ISSN (Print): 1411 – 8890, ISSN (Online): 2541 – 4518
http://journal.ums.ac.id/index.php/emitor

EMITOR: Jurnal Teknik Elektro

Analisis Perbandingan algoritme Bisecting K-Means dan Fuzzy C-Means pada Data Pengguna Kartu Kredit

Shinta Dwididanti*,[1], Dimas Aryo Anggoro[1], Muslich Hartadi Sutanto[2]
Program Studi Teknik Informatika/Fakultas Komunikasi dan Informatika – Universitas Muhammadiyah Surakarta[1]
Surakarta, Indonesia
Universiti Teknologi PETRONAS[2]
Seri Iskandar–Perak, Malaysia
*L200160124@student.ums.ac.id

Abstract– *In this digital era having a credit card is a common thing in society, with all the conveniences offered in every payment transaction, it is possible to attract public interest to using a credit card. With high public interest in credit cards, this can be used as a good indicator for credit card companies to develop a credit card business. In order to give consumer's needs for credit cards, companies are required to make decisions in determining the right marketing strategy to attract customers' interest. And one of the way is by segmenting the customers using the clustering method. Bisecting K-Means and Fuzzy C-Means are clustering algorithms that will be used in this study to segment the data about credit card user. Analysis will be performed to find out for the best performing algorithm based on validity measurement of both algorithm. From this research, it was found that Bisecting K-Means without normalization had a higher silhouette coefficient value than Fuzzy C-Means where the coefficient value of Bisecting K-Means silhouette is 0.588 and 0.579 with normalization, while the silhouette coefficient value of Fuzzy C-Means is 0.488 and 0.582 with normalization. Bisecting K-Means silhouette is 0.588 and 0.579 with normalization, while the silhouette coefficient value of Fuzzy C-Means is 0.488 and 0.582 with normalization.*

Abstrak– Di era digital seperti sekarang ini memiliki kartu kredit merupakan suatu hal yang wajar di masyarakat, dengan segala kemudahan yang ditawarkan dalam setiap transaksi pembayaran tidak menutup kemungkinan untuk menarik minat masyarakat dalam menggunakan kartu kredit. Dengan minat masyarakat yang tinggi terhadap kartu kredit, hal ini dapat dijadikan sebagai indikator yang baik bagi perusahaan kartu kredit untuk mengembangkan bisnis kartu kredit. Dalam rangka memenuhi kebutuhan konsumen akan kartu kredit, perusahaan dituntut untuk mengambil keputusan dalam menentukan strategi pemasaran yang tepat sehingga dapat menarik minat para pelanggan, salah satu caranya adalah dengan melakukan segmentasi pelanggan dengan metode *clustering*. *Bisecting K-Means* dan *Fuzzy C-Means* merupakan algoritme *clustering* yang akan digunakan pada penelitian ini untuk melakukan pengelompokan data pengguna kartu kredit. Analisis akan dilakukan untuk mengetahui algoritme dengan performa terbaik berdasarkan pengujian validitas dari kedua algoritme dengan menggunakan metode *silhouette coefficient*. Dari penelitian ini didapatkan hasil bahwa *Bisecting K-Means* tanpa normalisasi memiliki nilai *silhouette coefficient* yang lebih tinggi dibandingkan dengan *Fuzzy C-Means*. Nilai *silhouette coefficient Bisecting K-Means* sebesar 0,588 dan 0,579 dengan normalisasi, sedangkan nilai *silhouette coefficient Fuzzy C-Means* adalah 0,488 dan 0,582 dengan normalisasi.

Kata Kunci– *Credit Card; Data Mining; Bisecting K-Means; Fuzzy C-Means; Silhouette Coefficient*

I. PENDAHULUAN

SEJALAN dengan perkembangan teknologi yang kian maju di era modern ini, semakin banyak pula pengaruh teknologi di berbagai bidang tidak terkecuali di bidang finansial. Salah satu contohnya adalah dengan semakin maraknya penggunaan kartu kredit di kalangan masyarakat. Kartu kredit dapat didefinisikan sebagai transaksi modern dalam bidang ekonomi yang menggunakan jasa bank atau perusahaan untuk menarik uang tunai dari bank serta perusahaan pembiayaan [1].

Berdasarkan data dari Bank Sentral Republik Indonesia (BI), jumlah alat pembayaran dengan menggunakan kartu (APMK) kategori kartu kredit yang beredar mengalami peningkatan yang cukup signifikan dari tahun ke tahun. Seperti pada tahun 2018 dengan jumlah pengguna kartu kredit sebanyak 17.275.128 meningkat pada Desember 2019 menjadi 17.487.057 kartu kredit [2]. Hal ini merupakan suatu peluang bagi para pelaku bisnis kartu kredit untuk mengembangkan bis-

Naskah diterima 4 September 2021, diterima setelah revisi 11 Agustus 2022, terbit online 2 September 2022. Emitor merupakan jurnal Teknik Elektro Universitas Muhammadiyah Surakarta yang terakreditasi Sinta 4 dengan alamat Gedung H Lantai 2 UMS, Jalan Ahmad Yani Tromol Pos 1 Surakarta Indonesia 57165.

nis kartu kredit. Pengembangan bisnis tidak lepas dari proses pemasaran, oleh karena itu pentingnya suatu perusahaan untuk memiliki strategi pemasaran yang sesuai dengan kebutuhan pasar dan konsumen yang menjadi target pasar, untuk menarik minat para pelanggan. Salah satu cara untuk menentukan strategi pemasaran adalah dengan proses segmentasi pada pelanggan.

Salah satu metode untuk melakukan segmentasi data adalah dengan menggunakan metode pada *data mining* yaitu *clustering*. *Clustering* merupakan pendekatan klasifikasi tanpa pengawasan (*unsupervised*) untuk mengenali pola, yang didasarkan pada pengelompokan objek yang memiliki kemiripan secara bersama-sama. Pendekatan ini berguna untuk menemukan pola dalam kumpulan data tidak berlabel [3]. Pada *clustering* kemiripan data dalam satu kelompok akan bernilai maksimum sedangkan kemiripan data antar kelompok akan bernilai minimum [4]. Oleh karena itu, pada penelitian ini akan dilakukan analisis perbandingan algoritme *clustering* yaitu algoritme *Bisecting K-Means* dan *Fuzzy C-Means* untuk menentukan algoritme yang lebih tepat digunakan pada proses segmentasi data pengguna kartu kredit.

Pada analisis mengenai performa *K-Means* dan *Bisecting K-Means* di data *web log* didapatkan kesimpulan bahwa performa *Bisecting K-Means* relatif lebih unggul dan efisien daripada *K-Means* berdasarkan nilai performa dan akurasi. Nilai akurasi akhir dari *Bisecting K-Means* sebesar 84,52% dan K-Means 78,75%, selain itu nilai performa dari *Bisecting K-Means* juga lebih unggul dibandingkan dengan *K-Means* dengan rata-rata performa dari setiap *log files* berkisar 70-80% dan K-Means berkisar 60-70% [5]. Pada penelitian mengenai analisis perbandingan *K-Means* dan *Fuzzy C-Means* untuk pengelompokan data *user knowledge modeling* menjelaskan bahwa *Fuzzy C-Means* memiliki nilai validitas yang lebih tinggi dibandingkan dengan *K-Means* dimana nilai PCI dari *Fuzzy C-Means* sebesar 0,2854 sedangkan nilai *silhouette coefficient* dari *K-Means* hanya sebesar 0,1866.

Penelitian ini bertujuan untuk membandingkan performa dari algoritme *Bisecting K-Means* dan *Fuzzy C-Means* pada data pengguna kartu kredit guna mengetahui algoritme yang memiliki performa lebih baik. Hasil dari penelitian ini berupa perbandingan validitas dari kedua algoritme berdasarkan nilai *silhouette coefficient*, dimana algoritme yang memiliki nilai *silhouette coefficient* lebih tinggi diindikasikan sebagai algoritme yang memiliki kualitas performa lebih baik. Diharapkan penelitian ini dapat dimanfaatkan oleh masyarakat untuk menunjang pembelajaran mengenai algoritme yang sama dan dapat dijadikan referensi untuk penelitian selanjutnya.

II. METODE PENELITIAN

i. Data Collection

Pada penelitian ini dataset diperoleh dari *Kaggle Dataset* yang dibuat oleh seorang *data scientist* asal India bernama Arjun Bhasin. Dataset ini merangkum 8951 tingkah laku para pengguna kartu kredit aktif selama 6 bulan dan memiliki 18 atribut. Tabel 1 adalah penjelasan rinci mengenai atribut yang digunakan.

Tabel 1: Keterangan atribut data kartu kredit

Nama Atribut	Keterangan
Cust id	Identifikasi pemegang kartu kredit
Balance	Jumlah saldo yang tersisa di akun mereka untuk melakukan pembelian
Balance Frequency	Seberapa sering saldo diperbarui
Purchases	Jumlah pembelian yang dilakukan dari akun
One off Purchases	Jumlah pembelian maksimum yang dilakukan dalam sekali transaksi
Installment Purchases	Jumlah pembelian yang dilakukan dengan mencicil
Cash Advance	Uang muka yang diberikan oleh pengguna
Purchases Frequency	Seberapa sering melakukan pembelian
Cash advance trx	Jumlah transaksi yang dilakukan dengan tunai
One offpurchase frequency	Seberapa sering pembelian terjadi dalam sekali transaksi
Purchases installment frequency	Seberapa sering pembelian dalam cicilan sedang dilakukan
Cash advance frequency	Uang tunai yang dibayar di muka
Tenure	Masa berlaku layanan kartu kredit untuk pengguna
Purchases Trx	Banyaknya transaksi pembelian yang dilakukan
Credit Limit	Batas kartu kredit untuk pengguna
Payments	Jumlah pembayaran yang dilakukan oleh pengguna
Minimum Payments	Jumlah minimum pembayaran yang dilakukan oleh pengguna
Prc full payment	Persen dari pembayaran penuh yang dibayarkan oleh pengguna

ii. Data Preprocessing

Beberapa dataset memiliki kualitas yang kurang baik seperti tidak lengkap, tidak konsisten bahkan memiliki *noise* yang dapat mempengaruhi hasil akhir dari proses *data mining*. Salah satu cara untuk meningkatkan kualitas data adalah dengan melakukan pengolahan data. *Preprocessing* adalah langkah pengolahan data pada *data mining* yang mempersiapkan dan mentransformasi data agar sesuai dengan proses *data mining*. *Preprocessing* bertujuan untuk mereduksi data, menormalisasi data dan menghilangkan *outlier*, pada proses *preprocessing* terdapat beberapa teknik seperti *data cleaning* dan *reduction* [6]. Pada penelitian ini akan dilakukan dua proses *preprocessing* yaitu normalisasi dengan menggunakan min – max *Normalization* dan proses *dimentional reduction* dengan menggunakan *principal component analysis (PCA)*.

Normalisasi min – max merupakan metode proses data yang menggunakan nilai maksimum dan minimum dari suatu atribut untuk mentransformasikan data ke rentang baru secara linier dengan rentang nilai 0 sampai dengan 1 sehingga menghasilkan perbandingan antar data yang seimbang, baik sebelum atau sesudah proses

normalisasi [7]. Persamaan 1 merupakan persamaan min − max:

$$x' = \frac{\min_R + (x - x_{\min})(\max_R - \min_R)}{x_{\max} - x_{\min}} \quad (1)$$

dengan x' merupakan data hasil normalisasi, x merupakan data asli, $x_{\max}$ merupakan nilai maksimal pada atribut, $x_{\min}$ merupakan nilai minimum pada atribut,$\min_R$ merupakan *minimum range*, dan $\max_R$ merupakan *maximum range*.

Principal Component Analysis (PCA) adalah transformasi linier untuk menentukan sistem koordinat baru dari sebuah *dataset* [8]. PCA bertujuan untuk menyederhanakan data berdimensi tinggi menjadi data dengan dimensi yang lebih kecil dan berfungsi sebagai ringkasan dari keseluruhan data tanpa mengubah pola dan karakteristik data. PCA mengubah satu set variabel yang saling terkait menjadi satu set variabel tidak memiliki korelasi yang disebut sebagai *principal component* [9]. Berikut adalah langkah dari algoritme PCA [10]. Dengan X berupa *dataset* yang memiliki himpunan n vektor $(x_1, x_2, \ldots, x_n)$ dan setiap X_i merupakan titik data ke$-i$ dari *dataset*, lakukan perhitungan rata-rata dari setiap dimensi dengan Persamaan 2.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2)$$

dengan n merupakan jumlah data dan $\bar{x}$ berupa nilai rata-rata. Untuk mendapatkan matriks kovarian digunakan Persamaan 3.

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (X_i - \bar{x}) - (Y_i - \bar{y})}{n-1} \quad (3)$$

Inti dari PCA adalah nilai eigen dan vektor eigen dari matriks kovarian, di mana nilai eigen akan menentukan arah dari dimensi yang baru dan vektor eigen akan menentukan besarnya. Untuk mendapatkan nilai eigen digunakan Persamaan 4.

$$Av = \lambda v, \quad (4)$$

dengan A merupakan matriks $n \times n$, v berupa nilai eigen dari A dan λ adalah nilai eigen. Kemudian nilai eigen yang dikomputasi, ditransformasikan menggunakan matriks identitas (I) dengan menggunakan Persamaan 5.

$$A - \lambda I = 0 \quad (5)$$

Setelah didapatkan nilai eigen, urutkan dari nilai tertinggi sampai terendah. Pada tahap ini vektor eigen yang sesuai dengan nilai eigen yang sudah diurutkan akan menjadi *principal component*.

iii. Data Processing

Data processing merupakan metode proses data untuk mendapatkan fitur yang diperlukan untuk proses berikutnya. Berikut data processing yang akan diuraikan yaitu Fuzzy C-Means dan bisecting K-Means.

Fuzzy C-Means (FCM) merupakan teknik pengklasteran data dengan keberadaan setiap objek dalam suatu *cluster* ditentukan oleh derajat keanggotaan tertentu [4]. Tujuan utama *Fuzzy C-Means* adalah membagi objek yang berada pada beberapa dimensi ke dalam jumlah *cluster* tertentu agar mendapatkan *centroid* yang dapat meminimalisasi ketidaksamaan antar *cluster*. *Fuzzy C-Means* termasuk ke dalam *soft clustering* yang memungkinkan objek untuk dimiliki oleh lebih dari satu *cluster* dengan derajat keanggotaan yang berbeda. Objek yang berada pada perbatasan *cluster* tidak sepenuhnya termasuk dalam salah satu *cluster*, melainkan dapat menjadi anggota dari beberapa *cluster* dengan derajat keanggotaan parsial antara 0 sampai dengan 1 [3]. Berikut adalah algoritme dari *Fuzzy C-Means* [11].

Inisiasi data berupa matriks berukuran $n \times m$, dengan n merupakan jumlah sampel data dan m adalah atribut setiap data. Kemudian tentukan jumlah *cluster* (c), pangkat (w), maksimum iterasi (max Iter), *error* terkecil (ε), fungsi objektif awal ($P_0 = 1$) dan iterasi awal ($t = 1$). Buat bilangan acak dengan $i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, n$ dan c sebagai elemen matriks partisi awal U. Hitung jumlah setiap kolom dengan Persamaan 6 dan nilai matriks partisi dengan Persamaan 7.

$$Q_i = \sum_{k=1}^{c} \mu_{ik} \quad (6)$$

$$\mu_{ik} = \frac{\mu_{ik}}{Q_i} \quad (7)$$

Lakukan perhitungan untuk mendapatkan pusat *cluster* ke$-k$ dengan menggunakan Persamaan (8) dan fungsi objektif pada iterasi ke$-t$ dengan Persamaan (9).

$$V_{kj} = \frac{\sum_{i=1}^{n} \mu_{ik}^{w} x_{ij}}{\sum_{i=1}^{n} \mu_{ik}^{w}} \quad (8)$$

$$P_t = \sum_{t=1}^{n} \sum_{k=1}^{c} \left(\sum_{j=1}^{m} \left(X_{ij} - V_{kj} \right)^2 \left(\mu_{ik}^{w} \right) \right) \quad (9)$$

Kemudian hitung perubahan matriks partisi dengan menggunakan Persamaan (10).

$$\mu_{ik} = \frac{\left(\sum_{j=1}^{m} \left(x_{ij} - v_{kj} \right)^2 \right)^{\frac{-1}{w-1}}}{\sum_{k=1}^{c} \left(\sum_{j=1}^{m} \left(x_{ij} - v_{kj} \right)^2 \right)^{\frac{-1}{w-1}}} \quad (10)$$

Selanjutnya adalah memastikan kondisi berhenti, terdapat beberapa kriteria untuk menentukan berhentinya proses perhitungan algoritme *Fuzzy C-Means* yaitu:

1. Jika: $(|P - P_{t-1} < \varepsilon|)$ atau $(t > \max \text{Iter})$
2. Jika tidak: $t = t + 1$, ulang kembali proses perhitungan pusat *cluster*

Bisecting K-Means merupakan algoritme berbasis *K-Means* yang mengkombinasikan algoritme *K-Means* dan hirarki *clustering* yang memiliki kepekaan terhadap data pencilan [12]. Pada setiap langkah *bisecting* atau membelah diri, hanya titik data pada *cluster* dan dua *centroid* yang terlibat dalam proses komputasi. Selain itu, *Bisecting K-Means* efektif dalam mengatasi situasi dimana algoritme memasuki kondisi optimal lokal sampai batas tertentu [5, 13]. *Bisecting K-Means* termasuk ke dalam *hard clustering* dimana setiap objek hanya memiliki satu keanggotaan *cluster*. Berikut adalah persamaan *Bisecting K-Means*. Buat seluruh data menjadi satu *cluster* dan tentukan nilai k, kemudian lakukan langkah *bisecting* dengan mencari dua *sub cluster* menggunakan algoritme *K-Means*. Berikut adalah langkah dari algoritme K-Means.

1. Inisiasi 2 *centroid cluster*
2. Untuk setiap data objek pada *cluster*, hitung kemiripan dengan kedua *centroid* dan tempatkan objek dengan *centroid* yang lebih dekat dengan menggunakan Persamaan (11) *Euclidean Distance*, dengan $D(i, j)$ merupakan jarak data i ke pusat *cluster* j, x_{ki} merupakan data ke$-i$ pada atribut ke$-j$, dan x_{ji} merupakan titik pusat ke$-j$ pada atribut k.
3. Hitung ulang kedua *centroid* berdasarkan letak objek yang baru
4. Ulangi langkah 2 dan 3 sampai konvergensi. Kemudian ulangi langkah *bisecting* hingga didapatkan *cluster* sebanyak nilai k yang ditetapkan di awal, pilih hasil *clustering* dengan tingkat kemiripan tinggi.

$$D(i, j) = \sqrt{(x_{1i} - x_{1j})^2 + \ldots + (x_{ki} - x_{kj})^2} \quad (11)$$

iv. Evaluasi Model

Silhouette Coefficient merupakan gabungan dari metode separasi dan kohesi yang berguna untuk mengetahui kualitas dan kekuatan *cluster* berdasarkan seberapa baik penempatan objek pada *cluster* [14]. Nilai *silhouette coefficient* berkisaran antara -1 sampai dengan 1 dimana nilai tersebut menunjukkan ketepatan penempatan objek dan seberapa besar kemiripan data suatu *cluster*, nilai yang mendekati 1 menunjukkan objek data berada pada *cluster* yang tepat. Sebaliknya jika nilai mendekati -1 menunjukkan bahwa rata-rata antar objek jauh. Nilai = 0 menunjukkan bahwa data berada di antara dua *cluster*. Untuk menghitung nilai *silhouette coefficient* terdapat 2 komponen yaitu $a(i)$ yang merupakan rata-rata jarak data ke$-i$ dengan data lainnya yang ada pada satu *cluster* dan $b(i)$ adalah rata-rata jarak data ke$-i$ dengan semua data yang ada pada *cluster* lain [15]. Komponen $a(i)$ dan $b(i)$ ditentukan dengan menggunakan Persamaan (12) dan Persamaan (13).

$$a(i) = \frac{1}{m_j - 1} \sum_{r=1} d(x_i^j, x_r^j) \quad (12)$$

$$b(i) = min_{n=j}^{n=,\ldots,k} \left\{ \sum_{r \neq 1}^{r=1} d(x_i^j, x_r^j) \right\} \quad (13)$$

Persamaan 14 berikut merupakan persamaan dari *Silhouette Coefficient*.

$$S = \frac{b(i) - a(i)}{max\{b(i), a(i)\}}, \quad (14)$$

dengan j merupakan *cluster*, i merupakan indeks data ($i = 1, 2, \ldots, m_j$), dengan m_j merupakan jumlah data dalam *cluster* ke$-j$, $d(x_i^r, x_j^r)$ merupakan jarak data ke$-i$ dengan data ke$-r$ dalam satu *cluster* j, dan S merupakan nilai *silhouette coefficient*.

III. HASIL PENELITIAN DAN DISKUSI

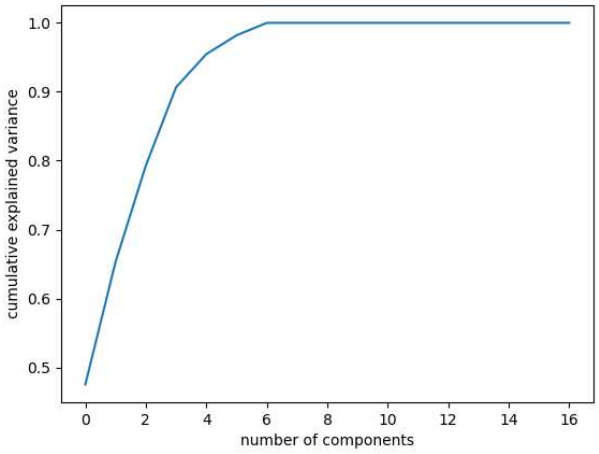
Dataset dengan jumlah keseluruhan data sebanyak 8951 data dan 18 atribut mengenai tingkah laku para pengguna kartu kredit aktif pada suatu bank digunakan pada penelitian ini. Sebelum dataset melalui tahap *preprocessing* dan *processing*, salah satu dimensi pada dataset akan dikurangi dikarenakan keberadaannya tidak diperlukan sehingga jumlah dimensi pada dataset berkurang menjadi 17 dimensi, dimensi tersebut adalah 'CUST-ID'. Tidak semua dataset memiliki kualitas yang baik. Terdapat dataset yang memiliki beberapa permasalahan seperti data yang tidak konsisten, adanya data *noise*, *outlier* dan penskalaan data yang berbeda dapat mempengaruhi hasil dari proses *data mining* sehingga dilakukan tahap *preprocessing* untuk meningkatkan kualitas dari data.

Tahap awal *preprocessing* adalah penanganan *missing values* untuk menyeimbangkan data. Selain dapat menyebabkan pendistribusian dan variasi data yang tidak seimbang, *missing values* juga menyebabkan analisa statistik yang tidak akurat [16]. *Missing values* adalah nilai data yang tidak tersimpan pada suatu variabel, *missing values* dapat mengurangi kekuatan statistik dari analisis yang dapat merusak validitas hasil dan

estimasi yang bias [17]. Dengan pertimbangan jumlah *missing values* sebanyak 314 data, di mana 313 data berasal dari dimensi 'TENURE' dan 1 data berasal dari dimensi 'CREDIT-LIMIT' yang kemudian didapatkan persentase *missing values* dari jumlah keseluruhan data sebesar 3,5%. Pada penelitian ini metode *imputation* digunakan untuk menangani *missing values*, di mana *missing values* akan diganti dengan nilai rata-rata yang didapat melalui perhitungan nilai rata-rata dari data *non-missing values* [18].

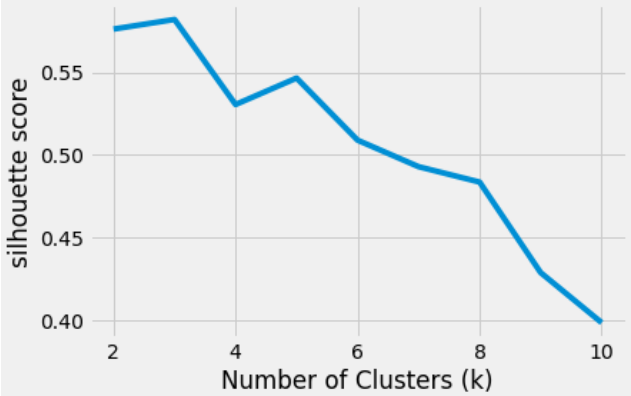
Proses selanjutnya adalah proses normalisasi data untuk menyamakan perbedaan skala agar data tidak menjadi timpang dan dapat menemukan rentang baru dari rentang yang ada [19]. Metode normalisasi Min-Max digunakan pada penelitian ini karena Min-Max dapat mengubah data yang kompleks dengan tetap mempertahankan hubungan antar nilai pada data [20, 21].

Data yang sudah dinormalisasi dilanjutkan dengan proses reduksi dimensi menggunakan metode PCA untuk menyederhanakan kompleksitas data berdimensi tinggi dengan tetap mempertahankan pola dan karakteristik data. Selain itu, PCA juga mengurangi data dan memproyeksikan secara geometris kepada dimensi yang lebih kecil dan disebut dengan PC, dengan tujuan untuk menemukan ringkasan terbaik dari data [22]. Sebanyak 17 dimensi dirangkum dan direduksi sesuai dengan nilai PC yang didapatkan yaitu 2 dimensi. Nilai PC diperoleh melalui pengujian nilai PC 1 sampai dengan 17 dengan menggunakan metode Silhouette untuk menemukan PC optimal yang memiliki Indeks Silhouette tertinggi. Kemudian didapatkan nilai PC optimal = 2, dengan nilai *cumulative variance* sebesar 80,2% yang dapat mengurangi 88,2% dari total dimensi yaitu sebanyak 15 dimensi, seperti yang terdapat pada Gambar 1.

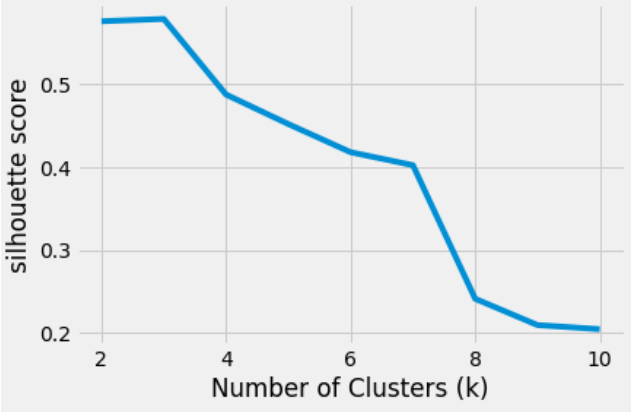


Gambar 1: Nilai *Principal Component*

Dataset yang sudah melalui tahap *preprocessing* dilanjutkan dengan tahap *processing* menggunakan algoritme *Fuzzy C-Means* dan *Bisecting K-Means*. Untuk mendapatkan hasil *clustering* terbaik dibutuhkan k yang optimal. Pada penelitian ini, k optimal didapatkan melalui pengujian nilai k pada rentang dengan menerapkan metode *silhouette coefficient*, k dengan nilai *silhouette coefficient* tertinggi diindikasikan sebagai k optimal. Pada *Fuzzy C-Means* dan *Bisecting K-Means* didapatkan k optimal dari setiap algoritme yaitu 3.

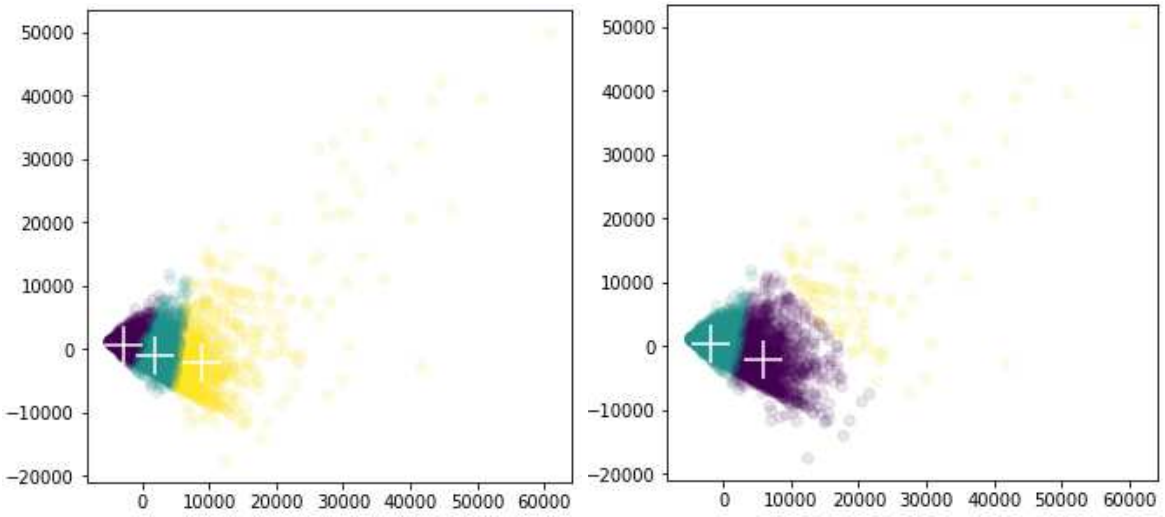


Gambar 2: Nilai k Optimal *Fuzzy C-Means*

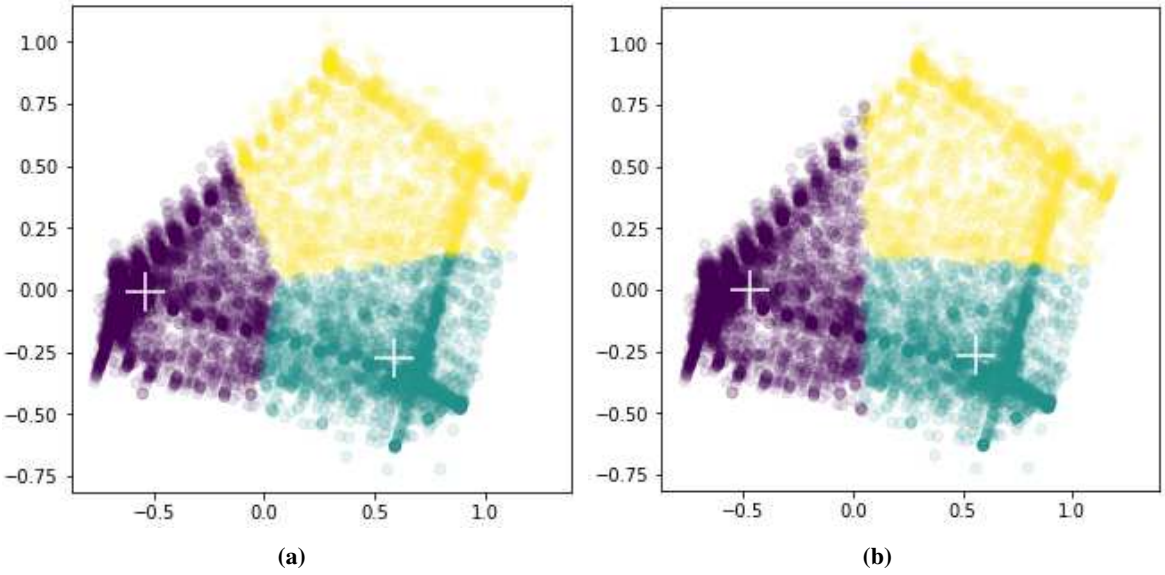


Gambar 3: Nilai k Optimal *Bisecting K-Means*

Gambar 2 dan 3 merupakan hasil dari pengujian nilai k optimal pada *Fuzzy C-Means* dan *Bisecting K-Means* yang menunjukkan adanya penurunan nilai *silhouette coefficient* secara berkala seiring bertambahnya nilai k, setelah mendapatkan nilai k optimal yaitu k = 3. Hal ini dapat diartikan bahwa semakin banyak *cluster* maka semakin rendah ketepatan penempatan data pada *cluster*. Selain itu, semakin tinggi nilai *silhouette coefficient* menunjukkan bahwa semakin baik proses kohesi dan separasi pada *cluster* [23]. Dalam pengujian ini juga dapat disimpulkan bahwa semakin banyak nilai k, maka jarak antar objek pada *cluster* semakin besar dan jarak antar *cluster* semakin kecil.



Gambar 4: Proses *clustering* sebelum normalisasi (a) *clustering Fuzzy C-Means* (b) *Bisecting K-Means*



Gambar 5: Proses *clustering* sesudah normalisasi (a) *clustering Fuzzy C-Means* (b) *Bisecting K-Means*

Pada tahap *processing* menggunakan *Fuzzy C-Means*, langkah awal yang dilakukan adalah melakukan inisiasi pada beberapa parameter, yaitu maksimal iterasi sebanyak 30 iterasi, $\varepsilon = 0,00001$, $P_0 = 0$. Lalu dilanjutkan dengan perhitungan *silhouette coefficient* dari *Fuzzy C-Means*. Didapatkan nilai *silhouette coefficient* sebelum dan sesudah normalisasi seperti pada Tabel 2.

Tabel 2: Nilai *Silhouette Coefficient Fuzzy C-Means*

	Nilai *Silhouette Coefficient*
Tanpa Normalisasi	0,488
Normalisasi	0,582

Nilai *silhouette coefficient* dari *Fuzzy C-Means* setelah proses normalisasi dengan menggunakan metode Min-Max mengalami peningkatan sebesar 0,094 yaitu dari 0,488 menjadi 0,582. Dikarenakan pada proses normalisasi menggunakan metode Min-Max, data ditransformasikan kepada interval 0 - 1 sehingga didapatkan hasil yang lebih baik dan akurat [24]. Gambar 4 (a) dan 5 (a) menunjukkan hasil persebaran data sebelum dan sesudah melalui proses normalisasi.

Pada *Bisecting K-Means* nilai *silhouette coefficient* dari data sebelum dan sesudah normalisasi mengalami penurunan sebesar 0,009, dengan nilai *silhouette coefficient* sebelum normalisasi sebesar 0,588 menjadi 0,579 setelah melalui proses normalisasi. Hal ini dapat

menandakan bahwa pada penelitian ini algoritme *Bisecting K-Means* memiliki ketepatan penempatan data pada *cluster* yang lebih baik dengan data tanpa normalisasi. Hasil pada Tabel 2 didapatkan melalui pengujian *Bisecting K-Means* dengan k optimal = 3 dan inisiasi maksimum iterasi sebanyak 30 iterasi. Gambar 4 (b) dan 5 (b) menunjukkan hasil persebaran data sebelum dan sesudah proses normalisasi pada *Bisecting K-Means*.

Tabel 3: Nilai *Silhouette Coefficient Bisecting K-Means*

	Nilai *Silhouette Coefficient*
Tanpa Normalisasi	0,588
Normalisasi	0,579

Berdasarkan pada Tabel 2 dan 3, *Bisecting K-Means* tanpa normalisasi memiliki nilai *silhouette coefficient* yang lebih tinggi dibandingkaan dengan *Fuzzy C-Means* baik sebelum maupun sesudah proses normalisasi. Hal ini dapat disimpulkan bahwa *Bisecting K-Means* memiliki kualitas dan kekuatan *clustering* berdasarkan ketepatan penempatan data yang lebih baik dibandingkan *Fuzzy C-Means*.

IV. KESIMPULAN

Berdasarkan penelitian mengenai perbandingan algoritme *Fuzzy C-Means* dan *Bisecting K-Means* pada data pengguna kartu kredit yang telah dilakukan didapatkan hasil bahwa algortima *Bisecting K-Means* memiliki nilai *silhouette coefficient* yang lebih tinggi dibandingkan dengan *Fuzzy C-Means*. Selain itu penggunaan metode normalisasi Min-Max juga mempengaruhi hasil dari *Fuzzy C-Means* dan *Bisecting K-Means*. Pada *Fuzzy C-Means* data yang telah melalui proses normalisasi mengalami peningkatan nilai *silhouette coefficient* sedangkan pada *Bisecting K-Means* mengalami penurunan. Pada penelitian ini terdapat beberapa hal yang dapat dikembangkan pada penelitian selanjutnya yaitu dengan menggunakan dataset yang lebih baik dengan sedikit data *noise*, *missing values*, serta *outlier*. Selain itu dapat dilakukan modifikasi pada algoritme dan proses *preprocessing* untuk mendapatkan hasil yang lebih optimal.

DAFTAR PUSTAKA

[1] R. N. Pramuhadi *et al.*, "Gaya hidup penggunaan kartu kredit masyarakat urban di surabaya," Ph.D. dissertation, Universitas Airlangga, 2019.

[2] B. Indonesia, "Jumlah apmk beredar," *Diambil dari www.bi.go.id/ id/ statistik/ sistempembayaran/ apmk/ contents/ jumlah apmk beredar.aspx*, 2020.

[3] S. Zahra, M. A. Ghazanfar, A. Khalid, M. A. Azam, U. Naeem, dan A. Prugel-Bennett, "Novel centroid selection approaches for kmeans-clustering based recommender systems," *Information sciences*, vol. 320, pp. 156–189, 2015.

[4] Y. Yohannes, "Analisis perbandingan algoritma fuzzy c-means dan k-means," in *Annual Research Seminar (ARS)*, vol. 2, no. 1, 2017, pp. 151–155.

[5] K. Abirami dan P. Mayilvahanan, "Performance analysis of k-means and bisecting k-means algorithms in weblog data," *Int. J. Emerg. Technol. Eng. Res*, vol. 4, no. 8, pp. 119–124, 2016.

[6] R. Tamilselvi, B. Sivasakthi, dan R. Kavitha, "An efficient preprocessing and postprocessing techniques in data mining," *Int. J. Res. Comput. Appl. Robot*, vol. 3, no. 4, pp. 80–85, 2015.

[7] D. A. Nasution, H. H. Khotimah, dan N. Chamidah, "Perbandingan normalisasi data untuk klasifikasi wine menggunakan algoritma k-nn," *CESS (Journal of Computer Engineering, System and Science)*, vol. 4, no. 1, pp. 78–82, 2019.

[8] A. R. Syakhala, D. Puspitaningrum, dan E. P. Purwandari, "Perbandingan metode principal component analysis (pca) dengan metode hidden markov model (hmm) dalam pengenalan identitas seseorang melalui wajah," *Rekursif: Jurnal Informatika*, vol. 3, no. 2, 2015.

[9] A. Jamal, A. Handayani, A. A. Septiandri, E. Ripmiatin, dan Y. Effendi, "Dimensionality reduction using pca and k-means clustering for breast cancer prediction," *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, pp. 192–201, 2018.

[10] C. Zhu, C. U. Idemudia, dan W. Feng, "Improved logistic regression model for diabetes prediction by integrating pca and k-means techniques," *Informatics in Medicine Unlocked*, vol. 17, p. 100179, 2019.

[11] F. Febrianti, M. Hafiyusholeh, dan A. H. Asyhar, "Perbandingan pengklusteran data iris menggunakan metode k-means dan fuzzy c-means," *Jurnal Matematika" MANTIK*, vol. 2, no. 1, pp. 7–13, 2016.

[12] Z. Zhou, A. Ran, S. Chen, X. Zhang, G. Wei, B. Li, F. Kang, X. Zhou, dan H. Sun, "A fast screening framework for second-life batteries based on an improved bisecting k-means algorithm combined with fast pulse test," *Journal of Energy Storage*, vol. 31, p. 101739, 2020.

[13] F. Zhang dan S. Wang, "Detecting group shilling attacks in online recommender systems based on bisecting k-means clustering," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 5, pp. 1189–1199, 2020.

[14] T. M. Kodinariya dan P. R. Makwana, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[15] E. Kuswantoro, "Komparasi gabungan algoritma average linkage dan k-means dengan kmeans clustering untuk analisa faktor pengangguran dan angkatan kerja-comparison of algorithms average linkage and k-means clustering with k-means clustering to analyze the unemployment factor on work force," Ph.D. dissertation, Institut Teknologi Sepuluh Nopember, 2016.

[16] E. G. Armitage, J. Godzien, V. Alonso-Herranz, Á. López-Gonzálvez, dan C. Barbas, "Missing value imputation strategies for metabolomics data," *Electrophoresis*, vol. 36, no. 24, pp. 3050–3060, 2015.

[17] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013.

[18] T. Aljuaid dan S. Sasi, "Proper imputation techniques for missing values in data sets," in *2016 international conference on data science and engineering (ICDSE)*. IEEE, 2016, pp. 1–5.

[19] S. G. K. Patro, "Kk sahu.(2015)," *Normalization: A Preprocessing Stage. IARJSET*, vol. 2, no. 3, pp. 20–22.

[20] D. Fenny, "Analisis perbandingan cosine normalization dan min-max normalization pada pengelompokan terjemahan ayat al quran menggunakan algoritma k-means clustering," B.S. thesis, Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah, 2019.

[21] B. K. Singh, K. Verma, dan A. Thoke, "Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification," *International Journal of Computer Applications*, vol. 116, no. 19, 2015.

[22] J. Lever, "Krzywinski., m., & altman, n.(2017). principal component analysis," *Nature Methods*, vol. 14, no. 7, pp. 641–642.

[23] D. M. Eler, J. B. M. Teixeira, P. A. Macanha, dan R. E. Garcia, "Simplified stress and simplified silhouette coefficient to a faster quality evaluation of multidimensional projection techniques and feature spaces," in *2015 19th International Conference on Information Visualisation*. IEEE, 2015, pp. 133–139.

[24] D. A. Anggoro dan N. D. Kurnia, "Comparison of accuracy level of support vector machine (svm) and k-nearest neighbors (knn) algorithms in predicting heart disease," *International Journal*, vol. 8, no. 5, 2020.