

See discussions, stats, and author profiles for this publication at: [https://www.researchgate.net/publication/370653667](https://www.researchgate.net/publication/370653667_Peningkatan_Akurasi_Algoritma_K-Means_dengan_Clustering_Purity_Sebagai_Titik_Pusat_Cluster_Awal_Centroid?enrichId=rgreq-d98d10148ad2ca7cc2b65b9399a7769b-XXX&enrichSource=Y292ZXJQYWdlOzM3MDY1MzY2NztBUzoxMTQzMTI4MTE1NzI1NDUzNkAxNjgzNzQyNzE4MjE5&el=1_x_2&_esc=publicationCoverPdf)

[Peningkatan Akurasi Algoritma K-Means dengan Clustering Purity Sebagai Titik Pusat Cluster Awal (Centroid)](https://www.researchgate.net/publication/370653667_Peningkatan_Akurasi_Algoritma_K-Means_dengan_Clustering_Purity_Sebagai_Titik_Pusat_Cluster_Awal_Centroid?enrichId=rgreq-d98d10148ad2ca7cc2b65b9399a7769b-XXX&enrichSource=Y292ZXJQYWdlOzM3MDY1MzY2NztBUzoxMTQzMTI4MTE1NzI1NDUzNkAxNjgzNzQyNzE4MjE5&el=1_x_3&_esc=publicationCoverPdf)

**Thesis** · July 2019

CITATIONS

5

READS

48

1. **author:**

[Sujacka Retno](https://www.researchgate.net/profile/Sujacka-Retno?enrichId=rgreq-d98d10148ad2ca7cc2b65b9399a7769b-XXX&enrichSource=Y292ZXJQYWdlOzM3MDY1MzY2NztBUzoxMTQzMTI4MTE1NzI1NDUzNkAxNjgzNzQyNzE4MjE5&el=1_x_5&_esc=publicationCoverPdf)

[Universitas Malikussaleh](https://www.researchgate.net/institution/Universitas_Malikussaleh?enrichId=rgreq-d98d10148ad2ca7cc2b65b9399a7769b-XXX&enrichSource=Y292ZXJQYWdlOzM3MDY1MzY2NztBUzoxMTQzMTI4MTE1NzI1NDUzNkAxNjgzNzQyNzE4MjE5&el=1_x_6&_esc=publicationCoverPdf)

**22** PUBLICATIONS **68** CITATIONS

[SEE PROFILE](https://www.researchgate.net/profile/Sujacka-Retno?enrichId=rgreq-d98d10148ad2ca7cc2b65b9399a7769b-XXX&enrichSource=Y292ZXJQYWdlOzM3MDY1MzY2NztBUzoxMTQzMTI4MTE1NzI1NDUzNkAxNjgzNzQyNzE4MjE5&el=1_x_7&_esc=publicationCoverPdf)

**Some of the authors of this publication are also working on these related projects:**

 Data Mining [View project](https://www.researchgate.net/project/Data-Mining-254?enrichId=rgreq-d98d10148ad2ca7cc2b65b9399a7769b-XXX&enrichSource=Y292ZXJQYWdlOzM3MDY1MzY2NztBUzoxMTQzMTI4MTE1NzI1NDUzNkAxNjgzNzQyNzE4MjE5&el=1_x_9&_esc=publicationCoverPdf)

All content following this page was uploaded by [Sujacka Retno](https://www.researchgate.net/profile/Sujacka-Retno?enrichId=rgreq-d98d10148ad2ca7cc2b65b9399a7769b-XXX&enrichSource=Y292ZXJQYWdlOzM3MDY1MzY2NztBUzoxMTQzMTI4MTE1NzI1NDUzNkAxNjgzNzQyNzE4MjE5&el=1_x_10&_esc=publicationCoverPdf) on 10 May 2023.

The user has requested enhancement of the downloaded file.

**Universitas Sumatera Utara**

**Repositori Institusi USU** [**http://repositori.usu.ac.id**](http://repositori.usu.ac.id/)

Departemen Teknologi Informasi Tesis Magister

2019

Peningkatan Akurasi Algoritma K-Means dengan Clustering Purity Sebagai Titik Pusat Cluster Awal (Centroid)

Retno, Sujacka

Universitas Sumatera Utara

<http://repositori.usu.ac.id/handle/123456789/16782>

*Downloaded from Repositori Institusi USU, Univsersitas Sumatera Utara*

# PENINGKATAN AKURASI ALGORITMA *K-MEANS* DENGAN

***CLUSTERING PURITY* SEBAGAI TITIK PUSAT**

***CLUSTER* AWAL (*CENTROID*) TESIS**

# SUJACKA RETNO 177038001



**PROGRAM STUDI S2 TEKNIK INFORMATIKA FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI**

# UNIVERSITAS SUMATERA UTARA MEDAN

**2019**

PENINGKATAN AKURASI ALGORITMA *K-MEANS* DENGAN

*CLUSTERING PURITY* SEBAGAI TITIK PUSAT

*CLUSTER* AWAL (*CENTROID*) TESIS

Diajukan untuk melengkapi tugas dan memenuhi syarat memperoleh ijazah Magister Teknik Informatika

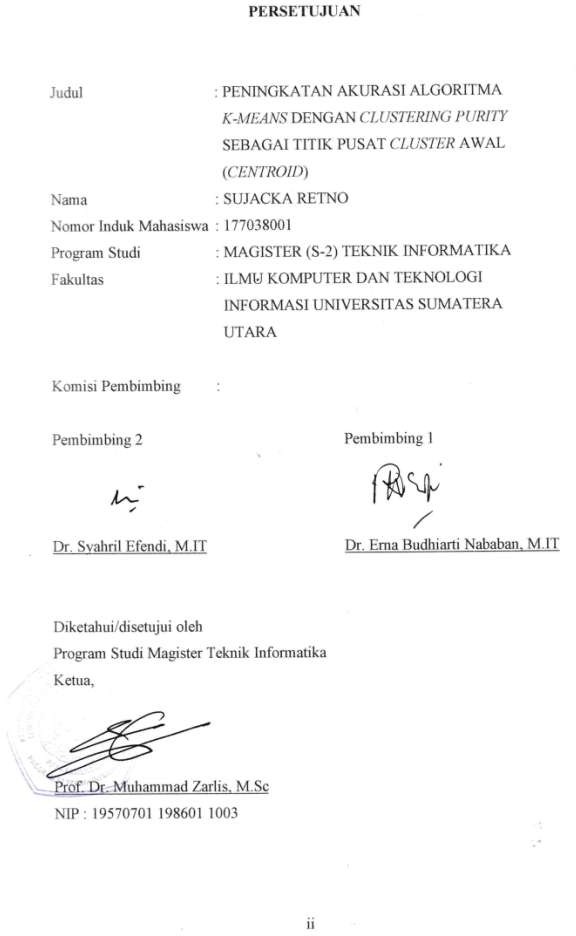
SUJACKA RETNO 177038001

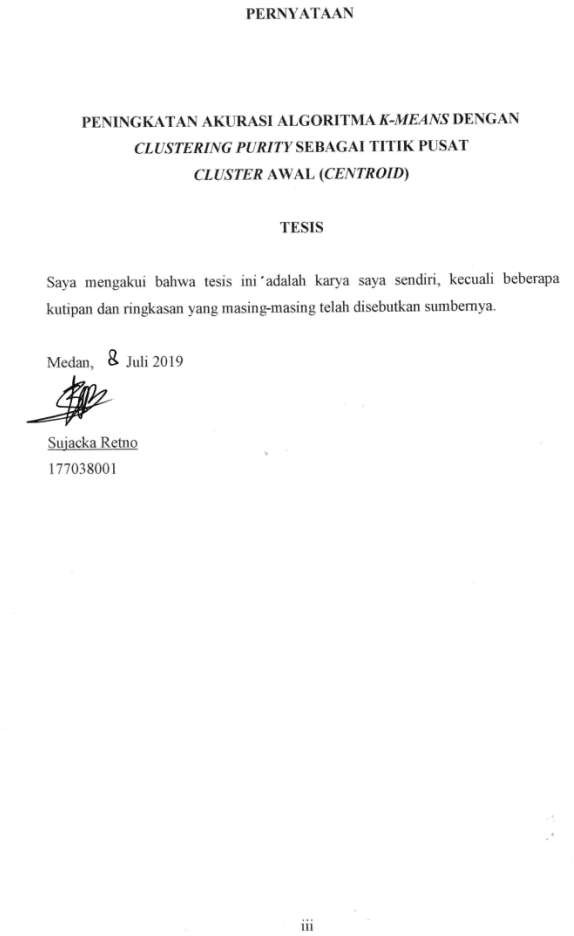


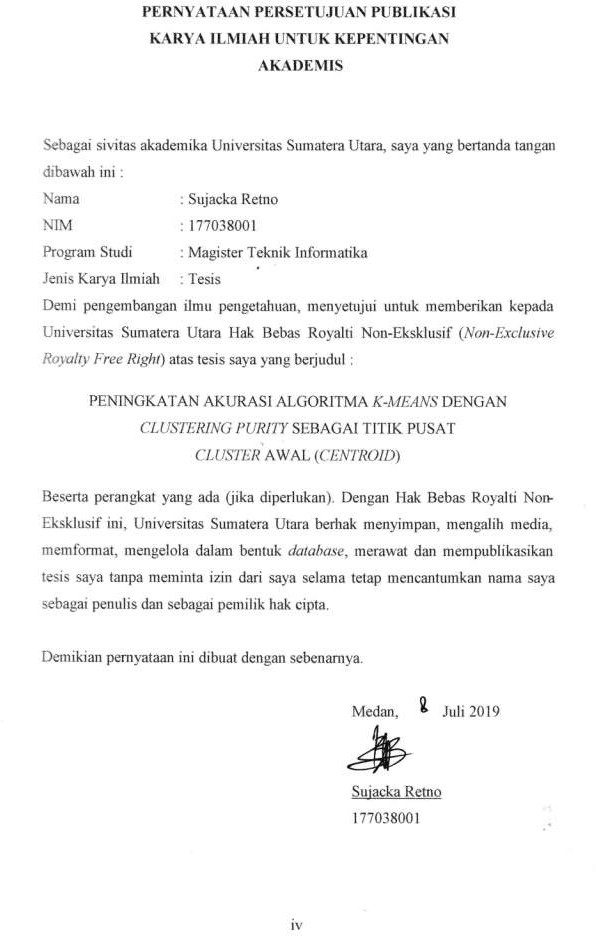
PROGRAM STUDI S2 TEKNIK INFORMATIKA FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS SUMATERA UTARA MEDAN

2019







Telah diuji pada Tanggal : 08 Juli 2019

PANITIA PENGUJI TESIS

Ketua : Dr. Erna Budhiarti Nababan, M.IT Anggota : 1. Dr. Syahril Efendi, M.IT

* 1. Prof. Dr. Muhammad Zarlis
  2. Prof. Dr. Herman Mawengkang

v

## RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap : Sujacka Retno

Tempat dan Tanggal Lahir : Lhokseumawe, 25 Juni 1996

Alamat Rumah : Dusun 1 Loskala, Blang Panyang, Aceh

Telepon/HP 082295574747

Email : [sujackaretno@gmail.com](mailto:sujackaretno@gmail.com) Instansi Tempat Bekerja : -

Alamat Kantor : -

DATA PENDIDIKAN

|  |  |  |
| --- | --- | --- |
| SD | : SD Negeri 2 Muara Dua | Tamat : 2007 |
| SLTP | : SMP Negeri 7 Cunda | Tamat : 2010 |
| SLTA | : SMK Negeri 1 Lhokseumawe | Tamat : 2013 |
| S1 | : Teknik Informatika Universitas Malikussaleh | Tamat : 2017 |
| S2 | : Teknik Informatika Universitas Sumatera Utara | Tamat : 2019 |

## UCAPAN TERIMAKASIH

Puji syukur kehadirat Allah SWT yang telah memberikan anugerah dan rahmat yang luar biasa sehingga penulis dapat menyelesaikan tesis yang berjudul “**Peningkatan Akurasi Algoritma *K-Means* dengan *Clustering Purity* sebagai Titik Pusat *Cluster* Awal (*Centroid*)**”. Juga Shalawat beserta salam kepada Baginda Rasulullah Muhammad, SAW yang telah membawa pedoman kepada sahabat serta seluruh ummat. Penulisan tesis ini merupakan salah satu syarat untuk memperoleh gelar Magister Teknik Informatika pada Program Studi S2- Teknik Informatika Universitas Sumatera Utara.

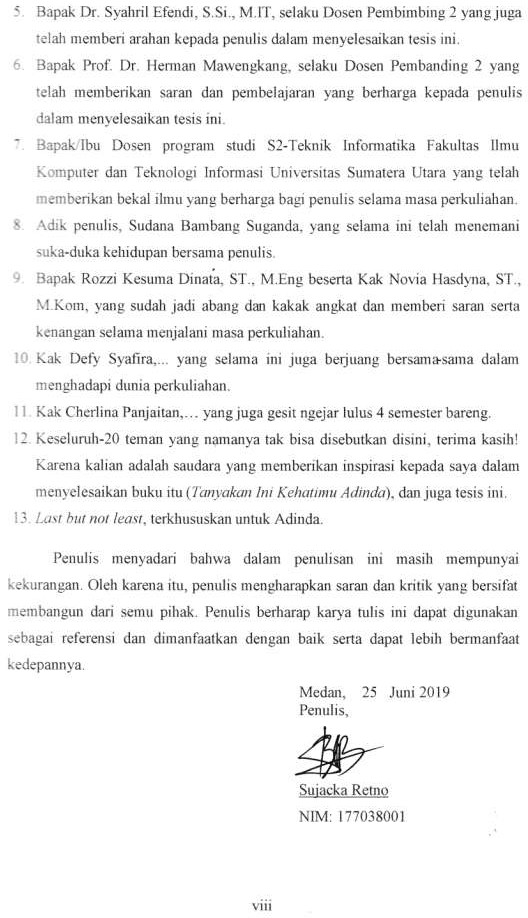
Pada kesempatan ini penulis ingin mengucapkan terima kasih sepenuhnya kepada kedua orang tua saya, Surito, ST., MSM, dan Seri Erlina yang selalu menyertai dan memberikan dukungan baik moril maupun materil kepada penulis, sehingga terselesaikannya proses pendidikan dari program studi S2-Teknik Infomatika, Fakultas Ilmu Komputer dan Teknik Informatika Universitas Sumatera Utara.

Penulis juga tidak lupa mengucapkan terima kasih banyak kepada semua pihak yang telah membantu penulis dalam menyelesaikan tesis ini, diantaranya :

1. Bapak Prof. Dr. Runtung Sitepu, S.H., M.Hum, selaku Rektor Universitas Sumatera Utara.
2. Bapak Prof. Dr. Opim Salim Sitompul, M.Sc, selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara.
3. Bapak Prof. Dr. Muhammad Zarlis, M.Sc, selaku Ketua Program Studi S2- Teknik Infomatika Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara, dan sekaligus juga selaku Dosen Pembanding 1 yang telah memberikan saran kepada penulis untuk menyelesaikan tesis ini.
4. Ibu Dr. Erna Budhiarti Nababan, selaku Dosen Pembimbing 1 yang selama ini tidak mengenal lelah memberikan banyak saran serta pembelajaran yang berharga bagi penulis sebagai arahan untuk menyelesaikan penulisan tesis ini.

vii

viii



## ABSTRAK

*Centroid* awal pada *K-Means Clustering* sangat mempengaruhi hasil dari algoritma *K-Means* itu sendiri. Penelitian ini membahas tentang hasil *clustering* dari *K-Means* dengan titik pusat *cluster* awal (*centroid*) secara acak dan dengan metode *Purity.* Titik pusat *cluster* awal secara acak pada dataset *Iris* menghasilkan jumlah rata-rata iterasi sebanyak 9 iterasi, sedangkan dengan menggunakan *Purity,* jumlah iterasinya hanya 5 iterasi. Tingkat akurasi yang diperoleh dengan menggunakan *K-Means* secara konvensional adalah sebesar 85.5%, sedangkan dengan menggunakan *Purity,* tingkat akurasi yang diperoleh adalah sebesar 88.7%. Metode ini menghasilkan tingkat akurasi sebesar 3.2% lebih baik daripada *K-Means* secara konvensional.

Kata Kunci : *K-Means, Purity, centroid, cluster.*

## ABSTRACT

*Initial centroid in K-Means Clustering greatly affects the results of the K-Means algorithm. This research discusses the results of K-Means Clustering with initial centroid determination with a random and Purity method. The initial of random centroid on the data set iris of 150 data with 4 attributes to analyze the average iterations for the dataset about 9 iterations and the accuracy is 85.5%, whereas the initial centroid using Purity has only 5 iterations and the accuracy is 88.7%. The proposed K-Means methods is 3.2% better than the original K-Means methods*.

Keywords : *K-Means, Purity, centroid, cluster.*

## DAFTAR ISI

Hal

HALAMAN JUDUL i

PERSETUJUAN ii

PERNYATAAN ORISINALITAS iii

PERNYATAAN PERSETUJUAN PUBLIKASI iv

PANITIA PENGUJI v

[RIWAYAT HIDUP vi](#_TOC_250006)

[UCAPAN TERIMAKASIH vii](#_TOC_250005)

[ABSTRAK ix](#_TOC_250004)

[ABSTRACT x](#_TOC_250003)

[DAFTAR ISI xi](#_TOC_250002)

[DAFTAR TABEL xiv](#_TOC_250001)

[DAFTAR GAMBAR xvi](#_TOC_250000)

|  |  |  |
| --- | --- | --- |
| **BAB I** | **PENDAHULUAN** | **1** |
| 1.1 | Latar Belakang | 1 |
| 1.2 | Rumusan Masalah | 3 |
| 1.3 | Batasan Masalah | 3 |
| 1.4 | Tujuan Penelitian | 4 |
| 1.5 | Manfaat Penelitian | 4 |
| **BAB II** | **LANDASAN TEORI** | **5** |
| 2.1 | Data *Mining* | 5 |
| 2.2 | Algoritma *K-Means* | 5 |
| 2.3 | *Purity* | 6 |
| 2.4 | *Davies-Bouldin Index* | 6 |
| 2.5 | Penelitian Terdahulu | 8 |
| 2.6 | Perbedaan Dengan Peneliti Yang Lain | 10 |
| **BAB III** | **METODOLOGI PENELITIAN** | **11** |
| 3.1 | Data Yang Digunakan | 11 |

xi

|  |  |  |
| --- | --- | --- |
| 3.2 | Arsitektur Algoritma Yang Diusulkan | 12 |
| 3.3 | Input Dataset | 13 |
| 3.4 | Inisialisasi Jumlah K | 14 |
| 3.5 | Perhitungan nilai *Purity* Setiap Data | 14 |
| 3.6 | Inisialisasi *Centroid* awal berdasarkan nilai maximum dan |  |
|  | Minimum *Purity* | 17 |
| 3.7 | Perhitungan Jarak Data terhadap *centroid* | 17 |
| 3.8 | Pengelompokan Data Berdasarkan Jarak Minimum | 20 |
| 3.9 | Perhitungan Titik Pusat *Cluster* Baru | 21 |
| 3.10 | Menghitung Nilai *Davies-Bouldin Index* | 22 |
| **BAB IV** | **HASIL DAN PEMBAHASAN** | **25** |
| 4.1 | Hasil Pengujian | 25 |
| 4.2 | Dataset Iris | 25 |
| 4.3 | Nilai *Purity* dataset Iris | 29 |
| 4.4 | Hasil Titik Pusat *Cluster* Dengan Nilai *Purity* | 36 |
| 4.5 | Hasil Pengujian | 43 |
| 4.6 | Jumlah Iterasi pada dataset Iris | 51 |
| 4.7 | Nilai *Purity* pada dataset lain | 52 |
| 4.8 | Hasil *Clustering* pada dataset lain | 54 |
|  | 4.8.1 Jumlah Iterasi Pada *Dataset* 2 | 54 |
|  | 4.8.2 Hasil *Clustering* Pada *Dataset* 2 | 55 |
|  | 4.8.3 Jumlah Iterasi Pada *Dataset* 3 | 56 |
|  | 4.8.4 Hasil *Clustering* Pada *Dataset* 3 | 56 |
|  | 4.8.5 Jumlah Iterasi Pada *Dataset* 4 | 57 |
|  | 4.8.6 Hasil *Clustering* Pada *Dataset* 3 | 58 |
| 4.9 | Hasil Evaluasi *Clustering* | 59 |
| 4.10 | Perbandingan Hasil Evaluasi *Clustering* | 62 |
| **BAB V** | **KESIMPULAN DAN SARAN** | **65** |
| 4.1 | Kesimpulan | 65 |
| 5.2 | Saran | 67 |

xii

## DAFTAR PUSTAKA LAMPIRAN

xiii

xii

## DAFTAR TABEL

|  |  |  |
| --- | --- | --- |
|  | | Hal |
| Tabel 2.1 | Penelitian Terdahulu | 9 |
| Tabel 2.2 | Penelitian Terdahulu (lanjutan) | 10 |
| Tabel 3.1 | Data Sampel dari Dataset *User Knowledge Modeling* | 14 |
| Tabel 3.2 | Hasil Perhitungan *Purity* setiap data | 16 |
| Tabel 3.3 | Titik Pusat *cluster* (*Centroid*) awal | 17 |
| Tabel 3.4 | Perhitungan jarak data dengan titik pusat *cluster* | 20 |
| Tabel 3.5 | Pengelompokan data berdasarkan jarak minimum | 20 |
| Tabel 3.6 | Pengelompokan data berdasarkan jarak minimum (lanjutan) | 21 |
| Tabel 3.7 | Perhitungan titik pusat *cluster* baru pada *cluster* 1 | 21 |
| Tabel 3.8 | Perhitungan titik pusat *cluster* baru pada *cluster* 2 | 21 |
| Tabel 3.9 | Titik pusat *cluster* (*Centroid*) baru | 22 |
| Tabel 3.10 | Ilustrasi *sample* yang telah di-*clustering* | 22 |
| Tabel 3.11 | Ilustrasi *sample* yang telah di-*clustering* | 23 |
| Tabel 3.12 | Ilustrasi titik pusat *cluster* (*Centroid*) akhir | 23 |
| Tabel 3.13 | Nilai rasio dari setiap *cluster* | 24 |
| Tabel 4.1 | Ringkasan Dataset Yang Digunakan | 25 |
| Tabel 4.2 | Dataset Iris | 26 |
| Tabel 4.3 | Dataset Iris (lanjutan) | 27 |
| Tabel 4.4 | Dataset Iris (lanjutan) | 28 |
| Tabel 4.5 | Dataset Iris (lanjutan) | 29 |
| Tabel 4.6 | Hasil pehitungan *Purity* setiap data | 31 |
| Tabel 4.7 | Hasil pehitungan *Purity* setiap data (lanjutan) | 32 |
| Tabel 4.8 | Hasil pehitungan *Purity* setiap data (lanjutan) | 33 |
| Tabel 4.9 | Hasil pehitungan *Purity* setiap data (lanjutan) | 34 |
| Tabel 4.10 | Hasil pehitungan *Purity* setiap data (lanjutan) | 35 |
| Tabel 4.11 | Titik Pusat *cluster* (*Centroid*) awal | 35 |
| Tabel 4.12 | Perhitungan jarak data dengan titik pusat *cluster* | 38 |

xiv

|  |  |  |
| --- | --- | --- |
| Tabel 4.13 | Perhitungan jarak data dengan titik pusat *cluster* (lanjutan) | 39 |
| Tabel 4.14 | Perhitungan jarak data dengan titik pusat *cluster* (lanjutan) | 40 |
| Tabel 4.15 | Perhitungan jarak data dengan titik pusat *cluster* (lanjutan) | 41 |
| Tabel 4.16 | Perhitungan jarak data dengan titik pusat *cluster* (lanjutan) | 42 |
| Tabel 4.17 | Hasil Akhir Pengujian *Purity K-Means* pada dataset Iris | 43 |
| Tabel 4.18 | Hasil Akhir Pengujian *Purity K-Means* pada dataset Iris |  |
|  | (lanjutan) | 44 |
| Tabel 4.19 | Hasil Akhir Pengujian *Purity K-Means* pada dataset Iris |  |
|  | (lanjutan) | 45 |
| Tabel 4.20 | Hasil Akhir Pengujian *Purity K-Means* pada dataset Iris |  |
|  | (lanjutan) | 46 |
| Tabel 4.21 | Hasil Akhir Pengujian *Purity K-Means* pada dataset Iris |  |
|  | (lanjutan) | 47 |
| Tabel 4.22 | Perbandingan hasil *clustering* pada dataset *Iris* | 47 |
| Tabel 4.23 | Perbandingan hasil *clustering* pada dataset *Iris* (lanjutan) | 48 |
| Tabel 4.24 | Perbandingan hasil *clustering* pada dataset *Iris* (lanjutan) | 49 |
| Tabel 4.25 | Perbandingan hasil *clustering* pada dataset *Iris* (lanjutan) | 50 |
| Tabel 4.26 | Perbandingan hasil *clustering* pada dataset *Iris* (lanjutan) | 51 |
| Tabel 4.27 | Hasil Perhitungan Nilai *Purity* pada dataset 2 | 52 |
| Tabel 4.28 | Hasil Perhitungan Nilai *Purity* pada dataset 3 | 53 |
| Tabel 4.29 | Hasil Perhitungan Nilai *Purity* pada dataset 4 | 53 |
| Tabel 4.30 | Perbandingan Nilai *Davies Bouldin Index* pada *dataset* 2 | 60 |
| Tabel 4.31 | Perbandingan Nilai *Davies Bouldin Index* pada *dataset* 3 | 60 |
| Tabel 4.32 | Perbandingan Nilai *Davies Bouldin Index* pada *dataset* 3 |  |
|  | (lanjutan) | 61 |
| Tabel 4.33 | Perbandingan Nilai *Davies Bouldin Index* pada *dataset* 4 | 61 |
| Tabel 4.34 | Hasil evaluasi *clustering* berdasarkan jumlah Iterasi | 62 |
| Tabel 4.35 | Hasil evaluasi *clustering* berdasarkan *Davies Bouldin Index* | 63 |

xv

## DAFTAR GAMBAR

|  |  |  |
| --- | --- | --- |
|  |  | Hal |
| Gambar 3.1 | Framework peningkatan akurasi K-Means dengan Purity | 12 |
| Gambar 4.1 | Grafik hasil *clustering* pada *dataset Iris* | 51 |
| Gambar 4.2 | Grafik hasil *clustering* pada *dataset* 2 | 54 |
| Gambar 4.3 | Grafik hasil *clustering* pada *dataset* 2 | 55 |
| Gambar 4.4 | Grafik hasil *clustering* pada *dataset* 3 | 56 |
| Gambar 4.5 | Grafik hasil *clustering* pada *dataset* 3 | 57 |
| Gambar 4.6 | Grafik hasil *clustering* pada *dataset* 4 | 58 |
| Gambar 4.7 | Grafik hasil *clustering* pada *dataset* 4 | 59 |
| Gambar 4.8 | Grafik perbandingan rata-rata jumlah iterasi pada keempat |  |
|  | *dataset* | 63 |
| Gambar 4.9 | Grafik perbandingan rata-rata nilai DBI pada ketiga *dataset* | 64 |

xvi

## BAB I PENDAHULUAN

1

**1.1. Latar Belakang**

*Clustering* adalah proses pengelompokan objek data menjadi beberapa *cluster* yang terpisah sehingga data yang ada di dalam masing-masing *cluster* tersebut menjadi sebuah kelompok data yang memiliki kemiripan yang relatif sama. *Cluster* adalah kumpulan objek data yang karakteristiknya sama satu sama lain yang ada di dalam *cluster* yang sama dan data tersebut memiliki karakteristik yang berbeda dengan objek data yang ada di *cluster* yang lain. Ada banyak teknik yang dapat digunakan untuk proses *clustering* seperti *Single Linkage, Complete Linkage, Average Linkage, Fuzzy C-Means, Kohonen SOM, LVQ dan K-Means. K-means* adalah salah satu teknik *clustering* tertua dan yang paling umum digunakan untuk proses *clustering*. Di dalam metode ini, jumlah cluster (k) telah ditentukan sebelumnya sebelum dianalisis dan kemudian pemilihan *centroid* awal akan ditentukan secara acak dan diikuti dengan proses iteratif untuk menentukan jarak dari setiap titik data ke *centroid* terdekatnya. Proses ini akan terus berulang hingga tidak ada anggota *cluster* yang berpindah ke *cluster* yang lain (Unnati & Chaita, 2016).

Titik pusat *cluster* (*centroid*) yaitu titik awal dimulainya sebuah pengelompokan didalam *cluster* pada algoritma *K-Means.* Pengelompokan data dilakukan dengan menghitung jarak terdekat dengan titik pusat *cluster* awal sebagai titik tengah dalam pembentukan kelompok atau *cluster*. Hasil output dari proses *clustering* menggunakan metode *K-Means* sangat bergantung pada penentuan titik pusat awal *cluster* (*centroid*). Penentuan titik pusat awal *cluster* (*centroid*) menggunakan berbagai metode dapat meningkatkan akurasi dari *output* yang proses *clustering* dengan menggunakan metode *K-Means* (Sirait, et al*.* 2016).

1

Pada penelitian (Hartono, et al. 2015), *Genetic Algorithm* (GA) digunakan untuk menentukan titik pusat *cluster* pada algoritma *K-Means*. Penelitian tersebut diuji pada satu dataset yaitu *Iris Dataset* dan dari pengujian tersebut diperoleh hasil *Average Mean Square Error* yang lebih baik dibandingkan dengan penentuan titik pusat *cluster* secara acak.

Qiang Zhan (2017), didalam penelitiannya membahas metode yang digunakan untuk menentukan titik pusat *cluster* awal dengan menggunakan *Structure Features.* Dari penelitian tersebut, metode tersebut memperoleh hasil yang lebih akurat dengan menggunakan *Minkowski Distance* untuk perhitungan jarak data terhadap *centroid* tersebut. Sedangkan Rishikesh (2016), di dalam penelitiannya menerapkan metode *Min-Max Normalization* untuk menentukan titik pusat *cluster* awal (*centroid*). Hasil dari penelitian tersebut didapatkan peningkatan waktu eksekusi pengelompokan untuk data yang variant.

Sirait, et al (2016), didalam penelitiannya menggunakan metode *KD-Tree* untuk menentukan titik pusat *cluster* awal pada algoritma *K-Means.* Hasil dari penelitian tersebut diperoleh hasil *Sum of Square Error* sebesar 214,37 untuk dataset *GPA.* Sedangkan Rose, et al (2014), didalam penelitiannya menggunakan *Modified Centroid Selection Method* pada algoritma *K-Means* memperoleh akurasi yang sama dengan *K-Means* secara konvensional.

Pada penelitian yang dilakukan oleh Bernard (2018), dengan penentuan titik pusat *cluster* awal algoritma *K-Means* menggunakan *Sum of Squared Error* diperoleh hasil kualitas *clustering* dengan nilai *Davies-Bouldin Index* sebesar 0.0378, atau 3,78% lebih sedikit dibandingkan dengan metode *K-Means* secara konvensional.

Sementara pada penelitian ini, titik pusat *cluster* awal ditentukan dengan menghitung nilai *Purity* dari setiap data yang akan diuji. Data yang memperoleh nilai *purity* tertinggi dan *purity* terendah akan digunakan sebagai titik pusat *cluster* awal untuk proses *clustering* algoritma *K-Means*. Untuk membuktikan bahwa adanya perbedaan pada sistem penyelesaian metode tersebut, maka akan dilakukan uji *Davies-Boulding Index* (DBI) untuk proses *clustering.*

*Davies-Bouldin index* adalah salah satu metode yang bertujuan untuk mengevaluasi proses *clustering* pada suatu metode peng-*cluster*an yang didasarkan pada nilai *kohesi* dan *separasi*. Dalam suatu pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap *centroid* dari *cluster* yang diikuti. Sedangkan separasi didasarkan pada jarak antar *centroid* dari *cluster* tersebut (Nurul, et al 2015). Hasil *clustering* yang diperoleh dari penentuan titik pusat *cluster* yang diusulkan akan di evaluasi dengan metode *Davies-Bouldin Index* (DBI). Sehingga dapat diketahui korelasi antar iterasi pada proses *clustering* dari metode penentuan titik pusat *cluster* berdasarkan *Purity* terhadap peningkatan kualitas *cluster* berdasarkan nilai *Davies-Bouldin Index* (DBI) yang diperoleh.

## 1.2. Rumusan Masalah

Titik pusat *cluster* awal yang ditentukan secara acak pada algoritma *K-Means*, akan menghasilkan nilai akurasi yang bervariasi dalam setiap proses peng-*cluster*an yang dilakukan. Oleh karena itu, perlu adanya suatu pendekatan dalam menentukan titik pusat *cluster* awal untuk digunakan sebagai titik pusat *cluster* awal pada proses *clustering* algoritma *K-Means* untuk menghasilkan nilai akurasi yang baik.

## 1.3. Batasan Masalah

Beberapa batasan masalah pada penelitian ini adalah sebagai berikut :

1. *Dataset* yang digunakan dalam penelitian ini yaitu
   1. *Dataset User Knowledge Modeling,* yang berjumlah 403 data dengan 5 atribut numerik yang diperoleh dari *UCI Machine Learning Repository.*
   2. *Dataset Birth and Death Rates,* yang berjumlah 70 data dengan 2 atribut numerik yang diperoleh dari *Harrigan Dataset.*
   3. *Dataset Wholesale Customers*, yang berjumlah total keseluruhan 440 data dengan 8 atribut numerik, yang diperoleh dari *UCI Machine Learning Repository*.
   4. *Iris,* yang berjumlah total 150 data dengan 4 atribut numerik, yang diperoleh dari *UCI Machine Learning Repository.*
2. Jumlah *cluster* pada penelitian ini ada 2, yaitu 2 *cluster* dan 3 *cluster* dan batas iterasi dari proses *clustering* adalah 100 iterasi.

## 1.4. Tujuan Penelitian

Tujuan penelitian ini adalah untuk menganalisa jumlah iterasi yang dibutuhkan untuk proses *clustering* mencapai konvergen dan menganalisa proses *clustering* dalam proses peng-*cluster*an algoritma *K-Means* dengan menentukan titik pusat *cluster* awal menggunakan *Purity* dan membandingkan hasil evaluasi *clustering* tersebut dengan metode *K-Means* secara konvensional melalui pengukuran nilai *Davies Bouldin Index* (DBI).

## 1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah untuk mengetahui seberapa besar peningkatan hasil evaluasi *clustering* yang diperoleh dari penentuan titik pusat *cluster* awal berdasarkan nilai *Purity* yang tertinggi dan terendah dan mengetahui perbandingan hasil evaluasi *clustering* antara metode penentuan titik pusat *cluster* awal secara konvensional dengan penentuan titik pusat *cluster* awal yang diusulkan. Sehingga dapat memberikan masukan dalam ilmu pengetahuan untuk penelitian selanjutnya dalam pengembangan ilmu penyelesaian *clustering* algoritma *K-Means*.

## BAB II LANDASAN TEORI

### Data Mining

Di era sekarang ini, informasi menjadi sumber daya yang memegang pengaruh penting dalam berbagai hal. Banyak organisasi yang telah mengumpulkan dan menyimpan data yang berskala besar. Namun, sebagian dari mereka tidak dapat menemukan informasi berharga yang tersembunyi dari banyaknya data tersebut, sehingga organisasi-organisasi tersebut memilih untuk menggunakan teknologi informasi untuk memanajemen data-data yang telah tertimbun. Pada dasarnya *data mining* adalah proses yang menggunakan alat (*tools*) untuk mengekstraksi atau menggali informasi-informasi yang berguna dari *dataset* yang besar; *data mining* adalah sebuah proses untuk menggali informasi di dalam database yang menghasilkan sesuatu yang berguna dari data tersebut. (Silwattananusarn & Tuamsuk 2012).

* 1. **Algoritma *K-Means***

Secara umum, *K-Means* adalah sebuah algoritma untuk proses *clustering. Clusters* sangat bergantung pada inisialisasi nilai titik pusat awal (*centroid*). Adapun langkah- langkah dari algoritma *K-Means* adalah sebagai berikut (Unnati & Chaita 2016) :

1. Tentukan jumlah *K*, *K* adalah jumlah *cluster.*
2. Tentukan nilai awal titik pusat *cluster* untuk dilakukannya proses *clustering*.
3. Hitung *distance measure* (jarak data) terhadap masing-masing *centroid*

menggunakan *Euclidean Distance.*

1. Alokasikan seluruh objek data yang telah dihitung ke dalam masing-masing

*cluster.*

1. Tentukan *centroid* baru dengan menggunakan persamaan berikut:

1 𝑛 (2.1)

𝑣̅ 𝑖𝑗 = 𝑁 ∑ 𝑥𝑘𝑗

𝑖

𝑘=0

5

Keterangan:

𝑣ij = *centroid/*rata-rata *cluster* ke-i untuk variabel ke-j *Ni* = jumlah data yang menjadi anggota *cluster ke-i i,k* = indeks dari *cluster*

*j* = indeks dari variabel

*xkj* = nilai data ke-*k* yang ada di dalam *cluster* tersebut untuk variabel ke-*j*

1. Ulang kembali langkah 3, 4 dan 5 hingga tidak ada lagi anggota *cluster* yang berpindah ke *cluster* lain.

### Purity

Sebuah *cluster* dikatakan murni (*pure*) apabila semua objek data dengan *class* yang sama berada pada *cluster* yang sama pula (Nurul, et al 2015). *Purity* digunakan untuk menghitung kemurnian suatu *cluster* yang direpresentasikan sebagai anggota *cluster* yang paling cocok pada suatu *class*. Untuk menghitung *purity* setiap *cluster*, digunakan rumus berikut:

𝑃𝑢𝑟𝑖𝑡(𝑗) = 1 𝑚𝑎𝑥(𝑛 )

Keterangan:

𝑁𝑗

𝑖𝑗

(2.2)

Purity(j) = nilai *purity* untuk variabel ke-j

*Nj* = jumlah data yang menjadi anggota *cluster ke-j i,j* = indeks dari *cluster*

### Davies-Bouldin Index

*Davies-Bouldin index* adalah salah satu metode yang bertujuan untuk mengevaluasi *cluster* pada suatu metode *clustering* yang didasarkan pada nilai kohesi dan separasi. Dalam suatu pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap *centroid* dari *cluster* yang diikuti. Sedangkan separasi didasarkan pada jarak antar *centroid* dari *cluster* tersebut (Nurul, et al 2015). Adapun langkah-langkah untuk menghitung *Davies-Bouldin Index* adalah sebagai berikut:

1. *Sum of Square Within cluster* (SSW)

*Sum of Square Within cluster* (SSW) merupakan persamaan yang digunakan untuk mengetahui matriks kohesi dalam sebuah *cluster* ke-*i* yang dirumuskan sebagai berikut :

𝑚𝑖

1

(2.3)

𝑆𝑆𝑊𝑖 = 𝑚𝑖 ∑ 𝑑(𝑥𝑗, 𝑐𝑖)

𝑗 =𝑖

Dari persamaan tersebut, *mi* merupakan jumlah data dalam *cluster* ke-*i*, *ci* adalah *centroid cluster* ke-*i*, dan *d*(*xj,ci*) merupakan jarak setiap data terhadap *centroid* yang dihitung menggunakan rumus.

1. *Sum of Square Between cluster* (SSB)

*Sum of Square Between cluster* (SSB) merupakan persamaan yang digunakan untuk mengetahui separasi antar *cluster* yang dihitung menggunakan persamaan sebagai berikut:

1. *Ratio* (Rasio)*.*

𝑆𝑆𝐵𝑖, 𝑗 = 𝑑(𝑐𝑖, 𝑐𝑗)

(2.4)

Setelah nilai SSW dan SSB diperoleh, kemudian dilakukan pengukuran rasio (𝑅𝑖𝑗) untuk mengetahui nilai perbandingan antara *cluster* ke-*i* dan *cluster* ke-*j*. Nilai rasio dihitung menggunakan persamaan sebagai berikut :

𝑅𝑖𝑗 = 𝑆𝑆𝑊𝑖 + 𝑆𝑆𝑊𝑗

𝑆𝑆𝐵𝑖𝑗

(2.5)

1. *Davies-Bouldin Index* (DBI)

Nilai rasio yang diperoleh tersebut digunakan untuk mencari nilai *Davies-Bouldin Index* (DBI) dengan menggunakan persamaan berikut :

1 𝑘 (2.6)

𝐷𝐵𝐼 = 𝑘 ∑ 𝑚𝑎𝑥𝑖≠𝑗 ( 𝑅𝑖,𝑗 )

𝑖=1

Dari persamaan tersebut, *k* merupakan jumlah *cluster* yang digunakan. Semakin kecil nilai DBI yang diperoleh (non-negatif >= 0), maka semakin baik *cluster* yang diperoleh dari pengelompokan *K-Means* yang digunakan (Qiang Zhan 2017).

## Penelitian Terdahulu

Dalam melakukan penelitian ini, penulis menggunakan beberapa penelitian terdahulu yang relevan dengan permasalahan yang akan diteliti yang berguna sebagai acuan agar penelitian ini dapat berjalan dengan lancar.

Sirait, et al (2016), pada penelitiannya menerapkan metode *KD-Tree* untuk menentukan titik pusat *cluster* awal (*centroid*). Penelitian tersebut diuji menggunakan dataset *Quality* dan GPA*.* Nilai *Sum of Square Error* yang diperoleh pada metode yang diusulkan tersebut adalah sebesar 504302 untuk dataset *Quality* dan 214,37 untuk dataset GPA, namun tidak dijelaskan jumlah iterasi yang dihasilkan untuk mencapai konvergen dan nilai DBI untuk proses *clustering* yang diperoleh saat proses *clustering* dilakukan.

Rishikesh (2016), di dalam penelitiannya menerapkan metode *Min-Max Normalization* untuk menentukan titik pusat *cluster* awal (*centroid*). Penelitian tersebut diuji pada 5 dataset yaitu *Transfusion dataset, Wavesurge Dataset, IQitems dataset, ds1.10 Dataset,* dan *Bfi Dataset.* Hasil dari penelitian tersebut diperoleh peningkatan waktu eksekusi pengelompokan untuk data yang variant (beragam) dibandingkan dengan algoritma *K-Means* secara konvensional.

Pada penelitian (Hartono, et al. 2018), *Perceptron* digunakan untuk menentukan titik pusat *cluster* pada algoritma *K-Means*. Pada penelitian tersebut, titik pusat *cluster* ditentukan dengan menghitung *Weight Value*. Penelitian tersebut diuji pada dataset *Balanced Scale* dan *Abalone*. Hasil penelitian tersebut menunjukkan metode tersebut mampu mengurangi masalah *Class Imbalance,* namun tidak dijelaskan korelasi antar iterasi pada proses *clustering* pada penelitian tersebut.

Sedangkan penelitian (Unnati, et al. 2016), nilai *Mean* dari keseleruhuan data digunakan untuk menentukan titik pusat *cluster* pada algoritma *K-Means.* Penelitian tersebut diuji pada 3 dataset yaitu *Bank Dataset, Credit Dataset,* dan *Adult Dataset.* Pada penelitian tersebut, hasil yang diperoleh dengan menggunakan metode tersebut

mampu mengurangi kompleksitas waktu dalam menyelesaikan proses *clustering,*

namun tidak dijelaskan jumlah iterasi yang dihasilkan untuk mencapai konvergen.

Pada penelitian yang dilakukan oleh Bernard (2018), dengan penentuan titik pusat *cluster* awal algoritma *K-Means* menggunakan *Sum of Squared Error* diperoleh hasil kualitas *clustering* dengan nilai *Davies-Bouldin Index* sebesar 0.0378, atau 3,78% lebih sedikit dibandingkan dengan metode *K-Means* secara konvensional.

Pada penelitian yang dilakukan oleh Qiang Zhan (2017), di dalam penelitiannya membahas metode yang digunakan untuk menentukan titik pusat *cluster* awal dengan menggunakan *Structure Features.* Dari penelitian tersebut, nilai DBI yang diperoleh dengan menggunakan *K-Means* yang diajukan adalah sebesar 0.683, sedangkan dengan menggunakan *K-Means* secara konvensional nilai DBI yang diperoleh adalah 0.666, atau selisih sebesar 0.017.

Adapun *summary* dari penelitian tersebut dapat dilihat pada table 2.1 sebagai berikut:

Tabel 2.1 Penelitian Terdahulu

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Nama Peneliti dan Tahun** | **Metode yang digunakan** | **Keterangan** |
| **1** | Sirait, at el. 2016 | *K-Means, KD-Tree* | Diperoleh SSE yang lebih baik namun tidak dijelaskan nilai DBI yang diperoleh untuk setiap iterasi pada  proses *clustering.* |
| **2** | Rishikesh. 2016 | *K-Means, Min-Max Normalization* | Hasil dari penelitian tersebut diperoleh peningkatan waktu eksekusi pengelompokan untuk data yang variant (beragam) dibandingkan dengan algoritma *K-Means*  secara konvensional |
| **3** | Hartono, at el. 2018 | *K-Means, Genetic Algorithm* | Hasil penelitian tersebut menunjukkan metode tersebut mampu mengurangi masalah *Class Imbalance,* namun tidak dijelaskan korelasi antar iterasi pada  proses *clustering* pada |

Tabel 2.2 Penelitian Terdahulu (lanjutan)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | penelitian tersebut. |
| **4** | Unnati, at el. 2016 | *K-Means, Mean* | Pada penelitian tersebut, hasil yang diperoleh dengan menggunakan metode tersebut mampu mengurangi kompleksitas waktu dalam menyelesaikan proses *clustering,* namun tidak dijelaskan jumlah iterasi  yang dihasilkan untuk mencapai konvergen |
| **5** | Bernard. 2018 | *K-Means, Sum of Squared Error* | Diperoleh hasil kualitas *clustering* dengan nilai *Davies-Bouldin Index* sebesar 0.0378, atau 3,78% lebih sedikit dibandingkan dengan metode *K-Means*  secara konvensional |
| **6** | Qiang Zhan. 2015 | *K-means, Structure Feature* | Dari penelitian tersebut, nilai DBI yang diperoleh dengan menggunakan *K-Means* yang diajukan adalah sebesar 0.683, sedangkan dengan menggunakan *K-Means* secara konvensional nilai DBI yang diperoleh adalah 0.666, atau selisih sebesar  0.017. |

## Perbedaan Dengan Peneliti Terdahulu

Pada penelitian ini peneliti menggunakan *Purity* yang akan digunakan sebagai titik pusat *cluster* awal (centroid) dalam proses *clustering* pada metode *K-Means.* Diharapkan dapat diketahui korelasi antar iterasi pada setiap proses *clustering* dari metode yang diusulkan terhadap peningkatan kualitas *cluster* berdasarkan nilai *Davies-Bouldin Index* (DBI) yang diperoleh.

## BAB III METODOLOGI PENELITIAN

* 1. **Data Yang Digunakan**

Data yang digunakan pada penelitian ini adalah beberapa *dataset* yang diperoleh dari berbagai sumber. Data tersebut direpresentasikan menjadi sebuah *file* yang berformat *text* (txt). Data yang digunakan pada penelitian ini adalah sebagai berikut :

1. *User Knowledge Modeling*

*User Knowledge Modeling* adalah *dataset* yang berisi tentang status pengetahuan pelajar tentang subjek *Electrical DC Machines. User Knowledge Modeling* ini terdiri dari 403 data dengan 5 attribut data (fitur) yaitu, *STG (The degree of study time for goal object materials), SCG, STR, LPR* dan *PEG. Dataset* ini diambil dari ([*https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling*](https://archive.ics.uci.edu/ml/datasets/User%2BKnowledge%2BModeling)*).*

1. *Birth and Death Rates*

*Birth and Death Rates* adalah *dataset* yang berisi angka kelahiran dan kematian pada 1000 orang di beberapa negara pada tahun 1966*. Dataset* ini terdiri dari 70 data dengan 2 attribut data (fitur) yaitu, angka kelahiran dan angka kematian yang bersumber dari ([*http://people.sc.fsu.edu/~jburkardt/datasets/hartigan/file26.txt*](http://people.sc.fsu.edu/~jburkardt/datasets/hartigan/file26.txt)*).*

1. *Wholesale Customers*

*Wholesale Customers* adalah dataset yang berisi tentang pengeluaran tahunan dalam satuan dari beberapa macam produk dari distributor. Dataset ini memiliki 440 jumlah data numerik dengan 8 buah attribute data (fitur), akan tetapi pada penelitian ini hanya menggunakan 6 buah attribut data (fitur) saja yaitu, *fresh, milk, grocery, frozen, detergents,* dan *delicaressen*. Data ini diperoleh dari ([*https://archive.ics.uci.edu/ml/datasets/Wholesale+customers*](https://archive.ics.uci.edu/ml/datasets/Wholesale%2Bcustomers)*).*

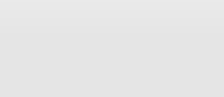
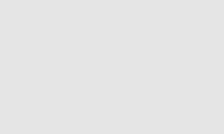
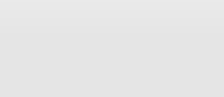
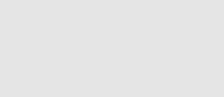
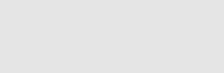
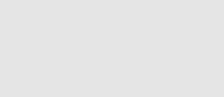
11

1. *Iris*

Dataset ini memiliki 150 jumlah data numerik dengan 4 attribut data (fitur),. Data ini diperoleh dari ([*https://archive.ics.uci.edu/ml/datasets/Iris*](https://archive.ics.uci.edu/ml/datasets/Iris)*).*

## Arsitektur Algoritma Yang Diusulkan

Penentuan titik pusat *cluster* awal pada algoritma *K-Means* merupakan hal yang sangat penting untuk menentukan kualitas *cluster* yang dihasilkan. Metode penentuan titik pusat *cluster* awal yang diusulkan pada penelitian ini adalah dengan mencari nilai *Purity* dari keseluruhan data terlebih dahulu yang kemudian data yang memiliki nilai *Purity* tertinggi dan terendah yang akan digunakan untuk menentukan titik pusat *cluster* awal pada algoritma *K-Means.* Arsitektur algoritma *K-Means* dengan metode penentuan titik pusat *cluster* awal yang diusulkan, dapat dilihat pada gambar 3.1.



Menghitung nilai

*Davies-Bouldin Index*

Menghitung Titik Pusat Cluster Baru

Input Dataset

Menghitung Jarak Data terhadap Centroid

Pengelompokan data berdasarkan jarak minimum

Mengitung nilai Purity Setiap Data

Inisialisasi Centroid awal berdasarkan *nilai maximum dan minimum Purity*

Inisialisasi Jumlah K

**Hasil Akurasi Clustering**

**Gambar 3.1. *Framework* peningkatan akurasi K-Means dengan *Purity***

Berdasarkan gambar 3.1, langkah pertama yang akan dilakukan adalah menginput dataset sebagai bahan yang akan digunakan dalam pengolahan data. Kemudian menentukan berapa *cluster* (K) data tersebut akan dikelompokkan. Dan berikutnya adalah menghitung nilai *Purity* dari setiap data yang akan diproses untuk mencari data yang memiliki nilai *Purity* tertinggi dan terendah yang nantinya akan digunakan sebagai titik pusat *cluster* awal. Langkah selanjutnya adalah mulai menghitung jarak data dengan titik pusat *cluster* awal dengan rumus *Euclidean Distance.* Setelah perhitungan jarak tersebut selesai, maka selanjutnya adalah pengelompokkan data berdasarkan jarak minimumnya. Setelah pengelompokkan selesai, maka akan dilakukan perhitungan ulang titik pusat *cluster* dengan mencari nilai tengah (*mean*) dari setiap anggota masing-masing *cluster* untuk mendapatkan titik pusat *cluster* yang baru menggunakan persamaan 2.1.

Langkah selanjutnya adalah menghitung ulang jarak data dengan *centroid* baru tersebut dan dilakukan pengelompokan kembali. Setelah tahap pengelompokan ini, dilakukan pengecekan apakan tidak ada lagi anggota *cluster* yang berpindah ke *cluster* yang lain. Jika masih ada, maka akan dilakukan perhitungan pada iterasi selanjutnya. Kondisi ini akan berhenti apabila tidak ada lagi anggota *cluster* yang berpindah ke *cluster* yang lain, atau akan berhenti secara otomatis apabila telah sampai pada ambang batas iterasi.

Setelah pengelompokan akhir dilakukan, maka langkah selanjutnya adalah melakukan perhitungan nilai *Davies-Bouldin Index* pada setiap *cluster* dalam setiap iterasi pada proses *clustering* untuk menganalisa hasil dari model *clustering* yang diperoleh. Index yang digunakan untuk pengujian adalah dengan menggunakan *Davies-Bouldin Index* (DBI).

## Input Dataset

Untuk mendukung kelancaran penelitian, berikut akan dipaparkan contoh untuk perhitungan manual dari proses *clustering* digunakan satu data sampel dengan jumlah 10 data didalamnya. Data sampel tersebut dapat dilihat pada tabel 3.1.

Tabel 3.1 Data Sampel dari Dataset *User Knowledge Modeling*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data ke-*i* (*pi*)** | **Atribut 1 (*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** |
| **1** | 0.08 | 0.08 | 0.24 | 0.9 |
| **2** | 0.06 | 0.06 | 0.25 | 0.33 |
| **3** | 0.1 | 0.1 | 0.65 | 0.3 |
| **4** | 0.08 | 0.08 | 0.98 | 0.24 |
| **5** | 0.09 | 0.15 | 0.1 | 0.66 |
| **6** | 0.1 | 0.1 | 0.29 | 0.56 |
| **7** | 0.15 | 0.02 | 0.4 | 0.01 |
| **8** | 0.2 | 0.14 | 0.72 | 0.25 |
| **9** | 0 | 0 | 0.2 | 0.85 |
| **10** | 0.18 | 0.18 | 0.3 | 0.81 |

## Inisialisasi Jumlah K

Setelah data di-*input* maka selanjutnya adalah penentuan jumlah *cluster* (K)*.* Jumlah *cluster* merupakan jumlah dari pembagian data kedalam beberapa *cluster.* Pada contoh perhitungan ini, jumlah *cluster* yang digunakan adalah 2 *cluster,* yang mana nantinya data akan dikelompokkan kedalam 2 *cluster* (kelompok).

## Perhitungan nilai *Purity* Setiap Data

Penentuan titik pusat *cluster* (*centroid*) awal dilakukan berdasarkan nilai *Purity*. Perhitungan *purity* dari data ke-1 sampai data ke-10 adalah sebagai berikut:

1. *Purity* data ke-1

𝑃𝑢𝑟𝑖𝑡(1) = 1

(0.08 + 0.08 + 0.24 + 0.9)

max(0.08, 0.08, 0.24, 0.9)

= 1

(1.3)

= 0.69

1. *Purity* data ke-2

(0.9)

𝑃𝑢𝑟𝑖𝑡(2) = 1

(0.06 + 0.06 + 0.25 + 0.33)

max(0.06, 0.06, 0.25, 0.33)

= 1

(0.7)

= 0.47

(0.33)

1. *Purity* data ke-3

𝑃𝑢𝑟𝑖𝑡(3) = 1

(0.1 + 0.1 + 0.65 + 0.3)

max(0.1, 0.1, 0.65, 0.3)

= 1

(1.15)

= 0.56

1. *Purity* data ke-4

(0.65)

𝑃𝑢𝑟𝑖𝑡(4) = 1

(0.08 + 0.08 + 0.98 + 0.24)

max(0.08, 0.08, 0.98, 0.24)

= 1

(1.38)

= 0.71

1. *Purity* data ke-5

(0.98)

𝑃𝑢𝑟𝑖𝑡(5) = 1

(0.09 + 0.15 + 0.1 + 0.66)

max(0.09, 0.15, 0.1, 0.66)

= 1

(1)

(0.66)

= 0.66

1. *Purity* data ke-6

𝑃𝑢𝑟𝑖𝑡(6) = 1

(0.1 + 0.1 + 0.29 + 0.56)

max(0.1, 0.1, 0.29, 0.56)

= 1

(1.05)

= 0.53

1. *Purity* data ke-7

(0.56)

𝑃𝑢𝑟𝑖𝑡(7) = 1

(0.15 + 0.02 + 0.4 + 0.01)

max(0.15, 0.02, 0.4, 0.01)

= 1

(0.58)

= 0.68

(0.4)

1. *Purity* data ke-8

𝑃𝑢𝑟𝑖𝑡(8) = 1

(0.2 + 0.14 + 0.72 + 0.25)

max(0.2, 0.14, 0.72, 0.25)

= 1

(1.31)

= 0.54

1. *Purity* data ke-9

(0.72)

𝑃𝑢𝑟𝑖𝑡(9) = 1

(0 + 0 + 0.2 + 0.85)

max(0, 0, 0.2, 0.85)

= 1

(1.05)

= 0.8

1. *Purity* data ke-10

(0.85)

𝑃𝑢𝑟𝑖𝑡(10) = 1

(0.18 + 0.18 + 0.3 + 0.81)

max(0.18, 0.18, 0.3, 0.81)

= 1

(1.47)

= 0.55

(0.81)

Adapun untuk hasil perhitungan *Purity* secara keseluruhan yang dihitung menggunakan persamaan (2.2) dirangkum pada tabel 3.2 dibawah ini :

Tabel 3.2 Hasil perhitungan *Purity* setiap data

|  |  |
| --- | --- |
| **Data ke-** | **Nilai *Purity*** |
| **1** | 0.69 |
| **2** | 0.47 |
| **3** | 0.56 |
| **4** | 0.71 |
| **5** | 0.66 |
| **6** | 0.53 |
| **7** | 0.68 |
| **8** | 0.54 |
| **9** | 0.8 |
| **10** | 0.55 |

* 1. **Inisialisasi *Centroid* Awal Berdasarkan nilai maximum dan minimum *Purity***

Berdasarkan tabel 3.2, data yang memiliki nilai *Purity* tertinggi adalah data ke-9 dengan nilai 0.8, dan data ke-2 merupakan data yang memiliki nilai *Purity* terendah dengan nilai 0.47. Maka data ke-9 dan ke-2 akan digunakan sebagai titik pusat *cluster* awal pada proses *clustering.* Titik pusat *cluster* awal dapat dilihat pada tabel 3.3 dibawah ini.

Tabel 3.3 Titik Pusat *cluster* (*Centroid*) awal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Titik pusat *cluster* (m) ke-*i*** | **Data Ke-** | **Atribut/Fitur** | | | |
| **Atribut 1 (*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** |
| ***m1*** | 9 | 0 | 0 | 0.2 | 0.85 |
| ***m2*** | 2 | 0.06 | 0.06 | 0.25 | 0.33 |

## Perhitungan Jarak Data terhadap *centroid*

Proses selanjutnya adalah perhitungan jarak data dengan titik pusat *cluster* menggunakan *Euclidean Distance*. Perhitungan jarak data ke-1 sampai data ke-10 terhadap titik pusat *cluster* adalah sebagai berikut:

1. Jarak data ke-1 dengan titik pusat *cluster*

𝐷(1,1) = √(0.08 − 0)2 + (0.08 − 0)2 + (0.24 − 0.2)2 + (0.9 − 0.85)2

= √(0.08)2 + (0.08)2 + (0.04)2 + (0.05)2

= √(0.0169) = 0.13

𝐷(1,2) = √(0.08 − 0.06)2 + (0.08 − 0.06)2 + (0.24 − 0.25)2 + (0.9 − 0.33)2

= √(0.02)2 + (0.02)2 + (−0.01)2 + (0.57)2

= √(0.3258) = 0.57

1. Jarak data ke-2 dengan titik pusat *cluster*

𝐷(2,1) = √(0.06 − 0)2 + (0.06 − 0)2 + (0.25 − 0.2)2 + (0.33 − 0.85)2

= √(0.06)2 + (0.06)2 + (0.05)2 + (−0.52)2

= √(0.2801) = 0.52

𝐷(2,2) = √(0.06 − 0.06)2 + (0.06 − 0.06)2 + (0.25 − 0.25)2 + (0.33 − 0.33)2

= √(0)2 + (0)2 + (0)2 + (0)2

= √(0) = 0

1. Jarak data ke-3 dengan titik pusat *cluster*

𝐷(3,1) = √(0.1 − 0)2 + (0.1 − 0)2 + (0.65 − 0.2)2 + (0.3 − 0.85)2

= √(0.1)2 + (0.1)2 + (0.45)2 + (−0.55)2

= √(0.525) = 0.72

𝐷(3,2) = √(0.1 − 0.06)2 + (0.1 − 0.06)2 + (0.65 − 0.25)2 + (0.3 − 0.33)2

= √(0.04)2 + (0.04)2 + (0.4)2 + (−0.03)2

= √(0.1641) = 0.4

1. Jarak data ke-4 dengan titik pusat *cluster*

𝐷(4,1) = √(0.08 − 0)2 + (0.08 − 0)2 + (0.98 − 0.2)2 + (0.24 − 0.85)2

= √(0.08)2 + (0.08)2 + (0.78)2 + (−0.61)2

= √(0.9933) = 0.99

𝐷(4,2) = √(0.08 − 0.06)2 + (0.08 − 0.06)2 + (0.98 − 0.25)2 + (0.24 − 0.33)2

= √(0.02)2 + (0.02)2 + (0.73)2 + (−0.09)2

= √(0.5418) = 0.73

1. Jarak data ke-5 dengan titik pusat *cluster*

𝐷(5,1) = √(0.09 − 0)2 + (0.15 − 0)2 + (0.1 − 0.2)2 + (0.66 − 0.85)2

= √(0.09)2 + (0.15)2 + (−0.1)2 + (−0.19)2

= √(0.0767) = 0.27

𝐷(5,2) = √(0.09 − 0.06)2 + (0.15 − 0.06)2 + (0.1 − 0.25)2 + (0.66 − 0.33)2

= √(0.03)2 + (0.09)2 + (−0.15)2 + (0.33)2

= √(0.1396) = 0.37

1. Jarak data ke-6 dengan titik pusat *cluster*

𝐷(6,1) = √(0.1 − 0)2 + (0.1 − 0)2 + (0.29 − 0.2)2 + (0.56 − 0.85)2

= √(0.1)2 + (0.1)2 + (0.09)2 + (−0.29)2

= √(0.1122) = 0.33

𝐷(6,2) = √(0.1 − 0.06)2 + (0.1 − 0.06)2 + (0.29 − 0.25)2 + (0.56 − 0.33)2

= √(0.04)2 + (0.04)2 + (0.04)2 + (0.23)2

= √(0.0577) = 0.24

1. Jarak data ke-7 dengan titik pusat *cluster*

𝐷(7,1) = √(0.15 − 0)2 + (0.02 − 0)2 + (0.4 − 0.2)2 + (0.01 − 0.85)2

= √(0.15)2 + (0.02)2 + (0.2)2 + (−0.84)2

= √(0.7685) = 0.87

𝐷(7,2) = √(0.15 − 0.06)2 + (0.02 − 0.06)2 + (0.4 − 0.25)2 + (0.01 − 0.33)2

= √(0.09)2 + (−0.04)2 + (0.15)2 + (−0.32)2

= √(0.1346) = 0.36

1. Jarak data ke-8 dengan titik pusat *cluster*

𝐷(8,1) = √(0.2 − 0)2 + (0.14 − 0)2 + (0.72 − 0.2)2 + (0.25 − 0.85)2

= √(0.2)2 + (0.14)2 + (0.52)2 + (−0.6)2

= √(0.69) = 0.83

𝐷(8,2) = √(0.2 − 0.06)2 + (0.14 − 0.06)2 + (0.72 − 0.25)2 + (0.25 − 0.33)2

= √(0.14)2 + (0.08)2 + (0.47)2 + (−0.08)2

= √(0.2533) = 0.50

1. Jarak data ke-9 dengan titik pusat *cluster*

𝐷(9,1) = √(0 − 0)2 + (0 − 0)2 + (0.2 − 0.2)2 + (0.85 − 0.85)2

= √(0)2 + (0)2 + (0)2 + (0)2

= √(0) = 0

𝐷(9,2) = √(0 − 0.06)2 + (0 − 0.06)2 + (0.2 − 0.25)2 + (0.85 − 0.33)2

= √(−0.06)2 + (−0.06)2 + (−0.05)2 + (0.52)2

= √(0.2801) = 0.52

1. Jarak data ke-10 dengan titik pusat *cluster*

𝐷(10,1) = √(0.18 − 0)2 + (0.18 − 0)2 + (0.3 − 0.2)2 + (0.81 − 0.85)2

= √(0.18)2 + (0.18)2 + (0.1)2 + (−0.04)2

= √(0.0764) = 0.27

𝐷(10,2) = √(0.18 − 0.06)2 + (0.18 − 0.06)2 + (0.3 − 0.25)2 + (0.81 − 0.33)2

= √(0.12)2 + (0.12)2 + (0.05)2 + (0.48)2

= √(0.2617) = 0.51

Perhitungan jarak dari setiap data dengan kedua titik pusat *cluster* dapat dilihat pada tabel 3.4 sebagai berikut.

Tabel 3.4 Perhitungan jarak data dengan titik pusat *cluster*

|  |  |  |
| --- | --- | --- |
| **Data ke-** | **Jarak data dengan titik pusat**  ***cluster* 1** | **Jarak data dengan titik pusat**  ***cluster* 2** |
| **1** | 0.13 | 0.57 |
| **2** | 0.52 | 0 |
| **3** | 0.72 | 0.4 |
| **4** | 0.99 | 0.73 |
| **5** | 0.27 | 0.37 |
| **6** | 0.33 | 0.24 |
| **7** | 0.87 | 0.36 |
| **8** | 0.83 | 0.50 |
| **9** | 0 | 0.52 |
| **10** | 0.27 | 0.51 |

## Pengelompokan Data Berdasarkan Jarak Minimum

Berdasarkan hasil perhitungan *Euclidean Distance* pada tabel 3.4, memperlihatkan bahwa jarak terdekat pada data ke-1 adalah dengan titik pusat *cluster* pertama (*m1*). Karena jarak data pertama sebesar 0.13 dan lebih kecil dibandingkan dengan jarak data pertama dengan titik pusat *cluster* kedua (*m2*) dengan jarak sebesar 0.57. Sehingga dapat disimpulkan bahwa data pertama menjadi anggota *cluster* dari *cluster* ke-1. Pengelompokan dari keseluruhan jarak minimum dapat dilihat pada tabel 3.5 dibawah ini :

Tabel 3.5 Pengelompokan data berdasarkan jarak minimum

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data ke-** | **Jarak data dengan titik**  **pusat *cluster* 1** | **Jarak data dengan titik**  **pusat *cluster* 2** | ***Min*** | ***Cluster*** | |
| C1 | C2 |
| **1** | 0.13 | 0.57 | 0.13 | C1 |  |
| **2** | 0.52 | 0 | 0 |  | C2 |
| **3** | 0.72 | 0.4 | 0.4 |  | C2 |
| **4** | 0.99 | 0.73 | 0.73 |  | C2 |
| **5** | 0.27 | 0.37 | 0.27 | C1 |  |

Tabel 3.6 Pengelompokan data berdasarkan jarak minimum (lanjutan)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **6** | **0.33** | **0.24** | **0.24** |  | **C2** |
| **7** | 0.87 | 0.36 | 0.36 |  | C2 |
| **8** | 0.83 | 0.50 | 0.50 |  | C2 |
| **9** | 0 | 0.52 | 0 | C1 |  |
| **10** | 0.27 | 0.51 | 0.27 | C1 |  |

* 1. **Perhitungan Titik Pusat *Cluster* Baru**

Setelah data dikelompokan berdasarkan jarak minimum, langkah selanjutnya adalah menghitung titik pusat *cluster* baru. Untuk mendapatkan titik pusat cluster baru, digunakan persamaan (2.1). Untuk contoh perhitungannya dapat dilihat pada tabel 3.7 dan 3.8 berikut ini.

Tabel 3.7 Perhitungan titik pusat *cluster* baru pada *cluster* 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Ke-** | **Atribut/Fitur** | | | |
| **Atribut 1**  **(*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** |
| **1** | 0.08 | 0.08 | 0.24 | 0.9 |
| **5** | 0.09 | 0.15 | 0.1 | 0.66 |
| **9** | 0 | 0 | 0.2 | 0.85 |
| **10** | 0.18 | 0.18 | 0.3 | 0.81 |
| **Total Jumlah** | 0.35 | 0.41 | 0.84 | 3.22 |
| ***Average*** | 0.35/4=0.08 | 0.41/4=0.10 | 0.84/4=0.21 | 3.22/4=0.80 |
| ***m1* baru** | 0.08 | 0.10 | 0.21 | 0.80 |

Tabel 3.8 Perhitungan titik pusat *cluster* baru pada *cluster* 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Ke-** | **Atribut/Fitur** | | | |
| **Atribut 1 (*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** |
| **2** | 0.06 | 0.06 | 0.25 | 0.33 |
| **3** | 0.1 | 0.1 | 0.65 | 0.3 |
| **4** | 0.08 | 0.08 | 0.98 | 0.24 |
| **6** | 0.1 | 0.1 | 0.29 | 0.56 |
| **7** | 0.15 | 0.02 | 0.4 | 0.01 |
| **8** | 0.2 | 0.14 | 0.72 | 0.25 |
| **Total Jumlah** | 0.69 | 0.5 | 3.29 | 1.69 |
| ***Average*** | 0.69/6=0.11 | 0.5/6=0.08 | 3.29/6=0.54 | 1.69/6=0.28 |
| ***m2* baru** | 0.11 | 0.08 | 0.54 | 0.28 |

Dari tabel 3.7 dan 3.8 diperoleh titik pusat *cluster* (*centroid*) yang baru yang akan digunakan untuk menghitung jarak data terhadap titik pusat *cluster* yang baru pada iterasi selanjutnya. Hasil perhitungan titik pusat *cluster* baru dapat dilihat pada tabel 3.9.

Tabel 3.9 Titik Pusat *cluster* (*Centroid*) baru

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***mj*** | **Atribut/Fitur** | | | |
| **Atribut 1 (*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** |
| ***m1* baru** | 0.08 | 0.10 | 0.21 | 0.80 |
| ***m2* baru** | 0.11 | 0.08 | 0.54 | 0.28 |

Setelah diperoleh nilai *centroid* yang baru, maka jarak data terhadap *centroid* kembali dihitung. Iterasi selanjutnya dilakukan dengan cara yang sama, dan akan berhenti sampai tidak ada lagi perubahan data anggota *cluster* didalam sebuah *cluster*, atau tidak terjadi lagi perpindahan data dari satu *cluster* ke *cluster* lainnya.

* 1. **Menghitung Nilai *Davies-Bouldin Index***

Perhitungan nilai *Davies-Bouldin Index* bertujuan untuk menganalisa kualitas *clusters* dari setiap proses *clustering* tersebut. Untuk mendapatkan nilai *Davies-Bouldin Index*, terlebih dahulu dihitung nilai *Sum of Square Within-cluster* (SSW), *Sum of Square Between-cluster* (SSB) dan *Ratio*. Untuk melakukan perhitungan manual *Davies- Bouldin Index*, dibutuhkan data yang sudah selesai di-*cluster* terlebih dahulu. Contoh data *sample* yang telah selesai di-*clustering* dapat dilihat pada tabel 3.10 sebagai berikut:

Tabel 3.10 Ilustrasi *sample* yang telah di-*clustering*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C1** | | | | | **C2** | | | | |
| **Data** | **x1** | **x2** | **x3** | **x4** | **Data** | **x1** | **x2** | **x3** | **x4** |
| **ke-** | **ke-** |
| **1** | 0.08 | 0.08 | 0.24 | 0.9 | **2** | 0.06 | 0.06 | 0.25 | 0.33 |
| **5** | 0.09 | 0.15 | 0.1 | 0.66 | **3** | 0.1 | 0.1 | 0.65 | 0.3 |
| **9** | 0 | 0 | 0.2 | 0.85 | **4** | 0.08 | 0.08 | 0.98 | 0.24 |

Tabel 3.11 Ilustrasi *sample* yang telah di-*clustering* (lanjutan)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | **0.18** | **0.18** | **0.3** | **0.81** | 6 | **0.1** | **0.1** | **0.29** | **0.56** |
|  |  |  |  |  | **7** | 0.15 | 0.02 | 0.4 | 0.01 |
|  |  |  |  |  | **8** | 0.2 | 0.14 | 0.72 | 0.25 |

Data yang telah selesai di-*clustering* tersebut kemudian digunakan sebagai *input-*an dalam menghitung *Davies-Bouldin Index.* Untuk perhitungan hasil evaluasi *Davies- Bouldin Index* adalah dengan menghitung nilai *Sum of Square Within-cluster* (SSW) menggunakan persamaan (2.3). Nilai SSW yang diperoleh dari perhitungan jarak setiap data terhadap titik pusat *cluster* akhir menggunakan *Euclidean Distance.* Contoh data *sample* yang untuk nilai *centroid* akhir dari proses *clustering* dapat dilihat pada tabel 3.12 sebagai berikut:

Tabel 3.12 Ilustrasi titik pusat *cluster* (*Centroid*) akhir

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***mj*** | **Atribut/Fitur** | | | |
| **Atribut 1 (*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** |
| ***m1* baru** | 0.08 | 0.09 | 0.23 | 0.68 |
| ***m2* baru** | 0.11 | 0.08 | 0.54 | 0.28 |

Nilai yang diperoleh dari keseluruhan perhitungan SSW untuk masing-masing *cluster*

adalah sebagai berikut :

𝑆𝑆𝑊1 = 1

(5)

= 1

(5)

∗(0.22 + 0.145 + 0.136 + 0.21 + 0.2)

(0.912)

= 0.182

𝑆𝑆𝑊1 = 1

(5)

= 1

(5)

∗(0.299 + 0.114 + 0.443 + 0.313 + 0.212)

(1.381)

= 0.276

Setelah diperoleh nilai SSW, maka langkah selanjutnya adalah menghitung nilai *Sum of Square Between-cluster* (SSB). Untuk menghitung nilai SSB digunakan persamaan (2.4). Adapun nilai yang diperoleh setelah menghitung SSB adalah sebagai berikut:

𝑆𝑆𝐵1,2 = √(0.08 − 0.11)2 + (0.09 − 0.08)2 + (0.23 − 0.54)2 + (0.68 − 0.28)2

= 0.507

Langkah selanjutnya setelah mendapatkan nilai SSW dan nilai SSB adalah menghitung nilai *Ratio* (Rasio). Nilai *ratio* dihitung dengan menggunakan persamaan (2.5).

𝑅1,2

= 0.182 + 0.276 = 0.904

0.507

Tabel 3.13 Nilai rasio dari setiap *cluster*

|  |  |  |  |
| --- | --- | --- | --- |
| ***R*** | **Data ke-*i*** | | ***R-Max*** |
| **1** | **2** |
| **1** | 0 | 0.904 | 0.904 |
| **2** | 0.904 | 0 | 0.904 |

Setelah mendapatkan nilai rasio, maka langkah terakhir adalah menghitung nilai *Davies-Bouldin Index* (DBI). Nilai *Davies-Bouldin Index* (DBI) diperoleh dengan menggunakan persamaan (2.6).

𝐷𝐵𝐼 = 1 (0.904 + 0.904) = 0.904

2

Berdasarkan nilai DBI diatas, maka dapat disimpulkan hasil akurasi *clustering* pada tabel 3.13 adalah cukup baik, karena semakin kecil nilai DBI yang diperoleh (non-negatif >= 0), maka semakin baik *cluster* yang diperoleh dari proses peng*clusteran* K-Means yang dilakukan (Qiang Zhan, 2017).

## BAB IV

**HASIL DAN PEMBAHASAN**

## Hasil Pengujian

Untuk menguji hasil penelitian ini, digunakan beberapa dataset yang diperoleh dari berbagai sumber. Adapun ringkasan dataset yang digunakan pada penelitian ini dapat dilihat pada table 4.1.

Tabel 4.1 Ringkasan Dataset Yang Digunakan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | ***Dataset*** | **Jumlah Data** | **Jumlah Atribut** | **Jenis Data** | **Tipe Data** |
| **1** | *Iris* | 150 | 4 | Numerik | *Real* |
| **2** | *User Knowledge*  *Modeling* | 395 | 5 | Numerik | *Real,*  *Integer* |
| **3** | *Birth and Death Rates* | 70 | 2 | Numerik | *Real,*  *Integer* |
| **4** | *Wholesale Customers* | 440 | 6 | Numerik | *Real* |

Dalam penelitian ini bertujuan untuk meminimalisasi proses iterasi yang dibutuhkan untuk mencapai konvergen dan meningkatkan akurasi dari setiap proses peng- *cluster*an yang dilakukan dengan menggunakan *Purity* dan mengevaluasinya dengan menggunakan *Davies-Bouldin Index* (DBI).

## Dataset Iris

Untuk mendukung kelancaran penelitian, berikut akan dipaparkan dataset *iris* untuk perhitungan manual dari proses *clustering* digunakan satu data sampel dengan jumlah 150 data didalamnya. Data tersebut tertera pada tabel 4.2 berikut.

25

Tabel 4.2 Dataset *Iris*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data ke-*i*  (*pi*) | Atribut 1 (*x1*) | Atribut 2 (*x2*) | Atribut 3 (*x3*) | Atribut 4 (*x4*) | Variety |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Setosa |
| 2 | 4.9 | 3 | 1.4 | 0.2 | Setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | Setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | Setosa |
| 5 | 5 | 3.6 | 1.4 | 0.2 | Setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | Setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | Setosa |
| 8 | 5 | 3.4 | 1.5 | 0.2 | Setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | Setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | Setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | Setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | Setosa |
| 13 | 4.8 | 3 | 1.4 | 0.1 | Setosa |
| 14 | 4.3 | 3 | 1.1 | 0.1 | Setosa |
| 15 | 5.8 | 4 | 1.2 | 0.2 | Setosa |
| 16 | 5.7 | 4.4 | 1.5 | 0.4 | Setosa |
| 17 | 5.4 | 3.9 | 1.3 | 0.4 | Setosa |
| 18 | 5.1 | 3.5 | 1.4 | 0.3 | Setosa |
| 19 | 5.7 | 3.8 | 1.7 | 0.3 | Setosa |
| 20 | 5.1 | 3.8 | 1.5 | 0.3 | Setosa |
| 21 | 5.4 | 3.4 | 1.7 | 0.2 | Setosa |
| 22 | 5.1 | 3.7 | 1.5 | 0.4 | Setosa |
| 23 | 4.6 | 3.6 | 1 | 0.2 | Setosa |
| 24 | 5.1 | 3.3 | 1.7 | 0.5 | Setosa |
| 25 | 4.8 | 3.4 | 1.9 | 0.2 | Setosa |
| 26 | 5 | 3 | 1.6 | 0.2 | Setosa |
| 27 | 5 | 3.4 | 1.6 | 0.4 | Setosa |
| 28 | 5.2 | 3.5 | 1.5 | 0.2 | Setosa |
| 29 | 5.2 | 3.4 | 1.4 | 0.2 | Setosa |
| 30 | 4.7 | 3.2 | 1.6 | 0.2 | Setosa |
| 31 | 4.8 | 3.1 | 1.6 | 0.2 | Setosa |
| 32 | 5.4 | 3.4 | 1.5 | 0.4 | Setosa |
| 33 | 5.2 | 4.1 | 1.5 | 0.1 | Setosa |
| 34 | 5.5 | 4.2 | 1.4 | 0.2 | Setosa |
| 35 | 4.9 | 3.1 | 1.5 | 0.2 | Setosa |
| 36 | 5 | 3.2 | 1.2 | 0.2 | Setosa |
| 37 | 5.5 | 3.5 | 1.3 | 0.2 | Setosa |
| 38 | 4.9 | 3.6 | 1.4 | 0.1 | Setosa |
| 39 | 4.4 | 3 | 1.3 | 0.2 | Setosa |
| 40 | 5.1 | 3.4 | 1.5 | 0.2 | Setosa |

Tabel 4.3 Dataset *Iris* (lanjutan)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 41 | 5 | 3.5 | 1.3 | 0.3 | Setosa |
| 42 | 4.5 | 2.3 | 1.3 | 0.3 | Setosa |
| 43 | 4.4 | 3.2 | 1.3 | 0.2 | Setosa |
| 44 | 5 | 3.5 | 1.6 | 0.6 | Setosa |
| 45 | 5.1 | 3.8 | 1.9 | 0.4 | Setosa |
| 46 | 4.8 | 3 | 1.4 | 0.3 | Setosa |
| 47 | 5.1 | 3.8 | 1.6 | 0.2 | Setosa |
| 48 | 4.6 | 3.2 | 1.4 | 0.2 | Setosa |
| 49 | 5.3 | 3.7 | 1.5 | 0.2 | Setosa |
| 50 | 5 | 3.3 | 1.4 | 0.2 | Setosa |
| 51 | 7 | 3.2 | 4.7 | 1.4 | Versicolor |
| 52 | 6.4 | 3.2 | 4.5 | 1.5 | Versicolor |
| 53 | 6.9 | 3.1 | 4.9 | 1.5 | Versicolor |
| 54 | 5.5 | 2.3 | 4 | 1.3 | Versicolor |
| 55 | 6.5 | 2.8 | 4.6 | 1.5 | Versicolor |
| 56 | 5.7 | 2.8 | 4.5 | 1.3 | Versicolor |
| 57 | 6.3 | 3.3 | 4.7 | 1.6 | Versicolor |
| 58 | 4.9 | 2.4 | 3.3 | 1 | Versicolor |
| 59 | 6.6 | 2.9 | 4.6 | 1.3 | Versicolor |
| 60 | 5.2 | 2.7 | 3.9 | 1.4 | Versicolor |
| 61 | 5 | 2 | 3.5 | 1 | Versicolor |
| 62 | 5.9 | 3 | 4.2 | 1.5 | Versicolor |
| 63 | 6 | 2.2 | 4 | 1 | Versicolor |
| 64 | 6.1 | 2.9 | 4.7 | 1.4 | Versicolor |
| 65 | 5.6 | 2.9 | 3.6 | 1.3 | Versicolor |
| 66 | 6.7 | 3.1 | 4.4 | 1.4 | Versicolor |
| 67 | 5.6 | 3 | 4.5 | 1.5 | Versicolor |
| 68 | 5.8 | 2.7 | 4.1 | 1 | Versicolor |
| 69 | 6.2 | 2.2 | 4.5 | 1.5 | Versicolor |
| 70 | 5.6 | 2.5 | 3.9 | 1.1 | Versicolor |
| 71 | 5.9 | 3.2 | 4.8 | 1.8 | Versicolor |
| 72 | 6.1 | 2.8 | 4 | 1.3 | Versicolor |
| 73 | 6.3 | 2.5 | 4.9 | 1.5 | Versicolor |
| 74 | 6.1 | 2.8 | 4.7 | 1.2 | Versicolor |
| 75 | 6.4 | 2.9 | 4.3 | 1.3 | Versicolor |
| 76 | 6.6 | 3 | 4.4 | 1.4 | Versicolor |
| 77 | 6.8 | 2.8 | 4.8 | 1.4 | Versicolor |
| 78 | 6.7 | 3 | 5 | 1.7 | Versicolor |
| 79 | 6 | 2.9 | 4.5 | 1.5 | Versicolor |
| 80 | 5.7 | 2.6 | 3.5 | 1 | Versicolor |
| 81 | 5.5 | 2.4 | 3.8 | 1.1 | Versicolor |
| 82 | 5.5 | 2.4 | 3.7 | 1 | Versicolor |
| 83 | 5.8 | 2.7 | 3.9 | 1.2 | Versicolor |

Tabel 4.4 Dataset *Iris* (lanjutan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 84 | 6 | 2.7 | 5.1 | 1.6 | Versicolor |
| 85 | 5.4 | 3 | 4.5 | 1.5 | Versicolor |
| 86 | 6 | 3.4 | 4.5 | 1.6 | Versicolor |
| 87 | 6.7 | 3.1 | 4.7 | 1.5 | Versicolor |
| 88 | 6.3 | 2.3 | 4.4 | 1.3 | Versicolor |
| 89 | 5.6 | 3 | 4.1 | 1.3 | Versicolor |
| 90 | 5.5 | 2.5 | 4 | 1.3 | Versicolor |
| 91 | 5.5 | 2.6 | 4.4 | 1.2 | Versicolor |
| 92 | 6.1 | 3 | 4.6 | 1.4 | Versicolor |
| 93 | 5.8 | 2.6 | 4 | 1.2 | Versicolor |
| 94 | 5 | 2.3 | 3.3 | 1 | Versicolor |
| 95 | 5.6 | 2.7 | 4.2 | 1.3 | Versicolor |
| 96 | 5.7 | 3 | 4.2 | 1.2 | Versicolor |
| 97 | 5.7 | 2.9 | 4.2 | 1.3 | Versicolor |
| 98 | 6.2 | 2.9 | 4.3 | 1.3 | Versicolor |
| 99 | 5.1 | 2.5 | 3 | 1.1 | Versicolor |
| 100 | 5.7 | 2.8 | 4.1 | 1.3 | Versicolor |
| 101 | 6.3 | 3.3 | 6 | 2.5 | Virginica |
| 102 | 5.8 | 2.7 | 5.1 | 1.9 | Virginica |
| 103 | 7.1 | 3 | 5.9 | 2.1 | Virginica |
| 104 | 6.3 | 2.9 | 5.6 | 1.8 | Virginica |
| 105 | 6.5 | 3 | 5.8 | 2.2 | Virginica |
| 106 | 7.6 | 3 | 6.6 | 2.1 | Virginica |
| 107 | 4.9 | 2.5 | 4.5 | 1.7 | Virginica |
| 108 | 7.3 | 2.9 | 6.3 | 1.8 | Virginica |
| 109 | 6.7 | 2.5 | 5.8 | 1.8 | Virginica |
| 110 | 7.2 | 3.6 | 6.1 | 2.5 | Virginica |
| 111 | 6.5 | 3.2 | 5.1 | 2 | Virginica |
| 112 | 6.4 | 2.7 | 5.3 | 1.9 | Virginica |
| 113 | 6.8 | 3 | 5.5 | 2.1 | Virginica |
| 114 | 5.7 | 2.5 | 5 | 2 | Virginica |
| 115 | 5.8 | 2.8 | 5.1 | 2.4 | Virginica |
| 116 | 6.4 | 3.2 | 5.3 | 2.3 | Virginica |
| 117 | 6.5 | 3 | 5.5 | 1.8 | Virginica |
| 118 | 7.7 | 3.8 | 6.7 | 2.2 | Virginica |
| 119 | 7.7 | 2.6 | 6.9 | 2.3 | Virginica |
| 120 | 6 | 2.2 | 5 | 1.5 | Virginica |
| 121 | 6.9 | 3.2 | 5.7 | 2.3 | Virginica |
| 122 | 5.6 | 2.8 | 4.9 | 2 | Virginica |
| 123 | 7.7 | 2.8 | 6.7 | 2 | Virginica |
| 124 | 6.3 | 2.7 | 4.9 | 1.8 | Virginica |
| 125 | 6.7 | 3.3 | 5.7 | 2.1 | Virginica |
| 126 | 7.2 | 3.2 | 6 | 1.8 | Virginica |

Tabel 4.5 Dataset *Iris* (lanjutan)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 127 | 6.2 | 2.8 | 4.8 | 1.8 | Virginica |
| 128 | 6.1 | 3 | 4.9 | 1.8 | Virginica |
| 129 | 6.4 | 2.8 | 5.6 | 2.1 | Virginica |
| 130 | 7.2 | 3 | 5.8 | 1.6 | Virginica |
| 131 | 7.4 | 2.8 | 6.1 | 1.9 | Virginica |
| 132 | 7.9 | 3.8 | 6.4 | 2 | Virginica |
| 133 | 6.4 | 2.8 | 5.6 | 2.2 | Virginica |
| 134 | 6.3 | 2.8 | 5.1 | 1.5 | Virginica |
| 135 | 6.1 | 2.6 | 5.6 | 1.4 | Virginica |
| 136 | 7.7 | 3 | 6.1 | 2.3 | Virginica |
| 137 | 6.3 | 3.4 | 5.6 | 2.4 | Virginica |
| 138 | 6.4 | 3.1 | 5.5 | 1.8 | Virginica |
| 139 | 6 | 3 | 4.8 | 1.8 | Virginica |
| 140 | 6.9 | 3.1 | 5.4 | 2.1 | Virginica |
| 141 | 6.7 | 3.1 | 5.6 | 2.4 | Virginica |
| 142 | 6.9 | 3.1 | 5.1 | 2.3 | Virginica |
| 143 | 5.8 | 2.7 | 5.1 | 1.9 | Virginica |
| 144 | 6.8 | 3.2 | 5.9 | 2.3 | Virginica |
| 145 | 6.7 | 3.3 | 5.7 | 2.5 | Virginica |
| 146 | 6.7 | 3 | 5.2 | 2.3 | Virginica |
| 147 | 6.3 | 2.5 | 5 | 1.9 | Virginica |
| 148 | 6.5 | 3 | 5.2 | 2 | Virginica |
| 149 | 6.2 | 3.4 | 5.4 | 2.3 | Virginica |
| 150 | 5.9 | 3 | 5.1 | 1.8 | Virginica |

Setelah data di-*input* maka selanjutnya adalah penentuan jumlah *cluster* (K)*..* Pada perhitungan ini, jumlah *cluster* yang digunakan adalah 3 *cluster.*

* 1. **Nilai *Purity* dataset *Iris***

Penentuan titik pusat *cluster* (*centroid*) awal dilakukan berdasarkan nilai *Purity*. Perhitungan *purity* dari data ke-1 sampai data ke-10 adalah sebagai berikut:

1. *Purity* data ke-1

𝑃𝑢𝑟𝑖𝑡(1) = 1

(5.1 + 3.5 + 1.4 + 0.2)

max(5.1, 3.5, 1.4, 0.2)

= 1

(10.2)

= 0.5

(5.1)

1. *Purity* data ke-2

𝑃𝑢𝑟𝑖𝑡(2) = 1

(4.9 + 3 + 1.4 + 0.2)

max(4.9, 3, 1.4, 0.2)

= 1

(9.5)

(4.9)

= 0.515789

1. *Purity* data ke-3

𝑃𝑢𝑟𝑖𝑡(3) = 1

(4.7 + 3.2 + 1.3 + 0.2)

max(4.7, 3.2, 1.3, 0.2)

= 1

(9.4)

= 0.5

1. *Purity* data ke-4

(4.7)

𝑃𝑢𝑟𝑖𝑡(4) = 1

(4.6 + 3.1 + 1.5 + 0.2)

max(4.6, 3.1, 1.5, 0.2)

= 1

(9.4)

(0.489362)

= 0.489362

1. *Purity* data ke-5

𝑃𝑢𝑟𝑖𝑡(5) = 1

(5 + 3.6 + 1.4 + 0.2)

max(5, 3.6, 1.4, 0.2)

= 1

(10.2)

(5)

= 0.490196

1. *Purity* data ke-6

𝑃𝑢𝑟𝑖𝑡(6) = 1

(5.4 + 3.9 + 1.7 + 0.4)

max(5.4, 3.9, 1.7, 0.4)

= 1

(11.4)

(5.4)

= 0.473684

1. *Purity* data ke-7

𝑃𝑢𝑟𝑖𝑡(7) = 1

(4.6 + 3.4 + 1.4 + 0.3)

max(4.6, 3.4, 1.4, 0.3)

= 1

(9.7)

(4.6)

= 0.474227

1. *Purity* data ke-8

𝑃𝑢𝑟𝑖𝑡(8) = 1

(5 + 3.4 + 1.5 + 0.2)

max(5, 3.4, 1.5, 0.2)

= 1

(10.1)

(5)

= 0.49505

1. *Purity* data ke-9

𝑃𝑢𝑟𝑖𝑡(9) = 1

(4.4 + 2.9 + 1.4 + 0.2)

max(4.4, 2.9, 1.4, 0.2)

= 1

(8.9)

(4.4)

= 0.494382

1. *Purity* data ke-10

𝑃𝑢𝑟𝑖𝑡(10) = 1

(4.9 + 3.1 + 1.5 + 0.1)

max(4.9, 3.1, 1.5, 0.1)

= 1

(9.6)

(4.9)

= 0.510417

Adapun untuk hasil perhitungan *Purity* secara keseluruhan yang dihitung menggunakan persamaan (2.2) dirangkum pada tabel 4.6 dibawah ini :

Tabel 4.6 Hasil perhitungan *Purity* setiap data

|  |  |
| --- | --- |
| **Data ke-** | **Nilai *Purity*** |
| **1** | 0.5 |
| **2** | 0.515789 |
| **3** | 0.5 |
| **4** | 0.489362 |
| **5** | 0.490196 |
| **6** | 0.473684 |
| **7** | 0.474227 |
| **8** | 0.49505 |
| **9** | 0.494382 |

Tabel 4.7 Hasil perhitungan *Purity* setiap data (lanjutan)

|  |  |
| --- | --- |
| **10** | 0.510417 |
| **11** | 0.5 |
| **12** | 0.48 |
| **13** | 0.516129 |
| **14** | 0.505882 |
| **15** | 0.517857 |
| **16** | 0.475 |
| **17** | 0,490909 |
| **18** | 0.495146 |
| **19** | 0.495652 |
| **20** | 0.476636 |
| **21** | 0.504673 |
| **22** | 0.476636 |
| **23** | 0.489362 |
| **24** | 0.481132 |
| **25** | 0.466019 |
| **26** | 0.510204 |
| **27** | 0.480769 |
| **28** | 0.5 |
| **29** | 0.509804 |
| **30** | 0.484536 |
| **31** | 0.494845 |
| **32** | 0.504673 |
| **33** | 0.477064 |
| **34** | 0.486726 |
| **35** | 0.505155 |
| **36** | 0.520833 |
| **37** | 0.52381 |
| **38** | 0.49 |
| **39** | 0.494382 |
| **40** | 0.5 |
| **41** | 0.49505 |
| **42** | 0.535714 |
| **43** | 0.483516 |
| **44** | 0.46729 |
| **45** | 0.455357 |
| **46** | 0.505263 |
| **47** | 0.476636 |
| **48** | 0.489362 |
| **49** | 0.495327 |
| **50** | 0.505051 |

Tabel 4.8 Hasil perhitungan *Purity* setiap data (lanjutan)

|  |  |
| --- | --- |
| **51** | 0.429448 |
| **52** | 0.410256 |
| **53** | 0.420732 |
| **54** | 0.419847 |
| **55** | 0.422078 |
| **56** | 0.398601 |
| **57** | 0.396226 |
| **58** | 0.422414 |
| **59** | 0.428571 |
| **60** | 0.393939 |
| **61** | 0.434783 |
| **62** | 0.40411 |
| **63** | 0.454545 |
| **64** | 0.403974 |
| **65** | 0.41791 |
| **66** | 0.429487 |
| **67** | 0.383562 |
| **68** | 0.426471 |
| **69** | 0.430556 |
| **70** | 0.427481 |
| **71** | 0.375796 |
| **72** | 0.429577 |
| **73** | 0.414474 |
| **74** | 0.412162 |
| **75** | 0.42953 |
| **76** | 0.428571 |
| **77** | 0.43038 |
| **78** | 0.408537 |
| **79** | 0.402685 |
| **80** | 0.445313 |
| **81** | 0.429688 |
| **82** | 0.436508 |
| **83** | 0.426471 |
| **84** | 0.38961 |
| **85** | 0.375 |
| **86** | 0.387097 |
| **87** | 0.41875 |
| **88** | 0.440559 |
| **89** | 0.4 |
| **90** | 0.413534 |
| **91** | 0.40146 |

Tabel 4.9 Hasil perhitungan *Purity* setiap data (lanjutan)

|  |  |
| --- | --- |
| **92** | 0.403974 |
| **93** | 0.426471 |
| **94** | 0.431034 |
| **95** | 0.405797 |
| **96** | 0.404255 |
| **97** | 0.404255 |
| **98** | 0.421769 |
| **99** | 0.435897 |
| **100** | 0.410072 |
| **101** | 0.348066 |
| **102** | 0.374194 |
| **103** | 0.392265 |
| **104** | 0.379518 |
| **105** | 0.371429 |
| **106** | 0.393782 |
| **107** | 0.360294 |
| **108** | 0.398907 |
| **109** | 0.39881 |
| **110** | 0.371134 |
| **111** | 0.386905 |
| **112** | 0.392638 |
| **113** | 0.390805 |
| **114** | 0.375 |
| **115** | 0.360248 |
| **116** | 0.372093 |
| **117** | 0.386905 |
| **118** | 0.377451 |
| **119** | 0.394872 |
| **120** | 0.408163 |
| **121** | 0.381215 |
| **122** | 0.366013 |
| **123** | 0.401042 |
| **124** | 0.401274 |
| **125** | 0.376404 |
| **126** | 0.395604 |
| **127** | 0.397436 |
| **128** | 0.386076 |
| **129** | 0.378698 |
| **130** | 0.409091 |
| **131** | 0.406593 |
| **132** | 0.393035 |

Tabel 4.10 Hasil perhitungan *Purity* keseluruhan data (lanjutan)

|  |  |
| --- | --- |
| **133** | 0.376471 |
| **134** | 0.401274 |
| **135** | 0.388535 |
| **136** | 0.403141 |
| **137** | 0.355932 |
| **138** | 0.380952 |
| **139** | 0.384615 |
| **140** | 0.394286 |
| **141** | 0.376404 |
| **142** | 0.396552 |
| **143** | 0.374194 |
| **144** | 0.373626 |
| **145** | 0.368132 |
| **146** | 0.389535 |
| **147** | 0.401274 |
| **148** | 0.389222 |
| **149** | 0.358382 |
| **150** | 0.373418 |

Berdasarkan tabel diatas, data yang memiliki nilai *Purity* tertinggi pada 50 data pertama adalah data ke-42 dengan nilai 0.535714, nilai *Purity* tertinggi pada 50 data ke-2 adalah data ke-63 dengan nilai 0.454545 dan nilai *Purity* tertinggi pada 50 data terakhir adalah data ke-130 dengan nilai 0.409091. Maka data ke-42, data ke-63 dan ke-130 akan digunakan sebagai titik pusat *cluster* awal pada proses *clustering.* Titik pusat *cluster* awal dapat dilihat pada tabel 4.11 dibawah ini.

Tabel 4.11 Titik Pusat *cluster* (*Centroid*) awal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Titik pusat *cluster* (m) ke-*i*** | **Data Ke-** | **Atribut/Fitur** | | | |
| **Atribut 1 (*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** |
| ***m1*** | 42 | 4.5 | 2.3 | 1.3 | 0.3 |
| ***m2*** | 63 | 6 | 2.2 | 4 | 1 |
| ***m3*** | 130 | 7.2 | 3 | 5.8 | 1.6 |

* 1. **Hasil Titik Pusat *Cluster* Dengan Nilai *Purity***

Proses selanjutnya adalah perhitungan jarak data dengan titik pusat *cluster* menggunakan *Euclidean Distance*. Perhitungan jarak data ke-1 sampai data ke-10 terhadap titik pusat *cluster* adalah sebagai berikut:

1. Jarak data ke-1 dengan titik pusat *cluster*

𝐷(1,1) = √(5.1 − 4.5)2 + (3.5 − 2.3)2 + (1.4 − 1.3)2 + (0.2 − 0.3)2

= √(1.82) = 1.349074

𝐷(1,2) = √(5.1 − 6)2 + (3.5 − 2.2)2 + (1.4 − 4)2 + (0.2 − 1)2

= √(9.9) = 3.146427

𝐷(1,3) = √(5.1 − 7.2)2 + (3.5 − 3)2 + (1.4 − 5.8)2 + (0.2 − 1.6)2

= √(25.98) = 5.097058

1. Jarak data ke-2 dengan titik pusat *cluster*

𝐷(2,1) = √(4.9 − 4.5)2 + (3 − 2.3)2 + (1.4 − 1.3)2 + (0.2 − 0.3)2

= √(0.67) = 0.818535

𝐷(2,2) = √(4.9 − 6)2 + (3 − 2.2)2 + (1.4 − 4)2 + (0.2 − 1)2

= √(9.25) = 3.041381

𝐷(2,3) = √(4.9 − 7.2)2 + (3 − 3)2 + (1.4 − 5.8)2 + (0.2 − 1.6)2

= √(26.61) = 5.158488

1. Jarak data ke-3 dengan titik pusat *cluster*

𝐷(3,1) = √(4.7 − 4.5)2 + (3.2 − 2.3)2 + (1.3 − 1.3)2 + (0.2 − 0.3)2

= √(0.86) = 0.927362

𝐷(3,2) = √(4.7 − 6)2 + (3.2 − 2.2)2 + (1.3 − 4)2 + (0.2 − 1)2

= √(10.62) = 3.258834

𝐷(3,3) = √(4.7 − 7.2)2 + (3.2 − 3)2 + (1.3 − 5.8)2 + (0.2 − 1.6)2

= √(28.5) = 5.338539

1. Jarak data ke-4 dengan titik pusat *cluster*

𝐷(4,1) = √(4.6 − 4.5)2 + (3.1 − 2.3)2 + (1.5 − 1.3)2 + (0.2 − 0.3)2

= √(0.7) = 0.83666

𝐷(4,2) = √(4.6 − 6)2 + (3.1 − 2.2)2 + (1.5 − 4)2 + (0.2 − 1)2

= √(9.66) = 3.108054

𝐷(4,3) = √(4.6 − 7.2)2 + (3.1 − 3)2 + (1.5 − 5.8)2 + (0.2 − 1.6)2

= √(27.22) = 5.217279

1. Jarak data ke-5 dengan titik pusat *cluster*

𝐷(5,1) = √(5 − 4.5)2 + (3.6 − 2.3)2 + (1.4 − 1.3)2 + (0.2 − 0.3)2

= √(1.96) = 1.4

𝐷(5,2) = √(5 − 6)2 + (3.6 − 2.2)2 + (1.4 − 4)2 + (0.2 − 1)2

= √(10.36) = 3.218695

𝐷(5,3) = √(5 − 7.2)2 + (3.6 − 3)2 + (1.4 − 5.8)2 + (0.2 − 1.6)2

= √(26.52) = 5.149757

1. Jarak data ke-6 dengan titik pusat *cluster*

𝐷(6,1) = √(5.4 − 4.5)2 + (3.9 − 2.3)2 + (1.7 − 1.3)2 + (0.4 − 0.3)2

= √(3.54) = 1.881489

𝐷(6,2) = √(5.4 − 6)2 + (3.9 − 2.2)2 + (1.7 − 4)2 + (0.4 − 1)2

= √(8.9) = 2.983287

𝐷(6,3) = √(5.4 − 7.2)2 + (3.9 − 3)2 + (1.7 − 5.8)2 + (0.4 − 1.6)2

= √(22.3) = 4.722288

1. Jarak data ke-7 dengan titik pusat *cluster*

𝐷(7,1) = √(4.6 − 4.5)2 + (3.4 − 2.3)2 + (1.4 − 1.3)2 + (0.3 − 0.3)2

= √(1.23) = 1.109054

𝐷(7,2) = √(4.6 − 6)2 + (3.4 − 2.2)2 + (1.4 − 4)2 + (0.3 − 1)2

= √(10.65) = 3.263434

𝐷(7,3) = √(4.6 − 7.2)2 + (3.4 − 3)2 + (1.4 − 5.8)2 + (0.3 − 1.6)2

= √(27.97) = 5.288667

1. Jarak data ke-8 dengan titik pusat *cluster*

𝐷(8,1) = √(5 − 4.5)2 + (3.4 − 2.3)2 + (1.5 − 1.3)2 + (0.2 − 0.3)2

= √(1.51) = 1.228821

𝐷(8,2) = √(5 − 6)2 + (3.4 − 2.2)2 + (1.5 − 4)2 + (0.2 − 1)2

= √(9.33) = 3.054505

𝐷(8,3) = √(5 − 7.2)2 + (3.4 − 3)2 + (1.5 − 5.8)2 + (0.2 − 1.6)2

= √(25.45) = 5.044799

1. Jarak data ke-9 dengan titik pusat *cluster*

𝐷(9,1) = √(4.4 − 4.5)2 + (2.9 − 2.3)2 + (1.4 − 1.3)2 + (0.2 − 0.3)2

= √(0.39) = 0.6245

𝐷(9,2) = √(4.4 − 6)2 + (2.9 − 2.2)2 + (1.4 − 4)2 + (0.2 − 1)2

= √(10.45) = 3.232646

𝐷(9,3) = √(4.4 − 7.2)2 + (2.9 − 3)2 + (1.4 − 5.8)2 + (0.2 − 1.6)2

= √(29.17) = 5.400926

1. Jarak data ke-10 dengan titik pusat *cluster*

𝐷(10,1) = √(4.9 − 4.5)2 + (3.1 − 2.3)2 + (1.5 − 1.3)2 + (0.1 − 0.3)2

= √(0.88) = 0.938083

𝐷(10,2) = √(4.9 − 6)2 + (3.1 − 2.2)2 + (1.5 − 4)2 + (0.1 − 1)2

= √(9.08) = 3.013304

𝐷(10,3) = √(4.9 − 7.2)2 + (3.1 − 3)2 + (1.5 − 5.8)2 + (0.1 − 1.6)2

= √(26.04) = 5.10294

Perhitungan jarak dari setiap data dengan kedua titik pusat *cluster* dapat dilihat pada tabel

4.12 sebagai berikut.

Tabel 4.12 Perhitungan jarak data dengan titik pusat *cluster*

|  |  |  |  |
| --- | --- | --- | --- |
| **Data ke-*i*** | **Jarak data dengan titik pusat *cluster* 1** | **Jarak data dengan titik pusat *cluster* 2** | **Jarak data dengan titik pusat *cluster* 3** |
| **1** | 1.349074 | 3.146427 | 5.097058 |
| **2** | 0.818535 | 3.041381 | 5.158488 |
| **3** | 0.927362 | 3.258834 | 5.338539 |
| **4** | 0.83666 | 3.108054 | 5.217279 |
| **5** | 1.4 | 3.218695 | 5.149757 |
| **6** | 1.881489 | 2.983287 | 4.722288 |
| **7** | 1.109054 | 3.263434 | 5.288667 |
| **8** | 1.228821 | 3.054505 | 5.044799 |
| **9** | 0.6245 | 3.232646 | 5.400926 |

Tabel 4.13 Perhitungan jarak data dengan titik pusat *cluster* (lanjutan)

|  |  |  |  |
| --- | --- | --- | --- |
| **10** | 0.938083 | 3.013304 | 5.10294 |
| **11** | 1.679286 | 3.082207 | 4.917316 |
| **12** | 1.183216 | 3.046309 | 5.051732 |
| **13** | 0.793725 | 3.106445 | 5.231635 |
| **14** | 0.781025 | 3.570714 | 5.722762 |
| **15** | 2.144761 | 3.429286 | 5.106858 |
| **16** | 2.428992 | 3.397058 | 4.913247 |
| **17** | 1.838478 | 3.301515 | 5.07346 |
| **18** | 1.345362 | 3.122499 | 5.070503 |
| **19** | 1.962142 | 2.903446 | 4.624932 |
| **20** | 1.627882 | 3.179623 | 5.022947 |
| **21** | 1.479865 | 2.780288 | 4.708503 |
| **22** | 1.53948 | 3.109662 | 4.982971 |
| **23** | 1.341641 | 3.682391 | 5.667451 |
| **24** | 1.249 | 2.749545 | 4.745524 |
| **25** | 1.292285 | 2.816026 | 4.805206 |
| **26** | 0.916515 | 2.835489 | 4.943683 |
| **27** | 1.249 | 2.925748 | 4.907138 |
| **28** | 1.407125 | 3.036445 | 4.969909 |
| **29** | 1.311488 | 3.078961 | 5.047772 |
| **30** | 0.974679 | 3.014963 | 5.088222 |
| **31** | 0.911043 | 2.941088 | 5.036864 |
| **32** | 1.438749 | 2.9 | 4.830114 |
| **33** | 1.951922 | 3.363034 | 5.094114 |
| **34** | 2.151743 | 3.41321 | 5.064583 |
| **35** | 0.921954 | 2.984962 | 5.074446 |
| **36** | 1.03923 | 3.237283 | 5.291503 |
| **37** | 1.565248 | 3.141656 | 5.034878 |
| **38** | 1.378405 | 3.277194 | 5.221111 |
| **39** | 0.714143 | 3.336165 | 5.481788 |
| **40** | 1.272792 | 3.023243 | 5.002 |
| **41** | 1.3 | 3.235738 | 5.199038 |
| **42** | 0 | 3.168596 | 5.451605 |
| **43** | 0.911043 | 3.38969 | 5.485435 |
| **44** | 1.367479 | 2.93428 | 4.871345 |
| **45** | 1.726268 | 2.853069 | 4.658326 |
| **46** | 0.768115 | 3.054505 | 5.177837 |
| **47** | 1.646208 | 3.1257 | 4.964877 |

Tabel 4.14 Perhitungan jarak data dengan titik pusat *cluster* (lanjutan)

|  |  |  |  |
| --- | --- | --- | --- |
| **48** | 0.916515 | 3.218695 | 5.302829 |
| **49** | 1.627882 | 3.103224 | 4.954796 |
| **50** | 1.126943 | 3.1 | 5.123475 |
| **51** | 4.453089 | 1.627882 | 1.153256 |
| **52** | 4.012481 | 1.28841 | 1.542725 |
| **53** | 4.560702 | 1.637071 | 0.959166 |
| **54** | 3.04795 | 0.591608 | 2.590367 |
| **55** | 4.071855 | 1.104536 | 1.407125 |
| **56** | 3.595831 | 0.888819 | 2.017424 |
| **57** | 4.182105 | 1.466288 | 1.452584 |
| **58** | 2.158703 | 1.319091 | 3.501428 |
| **59** | 4.081666 | 1.140175 | 1.378405 |
| **60** | 2.935984 | 1.029563 | 2.782086 |
| **61** | 2.381176 | 1.135782 | 3.38969 |
| **62** | 3.507136 | 0.969536 | 2.063977 |
| **63** | 3.168596 | 0 | 2.383275 |
| **64** | 3.96106 | 1.072381 | 1.571623 |
| **65** | 2.803569 | 0.948683 | 2.738613 |
| **66** | 4.037326 | 1.272792 | 1.50333 |
| **67** | 3.657868 | 1.140175 | 2.063977 |
| **68** | 3.190611 | 0.547723 | 2.302173 |
| **69** | 3.818377 | 0.734847 | 1.827567 |
| **70** | 2.941088 | 0.519615 | 2.582634 |
| **71** | 4.155719 | 1.513275 | 1.664332 |
| **72** | 3.331666 | 0.678233 | 2.140093 |
| **73** | 4.204759 | 1.113553 | 1.371131 |
| **74** | 3.896152 | 0.948683 | 1.618641 |
| **75** | 3.737646 | 0.911043 | 1.729162 |
| **76** | 3.964846 | 1.148913 | 1.536229 |
| **77** | 4.358899 | 1.341641 | 1.113553 |
| **78** | 4.580393 | 1.618641 | 0.948683 |
| **79** | 3.780212 | 0.994987 | 1.774824 |
| **80** | 2.61916 | 0.707107 | 2.839014 |
| **81** | 2.810694 | 0.583095 | 2.738613 |
| **82** | 2.694439 | 0.616441 | 2.83196 |
| **83** | 3.069202 | 0.583095 | 2.412468 |
| **84** | 4.30581 | 1.349074 | 1.421267 |
| **85** | 3.602777 | 1.224745 | 2.222611 |

Tabel 4.15 Perhitungan jarak data dengan titik pusat *cluster* (lanjutan)

|  |  |  |  |
| --- | --- | --- | --- |
| **86** | 3.923009 | 1.431782 | 1.813836 |
| **87** | 4.298837 | 1.428286 | 1.216553 |
| **88** | 3.721559 | 0.591608 | 1.830301 |
| **89** | 3.246537 | 0.948683 | 2.35372 |
| **90** | 3.054505 | 0.655744 | 2.543619 |
| **91** | 3.392639 | 0.781025 | 2.273763 |
| **92** | 3.8923 | 1.081665 | 1.640122 |
| **93** | 3.143247 | 0.489898 | 2.349468 |
| **94** | 2.177154 | 1.224745 | 3.455431 |
| **95** | 3.283291 | 0.734847 | 2.302173 |
| **96** | 3.339162 | 0.9 | 2.22935 |
| **97** | 3.348134 | 0.842615 | 2.215852 |
| **98** | 3.640055 | 0.842615 | 1.830301 |
| **99** | 1.982423 | 1.382027 | 3.570714 |
| **100** | 3.244996 | 0.74162 | 2.295648 |
| **101** | 5.58301 | 2.747726 | 1.322876 |
| **102** | 4.341659 | 1.519868 | 1.621727 |
| **103** | 5.625833 | 2.582634 | 0.519615 |
| **104** | 4.933559 | 1.944222 | 0.948683 |
| **105** | 5.324472 | 2.360085 | 0.921954 |
| **106** | 6.436614 | 3.342155 | 1.024695 |
| **107** | 3.521363 | 1.428286 | 2.690725 |
| **108** | 5.95399 | 2.847806 | 0.556776 |
| **109** | 5.23259 | 2.111871 | 0.734847 |
| **110** | 6.071244 | 3.17175 | 1.122497 |
| **111** | 4.705316 | 1.860108 | 1.086278 |
| **112** | 4.725463 | 1.705872 | 1.034408 |
| **113** | 5.163332 | 2.177154 | 0.707107 |
| **114** | 4.249706 | 1.476482 | 1.81659 |
| **115** | 4.559605 | 1.889444 | 1.769181 |
| **116** | 4.94166 | 2.130728 | 1.191638 |
| **117** | 4.937611 | 1.944222 | 0.787401 |
| **118** | 6.727555 | 3.765634 | 1.43527 |
| **119** | 6.759438 | 3.626293 | 1.452584 |
| **120** | 4.170132 | 1.118034 | 1.652271 |
| **121** | 5.470832 | 2.527845 | 0.793725 |
| **122** | 4.160529 | 1.526434 | 1.889444 |
| **123** | 6.52227 | 3.397058 | 1.122497 |

Tabel 4.16 Perhitungan jarak data dengan titik pusat *cluster* (lanjutan)

|  |  |  |  |
| --- | --- | --- | --- |
| **124** | 4.313931 | 1.337909 | 1.322876 |
| **125** | 5.332917 | 2.408319 | 0.774597 |
| **126** | 5.695612 | 2.660827 | 0.34641 |
| **127** | 4.2 | 1.296148 | 1.442221 |
| **128** | 4.273172 | 1.449138 | 1.43527 |
| **129** | 5.058656 | 2.071232 | 0.984886 |
| **130** | 5.451605 | 2.383275 | 0 |
| **131** | 5.853204 | 2.745906 | 0.509902 |
| **132** | 6.535289 | 3.595831 | 1.284523 |
| **133** | 5.095096 | 2.126029 | 1.03923 |
| **134** | 4.401136 | 1.382027 | 1.161895 |
| **135** | 4.727579 | 1.7 | 1.204159 |
| **136** | 6.14573 | 3.103224 | 0.911043 |
| **137** | 5.229723 | 2.459675 | 1.284523 |
| **138** | 4.913247 | 1.964688 | 0.883176 |
| **139** | 4.152108 | 1.385641 | 1.574802 |
| **140** | 5.142956 | 2.188607 | 0.714143 |
| **141** | 5.327288 | 2.412468 | 0.969536 |
| **142** | 4.983974 | 2.126029 | 1.03923 |
| **143** | 4.341659 | 1.519868 | 1.621727 |
| **144** | 5.591064 | 2.634388 | 0.83666 |
| **145** | 5.480876 | 2.615339 | 1.077033 |
| **146** | 4.953786 | 2.063977 | 1.048809 |
| **147** | 4.419276 | 1.410674 | 1.337909 |
| **148** | 4.752894 | 1.824829 | 1.004988 |
| **149** | 4.990992 | 2.26495 | 1.345362 |
| **150** | 4.374929 | 1.581139 | 1.489966 |

Langkah selanjutnya adalah menentukan *centroid* baru dengan menghitung nilai *mean* pada setiap *cluster* yang ada. Setelah diperoleh nilai *centroid* yang baru, maka jarak data terhadap *centroid* kembali dihitung. Iterasi selanjutnya dilakukan dengan cara yang sama, dan akan berhenti sampai tidak ada lagi perubahan data anggota *cluster* didalam sebuah *cluster*, atau tidak terjadi lagi perpindahan data dari satu *cluster* ke *cluster* lainnya.

## Hasil Pengujian

Setelah tidak terjadi lagi perpindahan data dari satu *cluster* ke *cluster* lainnya, maka proses *clustering* dihentikan. Pada pengujian yang dilakukan pada dataset *Iris* dengan *Purity K-Means* berhenti pada iterasi ke-5. Berikut merupakan hasil akhir pengujian *Purity K-Means* yang dilakukan pada dataset *Iris*.

Tabel 4.17 Hasil Akhir Pengujian *Purity K-Means* pada dataset *Iris*

|  |  |
| --- | --- |
| **Data**  **ke-*i*** | **Hasil Pengelompokan** |
| **1** | Setosa |
| **2** | Setosa |
| **3** | Setosa |
| **4** | Setosa |
| **5** | Setosa |
| **6** | Setosa |
| **7** | Setosa |
| **8** | Setosa |
| **9** | Setosa |
| **10** | Setosa |
| **11** | Setosa |
| **12** | Setosa |
| **13** | Setosa |
| **14** | Setosa |
| **15** | Setosa |
| **16** | Setosa |
| **17** | Setosa |
| **18** | Setosa |
| **19** | Setosa |
| **20** | Setosa |
| **21** | Setosa |
| **22** | Setosa |
| **23** | Setosa |
| **24** | Setosa |
| **25** | Setosa |
| **26** | Setosa |
| **27** | Setosa |

Tabel 4.18 Hasil Akhir Pengujian *Purity K-Means* pada dataset *Iris* (lanjutan)

|  |  |
| --- | --- |
| **28** | Setosa |
| **29** | Setosa |
| **30** | Setosa |
| **31** | Setosa |
| **32** | Setosa |
| **33** | Setosa |
| **34** | Setosa |
| **35** | Setosa |
| **36** | Setosa |
| **37** | Setosa |
| **38** | Setosa |
| **39** | Setosa |
| **40** | Setosa |
| **41** | Setosa |
| **42** | Setosa |
| **43** | Setosa |
| **44** | Setosa |
| **45** | Setosa |
| **46** | Setosa |
| **47** | Setosa |
| **48** | Setosa |
| **49** | Setosa |
| **50** | Setosa |
| **51** | Virginica |
| **52** | Versicolor |
| **53** | Virginica |
| **54** | Versicolor |
| **55** | Versicolor |
| **56** | Versicolor |
| **57** | Versicolor |
| **58** | Versicolor |
| **59** | Versicolor |
| **60** | Versicolor |
| **61** | Versicolor |
| **62** | Versicolor |
| **63** | Versicolor |
| **64** | Versicolor |
| **65** | Versicolor |

Tabel 4.19 Hasil Akhir Pengujian *Purity K-Means* pada dataset *Iris* (lanjutan)

|  |  |
| --- | --- |
| **66** | Versicolor |
| **67** | Versicolor |
| **68** | Versicolor |
| **69** | Versicolor |
| **70** | Versicolor |
| **71** | Versicolor |
| **72** | Versicolor |
| **73** | Versicolor |
| **74** | Versicolor |
| **75** | Versicolor |
| **76** | Versicolor |
| **77** | Versicolor |
| **78** | Virginica |
| **79** | Versicolor |
| **80** | Versicolor |
| **81** | Versicolor |
| **82** | Versicolor |
| **83** | Versicolor |
| **84** | Versicolor |
| **85** | Versicolor |
| **86** | Versicolor |
| **87** | Versicolor |
| **88** | Versicolor |
| **89** | Versicolor |
| **90** | Versicolor |
| **91** | Versicolor |
| **92** | Versicolor |
| **93** | Versicolor |
| **94** | Versicolor |
| **95** | Versicolor |
| **96** | Versicolor |
| **97** | Versicolor |
| **98** | Versicolor |
| **99** | Versicolor |
| **100** | Versicolor |
| **101** | Virginica |
| **102** | Versicolor |
| **103** | Virginica |

Tabel 4.20 Hasil Akhir Pengujian *Purity K-Means* pada dataset *Iris* (lanjutan)

|  |  |
| --- | --- |
| **104** | Virginica |
| **105** | Virginica |
| **106** | Virginica |
| **107** | Versicolor |
| **108** | Virginica |
| **109** | Virginica |
| **110** | Virginica |
| **111** | Virginica |
| **112** | Virginica |
| **113** | Virginica |
| **114** | Versicolor |
| **115** | Versicolor |
| **116** | Virginica |
| **117** | Virginica |
| **118** | Virginica |
| **119** | Virginica |
| **120** | Versicolor |
| **121** | Virginica |
| **122** | Versicolor |
| **123** | Virginica |
| **124** | Versicolor |
| **125** | Virginica |
| **126** | Virginica |
| **127** | Versicolor |
| **128** | Versicolor |
| **129** | Virginica |
| **130** | Virginica |
| **131** | Virginica |
| **132** | Virginica |
| **133** | Virginica |
| **134** | Versicolor |
| **135** | Virginica |
| **136** | Virginica |
| **137** | Virginica |
| **138** | Virginica |
| **139** | Versicolor |
| **140** | Virginica |
| **141** | Virginica |

Tabel 4.21 Hasil Akhir Pengujian *Purity K-Means* pada dataset *Iris* (lanjutan)

|  |  |
| --- | --- |
| **142** | Virginica |
| **143** | Versicolor |
| **144** | Virginica |
| **145** | Virginica |
| **146** | Virginica |
| **147** | Versicolor |
| **148** | Virginica |
| **149** | Virginica |
| **150** | Versicolor |

Tahap selanjutnya adalah membandingkan hasil *clustering Purity K-Means* dengan metode *K-Means* secara konvensional yang dilakukan pada dataset *Iris* dengan keseluruhan 10 kali percobaan. Adapun hasil perbandingan *clustering* tersebut dapat dilihat pada tabel berikut.

Tabel 4.22 Perbandingan hasil *clustering* pada dataset *Iris*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data  ke-i | Data Asli | Hasil K-Means  (Percobaan ke 1-9) | Hasil K-Means  (Percobaan ke-10) | Purity K-Means |
| 1 | Setosa | Setosa | Setosa | Setosa |
| 2 | Setosa | Setosa | Setosa | Setosa |
| 3 | Setosa | Setosa | Setosa | Setosa |
| 4 | Setosa | Setosa | Setosa | Setosa |
| 5 | Setosa | Setosa | Setosa | Setosa |
| 6 | Setosa | Setosa | Setosa | Setosa |
| 7 | Setosa | Setosa | Setosa | Setosa |
| 8 | Setosa | Setosa | Setosa | Setosa |
| 9 | Setosa | Setosa | Setosa | Setosa |
| 10 | Setosa | Setosa | Setosa | Setosa |
| 11 | Setosa | Setosa | Setosa | Setosa |
| 12 | Setosa | Setosa | Setosa | Setosa |
| 13 | Setosa | Setosa | Setosa | Setosa |
| 14 | Setosa | Setosa | Setosa | Setosa |
| 15 | Setosa | Setosa | Setosa | Setosa |
| 16 | Setosa | Setosa | Setosa | Setosa |
| 17 | Setosa | Setosa | Setosa | Setosa |
| 18 | Setosa | Setosa | Setosa | Setosa |
| 19 | Setosa | Setosa | Setosa | Setosa |
| 20 | Setosa | Setosa | Setosa | Setosa |
| 21 | Setosa | Setosa | Setosa | Setosa |
| 22 | Setosa | Setosa | Setosa | Setosa |

Tabel 4.23 Perbandingan hasil *clustering* pada dataset *Iris* (lanjutan)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 23 | Setosa | Setosa | Setosa | Setosa |
| 24 | Setosa | Setosa | Setosa | Setosa |
| 25 | Setosa | Setosa | Setosa | Setosa |
| 26 | Setosa | Setosa | Setosa | Setosa |
| 27 | Setosa | Setosa | Setosa | Setosa |
| 28 | Setosa | Setosa | Setosa | Setosa |
| 29 | Setosa | Setosa | Setosa | Setosa |
| 30 | Setosa | Setosa | Setosa | Setosa |
| 31 | Setosa | Setosa | Setosa | Setosa |
| 32 | Setosa | Setosa | Setosa | Setosa |
| 33 | Setosa | Setosa | Setosa | Setosa |
| 34 | Setosa | Setosa | Setosa | Setosa |
| 35 | Setosa | Setosa | Setosa | Setosa |
| 36 | Setosa | Setosa | Setosa | Setosa |
| 37 | Setosa | Setosa | Setosa | Setosa |
| 38 | Setosa | Setosa | Setosa | Setosa |
| 39 | Setosa | Setosa | Setosa | Setosa |
| 40 | Setosa | Setosa | Setosa | Setosa |
| 41 | Setosa | Setosa | Setosa | Setosa |
| 42 | Setosa | Setosa | Setosa | Setosa |
| 43 | Setosa | Setosa | Setosa | Setosa |
| 44 | Setosa | Setosa | Setosa | Setosa |
| 45 | Setosa | Setosa | Setosa | Setosa |
| 46 | Setosa | Setosa | Setosa | Setosa |
| 47 | Setosa | Setosa | Setosa | Setosa |
| 48 | Setosa | Setosa | Setosa | Setosa |
| 49 | Setosa | Setosa | Setosa | Setosa |
| 50 | Setosa | Setosa | Setosa | Setosa |
| 51 | Versicolor | Virginica | Versicolor | Virginica |
| 52 | Versicolor | Versicolor | Setosa | Versicolor |
| 53 | Versicolor | Virginica | Setosa | Virginica |
| 54 | Versicolor | Versicolor | Setosa | Versicolor |
| 55 | Versicolor | Versicolor | Setosa | Versicolor |
| 56 | Versicolor | Versicolor | Versicolor | Versicolor |
| 57 | Versicolor | Versicolor | Setosa | Versicolor |
| 58 | Versicolor | Versicolor | Setosa | Versicolor |
| 59 | Versicolor | Versicolor | Setosa | Versicolor |
| 60 | Versicolor | Versicolor | Setosa | Versicolor |
| 61 | Versicolor | Versicolor | Setosa | Versicolor |
| 62 | Versicolor | Versicolor | Setosa | Versicolor |
| 63 | Versicolor | Versicolor | Setosa | Versicolor |
| 64 | Versicolor | Versicolor | Setosa | Versicolor |
| 65 | Versicolor | Versicolor | Setosa | Versicolor |

Tabel 4.24 Perbandingan hasil *clustering* pada dataset *Iris* (lanjutan)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 66 | Versicolor | Versicolor | Setosa | Versicolor |
| 67 | Versicolor | Versicolor | Setosa | Versicolor |
| 68 | Versicolor | Versicolor | Setosa | Versicolor |
| 69 | Versicolor | Versicolor | Setosa | Versicolor |
| 70 | Versicolor | Versicolor | Setosa | Versicolor |
| 71 | Versicolor | Versicolor | Setosa | Versicolor |
| 72 | Versicolor | Versicolor | Setosa | Versicolor |
| 73 | Versicolor | Versicolor | Setosa | Versicolor |
| 74 | Versicolor | Versicolor | Setosa | Versicolor |
| 75 | Versicolor | Versicolor | Setosa | Versicolor |
| 76 | Versicolor | Versicolor | Setosa | Versicolor |
| 77 | Versicolor | Versicolor | Setosa | Versicolor |
| 78 | Versicolor | Virginica | Setosa | Virginica |
| 79 | Versicolor | Versicolor | Setosa | Versicolor |
| 80 | Versicolor | Versicolor | Setosa | Versicolor |
| 81 | Versicolor | Versicolor | Setosa | Versicolor |
| 82 | Versicolor | Versicolor | Setosa | Versicolor |
| 83 | Versicolor | Versicolor | Setosa | Versicolor |
| 84 | Versicolor | Versicolor | Setosa | Versicolor |
| 85 | Versicolor | Versicolor | Setosa | Versicolor |
| 86 | Versicolor | Versicolor | Setosa | Versicolor |
| 87 | Versicolor | Versicolor | Setosa | Versicolor |
| 88 | Versicolor | Versicolor | Setosa | Versicolor |
| 89 | Versicolor | Versicolor | Versicolor | Versicolor |
| 90 | Versicolor | Versicolor | Setosa | Versicolor |
| 91 | Versicolor | Versicolor | Setosa | Versicolor |
| 92 | Versicolor | Versicolor | Versicolor | Versicolor |
| 93 | Versicolor | Versicolor | Setosa | Versicolor |
| 94 | Versicolor | Versicolor | Setosa | Versicolor |
| 95 | Versicolor | Versicolor | Setosa | Versicolor |
| 96 | Versicolor | Versicolor | Setosa | Versicolor |
| 97 | Versicolor | Versicolor | Setosa | Versicolor |
| 98 | Versicolor | Versicolor | Setosa | Versicolor |
| 99 | Versicolor | Versicolor | Setosa | Versicolor |
| 100 | Versicolor | Versicolor | Setosa | Versicolor |
| 101 | Virginica | Virginica | Virginica | Virginica |
| 102 | Virginica | Versicolor | Virginica | Versicolor |
| 103 | Virginica | Virginica | Versicolor | Virginica |
| 104 | Virginica | Virginica | Virginica | Virginica |
| 105 | Virginica | Virginica | Versicolor | Virginica |
| 106 | Virginica | Virginica | Virginica | Virginica |
| 107 | Virginica | Versicolor | Virginica | Versicolor |
| 108 | Virginica | Virginica | Versicolor | Virginica |

Tabel 4.25 Perbandingan hasil *clustering* pada dataset *Iris* (lanjutan)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 109 | Virginica | Virginica | Versicolor | Virginica |
| 110 | Virginica | Virginica | Virginica | Virginica |
| 111 | Virginica | Virginica | Virginica | Virginica |
| 112 | Virginica | Virginica | Versicolor | Virginica |
| 113 | Virginica | Virginica | Virginica | Virginica |
| 114 | Virginica | Versicolor | Virginica | Versicolor |
| 115 | Virginica | Versicolor | Virginica | Versicolor |
| 116 | Virginica | Virginica | Versicolor | Virginica |
| 117 | Virginica | Virginica | Virginica | Virginica |
| 118 | Virginica | Virginica | Virginica | Virginica |
| 119 | Virginica | Virginica | Virginica | Virginica |
| 120 | Virginica | Versicolor | Versicolor | Versicolor |
| 121 | Virginica | Virginica | Versicolor | Virginica |
| 122 | Virginica | Versicolor | Virginica | Versicolor |
| 123 | Virginica | Virginica | Virginica | Virginica |
| 124 | Virginica | Versicolor | Virginica | Versicolor |
| 125 | Virginica | Virginica | Versicolor | Virginica |
| 126 | Virginica | Virginica | Versicolor | Virginica |
| 127 | Virginica | Versicolor | Virginica | Versicolor |
| 128 | Virginica | Versicolor | Virginica | Versicolor |
| 129 | Virginica | Virginica | Virginica | Virginica |
| 130 | Virginica | Virginica | Virginica | Virginica |
| 131 | Virginica | Virginica | Virginica | Virginica |
| 132 | Virginica | Virginica | Virginica | Virginica |
| 133 | Virginica | Virginica | Virginica | Virginica |
| 134 | Virginica | Versicolor | Virginica | Versicolor |
| 135 | Virginica | Virginica | Virginica | Virginica |
| 136 | Virginica | Virginica | Virginica | Virginica |
| 137 | Virginica | Virginica | Versicolor | Virginica |
| 138 | Virginica | Virginica | Versicolor | Virginica |
| 139 | Virginica | Versicolor | Virginica | Versicolor |
| 140 | Virginica | Virginica | Virginica | Virginica |
| 141 | Virginica | Virginica | Versicolor | Virginica |
| 142 | Virginica | Virginica | Versicolor | Virginica |
| 143 | Virginica | Versicolor | Virginica | Versicolor |
| 144 | Virginica | Virginica | Versicolor | Virginica |
| 145 | Virginica | Virginica | Virginica | Virginica |
| 146 | Virginica | Virginica | Virginica | Virginica |
| 147 | Virginica | Versicolor | Versicolor | Versicolor |
| 148 | Virginica | Virginica | Versicolor | Virginica |
| 149 | Virginica | Virginica | Versicolor | Virginica |
| 150 | Virginica | Versicolor | Virginica | Versicolor |
| Jumlah Salah | | 17 | 64 | 17 |

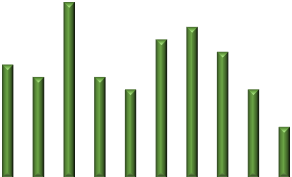
Tabel 4.26 Perbandingan hasil *clustering* pada dataset *Iris* (lanjutan)

|  |  |  |  |
| --- | --- | --- | --- |
| Jumlah Benar | 133 | 86 | 133 |
| Akurasi | 133 𝑥100% = 88.7% | 86 𝑥100% = 57.3% | 133 𝑥100% = 88.7% |
| 150 | 150 | 150 |

Pada tabel diatas, diketahui bahwa hasil *clustering* pada percobaan ke-1 sampai ke-9 dengan metode *K-Means* secara konvensional memperoleh hasil akurasi yang sama dengan percobaan yang dilakukan dengan metode *Purity K-Means* yaitu sebesar 88.7%. Sedangkan pada percobaan ke-10 menggunakan metode *K-Means* secara konvensional memperoleh hasil akurasi yang berbeda, yaitu hanya sebesar 57.3%. Untuk total akurasi keseluruhan dari 10 kali percobaan yang dilakukan dengan metode *K-Means* konvensional adalah sebesar 85.5%. *Purity K-Means* memperoleh tingkat akurasi yang lebih baik sebesar 3.2% dibandingkan *K-Means* secara konvensional dengan 10 kali percobaan yang dilakukan.

## Jumlah Iterasi pada dataset *Iris*

Pengujian *clustering* pada *dataset Iris* dilakukan sebanyak 10 kali pengujian. Grafik hasil *clustering* pada *dataset Iris* dapat dilihat pada gambar 4.1 dibawah ini.



Jumlah Iterasi

16

14

12

10

8

K-Means

K-Means Purity

6

4

2

0

Pengujian

1 2 3 4 5 6 7 8 9 10

**Gambar 4.1. Grafik hasil *clustering* pada *dataset Iris***

Berdasarkan gambar 4.1, jumlah iterasi yang dibutuhkan untuk 10 kali pengujian

*clustering* yang dilakukan pada *dataset Iris* yang berjumlah 150 data menggunakan

metode *K-Means* konvensional memiliki jumlah yang berbeda pada setiap pengujiannya. Pengujian ke-3 pada *K-Means* konvensional memiliki jumlah iterasi tertinggi untuk menyelesaikan proses *clustering* pada *dataset iris* dengan jumlah 14 iterasi, sedangkan pengujian ke-10 memperoleh jumlah iterasi terendah dengan 4 iterasi. Sedangkan dengan menggunakan metode *K-Means* yang diusulkan, iterasi yang dibutuhkan untuk proses *clustering dataset iris* berjumlah 5 iterasi.

* 1. **Nilai *Purity* pada dataset lain**

Untuk melakukan proses *clustering*, diperlukan titik pusat *cluster* (centroid) awal yang sesuai dengan banyak *cluster* (kelompok) yang diinginkan dari data yang akan diuji. Pada penelitian ini, untuk mencari dan menentukan titik pusat *cluster* (centroid) awal, seluruh data pada *dataset* yang akan diuji akan dihitung nilai *Purity-*nya secara keseluruhan terlebih. Adapun hasil perhitungan nilai *Purity* untuk masing-masing *dataset* dapat dilihat pada tabel berikut.

Tabel 4.27 Hasil Perhitungan Nilai *Purity* pada *dataset 2*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data ke-*i* (*pi*)** | **Atribut 1 (*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** | **Atribut 5 (*x5*)** | **Nilai**  ***Purity*** |
| **1** | 0.08 | 0.08 | 0.1 | 0.24 | 0.9 | **0.642857** |
| **2** | 0.06 | 0.06 | 0.05 | 0.25 | 0.33 | **0.44** |
| **3** | 0.1 | 0.1 | 0.15 | 0.65 | 0.3 | **0.5** |
| **4** | 0.08 | 0.08 | 0.08 | 0.98 | 0.24 | **0.671233** |
| **5** | 0.09 | 0.15 | 0.4 | 0.1 | 0.66 | **0.471429** |
| **...** | ... | ... | ... | ... | ... | **...** |
| **15** | 0.05 | 0.07 | 0.7 | 0.01 | 0.05 | **0.795455** |
| **...** | ... | ... | ... | ... | ... | **...** |
| **149** | 0.334 | 0.295 | 0.33 | 0.32 | 0.3 | **0.211526** |
| **...** | ... | ... | ... | ... | ... | **...** |
| **395** | 0.68 | 0.64 | 0.79 | 0.97 | 0.24 | **0.292169** |

Pada tabel 4.27 diatas, pada *dataset 2*, yaitu *dataset User Knowledge Modeling* diperoleh nilai *Purity* untuk keseluruhan data pada *dataset* tersebut. Data ke-15 memiliki nilai *Purity* tertinggi sebesar 0.795455, dan data ke-149 memiliki nilai

*Purity* terendah sebesar 0.211526. Maka dari itu, data ke-15 dan data ke-149 yang akan digunakan sebagai titik pusat *cluster* awal pada *dataset 2*.

Tabel 4.28 Hasil Perhitungan Nilai *Purity* pada *dataset 3*

|  |  |  |  |
| --- | --- | --- | --- |
| **Data ke-*i***  **(*pi*)** | **Atribut 1**  **(*x1*)** | **Atribut 2**  **(*x2*)** | **Nilai**  ***Purity*** |
| **1** | 36.4 | 14.6 | **0.713725** |
| **2** | 37.3 | 8 | **0.8234** |
| **3** | 42.1 | 15.3 | **0.733449** |
| **4** | 55.8 | 25.6 | **0.685504** |
| **5** | 56.1 | 33.1 | **0.628924** |
| **...** | ... | ... | **...** |
| **19** | 46.3 | 6.4 | **0.878558** |
| **...** | ... | ... | **...** |
| **50** | 17.6 | 19.8 | **0.529412** |
| **...** | ... | ... | **...** |
| **70** | 25.5 | 8.8 | **0.74344** |

Pada tabel 4.28, *dataset 3*, yaitu *dataset Death and Birth Rates* diperoleh nilai *Purity* untuk keseluruhan data pada *dataset* tersebut. Data ke-19 memiliki nilai *Purity* tertinggi sebesar 0.878558, dan data ke-50 memiliki nilai *Purity* terendah sebesar 0.529412. Maka dari itu, data ke-15 dan data ke-50 yang akan digunakan sebagai titik pusat *cluster* awal pada *dataset 3*.

Tabel 4.29 Hasil Perhitungan Nilai *Purity* pada *dataset 4*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Data ke-*i***  **(*pi*)** | **Atribut 1 (*x1*)** | **Atribut 2 (*x2*)** | **Atribut 3 (*x3*)** | **Atribut 4 (*x4*)** | **Atribut 5 (*x5*)** | **Atribut 6 (*x6*)** | **Nilai**  ***Purity*** |
| **1** | 12669 | 9656 | 7561 | 214 | 2674 | 1338 | **0.371394** |
| **2** | 7057 | 9810 | 9568 | 1762 | 3293 | 1776 | **0.294896** |
| **3** | 6353 | 8808 | 7684 | 2405 | 3516 | 7844 | **0.24059** |
| **4** | 13265 | 1196 | 4221 | 6404 | 507 | 1788 | **0.48446** |
| **5** | 22615 | 5410 | 7198 | 3915 | 1777 | 5185 | **0.490564** |
| **...** | ... | ... | ... | ... | ... | ... | **...** |
| **290** | 42786 | 286 | 471 | 1388 | 32 | 22 | **0.951117** |
| **...** | ... | ... | ... | ... | ... | ... | **...** |
| **440** | 2787 | 1698 | 2510 | 65 | 477 | 52 | **0.367242** |

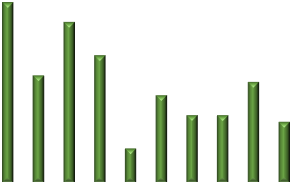
Pada tabel 4.29, *dataset 4*, yaitu *dataset Wholesale Customers* diperoleh nilai *Purity* untuk keseluruhan data pada *dataset* tersebut. Data ke-290 memiliki nilai *Purity* tertinggi sebesar 0.951117, dan data ke-3 memiliki nilai *Purity* terendah sebesar 0.24059. Maka dari itu, data ke-290 dan data ke-3 yang akan digunakan sebagai titik pusat *cluster* awal pada *dataset 4*.

* 1. **Hasil *Clustering* pada dataset lain**

Pengujian *clustering* untuk setiap dataset dilakukan masing-masing sebanyak 10 kali pengujian. Pengujian dilakukan dengan menggunakan metode *K-Means* konvensional sebanyak 10 kali lalu di uji dengan metode *K-Means* yang diusulkan. Hasil *clustering* adalah kumpulan data pada sebuah *dataset* yang telah di *cluster* (dikelompokkan) sesuai dengan jumlah *cluster* (kelompok) yang ditentukan.

* + 1. **Jumlah Iterasi Pada *Dataset 2***

Pengujian *clustering* pada *dataset User Knowledge Modeling* dilakukan sebanyak 10 kali pengujian. Grafik hasil *clustering* pada *dataset 2* dapat dilihat pada gambar 4.2 dibawah ini.



Jumlah Iterasi

30

25

20

15

K-Means

K-Means Purity

10

5

0

1 2 3 4 5 6 7 8 9

10

Pengujian

**Gambar 4.2. Grafik hasil *clustering* pada *dataset 2***

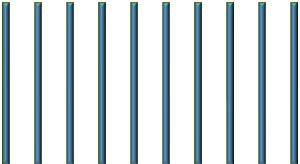
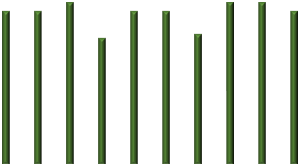
Berdasarkan gambar 4.2, jumlah iterasi yang dibutuhkan untuk 10 kali pengujian

*clustering* yang dilakukan pada *dataset User Knowledge Modeling* yang berjumlah

395 data menggunakan metode *K-Means* konvensional memiliki jumlah yang berbeda pada setiap pengujiannya. Pengujian pertama pada *K-Means* konvensional memiliki jumlah iterasi tertinggi untuk menyelesaikan proses *clustering* pada *dataset 2* dengan jumlah 27 iterasi, sedangkan pengujian ke-5 memperoleh jumlah iterasi terendah dengan 5 iterasi. Sedangkan dengan menggunakan metode *K-Means* yang diusulkan, iterasi yang dibutuhkan untuk proses *clustering dataset 2* berjumlah 9 iterasi.

* + 1. **Hasil *Clustering* Pada *Dataset 2***

Hasil *clustering* pada *dataset 2* dengan jumlah *cluster* 2 dapat dilihat pada grafik yang ada pada gambar 4.3.



Jumlah Data

250

200

150

100

KM Cluster 1

KM Cluster 2

KMP Cluster 1

KMP Cluster 2

50

0

1 2 3 4 5 6 7 8 9

10

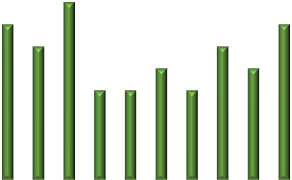
Pengujian

**Gambar 4.3. Grafik hasil *clustering* pada *dataset 2***

Berdasarkan grafik pada gambar 4.3, hasil *clustering K-Means* dengan penentuan titik pusat *cluster* secara acak (konvensional) dengan jumlah data yang sama dengan jumlah data yang dihasilkan menggunakan *K-Means* yang diusulkan pada penelitian ini ditunjukkan pada pengujian ke 1, 2, 5, 6, dan 10 dengan jumlah data di *cluster* 1 sebanyak 192 data dan pada *cluster* 2 sebanyak 203 data. Sedangkan pada pengujian ke 3, 4, 7, 8 dan 9, hasil yang diperoleh dari proses *clustering K-Means* konvensional menghasilkan jumlah data yang berbeda dari proses *K-Means* yang diusulkan.

* + 1. **Jumlah Iterasi Pada *Dataset 3***

Pengujian *clustering* pada *dataset Death and Birth Rates* dilakukan sebanyak 10 kali pengujian. Jumlah iterasi yang dibutuhkan untuk 10 kali pengujian *clustering* yang dilakukan pada *dataset Death and Birth Rates* menggunakan metode *K-Means* konvensional memiliki jumlah yang berbeda pada setiap pengujiannya. Grafik hasil *clustering* pada *dataset 3* dapat dilihat pada gambar 4.4.



Jumlah Iterasi

9

8

7

6

5

4

3

2

1

0

K-Means

K-Means Purity

Pengujian

1 2 3 4 5 6 7 8 9 10

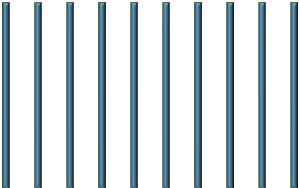
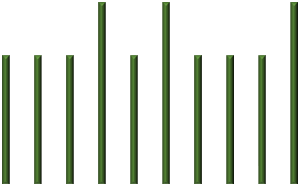
**Gambar 4.4. Grafik hasil *clustering* pada *dataset 3***

Berdasarkan gambar 4.4, jumlah iterasi yang dibutuhkan untuk 10 kali pengujian *clustering* yang dilakukan pada *dataset Death and Birth Rates* menggunakan metode *K-Means* konvensional memiliki jumlah yang berbeda pada setiap pengujiannya. Pengujian ke-3 pada *K-Means* konvensional memiliki jumlah iterasi tertinggi untuk menyelesaikan proses *clustering* pada *dataset 3* dengan jumlah 8 iterasi, sedangkan pengujian ke-4 dan ke-5 memperoleh jumlah iterasi terendah dengan 4 iterasi. Sedangkan dengan menggunakan metode *K-Means* yang diusulkan, iterasi yang dibutuhkan untuk proses *clustering dataset 3* berjumlah 2 iterasi.

* + 1. **Hasil *Clustering* Pada *Dataset 3***

Metode *K-Means* konvesional dan *K-Means* yang diusulkan untuk menentukan titik pusat *cluster* awal digunakan pada proses *clustering dataset 3*. Hasil *clustering* pada

*dataset 3* dengan jumlah *cluster* 2 dapat dilihat pada grafik yang ada pada gambar 4.5 berikut.



Jumlah Data

45

40

35

30

25

20

15

10

5

0

KM Cluster 1

KM Cluster 2

KMP Cluster 1

KMP Cluster 2

1 2 3 4 5 6 7 8 9 10

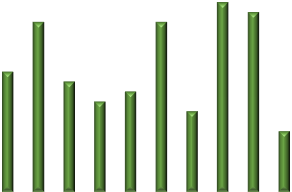
Pengujian

**Gambar 4.5. Grafik hasil *clustering* pada *dataset 3***

Berdasarkan grafik pada gambar 4.5, hasil *clustering K-Means* dengan penentuan titik pusat *cluster* secara acak (konvensional) dengan jumlah data yang sama dengan jumlah data yang dihasilkan menggunakan *K-Means* yang diusulkan pada penelitian ini tidak ada. Seluruh pengujian yang dilakukan dengan *K-Means* konvensional memiliki hasil yang berbeda dengan hasil *clustering* dari metode *K-Means* yang diusulkan. Hasil yang diperoleh menggunakan *K-Means* yang diusulkan adalah 28 data pada *cluster* 1 dan 42 data pada *cluster* 2.

* + 1. **Jumlah Iterasi Pada *Dataset 4***

Pengujian *clustering* pada *dataset Wholesale Customers* yang berjumlah 440 data dengan 6 attribut dilakukan sebanyak 10 kali pengujian. Jumlah iterasi yang dibutuhkan untuk 10 kali pengujian *clustering* yang dilakukan pada *dataset Wholesale Customers* menggunakan metode *K-Means* konvensional memiliki jumlah yang berbeda pada setiap pengujiannya. Grafik hasil *clustering* pada *dataset 4* dapat dilihat pada gambar 4.6.



Jumlah Iterasi

20

18

16

14

12

10

8

6

4

2

0

K-Means

K-Means Purity

1 2 3 4 5 6 7 8 9 10

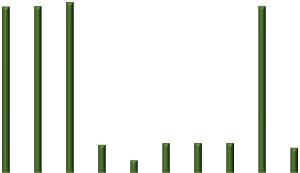
Pengujian

**Gambar 4.6. Grafik hasil *clustering* pada *dataset 4***

Berdasarkan gambar 4.6, Pengujian ke-8 pada *K-Means* konvensional memiliki jumlah iterasi tertinggi untuk menyelesaikan proses *clustering* pada *dataset 4* dengan jumlah 19 iterasi, sedangkan pengujian ke-10 memperoleh jumlah iterasi terendah dengan 6 iterasi. Sedangkan dengan menggunakan metode *K-Means* yang diusulkan, iterasi yang dibutuhkan untuk proses *clustering dataset 4* berjumlah 6 iterasi.

* + 1. **Hasil *Clustering* Pada *Dataset 4***

Metode *K-Means* konvesional dan *K-Means* yang diusulkan untuk menentukan titik pusat *cluster* awal digunakan pada proses *clustering dataset 4*. Hasil *clustering* pada *dataset 4* dengan jumlah *cluster* 2 dapat dilihat pada grafik yang ada pada gambar 4.7 sebagai berikut.



Jumlah Data

450

400

350

300

250

200

150

100

50

0

KM Cluster 1

KM Cluster 2

KMP Cluster 1

KMP Cluster 2

Pengujian

1 2 3 4 5 6 7 8 9 10

**Gambar 4.7. Grafik hasil *clustering* pada *dataset 4***

Berdasarkan grafik pada gambar 4.7, hasil *clustering K-Means* dengan penentuan titik pusat *cluster* secara acak (konvensional) dengan jumlah data yang sama dengan jumlah data yang dihasilkan menggunakan *K-Means* yang diusulkan pada penelitian ini ditunjukkan pada pengujian ke 1 dengan jumlah data di *cluster* 1 sebanyak 375 data dan pada *cluster* 2 sebanyak 65 data. Sedangkan pada pengujian lainnya, hasil yang diperoleh dari proses *clustering K-Means* konvensional menghasilkan jumlah data yang berbeda dari proses *K-Means* yang diusulkan.

* 1. **Hasil Evaluasi *Clustering***

Tahap selanjutnya setelah proses *clustering* adalah menghitung nilai DBI dari setiap masing-masing pengujian hasil *clustering*. Lalu dihitung nilai rata-rata DBI dari seluruh percobaan dan pada setiap jumlah *cluster*. Pada algoritma *K-Means* konvensional, pengujian dilakukan sebanyak 10 kali untuk mengetahui perbandingan dari kedua metode. Perbandingan dilakukan terhadap ketiga *dataset* dengan jumlah *cluster* 2. Titik pusat *cluster* awal dari algoritma *K-Means* konvensional dipilih secara acak. Nilai acak yang digunakan berada dalam rentang 1 sampai sebanyak jumlah data yang ada, berdasarkan jumlah dari data pada *dataset* itu sendiri. Perbandingan evaluasi *clustering* dilakukan terhadap hasil *clustering* kedua metode berdasarkan nilai *Davies-Bouldin Index* dari kedua metode tersebut. Hasil perbandingan nilai

*Davies-Bouldin Index* dari kedua metode pada *dataset 2* dapat dilihat pada tabel 4.30 dibawah ini.

Tabel 4.30 Perbandingan Nilai *Davies Bouldin Index* pada *dataset 2*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Uji ke-** | **Nilai *Davies-Bouldin Index*** | | | | | | | |
| **K-Means konvensional** | | | | **Purity K-Means** | | | |
| **Iterasi**  **ke-1** | **Iterasi**  **ke-5** | **Iterasi**  **ke-10** | 𝒙̅ | **Iterasi**  **ke-1** | **Iterasi**  **ke-5** | **Iterasi**  **ke-10** | 𝒙̅ |
| **1** | 2.1546 | 2.2319 | 2.1017 | **2.0787** | 1.4369 | 2.1139 | - | **1.9183** |
| **2** | 3.3801 | 2.0649 | 1.9356 | **2.0327** | 1.4369 | 2.1139 | - | **1.9183** |
| **3** | 3.2546 | 2.1581 | 2.0892 | **2.0832** | 1.4369 | 2.1139 | - | **1.9183** |
| **4** | 2.0592 | 2.2619 | 2.1394 | **2.1492** | 1.4369 | 2.1139 | - | **1.9183** |
| **5** | 1.6532 | 1.8517 | - | **1.8521** | 1.4369 | 2.1139 | - | **1.9183** |
| **6** | 1.1898 | 1.9952 | 1.8525 | **1.8809** | 1.4369 | 2.1139 | - | **1.9183** |
| **7** | 2.2336 | 2.1412 | 1.9740 | **1.9678** | 1.4369 | 2.1139 | - | **1.9183** |
| **8** | 1.4812 | 1.8704 | 1.8517 | **1.8406** | 1.4369 | 2.1139 | - | **1.9183** |
| **9** | 2.5717 | 2.0503 | 1.9206 | **2.0073** | 1.4369 | 2.1139 | - | **1.9183** |
| **10** | 1.1218 | 1.8634 | - | **1.8012** | 1.4369 | 2.1139 | - | **1.9183** |
|  |  |  |  |  |  |  |  |  |
| **Rata- Rata** | 2.1099 | 2.0489 | 1.9830 | **1.9693** | 1.4369 | 2.1139 | - | **1.9183** |

Berdasarkan tabel 4.30 diatas, diketahui bahwa nilai *Davies Bouldin Index* pada *K- Means* konvensional memperoleh nilai rata-rata 2.1099 untuk iterasi pertama, pada iterasi ke-5, nilai DBI yang diperoleh sebesar 2.0489 dan untuk iterasi ke-10, nilai DBI yang diperoleh adalah sebesar 1.9830. Secara keseluruhan, hasil dari 10 kali pengujian *K-Means* konvensional memperoleh nilai DBI rata-rata sebesar 1.9693. Sedangkan pada metode *K-Means* yang diusulkan, nilai rata-rata DBI yang diperoleh adalah sebesar 1.9183. Untuk hasil perbandingan nilai *Davies-Bouldin Index* dari kedua metode pada *dataset 3* dapat dilihat pada tabel 4.31.

Tabel 4.31 Perbandingan Nilai *Davies Bouldin Index* pada *dataset 3*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Uji ke-** | **Nilai *Davies-Bouldin Index*** | | | | | | | |
| **K-Means konvensional** | | | | **Purity K-Means** | | | |
| **Iterasi ke-1** | **Iterasi ke-2** | **Iterasi ke-5** | 𝒙̅ | **Iterasi ke-1** | **Iterasi ke-2** | **Iterasi ke-5** | 𝒙̅ |
| **1** | 8.0097 | 0.6782 | 0.5270 | **1.6231** | 0.6887 | 0.5600 | - | **0.6244** |
| **2** | 1.2672 | 0.6313 | 0.5251 | **0.6726** | 0.6887 | 0.5600 | - | **0.6244** |
| **3** | 0.5031 | 0.5656 | 0.5486 | **0.5690** | 0.6887 | 0.5600 | - | **0.6244** |

Tabel 4.32 Perbandingan Nilai *Davies Bouldin Index* pada *dataset 3* (lanjutan)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **4** | 1.8027 | 0.7855 | - | **0.9174** | 0.6887 | 0.5600 | - | **0.6244** |
| **5** | 39.218 | 1.1126 | - | **10.366** | 0.6887 | 0.5600 | - | **0.6244** |
| **6** | 0.5997 | 0.5486 | 0.5259 | **0.5453** | 0.6887 | 0.5600 | - | **0.6244** |
| **7** | 4.6548 | 1.0683 | - | **1.7038** | 0.6887 | 0.5600 | - | **0.6244** |
| **8** | 0.5945 | 0.5892 | 0.5251 | **0.5517** | 0.6887 | 0.5600 | - | **0.6244** |
| **9** | 0.6911 | 0.5362 | 0.5259 | **0.5610** | 0.6887 | 0.5600 | - | **0.6244** |
| **10** | 0.5960 | 0.7123 | 0.5270 | **0.5721** | 0.6887 | 0.5600 | - | **0.6244** |
|  |  |  |  |  |  |  |  |  |
| **Rata-**  **Rata** | 5.7936 | 0.7227 | 0.5292 | **1.8072** | 0.6887 | 0.5600 | - | **0.6244** |

Berdasarkan tabel 4.31 dan 4.32 diatas, diketahui bahwa nilai *Davies Bouldin Index* pada *K-Means* konvensional memperoleh nilai rata-rata sebesar 5.7936 untuk iterasi pertama, pada iterasi ke-2, nilai rata-rata DBI yang diperoleh sebesar 0.7227 dan untuk iterasi ke-5, nilai rata-rata DBI yang diperoleh adalah sebesar 0.5292. Secara keseluruhan, hasil dari 10 kali pengujian *K-Means* konvensional memperoleh nilai DBI rata-rata sebesar 1.8072. Sedangkan pada metode *K-Means* yang diusulkan, nilai rata-rata DBI yang diperoleh adalah sebesar 0.6244. Untuk hasil perbandingan nilai *Davies-Bouldin Index* dari kedua metode pada *dataset 4* dapat dilihat pada tabel 4.33.

Tabel 4.33 Perbandingan Nilai *Davies Bouldin Index* pada *dataset 4*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Uji ke-** | **Nilai *Davies-Bouldin Index*** | | | | | | | |
| **K-Means konvensional** | | | | **Purity K-Means** | | | |
| **Iterasi ke-1** | **Iterasi ke-5** | **Iterasi ke-10** | 𝒙̅ | **Iterasi ke-1** | **Iterasi ke-5** | **Iterasi ke-10** | 𝒙̅ |
| **1** | 8.2561 | 1.1380 | 1.1424 | **1.8910** | 0.8317 | 1.1153 | - | **1.0357** |
| **2** | 1.8066 | 1.2568 | 1.1594 | **1.2303** | 0.8317 | 1.1153 | - | **1.0357** |
| **3** | 4.3494 | 1.1429 | 1.1686 | **1.4798** | 0.8317 | 1.1153 | - | **1.0357** |
| **4** | 1.4450 | 1.2907 | - | **1.3161** | 0.8317 | 1.1153 | - | **1.0357** |
| **5** | 0.9588 | 0.9762 | 1.0515 | **0.9701** | 0.8317 | 1.1153 | - | **1.0357** |
| **6** | 4.6420 | 1.2142 | 1.1352 | **1.3845** | 0.8317 | 1.1153 | - | **1.0357** |
| **7** | 0.8346 | 1.1482 | - | **1.0973** | 0.8317 | 1.1153 | - | **1.0357** |
| **8** | 3.8521 | 1.2387 | 1.1697 | **1.3374** | 0.8317 | 1.1153 | - | **1.0357** |
| **9** | 4.1809 | 1.2800 | 1.1721 | **1.3833** | 0.8317 | 1.1153 | - | **1.0357** |
| **10** | 1.4546 | 1.1299 | - | **1.1832** | 0.8317 | 1.1153 | - | **1.0357** |
|  |  |  |  |  |  |  |  |  |
| **Rata-**  **Rata** | 3.1780 | 1.1815 | 1.1427 | **1.3273** | 0.8317 | 1.1153 | - | **1.0357** |

Berdasarkan tabel 4.33 diatas, diketahui bahwa nilai *Davies Bouldin Index* pada *K- Means* konvensional memperoleh nilai rata-rata sebesar 3.1780 untuk iterasi pertama, pada iterasi ke-5, nilai rata-rata DBI yang diperoleh sebesar 1.1815 dan untuk iterasi ke-10, nilai rata-rata DBI yang diperoleh adalah sebesar 1.1427. Secara keseluruhan, hasil dari 10 kali pengujian *K-Means* konvensional memperoleh nilai DBI rata-rata sebesar 1.3273. Sedangkan pada metode *K-Means* yang diusulkan, nilai rata-rata DBI yang diperoleh adalah sebesar 1.0357.

* 1. **Perbandingan Hasil Evaluasi *Clustering***

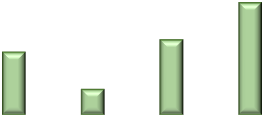
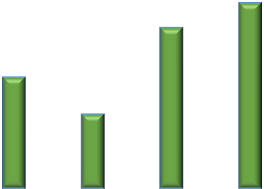
Untuk mengetahui efektifitas dari metode yang diusulkan yaitu penentuan titik pusat *cluster* berdasarkan *Purity minimum-maximum* dengan metode penentuan titik pusat *cluster* awal pada algoritma K-Means secara konvensional terhadap proses *clustering*, maka dilakukan perbandingan total rata-rata hasil evaluasi *clustering* dari keempat *dataset*. Perbandingan rata-rata hasil evaluasi *clustering* dari kedua metode terhadap keempat *dataset* yang digunakan dapat dilihat pada tabel 4.34 dan 4.35 dibawah ini.

Tabel 4.34 Hasil evaluasi *clustering* berdasarkan jumlah Iterasi

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Dataset** | **Rata-rata Jumlah Iterasi yang Dibutuhkan**  **Untuk Konvergen** | |
| *K-Means* | *K-Means with Purity*  *Min-Max* |
| **1** | *Iris* | **9** | **5** |
| **2** | *User Knowledge Modeling* | **15** | **9** |
| **3** | *Birth and Death Rates* | **6** | **2** |
| **4** | *Wholesale Customers* | **13** | **6** |

Pada tabel 4.34 diatas menunjukkan jumlah rata-rata iterasi yang dibutuhkan untuk mencapai konvergen dalam proses *clustering.* Pada *dataset Iris,* jumlah rata-rata iterasi yang dibutuhkan untuk *clustering* algoritma *K-Means* konvensional adalah sebanyak 9 iterasi, sedangkan dengan menggunakan metode *K-means* yang diusulkan, jumlah iterasi yang dibutuhkan untuk mencapai konvergen adalah 5 iterasi. Pada *dataset User Knowledge Modeling,* jumlah rata-rata iterasi yang dibutuhkan untuk *clustering* algoritma *K-Means* konvensional adalah sebanyak 15 iterasi, sedangkan dengan menggunakan metode *K-means* yang diusulkan, jumlah iterasi yang

dibutuhkan untuk mencapai konvergen adalah 9 iterasi. Hasil perbandingan selanjutnya pada *dataset Birth and Death Rates,* jumlah rata-rata iterasi yang dibutuhkan untuk *clustering* algoritma *K-Means* konvensional adalah sebanyak 6 iterasi, sedangkan dengan menggunakan metode *K-Means* yang diusulkan, jumlah iterasi yang dibutuhkan adalah 2 iterasi. Kemudian pada *dataset Wholesale Customers,* jumlah rata-rata iterasi yang dibutuhkan untuk *clustering* algoritma *K- Means* konvensional adalah sebanyak 13 iterasi, sedangkan dengan menggunakan metode *K-Means* yang diusulkan, jumlah iterasi yang dibutuhkan adalah 6 iterasi. Secara keseluruhan, rata-rata jumlah iterasi yang diperoleh dari kedua metode tersebut dapat dilihat pada grafik di gambar 4.8 dibawah ini.



Jumlah Iterasi

16

14

12

10

8

K-Means

K-Means Purity

6

4

2

0 Dataset

Iris BirthDeath Wholesale Modeling

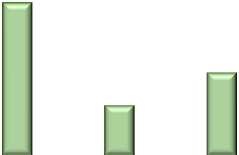
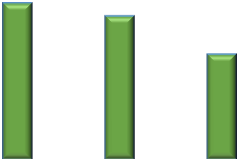
**Gambar 4.8. Grafik perbandingan rata-rata jumlah iterasi pada keempat *dataset***

Perbandingan rata-rata hasil evaluasi *clustering* berdasarkan index *Davies Bouldin Index* (DBI) dari kedua metode terhadap ketiga *dataset* yang digunakan dapat dilihat pada tabel 4.35 berikut ini.

Tabel 4.35 Hasil evaluasi *clustering* berdasarkan *Davies Bouldin Index*

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Dataset** | **Nilai *Davies Bouldin Index* (DBI)** | |
| *K-Means* | *K-Means with Purity*  *Min-Max* |
| **1** | *User Knowledge Modeling* | **1.9693** | **1.9183** |
| **2** | *Birth and Death Rates* | **1.8072** | **0.6244** |
| **3** | *Wholesale Customers* | **1.3273** | **1.0357** |

Berdasarkan tabel 4.35 diatas, diketahui bahwa pada *dataset User Knowledge Modeling,* jumlah rata-rata nilai DBI yang diperoleh untuk proses *clustering* algoritma *K-Means* konvensional adalah sebesar 1.9693, sedangkan dengan menggunakan metode *K-means* yang diusulkan, nilai DBI yang diperoleh adalah sebesar 1.9183. Hasil perbandingan selanjutnya pada *dataset Birth and Death Rates,* jumlah rata-rata nilai DBI yang diperoleh untuk proses *clustering* algoritma *K-Means* konvensional adalah sebesar 1.8072, sedangkan dengan menggunakan metode *K-means* yang diusulkan, nilai DBI yang diperoleh adalah sebesar 0.6244. Kemudian pada *dataset Wholesale Customers,* jumlah rata-rata nilai DBI yang diperoleh untuk proses *clustering* algoritma *K-Means* konvensional adalah sebesar 1.3273, sedangkan dengan menggunakan metode *K-means* yang diusulkan, nilai DBI yang diperoleh adalah sebesar 1.0357. Secara keseluruhan, rata-rata jumlah iterasi yang diperoleh dari kedua metode tersebut dapat dilihat pada grafik di gambar 4.9 dibawah ini.



NIlai DBI

2.5

2

1.5

1

K-Means

K-Means Purity

0.5

0

Dataset

Knowledge BirthDeath Wholesale

**Gambar 4.9. Grafik perbandingan rata-rata nilai DBI pada ketiga *dataset***

## BAB V KESIMPULAN DAN SARAN

* 1. **Kesimpulan**

Berdasarkan pengujian dan evaluasi yang telah dijabarkan pada bab 4, berikut dapat diambil beberapa kesimpulan yang diuraikan dibawah ini:

1. Pada *dataset Iris,* tingkat akurasi yang dihasilkan pada proses *clustering* yang dilakukan dalam 10 kali pengujian dengan menggunakan *K-Means* secara konvensional adalah sebesar 85.5%, sedangkan dengan menggunakan *Purity K- Means,* tingkat akurasi yang diperoleh adalah sebesar 88.7%, atau 3.2% lebih baik daripada hasil *K-Means* secara konvensional.
2. Dari 10 kali pengujian yang telah dipaparkan sebelumnya pada bab 4, pengujian pada 4 *dataset* masing-masing telah diperoleh hasilnya, jumlah rata-rata iterasi yang dibutuhkan untuk pengujian *dataset Iris* untuk mencapai konvergen dengan metode *K-Means* secara konvensional adalah 9 iterasi, sedangkan dengan metode yang diusulkan, jumlah rata-rata iterasi yang dibutuhkan untuk mencapai konvergen adalah 5 iterasi. Pada *dataset User Knowledge Modeling* jumlah rata- rata iterasi yang dibutuhkan dalam proses *clustering* metode *K-Means* konvensional adalah 15 iterasi, sedangkan dengan metode yang diusulkan, jumlah rata-rata iterasi yang dibutuhkan untuk mencapai konvergen adalah 9 iterasi. Pada *dataset Death and Birth Rates,* jumlah rata-rata iterasi yang dibutuhkan dalam proses *clustering* metode *K-Means* konvensional adalah 6 iterasi, sedangkan dengan metode yang diusulkan, jumlah rata-rata iterasi yang dibutuhkan untuk mencapai konvergen adalah 2 iterasi. Pada *dataset* ketiga, yaitu *Wholesale Customers,* jumlah rata-rata iterasi yang dibutuhkan dalam proses *clustering*

65

metode *K-Means* konvensional adalah 13 iterasi, sedangkan dengan metode yang diusulkan, jumlah rata-rata iterasi yang dibutuhkan adalah 6 iterasi.

1. Hasil evaluasi *clustering* berdasarkan jumlah iterasi pada keempat *dataset* yang diuji, pada *dataset Iris* mampu mempercepat proses *clustering K-Means* untuk mencapai konvergen, dengan selisih rata-rata sebanyak 4 iterasi atau sebesar 35% lebih cepat konvergen dibandingkan dengan metode *K-Means* secara konvensional. Pada *dataset User Knowledge Modeling* mampu mempercepat proses *clustering K-Means* untuk mencapai konvergen, dengan selisih rata-rata sebanyak 6 iterasi atau sebesar 40% lebih cepat konvergen dibandingkan dengan metode *K-Means* secara konvensional. Pada *dataset Birth and Death Rates* mampu mempercepat proses *clustering K-Means* untuk mencapai konvergen, dengan selisih rata-rata sebanyak 4 iterasi atau sebesar 66% lebih cepat konvergen dibandingkan dengan metode *K-Means* secara konvensional. Pada *dataset Wholesale Customers* mampu mempercepat proses *clustering K-Means* untuk mencapai konvergen, dengan selisih rata-rata sebanyak 7 iterasi atau sebesar 53% lebih cepat konvergen dibandingkan dengan metode *K-Means* secara konvensional.
2. Nilai rata-rata *Davies-Bouldin Index* (DBI) untuk *dataset User Knowledge Modeling* dengan metode penentuan titik pusat *cluster* awal secara konvensional adalah sebesar 1.9693 sedangkan dengan metode yang diusulkan, nilai rata-rata DBInya adalah sebesar 1.9183. Pada *dataset Birth and Death Rates*, nilai rata-rata DBI untuk hasil *clustering* metode penentuan titik pusat *cluster* awal secara konvensional adalah sebesar 1.8072 sementara nilai rata-rata DBI untuk hasil *clustering* metode yang diusulkan adalah sebesar 0.6244. Pada *dataset* yang terakhir yaitu *dataset Wholesale Customers*, nilai rata-rata DBI untuk hasil *clustering* dengan metode penentuan titik pusat *cluster* awal secara konvensional adalah sebesar 1.3273 sementara nilai rata-rata DBI untuk hasil *clustering* dari metode yang diusulkan adalah sebesar 1.0357.
3. Hasil Evaluasi *clustering* dengan *Davies Bouldin Index* (DBI) pada ketiga *dataset*

tersebut yaitu pada *dataset User Knowledge Modeling* menunjukkan hasil

*clustering* yang lebih baik dibandingkan dengan metode *K-Means* secara konvensional dengan selisih perbandingan nilai *Davies-Bouldin Index* sebesar 0.0510 atau lebih baik sebesar 2.58%. Pada *dataset Birth and Death Rates* menunjukkan hasil *clustering* yang lebih baik dibandingkan dengan metode *K- Means* secara konvensional dengan selisih perbandingan nilai *Davies-Bouldin Index* sebesar 1.1828 atau lebih baik sebesar 65.44%. Pada *dataset Wholesale Customers* menunjukkan hasil *clustering* yang lebih baik dibandingkan dengan metode *K-Means* secara konvensional dengan selisih perbandingan nilai *Davies- Bouldin Index* sebesar 0.2916 atau lebih baik sebesar 21.96%.

## Saran

Untuk penelitian kedepan selanjutnya, saran yang diajukan penulis yaitu sebagai berikut:

1. Kombinasi metode penentuan titik pusat *Purity* dengan metode yang lain untuk menentukan dan memilih titik pusat *cluster* awal yang lebih baik untuk *dataset* yang sama ataupun *dataset* yang berbeda.
2. Penerapan metode evaluasi *clustering* yang lain pada *dataset* yang memiliki jumlah data lebih besar untuk mengevaluasi hasil yang lebih baik terhadap hasil *clustering*.

# DAFTAR PUSTAKA

Archana, S, Avantika, Y & Ajay, R. 2013. *K-Means with Three different Distance Metrics*. IJCA. **Vol 67,** No 10.

Bates, A & Kalita,J. 2016. *Counting Clusters in Twitter Posts. Proceedings of the 2nd International Conference on Information Technology for Competitive Strategies*, pp. 85.

Bernard J. D. S. 2018. Peningkatan Hasil Evaluasi *Cluster Davies-Bouldin Index* Dengan Penentuan Titik Pusat *cluster* awal Algoritma *K-Means*. Tesis. Universitas Sumatera Utara.

Dibya, J.B & Anil, K.G. 2014. *Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab*. IJCBIT. **Vol 5,** No 2.

Hartono, Ongko, E. & Abdullah, D. 2015. *Determining a cluster centroid of k- means clustering using genetic algorithm.* IJCSSE. **4(6)**: 160-164.

Nurul,R.W, Defiyanti, S & Jajuli, M. 2015. *Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa.* JITTER **Vol 1,** No 2.

Oyelade, O. J, Oladipupo, O.O & Obagbuwa, I.C. 2010. *Application of K-Means Clustering Algorithm for Prediction of Students’s Academic Performance*. IJCSIS. **Vol 7,** No 1.

Rishikesh, & Shubha, P. 2016. *A Novel Approach for Data Clustering using Improved K-Means Algorithm*. IJCA. **Vol 142**, No 12, May 2016.

Sirait, K, Tulus, & Erna, B.N. 2016. *K-Means Algorithm Performance Analysis With Determining The Value Of Starting Centroid With Random And KD- Tree Method*. IconICT. **Series 930**.

Silwattananusarn, T & Tuamsuk, K. 2012. *Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012*. IJDKP. **Vol 2,** No 5.

Seung, S.C, Sung, H.C & Charles, C.T. 2010. *A Survey of Binary Similarity and Distance Measures*. SCI. **Vol 8,** No 1.

Unnati, R.R. & Chaita, J. 2015. *Implementing & Improvisation of K-means Clustering*. IJCSMC. **Vol 4**, 72-76, Issue 11.

Unnati, R.R. & Chaita, J. 2016. *Implementing & Improvisation of K-means Clustering Algorithm*. IJCSMC. **Vol 5**, 191-203, Issue 5.

Widiarina & Romi, S.W. 2015. *Algoritma Cluster Dinamik untuk Optimasi Cluster pada Algoritma K-Means dalam Pemetaan Nasabah Potensial*. JIS. **Vol 1,** No 1.

Zhan, Qiang. 2017*. An Improved K-Means Algorithm Based on Structure Features.* Journal of Software. **Vol 12,** No 1.