Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Program Studi Teknik Informatika Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 6, No. 3, September 2021 (545-549) **10.32493/informatika.v6i3.11629**

Klasterisasi Pengunjung Mall untuk Menentukan Target Pasar Ponsel Terbaru Menggunakan Algoritma K-Means Clustering

Irvansah Satria Pamungkas[1], Dhimas Maulana Hadi[2,] Gilang Aditya[3], Rosyidah Astari Nur[4], Aries Saifudin[5], Teti Desyani[6]

Teknik Informatika, Universitas Pamulang, Jl. Raya Puspitek No. 46 Buaran, Serpong, Tangerang Selatan, Banten, Indonesia, 15417
e-mail: [1]sahirvan19@gmail.com, [2]dhimasm.hadi@gmail.com, [3]rosyidahastarinur@gmail.com, [4]gilang.oconner123@gmail.com, [5]aries.saifudin@unpam.ac.id, [6]dosen00839@unpam.ac.id

Submitted Date: July 05th, 2021 Reviewed Date: July 22nd, 2021
Revised Date: November 10th, 2021 Accepted Date: November 28th, 2021

Abstract

Every day there are always people who visit the mall with different needs, there are some people who visit the mall to buy cell phones that have the latest features to make them look more updated, some just look around and some are not recommend to buy a new one because he still thinks his phone can still be used. To overcome this, electronic goods sales shops, especially those selling cellphones, need an analysis when they want to sell the latest products. They can group those people with expenditure information and their stages using the Kmeans Clustering algorithm. The results of these people will be divided into several clusters, namely people who will buy, people who might buy or people who will not buy, so when these cellphone distributors want to sell the latest cellphones, they will know how many people have the opportunity to buy the cellphone.

Keywords: *K-means*; Analysis: Cellphone; People; Clusters

Abstrak

Setiap harinya selalu ada orang-orang yang berkunjung ke mall dengan kebutuhan yang berbeda-beda, ada beberapa orang yang berkunjung ke mall untuk membeli ponsel-ponsel yang memiliki fitur terbaru agar terlihat lebih *update*, ada yang hanya melihat-lihat saja dan ada juga yang tidak berniat untuk membeli yang baru karena dia masih berpikir ponselnya masih bisa dipakai. Untuk mengatasi hal ini, para toko penjual barang elektronik terutama yang menjual ponsel membutuhkan suatu analisa ketika ingin menjual produk terbaru. Mereka bisa mengelompokan orang-orang tersebut berdasarkan skor pengeluaran dan penghasilan mereka menggunakan algoritma Kmeans Clustering. Hasilnya orang-orang ini akan dibagi menjadi beberapa klaster yaitu orang yang akan membeli, orang yang mungkin membeli atau orang yang tidak akan membeli, dengan begitu ketika para distributor ponsel ini ingin menjual ponsel terbaru dia akan tau ada berapa orang yang berpeluang untuk membeli ponsel tersebut.

Keywords: *K-means*; Analisis; Ponsel; Orang; Klaster

1 Pendahuluan

Pentingnya suatu Analisa untuk menentukan target pasar adalah untuk mengetahui apakah orang-orang yang berkunjung ke mall ini memungkinkan membeli produk kita atau tidak, karena setiap harinya ada banyak orang-orang berkunjung ke mall dengan kebutuhan yang berbeda-beda.

Untuk Analisa Klaster ini digunakan 3 variabel dari dataset yang digunakan yaitu umur, penghasilan pertahun, skor pengeluaran.

Para penjual ponsel belum bisa memastikan dengan tepat apakah orang-orang yang berkunjung ke mall akan membeli ponsel atau tidak, banyak orang yang dinilai akan membeli ponsel ternyata hanya melihatnya saja

atau ada orang yang dinilai tidak membeli tetapi dia membeli ponsel yang dijual.

banyak metode/strategi pemasaran yang digunakan tetapi belum bisa memastikan apakah ponsel baru yang dijual akan menutup target penjualan atau tidak. Pada penelitian ini diusulkan untuk menerapkan Algoritma *K-Means Clustering* untuk memprediksi orang-orang yang berkunjung ke mall membeli ponsel terbaru atau tidak, Algoritma ini memiliki tingkat akurasi sebesar 85,5% (Retno, 2019).

2 Metodologi

Metode yang digunakan adalah *Algoritma K-Means Clustering* (Sardar & Ansari, 2018), yaitu suatu metode penganalisaan yang mengelompokan data berdasarkan titik pusat klaster (*centroid*) terdekat dengan data. *K-Means* merupakan salah satu metode clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster. (Arofah & Marisa, 2018). Untuk implementasi dapat menggunakan bahasa Python dari library scikit-learn (scikit-learn developers (BSD License), 2021).

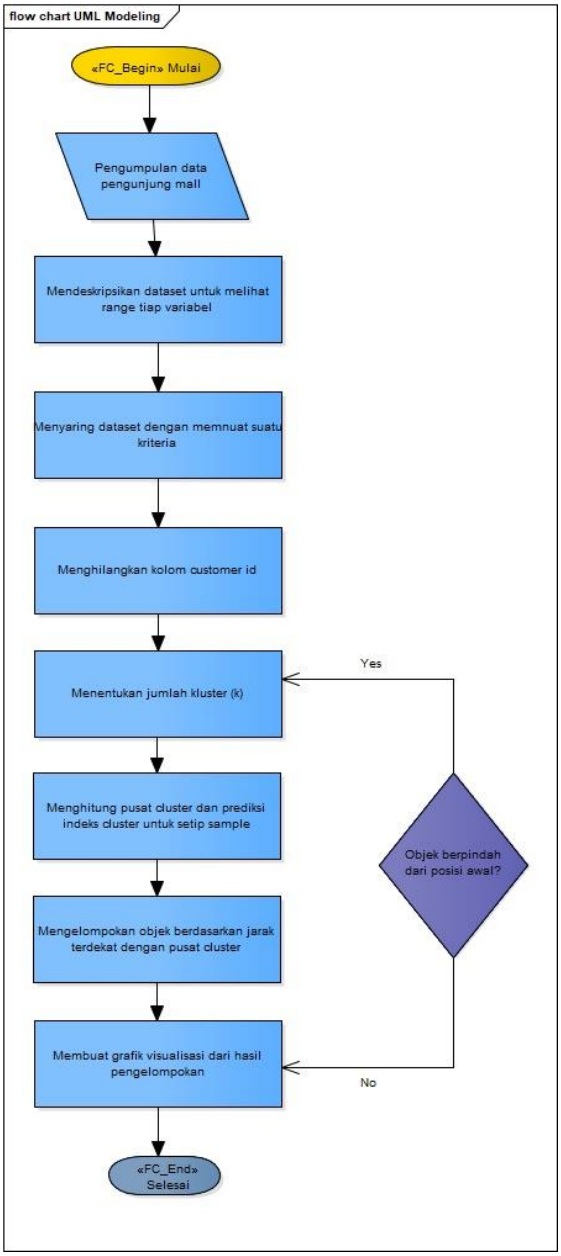
Cara kerja algoritma *K-Means* yaitu sebagai berikut:

1. Tentukan nilai *k* sebagai jumlah klaster yang ingin dibentuk
2. Inisiasi *k* sebagai centroid yang dapat dibangkitkan secara random
3. Hitung jarak setiap data ke masing-masing centroid menggunakan persamaan *Euclidean Distance*:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

4. Kelompokan setiap data berdasarkan jarak terdekat antara data dengan *centroid*-nya
5. Temukan posisi *centroid* baru (k)
6. Kembali ke langkah ke 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama

Penelitian ini akan menggunakan dataset yang diperoleh dari website Kaggle.com pada pemrosesan awal data, saya mendeskripsikan dataset tersebut untuk melihat range dari beberapa variable seperti umur, penghasilan dan skor pengeluaran, lalu menghilangkan kolom yang tidak ada kaitannya dengan Analisa yang akan dibuat, membuat kriteria, dan menerapkan algoritmanya.



Dataset yang digunakan diperoleh dari situs web Kaggle.com yang berisi data pengunjung mall seperti:

```
In [33]: data = pd.read_csv ("Mall_Customers.csv")
         data.head()
```

Out[33]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

customer ID, Gender, Age, Annual Income k$, Spending Score (1-100). ID Pelanggan, Jenis Kelamin tiap pengunjung, Umur, Penghasilan Pertahun dalam ribu dollar US, serta skor pengeluaran dalam skala 1-100.

Dataset dideskripsikan untuk melihat *range* tiap variable.

```
In [34]: data.describe()
```

Out[34]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

Pada dataset yang digunakan jumlah keseluruhan data sebanyak 200 data, pada variabel usia minimal 18 tahun sedangkan maksimal 70 tahun, pada variabel Annual income atau penghasilan pertahun minimal 15k$ sedangkan maksimalnya 137k$, dan pada variabel Spending score atau skor pengeluaran tiap pengunjung minimal 1 sedangkan maksimalnya 99. Sebelum dataset dianalisa peneliti melakukan data *preprocessing* untuk mengetahui apakah ada kolom yang kosong atau tidak dan menghilangkan variabel yang tidak ada kaitannya dengan Analisa yang akan dilakukan yaitu variabel *Customer ID*

```
del kr1['CustomerID']
```

```
In [9]: kr1.isnull().sum()
```

```
Out[9]: Gender                  0
        Age                     0
        Annual Income (k$)      0
        Spending Score (1-100)  0
        cluster                 0
        dtype: int64
```

Dari gambar di atas terlihat tidak ada kolom yang kosong pada dataset, dan pada dataset juga variabel *Customer ID* sudah dihilangkan.

3 Hasil dan Pembahasan

Peneliti menyaring kembali dataset dengan membuat suatu kriteria dimana umur kurang atau sama dengan 45 tahun dan skor pengeluaran yang lebih atau sama dengan 50.

```
In [5]: kr1 = data [
            (data['Age']<45)&
            (data['Annual Income (k$)'])&
            (data['Spending Score (1-100)'] >= 50)

        ]
        kr1.count()
```

```
Out[5]: CustomerID              42
        Gender                  42
        Age                     42
        Annual Income (k$)      42
        Spending Score (1-100)  42
        dtype: int64
```

Ketika sudah dibuat kriteria, peneliti menghitung jumlah data yang sesuai dengan kriteria yang dibuat dengan menggunakan fungsi *count.*

```
kr1.count()
```

```
Out[5]: CustomerID              42
        Gender                  42
        Age                     42
        Annual Income (k$)      42
        Spending Score (1-100)  42
        dtype: int64
```

Pada gambar di atas terlihat data yang sesuai dengan kriteria sebanyak 42 data, maka dengan begini proses analisa akan lebih mudah, kenapa harus dibuat kriteria? karena untuk menghubungkan data dengan apa yang terjadi di kenyataan saat ini. Setelah mengetahui berapa jumlah data yang akan dianalisa maka dilakukan perhitungan untuk menentukan nilai k sebagai jumlah klaster yang ingin dibentuk kemudian dilakukan perhitungan pusat cluster dan prediksi indeks cluster untuk setiap sampel menggunakan fungsi *fit_predict.*

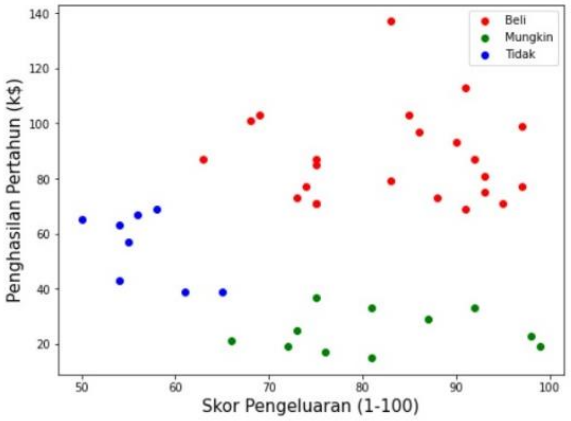
```
In [6]: km = KMeans(n_clusters = 3)
        km
Out[6]: KMeans(n_clusters=3)

In [11]: y_predicted = km.fit_predict(km1[['Age','Annual Income (k$)','Spending Score (1-100)']])
         y_predicted
Out[11]: array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Pada gambar di atas jumlah klaster sebanyak 3 klaster yang akan dibentuk dan pada perhitungan kedua menghasilkan sebuah array yang berisikan nomor 0-2 yang di mana itu menandakan tiap klasternya.

Out[18]:

	Gender	Age	Annual Income (k$)	Spending Score (1-100)	cluster
1	Male	21	15	81	1
5	Female	22	17	76	1
9	Female	30	19	72	1
11	Female	35	19	99	1
17	Male	20	21	66	1
19	Female	35	23	98	1
23	Male	31	25	73	1
29	Female	23	29	87	1
33	Male	18	33	92	1
35	Female	21	33	81	1
39	Female	20	37	75	1
43	Female	31	39	61	2
45	Female	24	39	65	2
52	Female	31	43	54	2
87	Female	22	57	55	2
111	Female	19	63	54	2
115	Female	19	65	50	2
120	Male	27	67	56	2
122	Female	40	69	58	2
123	Male	39	69	91	0
127	Male	40	71	95	0
129	Male	38	71	75	0
131	Male	39	71	75	0
135	Female	29	73	88	0
137	Male	32	73	73	0
141	Male	32	75	93	0

Untuk memudahkan, peneliti membuat visualisasi dari hasil analisa tadi menggunakan grafik plot (Matplotlib development team, 2021), peneliti mengelompokan tiap data berdasarkan nomor klasternya dan memberi warna di setiap kategorinya.



Pada titik yang berwarna merah menunjukan orang-orang yang berpenghasilan lebih dari 60k$ dan memiliki skor pengeluaran lebih dari 50, maka orang-orang ini memungkinkan sekali untuk membeli ponsel terbaru. Pada titik yang berwarna biru menunjukan orang-orang yang bisa dibilang kelas menengah tetapi memiliki skor pengeluaran yang cukup rendah, jadi bisa dibilang orang-orang ini adalah orang yang cukup hati-hati ketika berbelanja atau hemat. Berbeda dengan klaster yang berwarna hijau orang-orang dengan penghasilan yang bisa dibilang cukup rendah karena kurang dari 50k$ tetapi memiliki skor pengeluaran yang cukup tinggi yaitu lebih dari 50 jadi orang-orang ini bisa dibilang mungkin akan membeli ponsel terbaru.

4 Kesimpulan

Dari sekian banyaknya orang yang berkunjung ke mall dengan kriteria yang berbeda-beda analisa ini bisa digunakan untuk mengelompokan orang-orang tersebut sesuai dengan kesamaan-kesamaan tertentu yang nantinya akan dibuat beberapa klaster dan dicocokan apakah orang ini akan membeli produk atau tidak sesuai dengan kriteria yang dibuat, dengan menggunakan algoritma ini para distributor bisa

memperhitungkan apakah dia harus menjual suatu produk baru atau tidak ke depannya, tetapi karena analisa ini hanya memiliki akurasi 85,5% (Retno, 2019), maka analisa ini tidak bisa dijadikan satu-satunya acuan untuk menjual suatu produk baru dalam hal ini ponsel baru.

5 Saran

Jadi, menurut saran saya untuk meneliti target pasar ponsel kita menggunakan metode algoritma K-Means Clustering dikarenakan mampu mengelompokan objek besar dan kecil dengan sangat cepat sehingga mempercepat proses pengelompokan tetapi dalam metode ini juga memiliki kekurangannya yaitu sangat sensitif pada pembangkitan titik pusat awal secara random. Contoh kekurangannya kita bisa lihat pada grafik di atas, klaster yang berwarna hijau orang-orangnya memliki penghasilan yang cukup rendah tetapi memiliki pengeluaran yang cukup tinggi yaitu yaitu 50, orang-orang ini memungkinkan membeli ponsel baru. Jadi analisa ini belum bisa dijadikan acuan satu-satunya atau masih memiliki kurangan dengan kata lain masih memprediksi.

References

Arofah, S. N., & Marisa, F. (2018, Mei). Penerapan Data Mining untuk Mengetahui Minat Siswa pada Pelajaran Matematika menggunakan Metode K-Means Clustering. *JOINTECS (Journal of Information Technology and Computer Science), 3*(2), 85-90. doi:10.31328/jointecs.v3i2.787

Fathia, A. N., Rahmawati, R., & Tarno. (2016). Analisis Klaster Kecamatan di Kabupaten Semarang Berdasarkan Potensi Desa Menggunakan Metode Ward dan Single Linkage. *Jurnal Gaussian, 5*(4), 801-810. doi:10.14710/j.gauss.v5i4.17109

Matplotlib development team. (2021, August 30). *matplotlib.pyplot.bar*. Retrieved from Matplotlib: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.bar.html

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *JPIT (Jurnal Informatika: Jurnal Pengembangan IT), 4*(1), 20-24. doi:10.30591/jpit.v4i1.1253

Retno, S. (2019, Juli 8). *Peningkatan Akurasi Algoritma K-Means dengan Clustering Purity Sebagai Titik Pusat Cluster Awal (Centroid).* Medan: Universitas Sumatera Utara.

Sardar, T. H., & Ansari, Z. (2018). An Analysis of MapReduce Efficiency in Document Clustering Using Parallel K-means. *Future Computing and Informatics Journal, 3*(2), 200-209. doi:10.1016/j.fcij.2018.03.003

scikit-learn developers (BSD License). (2021, August 30). *sklearn.cluster.KMeans*. Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Waskom, M. (2021, August 30). *seaborn.scatterplot*. Retrieved from Seaborn: https://seaborn.pydata.org/generated/seaborn.scatterplot.html