Pengelompokan Transaksi Pelanggan Kartu Kredit Menggunakan Algoritma Clustering K-Means

1st Bambang Ary Nugroho
Fakultas Ilmu Terapan
Universitas Telkom
Bandung, Indonesia
bambangarynugroho@student.telkomuniversity.ac.id

2nd Indrarini Dyah Irawati
Fakultas Ilmu Terapan
Universitas Telkom
Bandung, Indonesia
indrarini@telkomuniversity.ac.id

3rd Aldo Lionel Saonard
Pt Hacktivate Teknologi Indonesia
Hacktive8
Jakarta, Indonesia
jobs.aldolionel@gmail.com

Abstrak*—Kartu kredit adalah salahsatu layanan yang terdapat di bank, hal ini menjadikan sebagai peluang dan tantangan bagi bank, seiring bertambahnya data maka semakin sulit juga bagaimana mendapatkan informasi yang bisa berguna, maka pengolahan data pelanggan akan sangat berguna untuk informasi strategi bisnis. Panelitian ini bertujuan untuk mengelompokan pelanggan kartu kredit berdasarkan perilaku penggunaan kartu kreditnya dengan menggunakan algoritma K-means. Sedangkan proses evaluasi dan penentuan jumlah cluster menggunakan sillhoette index dan elbow method, kemudian PCA akan di terapkan dalam proses reduksi dimensi. Berdasarkan hasil percobaan, diperoleh jumlah cluster terbaik adalah 2 cluster dengan silhouette score tertinggi. Dari 8950 pelanggan, data tersebut terbagi ke dalam cluster 0 dan cluster 1

Kata kunci*— K-Means, Principal Component Analysis, Cluster,silhouette score, elbow method

I. PENDAHULUAN

Pada era digital ini, data menjadi inti dalam kelangsungan sebuah bisnis. Manusia dihadapkan pada melimpahnya data yang bisa didapatkan tanpa menguasai kemampuan untuk ekstraksi informasi di dalamnya. Data Science adalah sebuah studi interdisipliner yang mengeksplorasi metode ilmiah dan cara mengekstraknya pengetahuan atau wawasan dari klan data di berbagai bentuk, tidak hanya terstruktur tetapi juga tidak terstruktur[6]. Dengan begitu banyaknya data dan sulit untuk dianalisis, maka dengan menerapkan data science data bisa diolah sehingga memberikan informasi yang berharga kepada perusahaan dalam mengambil keputusan bisnis. Penerapan bidang data science bisa diberbagai industri seperti perbankan. Perusahaan di industri seperti itu akan memproduksi data yang sangat besar dari pelanggan-pelanggannya yang memerlukan pengolahan data yang baik agar memberikan informasi terhadap keputusan bisnis. Dataset Credit card adalah data yang akan digunakan pada penelitian ini, data ini berisikan data-data pelanggan kartu kredit dengan berbagai perilaku penggunanya, data ini berisikan 18 features dan 8950 records.

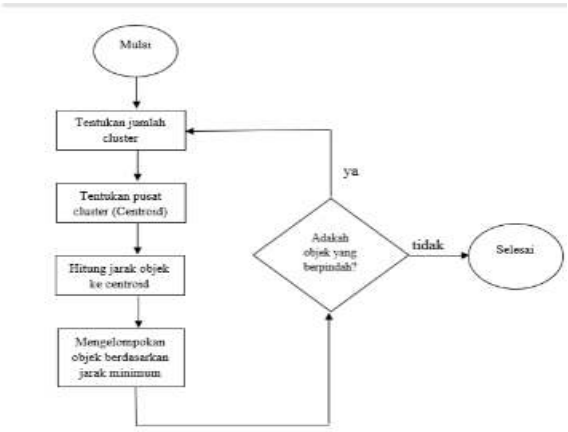
Dari hasil tahap data preprocessing yang dilakukan, data ini memiliki missing value pada features MINIMUM_PAYMENTS sebanyak 313 dan pada features CREDIT_LIMIT sebanyak 1, input median akan digunakan untuk mengatasi missing value tersebut. Algoritma K-means adalah salah satu algoritma clustering yang bersifat iteratif yang mencoba untuk mempartisi dataset menjadi subkelompok non-overlapping berbeda yang ditentukan oleh K (cluster) di mana setiap titik data hanya dimiliki oleh satu kelompok. K-Means menetapkan poin data ke kluster sedemikian rupa sehingga jumlah jarak kuadrat antara titik data dan pusat massa cluster (rata-rata aritmatika dari semua titik data yang termasuk dalam cluster itu) minimal. Semakin sedikit variasi yang kita miliki dalam cluster, semakin homogen (serupa) titik data dalam cluster yang sama

Pada Proyek akhir ini akan dilakukan clustering menggunakan algoritma K-Means, kemudian penerapan Principal Component Analysis (PCA) untuk reductions dimention, serta elbow method dan silhouetts score untuk mengevaluasi berapa banyak jumlah cluster (K) yang tepat untuk di terapkan pada modeling

II. KAJIAN TEORI

A. K-Means

Algoritma K-Means termasuk kedalam jenis unsupervised machine learning algorithm yang digunakan untuk mengelompokan data yang tidak terstruktur menurut kesamaan dan pola yang berbeda dalam kumpulan data, teknik clustering sederhana namun efektif. Algoritma K-means adalah salah satu algoritma clustering yang bersifat iteratif yang mencoba untuk mempartisi dataset menjadi subkelompok non-overlapping berbeda yang ditentukan oleh K (cluster) di mana setiap titik data hanya dimiliki oleh satu kelompok[3]. K-Means akan menetapkan poin data ke cluster sedemikian rupa sehingga jumlah jarak kuadrat antara titik data dan titik pusat massa cluster (rata-rata aritmatika dari semua titik data yang termasuk dalam cluster itu) minimal, semakin sedikit variasi yang kita miliki dalam cluster, semakin homogen (serupa) titik data dalam cluster yang sama



Gambar 1. Diagram alir K-Means

1. Tentukan nilai k sebagai jumlah cluster yang ingin dibentuk, menentukan jumlah cluster yang tepat dapat menggunakan elbow method, inertia dan Silhouette Score
2. Membuat k centroid (titik pusat cluster) awal secara random. K-Means akan membuat titik cenroid untuk masing-masing cluster
3. Hitung jarak setiap data ke masing-masing centroid menggunakan Rumus korelasi antar dua objek yaitu Euclidean Distance dan kesamaan Cosine
4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centroidnya
5. Tentukan posisi centroid baru (Ck) dengan cara menghitung nilai rata-rata dari data-data yang ada pada centroid yang sama. Jika terdapat penambahan atau pengurangan pada data, maka jumlah cluster juga akan berubah

B. Optimasi K-Means

Salah satu faktor krusial baik tidaknya algoritma K-Means adalah saat menentukan jumlah klusternya (nilai K)Karena hasil pengelompokan akan menghasilkan analisis yang berbeda untuk jumlah klaster yang berbeda juga. Untuk mengetahui jumlah klaster yang paling baik untuk studi kasus yang diuji adalah menggunakan metode elbow dan silhouetee score

1. Metode Elbow

metode elbow merupakan suatu heuristik yang digunakan dalam menentukan jumlah klaster dalam suatu kumpulan data. Metode ini terdiri dari memplot variasi yang dijelaskan sebagai fungsi dari jumlah cluster dan memilih siku kurva sebagai jumlah cluster yang akan digunakan[5].

2. Silhouette Score

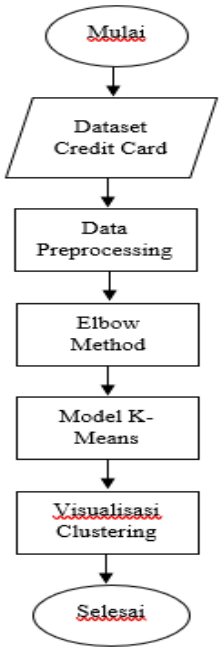
Metode Silhouetee Coefficient merupakan gabungan dari metode cohesion dan separation. Metode ini sering figunakan untuk melihat kualitas dan kekuatan cluster yaitu seberapa baik suatu objek ditempatkan dalam suatu cluster. Selain itu dapat juga digunakan untuk mengukur seberapa dekat relasi antara objek dalam sebuah cluster. Metode separation yang berfungsi untuk mengukur seberapa jauh sebuah cluster terpisah dengan cluster lain

3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah teknik linear dimensionality reduction yang dapat digunakan untuk mengekstraksi informasi dari ruang dimensi tinggi dengan memproyeksikannya ke dalam sub-ruang berdimensi lebih rendah. PCA mencoba untuk mempertahankan bagian penting yang memiliki lebih banyak variasi data dan menghapus bagian yang tidak penting dengan variasi yang lebih sedikit. Dimensionality rediction ini adalah teknik untuk mengurangi jumlah feature dengan tetap menjaga kualitas informasi yang ada pda data set. Satu hal

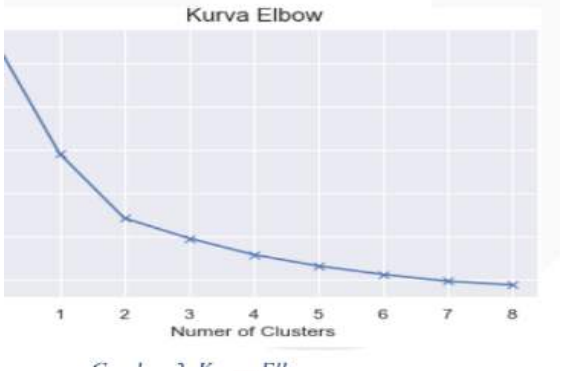
III. METODE

A. Diagram Alir Kerangka Kerja Penelitian



Gambar 2. Diagram alir kerangka kerja

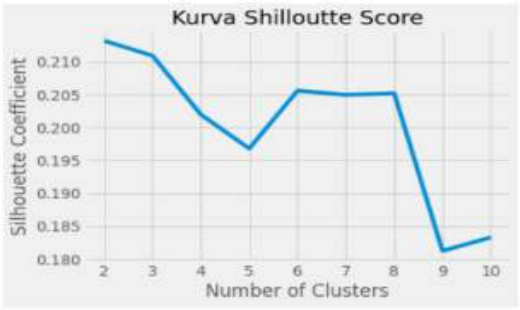
B. Menentukan jumlah cluster



Gambar 3. Kurva Elbow

Gambar 2 diatas menunjukan kurva elbow, pada titik K=2 sudut siku terbentuk dan setelah itu penurunan skornya tidak menurun tajam melainkan sudah melandai, jadi dapat disimpulkan bahwa jumlah cluster yang optimal adalah sebanyak 2 cluster.

C. Silhouette score



Gambar 4. Kurva Silhouete

Gambar 4 diatas menunjukan silhouette score untuk masing-masing jumlah cluster, semakin tinggi score yang didapatkan maka itulah jumlah cluster (K) yang terbaik, dilihat bahwa jumlah cluster terbaik adalah sebanyak 2 cluster dengan silhouetee score tertinggi

IV. HASIL DAN PEMBAHASAN

Analisis hasil dari clustering bisa dilakukan dengan memperhatikan bagaimana distribusi nilai pada setiap features yang sudah di kelompokan menjadi 2 cluster

Table 1. Hasil Clustering

Nama Cluster	Jumlah pengguna
Cluster 0	Lebih dari 5000 Pengguna
Cluster 1	Lebih dari 3000 Pengguna

Gambar 5. Hasil Clustering

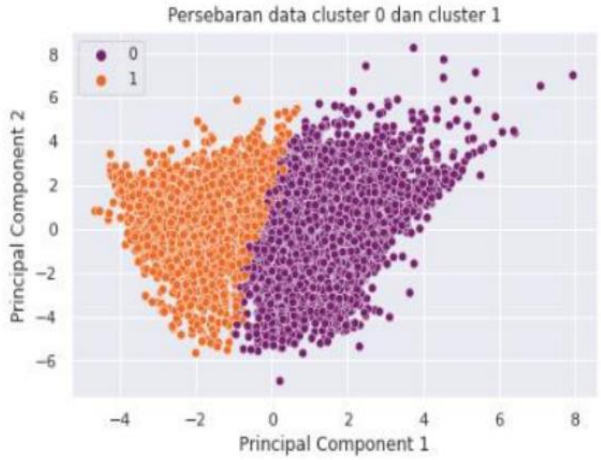
Cluster 0:

1. Jumlah pengguna sebanyak 5315, lebih besar dibandingkan dengan cluster 1
2. Nilai balance frequency lebih besar dibandingkan cluster 1, yang berarti saldo pengguna pada cluster lebih sering perbaharui
3. Pengguna sering melakukan purchases atau pembelian
4. Jumlah pengguna yang melakukan payment (pembayaran) kartu kredit tinggi
5. Pengguna yang membayar tenure berjumlah kurang lebih 5000 pengguna
6. Credit limit pengguna lebih tinggi dibandingkan dengan cluster 1

Cluster 1:

1. Jumlah pengguna sebanyak 3635
2. Nilai balance frequency lebih kecil dibandingkan dengan pengguna di cluster 0, artinya bahwa pada cluster ini saldo kartu kredit lebih jarang di perbaharui
3. Pengguna lebih jarang melakukan purchases atau pembelian bahkan lebih 2000 pengguna tidak melakukan pembelian
4. Jumlah pengguna yang melakukan payment (pembayaran) kartu kredit lebih kecil dari dibandingkan cluster 0
5. Pengguna yang membayar tenure berjumlah kurang lebih 2500 pengguna
6. Credit limit pengguna lebih rendah daripada cluster 0

Analisis hasil pengelompokan dapat dilakukan pada setiap features pada data, hal ini di sesuaikan dengan kebutuhan dan strategi pada bisnis yang akan dihadapi oleh perusahaan



Gambar 6.Persebaran cluster

IV. KESIMPULAN

Clustering pelanggan kartu kredit dapat dilakukan dengan menggunakan model clustering dengan algoritma K-means. Dalam penelitian ini, dilakukan proses clustering pelanggan berdasarkan perilaku penggunaan kartu kredit, dengan algoritma k-means dan metode PCA dalam proses visualisasinya maka diperoleh jumlah cluster paling optimal adalah 2 dilihat dari hasil plot kurva metode elbow dan silhouette score tertinggi. Karakteristik pelanggan pada cluster 0 adalah sering melakukan pembelian, credit limit lebih tinggi dari cluster 1, saldo sering di perbarui. Karakteristik pelanggan pada cluster 1 adalah bertolak belakang dengan perilaku pelanggan pada cluster 0, pelanggan pada cluster 1 tidak sering melakukan pembelian, saldo jarang di perbarui, credit limit lebih rendah dibandingkan pelanggan di cluster 0. Analisis dapat dilakukan pada setiap features sesuai dengan kebutuhan pada strategi bisnis

REFERENSI

ECTRONIC REFERENCES

[1] S. Z. A. S. Sri Rahmayani, "Analisis Algoritma K-Means untuk Klustering Penerima Bantuan Sosial," *JOMLAI: Journal of Machine Learning and Artificial Intelligence,* Vols. 1, no1, pp. 77-84, 2022.

[2] N. H. N. A. Rozzi Kesuma Dinata Sfawandi, "Analisis K-Means Clustering Pada Data Sepeda Motor," *Informatics Jurnal, Unej,* Vols. 5, No 1, 2020.

[3] W. M. P. Dhuhita, "CLUSTERING MENGGUNAKAN METODE K-MEANS UNTUK," *Jurnal Informatika,* Vols. Vol. 15, No. 2, pp. 6-7, 2015.

[4] P. D. Windha Mega, "Clustering Menggunakan Metode K-Means Untuk Menentukan Status Gizi Balita," *Jurnal Informatika,* Vols. 15, No 2, 2015.

[5] D. R. M. Nainggolan1, "DATA SCIENCE, BIG DATA, AND PREDICTIVE ANALYTICS:," *Jurnal Petahanan dan Bela Negara,* vol. II, p. 2, 2017.

[6] G. M. P. M. F. L. Dewi Sinta Saputri, "Implementasi Algoritma K-Means Clustering untuk Desa Tervaksinasi Covid-19 pada Kecamatan Ujung Padang," *Jurnal Teknik Informatika (JUTIF),* Vols. 3, no 2, pp. 261-267, 2022.

[7] A. Chaerudin, "Implementasi Algoritma K-Means++ untuk Segmentasi Pelanggan Toko dengan Menggunakan Neo4J," Bandung, 2021.

[8] S. J. P. G. M. P. N. P. H. S. P. Haviludin, "Implementasi Metode K-Means Untuk Pengelompokan Rekomendasi Tugas Akhir," *Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer,* Vols. 16, no 1, 2021.

[9] F. S. D. H. Hendro Priyatman, "Klasterisasi Menggunakan Algoritma K-Means Clustering Untuk Mempredikasi Waktu Kelulusan Mahasiswa," *Jurnal Edukasi dan Penelitian Informatika,* 2019.

[10] D. Dikti, "kursus saya: Data Science Associatte," Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, [Online]. Available: https://spadadikti.id. [Accessed 5 Agustus 2022].