Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9394

Peningkatan Kinerja K-Means Clustering Berdasarkan Pembobotan Jarak Menggunakan Metode Principal Component Analysis

Alfian Agusnady*, Opim Salim Sitompul, Tulus

Fakultas Ilmu Komputer dan Teknologi Informasi, Magister Teknik Informatika, Universitas Sumatera Utara, Indonesia
Email: 1*alfianagusnady@gmail.com, 2 opim@usu.ac.id, 3 tulus@usu.ac.id
Email Penulis Korespondensi: alfianagusnady@gmail.com*
Submitted: **30/11/2025**; Accepted: **19/12/2025**; Published: **31/12/2025**

Abstrak–Algoritma K-Means memiliki beberapa kelemahan, salah satunya terletak pada model jarak yang digunakan dalam penentuan kemiripan antar data yang memberikan perlakuan yang sama terhadap setiap atribut data, sehingga atribut yang kurang relevan dan memiliki sedikit kontribusi terhadap variasi data dapat memberikan dampak yang cukup berpengaruh terhadap hasil clustering. Hal ini tentu saja dapat menurunkan kinerja algoritma K-Means. Pembobotan atribut merupakan salah satu cara yang dapat digunakan untuk mendapatkan korelasi atribut data terhadap variasi data. Semakin tinggi nilai bobot dari suatu atribut maka semakin besar korelasinya terhadap variasi data, sehingga nilai bobot yang rendah dari suatu atribut tentunya memiliki sedikit kontribusi terhadap variasi data dan dapat memberikan dampak yang cukup berpengaruh terhadap kinerja dan hasil clustering. Pada penelitian ini, metode yang digunakan dalam perhitungan bobot atribut data yaitu Principal Component Analysis (PCA). Untuk melakukan pengujian terhadap metode yang diusulkan, maka penelitian ini menggunakan dataset dari UCI Machine Learning yang terdiri dari 351 data Ionosphere, 4177 data Abalone serta 1096 data kualitas udara dari Laboratorium Udara Kota Pekanbaru dan 120 data kualitas air. Evaluasi kinerja Clustering yang diusulkan berdasarkan nilai Sum of Square Error (SSE). Hasil pengujian pada penelitian ini terlihat bahwa dengan metode yang diusulkan dapat menghasilkan nilai SSE yang signifikan lebih kecil.

Kata Kunci: Algoritma K-Means; Pembobotan atribut; Principal Component Analysis; Clustering; Sum of Square Error (SSE).

Abstract–The K-Means algorithm has several weaknesses, one of which lies in the distance model used in determining the similarity between data which provides the same treatment for each data attribute, so that attributes that are less relevant and have little contribution to data variation can have a significant impact on Clustering results. This of course can reduce the performance of the K-Means algorithm. Attribute weighting is one way that can be used to get the correlation of data attributes to data variations. The higher the weight value of an attribute, the greater the correlation to data variation, so that the low weight value of an attribute certainly has little contribution to data variation and can have a significant impact on performance and Clustering results. In this study, the method used in calculating the weight of data attributes is Principal Component Analysis (PCA). To test the proposed method, this study uses a dataset from UCI Machine Learning which consists of 351 Ionosphere data, 4177 Abalone data and 1096 air quality data from Pekanbaru City Air Laboratory and 120 water quality data. The evaluation of the proposed Clustering performance is based on the Sum of Square Error (SSE) value. The test results in this study show that the proposed method can produce a significantly smaller SSE value.

Keywords: K-Means algorithm; Attribute weighting; Principal Component Analysis; Clustering; Sum of Square Error (SSE)

1. PENDAHULUAN

Algoritma *K-Means* merupakan salah satu metode *Clustering* yang paling banyak digunakan dalam analisis data karena kesederhanaan dan efisiensinya dalam mengelompokkan data berdasarkan tingkat kemiripan. Namun, salah satu permasalahan utama pada algoritma *K-Means* terletak pada penggunaan perhitungan jarak yang memperlakukan setiap atribut secara setara tanpa mempertimbangkan tingkat kepentingan masing-masing atribut terhadap struktur data. Kondisi ini menyebabkan atribut yang kurang relevan atau memiliki variasi rendah tetap memberikan kontribusi yang sama besar terhadap proses pengelompokan, sehingga berpotensi menurunkan kualitas hasil clustering. Dalam data mining, data dari suatu himpunan data dikelompokkan ke dalam kategori atau kelompok tertentu berdasarkan kemiripan terhadap centroid cluster[1]-[3], dan proses ini disebut sebagai *Clustering* yang diklasifikasikan sebagai unsupervised learning karena hanya data masukan dan parameter yang digunakan tanpa adanya kelas. Objek-objek dikelompokkan ke dalam cluster sehingga tiap kelompok memiliki kemiripan yang lebih tinggi dibandingkan dengan objek pada kelompok lain, di mana kemiripan diukur menggunakan model pengukuran jarak dan teknik *Clustering* dibagi menjadi pendekatan hierarchical dan non-hierarchical seperti *Partition Clustering* dengan algoritma *K-Means* [4].

Dalam algoritma *K-Means*, selisih jarak setiap data ke centroid dihitung dan diperbarui secara berulang sampai perpindahan data antar cluster tidak lagi terjadi atau batas iterasi tercapai, dan dengan cara ini jarak dalam-cluster diminimalkan[5]-[8]. Namun, pemilihan centroid awal yang dilakukan secara acak dan penggunaan model jarak yang memberi pengaruh sama pada setiap atribut menyebabkan hasil *Clustering* berpotensi menjadi tidak stabil, tidak akurat, dan kinerja *K-Means* menurun ketika atribut kurang relevan tetap diberi bobot yang sama[9]. Untuk mengatasi kelemahan tersebut, pembobotan atribut diusulkan agar atribut dengan korelasi lebih tinggi terhadap variasi data diberikan bobot yang lebih besar, sementara atribut dengan kontribusi kecil diberi bobot rendah[10]-[11], misalnya dengan pendekatan Gain Ratio atau *Principal Component Analysis* (PCA).

Berbagai penelitian telah dilakukan, seperti pendekatan object weighting untuk mendeteksi outlier dan meningkatkan kestabilan cluster, penerapan PCA sebelum *Clustering* untuk mereduksi atribut kurang relevan dan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9394

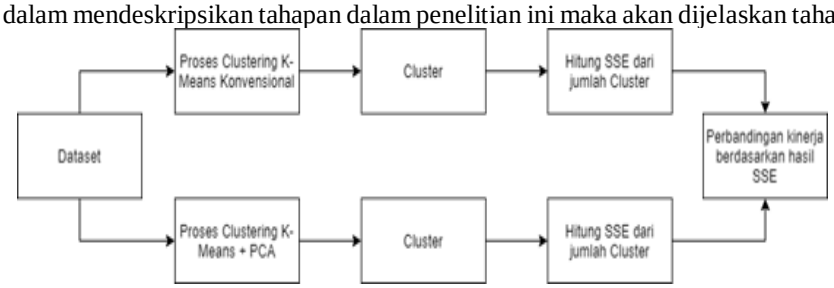
meningkatkan kinerja, serta algoritma Feature-Reduction Multi-View *K-Means* (FRMVK) yang memberikan bobot atribut berbasis seleksi fitur[12]. Dalam penelitian yang diringkas, korelasi antar atribut dimanfaatkan sebagai dasar pemberian bobot pada perhitungan jarak di algoritma *K-Means* sehingga diharapkan kelemahan *K-Means* dapat diatasi[13]-[15]. Pengujian dibatasi pada dataset Ionosphere dan Abalone dari UCI Machine Learning, data kualitas air, dan data kualitas udara, dengan pembobotan atribut menggunakan PCA dan kinerja *Clustering* dievaluasi berdasarkan nilai *Sum of Square Error* (SSE), di mana nilai SSE yang lebih kecil diharapkan dapat diperoleh oleh metode yang diusulkan.

Pada akhir penelitian ini, tujuan utama yang ingin dicapai adalah mengembangkan dan mengevaluasi pendekatan *K-Means Clustering* berbasis pembobotan atribut menggunakan metode *Principal Component Analysis* (PCA) guna meningkatkan kualitas hasil pengelompokan data. Penelitian ini bertujuan untuk membuktikan bahwa pemberian bobot atribut berdasarkan tingkat kontribusinya terhadap variasi data mampu mengurangi pengaruh atribut yang kurang relevan, sehingga menghasilkan nilai *Sum of Square Error* (SSE) yang lebih kecil dibandingkan metode K-Means konvensional. Selain itu, penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam mengenai peran pembobotan atribut dalam meningkatkan stabilitas dan akurasi proses clustering, serta menjadi referensi bagi pengembangan metode pengelompokan data yang lebih efektif pada berbagai jenis dataset di masa mendatang.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Algoritma *K-Means Clustering* umumnya menggunakan perhitungan jarak terdekat (*nearest distance*) antar data kedalam masing-masing *centroid* cluster berdasarkan persamaan jarak seperti *Euclidean Distance*. Teknik *Clustering* yang diusulkan sebagai dasar dalam pemberian nilai bobot terhadap model perhitungan jarak terdekat[16]- [20]. Semakin tinggi nilai bobot dari suatu atribut maka semakin besar korelasinya terhadap variasi data, Sehingga nilai bobot yang rendah dari suatu atribut tentunya memiliki sedikit kontribusi terhadap variasi data dan dapat memberikan dampak yang cukup berpengaruh terhadap hasil *clustering*. Tahapan PCA bertujuan untuk memperbaiki kelemahan *K-Means clustering*, khususnya pada penentuan kemiripan data terhadap *centroid cluster*. Untuk lebih jelas dalam mendeskripsikan tahapan dalam penelitian ini maka akan dijelaskan tahapan demi tahapan.



Gambar 1. Metode Penelitian

Berikut adalah tahapan – tahapan penelitian dapat dijelaskan, yaitu: Lakukan proses awal persiapan data, Lakukan perhitungan bobot menggunakan PCA, Lakukan proses *Clustering* menggunakan algoritma *K-Means* clustering[21], Analisis hasil *Clustering* berdasarkan total selisih jarak menggunakan persamaan Sum of Square-Error.

2.2 Data

Terdapat 4 dataset yang digunakan terdiri dari 2 dataset yang berasal dari UCI *Machine Learning Repository*, 1 dataset hasil penelitian Denades *et al*. (2016) dan 1 dataset dari Pengolahan Data Laboratorium Udara Pemerintah Pekanbaru. Dataset yang berasal dari UCI *Machine Learning Repository* adalah dataset *ionosphere* dan *abalone*. Dataset *ionosphere* berasal Goose Bay, Labrador, dimana data tersebut merupakan sinyal radar yang diterima dan diproses menggunakan fungsi autokorelasi. Dataset *abalone* berasal dari Marine Research Laboratories – Taroona dimana data tersebut digunakan untuk memprediksi umur *abalone* dari penampakan fisik[22]-[24].

2.2 Software dan Tools

Untuk mempermudah implementasi dan perhitungan dalam penelitian, Maka penulis menggunakan bahasa pemrograman *Python* versi 3.7, Menggunakan spesifikasi *processor* Intel Core i5 dan RAM 4 GB.

3. HASIL DAN PEMBAHASAN

Data yang digunakan diperoleh dari UCI *Machine Learning Repository*, yaitu: *Ionosphere* dan *Abalone*. Dalam penelitian ini juga menggunakan dataset *Water Quality* yang berasal dari hasil penelitian Denades *et al.* (2016), dimana data tersebut merupakan hasil pengumpulan data Kementerian Lingkungan Hidup tentang Ketentuan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9394

Kualitas Air. Kemudian Dataset Kualitas Udara yang diperoleh dari Pengolahan Data Laboratorium Udara Kota Pekanbaru selama 3 tahun terakhir yakni Tahun 2014, 2015 dan 2016 (*http://iku.menlhk.go.id*), merupakan hasil pengumpulan data oleh Laboratorium Udara Pemerintah Kota Pekanbaru tentang Ketentuan Kualitas Udara Kota Pekanbaru yang digolongkan kedalam enam kategori.

3.1 Data Preprocessing

Hasil dari proses data preprocessing dataset *Ionosphere* menggunakan bahasa pemrograman Python.

```
        SR1     SR2    SR3    SR4    SR5  ...    SR28   SR29   SR30   SR31   SR32
0     1.000 -0.060  0.850  0.020  0.830  ... -0.341  0.423 -0.545  0.186 -0.453
1     1.000 -0.190  0.930 -0.360 -0.110  ... -0.116 -0.166 -0.063 -0.137 -0.024
2     1.000 -0.030  1.000  0.000  1.000  ... -0.174  0.604 -0.242  0.560 -0.382
3     1.000 -0.450  1.000  1.000  0.710  ... -0.201  0.257  1.000 -0.324  1.000
4     1.000 -0.020  0.940  0.070  0.920  ... -0.622 -0.057 -0.596 -0.046 -0.657
..      ...    ...    ...    ...    ...  ...    ...    ...    ...    ...    ...
345   0.667 -0.014  0.974  0.068  0.496  ...  0.078  0.553  0.246  0.139  0.481
346   0.835  0.083  0.737 -0.147  0.843  ...  0.128  0.867 -0.107  0.905 -0.043
347   0.951  0.004  0.952 -0.027  0.934  ...  0.082  0.941  0.000  0.915  0.047
348   0.947  0.000  0.932 -0.032  0.952  ...  0.022  0.925  0.004  0.927 -0.006
349   0.906 -0.017  0.981 -0.020  0.957  ... -0.172  0.960 -0.038  0.874 -0.162
350   0.800  0.100  0.700 -0.100  0.900  ... -0.007  0.757 -0.067  0.858 -0.062

[350 rows x 32 columns]
#Record(Before): 351
#Record(After): 350
#Record Remove: 1
```

Gambar 2. *Output Data Preprocessing* (Dataset Ionosphere)

Proses awal melibatkan preprocessing data pada empat dataset: Ionosphere, Abalone, Kualitas Udara, dan Water Quality. Langkah utamanya adalah pembersihan data (*data cleaning*) dengan menghapus *record* yang duplikat. Hasilnya, jumlah data observasi pada dataset Ionosphere berkurang menjadi 350, dataset Kualitas Udara menjadi 1.080, dan dataset Water Quality menjadi 117 *record*. Dataset Abalone tidak mengalami perubahan jumlah data karena tidak ditemukan duplikasi. Selain itu, dilakukan penghapusan kolom target ('Class' pada Water Quality dan 'Kategori' pada Kualitas Udara) untuk menyederhanakan format kolom.

```
      Length  Diameter  Height  ...  Shucked_weight  Viscera_weight  Shell_weight
0      0.455     0.365   0.095  ...          0.2245          0.1010        0.1500
1      0.350     0.265   0.090  ...          0.0995          0.0485        0.0700
2      0.530     0.420   0.135  ...          0.2565          0.1415        0.2100
3      0.440     0.365   0.125  ...          0.2155          0.1140        0.1550
4      0.330     0.255   0.080  ...          0.0895          0.0395        0.0550
...      ...       ...     ...  ...             ...             ...           ...
4174   0.600     0.475   0.205  ...          0.5255          0.2875        0.3080
4175   0.625     0.485   0.150  ...          0.5310          0.2610        0.2960
4176   0.710     0.555   0.195  ...          0.9455          0.3765        0.4950

[4177 rows x 7 columns]
#Record(Before): 4177
#Record(After): 4177
#Record Remove: 0
```

Gambar 3. *Output Data Preprocessing* (Dataset *Abalone*)

Terlihat hasil dari proses *preprocessing* pada dataset *Ionosphere*, dimana terdapat sebanyak 1 record yang duplikat sehingga proses *cleaning* dilakukan dengan cara membuang duplikasi data tersebut sehingga jumlah data observasi pada dataset *Ionosphere* yang semula sebanyak 351 record menjadi 350 *record*.

```
-- Hasil Preprocessing Dataset Udara --
      PM10  SO2  CO  O3  NO2
0       47   51   8  67    2
1       48   51   9  37    2
2       37   51   9  26    2
3       24   50   2  51    1
4       25   50   3  36    1
...    ...  ...  ..  ...  ...
1093    40   13   9  32    8
1094    33   10   7  21    7
1095    37   11   7  23    7

[1080 rows x 5 columns]
#Record(Before): 1096
#Record(After): 1080
#Record Remove: 16
```

Gambar 4. *Output Data Preprocessing* (Dataset Kualitas Udara)

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9394

Terlihat hasil standarisasi data dataset Kualitas Udara, dimana interval atau rentang data menjadi lebih proporsional. Nilai *standard score* atribut PM10 pada record data ke-1 adalah sebesar -0.1389 atau dibulatkan menjadi -0.14. Nilai *standard score* atribut SO2 pada record data ke-1 adalah sebesar 2.8057 atau dibulatkan menjadi 2.81. Hal yang sama juga berlaku untuk atribut lainnya pada record data selanjutnya.

```
====================== Normal Z-Score (Water Quality) ==========================
          TSS        DO       COD  ...  Fecal_Coliform  Total_Coliform       Pij
0   -0.908499 -0.455460 -0.442992  ...       -0.160822       -0.390098 -1.040001
1   -0.890764 -0.083938 -0.202853  ...       -0.160822       -0.390098 -1.014993
2   -0.890764 -0.158243 -0.271464  ...       -0.158512       -0.389891 -1.008741
3   -0.873029 -0.381156 -0.511753  ...       -0.158043       -0.389983 -1.017077
4   -0.873029 -0.306851 -0.442992  ...       -0.160441       -0.390043 -1.029581
..        ...       ...       ...  ...             ...             ...       ...
112  0.386152 -0.202825  0.773998  ...       -0.127997       -0.197077  0.987723
113  0.563501 -1.503151  0.479828  ...        0.015341        0.712941  1.554569
114  0.634441  1.677074 -0.049550  ...       -0.149222        0.464754  1.423277
115  0.740851 -1.183642  0.443378  ...       -0.119728        2.827126  1.881756
116  0.776320  1.253539  0.573311  ...       -0.156940       -0.132733  1.039823

[117 rows x 8 columns]
```

Gambar 5. *Output Data Normalization* (Dataset *Water Quality*)

Nilai *standard score* atribut TSS pada record data ke-1 adalah sebesar -0.9085 atau dibulatkan menjadi -0.91. Nilai *standard score* atribut DO pada record data ke-1 adalah sebesar -0.4555 atau dibulatkan menjadi -0.46.

3.2. *Bobot menggunakan PCA*

PCA digunakan sebagai tolak ukur untuk menguji korelasi antar atribut. PCA dalam penelitian ini digunakan sebagai dasar pemberian bobot terhadap karakteristik jarak antar data. Agar mempermudah perhitungan PCA terhadap seluruh data maka digunakan dukungan bahasa pemrograman *Python. pembentukan Covariance Matrix dataset Ionosphere*.

```
----------------- Matrik Kovarian(Ionosphere) -----------------
          SR1       SR2       SR3  ...      SR30      SR31      SR32
SR1    1.002865  0.143650  0.475641  ... -0.009369  0.262231  0.000465
SR2    0.143650  1.002865  0.000777  ... -0.123125 -0.154940  0.034689
SR3    0.475641  0.000777  1.002865  ...  0.025872  0.383193 -0.100011
SR4    0.024948 -0.190082  0.037606  ...  0.317013  0.016236  0.185794
SR5    0.439070 -0.054640  0.597152  ... -0.008364  0.546735 -0.076933
SR6    0.007912  0.255693 -0.030654  ...  0.152816 -0.201867  0.361582
SR7    0.470930 -0.304086  0.449773  ... -0.067767  0.344604 -0.096057
SR8    0.046905  0.208346 -0.035551  ... -0.015657 -0.205119  0.098320
SR9    0.323820 -0.191208  0.449535  ...  0.024010  0.339384 -0.152906
SR10   0.169730  0.316789  0.042109  ...  0.010335 -0.182339  0.066738
SR11   0.216336 -0.150073  0.482459  ... -0.041801  0.474483 -0.065416
SR12   0.164749  0.237282  0.127268  ... -0.067932 -0.096772  0.096726
SR13   0.197348 -0.254290  0.399089  ... -0.013220  0.491766 -0.120373
SR14   0.094266  0.186421  0.087969  ...  0.101762 -0.180816  0.242354
SR15   0.220261 -0.252340  0.277221  ...  0.015922  0.360262 -0.063695
SR16   0.172968 -0.147850  0.027741  ...  0.257750 -0.059270  0.054280
SR17   0.284577 -0.333716  0.220615  ...  0.125073  0.446693  0.001417
SR18   0.151795  0.167751  0.042309  ...  0.125751 -0.108102  0.189433
SR19   0.147981 -0.282382  0.325931  ...  0.122303  0.618011 -0.002618
SR20   0.138821 -0.035437  0.164483  ...  0.208327 -0.090751  0.133228
SR21   0.249945 -0.144522  0.503486  ...  0.161042  0.504270  0.049418
SR22  -0.012343  0.164649  0.099073  ...  0.084536 -0.212159  0.095002
SR23   0.303993 -0.105378  0.242017  ...  0.134923  0.461285  0.111414
SR24  -0.073077 -0.237628 -0.031917  ...  0.341657 -0.086038  0.222272
SR25   0.081578 -0.046980  0.144612  ...  0.160935  0.473156  0.083420
SR26   0.118177  0.000268  0.180531  ...  0.504346 -0.129067  0.379271
SR27   0.343852 -0.041602  0.256745  ...  0.056176  0.580795  0.082750
SR28   0.058372  0.343302  0.051513  ...  0.297940 -0.101123  0.392431
SR29   0.245629 -0.173247  0.399015  ... -0.028999  0.694032 -0.037751
SR30  -0.009369 -0.123125  0.025872  ...  1.002865 -0.013065  0.516484
SR31   0.262231 -0.154940  0.383193  ... -0.013065  1.002865 -0.132354
SR32   0.000465  0.034689 -0.100011  ...  0.516484 -0.132354  1.002865

[32 rows x 32 columns]
```

Gambar 6. *Output Covariance Matrix* (Dataset *Ionosphere*)

Atribut SR1 memiliki korelasi ragam peragam (*varians covariance*) sebesar 1, Sementara korelasi atribut SR1 dengan SR2 dan sebaliknya atribut SR2 dengan SR1 memiliki korelasi (*varians covariance*) sebesar 0.143 dan seterusnya dengan cara yang sama juga berlaku untuk setiap pasangan atribut. *output* pembentukan *Covariance Matrix* dataset *Abalone*.

```
----------------- Matrik Kovarian(Abalone) -----------------
                  Length  Diameter  ...  Viscera_weight  Shell_weight
Length          1.000239  0.987048  ...        0.903234      0.897921
Diameter        0.987048  1.000239  ...        0.899940      0.905547
Height          0.827752  0.833883  ...        0.798510      0.817534
Whole_weight    0.925483  0.925674  ...        0.966606      0.955584
Shucked_weight  0.898129  0.893376  ...        0.932184      0.882828
Viscera_weight  0.903234  0.899940  ...        1.000239      0.907874
Shell_weight    0.897921  0.905547  ...        0.907874      1.000239

[7 rows x 7 columns]
```

Gambar 7. *Output Covariance Matrix* (Dataset *Abalone*)

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9394

Terlihat hasil perolehan nilai korelasi antar atribut dataset *Abalone* menggunakan persamaan Atribut *Length* memiliki korelasi ragam peragam (*varians covariance*) sebesar 1, Sementara korelasi atribut *Length* dengan *Diameter* dan sebaliknya atribut *Diameter* dengan *Length* memiliki korelasi (*varians covariance*) sebesar 0.987 dan seterusnya dengan cara yang sama juga berlaku untuk setiap pasangan atribut. perolehan *eigenvalue* dataset *Ionosphere* menggunakan pemrograman *Python*

```
-------------------- Nilai Eigen(Ionosphere) ------------------
[8.8192476  4.24898003 2.60682746 2.37686674 1.89432707 1.15049698
 1.02456892 0.93589799 0.88935088 0.81901965 0.72563565 0.61569827
 0.57915467 0.54307015 0.49194485 0.47473106 0.44504096 0.07205123
 0.41115525 0.09748716 0.37409399 0.12966365 0.14470289 0.16937935
 0.18037646 0.20531661 0.21571365 0.22908161 0.26167616 0.33947187
 0.30580596 0.31485579]
```

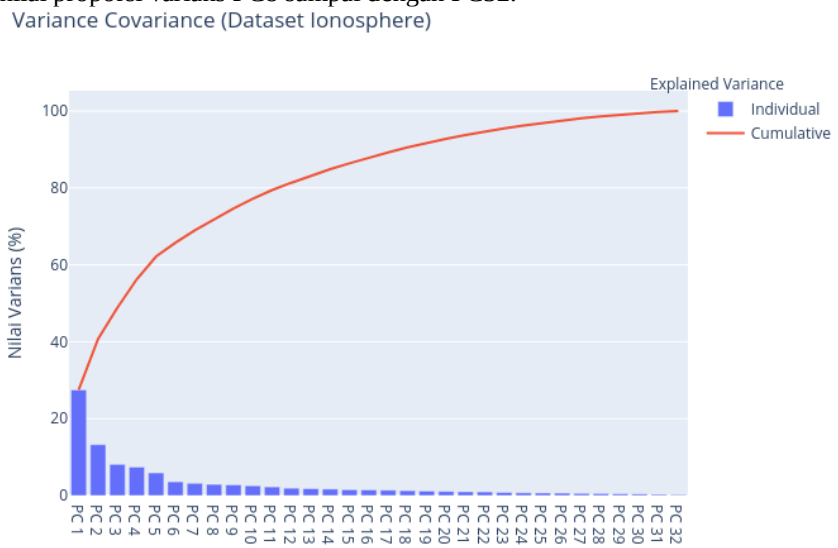
Gambar 8. *Output Eigenvalue* (Dataset *Ionosphere*)

Untuk memperoleh nilai proporsi, maka terlebih dahulu menghitung nilai ragam peragam (*Variance Covariance*).

```
-------- Nilai Variansi Kumulatif(%)-Ionosphere -------
      PC1    PC2    PC3    PC4    PC5  ...   PC28   PC29   PC30   PC31   PC32
0   27.48  40.72  48.84  56.25  62.15  ...  98.62  99.07  99.47  99.78  100.0
```

Gambar 9. *Output Proportion Cumulative*

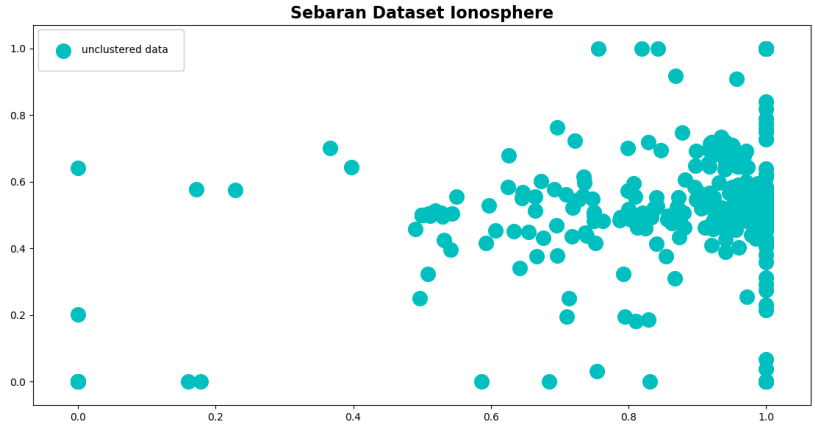
Dari gambar 4.17, Terlihat hasil perolehan persentase nilai proporsi varians dataset *Ionosphere* secara *cumulative*, artinya persentase nilai proporsi *cumulative* PC1 sebesar 27.48%, PC2 sebesar 40.72%, PC3 sebesar 48.84%, PC4 sebesar 56.25%, PC5 sebesar 62.15%, PC6 sebesar 65.74%, PC7 sebesar 68.93%, Hal yang sama juga berlaku untuk persentase nilai proporsi varians PC8 sampai dengan PC32.



Gambar 10. Grafik *Variance Covariance* (Dataset *Ionosphere*)

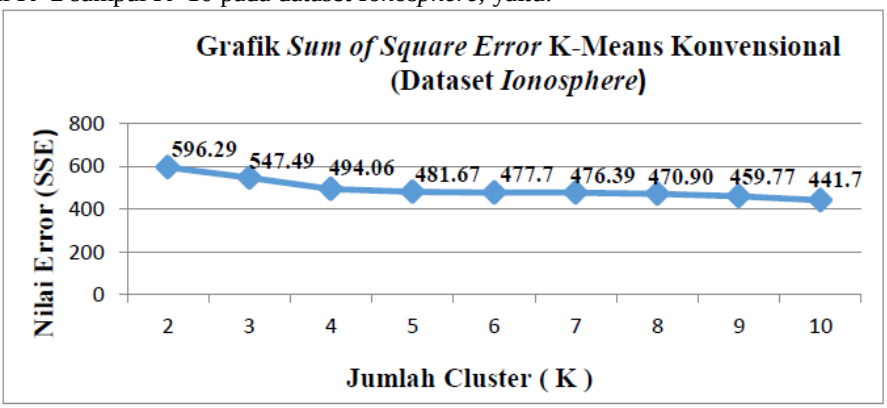
Implementasi hasil pengujian dari model *Clustering* dibagi atas dua bagian yaitu: model *K-Means* konvensional dan model *Distance Weight K-Means* menggunakan pendekatan PCA.
Berdasarkan hasil *data preprocessing* dataset *Ionosphere* memiliki 32 atribut, 2 kelas dan 350 *instances*. Distribusi kelas yaitu bernilai 1 sebanyak 313 *instances* dan bernilai 0 sebanyak 37 *instances*. Proses *Clustering* diawali dengan melakukan standarisasi data (*Min-Max normalization)*.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9394



Gambar 11. Sebaran Dataset *Ionosphere* (*Unclustered Data*)

Terlihat hasil tampilan dari sebaran informasi akan dataset *Ionosphere* sebelum dilakukannya proses *clustering*. Gambar menunjukkan simbol lingkaran berwarna *cyan* menunjukkan titik nilai yang belum dilakukan *Clustering* dari masing-masing atribut pada dataset *Ionosphere* dan rentang nilai masing-masing atribut bernilai terendah adalah 0.1 dan nilai tertinggi adalah 1. Hal ini dilakukan agar mempermudah proses *clustering.*
Berikut adalah grafik kinerja *Clustering K-Means* konvensional berdasarkan perolehan nilai *Sum of Square Error* (SSE), saat nilai K=2 sampai K=10 pada dataset *Ionosphere*, yaitu:



Gambar 12. *Sum of Square Error K-Means Konvensional* (*Ionosphere*)

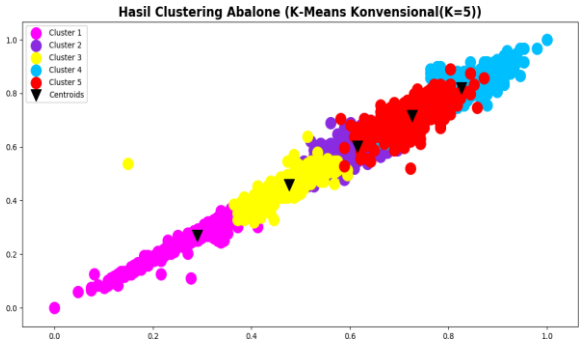
grafik rekapitulasi SSE *K-Means* konvensional dataset *Ionosphere*, saat nilai K=2 sampai K=10. Saat nilai K=2 terjadi error senilai 596.29, Saat K=3 terjadi error senilai 547.49, Saat K=4 terjadi error senilai 494.06, Saat K=5 terjadi error senilai 481.67, Saat K=6 terjadi error senilai 477.7, Saat K=7 terjadi error senilai 476.39, Saat K=8 terjadi error senilai 470.90, Saat K=9 terjadi error senilai 459.77 dan K=10 terjadi error senilai 441.7.
dataset *Abalone* memiliki 8 atribut dan 4.177 *instances*. Berikut adalah informasi nilai per atribut dari dataset *Abalone*, yaitu:

Tabel 1. Informasi Atribut Dataset *Abalone*

No.	Atribut	Nilai
1	*Sex*	{M, F, I}
2	*Length*	[0.075, 0.815]
3	*Diameter*	[0.055, 0.65]
4	*Height*	[0.0, 1.13]
5	*Whole_weight*	[0.0020, 2.8255]
6	*Shucked_weight*	[0.0010, 1.488]
7	*Viscera_weight*	[0.0005, 0.76]
8	*Shell_weight*	[0.0015, 1.005]

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9394

Proses *Clustering* diawali dengan melakukan standarisasi data (*Min-Max normalization)*, agar interval atau rentang data menjadi lebih proporsional



Gambar 13. Sebaran Centroid Akhir *Abalone* (*K-Means Konvensional*)

dari hasil yang diperoleh melalui selisih derajat error dari masing- masing dataset, maka penelitian ini telah mampu menjawab tujuan penelitian diawal yakni meminimalkan nilai derajat *error K-Means Clustering* dengan cara memberikan bobot terhadap model jarak (*distance weight*) menggunakan PCA, Sehingga meningkatkan kinerja *K-Means Clustering* dalam menentukan kemiripan data terhadap centroid cluster.

4. KESIMPULAN

Penggunaan metode PCA sebagai dasar pemberian bobot terhadap karakteristik jarak antar data telah memiliki kontribusi yang dapat meminimalkan nilai derajat *error pada metode K-Means clustering*. Pemberian nilai bobot menggunakan PCA diperoleh mulai dari menghitung korelasi antar atribut, persentase nilai proporsi varians data, memilih kontribusi maksimum terhadap variasi data, kemudian memilih nilai maksimum yang dijadikan sebagai nilai bobot atribut. Dan, Pembobotan jarak pada *K-Means* menggunakan PCA terbukti dapat meminimalkan nilai derajat *error pada K-Means clustering*, dimana perubahan nilai derajat error yang signifikan banyak terjadi saat menggunakan metode pembobotan jarak (nilai K=5) yaitu dataset *Water Quality* sebesar 11.80, sedangkan nilai derajat error tanpa pembobotan jarak sebesar 13.57. Jika dibandingkan dengan dataset *Ionosphere* dimana perubahan nilai derajat error yang relatif masih sedikit saat menggunakan metode pembobotan jarak (nilai K=5) yaitu sebesar 468.32, sedangkan nilai derajat error tanpa pembobotan jarak sebesar 481.67. Kemudian, *K-Means* merupakan metode pengelompokkan data yang cukup baik, namun dapat menghasilkan nilai *centroid* yang bersifat *outlier*, sehingga metode penentuan nilai *centroid* menjadi sangat penting.

REFERENCES

[1] Y. Chen, P. Hu, and W. Wang, "Improved *K-Means* Algorithm and its Implementation Based on Mean Shift," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, Oct. 2018, pp. 1–5. doi: 10.1109/cisp-bmei.2018.8633100.

[2] A. Danades, D. Pratama, D. Anggraini, and D. Anggriani, "Comparison of accuracy level K-Nearest Neighbor algorithm and Support Vector Machine algorithm in classification water quality status," in *2016 6th International Conference on System Engineering and Technology (ICSET)*, IEEE, Oct. 2016, pp. 137–141. doi: 10.1109/icsengt.2016.7849638.

[3] V. Divya and K. N. Devi, "An Efficient Approach to Determine Number of Clusters Using *Principal Component Analysis*," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, IEEE, Mar. 2018, pp. 1–6. doi: 10.1109/icctct.2018.8551182.

[4] P. Fränti and S. Sieranoja, "How much can *K-Means* be improved by using better initialization and repeats?," *Pattern Recognition*, vol. 93, pp. 95–112, Sep. 2019, doi: 10.1016/j.patcog.2019.04.014.

[5] A. Gondeau, Z. Aouabed, M. Hijri, P. Peres-Neto, and V. Makarenkov, "Object Weighting: A New *Clustering* Approach to Deal with Outliers and Cluster Overlap in Computational Biology," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 18, no. 2, pp. 633–643, Mar. 2021, doi: 10.1109/tcbb.2019.2921577.

[6] V. Kotu and B. Deshpande, "Data Mining Process," in *Predictive Analytics and Data Mining*, Elsevier, 2015, pp. 17–36. doi: 10.1016/b978-0-12-801460-8.00002-1.

[7] Y. Li, J. Cai, H. Yang, J. Zhang, and X. Zhao, "A Novel Algorithm for Initial Cluster Center Selection," *IEEE Access*, vol. 7, pp. 74683–74693, 2019, doi: 10.1109/access.2019.2921320.

[8] D. Lase and T. S. Alasi, "Penerapan Web untuk Pengolahan Data Pegawai Kantor Desa Menggunakan Bahasa Pemrograman PHP dan UML," *Jurnal Mahajana Informasi*, vol. 9, no. 1, pp. 1–6, 2024.

[9] A. A. Nababan, O. S. Sitompul, and Tulus, "Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio," *Journal of Physics: Conference Series*, vol. 1007, p. 12007, Apr. 2018, doi: 10.1088/1742-6596/1007/1/012007.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 2, Desember 2025
e-ISSN 2685-998X
DOI 10.30865/json.v7i2.9394

[10] J. Ortiz-Bejar, E. S. Tellez, M. Graff, J. Ortiz-Bejar, J. C. Jacobo, and A. Zamora-Mendez, "Performance Analysis of *K-Means* Seeding Algorithms," in *2019 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, IEEE, Nov. 2019, pp. 1–6. doi: 10.1109/ropec48299.2019.9057044.

[11] A. S. Edu, D. Agozie, and M. Agoyi, "Digital security vulnerabilities and threats implications for financial institutions deploying digital technology platforms and application: FMEA and FTOPSIS analysis," *PeerJ Computer Science*, vol. 7, pp. 1–26, 2021, doi: 10.7717/PEERJ-CS.658.

[12] "No Title," in *Applied Multivariate Statistical Analysis*, Springer Berlin Heidelberg, pp. 93–146. doi: 10.1007/978-3-540-72244-1_4.

[13] D. Tanir and F. Nuriyeva, "On selecting the Initial Cluster Centers in the *K-Means* Algorithm," in *2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, Sep. 2017, pp. 1–5. doi: 10.1109/icaict.2017.8687081.

[14] C. Xiong, Z. Hua, K. Lv, and X. Li, "An Improved *K-Means* Text *Clustering* Algorithm by Optimizing Initial Cluster Centers," in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, IEEE, Nov. 2016, pp. 265–268. doi: 10.1109/ccbd.2016.059.

[15] M.-S. Yang and K. P. Sinaga, "A Feature-Reduction Multi-View *K-Means* *Clustering* Algorithm," *IEEE Access*, vol. 7, pp. 114472–114486, 2019, doi: 10.1109/access.2019.2934179.

[16] Z. Sitorus, A. P. U. Siahaan, B. S. Ibrahim, A. O. Sari, and A. Ibezato, "Strategi Digital Marketing Dengan Metode SEO (Search Engine Optimization) untuk UMKM di Desa Klambir 5 Kebun," 2022.

[17] A. I. Zalukhu, I. Syahputra, Z. Sitorus, and others, "Penerapan Metode Certainty Factor Pada Sistem Pakar Diagnosa Penyakit Gigi Dan Mulut," *Bulletin of Information Technology (BIT)*, vol. 4, no. 4, pp. 544–553, 2023.

[18] A. Astel, S. Tsakovski, P. Barbieri, and V. Simeonov, "Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets," *Water Res*, vol. 41, no. 19, pp. 4566–4578, 2007.

[19] M. K. Islam, M. S. Ali, M. S. Miah, M. M. Rahman, M. S. Alam, and M. A. Hossain, "Brain tumor detection in MR image using superpixels, *Principal Component Analysis* and template based *K-Means* *Clustering* algorithm," *Machine Learning with Applications*, vol. 5, p. 100044, 2021.

[20] L. Yin, L. Lv, D. Wang, Y. Qu, H. Chen, and W. Deng, "Spectral *Clustering* approach with K-nearest neighbor and weighted mahalanobis distance for data mining," *Electronics (Basel)*, vol. 12, no. 15, p. 3284, 2023.

[21] K. Honda, A. Notsu, and H. Ichihashi, "PCA-guided *K-Means* with variable weighting and its application to document clustering," in *International Conference on Modeling Decisions for Artificial Intelligence*, 2009, pp. 282–292.

[22] S. M. Shaharudin, N. Ahmad, and F. Yusof, "Improved cluster partition in *Principal Component Analysis* guided clustering," *Int J Comput Appl*, vol. 75, no. 11, 2013.

[23] S. Surono and R. D. A. Putri, "Optimization of fuzzy c-means *Clustering* algorithm with combination of minkowski and chebyshev distance using *Principal Component Analysis*," *International journal of fuzzy systems*, vol. 23, no. 1, pp. 139–144, 2021.

[24] S. Mulyaningsih and J. Heikal, "*K-Means* *Clustering* Using *Principal Component Analysis* (PCA) Indonesia Multi-Finance Industry Performance Before and During Covid-19," *APMBA (Asia Pacific Management and Business Application)*, vol. 11, no. 2, pp. 131–142, 2022.