Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1309−1316
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2458

Optimasi Cluster Pada K-Means Clustering Dengan Teknik Reduksi Dimensi Dataset Menggunakan Gini Index

Muhammad Imam Zarkasyi[1], Herman Mawengkang[2,*], Opim Salim Sitompul[3]

[1,3]Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Medan, Indonesia
[2]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sumatera Utara, Medan, Indonesia
Email: [1]m.imamzarkasyi96@gmail.com, [2*] hmawengkang@yahoo.com, [3]opim@usu.ac.id
Email Penulis Korespondensi: hmawengkang@yahoo.com
Submitted: **01/11/2022**; Accepted: **05/12/2022**; Published**: 30/12/2022**

Abstrak−Pada K-Means Clustering, jumlah atribut dari sebuah data dapat mempengaruhi jumlah iterasi yang dihasilkan pada proses pengelompokan data. Solusi untuk mengatasi permasalahan tersebut salah satunya dengan teknik reduksi pada dimensi dataset. Pada penelitian ini, penulis bertujuan untuk menerapkan Gini Index dalam melakukan reduksi atribut pada data set untuk mengurangi atribut yang tidak berpengaruh pada dataset sebelum dilakukan clustering dengan K-Means Clustering. Dataset yang digunakan untuk diujikan sebagai instrumen pengujian pada riset ini adalah Absenteeism at work yang diperoleh dari UCI Machine Learning Repository, dengan 20 atribut, 740 records data dan 4 kelas atribut. Hasil dari pengujian pada riset ini menunjukkan bahwa jumlah iterasi yang diperoleh dari perbandingan pengujian dengan menggunakan K-Means secara Konversional (Tanpa Reduksi Atribut) yaitu diperoleh jumlah 9 iterasi, sedangkan K-Means dengan reduksi atribut dengan Gini Index diperoleh jumlah iterasi jumlah 6 iterasi. Evaluasi clustering dihitung menggunakan Sum of Square Error (SSE). Nilai SSE pada K-Means Clustering secara Konversional (Tanpa Reduksi Atribut) yaitu sebesar 1391.613, sedangkan pada K-Means Clustering dengan reduksi atribut dengan Gini Index yaitu sebesar 440.912. Dari hasil metode yang diusulkan mampu dalam menurunkan persentase error serta meminimalkan jumlah iterasi pada K-Means Clustering dengan reduksi dimensi dataset menggunakan Gini Index.

Kata Kunci: Reduksi Dimensi; Clustering; K-Means Clustering; Gini Index; Sum of Square Error

Abstract−In K-Means Clustering, the number of attributes of a data can affect the number of iterations generated in the data grouping process. One of the solutions to overcome these problems is by using a reduction technique on the dimensions of the dataset. In this study, the authors apply the Gini Index to perform attribute reduction on the data set to reduce attributes that have no effect on the dataset before clustering with K-Means Clustering. The dataset used to be tested as a testing instrument in this research is Absenteeism at work obtained from the UCI Machine Learning Repository, with 20 attributes, 740 data records and 4 attribute classes. The results of the tests in this research indicate that the number of iterations obtained from the comparison of tests using the K-Means in a Conversional (Without Attribute Reduction) is obtained by the number of 9 iterations, while the K-Means with attribute reduction with the Gini Index obtains the number of iterations totaling 6 iterations. Clustering evaluation was calculated using Sum of Square Error (SSE). The SSE value in K-Means Clustering in a Conversional (Without Attribute Reduction) is 1391.613, while in K-Means Clustering with attribute reduction with a Gini Index, it is 440.912. From the results of the proposed method, it is able to reduce the percentage of errors and minimize the number of iterations in K-Means Clustering by reducing the dimensions of the dataset using the Gini Index.

Keywords: Dimensional Reduction; Clustering; K-Means Clustering; Gini Index; Sum of Square Error

1. PENDAHULUAN

Pengelompokan data (*Clustering)* adalah bagian dari kelompok metode pembelajaran tanpa pengawasan. *Clustering* dapat mengelompokkan data menjadi beberapa cluster atau kelompok berdasarkan kesamaan [1]. Ada 6 persyaratan, yaitu skalabilitas, kemampuan menganalisis berbagai bentuk data, menemukan cluster yang bentuknya tidak terduga, kemampuan menangani noise, kepekaan terhadap variasi input, kemampuan melakukan analisis berdimensi tinggi saat menerapkan metode clustering. kegunaan.[2].

K-Means clustering merupakan salah satu metode yang populer diterapkan untuk clustering di berbagai bidang. K-Means clustering memiliki banyak keunggulan, antara lain kemudahan implementasi dan pengoperasian, waktu proses yang relatif cepat, adaptasi yang lebih mudah dengan metode lain, dan sangat umum digunakan oleh para peneliti [3]. Namun K-Means Clustering tentu memiliki kekurangan, misalnya nilai centroid awal inisialisasi yang acak sehingga sangat mempengaruhi hasil pengelompokan sehingga kurang optimal, penentuan nilai k masih menggunakan metode trial and error, dan data outlier [4].

Kemudian pada K-Means juga dihadapkan pada persoalan *curse of dimensionality* dalam persoalan data dengan fitur yang besar ataupun dimensi tinggi dengan persoalan berupa menurunnya nilai performa dari kualitas cluster dan juga berpengaruh pada waktu komputasi [5]. Permasalahan ini dapat diatasi dengan melakukan reduksi dimensi [6].

Banyak keuntungan yang dapat diperoleh dengan mengurangi dimensi dataset. Banyak algoritma data mining bekerja dengan baik jika dimensi (jumlah atribut/fitur dalam data) rendah. Alasannya adalah pengurangan dimensi dapat menghilangkan fitur yang tidak relevan, mengurangi kebisingan, dan juga mengurangi *curse of dimensionality* [7]. Banyak metode yang bisa digunakan untuk reduksi dimensi pada dataset. Salah satunya adalah metode *Gini Index*. *Gini Index* dapat mereduksi dimensi data yang tinggi menjadi dimensi data yang lebih rendah dengan resiko kehilangan informasi yang sangat kecil [8].

Beberapa penelitian terdahulu yang terkait pada penelitian ini salah satunya yaitu pada penelitian dari [2] meneliti tentang kombinasi *K-Means* dengan *Decision Tree* dengan hasil penelitian tersebut yaitu menemukan kriteria mana yang menghasilkan pohon keputusan dan performa terbaik berdasarkan nilai akurasi tertinggi dari masing-

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1309−1316
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2458

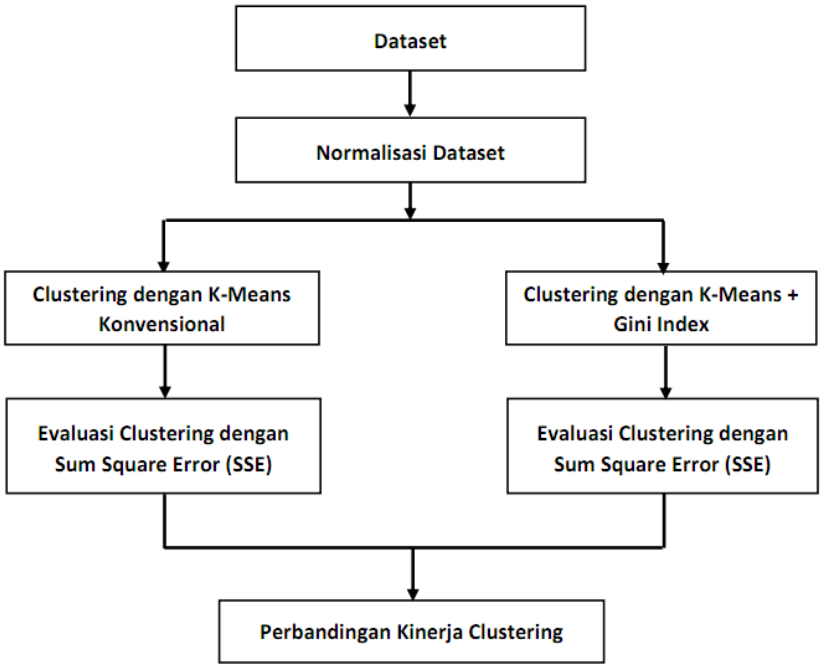
masing kelompok data. Penelitian dari [9] meneliti tentang peningkatan kinerja K-Means menggunakan *Rank Order Centroid* (ROC) dan *Braycurtis Distance* dengan hasil yang diperoleh yaitu dengan menggunakan pengujian beberapa dataset memperoleh peningkatan kinerja setelah *K-Means* dikombinasikan dengan ROC dan *Braycurtis Distance* dibandingkan dengan K-Means Konvensional. Penelitian dari [10] yang melakukan kombinasi *K-Nearest Neighbor* dengan *Gini Index* untuk klasifikasi tingkat kognitif soal pada *Taksonomi Bloom* dengan hasil yang diperoleh yaitu *Gini Index* mampu mengurangi dimensi fitur yang tinggi, sehingga meningkatkan kinerja KNN dan meningkatkan tingkat akurasi klasifikasi tingkat kognitif soal pada Taksonomi Bloom dari akurasi 59.97 % menjadi 68.37 %. Penelitian dari [11] melakukan reduksi atribut dengan *Gini Index* pada *K-Nearest Neighbor* pada klasifikasi kinerja siswa dengan hasil yang diperoleh yaitu peningkatan akurasi K-NN setelah reduksi atribut dari 74.068 % menjadi 76.516 %. Penelitian dari [12] melakukan kombinasi *K-Means* dengan *Bee Colony Optimization* (BCO) pada pengelompokan data dengan hasil yang diperoleh yaitu peningkatan performa K-Means setelah dikombinasikan dengan BCO yaitu dari 83.09 % menjadi 83.30%.

Berdasarkan dari beberapa penelitian terkait yang dipaparkan sebelumnya, maka pada penelitian ini penulis mengusulkan metode *Gini Index* untuk mereduksi dimensi dataset dengan tujuan untuk mengoptimasi kualitas kinerja cluster 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

Pada Gambar 1 di atas menggambarkan framework pada penelitian ini secara bertahap. Adapun penjelasan pada Gambar 1 secara detail dijelaskan sebagai berikut:

a. Dataset
Dataset yang dijadikan sebagai instrument pengujian pada penelitian ini yaitu dataset faktor yang mempengaruhi ketidakhadiran karyawan yang diambil dari *UCI Machine Learning Repository* yang berasal dari Universidade Nove de Julho' Program Pascasarjana di Informatika dan Manajemen Pengetahuan.

b. Normalisasi Data

Normalisasi data berfungsi untuk mempersiapkan data yang benar-benar valid sebelum diproses pada tahap berikutnya [13]. Pada penelitian ini, normalisasi data dilakukan menggunakan metode *Min-Max* dengan rumus berikut [14].

$$\frac{(Data-Min)*(NewMax-NewMin)}{(Max-Min)} + NewMin \quad (1)$$

c. K-Means Clustering

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1309−1316
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2458

K-Means Clustering merupakan metode umum dan paling sederhana dalam *clustering* [15]. *K-Means* digunakan dalam mengelompokkkan data menjadi beberapa kelompok tanpa mengetahui target kelasnya [16]. Hasil proses *cluster* dipengaruhi oleh pada nilai *centroid* awal. Proses *K-Means* sebagai berikut [17]:

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

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

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

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

d. Gini Index

Gini Index merupakan probabilitas dari dua data yang dipilih secara acak yang memiliki *class* yang berbeda. *Gini Index* digunakan oleh Breiman pada tahun 2001 untuk menghasilkan pohon klasifikasi pada *decision tree* [10]. Misalkan S adalah 1 set dari sejumlah *s* data. Data ini memiliki sejumlah *m class* yang berbeda (*Ci*, i= 1, ..., *m*). Berdasarkan pada *class* tersebut, kita dapat membagi *S* ke dalam sejumlah *m subset* (*Si*, i= 1, ..., *m*) misalkan *Si* adalah dataset yang tergabung di dalam *class Ci*, *si* adalah jumlah data dari *Si*, maka *Gini Index* dapat dirumuskan sebagai berikut:

$$Gini\ Index\ (S) = 1 - \sum_{i=1}^{m}\left(\frac{S_i}{S}\right)^2 \quad (4)$$

2.2 Pengukuran Kinerja Clustering

Dalam mengukur kinerja hasil pengujian *clustering* pada metode yang diterapkan pada penelitian ini, menggunakan perhitungan dari *Sum of Square Error* (SSE). SSE merupakan cara dalam melakukan validasi *cluster* melalui jumlah kuadrat setiap anggota *cluster* menuju pusatnya. Semakin banyak jumlah *k* yang diujikan, maka nilai SSE. Kemudian semakin kecil Semakin jauh jarak yang membentuk titik siku, maka jumlah *cluster* tersebut menjadi yang paling optimal [18]. Rumus SSE adalah sebagai berikut [19]:

$$SSE = \sum_{k=1}^{K}(x_i - C_k)^2 \quad (5)$$

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

Dalam memperoleh informasi dari hasil pada penelitian ini, dilakukan tahapan pengujian yang sebelumnya telah dipaparkan pada bagian sebelumnya. Tahapan pengujian yang dilakukan yaitu menormalisasi dataset terlebih dahulu sebelum pengujian *clustering*, pengujian teknik reduksi dimensi dataset yang diujikan, pengujian clustering pada dataset, dan menganalisis hasil clustering dari pengujian dataset.

3.1.1 Tahapan Normalisasi Dataset

Adapun pada penelitian ini data yang digunakan sebagai instrument pengujian dari metode yang diusulkan yaitu menggunakan dataset faktor yang mempengaruhi ketidakhadiran karyawan yang diambil dari *UCI Machine Learning Repository* pada penelitian akademik di Universidade Nove de Julho' Program Pascasarjana di Informatika dan Manajemen Pengetahuan. Data tersebut telah digunakan pada penelitian – penelitian sebelum dalam memprediksi ketidakhadiran karyawan. Pada dataset memiliki 740 data dan 21 attribut dengan target kelas dalam satuan jam. Adapun informasi dari atribut dataset tersebut yaitu pada Tabel 1 berikut:

Tabel 1. Atribut *Absenteeism at Work Dataset*

No.	Atribut	Range Nilai	Keterangan
1	ID	1 – 36	Individual Identification
2	Reason for Absence	1 – 28	Terdapat 21 kategori penyakit dan 7 kategori tanpa
3	Month of Absence	Bulan 1 – 12	CID (Code of Diseases)
4	Day of the Week	1–5 Hari	Setahun terdapat 12 bulan
5	Seasons	1–4 Musim	Senin, selasa, rabu, kamis, dan jumat
6	Transportation Expense	118 s.d 388	Terdapat 4 musim yaitu; musim panas, musim semi, musim gugur, dan musim dingin

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1309−1316
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2458

7	Distance from Residence to Work	5–52 Km	Biaya transportasi Satuan kilometer
8	Service Time	1–29 Menit	Waktu pelayanan
9	Age	27 s.d 58 Tahun	Umur karyawan
10	Work Load Average/Day	205.917 s.d 378.884	Rata – rata waktu kerja (Menit)
11	Hit Target	81 s.d 100 Target	Pencapaian target
12	Disciplinary Failure	0 dan 1	Disiplin kerja
13	Education	1 s.d 4	High school, graduate, postgraduate, master and
14	Son	0 s.d 4 Anak	Doctor
15	Social Drinker	0 dan 1	Jumlah anak
16	Social Smoker	0 dan 1	Peminum
17	Pet	0 s.d 8 Hewan	Perokok
18	Weight	56 s.d 108 kg	Hewan peliharaan
19	Height	163 s.d 196 Cm	Berat badan karyawan
20	Body Mass Index	19 s.d 38 kg/m2	Tinggi badan karyawan
21	Absenteeism Time in Hours (Target)	0 s,d 120 jam	Index masa tubuh karyawan Klasifikasi karyawan

Normalisasi dataset dilakukan dengan perhitungan normalisasi *min-max* dengan range nilai *records* data diantara 0 sampai dengan 1. Perhitungan normalisasi *min-max* pada dataset yang digunakan pada penelitian ini berdasarkan rumus persamaan (1). Adapun hasil normalisasi dataset dengan perhitungan normalisasi min-max pada Tabel 1 berikut:

Tabel 2. Hasil Normalisasi Dataset

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	… …	X19	X20
1	0,928	0,583	0,25	0	0,633	0,659	0,428	0,193	0,194	0,842	0	0	0,5	… …	0,272	0,578
2	0	0,583	0,25	0	0	0,170	0,607	0,741	0,194	0,842	1	0	0,25	… …	0,454	0,631
3	0,821	0,583	0,5	0	0,225	0,978	0,607	0,354	0,194	0,842	0	0	0	… …	0,212	0,631
4	0	0,583	0,75	0	0,596	0	0,464	0,387	0,194	0,842	0	0	0,5	… …	0,151	0,263
5	0,25	0,583	0,75	0	0,633	0,659	0,428	0,193	0,194	0,842	0	0	0,5	… …	0,272	0,578
6	0,821	0,583	1	0	0,225	0,978	0,607	0,354	0,194	0,842	0	0	0	… …	0,212	0,631
7	0,821	0,583	1	0	0,9	1	0,071	0,032	0,194	0,842	0	0	0,25	… …	0,272	0,421
8	0,785	0,583	1	0	0,525	0,957	0,357	0,290	0,194	0,842	0	0	1	… …	0,151	0,210
9	0,821	0,583	0	0	0,137	0,148	0,464	0,225	0,194	0,842	0	0	0,5	… …	1	0,315
10	0,678	0,583	0	0	0,433	0,127	0,464	0,322	0,194	0,842	0	0,666	0,25	… …	0,272	0,526
11	0,785	0,583	0	0	0,525	0,957	0,357	0,290	0,194	0,842	0	0	1	… …	0,151	0,210
12	0,035	0,583	0,25	0	0,525	0,957	0,357	0,290	0,194	0,842	0	0	1	… …	0,151	0,210
13	0,035	0,583	0,5	0	0,525	0,957	0,357	0,290	0,194	0,842	0	0	1	… …	0,151	0,210
14	0,392	0,583	0,5	0	0,225	0,978	0,607	0,354	0,194	0,842	0	0	0	… …	0,212	0,631
15	0,392	0,583	0,5	0	0,225	0,978	0,607	0,354	0,194	0,842	0	0	0	… …	0,212	0,631
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
736	0,5	0,583	0,25	0	0,633	0,659	0,428	0,193	0,339	0,631	0	0	0,5	… …	0,272	0,578
737	0,392	0,583	0,25	0	0,433	0,127	0,464	0,322	0,339	0,631	0	0,666	0,25	… …	0,272	0,526
738	0	0	0,25	0	0	0,191	0,428	0,419	0,377	0,736	0	0	0,25	… …	0,212	0,789
739	0	0	0,5	0,333	0,418	0,638	0,464	0,387	0,377	0,736	0	0	0,5	… …	0,212	0,842
740	0	0	1	0,666	0,225	0,851	0,464	0,838	0,377	0,736	0	0	0,25	… …	0,363	0,315

3.1.2 Reduksi Dimensi Dataset

Pada tahap ini dilakukan reduksi dimensi dataset menggunakan *Gini Index*. Penelitian ini mereduksi dimensi dataset dari atribut yang mempunyai persentase rendah, sebab dipastikan bahwa atribut yang mempunyai persentase rendah maka tentu memiliki pengaruh yang rendah pada dataset. Dalam menghitung persentase pengaruh atribut pada dataset, pada penelitian ini menggunakan *RapidMiner Studio* untuk mempersingkat proses perhitungan nilai persentase atribut pada dataset. Hasil reduksi dimensi dataset ditunjukkan pada Tabel 3.

Tabel 3. Reduksi Dataset dengan Gini Index

No	Atribut	Nilai Gini Index	Bobot Gini Index (%)	Keterangan
X01	ID	-	-	-
X02	Reason for Absence	0.10003	100	-
X03	Month of Absence	0.00660	5.88	-
X04	Day of the Week	0.00181	1.06	-
X05	Seasons	0.00143	0.67	Reduksi
X06	Transportation Expense	0.00254	1.79	-
X07	Distance from Residence to Work	0.00314	2.39	-
X08	Service Time	0.00256	1.82	-

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1309−1316
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2458

X09	Age	0.00306	2.32	-
X10	Work Load Average/Day	0.00353	2.79	-
X11	Hit Target	0.00372	2.98	-
X12	Disciplinary Failure	0.09265	92.56	-
X13	Education	0.00075	0	Reduksi
X14	Son	0.00184	1.09	-
X15	Social Drinker	0.00239	1.64	-
X16	Social Smoker	0.00111	0.36	Reduksi
X17	Pet	0.00109	0.34	Reduksi
X18	Weight	0.00547	4.75	-
X19	Height	0.00317	2.43	-
X20	Body Mass Index	0.00317	3.78	-
X21	Absenteeism Time in Hours (Target)	-	-	-

Setelah bobot atribut dari data set diperoleh seperti pada Tabel 3 sebelumnya, maka selanjutnya melakukan reduksi terhadap atribut yang memiliki bobot persentase dengan pengaruh yang rendah atau bobot. Adapun atribut yang direduksi pada *Absenteeism at Work Dataset* yaitu sebanyak 4 atribut yaitu Seasons (X05), Education (X13), Social Smoker (X16), dan Pet (X17) dinyatakan direduksi karena memiliki persentase pengaruh yang kurang signifikan karena mempunyai persentase dibawah 1 % sehingga dapat dipastikan bahwa ke empat atribut tersebut memiliki pengaruh yang kurang signifikan terhadap *Absenteeism at Work Dataset*. Selain dari empat atribut tersebut, ada 15 atribut yang terpilih untuk dijadikan sebagai atribut dari *Absenteeism at Work Dataset* yang akan digunakan pada proses *clustering* dikarenakan ke lima belas atribut tersebut mempunyai persentase di atas 1 % sehingga dipastikan atribut-atribut tersebut mempunyai pengaruh yang cukup signifikan terhadap *Absenteeism at Work Dataset*

Kemudian setelah dilakukan reduksi dimensi dataset menggunakan *Gini Index* selesai dilakukan, maka tahapan berikutnya adalah tahap *clustering*.

3.1.3 Pengujian dengan K-Means Clustering Tanpa Reduksi Dimensi Dataset

Selanjutnya yaitu langkah dalam melakukan *clustering* pada data set dengan *K-Means Clustering* tanpa reduksi atribut (Konvensional). Pengujian dengan *K-Means Clustering* menggunakan nilai *k* yang beragam dimulai dari nilai k=2 s,d k=10. Pengujian clusterin diukur berdasarkan waktu proses (*Time*) per nilai *k*, kemudian menghitung jumlah iterasi yang dihasilkan dari masing-masing pengujian nilai *k*. Kemudian melakukan perhitungan jarak dari *clustering* sampai iterasi terakhir dari masing-masing pengujian nilai *k*, maka selanjutnya dilakukan perhitungan nilai kinerja dari jumlah iterasi yang dihasilkan dengan perhitungan *Sum of Square Error* (SSE) berdasarkan persamaan (5) dari pengujian dengan nilai k dari 2 sampai dengan 10. Hasil perolehan iterasi dan perhitungan *Sum of Square Error* (SSE) dari seluruh pengujian nilai *k* yang dihasilkan dapat dilihat pada Tabel 4.

Tabel 4. Hasil Pengujian dengan K-Means Clustering Tanpa Reduksi Atribut

K	Time (s)	Iteration	SSE
2	0.02	6	1684.654
3	0.02	5	1625.761
4	0.05	14	1505.564
5	0.03	9	1441.361
6	0.04	9	1376.944
7	0.03	10	1301.335
8	0.02	8	1224.783
9	0.09	9	1188.716
10	0.06	7	1175.395
Mean	**0.04**	**9**	**1391.613**

Pada Tabel 4 dapat disimpulkan untuk hasil pengujian clustering dengan K-Means Clustering tanpa reduksi atribut diperoleh iterasi dan nilai SSE yang beragam dari pengujian nilai k=2 sampai dengan k=10. Jumlah iterasi tertinggi yaitu pada saat nilai k=4 dengan jumlah 14 iterasi dan terendah yaitu pada saat nilai k=2 yaitu 5 iterasi. Adapun nilai SSE terbaik dari seluruh pengujian nilai k yaitu pada saat pengujian nilai k=10 yaitu sebesar 1175.395. Kemudian rata-rata yang diperoleh dari iterasi pada seluruh pengujian nilai *k* yaitu 9 iterasi dan rata-rata yang diperoleh dari SSE pada seluruh pengujian nilai *k* yaitu 1391.613.

3.1.4 Pengujian dengan K-Means Clustering + Gini Index

Selanjutnya yaitu langkah dalam melakukan *clustering* pada data set dengan *K-Means Clustering* dengan reduksi atribut menggunakan *Gini Index*. Pengujian dengan *K-Means Clustering* menggunakan nilai *k* yang beragam dimulai dari nilai k=2 s,d k=10. Pengujian clustering diukur berdasarkan waktu proses (*Time*) per nilai *k*, kemudian menghitung jumlah iterasi yang dihasilkan dari masing-masing pengujian nilai *k*. Kemudian melakukan perhitungan jarak dari *clustering* sampai iterasi terakhir dari masing-masing pengujian nilai *k*, maka selanjutnya dilakukan

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1309−1316
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2458

perhitungan nilai kinerja dari jumlah iterasi yang dihasilkan dengan perhitungan *Sum of Square Error* (SSE) berdasarkan persamaan (5) dari pengujian dengan nilai k dari 2 sampai dengan 10. Hasil perolehan iterasi dan perhitungan *Sum of Square Error* (SSE) dari seluruh pengujian nilai *k* yang dihasilkan dapat dilihat pada Tabel 5.

Tabel 5. Hasil Pengujian dengan K-Means Clustering dan Gini Index

K	Time (s)	Iteration	SSE
2	0.04	4	533.461
3	0.03	3	492.344
4	0.06	8	468.015
5	0.08	6	451.036
6	0.05	7	442.351
7	0.03	7	431.810
8	0.03	5	396.048
9	0.03	6	381.709
10	0.07	6	371.432
Mean	**0.04**	**6**	**440.912**

Pada Tabel 4 dapat disimpulkan untuk hasil pengujian *clustering* dengan K-Means Clustering tanpa reduksi atribut diperoleh iterasi dan nilai SSE yang beragam dari pengujian nilai k=2 sampai dengan k=10. Jumlah iterasi tertinggi yaitu pada saat nilai k=4 dengan jumlah 8 iterasi dan terendah yaitu pada saat nilai k=3 yaitu 3 iterasi. Adapun nilai SSE terbaik dari seluruh pengujian nilai k yaitu pada saat pengujian nilai k=10 yaitu sebesar 371.432. Kemudian rata-rata yang diperoleh dari iterasi pada seluruh pengujian nilai *k* yaitu 6 iterasi dan rata-rata yang diperoleh dari SSE pada seluruh pengujian nilai *k* yaitu 440.912.

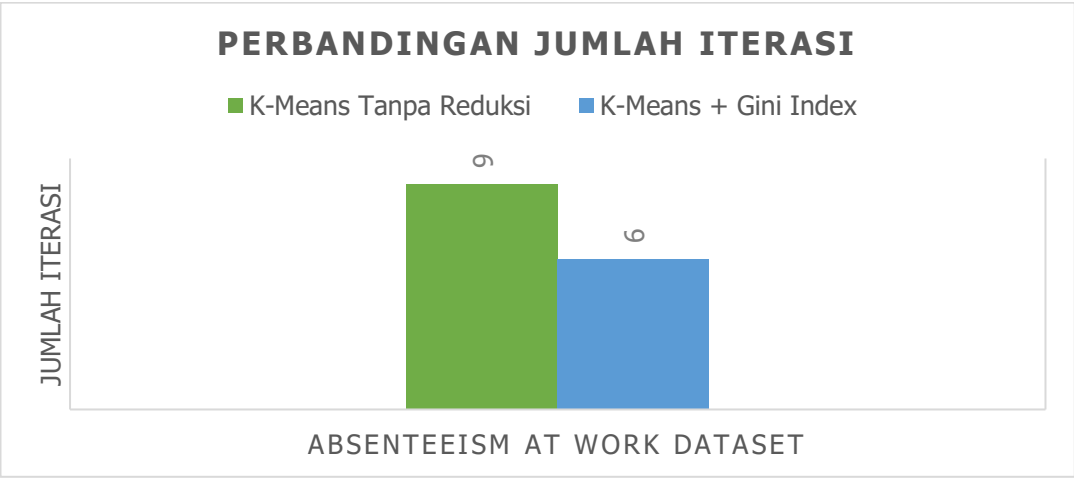
3.2 Pembahasan

Pada bagian ini dilakukan pembahasan mengenai hasil perbandingan yang diperoleh dari metode yang diusulkan yaitu perhitungan K-Means tanpa reduksi atribut dan *K-Means + Gini Index*, 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 6 dan Tabel 7.

Tabel 6. Hasil Perbandingan Jumlah Iterasi

Metode	Jumlah Rata-Rata Iterasi Yang Dihasilkan
K-Means Tanpa Reduksi	9
K-Means + Gini Index	6

Pada Tabel 6 menunjukkan jumlah rata-rata iterasi dari kedua metode yang telah diujikan. Pada data set yang dijukan, rata-rata jumlah iterasi yang dibutuhkan untuk *K-Means* tanpa reduksi dengan sebanyak 9 iterasi, kemudian setelah menggunakan kombinasi dari *K-Means + Gini Index*, rata-rata jumlah iterasi yang dibutuhkan untuk mencapai konvergen yaitu menjadi 6 iterasi. Secara keseluruhan, jumlah iterasi yang diperoleh dari kedua metode tersebut dapat dilihat pada grafik di Gambar 2 berikut.



Gambar 2. Grafik Perbandingan Jumlah Iterasi

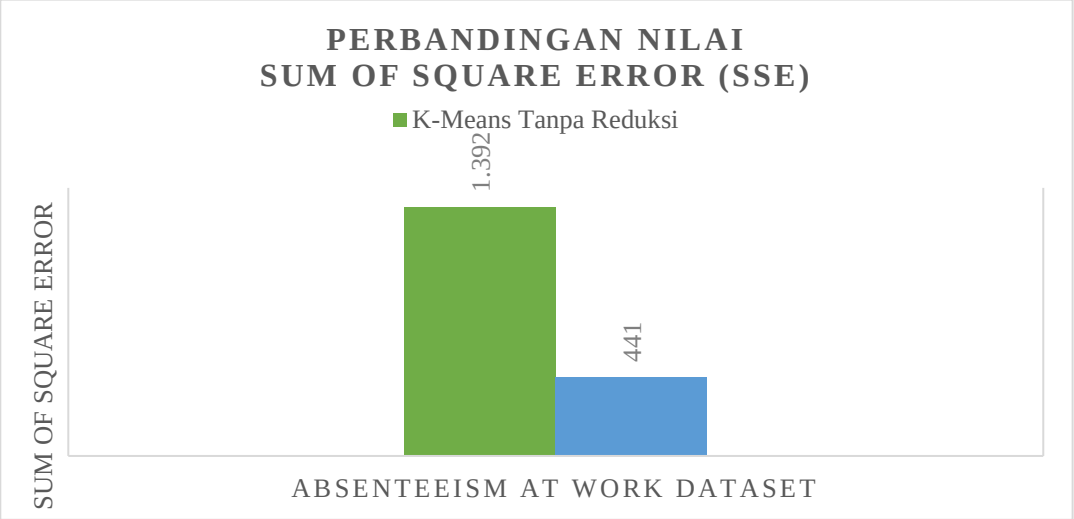
Kemudian, untuk perbandingan dari hasil evaluasi *clustering* berdasarkan *Sum of Square Error* (SSE) dari kedua metode terhadap data set yang digunakan, dapat dilihat pada Tabel 7.

Tabel 7. Hasil Perbandingan *Sum of Square Error* (SSE)

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1309−1316
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2458

Metode	Jumlah Rata-Rata Nilai *Sum of Square Error* (SSE)
K-Means Tanpa Reduksi	1391.613
K-Means + Gini Index	440.912

Berdasarkan Tabel 7, diketahui bahwa pada data set yang diujikan, jumlah rata-rata nilai SSE yang diperoleh dari perhitungan K-Means tanpa reduksi yaitu sebesar 1391.613. Kemudian setelah mengujikan dengan *K-Means Clustering + Gini Index*, maka diperoleh hasil SSE menjadi 440.912. Secara keseluruhan, rata-rata nilai *Sum of Square Error* yang diperoleh dari kedua metode tersebut dapat dilihat pada grafik di Gambar 3.



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

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, maka dapat disimpulkan bahwa *K-Means + Gini Index* lebih unggul dibandingkan dengan *K-Means* Konvensional pada dataset yang diujikan. Hal tersebut dibuktikan dengan hasil iterasi yang diperoleh sampai dalam kondisi *convergen* pada pengujian dengan *K-Means + Gini Index* memperoleh rata-rata iterasi yaitu 6 iterasi, sedangkan pada K-Means secara konvensional atau tanpa reduksi atribut memperoleh rata-rata iterasi yaitu 9 iterasi. Kemudian dibuktikan juga berdasarkan hasil pengujian pada *Absenteeism_at_work Dataset* yaitu dengan pengujian *K-Means + Gini Index* mampu memperkecil nilai *Sum of Square Error* (SSE) jika dibandingkan dengan pengujian K-Means secara konvensional atau tanpa reduksi atribut yang dimana nilai SSE sebelum reduksi atribut yaitu 1391.613 menjadi 440.912. Maka dapat disimpulkan bahwa penerapan reduksi atribut dengan metode Gini Index pada K-Means Clustering mampu dalam meminimalkan error pada cluster yang dihasilkan serta mampu dalam memilih atribut yang relevan dan mempunyai pengaruh signifikan dalam proses clustering pada dataset yang diujikan dalam penelitian ini.

REFERENCES

[1] I. Alpiana and L. Anifah, "Penerapan Metode KnA (Kombinasi K-Means dan Agglomerative Hierarchical Clustering) dengan Pendekatan Single Linkage untuk Menentukan Status Gizi pada Balita," *Indones. J. Eng. Technol.*, vol. 1, no. 2, pp. 2623–2464, 2019, [Online]. Available: https://journal.unesa.ac.id/index.php/inajet

[2] E. Muningsih, "Kombinasi Metode K-Means Dan Decision Tree Dengan Perbandingan Kriteria Dan Split Data," *J. Teknoinfo*, vol. 16, no. 1, p. 113, 2022, doi: 10.33365/jti.v16i1.1561.

[3] N. K. Zuhal, "Study Comparison K-Means Clustering dengan Algoritma Hierarchical Clustering," *Univ. Nusant. PGRI Kediri. Kediri*, vol. 1, no. 1, pp. 200–205, 2022.

[4] M. Arief Soeleman and F. Ilmu Komputer, "Penentuan Centroid Awal Pada Algoritma K-Means Dengan Dynamic Artificial Chromosomes Genetic Algorithm Untuk Tuberculosis Dataset Pre-Centroid Determination in K-Means Algorithm using Dynamic Artificial Chromosomes Genetic Algorithm for Tuberculosis Datas," *Februari*, vol. 20, no. 1, pp. 97–108, 2021.

[5] G. Rahayu and Mustakim, "Principal Component Analysis Untuk Dimensi Reduksi Data Clustering Sebagai Pemetaan Persentase Sertifikasi Guru Di Indonesia," *Semin. Nas. Teknol. Inf. Komun. dan Ind.*, vol. 0, no. 0, pp. 201–208, 2017, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/SNTIKI/article/view/3265

[6] A. Izzuddin, "Optimasi Cluster pada Algoritma K-Means dengan Reduksi Dimensi Dataset Menggunakan Principal Component Analysis untuk Pemetaan Kinerja Dosen," *Ed. Nop.*, vol. 5, no. 2, pp. 41–46, 2015.

[7] D. Hediyati and I. M. Suartana, "Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro," *J. Inf. Eng. Educ. Technol.*, vol. 5, no. 2, pp. 49–54, 2021.

[8] M. Mauludin Rohman and S. Adinugroho, "Analisis Sentimen pada Ulasan Aplikasi Mobile JKN Menggunakan Metode Maximum Entropy dan Seleksi Fitur Gini Index Text," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 6, pp. 2646–2654, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

Building of Informatics, Technology and Science (BITS)
Volume 4, No 3, Desember 2022 Page: 1309−1316
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i3.2458

[9] H. Irwandi, O. S. Sitompul, and S. Sutarman, "K-Means Performance Optimization Using Rank Order Centroid (ROC) And Braycurtis Distance," *SinkrOn*, vol. 7, no. 2, pp. 472–478, 2022, doi: 10.33395/sinkron.v7i2.11371.

[10] T. Setiyorini and R. T. Asmono, "Penerapan Metode K-Nearest Neighbor Dan Gini Index Pada Klasifikasi Kinerja Siswa," *J. Techno Nusa Mandiri*, vol. 16, no. 2, pp. 121–126, 2019, doi: 10.33480/techno.v16i2.747.

[11] T. Setiyorini and R. T. Asmono, "Penerapan Gini Index dan K-Nearest Neighbor untuk Klasifikasi Tingkat Kognitif Soal Pada Taksonomi Bloom," *Pilar Nusa Mandiri*, vol. 13, no. 2, pp. 209–216, 2017, [Online]. Available: https://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/239

[12] I. Arfiani, H. Yuliansyah, and M. D. Suratin, "Implementasi Bee Colony Optimization Pada Pemilihan Centroid (Klaster Pusat) Dalam Algoritma K-Means," *Build. Informatics, Technol. Sci.*, vol. 3, no. 4, pp. 756–763, 2022, doi: 10.47065/bits.v3i4.1446.

[13] A. I. Lubis, U. Erdiansyah, and R. Siregar, "Komparasi Akurasi pada Naive Bayes dan Random Forest dalam Klasifikasi Penyakit Liver," *J. Comput. Eng. Syst. Sci.*, vol. 7, no. 1, pp. 81–89, 2022.

[14] U. Erdiansyah, A. Irmansyah Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 208, 2022, doi: 10.30865/mib.v6i1.3373.

[15] N. Putu, E. Merliana, and A. J. Santoso, "Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means," *Pros. Semin. Nas. MULTI DISIPLIN ILMU&CALL Pap. UNISBANK*, pp. 978–979, 2016.

[16] A. I. Lubis, P. Sihombing, and E. B. Nababan, "Comparison SAW and MOORA Methods with Attribute Weighting Using Rank Order Centroid in Decision Making," *Mecn. 2020 - Int. Conf. Mech. Electron. Comput. Ind. Technol.*, no. February 2022, pp. 127–131, 2020, doi: 10.1109/MECnIT48290.2020.9166640.

[17] L. Zahrotun, "Analisis Pengelompokan Jumlah Penumpang Bus Trans Jogja Menggunakan Metode Clustering K-Means Dan Agglomerative Hierarchical Clustering (Ahc)," *J. Inform.*, vol. 9, no. 1, pp. 1039–1047, 2015, doi: 10.26555/jifo.v9i1.a2045.

[18] D. Jollyta, S. Efendi, M. Zarlis, and H. Mawengkang, "Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum of Square Error dan Davies Bouldin Index," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 918, 2019, doi: 10.30645/senaris.v1i0.100.

[19] 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, doi: 10.33633/tc.v20i2.4572.