KLASIFIKASI *MALL CUSTOMER SEGMENTATION* MENGGUNAKAN METODE *K-MEANS CLUSTERING*

[1]Ivan Maurits, [2]Betty Suswati, [3]Cahyawati Diah Kusumarini

[1,2,3] Universitas Gunadarma, Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
[1]ivan_maurits@staff.gunadarma.ac.id
[2]betty_s@staff.gunadarma.ac.id
[3]cahyawati@staff.gunadarma.ac.id

Abstract

Maintaining relationships with customers is the key to success in today's competitive business. Every customer has different behaviors, they have different expectations, fears, ambitions and shopping behaviors. However, these differences are needed to build a relationship and customer segmentation is used to identify groups who have the same behaviors, likes and habits so that they will respond to the same marketing elements such as price, promotion and sales. Segmentation can inform strategic roadmaps that aim to take advantage of opportunities for profit within each customer group and can shorten the buying cycle, encourage higher spending and build greater customer loyalty.

***Keyword**: Customer Segmentation, Classification, RFM Model, K-means Clustering Algorithm, EM Clustering Algorithm, Generalized Differential RFM Method (GDRFM)*

PENDAHULUAN

Dalam penelitian ini perlu untuk melakukan segmentasi pelanggan berdasarkan profil dari masing-masing pelanggan kemudian menganalisis hasil segmentasinya untuk diketahui pelanggan yang profitable maupun sebaliknya.Untuk melakukan profiling pada pelanggan maka diperlukan sebuah model yang memberikan gambaran segala aktivitas pelanggan, kebutuhan, keinginan, dan juga konsentrasi terhadap produk dan layanan perusahaan. Model yang umum dalam mengelompokkan pelanggan adalah model *Recency*, *Frequency*, *Monetary* (*RFM*), yaitu melakukan mengelompokkan pelanggan berdasarkan interval waktu kunjungan terakhir pelanggan, frekuensi kunjungan, dan besaran nilai yang dikeluarkan sebagai *royalty* perusahaan (Aggelis et al., 2005 dan Chen et al., 2009). Metode yang sering digunakan dalam pengelompokan pelanggan adalah data mining terutama dengan teknik clustering.Metode clustering digunakan untuk melakukan segmentasi pelanggan pada Belle Crown adalah *K-Means Clustering*.Alasan metode ini digunakan adalah metode ini merupakan metode interaktif yang mudah diinterpretasikan, diterapkan, dan bersifat dinamis pada data yang tersebar (Hughes, 1994). Pada beberapa penelitian terdahulu oleh Atyanto et al. (2011) dan Angelie (2017) telah dilakukan penelitian terkait segmentasi dengan melibatkan tiga buah variabel, yaitu R, F, dan M. Dari penelitian tersebut dijelaskan bahwa terdapat kekurangan bahwa pengelompokan dengan hanya menggunakan tiga variabel dianggap belum merepresentasikan karakteristik pelanggan yang sebenarnya. Oleh karena itu, pada penelitian ini melibatkan sejumlah atribut tambahan selain R, F, dan M yakni dengan membandingkan hasil R, F, dan M dengan layanan yang ditawarkan pada Klinik Kecantikan

Belle Crown Malang. Kemudian hasilnya divisualisasikan ke dalam bentuk grafik yang mudah untuk dipahami.*Relationship Marketing* mengalami perkembangan dalam hal konsep menjadi *Customer Relationship Management* (*CRM*).*CRM* merupakan kombinasi dari proses dan teknologi yang bertujuan memahami pelanggan dalam hal perbedaan produk dan jasa yang digunakan. *CRM* memungkinkan pemberian layanan yang unggul pada pelanggan berdasarkan penggunaan informasi yang efektif dari pelanggan (Kotler and Keller, 2007). Menurut Kotler dan Keller (2007:35), terdapat empat aktivitas *CRM* yaitu mengidentifikasi (*identify*), mengakuisisi (*acquire*), mempertahankan (*retain*), mengembangkan (*develop*).

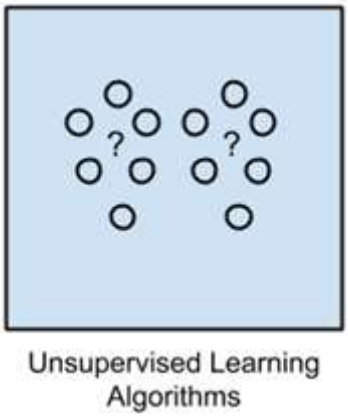
LANDASAN KEPUSTAKAAN

Machine Learning

Tom M. Mitchell pada buku *Machine Learning* (1997) menyatakan bahwa suatu program komputer dikatakan belajar dari pengalaman *E* yang berhubungan dengan beberapa tugas *T* dan ukuran performansi *P*, jika performansinya pada tugas-tugas *T*, sebagaimana diukur menggunakan *P*, meningkat dengan pengalaman *E*.

Sejak tahun 1980-an hingga saat ini, berbagai program komputer yang memiliki kemampuan belajar telah banyak diperkenalkan. Beberapa diantaranya adalah: *ALVINN* (*Autonomous Land Vehicle in Neural Networks*), sebuah kendaraan yang mampu mempelajari tingkah laku seorang sopir (manusia). Setelah beberapa menit belajar menggunakan metode *Artificial Neural Networks* (*ANN), ALVINN* mampu berjalan secara otomatis (tanpa sopir manusia) dengan kecepatan hingga 80 km/jam (Pomerlaeau, 1989); *ImageNet*, sebuah basis data citra (*image*) yang berisi jutaan citra terkelompok ke dalam ribuan kelas yang digunakan untuk pembelajaran mesin berskala besar sehingga mampu melakukan klasifikasi citra dengan akurasi tinggi. Sejak 2010, *ImageNet* telah digunakan untuk menghasilkan ratusan program komputer yang mampus mempelajari karakteristik juta citra tersebut dan mengklasifikasinya ke dalam ribuan kelas (VisionLab, 2017); *Dragon Speak*, sebuah program komputer yang mampu belajar mengenali sinyal ucapan manusia dengan akurasi sangat tinggi (Nuance, 2017). Di masa depan, komputer-komputer dengan kemampuan belajar diprediksi akan berkembang semakin pesat dengan dukungan teknologi perangkat keras komputer dan *internet of things* (*IoT*) yang semakin kuat.

Unsupervised Learning



Gambar 1. Algoritma *Unsupervised Learning*

Algoritma ini memodelkan sekumpulan *input* secara otomatis tanpa ada pandudan (yang berupa *output* yang diinginkan). Artinya, data-data yang dipelajari hanya berupa *input* tanpa label kelas. Algoritma ini biasanya digunakan untuk masalah klasterisasi (*clustering*).

Diberikan sebuah himpunan data masukkan, kelompokkan data tersebut ke dalam sejumlah klaster berdasarkan

kriteria tertentu.Jika diberikan sekumpulan data masukkan, algoritma ini mampu secara otomatis membagi data tersebut ke dalam sejumlah klaster berdasarkan, misalnya, tingkat kemiripan dalam suatu kelas.

K-Means Clustering

Pengklasteran K-Means adalah sebuah metode yang dikembangkan oleh Stuart Lloyd dari Bell Labs pada tahun 1957. Lloyd menggunakan metode ini untuk mengubah sinyal analog menjadi sinyal digital. Proses pengubahan sinyal ini juga dikenal sebagai Pulse Code Modulation.

Pada awalnya metode K-means hanya dipakai untuk internal perusahaan.Metode ini baru dipublikasikan sebagai jurnal ilmiah pada tahun 1982. Pada tahun 1965, Edward W. Forgy mempublikasikan metode yang sama dengan K-Means sehingga K-Means juga dikenal sebagai metode Lloyd-Forgy.

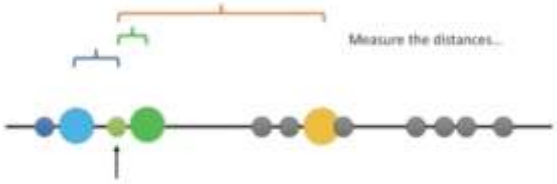
Menurut Mustapha (2020), *K-Means Clustering* bekerja seperti berikut yaitu terdapat dua belas sampel dan data ini merupakan data satu dimensi.

Gambar 2.*K-Means Cluster* - Contoh 1

Hal yang paling pertama *K-Means* lakukan adalah memilih sebuah sampel secara acak untuk dijadikan *centroid*.*Centroid* adalah sebuah sampel pada data yang menjadi pusat dari sebuah klaster.Dapat dilihat pada gambar dibawah ini terdapat tiga sampel yang dijadikan *centroid* diberi warna biru, hijau, dan kuning.

Gambar 3.*K-Means Cluster* - Contoh 2

Kedua, karena *centroid* adalah pusat dari sebuah klaster, setiap sampel akan masuk ke dalam klaster. Ini bermula dari *centroid* terdekat dengan sampel tersebut.Pada contoh dibawah, sampel yang ditunjuk anak panah memiliki jarak terdekat dengan *centroid* warna hijau.Alhasil, sampel tersebut masuk ke dalam klaster hijau.



Gambar 4.*K-Means Cluster* - Contoh 3

Berikut adalah hasil ketika tahap kedua selesai.

Gambar 5.*K-Means Cluster* - Contoh 4

Ketiga, setelah setiap sampel dimasukkan pada klaster dari *centroid* terdekat, *K-Means*akan menghitung rata-rata dari setiap sampel dan menjadikan rata-rata tersebut sebagai *centroid* baru. Rata-rata di sini adalah titik tengah dari setiap sampel pada sebuah klaster.Pada gambar dibawah, rata-rata yang menjadi *centroid* baru digambarkan sebagai garis tegak lurus.

Gambar 6.*K-Means Cluster* - Contoh 5

Keempat, langkah kedua diulang kembali. Sampel akan dimasukkan ke dalam klaster dari *centroid* baru yang paling dekat dengan sampel tersebut.

Gambar 7.*K-Means Cluster* - Contoh 6

Pada tahap ini mengulangi langkah ketiga, yaitu menemukan rata-rata dari klaster terbaru. Pada tahap ini akan menemukan rata-rata setiap klaster di tahap keempat akan sama dengan rata-rata tiap klaster pada tahap ketiga sehingga *centroid*nya tidak berubah. Ketika *centroid* baru tidak ditemukan, maka proses *clustering* berhenti.

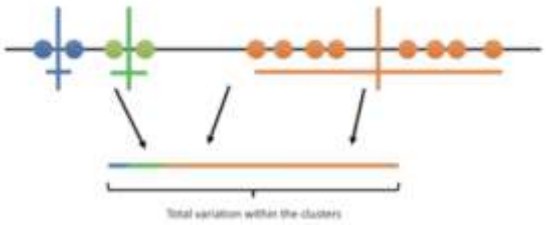
Pada proses pengklasteran dari tahapan sebelumnya belum terlihat optimal. Untuk mengukur kualitas dari pengklasteran, *K-Means*akan melakukan iterasi lagi dan mengulangi lagi tahap pertama yaitu memilih sampel secara acak untuk dijadikan *centroid*. Gambar dibawah menunjukkan *K-Means* pada iterasi kedua mengulangi kembali langkah pertama yaitu *centroid* secara acak.

Gambar 8.*K-Means Cluster* - Contoh 7

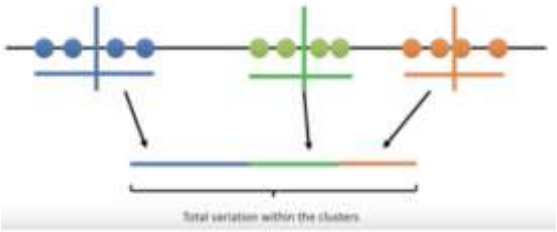
Untuk iterasi kedua, dapat mempraktekkan langkah yang sudah dijelaskan sebelumnya untuk menguji pemahaman.Hasil dari iterasi kedua adalah sebagai berikut.

Gambar 9.*K-Means Cluster* - Contoh 8

Hasil dari iterasi kedua terlihat lebih baik dibanding iterasi pertama. Untuk membandingkan klaster setiap iterasi, *K-Means*akan menghitung *variance* dari setiap iterasi. *Variance* adalah persentase jumlah sampel pada setiap klaster.Gambar dibawah menunjukkan *variance* pada iterasi pertama.



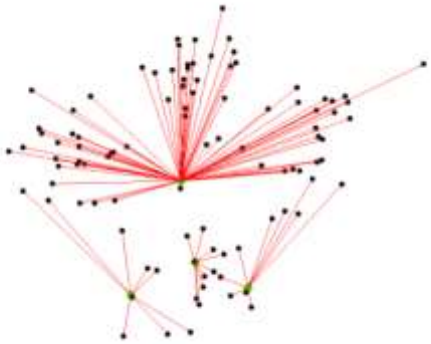
Gambar 10.*K-Means Cluster* - Contoh 9



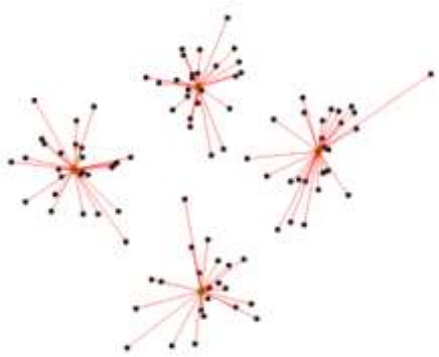
Gambar 11.*K-Means Cluster* - Contoh 10

Pada iterasi kedua atau gambar 10, *variance* nya lebih seimbang dan tidak condong pada klaster tertentu.Sehingga, hasil dari klaster iterasi kedua lebih baik dari iterasi pertama. Jumlah iterasi dari *K-Means* ditentukan oleh *programmer* dan *K-Means*akan berhenti melakukan iterasi sampai batas yang telah ditentukan.

Untuk data yang memiliki dua dimensi atau lebih, *K-Means* dengan sama yaitu menentukan *centroid* secara acak, lalu memindahkan *centroid* sampai posisi *centroid* tidak berubah. Gambar dibawah ini akan membantu untuk melihat *K-Means* bekerja pada data dua dimensi.



Gambar 12.*K-Means Cluster* - Contoh 11

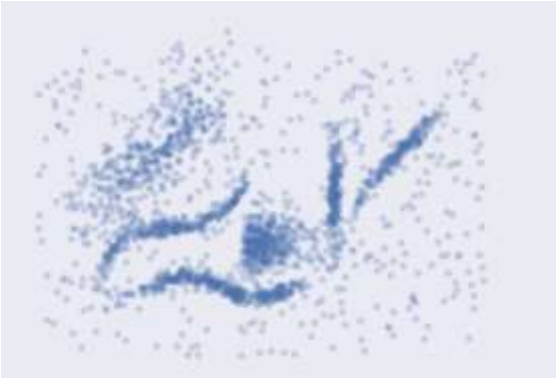


Gambar 13.*K-Means Cluster* - Contoh 12

Metode *Elbow*

Cara paling mudah untuk menentukan jumlah K atau klaster pada *K-Means* adalah dengan melihat langsung persebaran data.Otak manusia bisa mengelompokkan data-data yang berdekatan dengan sangat cepat. Tetapi cara ini hanya bekerja dengan baik pada data yang sangat sederhana.

Ketika masalah *clustering* lebih kompleks seperti gambar dibawah ini. Manusia akan bingung menentukan jumlah klaster yang pas. Untuk mengatasi itu dapat menggunakan metode *Elbow*.

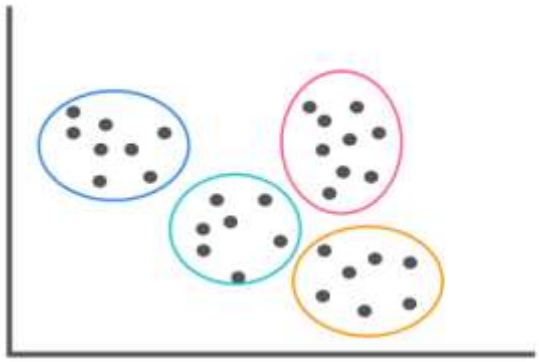


Gambar 14. Metode *Elbow* - Contoh 1

Ide mendasar dari metode *elbow* adalah untuk menjalankan *K-Means* pada dataset dengan nilai K pada jarak tertentu (1, 2, 3, …, N). Kemudian hitung inersia pada setiap nilai K. Inersia memberi tahu seberapa jauh jarak setiap sampel pada sebuah klaster.Semakin kecil inersia maka semakin baik karena jarak setiap sampel pada sebuah klaster lebih berdekatan.

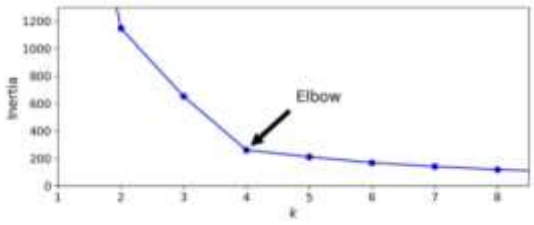
Metode *elbow* bertujuan untuk menentukan *elbow*, yaitu jumlah K yang optimal. Untuk menentukan *elbow*, perlu melakukannya secara manual, yaitu dengan melihat titik dimana penurunan inersia tidak lagi signifikan.

Pada gambar dibawah dapat dilihat data yang dibagi menjadi empat klaster. Lalu bagaimana metode *elbow* dapat menentukan jumlah klaster tersebut optimal?



Gambar 15. Metode *Elbow* - Contoh 2

Elbow berada dinilai K sama dengan 4 (empat), karena penurunan inersia pada K seterusnya tidak lagi signifikan atau perubahan nilainya kecil. Sehingga jumlah klaster yang optimal adalah 4 (empat).



Gambar 16. Metode *Elbow* - Contoh 3

METODE PENELITIAN

Pada penelitian ini, penulis menggunakan tahapan sebagai berikut:

1. Mengkonversi data menjadi *Dataframe*.
2. Melakukan *preprocessing* data.
3. Menghilangkan kolom '*CustomerID*' dan '*gender*'.
4. Melatih model *K-Means*.
5. Membuat *plot* untuk *Elbow* dan *Cluster*.

HASIL DAN PEMBAHASAN

Penulis mendapatkan *datasetMall Customers* dari *Kaggle* dengan dokumennya memiliki ekstensi *CSV* (*Comma Separated Values*). Hal yang dilakukan pertama kali adalah *importdataset* yang sudah diunduh dari *Kaggle* ke dalam *Google Collab*. Lalu konversi *dataset* tersebut menjadi sebuah *Data Frame* seperti gambar dibawah.

	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

Gambar 17. Konversi *dataset* menjadi *Dataframe*

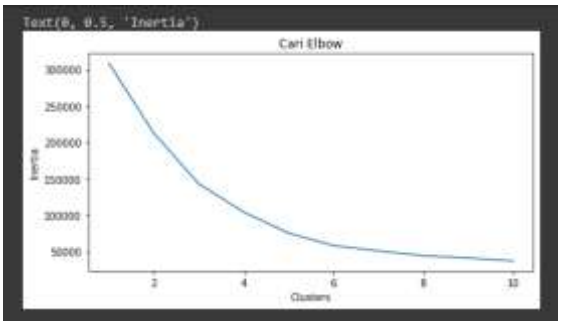
Selanjutnya, penulis melakukan *preprocessing data* yaitu mengubah nama kolom supaya lebih seragam. Lalu kolom *gender* adalah kolom kategorik, maka penulis akan mengubah data tersebut menjadi data numerik. Setelah dilakukan *preprocessing* dengan mengubah nama kolom supaya lebih seragam, maka hasilnya seperti gambar dibawah ini.

	CustomerID	gender	age	annual_income	spending_score
0	1	1	19	15	39
1	2	1	21	15	81
2	3	0	20	16	6
3	4	0	23	16	77
4	5	0	31	17	40

Gambar 18.*Preprocessing Data*

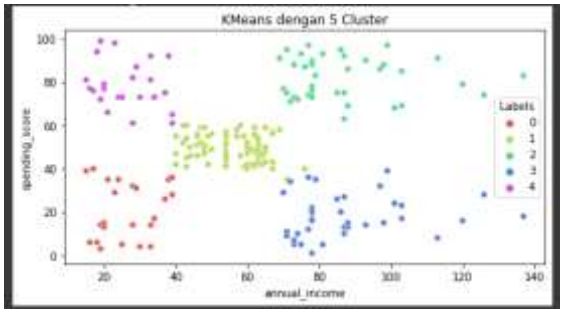
Pada tahap berikutnya, penulis melakukan *importK-Means* dari *library SKLearn*. Di tahap ini penulis menghilangkan kolom *Customer ID* dan *gender* karena kurang relevan untuk melakukan proses *clustering*. Selanjutnya, penulis akan menentukan nilai K yang optimal dengan metode *Elbow*. *K-Means* dari *library SKLearn* menyediakan fungsi untuk menghitung inersia dari *K-Means* dengan jumlah K tertentu.Disini penulis membuat *list* yang berisi inersia dari nilai K antara 1 (satu) sampai dengan 11 (sebelas).

Tahap berikutnya, penulis membuat plot dengan menggunakan *library matplotlib* dari inersia setiap K berbeda. Sesuai plot pada gambar dibawah, dapat dilihat bahwa *elbow* berada di nilai K sama dengan 5 (lima), dimana penurunan inersia tidak lagi signifikan setelah nilai K sama dengan 5 (lima).



Gambar 19.*PlotElbow*

Tahap terakhir, penulis melatih kembali *K-Means* dengan jumlah K yang didapat dari metode *Elbow*. Lalu penulis membuat *plot* menggunakan *library matplotlib* hasil dari pengklasteran *K-Means* dengan 5 (lima) klaster seperti gambar dibawah ini.



Gambar 20.*Plot Cluster*

KESIMPULAN

Segmentasi pelanggan merupakan metode untuk mengelompokkan pelanggan berdasarkan kemiripan yang diperlihatkan mereka dari segi apapun baik dari segi kebutuhan pelanggan, minat dalam fitur produk tertentu, profitabilitas pelanggan, dan sebagainya. Tujuan dari segmentasi pelanggan yang paling umum adalah mengembangkan produk dan layanan baru, menciptakan komunikasi pemasaran yang berbeda untuk setiap kelompok pelanggan, mengembangkan layanan pelanggan dan strategi retensi yang berbeda dan mengupayakan potensi keuntungan terbesar bagi perusahaan. Terdapat beberapa model dari implementasi segmentasi pelanggan seperti RFM, matriks nilai pelanggan, CLV dan metode data mining. Perlu ada pertimbangan bahwa terdapat nilai yang besar dalam sebuah kesederhanaan, terutama untuk usaha kecil dan menengah. Metode yang kompleks dapat memberikan informasi yang lebih berkualitas akan tetapi sulit untuk diterapkan dalam bisnis dan dapat menghadirkan tantangan bagi pengembang dan para strategi implementasi.

DAFTAR PUSTAKA

Brownlee, J. (2019, August 12). *A Tour of Machine Learning Algorithms*. Retrieved April 29, 2021, from Machine Learning Mastery: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Dicoding Indonesia.(n.d.).*Belajar Machine Learning untuk Pemula*. Retrieved April 29, 2021, from Dicoding Indonesia: https://www.dicoding.com/academies/184

Choudhary, V. (2018).*Mall Customer Segmentation Data*. Retrieved April 21, 2021, from Kaggle: https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Mustapha. (2020, March 11). *K-means Cluster, K-means clustering – How it works*. Retrieved April 29, 2021, from A State of Data: https://www.astateofdata.com/machine-learning/k-means-cluster-k-means-clustering-how-it-works/

Primartha, R. (2018). *Belajar Machine Learning Teori dan Praktik*. Bandung: Penerbit Informatika.

Purnama, B. (2019). *Pengantar Machine Learning Konsep dan Praktikum dengan Contoh Latihan Berbasis R dan Python*. Bandung: Penerbit Informatika.

Suyanto.(2018). *Machine Learning Tingkat Dasar dan Lanjut*. Bandung: Penerbit Informatika.