**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 1010−1017
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2022

# Seleksi Fitur Menggunakan Eigen Vector Untuk Peningkatan Kinerja K-Means Clustering Dalam Pengelompokan Data

**Nugroho Syahputra, Muhammad Zarlis*, Syahril Efendi**

Fakultas Ilmu Komputer dan Teknologi Informasi, Prodi S2 Teknik Informatika, Universitas Sumatera Utara, Medan, Indonesia
Email: [1]nughasyah@gmail.com, [2,*]m.zarlis@usu.ac.id, [3]syahril1@usu.ac.id
Email Penulis Korespondensi: m.zarlis@usu.ac.id
Submitted:**01/08/2022**; Accepted:**10/09/2022**; Published: **30/09/2022**

**Abstrak**−Banyaknya jumlah atribut data set dari proses pengelompokan data dengan K-Means Clustering dapat mempengaruhi besaran jumlah iterasi yang dihasilkan. Pada riset ini, Eigen Vector digunakan untuk melakukan seleksi fitur pada data set. Data set yang telah diseleksi selanjutnya dilakukan proses clustering dengan K-Means Clustering. Data set yang digunakan pada riset ini adalah Wine Quality Dataset yang diperoleh dari UCI Machine Learning Repository, dengan 11 atribut, 4898 records data dan 7 kelas atribut. Kemudian South German Credit Dataset diperoleh dari kaggle.com dengan 20 atribut, 1000 records data dan 2 kelas atribut. Hasil dari riset ini menunjukkan bahwa jumlah iterasi yang diperoleh dari perbandingan pengujian dengan menggunakan K-Means tanpa seleksi fitur yaitu pada Wine Quality Dataset diperoleh jumlah 11 iterasi, dan pada South German Credit Dataset diperoleh jumlah 10 iterasi. Sedangkan K-Means dengan seleksi fitur Eigen Vector diperoleh jumlah iterasi pada Wine Quality Dataset dengan jumlah 5 iterasi, dan pada South German Credit Dataset yaitu dengan jumlah 4 iterasi. Evaluasi clustering dihitung menggunakan Sum of Square Error (SSE). Nilai SSE pada K-Means Clustering tanpa seleksi fitur dari Wine Quality Dataset yaitu sebesar 678.5735, dan pada South German Credit Dataset yaitu sebesar 1534.3167. Sedangkan pada K-Means Clustering dengan Eigen Vector dari Wine Quality Dataset yaitu sebesar 383.0517, dan pada South German Credit Dataset yaitu sebesar 469.0698. Dari hasil metode yang diusulkan mampu dalam menurunkan persentase error serta meminimalkan jumlah iterasi pada K-Means Clustering dengan seleksi fitur menggunakan Eigen Vector.

**Kata Kunci:** Clustering; K-Means Clustering; Eigen Vector; Sum of Square Error

**Abstract**−The large number of data set attributes from the data grouping process with K-Means Clustering can affect the number of iterations produced. In this research, Eigen Vector is used to perform feature selection on the data set. The selected data set is then clustered using K-Means Clustering. The data set used in this research is the Wine Quality Dataset obtained from the UCI Machine Learning Repository, with 11 attributes, 4898 data records and 7 attribute classes. Then the South German Credit Dataset was obtained from kaggle.com with 20 attributes, 1000 data records and 2 attribute classes. The results of this research indicate that the number of iterations obtained from the comparison of tests using K-Means without feature selection is that in the Wine Quality Dataset, 11 iterations are obtained, and in the South German Credit Dataset, there are 10 iterations. Meanwhile, K-Means with Eigen Vector feature selection obtained the number of iterations in the Wine Quality Dataset with a total of 5 iterations, and in the South German Credit Dataset with a total of 4 iterations. Clustering evaluation was calculated using Sum of Square Error (SSE). The SSE value in K-Means Clustering without feature selection from the Wine Quality Dataset is 678.5735, while in the South German Credit Dataset it is 1534.3167. While the K-Means Clustering with Eigen Vector from the Wine Quality Dataset is 383.0517, and the South German Credit Dataset is 469.0698. From the results of the proposed method is able to reduce the percentage of errors and minimize the number of iterations on K-Means Clustering with feature selection using Eigen Vector.

**Keywords**: Clustering; K-Means Clustering; Eigen Vector; Sum of Square Error

## 1. PENDAHULUAN

Secara prinsip, *K-Means Clustering* adalah metode untuk pengelompokan sekumpulan data atau biasa juga disebut metode *clustering* [1]. *Clustering* sendiri merupakan proses untuk mengelompokkan data menjadi sejumlah kelompok (*cluster*) dari data multidimensi berdasarkan ukuran kesamaan atau kemiripan. Dan pada umumnya proses perhitungan pada metode *K-Means Clustering* yang sangat sederhana dan sering digunakan [2].

Akan tetapi pada beberapa riset ditemukan kelemahan dari *K-Means Clustering* dalam melakukan proses klasterisasi pada suatu kumpulan data. Salah satu permasalahan tersebut yaitu banyaknya atribut dataset yang diujikan dan berpengaruh pada jumlah iterasi yang dihasilkan dalam proses *clustering* sampai kondisi *convergen* dan hal tersebut berimplikasi pada kinerja *K-Means* sehingga kinerja yang diperoleh kurang maksimal [3].

Solusi yang dapat dilakukan dalam permasalahan optimasi kinerja pada pengelompokan data salah satunya yaitu dengan cara melakukan seleksi atribut pada data yang diujikan dengan membuang atribut yang memiliki pengaruh kecil pada data yang diujikan [4].

Beberapa penelitian terdahulu yang meneliti tentang seleksi fitur ataupun reduksi atribut pada *K-Means Clustering* seperti penelitian [5] yang meneliti tentang reduksi atribut dataset dengan menggunakan *Information Gain* pada algoritma *K-Means Clustering* dengan perhitungan kinerja dihitung berdasarkan nilai *Davies Bouldin Index* (DBI). Hasil yang diperoleh dari penelitian tersebut yaitu *Information Gain* mampu dalam meminimalkan jumlah iterasi clustering sampai dengan kondisi convergen jika dibandingkan dengan *K-Means Clustering* tanpa reduksi atribut. Kemudian penelitian dari [6] yang meneliti tentang pengoptimasian pada *K-Means Clustering* dengan reduksi atribut menggunakan *Principal Component Analysis* dalam proses pemetaan kinerja dosen. Hasil yang diperoleh diketahui bahwa metode PCA terbukti dapat meningkatkan kualitas *cluster* dengan pengukuran validitas

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 1010−1017
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2022

menggunakan *Davies Bouldin Index* yang menunjukkan bahwa kombinasi PCA dan *K-Means* menghasilkan nilai DBI paling kecil dibandingkan dengan *K-Means* tanpa reduksi atribut.

Setiap fitur pada data set memiliki nilai pengaruh yang berbeda-beda, dan juga data dengan atribut yang terlalu banyak tentunya akan memperlambat kinerja metode *clustering* [7]. Adapun metode seleksi fitur yang diusulkan untuk menyeleksi fitur pada data set yaitu dengan menggunakan perhitungan *Eigen Vector* yang diperoleh dari metode *Analytic Hierarchy Process* (AHP). Berdasarkan penelitian [8] bahwa *Analytical Hierarchy Process* diusulkan untuk menyelesaikan masalah dalam kasus *Multi-Criteria Group Decision-Making* di bawah *Probabilistic Dual Hesitant Fuzzy Information*, sehingga fitur yang banyak dalam suatu data dapat dibobotkan sesuai tingkat kepentingannya dalam pengambilan keputusan.

Kemudian pada penelitian ini metode *Eigen Vector* digunakan untuk mengukur bobot pengaruh dari suatu data sebelum dilakukan proses klasterisasi. Kemudian bobot yang dihasilkan dari perhitungan Eigen Vector tersebut nantinya akan dinormalisasikan menggunakan perhitungan normalisasi *min-max*, dengan bobot terendah yaitu 0 dan bobot tertinggi yaitu adalah 1 [9]. Sehingga dengan menggunakan *Eigen Vector*, nantinya dapat memberikan korelasi yang baik bagi data sebelum dilakukannya *clustering* dan memberikan pengaruh yang signifikan untuk optimasi kinerja pada *K-Means Clustering* dan diharapkan mampu dalam meningkatkan kinerja *K-Means Clustering*.

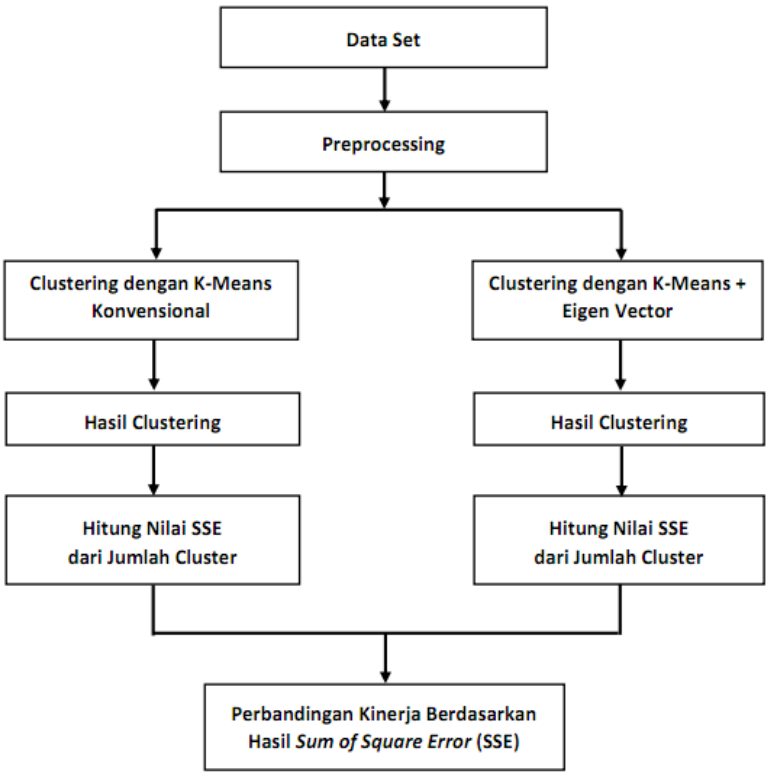
Untuk membuktikan hasil dari metode yang diusulkan pada riset ini dalam melihat tingkat keberhasilan penyelesaian metode yang diusulkan tersebut, dibuktikan dengan evaluasi kinerja *clustering* dengan menggunakan perhitungan *Sum of Square Error* (SSE) pada proses *clustering*. *Sum of Square Error* (SSE) merupakan hasil penjumlahan dari seluruh jarak masing-masing data dengan titik pusat clusternya. Semakin kecil nilai SSE yang didapat, semakin seragam data yang ada didalam masing-masing cluster, semakin baik cluster yang dihasilkan [10].

Pada riset ini, sampel data yang akan diujikan pada metode yang diusulkan yaitu menggunakan *Wine Quality Dataset* yang diperoleh daris *UCI Machine Learning Repository* dan *South German Credit Dataset* yang diperoleh dari *kaggle.com*. Pada data set tersebut akan dilakukan seleksi atribut dan kemudian menganalisis jumlah iterasi dan hasil kinerja *clustering* pada *K-Means Clustering*.

## 2. METODOLOGI PENELITIAN

### 2.1 Tahapan Penelitian

Tahapan pada penelitian ini dilakukan dengan tahapan penelitian pada Gambar 1 berikut.




**Gambar 1.** Tahapan Penelitian

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 1010−1017
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2022

### 2.1.1 Dataset

Data yang digunakan berupa file dataset berformat .csv yaitu Wine Quality Dataset dan South German Credit Dataset. *Wine Quality Dataset* yang diperoleh dari *UCI Machine Learning Repository*. Dataset ini berfokus pada klasifikasi dari kualitas *wine* berdasarkan tingkat kualitas dari atribut-atribut yang ada pada data set tersebut. Jumlah *records data* pada data set tersebut berjumlah 4898 records data. Kemudian jumlah fitur atau atribut pada *Wine Quality Dataset* berjumlah 11 atribut dan 1 kelas atribut dengan banyak kelas atribut yaitu 7 kelas yang menyatakan kualitas dari masing-masing *record* data *wine* pada data set tersebut. Kemudian dataset yang lain yaitu diperoleh dari *Kaggle.com* yaitu *South Germany Credit* yang merupakan dataset pengajuan kredit. Jumlah records data yang ada pada data set tersebut yaitu 1000 record data dengan jumlah atribut data yaitu 20 atribut dan 1 output atribut dengan 2 kelas atribut.

### 2.1.2 Preprocessing

*Preprocessing* bertujuan untuk mengolah data agar diperoleh data yang relevan [11] serta mengurangi *noise* pada data yang dilakukan dengan menggunakan Normalisasi Z-Score [12] dengan persamaan berikut.

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Adapun keterangan pada rumus persamaan (1) yaitu z menunjukkan nilai *standard score*, x menunjukkan data observasi, kemudian μ menunjukkan nilai *mean* per *variabel* dan σ menunjukkan nilai standar deviasi per variable.

### 2.1.3 K-Means Clustering

*K-Means Clustering* merupakan metode umum dan paling sederhana dalam *clustering* [13]. *K-Means* digunakan dalam mengelompokkkan data menjadi beberapa kelompok tanpa mengetahui target kelasnya. Hasil proses *cluster* dipengaruhi oleh pada nilai *centroid* awal. Proses *K-Means* sebagai berikut:

a. Penentuan nilai jumlah *cluster* (*k*)
b. Pemilihan titik awal *cluster* (*centroid*) berdasarkan nilai k
c. Perhitungan jarak antar data (*Euclidean Distance*) berdasarkan persamaan (2) berikut:

$$dist(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

d. Mengelompokkan data berdasarkan kedekatan data dengan *cluster* awal.
e. Menghitung *mean* dari data yang berada pada *centroid* yang sama untuk menentukan *cluster centroid* baru dengan cara:

$$v_{ij} = \frac{1}{N_i}\sum_{k=0}^{n} x_{kj} \quad (3)$$

f. Mengulangi langkah tiga jika sampai memenuhi kondisi *convergen*.

### 2.1.4 Eigen Vector

*Eigen Vector* dihitung berdasarkan *Analytical Hierarchy Process* yang merupakan metode pendukung keputusan yang dikembangkan oleh Thomas L. Saaty yang menguraikan masalah berbasis multi kriteria yang kompleks menjadi suatu hierarki dari sebuah permasalahan yang kompleks dalam suatu struktur multilevel dengan level pertama adalah tujuan, yang diikuti level faktor, kriteria, sub kriteria, dan seterusnya hingga level terakhir dari alternatif [14].

Secara umum perhitungan dengan metode AHP didasarkan pada langkah langkah berikut [15]:

a. Mendefinisikan permasalahan dan menentukan solusi yang diinginkan.
b. Membuat struktur hirarki, dilanjutkan dengan kriteria-kriteria dan alternaif pilihan yang ingin di rangking.
c. Membentuk matriks perbandingan berpasangan yang dilakukan berdasarkan pilihan atau judgement dari pembuat keputusan dengan menilai tingkat kepentingan elemen dibandingkan elemen lainnya.
d. Hitung normalisasi bobot dengan membagi nilai bobo telemen dengan nilai total setiap kolom berdasarkan persamaan (4):

$$Normalisasi\ Bobot = \frac{\sum Bobot\ Elemen}{\sum Sel\ Kolom} \quad (4)$$

e. Hitung nilai *eigen vector* berdasarkan persamaan (5):

$$Bobot\ Eigen\ Vector = \frac{\sum Sel\ Baris}{\sum Sel\ Kolom} \quad (5)$$

f. Hitung nilai eigen maksimum dengan menjumlahkan hasil perkalian jumlah kolom dengan *vector eigen* berdasarkan persamaan (6):

$$\sum Eigen\ Maks\ = \sum Kolom\ x\ Vector\ Eigen \quad (6)$$

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 1010−1017
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2022

g. Menguji nilai rasio konsistensi dengan syarat CR < 0,1 berdasarkan persamaan (7):

$$CR = \frac{CI}{RI} \quad (7)$$

Kemudian untuk menentukan CI (*Consistency Ratio*) berdasarkan persamaan (8):

$$CI = \frac{\lambda\, maks - n}{n-1} \quad (8)$$

Keterangan:
CI = *Consistency Index*
λ maks = Nilai eigen maksimal
n = Banyak unsur kriteria / sub kriteria

Untuk Nilai Random Index (RI) yang dipergunakan adalah berdasarkan yang ditetapkan oleh Saaty yaitu sebagai berikut

**Tabel 1.** Nilai Random Index (Tabel Skala *Saaty*) [15]

| N | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RI | 0 | 0 | 0.58 | 0.9 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 | 1.49 | 1.51 | 1.48 | 1.56 | 1.57 | 1.59 |

### 2.2 Sum of Square Error (SSE)

Sum of Square Error berfungsi dalam mengukur kualitas dari keanggotaan cluster serta merupakan hasil penjumlahan dari seluruh jarak masing-masing data dengan titik pusat clusternya [16], Jika semakin kecil nilai SSE maka semakin tinggi tingkat similaritas data yang ada di dalam masing-masing cluster atau semakin baik cluster yang dihasilkan [17]. *Sum-of-Square-Error* (SSE) dihitung dengan persamaan (9) berikut:

$$SSE = \sum_{k=1}^{K} (x_i - C_k)^2 \quad (9)$$

*SSE* merupakan nilai derajat error antara data kedalam masing-masing centroid cluster, *K* merupakan banyaknya jumlah cluster, *Ck* merupakan nilai centroid dari cluster ke- *k* dan *xi* adalah nilai dari data ke-*I*.

## 3. HASIL DAN PEMBAHASAN

### 3.1 Hasil Penelitian

Untuk membuktikan dari kinerja metode yang diusulkan pada penelitian ini, nantinya akan diujikan serta membandingkan hasil kinerja *clustering* yang dihasilkan dari pengujian *K-Means Clustering* tanpa seleksi fitur dengan *K-Means Clustering* menggunakan seleksi fitur berdasarkan *Eigen Vector*.

Langkah pertama yang dilakukan yaitu melakukan proses *preprocessing* untuk menormalisasi data set yang akan di ujikan pada metode yang diusulkan. Normalisasi data set dilakukan dengan perhitungan normalisasi *z-score* sesuai dengan rumus persamaan (1). Dan hasil normalisasi data set

**Tabel 2.** Normalisasi Wine Quality Dataset

| No. | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.172 | -0.08 | 0.213 | 2.821 | -0.03 | 0.569 | 0.744 | 2.331 | -1.24 | -0.34 | -1.39 | 6 |
| 2 | -0.65 | 0.215 | 0.047 | -0.94 | 0.147 | -1.25 | -0.14 | -0.01 | 0.739 | 0.001 | -0.82 | 6 |
| 3 | 1.475 | 0.017 | 0.543 | 0.100 | 0.193 | -0.31 | -0.97 | 0.358 | 0.475 | -0.43 | -0.33 | 6 |
| 4 | 0.409 | -0.47 | -0.11 | 0.415 | 0.559 | 0.687 | 1.120 | 0.525 | 0.011 | -0.78 | -0.49 | 6 |
| 5 | 0.409 | -0.47 | -0.11 | 0.415 | 0.559 | 0.687 | 1.120 | 0.525 | 0.011 | -0.78 | -0.49 | 6 |
| 6 | 1.475 | 0.017 | 0.543 | 0.100 | 0.193 | -0.31 | -0.97 | 0.358 | 0.475 | -0.43 | -0.33 | 6 |
| 7 | -0.77 | 0.414 | -1.43 | 0.119 | -0.03 | -0.31 | -0.05 | 0.291 | -0.05 | -0.17 | -0.74 | 6 |
| 8 | 0.172 | -0.08 | 0.213 | 2.821 | -0.03 | 0.569 | 0.744 | 2.331 | -1.24 | -0.34 | -1.39 | 6 |
| 9 | -0.65 | 0.215 | 0.047 | -0.94 | 0.147 | -1.25 | -0.14 | -0.01 | 0.739 | 0.001 | -0.82 | 6 |
| 10 | 1.475 | -0.57 | 0.791 | -0.96 | -0.08 | -0.42 | -0.22 | -0.07 | 0.210 | -0.34 | 0.394 | 6 |
| 11 | 1.475 | -0.47 | 0.626 | -0.97 | -0.58 | -1.42 | -1.77 | -1.07 | -1.31 | 0.614 | 1.207 | 5 |
| 12 | 2.068 | -0.97 | 0.543 | -0.43 | -0.49 | -1.07 | -0.69 | 0.224 | -0.31 | 0.351 | -0.66 | 5 |
| 13 | 1.238 | -1.17 | 0.295 | -1.02 | -0.26 | -1.13 | -1.49 | -0.67 | -0.05 | 1.228 | 0.232 | 5 |
| 14 | -0.30 | 1.406 | 0.543 | -0.96 | -0.08 | 0.746 | 0.109 | -0.94 | 2.329 | 0.264 | 1.532 | 7 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 4898 | 6 | 0.21 | 0.38 | 0.8 | 0.2 | 22 | 98 | 0.989 | 3.26 | 0.32 | 11.8 | 6 |

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 1010−1017
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2022

**Tabel 3.** Normalisasi South German Credit Dataset

| No. | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | ... | X20 | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -1.253 | -0.240 | 1,343 | -0.301 | -0.787 | -1,145 | 0.044 | -0.821 | 1,343 | ... | -0.738 | Good |
| 2 | -1.253 | -0.987 | 1,343 | -1,030 | -0.167 | -0.317 | -0.676 | 0.754 | 1,343 | ... | -0.738 | Good |
| 3 | -0.458 | -0.738 | -0.503 | 2,248 | -0.860 | 0.509 | 0.062 | 1,235 | 0.420 | ... | -0.738 | Good |
| 4 | -1.253 | -0.738 | 1,343 | -1,030 | -0.407 | -0.317 | 0.062 | 2,247 | 0.509 | ... | -0.904 | Good |
| 5 | -1.253 | -0.738 | 1,343 | -1,030 | -0.389 | -0.317 | 0.224 | -1,030 | -0.317 | ... | -1,070 | Good |
| 6 | -1.253 | -0.904 | 1,343 | -1,030 | -0.364 | -1,145 | -0.738 | 0.044 | -0.317 | ... | 0.044 | Good |
| 7 | -1.253 | -1,070 | 1,343 | 0.044 | -0.699 | 0.509 | -0.904 | -0.676 | -1,145 | ... | -0.676 | Good |
| 8 | -1.253 | -1,235 | 1,343 | -0.676 | -0.699 | -1,145 | -1,070 | 0.062 | 0.509 | ... | 0.062 | Good |
| 9 | -1.253 | -0.240 | 1,343 | 0.062 | -0.769 | -0.699 | -1,235 | 0.062 | -1,145 | ... | 0.062 | Good |
| 10 | 1.131 | 0.256 | -0.503 | 0.062 | 0.172 | 0.566 | -0.240 | 0.224 | -1.253 | ... | 0.224 | Good |
| 11 | -0.458 | -0.821 | 1,343 | 0.224 | -0.699 | -0.317 | -0.699 | 0.754 | -1.253 | ... | -0.666 | Good |
| 12 | -1.253 | 0.754 | 1,343 | -0.666 | -0.699 | 0.509 | -0.699 | 1,235 | -1.253 | ... | 0.062 | Good |
| 13 | -1.253 | 1,235 | 1,343 | 0.062 | -0.699 | 0.509 | -0.769 | 2,247 | -1.253 | ... | -1,070 | Good |
| 14 | -1.253 | 2,247 | 0.420 | 2,613 | -0.699 | -1,973 | 0.172 | 0.062 | -1.253 | ... | -1,235 | Good |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 1000 | -1.253 | 0.754 | -0.503 | -0.301 | 1.091 | 1.832 | 1.337 | 0.918 | 0.449 | ... | 0.460 | Good |

Langkah selanjutnya yaitu mengukur tingkat pengaruh fitur terhadap data set yang diujikan dengan menghitung bobot fitur. Dalam memperoleh nilai bobot pengaruh *Eigen Vector* pada data set yang digunakan, langkah-langkahnya yaitu normalisasi bobot seluruh atribut data menggunakan persamaan (4). Selanjutnya, menghitung bobot *eigen vector* dari setiap atribut berdasarkan persamaan (5). Selanjutnya menghitung nilai eigen maksimum berdasarkan persamaan (6). Kemudian menghitung nilai bobot *Eigen Vector* untuk setiap atribut berdasarkan normalisasi *min-max* dari nilai *Eigen Vector* menggunakan persamaan (7). Adapun hasil dari perolehan bobot atribut menggunakan *Eigen Vector* pada data set yang diujikan dalam hal ini yaitu *Wine Quality Dataset* dan *South German Credit Dataset* dapat dilihat pada Tabel 4 dan Tabel 5 berikut.

**Tabel 4.** Bobot *Eigen Vector* Pada *Wine Quality Dataset*

| No. | Atribut | Nilai *Eigen Vector* | Bobot *Eigen Vector* | Keterangan |
|---|---|---|---|---|
| 1 | *Fixed Accidity* (X1) | 0.01145 | 3.1 % | Terseleksi |
| 2 | *Volatile Accidity* (X2) | 0.03241 | 29 % | Terpilih |
| 3 | *Citric Acidity* (X3) | 0.03468 | 31.8 % | Terseleksi |
| 4 | *Residual Sugar* (X4) | 0.03178 | 28.2 % | Terseleksi |
| 5 | *Chlorides* (X5) | 0.04878 | 49.2 % | Terpilih |
| 6 | *Free Sulfur Dioxide* (X6) | 0.03376 | 30.7 % | Terpilih |
| 7 | *Total Sulfur Dioxide* (X7) | 0.03513 | 32.3 % | Terpilih |
| 8 | *Density* (X8) | 0.06524 | 69.5 % | Terpilih |
| 9 | *Ph* (X9) | 0.01171 | 3.5 % | Terseleksi |
| 10 | *Sulphates* (X10) | 0.00892 | 0 % | Terseleksi |
| 11 | *Alcohol* (X11) | 0.08998 | 100 % | Terpilih |

**Tabel 5.** Bobot *Eigen Vector* Pada *South German Credit Dataset*

| No. | Atribut | Nilai *Eigen Vector* | Bobot *Eigen Vector* | Keterangan |
|---|---|---|---|---|
| 1 | *Status (X1)* | 0,5721 | 87,31 % | Terseleksi |
| 2 | *Duration (X2)* | 0,3651 | 57,55 % | Terpilih |
| 3 | *Credit History (X3)* | 0,2569 | 41,98 % | Terseleksi |
| 4 | *Purpose (X4)* | 0,6604 | 100 % | Terseleksi |
| 5 | *Amount (X5)* | 0,0344 | 0,1 % | Terpilih |
| 6 | *Savings (X6)* | 0,5695 | 86,93 % | Terpilih |
| 7 | *Employment Duration (X7)* | 0,1664 | 28,98 % | Terpilih |
| 8 | *Installment Rate (X8)* | 0,4890 | 75,36 % | Terpilih |
| 9 | *Personal Status Sex (X9)* | 0,3637 | 57,35 % | Terseleksi |
| 10 | *Other Debtors (X10)* | 0,4598 | 71,15 % | Terseleksi |
| 11 | *Present Residence (X11)* | 0,4175 | 65,07 % | Terpilih |
| 12 | *Property (X12)* | 0,4196 | 65,37 % | Terseleksi |
| 13 | *Age (X13)* | 0,3638 | 57,35 % | Terseleksi |
| 14 | *Other Installment Plants (X14)* | 0,2718 | 44,14 % | Terseleksi |
| 15 | *Housing (X15)* | 0,5277 | 80,92 % | Terseleksi |
| 16 | *Number Credits (X16)* | 0,4114 | 64,21 % | Terseleksi |

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 1010−1017
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2022

| | | | | |
|---|---|---|---|---|
| 17 | *Job (X17)* | 0.03274 | 57,78 % | Terseleksi |
| 18 | *People Liable (X18)* | 0.00301 | 73,44 % | Terseleksi |
| 19 | *Telephone (X19)* | 0.03647 | 17,99 % | Terseleksi |
| 20 | *Foreign Worker (X20)* | 0.08208 | 57,67 % | Terseleksi |

Kemudian setelah bobot atribut dari data set diperoleh seperti pada Tabel 4 dan Tabel 5 sebelumnya, maka selanjutnya melakukan seleksi terhadap fitur yang memiliki bobot dengan pengaruh yang rendah atau bobot fitur terkecil. Adapun atribut yang terpilih pada Wine Quality Dataset yaitu sebanyak 3 atribut yaitu *Residual Sugar* (X4), *Free Sulfur Dioxode* (X6), dan *Ph* (X9) dan 8 atribut lainnya dinyatakan terseleksi karena memiliki nilai bobot pengaruh yang kurang signifikan. Kemudian pada South German Credit Dataset, atribut yang dipilih sebanyak 4 atribut yaitu *Status* (X1), *Purpose* (X4), *Savings* (X6), dan *Housing* (X15) dan 16 atribut lainnya dinyatakan terseleksi karena memiliki nilai bobot pengaruh yang kurang signifikan.

Kemudian Langkah selanjutnya yaitu melakukan *clustering* pada data set berdasarkan perhtitungan *K-Means Clustering*. Terlebih dahulu langkah yang dilakukan yaitu penentuan *centroid* dengan banyaknya jumlah *centroid* yang ditentukan berdasarkan banyassknya jumlah *class* pada data set. Pada *Wine Quality Dataset* mempunyai tujuh *class* yaitu *Class 3, Class 4, Class 5, Class 6, Class 7, Class 8 dan Class 9*. Maka dari itu untuk jumlah *centroid* yang ditentukan pada data set yaitu K = 7. Sedangkan pada South German Credit Dataset mempunyai dua kelas saja yaitu Good dan Bad sehingga nilai K = 2.

Kemudian, setelah diperoleh hasil perhitungan jarak dari *clustering* sampai iterasi terakhir, maka selanjutnya dilakukan perhitungan nilai kinerja dari jumlah iterasi yang dihasilkan dengan perhitungan *Sum of Square Error* (SSE) berdasarkan persamaan (9). Adapun hasil perhitungan *Sum of Square Error* (SSE) dari seluruh jumlah iterasi yang dihasilkan dapat dilihat pada Tabel 5 dan 6 berikut.

**Tabel 6.** Hasil Pengujian K-Means Tanpa Seleksi Fitur

| Dataset | Jumlah Iterasi | Nilai SSE |
|---|---|---|
| *Wine Quality Dataset* | 11 | 678.5735 |
| *South German Credit Dataset* | 10 | 1534.3167 |

**Tabel 6.** Hasil Pengujian K-Means dengan Seleksi Fitur Eigen Vector

| Dataset | Jumlah Iterasi | Nilai SSE |
|---|---|---|
| *Wine Quality Dataset* | 5 | 383.0517 |
| *South German Credit Dataset* | 4 | 469.0698 |

### 3.2 Pembahasan

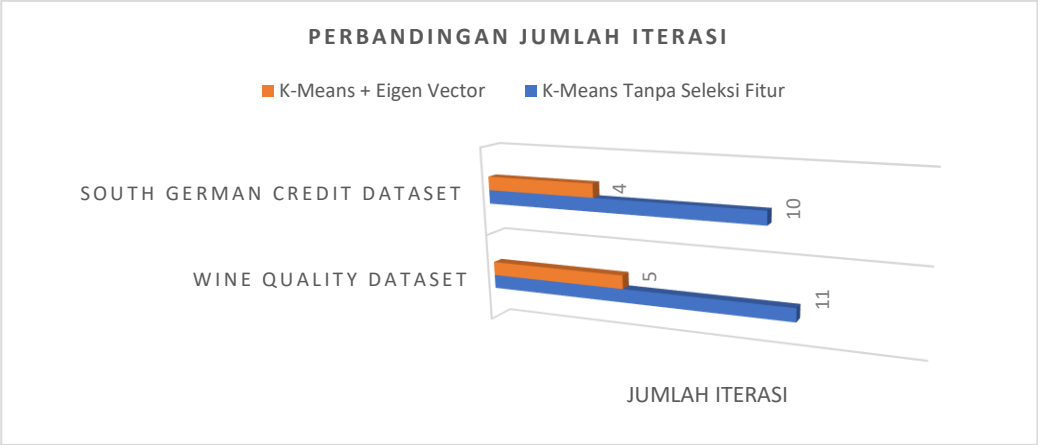
Pada bagian ini dilakukan pembahasan mengenai hasil perbandingan yang diperoleh dari metode yang diusulkan yaitu perhitungan K-Means Clustering tanpa seleksi fitur dan seleksi fitur menggunakan *Eigen Vector* pada *K-Means Clustering*, maka dilakukan perbandingan hasil evaluasi *clustering* dari data set yang diujikan. Perbandingan hasil evaluasi *clustering* dari metode yang diusulkan terhadap data set yang diujikan dapat dilihat pada Tabel 7 dan Tabel 8 berikut.

**Tabel 7.** Perbandingan Jumlah Iterasi Yang Dihasilkan

| No. | Data Set | Jumlah Iterasi Yang Dihasilkan | | |
|---|---|---|---|---|
| | | *K-Means Clustering* | *K-Means Clustering + Eigen Vector* | *Selisih Iterasi* |
| 1. | *Wine Quality Dataset* | 11 | 5 | 6 |
| 2. | *South German Credit Dataset* | 10 | 4 | 6 |

Pada Tabel 7 menunjukkan jumlah iterasi yang dibutuhkan untuk mencapai kondisi *convergen*. Pada data set yang dijukan, jumlah iterasi yang dibutuhkan untuk *K-Means Clustering* tanpa seleksi fitur pada *Wine Quality* yaitu sebanyak 11 iterasi dan pada *South German Credit* yaitu sebesar 10 iterasi, kemudian setelah menggunakan kombinasi dari *K-Means Clustering + Eigen Vector* yang diusulkan, jumlah iterasi yang dibutuhkan untuk mencapai konvergen pada *Wine Quality* yaitu menjadi 5 iterasi dan pada *South German Credit* menjadi 4 iterasi. Secara keseluruhan, jumlah iterasi yang diperoleh dari kedua metode tersebut dapat dilihat pada grafik di Gambar 2 berikut.

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 1010−1017
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2022



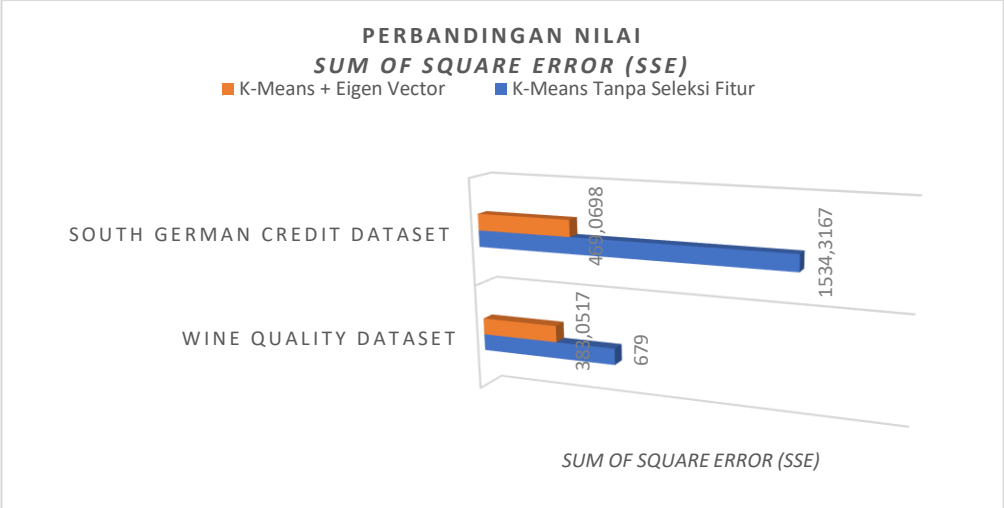

**Gambar 2.** Grafik Perbandingan Jumlah Iterasi Yang Dihasilkan

Kemudian, untuk perbandingan dari hasil evaluasi *clustering* berdasarkan *Sum of Square Error* (SSE) dari kedua metode terhadap data set yang digunakan dapat dilihat pada Tabel 8 berikut.

**Tabel 8.** Perbandingan Nilai *Sum of Square Error* (SSE) Yang Dihasilkan

| No. | Data Set | Nilai *Sum of Square Error* (SSE) | | |
|---|---|---|---|---|
| | | ***K-Means* Clustering Tanpa Seleksi Fitur** | ***K-Means Clustering +Eigen Vector*** | ***Selisih Error*** |
| 1. | *Wine Quality Dataset* | 678.5735 | 383.0517 | 1065.2469 |
| 2. | *South German Credit Dataset* | 1534.3167 | 469.0698 | 295.5218 |

Berdasarkan Tabel 8, diketahui bahwa pada data set yang diujikan, jumlah rata-rata nilai SSE yang diperoleh dari perhitungan K-Means Tanpa Seleksi Fitur pada *Wine Quality* yaitu sebesar 678.5735 dan *pada South German Credit* yaitu sebesar 1534.3167. Kemudian setelah mengujikan dengan kombinasi *K-Means Clustering + Eigen Vector*, maka diperoleh hasil SSE pada *Wine Quality* menjadi 383.0517 dan pada *South German Credit* menjadi 469.0698. Secara keseluruhan, rata-rata nilai *Sum of Square Error* (SSE) yang diperoleh dari kedua metode tersebut dapat dilihat pada grafik di Gambar 3 berikut.




**Gambar 3.** Grafik Perbandingan Nilai *Sum of Square Error* (SSE) Yang Dihasilkan

## 4. KESIMPULAN

Berdasarkan hasil pengujian yang dilakukan, *K-Means Clustering* dengan seleksi fitur menggunakan *Eigen Vector* dapat meningkatkan kinerja pada *K-Means Clustering*. Dengan kombinasi seleksi fitur *Eigen Vector* terbukti dapat meminimalkan jumlah iterasi pada *K-Means Clustering* yang dimana pada data set yang diujikan memperoleh jumlah iterasi pada *K-Means Clustering* tanpa seleksi fitur pada *Wine Quality Dataset* yaitu sebanyak 11 iterasi dan pada *South German Credit Dataset* yaitu sebanyak 10 iterasi, kemudian setelah diujikan dengan *Eigen Vector* memperoleh jumlah iterasi pada *Wine Quality Dataset* menjadi 5 iterasi dan pada *South German Credit Dataset*

**Building of Informatics, Technology and Science (BITS)**
Volume 4, No 2, September 2022 Page: 1010−1017
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2022

menjadi 4 iterasi. Pada data set yang diujikan, perolehan jumlah nilai *Sum of Square Error* (SSE) yang diperoleh dari proses K-Means tanpa seleksi fitur pada *Wine Quality* yaitu sebesar 678.5735 dan pada *South German Credit* yaitu 1534.3167, sedangkan dengan menggunakan *K-Means Clustering + Eigen Vector* yang diusulkan, nilai *Sum of Square Error* (SSE) yang diperoleh pada *Wine Quality* yaitu 383.0517 dan pada *South German Credit* yaitu 469.0698. Maka berdasarkan dengan hasil pengujian penelitian ini, seleksi fitur untuk mengurangi ataupun membuang fitur yang memiliki pengaruh rendah terhadap dataset dapat meningkatkan kinerja untuk meminimalkan jumlah error dari proses *clustering* pada *K-Means Clustering* dengan peningkatan kinerja yang cukup maksimal.

# REFERENCES

[1] N. Arunkumar, M. A. Mohammed, M. K.A Ghani, D. A. Ibrahim, "K-means clustering and neural network for osbject detecting and identifying abnormality of brain tumor". Soft Computing, vol. 23, no. 19, pp. 9083-9096, 2019.
[2] U. R. Raval, C. Jani, "Implementing & Improvisation of K-means Clustering Algorithm", IJCSMC, vol. 5, no. 5, 2016
[3] M. Bora, D. Jyoti, D. Gupta, A. Kumar, "Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab", IJCBIT, vol. 5, no. 2, 2014.
[4] M. Kuhkan, "A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm," *International Journal of Computer Engineering and Information Technology*, vol. 8, no. 6, pp. 90-95, 2016.
[5] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, "Reduksi atribut menggunakan information gain untuk optimasi cluster algoritma k-means," *Jurnal Edukasi dan Penelitian Informatika*, vol. 6, no. 1, pp. 48-53. 2020.
[6] A. Izzuddin, "Optimasi Cluster pada Algoritma K-Means dengan Reduksi Dimensi Dataset Menggunakan Principal Component Analysis untuk Pemetaan Kinerja Dosen," *Energy-Jurnal Ilmiah Ilmu-Ilmu Teknik*, vol. 5, no. 2, pp.41-46, 2015.
[7] T. Silwattananusarn, K. Tuamsuk, "Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012", IJDKP, vol. 2, no. 5, 2012.
[8] Z. Ren, Z. Xu, and H. Wang, "The strategy selection problem on artificial intelligence with an integrated VIKOR and AHP method under probabilistic dual hesitant fuzzy information," *IEEE Access*, vol. *7*, pp. 103979-103999, 2019.
[9] C. Saranya, and G. Manikandan, "A Study on Normalization Techniques for Privacy Preserving Data Mining," *International Journal of Engineering and Technology (IJET)*, vol. 5, no. 3, pp. 2701-2704, 2013.
[10] L. P. Refialy, H. Maitimu, and M. S. Pesulima, "Perbaikan Kinerja Clustering K-Means pada Data Ekonomi Nelayan dengan Perhitungan Sum of Square Error (SSE) dan Optimasi nilai K cluster," *Techno. Com*, vol. 20, no. 2, pp. 321-329, 2021.
[11] A. I. Lubis, U. Erdiansyah, and R. Siregar, "Comparison of Accuracy in Naïve Bayes and Random Forests in Classification of Liver Disease," *CESS (Journal of Computer Engineering, System and Science)*, vol. 7, no. 1, pp. 81-89, 2022.
[12] A.E. Munthafa, and H. Mubarok, "Penerapan Metode Analytical Hierarchy Process Dalam Sistem Pendukung Keputusan Penentuan Mahasiswa Berprestasi," *Jurnal Siliwangi*, vol.3, no.2, 2017.
[13] O. J. Oyelade, O. O. Oladipupo, I. C. Obagbuwa, "Application of K-Means Clustering Algorithm for Prediction of Students's Academic Performance", IJCSIS, Vol 7, No 1, 2010
[14] W. Wijayanti, R. Ayu, M. T. Furqon, and S. Adinugroho. "Penerapan Algoritme Support Vector Machine Terhadap Klasifikasi Tingkat Risiko Pasien Gagal Ginjal." Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN 2548 (2018): 964X.
[15] G. Tian, H. Zhang, M. Zhou, and Z. Li, "AHP, Gray Correlation, and TOPSIS Combined Approach to Green Performance Evaluation of Design Alternatives," IEEE Transaction on Systems, MAN, and Cybernetics, pp. 1-13, 2007
[16] H. Haviluddin, S. J. Patandianan, G. M. Putra, N. Puspitasari, and H. S. Pakpahan, "Implementasi Metode K-Means Untuk Pengelompokkan Rekomendasi Tugas Akhir," *Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer*, vol. 16, no. 1, pp. 13-18, 2021.s
[17] R. Nainggolan, and G. Lumbantoruan, "Optimasi performa cluster K-Means menggunakan Sum of Squared Error (SSE)," *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, vol. 2, no. 2, pp. 103-108, 2018.