Cluster Analysis of Facebook Ads User For Digital Marketing Using K-Means Algorithm

M. Aldan Adiar F
Information Systems, Faculty of Computer Science
East Java "Veteran" Pembangunan Negeri University
Surabaya, East Java, Indonesia
19082010063@student.upnjatim.ac.id

Rafli Fahreza
Information Systems, Faculty of Computer Science
East Java "Veteran" Pembangunan Negeri University
Surabaya, East Java, Indonesia
)19082010079@student.upnjatim.ac.id

Rhifky Arfiansyah
Information Systems, Faculty of Computer Science
East Java "Veteran" Pembangunan Negeri University
Surabaya, East Java, Indonesia
19082010107@student.upnjatim.ac.id

Dhian Satria Yudha Kartika
Information Systems, Faculty of Computer Science
East Java "Veteran" Pembangunan Negeri University
Surabaya, East Java, Indonesia
dhian.satria@upnjatim.ac.id

Facebook provides a digital advertising feature called Facebook Ads. Facebook Ads was developed in 2013 and started operating in 2014, but the advertising system at that time was only limited to advertisers. Facebook at that time had not opened up to mobile application developers or website publishers. Until finally Facebook Ads can be used or accessed by anyone. Facebook Ads are very popular with business people, complete features and clear information make it easier for business people to market their products. From the Facebook Ads process, Facebook user data can be retrieved starting from the number of ads that appear, the ads clicked, age range, and gender, to the amount of money spent on these advertising products/services. In this study, Facebook Ads data clustering was carried out to be analyzed. The final visualization results describe the level of clustering according to the attributes used in the study.

Keywords: Clustering, K-Means, Data, advertising, Facebook Ads

I. INTRODUCTION

Facebook is one of the largest social media platforms. Over time, Facebook began to spread to various campuses and has become what we know today. Facebook provides a digital advertising feature called Facebook Ads. Facebook Ads was developed in 2013 and started operating in 2014, but advertisers can only place ads in their circle of people. At that time, Facebook Ads had not yet developed an advertising system with mobile application developers and website publishers. Until finally Facebook Ads can be used or accessed by anyone.

Facebook Ads are very popular with business people, complete features and clear information make it easier for business people to market their products. [1] For example, if someone advertises a shoe product, Facebook Ads will filter only users who like or have an interest in shoes. The Facebook Ads party displays ads that are installed according to Facebook users. After that, Facebook users who have an interest in the shoe product ad displayed on their page will click on the ad. Later the Facebook user will be forwarded to the e-Commerce page or website according to the ads that appear. When a Facebook user clicks on an ad area, each click is calculated as advertising costs that must be paid by the advertiser.

From the Facebook Ads process, Facebook user data can be retrieved starting from the number of ads that appear, the ads clicked, age range, and gender, to the amount of money spent on these advertising products/services. Looking at the data, it can be grouped based on the similarity of their characteristics. One of the cluster methods is the K-Means Algorithm. [2] This algorithm works by dividing the data into several clusters to analyze the similarities and dissimilarity factors attached to the data set. Then explored the pattern of connectivity between the data.

This study discusses the use of the K-Means algorithm to group Facebook Ads based on similar characteristics of the size of three indicators, such Impression (number of ads served), Clicks (number of clicks for the ad), Spent (amount paid by advertisers to Facebook Ads to display these ads). Facebook Ads data.

II. RESEARCH METHODS

This study uses several stages, starting from searching for datasets, applying the concept of preprocessing data before the data is clustered, then clustering several k-clusters, comparing data from clusters of data patterns using the K-Means method and the Elbow method, Performing cluster analysis resulting from this research [3].

A. *Searching the Dataset*

At this stage, the data to be taken is customer data from Facebook users through Facebook Ads. The data used in this study is crammed with 1135 lines. The columns in this table contain ad_id, xyzcampaignid, fbcampaignid, age, gender, interest, impressions, Clicks, Spent, Total conversion, Approved conversion.

B. *K-Means Clustering Algorithm and Elbow Method*

The k-means algorithm is an algorithm that partitions data into clusters so that data with similarities are in the same cluster and data with dissimilarities are in other clusters [4]. In the K-Means algorithm, each data must belong to a certain cluster at one stage of the process, at the next stage of the process it can move to another cluster [5].

In addition to using the K-Means algorithm, the Elbow method is also used in clustering in this study. This method displays the results of the range of the specified number of clusters and displays the average silhouette value of each number of clusters.

C. *Melakukan Analisis Tiap Klaster*

Alur penelitian yang dilakukan pada tahap ini adalah melakukan analisa terhadap hasil pengolahan data dan laporan yang dihasilkan, melakukan perhitungan validitas cluster dengan membandingkan data hasil cluster lainnya serta mengetahui nilai validitas cluster [3].

III. RESULTS AND DISCUSSION

A. *Preparation*

```
In [5]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.preprocessing import StandardScaler
        from sklearn.cluster import KMeans
        from sklearn.metrics import silhouette_samples, silhouette_score
        import plotly.express as px
        import warnings
        warnings.filterwarnings('ignore')

In [6]: #data reading
        df = pd.read_csv('D:\Tugas\Buku kuliah smt 6\Data Mining\KAG_conversion_data.csv')
        print(df.columns)
        print(df.shape)
        df.head()

        Index(['ad_id', 'xyz_campaign_id', 'fb_campaign_id', 'age', 'gender',
               'interest', 'Impressions', 'Clicks', 'Spent', 'Total_Conversion',
               'Approved_Conversion'],
              dtype='object')
        (1143, 11)
```

Out[6]:

	ad_id	xyz_campaign_id	fb_campaign_id	age	gender	interest	Impressions	Clicks	Spent	Total_Conversion	Approved_Conversion
0	708746	916	103916	30-34	M	15	7350	1	1.43	2	1
1	708749	916	103917	30-34	M	16	17861	2	1.82	2	0
2	708771	916	103920	30-34	M	20	693	0	0.00	1	0
3	708815	916	103928	30-34	M	28	4259	1	1.25	1	0
4	708818	916	103928	30-34	M	28	4133	1	1.29	1	1

Fig 1. Prepare data and input library

Import libraries for data analysis and data visualization. Then load the dataset according to the saved file, display data from the dataset, and and display the number of data and columns. It can be seen that the amount of data is 1143 and has 11 columns.

Explanation of the function of each column is as follows:

- Add_id : Unique ID for each ad
- Xyz_campaign_id : ID associated with each XYZ company ad campaign.
- Fb_campaign_id : ID associated with how Facebook tracks each campaign.
- Age: the age of the person receiving the advertisement.
- Gender: the gender of the person you want to add
- Interest : a code that defines the category that the person is interested in (interests as stated in the person's Facebook public profile).
- Impressions: the number of times the ad was shown.
- Clicks: the number of clicks for the ad.
- Spent : The amount paid by company xyz to Facebook, to display that ad.
- Total_Conversion : The total number of people who asked about the product after viewing the ad.
- Approved_conversion : Total number of people who bought the product after viewing the ad

Use the Times New Roman typeface in all manuscripts, with the font size as shown in this writing guide. The spacing is single and the contents of the text or manuscript use the left-right alignment (justified).

B. *Exploration and display of data*

```
In [7]: df.columns = df.columns.str.upper()
        df.columns

Out[7]: Index(['AD_ID', 'XYZ_CAMPAIGN_ID', 'FB_CAMPAIGN_ID', 'AGE', 'GENDER',
               'INTEREST', 'IMPRESSIONS', 'CLICKS', 'SPENT', 'TOTAL_CONVERSION',
               'APPROVED_CONVERSION'],
              dtype='object')
```

Fig 2. Change the font of column headings to capitals

```
In [8]: df.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 1143 entries, 0 to 1142
        Data columns (total 11 columns):
         #   Column               Non-Null Count  Dtype
        ---  ------               --------------  -----
         0   AD_ID                1143 non-null   int64
         1   XYZ_CAMPAIGN_ID      1143 non-null   int64
         2   FB_CAMPAIGN_ID       1143 non-null   int64
         3   AGE                  1143 non-null   object
         4   GENDER               1143 non-null   object
         5   INTEREST             1143 non-null   int64
         6   IMPRESSIONS          1143 non-null   int64
         7   CLICKS               1143 non-null   int64
         8   SPENT                1143 non-null   float64
         9   TOTAL_CONVERSION     1143 non-null   int64
         10  APPROVED_CONVERSION  1143 non-null   int64
        dtypes: float64(1), int64(8), object(2)
        memory usage: 98.4+ KB
```

Fig 3. Check the dataset if there is a null value

It can be seen that all data does not have a nutt value and the data type can be known through the script.

```
In [9]: df_c = df.copy()
        df_c.head()
```

Out[9]:

	AD_ID	XYZ_CAMPAIGN_ID	FB_CAMPAIGN_ID	AGE	GENDER	INTEREST	IMPRESSIONS	CLICKS	SPENT	TOTAL_CONVERSION	APPROVED_CONVERSION
0	708746	916	103916	30-34	M	15	7350	1	1.43	2	1
1	708749	916	103917	30-34	M	16	17861	2	1.82	2	0
2	708771	916	103920	30-34	M	20	693	0	0.00	1	0
3	708815	916	103928	30-34	M	28	4259	1	1.25	1	0
4	708818	916	103928	30-34	M	28	4133	1	1.29	1	1

```
].mean()
```

103916	7350.0	1.0	1.430000
103917	17861.0	2.0	1.820000
103920	693.0	0.0	0.000000
103928	4196.0	1.0	1.270000
103929	1915.0	0.0	0.000000
...	...	...	...
179977	1129773.0	252.0	358.189997
179978	637549.0	120.0	173.880003
179979	151531.0	28.0	40.289999
179981	790253.0	135.0	198.710001
179982	513161.0	114.0	165.609999

691 rows × 3 columns

Fig 5. Grouping columns to do cluster

Next is to group the columns to be clustered. The column fulfills the clustering because it has similarities and is of type integer. The columns that have been grouped according to FB_Campaign_ID are then searched for the average value with the mean() function.

C. *Preprocessing*

```
In [11]: features = total_conversion_df.values
```

```
In [37]: # Standardization
         scaler = StandardScaler()
         scaled_features
```

```
Out[37]: array([[-0.54421246, -0.52820627, -0.53751754],
                [-0.49697268, -0.50480229, -0.53131265],
                [-0.57413114, -0.55161025, -0.56026878],
                ...,
                [ 0.10378284,  0.1037012 ,  0.08074336],
                [ 2.97440278,  2.60792708,  2.60119867],
                [ 1.72906315,  2.1164435 ,  2.07457908]])
```

Fig 6. Data Preprocessing

At this stage, data normalization is carried out so that the data used does not have large deviations by using the StandardScaler() function.

D. *Determine the most optimal number of K*

```
In [71]: MODEL = KMeans(n_clusters=3)
         MODEL.fit(scaled_features)
```

```
Out[71]: KMeans(n_clusters=3)
```

```
In [88]: data['Cluster'] = MODEL.predict(scaled_features)
         data.head()
```

Out[88]:

	IMPRESSIONS	CLICKS	SPENT	Cluster
FB_CAMPAIGN_ID				
103916	7350.0	1.0	1.43	2
103917	17861.0	2.0	1.82	2
103920	693.0	0.0	0.00	2
103928	4196.0	1.0	1.27	2
103929	1915.0	0.0	0.00	2

```
In [38]: ks = range(2, 6)
         inertias = []

         for k in ks:
             model = KMeans(n_clusters= k)
             clusters =  model.fit(scaled_features)
             inertias.append(model.inertia_)

         # Plot ks vs inertias
         plt.figure(figsize=(12,6))
         plt.plot(ks, inertias, '-o')
         plt.xlabel('number of clusters, k')
         plt.ylabel('inertia')
         plt.xticks(ks)
         plt.show()
```

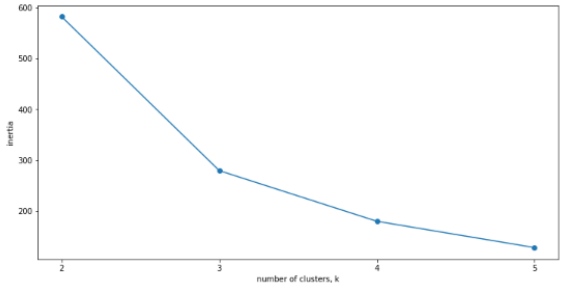

Fig 7. Determining the value of K using the Kmeans . method

By using a range of k 2 to 6, it can be seen in the graph above, that the decrease in the value of inertia at k2 is the largest of the others, after that at k3 there is a relatively constant decrease with the next decrease in k. Then it can be concluded that the optimal number of k is 3.

E. *Cluster Validation*

```
In [14]: from sklearn.metrics import silhouette_score
```

```
In [15]: x= df.iloc[:,[6,7]].values
```

```
In [16]: range_n_clusters = [2,3,4,5]

         for n_clusters in range_n_clusters:
             clusterer = KMeans(n_clusters = n_clusters, init = 'k-means++',
                                max_iter = 300, n_init = 6, random_state = 0)
             y_means= clusterer.fit_predict(x)
             silhouette_avg = silhouette_score(x,y_means)

             print ("Jumlah klaster =", n_clusters,
                    "nilai rata-rata silhouette =", silhouette_avg)
```

```
Jumlah klaster = 2 nilai rata-rata silhouette = 0.7942984011619919
Jumlah klaster = 3 nilai rata-rata silhouette = 0.7327039472049176
Jumlah klaster = 4 nilai rata-rata silhouette = 0.7226880465369411
Jumlah klaster = 5 nilai rata-rata silhouette = 0.6938479155466735
```

Fig 8. Validate the number of clusters

After using the K-mean method, the Elbow method was used to validate the number of clusters. It can be seen from the results of the silhouette average value of each number of clusters that has the largest value is k=2, but the gap between k=2 and the other k values is quite large compared to the others. So it can be concluded that the value of k is 3 because the gap in the next value is relatively small.

F. *Cluster result visualization*

Fig 9. Cluster result table

The resulting table contains cluster columns that correspond to values 1 to 3. After that, it displays the visualization.

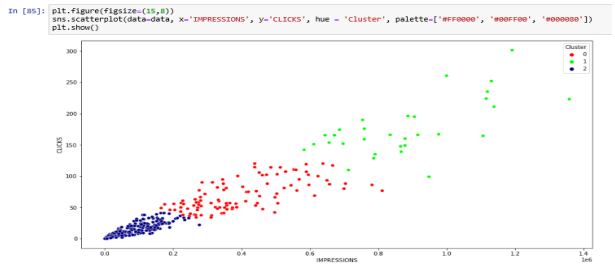

Fig 10. Cluster visualization based on impressions and clicks

The image above shows a cluster visualization based on clicks and impressions. There is a diagonal straight line pattern that is generated, indicating that the value of clicks is directly proportional to the value of impressions. It can be seen that the blue dot dominates at the smallest value, followed by the red dot and then the green dot. This means that users have a low level of interest in Facebook Ads that appear as evidenced by the number of users who are reluctant to click on Facebook Ads ads.

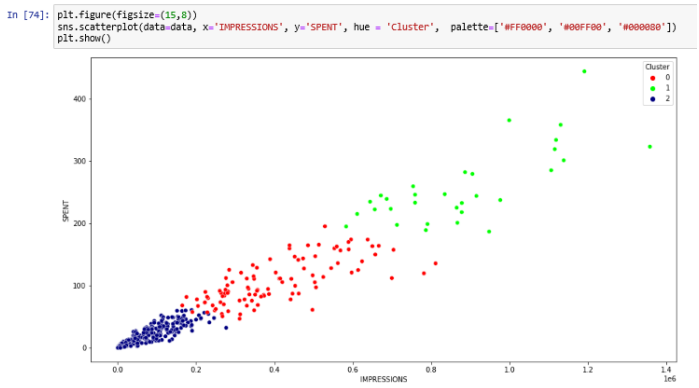

Fig 11. Cluster visualization is based on impressions and spent

The image above shows a cluster visualization based on spent and impressions. There is a straight diagonal line pattern that is generated, indicating that the value spent is directly proportional to the value of impressions. It can be seen that the blue dot dominates at the smallest value, followed by the red dot and then the green dot. This means that the xyz company still spends a lot of money with small amounts for Facebook Ads because their ads are displayed in small quantities.

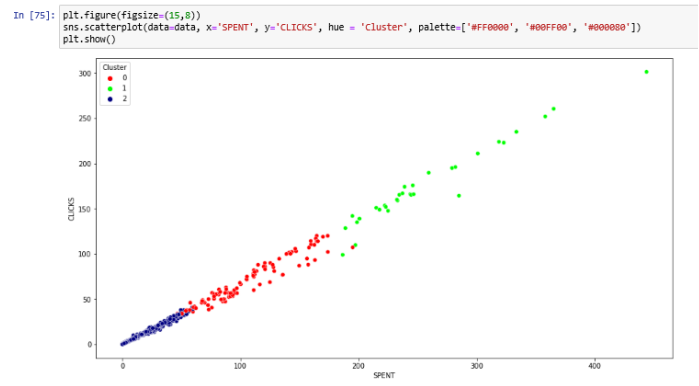

Fig 12. Cluster visualization based on clicks and spent

The image above shows a cluster visualization based on spent and clicks. There is a neat diagonal straight line pattern that is generated, indicating that the value spent is directly proportional to the value of clicks. It can be seen that the blue dot dominates at the smallest value, followed by the red dot and then the green dot. This means that the xyz company still spends a lot of money with a small amount for Facebook Ads because their ads that appear/appear are slightly clicked by users.

IV. CONCLUSION AND SUGGESTION

In this study, three clusters were found based on the attributes clicks,spent and impressions. From the results of the resulting visualization, the blue dot dominates in the lower-left corner, indicating that the ads on Facebook Ads are still underappreciated by most users. The reasons are various, from the forms of advertisements offered are less attractive and the Facebook Ads algorithm is still lacking to support the advertisements displayed so that more users can see and access them. Suggestions for future research are to cluster with different datasets and produce new conclusions later.

REFERENCES

[1] SEOMuda, "SEOMuda," SEOMuda, 18 October 2017. [Online]. Available: https://seomuda.id/sejarah-facebook-ads/. [Accessed 15 June 2022].

[2] M. W. Talakua, Z. A. Leleury and A. W. Talluta, "ANALISIS CLUSTER DENGAN MENGGUNAKAN METODE K-MEANS UNTUK PENGELOMPOKKAN KABUPATEN/KOTA DI PROVINSI MALUKU BERDASARKAN INDIKATOR INDEKS PEMBANGUNAN MANUSIA TAHUN 2014," Jurnal Ilmu Matematika dan Terapan , vol. 11, no. 2, pp. 119-128, 2017.

[3] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," JURNAL MATRIX, vol. 9, no. 3, pp. 102-109, 2019.

[4] S. D. M. J. N. Rohmawati, "Implementasi Algoritma K-Means dalam Pengklasteran Mahasiswa Pelamar Beasiswa," Jurnal Ilmiah Teknologi Informasi, vol. 1, no. 2, 2015.

[5] S. K. D. S. Ghosh, Comparative Analysis of K-Means and Fuzzy C-Means Algorithm, India, 2013.

[6] H. F. E. P. ASRONI, "Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik)," *SEMESTA TEKNIKA,* vol. 21, no. 1, pp. 60-64, 2018.

[7] G.Abdillah, F.A. Putra, F. Renaldi. "Penerapan Data Mining Pemakaian Air Pelanggan untuk Menentukan Klasifikasi Potensi Pemakaian Air Pelanggan Baru di PDAM Tirta Raharja Menggunakan Algoritma K-Means", Seminar Nasional Teknologi Informasi dan Komunikasi 2016 (SENTIKA 2016), Yogyakarta, 18-19 Maret 2016.

[8] S. T. Siska, "Analisa dan Penerapan Data Mining untuk Menentukan Kubikasi Air Terjual Berdasarkan Pengelompokan Pelanggan Menggunakan Algoritma K-Means Clustering", Jurnal Teknologi Informasi & PendidikaN, VOL. 9 NO. 1 April 2016.

[9] N. Rohmawati, S. Defiyanti, M. Jajuli, "Implementasi Algoritma K-Means dalam Pengklasteran Mahasiswa Pelamar Beasiswa",

Jurnal Ilmiah Teknologi Informasi Terapan, Volume I, No 2, 30 April 2015.

[10] S. Ghosh, S. K. Dubey. ”Comparative Analysis of K-Means and Fuzzy C-Means Algorithm”, India, 2013.