JISKa, Vol. 6, No. 2, MEI, 2021, Pp. 70 – 77
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Segmentasi Pelanggan Berdasarkan Perilaku Penggunaan Kartu Kredit Menggunakan Metode *K-Means Clustering*

Fatimah Defina Setiti Alhamdani [(1)], Ananda Ayu Dianti [(2)*], Yufis Azhar [(3)]
Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Malang, Malang
e-mail : {defina.a19,anandadianti8}@gmail.com, yufis@umm.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 6 Juni 2020, direvisi 25 Juni 2020, diterima 1 Juli 2020, dan dipublikasikan 3 Mei 2021.

Abstract

Credit card is one of the payment media owned by banks in conducting transactions. Credit card issuers provide benefits for banks with interest that must be paid. Credit card issuers also provide losses to banks that have agreed to pay not to pay their credit card bills. To request a loan from the bank, a cluster model is needed. This study, proposing a segmentation system in research using credit cards to determine marketing strategies using the K-Means Clustering method and conducting experiments using the 4 methods namely K-Means, Agglomerative Clustering, GMM, and DBSCAN. Clustering is done using 9000 active credit card user data at banks that have 18 characteristic features. The results of cluster quality accuracy obtained by using the K-Means method are 0.207014 with the number of clusters 3. Based on the results obtained by considering 4 of these methods, the best method for this case is K-Means.

Keywords: Credit Card, Cluster, K-Means, Elbow Method, Silhouette Method

Abstrak

Kartu kredit merupakan salah satu media pembayaran yang dimiliki oleh nasabah bank dalam melakukan sebuah transaksi. Penerbitan kartu kredit memberikan keuntungan bagi pihak bank dengan adanya bunga yang harus dibayar. Penerbitan kartu kredit juga memberikan kerugian pada pihak bank apabila nasabah tidak membayar tagihan kartu kreditnya. Untuk mengantisipasi kerugian pada pihak bank diperlukan sebuah model *cluster.* Pada penelitian ini diusulkan sistem segmentasi pelanggan berdasarkan perilaku penggunaan kartu kredit untuk menentukan strategi pemasaran efektif dengan menggunakan metode *K-Means Clustering* dan melakukan sebuah percobaan dengan menguji 4 metode yaitu *K-Means, Agglomerative Clustering,* GMM, dan DBSCAN. *Clustering* dilakukan menggunakan 9000 data pengguna aktif kartu kredit pada sebuah bank yang memiliki 18 fitur karakteristik. Nilai *silhouette coefficient* yang didapatkan dengan menggunakan metode *K-Means* adalah 0.207014 dengan jumlah *cluster* sama dengan 3. Berdasarkan hasil yang didapatkan dengan menguji 4 metode tersebut, metode yang paling baik untuk kasus ini adalah *K-Means*.

Kata Kunci: Kartu Kredit, *Cluster*, *K-Means*, Metode *Elbow*, Metode Silhouette

1. PENDAHULUAN

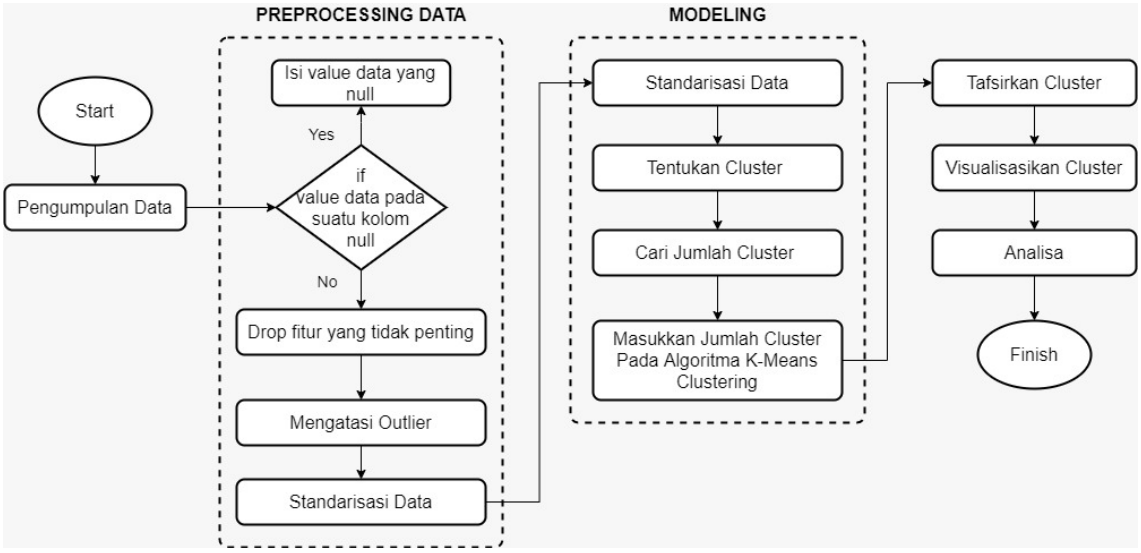
Kartu kredit merupakan salah satu media pembayaran yang dimiliki oleh nasabah bank dalam melakukan sebuah transaksi. Dengan berkembangnya bisnis kartu kredit saat ini fitur-fitur pada kartu kredit semakin beragam sehingga semakin banyak peminat kartu kredit (Sumarto et al., 2012). Namun, pemegang yang bersangkutan juga disertai dengan syarat dan ketentuan yang berlaku. Pada akhir tahun 2013, AKKI mencatat jumlah kartu yang telah beredar sebanyak 15.007.492 buah kartu. Dari informasi pertumbuhan tersebut telah menunjukkan bahwa pertumbuhan penggunaan kartu kredit di Indonesia berkembang sangat pesat (Ramadani, 2019). Penerbitan kartu kredit memberikan keuntungan bagi pihak bank dengan adanya bunga yang harus dibayar oleh nasabah. Penerbitan kartu kredit juga memberikan kerugian pada pihak bank apabila nasabah tidak membayar tagihan kartu kreditnya. Namun, untuk mengantisipasi kerugian pada pihak bank diperlukan sebuah model *cluster* untuk menganalisa pelanggan berdasarkan perilaku penggunaan kartu kredit sehingga pihak bank dapat menentukan strategi pemasaran kartu kredit.

Pada penelitian sebelumnya dalam menganalisis kartu kredit terdapat beberapa topik penelitian salah satunya mendeteksi penipuan transaksi kartu kredit menggunakan pendekatan *clustering* metode *K-Means* dengan melakukan *clustering* menjadi 4 *cluster* yaitu *low, high, risky,* dan *high risky* (Vaishali, 2014). Selain itu, juga telah dilakukan klasifikasi untuk *fraud detection* pada *e-tail merchant* menggunakan metode seperti *random forest, logistic regression,* dan *support vector machine* dan menghasilkan hasil akurasi yang baik (Carneiro et al., 2017). Pada penelitian sebelumnya dilakukan penelitian dalam *behavioral analysis* untuk prediksi penggunaan kartu kredit oleh pemilik kartu dengan beberapa faktor seperti pekerjaan dan kebiasaan pemegang kartu menggunakan kartu kredit. Penelitian ini fokus pada ketertarikan antara pendapatan *customer* dan penggunaan penuh *credit limit* yang dimiliki oleh *customer* (Dewri et al., 2016). Metode yang digunakan adalah *K-Means* untuk *clustering personal credit analysis* untuk menghasilkan analisa *trend* kebiasaan *customer* yang digunakan untuk menganalisa kelompok dari *customer* (Han & Chai, 2012).

Berdasarkan permasalahan di atas, maka dalam penelitian ini diusulkan untuk membuat sistem segmentasi pelanggan berdasarkan perilaku penggunaan kartu kredit untuk menentukan strategi pemasaran efektif dengan menggunakan metode *K-Means Clustering*. Algoritma *K-Means* merupakan suatu metode *clustering* yang dinilai paling sederhana dikarenakan dapat mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputas yang relative cepat dan efisien (Siregar, 2018). Metode yang diusulkan pada penelitian ini adalah melakukan sebuah percobaan dengan menguji 4 metode yaitu *K-Means, Agglomerative Clustering,* GMM (*Gaussian Mixture Model*), dan DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)*. Clustering* dilakukan menggunakan 9000 data pengguna aktif kartu kredit pada sebuah bank. Dengan menggunakan 4 metode *clustering* tersebut dapat menentukan hasil perbandingan yang dilakukan untuk menyakinkan bahwa metode yang diusulkan merupakan metode yang tepat untuk diterapkan.

2. METODE PENELITIAN

Pada penelitian ini menggunakan metode *data mining* dengan beberapa tahapan: pengumpulan data, *preprocessing*, *clustering*, dan analisis. Alur tahapan penelitian tertera pada Gambar 1.



Gambar 1. *Flowchart* Tahapan Penelitian.

2.1. Pengumpulan Data

Pada penelitian ini *dataset* yang digunakan bersifat *public*. Sumber data berasal dari https://www.kaggle.com/arjunbhasin2013/ccdata. Data yang diambil berisi 9000 data dari pengguna aktif kartu kredit pada sebuah bank. Data tersebut berupa angka pada setiap fiturnya kecuali fitur "CUSTID" yang memiiki kombinasi huruf dan angka. *Credit Card Dataset* ini memiliki 18 fitur karakteristik untuk setiap penggunanya, yaitu:

a) **CUSTID**: Identifikasi *customer* kartu kredit.
b) **BALANCE**: jumlah saldo yang tersisa pada akun masing-masing *customer* untuk melakukan pembelian.
c) **BALANCEFREQUENCY**: Seberapa sering saldo pada akun *customer* diperbarui, dengan skor antara 0 dan 1 (1: sering diperbarui, 0: tidak sering diperbarui).
d) **PURCHASES**: jumlah pembelian yang dilakukan dari akun milik *customer.*
e) **ONEOFFPURCHASES**: jumlah pembelian maksimum yang dilakukan dalam sekali jalan.
f) **INSTALLMENTSPURCHASES**: jumlah pembelian yang dilakukan dengan mencicil.
g) **CASHADVANCE**: Uang muka yang diberikan oleh *customer* dalam bentuk tunai.
h) **PURCHASESFREQUENCY**: seberapa sering pembelian dilakukan, dengan skor antara 0 dan 1 (1: sering, 0: tidak sering).
i) **ONEOFFPURCHASESFREQUENCY**: seberapa sering pembelian dilakukan dalamsekali jalan, dengan skor antara 0 dan 1 (1: sering, 0: tidak sering).
j) **PURCHASESINSTALLMENTSFREQUENCY**: seberapa sering pembelian dalam angsuran dilakukan, dengan skor antara 0 dan 1 (1: sering, 0: tidak sering).
k) **CASHADVANCEFREQUENCY**: seberapa sering uang muka dibayarkan.
l) **CASHADVANCETRX**: jumlah transaksi yang dilakukan dengan uang tunai.
m) **PURCHASESTRX**: banyaknya transaksi pembelian yang dilakukan.
n) **CREDITLIMIT**: batas penggunaan kartu kredit.
o) **PAYMENTS**: jumlah pembayaran yang dilakukan oleh *customer.*
p) **MINIMUM_PAYMENTS**: jumlah pembayaran minimum yang dilakukan oleh pengguna.
q) **PRCFULLPAYMENT**: pembayaran penuh yang dibayarkan oleh *customer* dalam bentuk persen.
r) **TENURE**: masa berlaku layanan kartu kredit.

2.2. *Preprocessing* Data

Penelitian ini memiliki beberapa tahapan *preprocessing* yang meliputi:

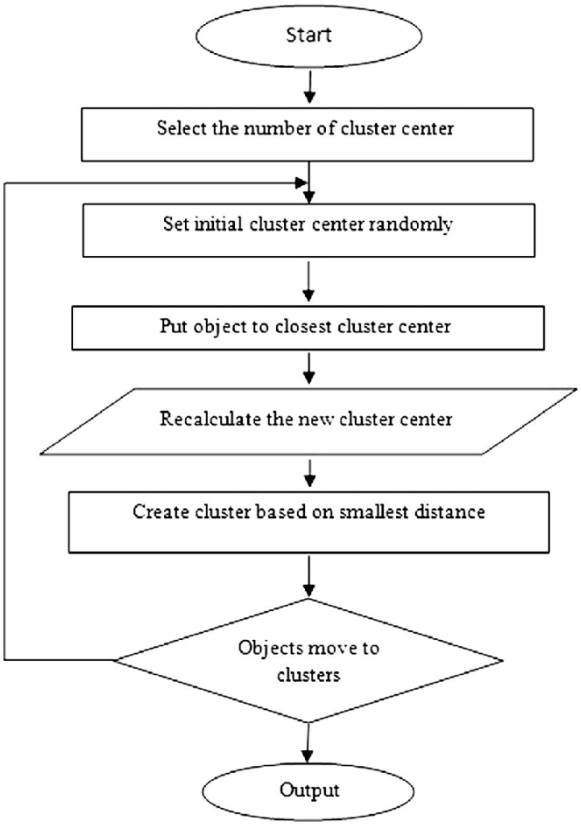
1) *Data Cleansing*
Pada tahapan ini dilakukan pengecekan data yang kosong pada *dataset* yang digunakan. Ketika ditemukan data yang kosong maka data tersebut akan diisi sesuai dengan tipe datanya menggunakan metode yang dipilih. Selain diisi dengan sebuah nilai, pembersihan data juga dapat dilakukan dengan menghapus fitur yang tidak penting atau tidak relevan pada saat pemrosesan *dataset*.

2) Menemukan Korelasi Antar Fitur
Langkah awal yang dilakukan pada tahapan ini adalah dengan menghitung korelasi kolom fitur secara berpasangan. Lalu dilanjutkan dengan memvisualisasikan hasil perhitungan korelasi antar kolom fitur agar lebih mudah untuk diamati dan dipahami. Pada proses ini dapat terlihat keterkaitan antara fitur yang digunakan dan fitur yang tidak digunakan. Fitur yang tidak digunakan tersebut adalah CUST_ID.

3) Mengatasi *Outlier*
Data *outlier* disebut juga dengan data pencilan. Pengertian dari *outlier* adalah data observasi yang memiliki nilai ekstrim secara *univariate* dan *multivariate* (Hidayat, 2016). Nilai ekstrim pada data observasi adalah nilai yang berbeda dengan sebagian nilai lain dalam suatu kelompok. Pada tahap ini dilakukan menghitung nilai absolut pada z *score* dari setiap nilai dalam sampel, *relative* terhadap *mean* sampel dan standar deviasi. Hal ini dilakukan untuk mentransformasikan data agar nilai ekstrim bisa dikurangi jaraknya dengan kelompok yang lain. Lalu filter data menjadi data *outlier free*.

4) Standarisasi Data
Standarisasi fitur dengan menghapus *mean* dan *scaling* ke varian unit. Standarisasi data adalah persyaratan umum bagi banyak penduga pembelajaran mesin. Misalnya banyak elemen yang digunakan dalam fungsi objektif dari algoritma pembelajaran (seperti kernel RBF dari *Support Vector Machines* atau L1 dan L2 regularizer model linier) mengasumsikan bahwa semua fitur berpusat di sekitar 0 dan memiliki varian dalam urutan yang sama. Jika suatu fitur memiliki varians yang urutan besarnya lebih besar dari yang lain, itu mungkin mendominasi fungsi objektif dan membuat estimator tidak dapat belajar dari fitur lain dengan benar seperti yang diharapkan.

2.3. Tahap Clustering



Gambar 2. *Flowchart* Tahapan *Clustering* (Younus et al., 2015).

Dalam penelitian ini dibentuk sebuah model dengan menggunakan algoritma *K-Means Clustering*. *K-Means* dapat diartikan dengan metode *clustering* data non-hirarki yang menggunakan metode partisi (*apportioning strategy*) berbasis *centroid* yang mengelompokkan suatu data menjadi satu atau lebih (Tendean et al., 2020). *Centroid* merupakan sebuah nilai yang digunakan untuk menghitung jarak pada suatu objek data dengan membuat penentuan nilai awalnya dilakukan secara acak dan pada nilai tiap iterasinya menggunakan rumus (Hajar et al., 2020).

$$\bar{v}_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} x_{kj} \qquad (1)$$

Menghitung jarak antara titik *Centroid* dengan titik tiap objek.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \qquad (2)$$

1) Pengelompokan objek untuk menentukan anggota *Cluster* adalah dengan memperhitungkan jarak minimal.

2) Kembali ke tahap 2, melakukan perulangan hingga nilai *Centroid* yang didapatkan tetap dan anggota *Cluster.*

Langkah awal yang perlu dilakukan adalah menentukan *cluster* menggunakan *elbow method*. *Elbow method* adalah suatu metode untuk melihat perbedaan persentase pada jumlah *cluster* (Purnima & Arvind, 2014). Metode ini diperlukan untuk menentukan jumlah *cluster* terbaik yang akan membentuk siku pada suatu titik. *Elbow criterion* adalah suatu *modelling criterion* yang bisa digunakan untuk menentukan jumlah cluster dengan melihat perubahan perbandingan antara nilai RMSSTD (*Root Mean Square Standartd Deviaton* dan RS (*R-Square*). Hal ini dilihat dengan membandingkan persentase tingkat perubahan kedua nilai (RMSSTD dan RS). RMSSTD untuk mengukur kemiripan pada cluster yang harus bernilai rendah sedangkan RS untuk mengukur perbedaan pada cluster yang harus bernilai tinggi (Sharma, 1996).

Jika terdapat suatu kondisi yang berlawanan dengan kondisi sebelumnya, maka titik sebelum terjadinya perubahan tersebut akan dianggap sebagai jumlah *cluster* yang paling tepat. Setelah menentukan *cluster*, proses dilanjutkan dengan mencari jumlah *cluster* menggunakan rata-rata metode *Silhouette Coefficient*. *Silhouette coefficient* digunakan untuk melihat seberapa baik kualitas dan kekuatan *cluster,* seberapa baik suatu objek ditempatkan dalam suatu *cluster* (Anggara et al., 2016). Lalu masukkan jumlah *cluster* yang sudah ditemukan kedalam fungsi *K-Means Clustering*. Setelah itu beri penafsiran menggunakan data cluster yang telah terbentuk dari proses *K-Means Clustering*. Untuk memudahkan proses analisis dilakukan visualisasi terhadap *cluster* yang terbentuk. Visualisasi ini dapat dilakukan dengan mereduksi dimensi data. Hal tersebut dilakukan karena pada sejumlah fitur yang digunakan terdapat kemungkinan fitur yang tidak relevan dan redundan.

Teknik yang digunakan dalam penelitian ini adalah T-SNE (*T-Distributed Stochastic Neighbor Embedding*) Tidak seperti metode seleksi fitur yang mengurangi jumlah fitur dengan cara menghilangkan fitur yang dianggap tidak penting tanpa membentuk fitur baru, "T-SNE mengurangi dimensi data dengan cara meminimalkan dua perbedaan yaitu distribusi yang mengukur kemiripan berpasangan dari objek input dan distribusi yang mengukur kemiripan berpasangan dari titik dimensi rendah yang sesuai dalam penyematan. T-SNE berupaya mengidentifikasi kelompok berdasarkan kesamaan titik data dengan banyak fitur" (Pathak, 2018).

Dalam penelitian ini dilakukan perbandingan *silhouette coefficient score* pada beberapa metode *clustering* lainnya yaitu DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), GMM (*Gaussian Mixture Models*), Aglomerative Clustering. Perbandingan tersebut dilakukan untuk menyakinkan bahwa metode yang diusulkan merupakan metode yang tepat untuk diterapkan.

2.4. Analisis Data

Tahapan analisis data dilakukan dengan mengamati data hasil visualisasi dari pemrosesan *K-Means clustering* yang direduksi dimensinya menggunakan T-SNE. Setelah proses analisis selesai dapat dijadikan sebagai acuan dalam menentukan strategi *marketing* yang dapat diambil sebagai hasil.

3. HASIL DAN PEMBAHASAN

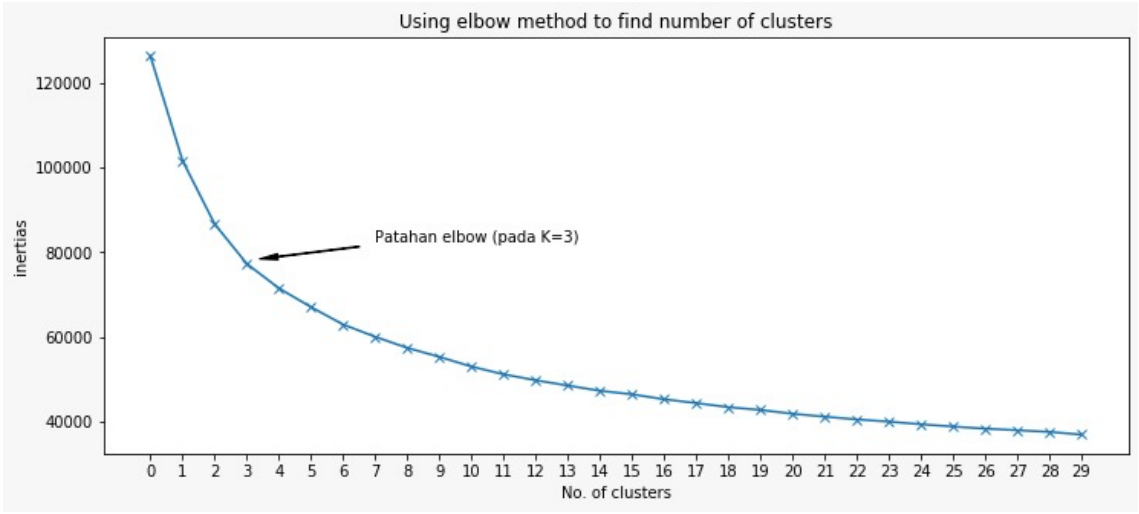
Tabel 1. Hasil Perbandingan *Silhouette Score* dan Jumlah *Cluster*.

No	Metode *Clustering*	*Silhouette Score*	Jumlah *Cluster* yang Dihasilkan
1	DBSCAN	-0.351371	9
2	GMM	0.003558	12
3	*Agglomerative clustering*	0.137499	9
4	*K-Means*	0.207014	3

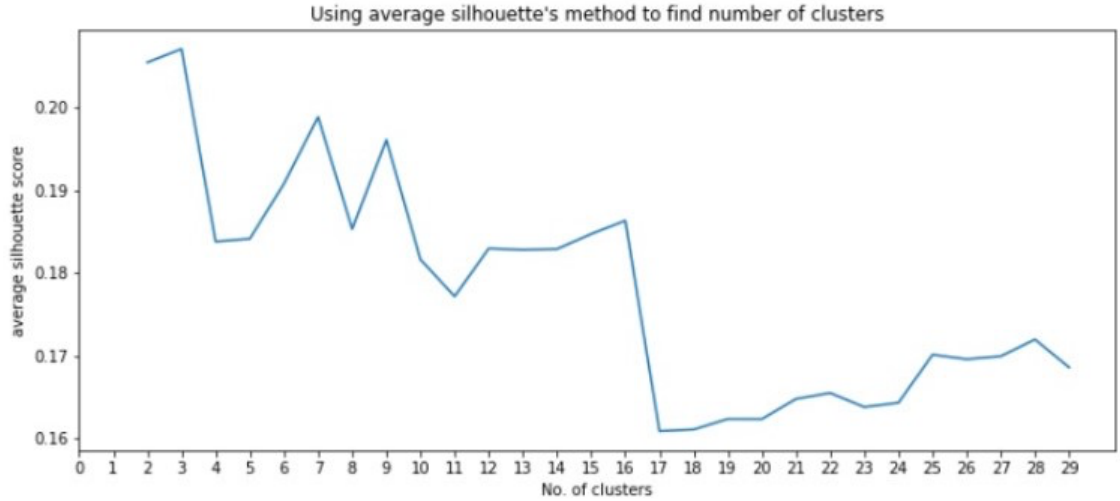
Hasil dari proses perbandingan *silhouette score* pada DBSCAN, GMM, *Aglomerative Clustering*, dan *K-Means* tertera dalam Tabel 1. Berdasarkan hasil perbandingan empat metode *clustering*

tersebut terbukti bahwa metode yang diusulkan pada penelitian ini menghasilkan *silhouette score* terbaik yaitu 0,207014.

3.1. Hasil Uji Elbow Method dan Silhouette Method pada Metode K-Means



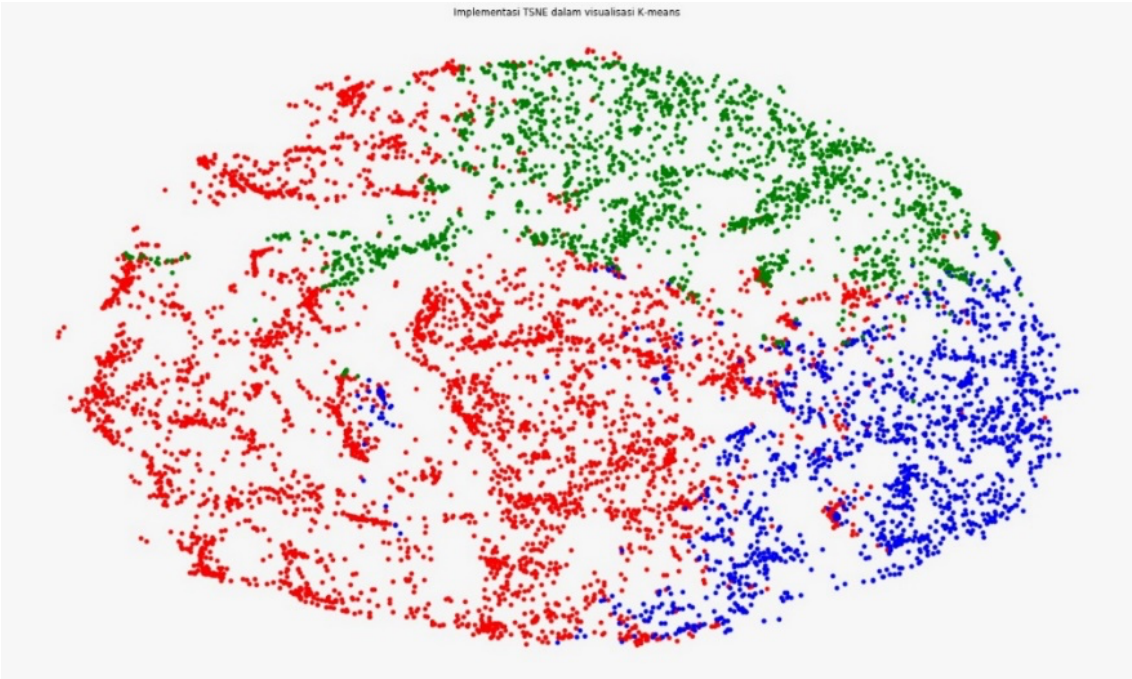
Gambar 3. Hasil Uji *Elbow Method.*



Gambar 4. Hasil Uji *Silhouette Method.*

Dari grafik hasil uji *elbow method* yang telah tersaji pada Gambar 3 terlihat patahan atau siku yang terbentuk terdapat pada nomor *cluster* 3. Lalu pada Gambar 4 grafik hasil uji *silhouette method* terlihat nilai yang paling tinggi adalah nomor *cluster* 3 dengan nilai 0.207014. Penggabungan antara hasil analisa pada Gambar 3 dan Gambar 4 menghasilkan keputusan nilai K terbaik untuk *K-Means* yang cocok untuk digunakan pada penelitian ini adalah 3.

3.2. Hasil Visualisasi T-SNE Menggunakan *K-means*



Gambar 5. Implementasi T-SNE untuk Visualisasi *K-Means.*

Pada Gambar 5. dilakukan visualisasi data hasil pengelompokan menggunakan algoritma *K-Means*. Hal itu dilakukan untuk memudahkan dalam melihat letak persebaran dari masing-masing *cluster* yang telah terbentuk. Dari visualisasi terlihat bahwa *K-Means* menghasilkan tiga *cluster* perilaku *customer credit card* dengan jumlah anggota dari masing-masing *cluster* berbeda-beda tertera pada Tabel 2.

Tabel 2. Keterangan pada Gambar 5.

Warna Titik	Keterangan	Jumlah Titik
Merah	*Customer* dengan penggunaan kartu kredit yang moderat	1696
Biru	*Customer* dengan penggunaan kartu kredit paling sedikit	1341
Hijau	*Customer* dengan lebih banyak menggunakan kartu kredit dan melakukan pembelian produk lebih sering	824

4. KESIMPULAN

Dataset pada penelitian ini memiliki jumlah data yang besar dan memiliki kesamaan pada tiap datanya merupakan hal yang tidak bisa diremehkan dalam menentukan metode *clustering* yang akan digunakan. Maka dari itu dibentuklah beberapa percobaan untuk membandingkan satu sama lain yaitu *K-Means*, *Agglomerative Clustering*, GMM dan DBSCAN dengan melihat *silhouette score* yang dihasilkan oleh masing-masing metode. Setelah melakukan perbandingan dengan 4 metode tersebut, metode terbaik untuk dataset kartu kredit ini adalah *K-Means*. Dari proses *clustering* yang dijalankan dihasilkan 3 *cluster* sehingga bisa digunakan untuk memahami segmentasi perilaku *customer* dalam menggunakan kartu kredit. Nilai *silhouette coefficient* yang didapatkan dengan menggunakan metode *K-Means* adalah 0.207014.

DAFTAR PUSTAKA

Anggara, M., Sujiani, H., & Helfi, N. (2016). Pemilihan Distance Measure Pada K-Means Clustering Untuk Pengelompokkan Member Di Alvaro Fitness. *Jurnal Sistem Dan Teknologi Informasi*, *1*(1), 1–6.

Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems*, *95*(June 2019), 91–101. https://doi.org/10.1016/j.dss.2017.01.002

Dewri, L. V., Islam, M. R., & Saha, N. K. (2016). Behavioral Analysis of Credit Card Users in a Developing Country: A Case of Bangladesh. *International Journal of Business and Management*, *11*(4), 299. https://doi.org/10.5539/ijbm.v11n4p299

Hajar, S., Novany, A. A., Windarto, A. P., Wanto, A., & Irawan, E. (2020). *Penerapan K-Means Clustering Pada Ekspor Minyak Kelapa Sawit Menurut Negara Tujuan*. 314–318.

Han, P., & Chai, J. (2012). The application of K-means in personal credit analysis. *Advanced Materials Research*, *403–408*, 2461–2464. https://doi.org/10.4028/www.scientific.net/AMR.403-408.2461

Hidayat, A. (2016). *Pengertian Data Outlier Univariat dan Multivariat*.

Pathak, M. (2018). *Introduction to t-SNE*.

Purnima, B., & Arvind, K. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, *105*(9), 17–24.

Ramadani, M. (2019). PENGARUH ATTITUDE TOWARD MONEY TERHADAP COMPULSIVE BUYING BEHAVIOUR PENGGUNA KARTU KREDIT. *Jurnal Ekonomi Vokasi*, *2*(9), 1689–1699.

Sharma, S. (1996). *Applied Multivariate Techniques Subhash Sharma* (pp. 1–5).

Siregar, M. H. (2018). Data Mining Klasterisasi Penjualan Alat-Alat Bangunan Menggunakan Metode K-Means (Studi Kasus Di Toko Adi Bangunan). *Jurnal Teknologi Dan Open Source*, *1*(2), 83–91. https://doi.org/10.36378/jtos.v1i2.24

Sumarto, S., Subroto, A., & Arianto, A. (2012). Penggunaan Kartu Kredit Dan Perilaku Belanja Kompulsif: Dampaknya Pada Risiko Gagal Bayar. *Jurnal Manajemen Pemasaran*, *6*(1). https://doi.org/10.9744/pemasaran.6.1.1-7

Tendean, T., Purba, W., & Kom, M. (2020). *Analisis Cluster Provinsi Indonesia Berdasarkan Produksi Bahan Pangan Menggunakan Algoritma K-Means*. *1*(2), 5–11.

Vaishali, V. (2014). Fraud Detection in Credit Card by Clustering Approach. *International Journal of Computer Applications*, *98*(3), 29–32. https://doi.org/10.5120/17164-7225

Younus, Z. S., Mohamad, D., Saba, T., Alkawaz, M. H., Rehman, A., Al-Rodhaan, M., & Al-Dhelaan, A. (2015). Content-based image retrieval using PSO and k-means clustering algorithm. *Arabian Journal of Geosciences*, *8*(8), 6211–6224. https://doi.org/10.1007/s12517-014-1584-7