Dan White Box Sistem Informasi. *Jurnal Testing Dan Implementasi Sistem Informasi*, 1-16.
- Rismayadi, Fatonah, & Junianto. (2021). Algoritma K-Means Clustering Untuk Menentukan Strategi Pemasaran Di Cv. Integreet Konstruksi. *J. Responsif Ris. Sains Dan Inform.*
- Santoso, D. &. (2016). Penerapan Klasterisasi Menggunakan Algoritma K-Means Dalam Segmentasi Pasar. 112-120.
- Sari, K., Utami, N., & Wijaya, I. (2024). Penerapan Data Mining Dalam Penentuan Strategi Marketing Menggunakan Algoritma K-Means (Studi Kasus: Pt Bpr Hoki) . *Jurnal Mahasiswa Teknik Informatika*.
- Siyamto, Y. (2017). Pemanfaatan Data Mining Dengan Metode Clustering Untuk Evaluasi Biyadokumen Ekspordi Pt Winstar Batam. 2.
- Wahyono, T. (2018). *Fundamental Of Python For Machine Learning*. Yogyakarta: Gava Media.
- Wahyuni, W. A., & Saepudin, S. (2021). Penerapan Data Mining Clustering Untuk Mengelompokkan Berbagai Jenis Merk Mesin Cuci. *Sismatik (Seminar Nasional Sistem Informasi Dan Manajemen Informatika)*, 306-313.
- Xie, W. B., Lee, Y. L., Wang, C., Chen, D. B., & Zhou, T. (2020). Hierarchical Clustering Supported By Reciprocal. *Information Science*, <https://doi.org/10.1016/j.ins.2020.04.016>.
- Yuliawati, E., & Widiyanto, E. (2018). Segmentasi Nasabah Bank Berdasarkan Perilaku Menggunakan Teknik Klasterisasi. 135-145.
- Zen, H. R., & Nuryasin, I. (2024). Penerapan Whitebox Testing Pada Pengujian Sistem Menggunakan Teknik Basis Path. *Joisie (Journal Of Information Systems And Informatics Engineering)*, 101-111.