Cenderawasih : Journal of Statistics and Data Science
E-ISSN: 3026-0981

CLUSTERING DATASET CUSTOMERS DENGAN MENGGUNAKAN METODE K MEANS

NISWA NILHAYA M.[1], ROSY M. LATUNUSA 2[2], CAECILIA BINTANG GIRIKALLO[3]
[1,2,3)]Program Studi Statistika Fakultas MIPA Universitas Cenderawasih Jayapura, Indonesia
e-mail: bintanggirikallo@gmail.com

ABSTRAK

Keunggulan kompetitif yang berkelanjutan sangat penting bagi sebagian besar perusahaan yang ingin mempertahankan atau memperluas posisinya dan mengoptimalkan probabilitas dan pertumbuhan keuntungannya. Keunggulan yang baik adalah yang dapat menarik klien/pelanggan. Segmentasi pelanggan sangat penting untuk mencapai misi ini. Teknik clustering dapat membantu dalam pengelompokan pelanggan dalam membuat kelompok yang terdiri objek target berdasarkan informasi dalam data yang membedakan objek dan hubungan di antara objek tersebut. Dengan mengelompokkan pelanggan ke dalam beberapa kelompok dapat membedakan antara pelanggan yang lebih disukai dan yang kurang disukai. Pada hal ini kami menyelidiki masalah audit internal yang terkait dengan protokol dengan tujuan untuk melihat nilai yang diberikan oleh mall kepada customer berdasarkan perilaku customer. Algoritma clustering yang digunakan adalah algoritma K-means. Pada metode algoritma K-Means diperoleh jumlah cluster dengan metode elbow yaitu $k = 2$. Cluster yang terbentuk yaitu cluster 1 dengan jumlah 11 pelanggan yang dan cluster 2 dengan jumlah 14 pelanggan.

Kata Kunci: Clustering, K-Means,Customers

1. PENDAHULUAN

Di era sekarang ini, persaingan pasar sangat ketat. Oleh karena itu, keunggulan kompetitif yang berkelanjutan sangat penting bagi sebagian besar perusahaan yang ingin mempertahankan atau memperluas posisinya dan mengoptimalkan probabilitas dan pertumbuhan keuntungannya. Keunggulan datang dalam berbagai bentuk, namun keunggulan yang baik adalah yang dapat menarik klien/pelanggan. Oleh karena itu, perusahaan perlu memiliki pengetahuan tentang pelanggannya agar dapat menyesuaikan aktivitas bisnisnya untuk mengoptimalkan nilai yang mereka berikan dan kepuasan yang mereka berikan kepada pelanggannya. Secara umum, pelanggan merupakan aset yang sangat berharga bagi perusahaan. Oleh karena itu, pelaku bisnis perlu melakukan upaya ekstra untuk mempertahankan pelanggannya.

Segmentasi pelanggan sangat penting untuk mencapai misi ini. Clustering adalah teknik analisis data yang membantu dalam mengelompokkan pelanggan. Teknik ini membuat kelompok yang terdiri dari objek target berdasarkan informasi dalam data yang membedakan objek dan hubungan di antara objek tersebut. Kondisi suatu cluster tertentu serupa satu sama lain tetapi berbeda dengan kondisi cluster lainnya. Dengan mengelompokkan pelanggan ke dalam beberapa kelompok dapat membedakan antara pelanggan yang lebih disukai dan yang kurang disukai. Setelah melakukan segmentasi pelanggan, auditor dapat memeriksa efisiensi dan efektivitas kegiatan bisnis perusahaan saat ini dalam memberikan layanan yang tepat kepada berbagai jenis pelanggan dengan harga yang tepat. Pada hal ini kami menyelidiki masalah audit internal yang terkait dengan protokol dengan tujuan untuk melihat nilai yang diberikan oleh mall kepada customer berdasarkan perilaku customer. Untuk mencapai tujuan ini maka kelompokkan pelanggan menggunakan teknik pengelompokan (sering disebut clustering). Setelah kumpulan data memasuki tahap prapemrosesan, data dapat dikelompokkan. Untuk clustering, fase ini menggunakan metode clustering K-means yang dioptimalkan.

Penelitian terdahulu oleh Abhinav Sagar yang berjudul “Customer Segmentation Using K Means Clustering” .Algoritma k-means digunakan untuk segmentasi pelanggan berdasarkan skor pengeluaran dan pendapatan tahunan. Hasilnya menunjukkan bahwa nilai K optimal adalah 5 menggunakan metode elbow, dan visualisasi 3D dari skor pengeluaran pelanggan dengan pendapatan tahunan menunjukkan pemisahan data menjadi 5 kelas yang direpresentasikan dengan warna yang berbeda.

Penelitian terdahulu dilakukan oleh Rina Yuliana Sari, Hardian Oktavianto, Henny Wahyu Sulistyo yang berjudul ”Algoritma K-Means Dengan Metode Elbow Untuk Mengelompokkan Kabupaten/Kota Di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia”. Penelitian ini bertujuan untuk membantu pemerintah untuk dapat mengetahui permasalahan serta mempertimbangkan pengambilan kebijakan pada wilayah kabupaten/kota diprovinsi Jawa Tengah berdasarkan variabel variabel IPM dengan memanfaatkan metode Clustering. Penelitian ini menggunakan metode K-Means, yang merupakan algoritma efektif untuk menganalisis data dalam jumlah besar.

2. METODE PENELITIAN

Metode pada penelitian ini dibagi menjadi dua yaitu:

2.1 Tahap Pengumpulan Data

Dalam penelitian ini peneliti menggunakan dataset *Shop Customer* yang tersedia pada website www.kaggle.com. Dataset ini berisi delapan atribut antara lain *CustomerID, Gender, Age, Annual Income, Spending Score, Profession, Work Experience dan Family Size* dengan jumlah 2000 data. Setelah data diperoleh selanjutnya dilakukan *data selection*. Dari delapan atribut pada data awal, dilakukan penyeleksian atribut, atribut yang digunakan adalah atribut yang bersifat numerik dan saling berhubungan maka diperoleh 3 (tiga) atribut yaitu *CustomerID*, *Spending Score* dan *Annual Income*. Tiga atribut tersebut dilakukan data cleansing, dari jumlah 2000 data kemudia direduksi menjadi 25 data. Pada data tidak ditemukan *missing value* sehingga tidak perlu dilakukan penginputan *missing value*.

2.2 Tahap Metode Data Mining

Metode da-ta mining yang dilakukan dalam penelitaian adalah K-Means Clustering dengan menggunakanbahasa pemrograman Python. Tahapan pengolahan data meliputi:

a. Data Selection
   Proses seleksi atribut dilakukan untuk menentukan atribur-atribut apa yang akan digunakan dalam proses clustering.

b. Penetuan Jumlah Cluster
   Jumlah cluster pada algoritma K-Means dapat ditentukan dengan menggunakan metode Elbow yang menghasilkan plot varians yang diperiksa setelah peningkatan jumlah cluster diplotkan terhadap jumlah cluster (Naeem dan Wumaier, 2018). Jumlah cluster dievaluasi menggunakan Sum Squared Error (SSE) dengan persamaan sebagai berikut (Hassan et al., 2021):

$$SSE = \Sigma\Sigma \| xj - \bar{xi} \| 2 xj \in C1 ki = 1 \quad (1)$$

   dimana x_j adalah objek disetiap cluster dan C_i adalh pusat cluster. Jika diagram garis terlihat sepertilengan, maka “siku” pada lengan tersebut adalah nilai *k* yang sesuai.

c. Pengelompokan dengan algoritma K-Means
   Untuk *k* cluster, J-Means didasarkan pada algoritma berulang yang meminimalkan jumlah jarak dari setiap objek ke pusat cluster. Objek-objek dipindahkan diantara klaster-klaster sampai jumlah tersebut tidak dapat dikurangi lagi, Algoritma K-Means melibatkan langkah-langkah berikut (Cebeci dan Yildiz,2015):

   a) Pusat-pusat dari *k* cluster dipilih secara acak *X* secara acak.

   b) Jarak antara titik data dan pusat klaster dihitung, Jarak diukur dengan norma Euclidean denganpersamaan berikut.

$$Dij2 = \| xj - vi \| \quad (2)$$

   Dimana x_{ij} mewakili jumlah titik data dalam klaster *i*

   c) Setiap titik data ditugaskan ke klaster yang memiliki centroid paling dekat dengannya.

   d) Pusat klaster diperbarui dengan menggunakan rumus berikut

$$V_i = \frac{\sum_{i=1}^{n} x_{ij}}{n_i} \quad (3)$$

   dimana v_i adalah centroid dari cluster i, *xij* adalah objek dalam cluster i, dan n_i adalah jumlah objek dalam cluster *i*.

   e) Jarak dari pusat klaster yang diperbarui dihitung ulang.

f) Jika tidak ada titik data yang ditetapkan ke klaster baru, maka proses algoritma akan dihentikan. Jika tidak, langkah-langkah dari (1) hingga (3) diulang untuk kemungkinan pergerakan titik data diantara cluster.

d. Menggambarkan klaster-klaster

Cluster yang telah terbentuk kemudian dideskripsikan sesuai dengan karakteristik dari objek-objek yang ada di dalam cluster tersebut. Karakteristik ini dapat dilihat dengan mengacu pada atribut-atribut yang digunakan sebagai dasar pengelompokan.

3. HASIL DAN PEMBAHASAN

Hasil dan pembahasan dari dataset Shop Customer dengan metode K-Means.

3.1 Tahap Pengumpulan Data

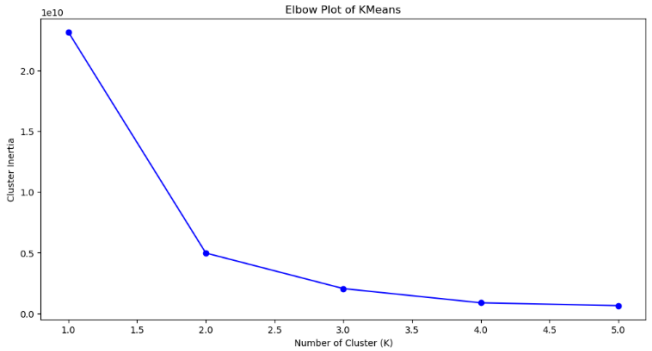
Dari delapan atribut pada data awal, dilakukan penyeleksian atribut, atribut yang digunakan adalah atribut yang bersifat numerik dan saling berhubungan maka diperoleh 3 (tiga) atribut yaitu *CustomerID*, *Spending Score* dan *Annual Income*.

3.2 Tahap Metode Data Mining

Atribut yang telah diseleksi kemudian dilanjtukan pada metode data mining yaitu clustering dengan algoritma K-Means.

1. Menentukan jumlah cluster

   Tahap berikutnya mendefinisikan algoritma K-Means Clustering pada python dan penentuan jumlah cluster. Penentuan jumlah cluster ditentukan dengan metode elbow Dimana untuk setiap jumlah cluster dievaluasi nilai SSE-nya



Dari gambar diatas dilihat bahwa nilai SSE menurun seiring bertambahnya jumlah cluster dan akan membentuk siku. Diperoleh titik siku terletak pada K= 2 sehingga jumlah cluster yang akan digunakan adalah 2 cluster.

2. Pengelompokkan Cluster

   Setelah diperoleh jumlah cluster, data kemudian dibagi menjadi 2 cluster dengan hasil sebagai berikut :

No	Spending Score (1-100)	Annual Income ($)	Clusters	Customer ID	**No**	Spending Score (1-100)	Annual Income ($)	Clusters	Customer ID
1	39	15000	1	1	**14**	77	91000	2	14
2	81	35000	1	2	**15**	13	19000	1	15
3	6	86000	2	3	**16**	79	51000	1	16
4	77	59000	2	4	**17**	35	29000	1	17
5	40	38000	1	5	**18**	66	89000	2	18
6	76	58000	2	6	**19**	29	20000	1	19
7	6	31000	1	7	**20**	98	62000	2	20

8	94	84000	2	8	**21**	35	96000	2	21
9	3	97000	2	9	**22**	73	4000	1	22
10	72	98000	2	10	**23**	5	42000	1	23
11	14	7000	1	11	**24**	73	71000	2	24
12	99	93000	2	12	**25**	14	67000	2	25
13	15	80000	2	13					

3. Deskripsi cluster

Dari hasil pengelompokkan dengan algoritma K-Means diperoleh 2 cluster dengan masing-masing karateristik yang berbeda sesuai dengan karateristik objek yang terdapat didalamnya. Anggota masing-masing cluster dapat dilihat pada table dibawah ini.

Cluster 1 diperoleh:

No	Customer ID	Spending score	Annual income
1	1	39	$ 15.000,00
2	2	81	$ 35.000,00
3	5	40	$ 38.000,00
4	7	6	$ 31.000,00
5	11	14	$ 7.000,00
6	15	13	$ 19.000,00
7	16	79	$ 51.000,00
8	17	35	$ 29.000,00
9	19	29	$ 20.000,00
10	22	73	$ 4.000,00
11	23	5	$ 42.000,00
Rata-rata		38	$ 26.454,55
Label		Rendah	Rendah

Cluster 1 terdiri dari berjumlah 11, customer yang masuk dalam cluster 1 yaitu customer dengan CustomerID (1,2,5,7,11,15,16,17,19,22,23). Cluster 1 adalah cluster dengan rata-rata *spending score* rendah dan *annual income* rendah.

Cluster 2 diperoleh :

No	Customer ID	Spending score	Annual income
1	3	6	$ 86.000,00
2	4	77	$ 59.000,00
3	6	76	$ 58.000,00
4	8	94	$ 84.000,00
5	9	3	$ 97.000,00
6	10	72	$ 98.000,00
7	12	99	$ 93.000,00
8	13	15	$ 80.000,00
9	14	77	$ 91.000,00
10	18	66	$ 89.000,00
11	20	98	$ 62.000,00

12	21	35	$	96.000,00
13	24	73	$	71.000,00
14	25	14	$	67.000,00
Rata-rata		57,5	$	80.785,71
Label		Tinggi		Tinggi

Cluster 2 terdiri dari berjumlah 12, customer yang masuk dalam cluster 2 yaitu customer dengan CustomerID (3,4,6,8,9,10,12,13,14,18,20,21,24,25). Cluster 2 adalah cluster dengan rata-rata *spending score* tinggi dan *annual income* tinggi.

4. SIMPULAN DAN SARAN

Pengolahan data dengan menggunakan *K-Means Clustering* dibagi menjadi dua cluster. Setiap klaster memiliki berbeda-beda tingkatan. Ada yang tinggi dan rendah. Cluster 1 terdiri dari berjumlah 11, customer yang masuk dalam cluster 1 yaitu customer dengan *CustomerID* (1,2,5,7,11,15,16,17,19,22,23). Cluster 1 adalah cluster dengan rata-rata *spending score* rendah dan *annual income* rendah. Cluster 2 terdiri dari berjumlah 12, customer yang masuk dalam cluster 2 yaitu customer dengan CustomerID (3,4,6,8,9,10,12,13,14,18,20,21,24,25). Cluster 2 adalah cluster dengan rata-rata *spending score* tinggi dan *annual income* tinggi.

DAFTAR PUSTAKA

Ahmad Fawaid Ridwan, S. S. (2021). IDX30 Stocks Clustering with K-Means Algorithm based on Expected Return and Value at Risk. *International Journal of Quantitative Research and Modeling*, Vol. 2, No. 44, pp. 201-208.

Bryar a Hassan, T. A. (2021). A novel cluster detection of COVID-19 patients and medical disease conditions using improved evolutionary clustering algorithm star. *National Centerfor Biotechnology Information*.

Cebeci Z., F. Y. (2015). Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures. *Journal of Agricultural Informatics*.

Dodi Alexsander Manalu, G. G. (2022). IMPLEMENTASI METODE DATA MINING K-MEANS CLUSTERING TERHADAP DATA PEMBAYARAN TRANSAKSI MENGGUNAKAN BAHASA PEMROGRAMAN PYTHON PADA CV DIGITAL DIMENSI. *Jornal of Technology Information*.

Irvansah Satria Pamungkas, D. M. (2021). Klasterisasi Pengunjung Mall untuk Menentukan Target Pasar Ponsel Terbaru Menggunakan Algoritma K-Means Clustering. *Jurnal Informatika Universitas Pamulang*, Vol.6, No.3.

Rina Yuliana Sari, H. O. (2022). ALGORITMA K-MEANS DENGAN METODE ELBOW UNTUK MENGELOMPOKKAN KABUPATEN/KOTA DI JAWA TENGAH BERDASARKAN KOMPONEN PEMBENTUK INDEKS PEMBANGUNAN MANUSIA. *Jurnal Smart Teknologi*, Vol. 3, No. 2.

Sagar, A. (2019). Customer Segmentation Using K Menas Clustering. *Towards Data Science*.

Teuku Muhammad Dista, F. F. (2022). Clustering Pengunjung Mall Menggunakan Metode K-Means dan Particle Swarm Optimization. *Jurnal Media Informatika Budidarma*.

Wahyu Wijaya Kristianto, C. R. (2022). Penerapan Data Mining Pada Penjualan Produk Menggunakan Metode K-Means Clustering (Studi Kasus Toko Sepatu Kakikaki. *Journal Pendidikan Teknologi Informasi (JUKANTI)*.