Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

Perbandingan Perhitungan Jarak *Euclidean Distance*, *Manhattan Distance*, dan *Cosine Similarity* dalam Pengelompokan Data Bibit Padi Menggunakan Algoritma K-Means

Comparison of Euclidean Distance, Manhattan Distance, and Cosine Similarity Calculations on Rice Seed Data Grouping Using the K-Means Algorithm

Mohamad Sugeng Pangestu, Maulida Ayu Fitriani*

Teknik Informatika, Universitas Muhammadiyah Purwokerto, Indonesia
*corr_author: maulidaayuf@gmail.com

ABSTRAK

Bibit padi yang mempunyai kualitas unggul memiliki peran penting dalam peningkatan produktivitas pada sektor pertanian. Banyaknya bibit padi yang dikembangkan oleh Balai Besar Penelitian Tanaman Padi menghasilkan karakteristik bibit padi baru serta mempunyai kemiripan karakteristik yang hampir sama. Bibit padi yang memiliki kemiripan berdasarkan karakteristiknya dapat dikelompokkan dengan menggunakan metode *Clustering* dimana dalam proses perhitungannya menggunakan metode pengukuran jarak. Pada penelitian ini menggunakan algoritma K-Means dengan metode *Euclidean Distance*, *Manhattan Distance*, dan *Cosine Similarity* sebagai pengukuran jarak pada proses pengelompokkan data bibit padi varietas unggul dengan 119 data, dan menggunakan *Davies Bouldin Index* sebagai teknik evaluasinya. Hasil penelitian yang telah dilakukan menghasilkan nilai *Davies Bouldin Index* sebesar 0.307 pada metode *Euclidean Distance* dan metode *Cosine Similarity*, sedangkan metode Manhattan Distance mendapat nilai *Davies Bouldin Indek* sebesar 0.318, Hasil tersebut dapat disimpulkan bahwa metode *Euclidean Distance* dan metode *Cosine Similarity*, sama-sama merupakan metode perhitungan jarak yang baik digunakan dalam melakukan pengelompokkan data bibit varietas unggul padi berdasarkan karakteristik benih padi karena menghasilkan nilai *Davies Bouldin Indek* yang kecil.

Kata Kunci: *Clustering,* K-Means, Perhitungan Jarak, *Davies Bouldin Index*

ABSTRACT

Rice seeds that have superior quality have an important role in increasing productivity in the agricultural sector. The number of rice seeds developed by the Indonesian Center for Rice Research has resulted in the characteristics of new rice seeds and have almost the same characteristics. Rice seeds that have similarities based on their characteristics can be grouped using the Clustering method which in the calculation process uses the distance measurement method. In this study using the K-Means algorithm with Euclidean Distance, Manhattan Distance, and Cosine Similarity methods as a distance measurement in the process of grouping high-yielding rice seed data with 119 data, and using the Davies Bouldin Index as an evaluation technique. The results of the research that have been carried out have resulted in a Davies Bouldin Index value of 0.307 in the Euclidean Distance method and the Cosine Similarity method, while the Manhattan Distance method

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

has a Davies Bouldin Index value of 0.318. These results can be concluded that the Euclidean Distance method and the Cosine Similarity method are both A good distance calculation method is used in grouping the data of superior rice varieties based on the characteristics of the rice seeds because it produces a small Davies Bouldin Index value.

Keywords: Clustering, K-Means, Distance Measure, Davies Bouldin Index

PENDAHULUAN

Balai Besar Penelitian Tanaman Padi merupakan Lembaga penelitian tanaman padi penghasil teknologi serta inovasi pada pertanaman padi yang modern. Banyaknya bibit padi yang dikembangkan menghasilkan karakteristik bibit padi yang hampir sama dengan karakteristik bibit yang lain, Bibit padi yang memiliki kemiripan berdasarkan karakteristiknya dapat dikelompokkan dengan menggunakan algoritma K-Means *Clustering*. Algoritma *K-Means* juga menggunakan metode perhitungan jarak dalam menentukan tingkat kemiripan diantara data. Metode perhitungan jarak ini telah dilakukan oleh peneliti-peneliti sebelumnya dengan studi kasus dan metode yang berbeda-beda, namun hasil yang didapat berbeda-beda, seperti penelitian yang dilakukan oleh Nishom (2019) mengunakan 3 metode perhitungan jarak yaitu *Euclidean, Manhattan,* dan *Minkowski* dengan dataset yang digunakan dalam penalitian ini adalah data pokok pendidikan tingkat dasar dan menengah dikota tegal. Hasil yang didapatkan bahwa dari ketiga metode yang dibandingkan menunjukkan metode pengukuran jarak yang baik, yaitu menggunakan metode *Euclidean Distance* pada penelitian yang bertujuan untuk mengetahui status disparitas kebutuhan guru di kota Tegal.

Berbeda dengan penelitian yang telah dilakukan oleh Ismail et al. (2019) metode yang diterapkan yaitu mengenai pengenalan wajah berbasis perhitungan jarak Fitur LBP Menggunakan *Euclidean, Manhattan, Chi Square Distance* yang bertujuan untuk membangun sistem pengenalan wajah dengan membandingkan performa 3 metode tersebut yang mendapatkan hasil terbaik yaitu menggunakan metode *Manhattan Distance*. Penelitian lainnya dilakukan Azmi et al. (2020) terkait optimasi *centroid* awal algoritma K-Means menggunakan *Cosine Similarity*, dengan membandingkan K-Means dengan K-Means *Cosine Similarity* menggunakan data acak, hasil yang diperoleh berdasarkan perhitungan akurasi bahwa K-Means *Cosine Similarity* mendapat hasil yang cukup baik dibandingkan dengan K-Means.

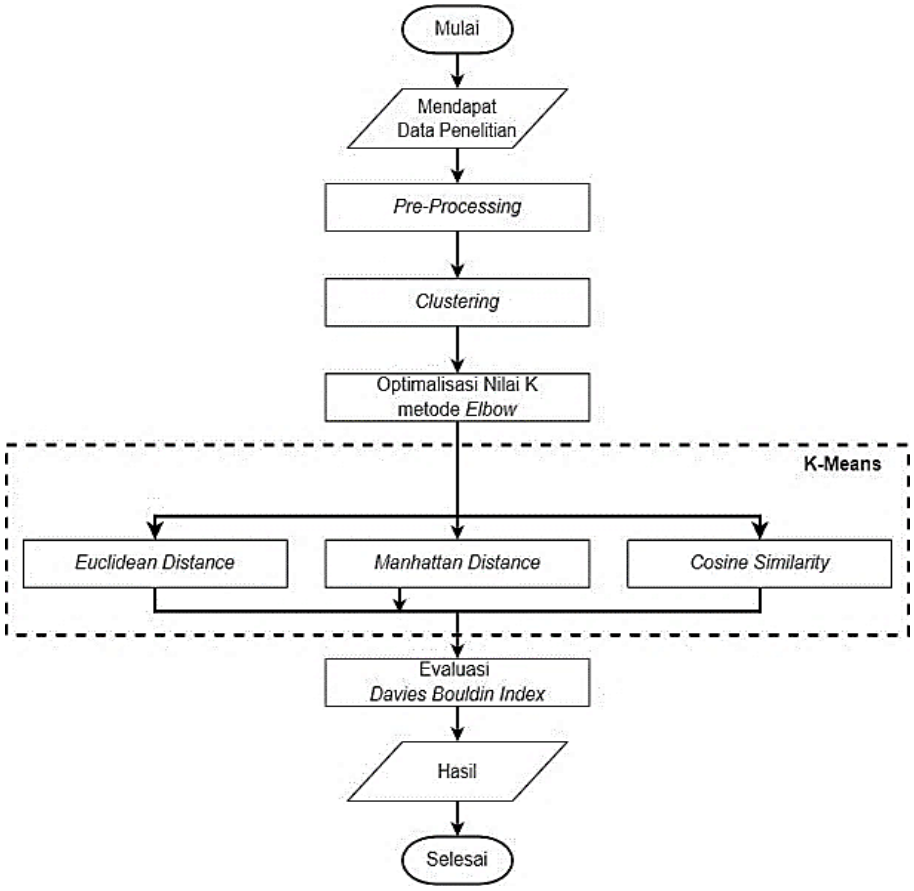
Berdasarkan permasalahan tersebut, dalam penelitian ini yaitu meimplementasikan metode perhitungan jarak pada pengelompokkan data bibit padi menggunakan teori jarak yaitu *Euclidean Distance*, *Manhattan Distance*, dan *Cosine Similarity* dalam menglompokkan data bibit padi berdasarkan karakteristiknya, dari hasil yang didapat kemudian dievaluasi menggunakan metode *Davies Bouldin Index* yang bertujuan untuk memaksimalkan jarak antar *Claster*.

METODE PENELITIAN

Penelitian ini diawali dengan melakukan kajian penelitian terdahulu yang masih berhubungan dengan apa yang akan diteliti. Penelitian ini menggunakan algoritma K-Means *Clustering* dengan 3 metode *Euclidean Distance*, *Manhattan Distance*, dan *Cosine Similarity* sebagai perhitungan jarak pada proses pengelompokkan data bibit padi dan menggunakan *Davies Bouldin Index* sebagai teknik evaluasinya. Selanjutnya pengumpulan data yang akan digunakan yaitu menggunakan data skunder yang diperoleh dari website Balai Besar Penelitian Tanaman Padi Balitbangtan Kementrian Pertanian dan e-book

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

deskripsi varietas padi (Sasmita *et al*., 2019). Data yang diperoleh yaitu sebanyak 119 dengan atribut sebanyak 16, kemudian di *Pre-Processing menggunakan metode min-max normalization*, untuk memperoleh jumlah kelompok yang optimal dilakukannya terlebih dahulu optimasi data menggunakan metode *elbow*, kemudian masuk ke tahap K-means dengan 3 metode perhitugan jarak yang digunakan dan Davies Bouldin Index sebagai teknik evaluasinya. Gambar 1 merupakan alur dalam penelitian.



Gambar 1. Diagram alur penelitian

1. *Pre-Processing*

Tahapan pre-processing data meliputi data reduction, data cleaning, dan data transformation. Proses normalisasi pada tahap data transformation menggunakan metode Min-Max Normalization dengan rumus yang digunakan yaitu pada persamaan (1) (Adeyemo et al., 2019).

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Keterangan:

x_{new} : data baru yang dinormalisasikan
x : data yang akan dinormalisasikan
x_{min} : nilai terkecil pada suatu atribut
x_{max} : nilai terbesar pada suatu atribut

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

2. Metode *Elbow*

Metode *Elbow* digunakan untuk menentukan jumlah K (*Cluster*) yang optimal dengan menghitung nilai SSE. Adapun untuk menghitung nilai SSE dapat menggunakan rumus pada persamaan (2) (Rahman et al., 2017).

$$SSE = \sum_{\mathrm{K}=1}^{\mathrm{K}} \sum_{x_i} sS_k \ \|X_i - C_k\|^2 \quad (2)$$

Keterangan:

K : jumlah *Cluster*
X_i : atribut dari data ke-i, (i= 1, 2, 3, …, …., n)
C_k : atribut titik pusat *Cluster* ke-i, (i = 1, 2, 3, …, …., n)

3. K-Means

Menurut Sadewo et al. (2018) K-Means adalah salah satu algoritma data mining yang mampu melakukan Pengelompokkan secara partisi dan memisahkan data menjadi kelompok yang berbeda. Sari et al. (2018) menyatakan Algoritma K-Means ini mampu meminimalkan jarak antara data ke *Cluster*-nya. Penelitian ini menggunakan algoritma K-Means dengan 3 metode perhitungan jarak.

Adapun tahapan algoritma K-Means sebagai berikut.

- Menentukan jumlah *Cluster* (K) yang akan di bentuk.
- Menentukan pusat *Cluster* (*Centroid*) sebanyak jumlah (K) pada dataset yang akan diteliti.
- Menghitung jarak antara data ke-pusat *Cluster*, menggunkan metode perhitungan jarak.

Euclidean Distance

Euclidean Distance digunakan untuk mengukur tingkat kemiripan jarak antara data dengan rumus *euclidean* (Nishom 2019). Rumus yang dapat digunakan dapat dilihat pada persamaan (3).

$$d(x,y) = |x - y| \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

Keterangan:

i : index dari atribut
n : jumlah data
x_i : atribut dari data ke-i, (i = 1, 2, 3, …, …., n)
y_i : atribut dari pusat *Cluster* ke-i, (i = 1, 2, 3, …, …., n)

Manhattan Distance

Manhattan Distance digunakan sebagai metode pengukuran jarak yang dihasilkan berdasarkan jumlah selisih antara dua objek data. Proses perhitungan dapat dilakukan dengan menggunakan rumus pada persamaan (4) (Nishom, 2019).

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (4)$$

Keterangan:

d : jarak antara x dan y
x_i : atribut dari data ke-i, (i = 1, 2, 3, …, …., n)
y_i : atribut pusat *Cluster* ke-i, (i = 1, 2, 3, …, …., n)

Cosine Similarity

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

Cosine Similarity digunakan untuk mengitung kemiripan antara dua objek data. Perhitungan metode *Cosine Similarity* dapat dilakukan menggunakan rumus pada persamaan (5) (Nosra, Arifianto and Rahman, 2021).

$$cos(\theta_{ij}) = \frac{\sum_{i=1}^{n}(d_i * d_j)}{\sqrt{\sum_{k=1}^{n} d_i^2 \sqrt{\sum_{k=1}^{n} d_k^2}}} \quad (5)$$

Keterangan:

$\sum_{i=1}^{n}(d_i * d_j)$: Jumlah bobot dokumen pertama dikalikan dokumen kedua
$\sum_{k=1}^{n} d_i^2$: Akar jumlah dari bobot dokumen pertama yang sudah di kuadratkan
$\sum_{k=1}^{n} d_k^2$: Akar jumlah dari bobot dokumen kedua yang sudah di kuadratkan

- Memperbarui nilai titik *Centroid* dan mengulangi langkah 3 sampai nilai dari titik *Centroid* dan *Cluster* tersebut tidak berubah atau berpindah lagi. Untuk memperbaharui *pusat Centroid* (*Cluster*) dapat dilakukan menggunakan persamaan (6) (Setiawan, 2019).

$$v = \frac{\sum_{i=1}^{n} xi}{n}; i = 1,2,3, \dots n \quad (6)$$

Keterangan:
v : *Centroid* pada *Cluster*
n : banyaknya data pada *Cluster*
x_i : data ke-i (i = 1, 2, 3, …, …., n)

4. Evaluasi

Davies Bouldin Index digunakan untuk mengevaluasi hasil dari algoritma K-Means dengan metode jarak *Euclidean Distance*, *Manhattan Distance*, dan *Cosine Similarity*. dan dilakukan dengan *software* Rapidminer Studio. Skema *Clustering* yang optimal dengan pengukuran *Davies Bouldin Index* memiliki nilai index terkecil namun nilai tidak *negative*.

HASIL DAN PEMBAHASAN

1. Data Penelitian

Data bibit padi diambil dari tahun 2002 sampai tahun 2020 dengan atribut yang diperoleh yaitu berjumlah 16 atribut, dan jumlah data pada masing – masing atribut adalah 119 data.

2. *Pre-Processing* Data

Tahapan ini dilakukan untuk mendapatkan kualitas data *output* yang baik dari data input yang akan di proses kedalam perhitungan, serta menghindari adanya kesalahan *erorr*, *missing value*, dan ketidakkonsistennya data yang dapat menyebabkan permasalahan pada saat proses penelitian. *Pre-processing* data pada penelitian ini meliputi beberapa tahap yaitu *data reduction* dan *data cleaning* dan *data transformation.*

a. Data Reduction

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

Data file yang diperoleh terdapat data atribut yang digunkan dan tidak digunakan dalam penelitian. Data atribut yang *digunakan* yaitu sejumlah 7 atribut dan data atribut yang *tidak digunakan* yaitu 9 atribut.

b. Data Cleaning

Data Cleaning diterapkan pada beberapa data bibit padi baik yang mempunyai format pengetikan yang salah (*typo*) atau data yang memiliki *sepasi* pada data tabel yang dapat menyebabkan *erorr* pada saat *system* dijalankan, adapun data *Cleaning* dilakukan secara manual menggunakan *software* Microsoft excel dengan menyeleksi semua data yang diperoleh seperti pada data atribut umur tanaman, tinggi tanaman, kadar amilosa, berat per 1000 butir, dan rata-rata hasil.

c. Data Transformation

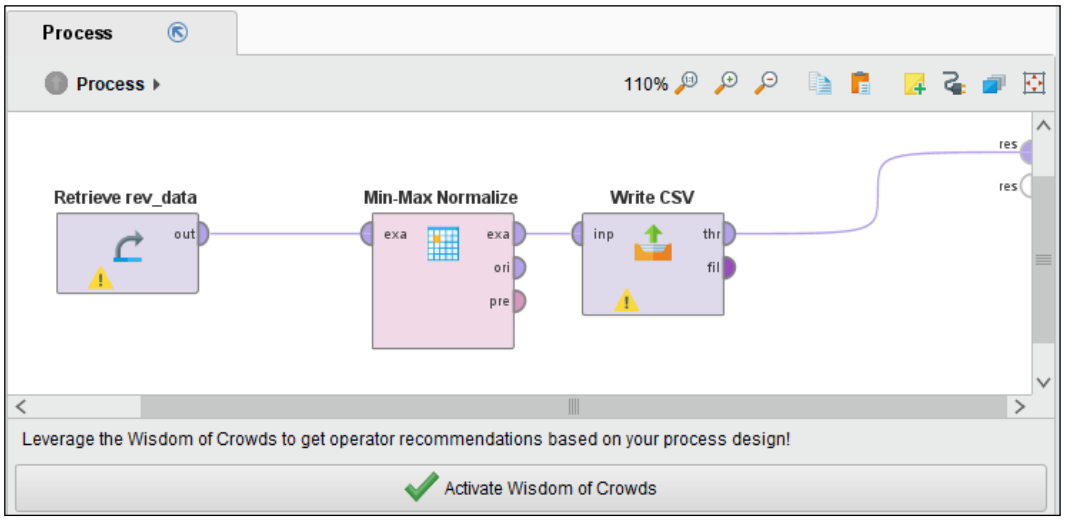
Penelitian ini *transformation data* digunakan pada atribut "kerontokan". Adapun data kriteria yang digunakan dalam pengelompokkan dapat dilihat pada Tabel 1 (Afiani, 2018).

Tabel 1. Kriteria Pengelompokkan

No	Kriteria	Satuan
1	Nama bibit	Nama bibit
2	Umur tanaman	hari
3	Tinggi tanaman	cm
4	Kerontokan	Skala: 1= Tahan 2= Sedang 3= Mudah
5	Kadar amilosa	%
6	Berat per 1000 butir	gram
7	Rata-rata hasil	(t/ha GKG) ton/hektar Gabah Kering Giling

Hasil dataset atribut yang telah dikriteriakan kemudian di normalisasikan menggunakan metode *Min-Max Normalization* untuk menghasilkan keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses (Hanifa, Adiwijaya and Al-Faraby, 2017). Adapun proses *normalisasi* data menggunakan metode *Min-Max Normalization* dilakukan dengan *software* RapidMinner Studio dan dapat dilihat pada Gambar 1.

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)



Gambar 1. Pemodelan *Min-Max normalization*

Hasil dari proses *normalisasi* dapat dilihat pada Tabel 2.

Tabel 2 Data Penelitian

Kode padi	Umur tanaman	Tinggi tanaman	Kerontokan	Kadar amilosa	Berat per 1000 butir	Rata-rata hasil
1	0.528	0.690	0.500	0.919	0.481	0.675
2	0.611	0.759	0.500	0.785	0.802	0.656
3	0.611	0.534	0.500	0.682	0.457	0.668
4	0.389	0.672	0.500	0.490	0.473	0.389
5	0.611	0.621	0.500	0.473	0.406	0.435
6	0.556	0.690	0.500	0.409	0.420	0.705
7	0.333	0.362	0.500	0.542	0.074	0.498
8	0.472	0.690	0.500	0.352	0.494	0.579
..	…..	…..	…..	…..	…..	…..
119	0.889	0.534	0.500	0.958	0.556	0.307

Data pada Tabel 2 ini merupakan hasil yang diperoleh dari tahap *pre-processing* data menggunkan 7 atribut dengan 119 data yang dinormalisasikan menggunakan *software* RapidMinner dengan metode *Min-Max Normalization* dengan *range* nilai minimal yaitu 0.0 dan nilai maksimal 1.0.

3. Metode *Elbow*

Perhitungan metode *Elbow* dilakukan dengan menggunakan *Software* RStudio dimana proses *pen-codean* (*syntak*) menggunkan bahasa R Untuk medapatkan perbandingan nilai K optimal yaitu dengan menghitung nilai (SSE) *Sum of Square Erorr* pada masing-masing nilai K yang dibentuk menggunakan persamaan (2), Semakin besar jumlah klaster (K) maka nilai SSE akan semakin kecil dan nilai SSE yang mengalami penurunan yang paling besar, maka nilai K tersebut merupakan jumlah K yang optimal. Adapun hasil *Sum of Square Erorr* menggunakan *Software* RStudio dapat dilihat pada Tabel 3.

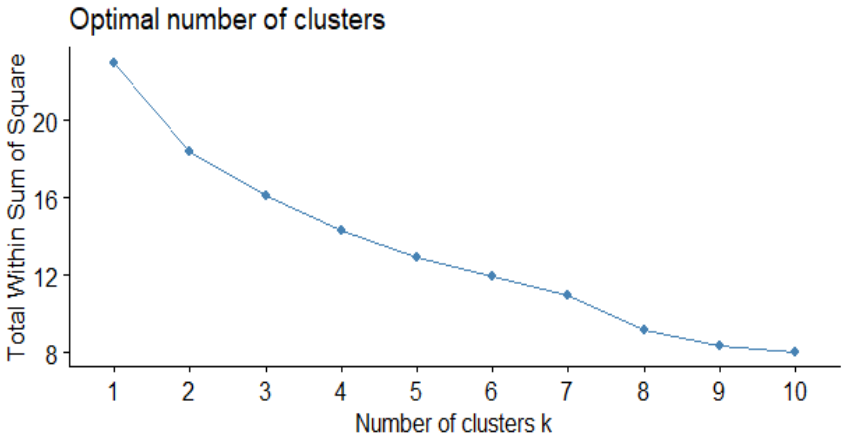
Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

Tabel 3. Hasil SSE

K	Hasil *Sum Square Error*
K1	23,003235
K2	**18,399305**
K3	16,342201
K4	14,198970
K5	12,592865
K6	10,989340
K7	10,136160
K8	9,714738
K9	8,949178
K10	7,918626

Berdasarkan Table 3 dapat disimpulkan hasil perbandingan masing-masing nilai yang dihasilakan dari perhitungan SSE dapat diketahui K1 ke K2 nilainya mengalami penurunan yang baling besar serta mempunyai selisih yang paling besar dibandingkan K lainnya, sehingga dapat diketahui bahwa K2 merupakan jumlah K yang optimal pada dataset yang digunakan pada algoritma K-Means.

Hasil Visualisasi metode *Elbow* (Gambar 2).



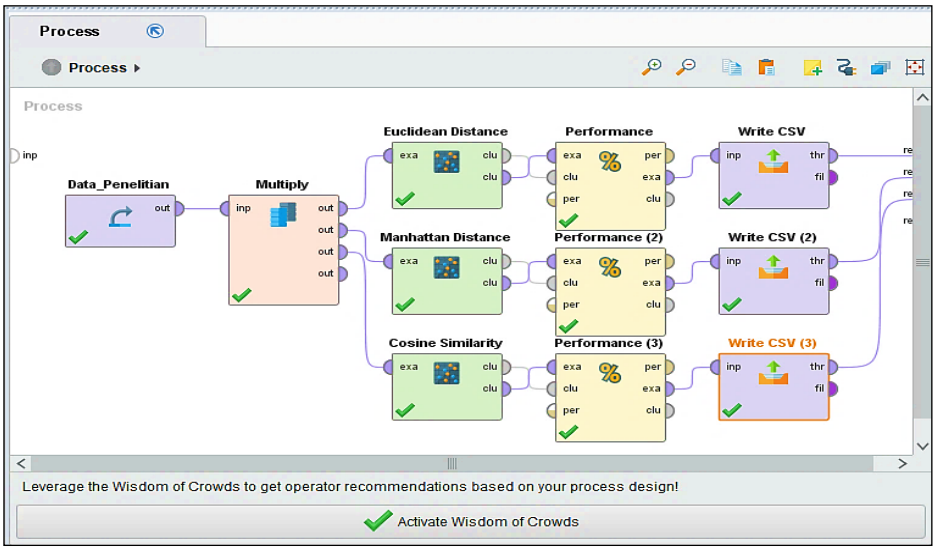
Gambar 2. Grafik hasil metode *Elbow*

Gambar 2 merupakan hasil visualisasi grafik *Elbow* dari perhitungan *Sum of Square Error* (SSE) yang sudah dihitung, dimulai dari k=1 sampai k=10 kemudian divisualisasikan menggunakan grafik untuk melihat hasil presentase nilai K yang optimal dari metode *Elbow* berupa grafik yang mengalami penurunan yang besar dan membentuk *Elbow* diantara dua titik.

4. K-Means

Berikut merupakan pemodelan proses perhitungan algoritma K-Means menggunakan *software* RapidMinner yang dapat dilihat pada Gambar 3.

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

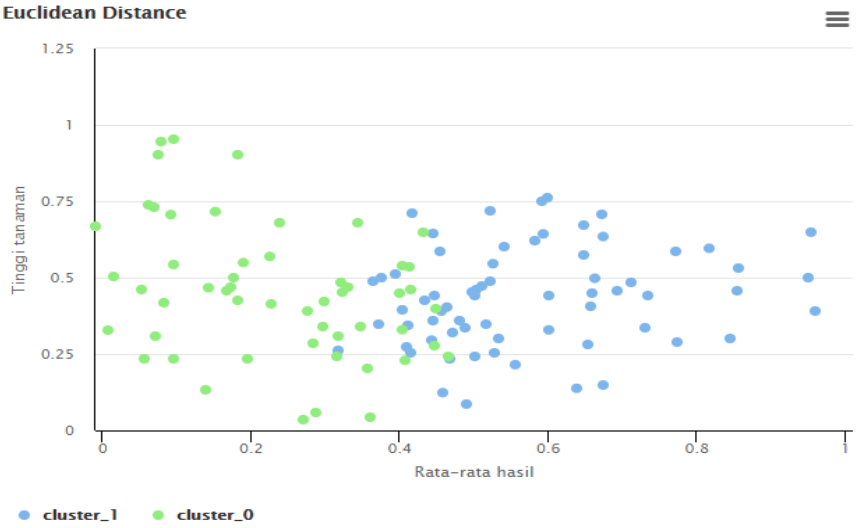


Gambar 3. Grafik hasil metode *Elbow*

Berdasarkan dari pemodelan yang ditunjukan pada Gambar 3 tahap proses algoritma K-Means menggunakan inisialisasi (K) sebanyak 2 berdasarakan metode *Elbow* yang ditentukan sebelumnya, kemudian pada proses menentukan pusat *Centroid* dilakukan oleh *system* secara acak dengan menggunakan 3 metode perhitungan jarak yaitu *Euclidean Distance*, *Manhattan Distance*, dan *Cosine Similarity* diperoleh hasil dengan *Cluster* yang terbentuk yaitu 2 *Cluster* di masing-masing metode perhitungan jarak yang dipakai. Penggunaan *software* rapidminer dalam proses K-Means *Clustering* memperoleh informasi grafik persebaran atau kedekatan jarak antar data hasil *Clustering* yang ditunjukkan pada Gambar 4, 5, 6 serta dua *output* yaitu berupa titik *Centroid* akhir dan hasil *Cluster* yang terbentuk yaitu 2 *Cluster* di masing-masing metode perhitungan jarak yang dipakai menggunakan algoritma K-Means yang ditunjukkan pada Tabel 5, 7, 9.

a. Hasil grafik perhitungan jarak Euclidean Distance

Perhitungan jarak *Euclidean Distance* menggunakan RapidMiner menghasilkan visualisasi grafik yang dapat dilihat pada Gambar 4.



Gambar 4. Grafik hasil *clustering* metode *Euclidean Distance*

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

Berdasarkan Gambar 4 dapat dianalisis secara visualisasi pesebaran data pada masing-masing *Cluster* yang dibedakan berdasarkan *warna*, pada metode *Euclidean Distance*, *Cluster* 1 di gambarkan dengan warna hijau, dan *Cluster* 2 digambarkan dengan warna biru, grafik ini juga dapat dianalisis kedekatan data satu dengan data yang lain yang terhimpun pada suatu *Cluster*.

b. Hasil pusat Cluster akhir metode Euclidean Distance

Perhitungan jarak pada algoritma K-Means juga menghasilkan pusat *Cluster* akhir, dimana proses perulangan *iterasi* berhenti ketika tidak ada data pada suatu *Cluster* yang berpindah. Pusat *Cluster* akhir metode *Euclidean Distance* dapat dilihat pada Tabel 4.

Tabel 4. Hasil *pusat Cluster* akhir *Euclidean Distance*

Atibut	C_1	C_2
Umur Tanaman	0.524	0.355
Tinggi Tanaman	0.454	0.435
Kerontokan	0.500	0.545
Kadar Amilosa	0.729	0.653
Berat per_1000 butir	0.376	0.349
Rata_rata hasil	0.231	0.576

Tabel 4 menunjukkan hasil pusat *Cluster* akhir yang terbentuk pada metode *Euclidean Distance*. Nilai pusat *Cluster* akhir ini merupakan nilai rata-rata dari data yang terhimpun pada suatu *Cluster*, nilai rata-rata ini dapat dianalisis berdasarkan karakteristik yang dihasilkan, bahwa C2 menunjukkan produktifitas padi yang cukup banyak dibandingkan dengan C1.

c. Hasil jumlah data di setiap Cluster

Hasil *Clustering* menggunakan algoritma K-Means dengan metode perhitungan jarak *Euclidean Distance* menghasilkan 2 *Cluster* dan data yang terkelompok pada masing-masing *Cluster*, hasil dapat dilihat pada Tabel 5.

Tabel 5. Hasil jumlah data disetiap *Cluster*

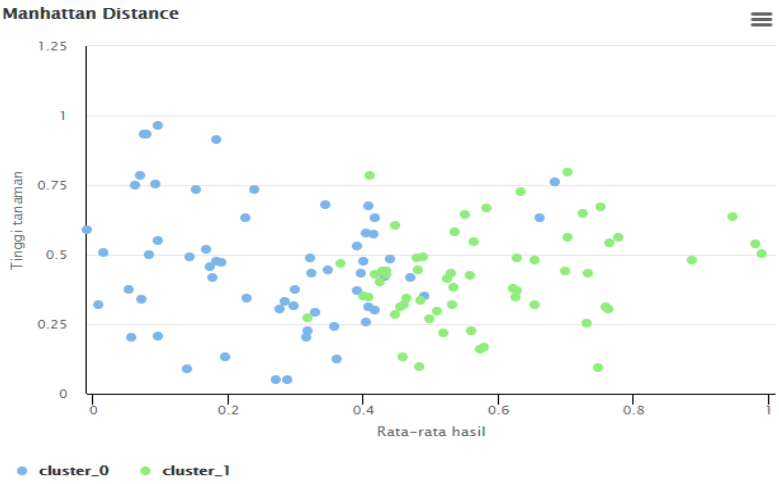
Klaster	Jumlah data padi
C_1	53
C_2	66

Tabel 5 menunjukkan hasil jumlah data bibit padi yang terhimpun pada suatu *Cluster* ini memiliki jarak kedekatan data satu dengan pusat *Cluster*, sehingga data bibit padi yang dikelompokkan memiliki tingkat kemiripan data karakteristik terhadap pusat *Cluster* yang ditentukan oleh *system* secara acak, kemudian di pisahkan bedasarkan tingkat kesamaan antar data.

d. Hasil grafik perhitungan jarak Manhattan Distance

Perhitungan jarak *Manhattan Distance* menggunakan RapidMiner menghasilkan visualisasi grafik yang dapat dilihat pada Gambar 5.

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)



Gambar 5 Grafik hasil *Clustering* metode *Manhattan Distance*

Berdasarkan Gambar 5 dapat dianalisis secara visualisasi pesebaran data pada masing-masing *Cluster* yang dibedakan berdasarkan *warna*, pada metode *Manhattan Distance* *Cluster* 1 di gambarkan dengan warna biru, dan *Cluster* 2 digambarkan dengan warna hijau, grafik ini juga dapat dianalisis kedekatan data satu dengan data yang lain yang terhimpun pada suatu *Cluster*.

e. Hasil pusat Cluster akhir metode Manhattan Distance

Perhitungan jarak pada algoritma K-Means juga menghasilkan pusat *Cluster* akhir, dimana proses perulangan *iterasi* berhenti ketika tidak ada data pada suatu *Cluster* yang berpindah. Pusat *Cluster* akhir metode *Manhattan Distance* dapat dilihat pada Tabel 6.

Tabel 6. Hasil *pusat Cluster* akhir *Manhattan Distance*

Atibut	C_1	C_2
Umur Tanaman	0.518	0.341
Tinggi Tanaman	0.465	0.421
Kerontokan	0.500	0.551
Kadar Amilosa	0.722	0.650
Berat per_1000 butir	0.382	0.340
Rata_rata hasil	0.261	0.586

Tabel 6 menunjukkan hasil pusat *Cluster* akhir yang terbentuk pada metode *Manhattan Distance*. Nilai pusat *Cluster* akhir ini merupakan nilai rata-rata dari data yang terhimpun pada suatu *Cluster*, nilai rata-rata ini dapat dianalisis berdasarkan karakteristik yang dihasilkan, bahwa C2 menunjukkan produktifitas padi yang cukup banyak dibandingkan dengan C1

f. Hasil jumlah data di setiap Cluster

Hasil *Clustering* menggunakan algoritma K-Means dengan metode perhitungan jarak *Manhattan Distance* menghasilkan 2 *Cluster* dan data yang terkelompok pada masing-masing *Cluster*, hasil dapat dilihat pada Tabel 7.

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

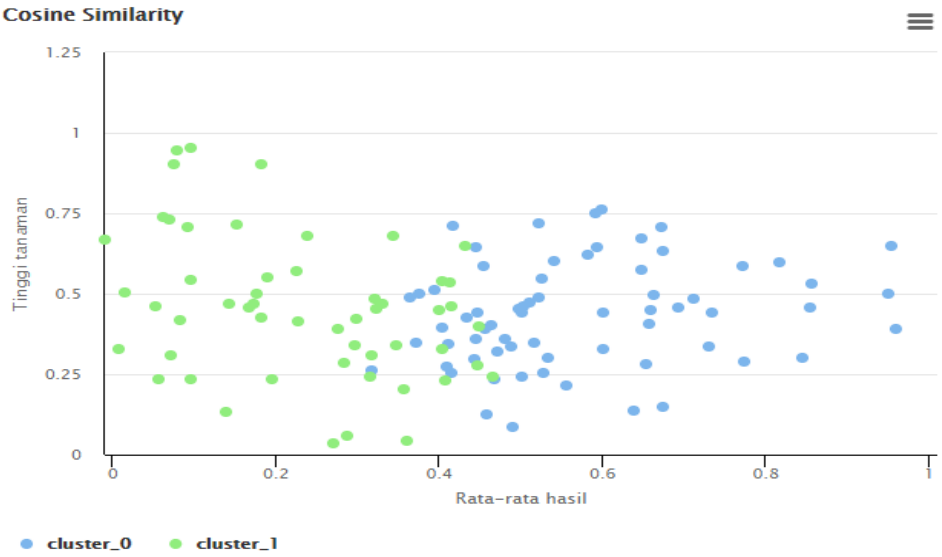
Tabel 7. Hasil jumlah data di setiap *Cluster*

Klaster	Jumlah data padi
C_1	60
C_2	59

Tabel 7 menunjukkan hasil jumlah data bibit padi yang terhimpun pada suatu *Cluster* ini memiliki jarak kedekatan data satu dengan pusat *Cluster*, sehingga data bibit padi yang dikelompokkan memiliki tingkat kemiripan data karakteristik terhadap pusat *Cluster* yang ditentukan oleh *system* secara acak, kemudian di pisahkan bedasarkan tingkat kesamaan antar data.

g. Hasil grafik perhitungan jarak Cosine Similarity

Perhitungan jarak *Cosine Similarity* menggunakan RapidMiner menghasilkan visualisasi grafik yang dapat dilihat pada Gambar 6.



Gambar 6. Grafik hasil *Clustering* metode *Cosine Similarity*

Berdasarkan Gambar 6 dapat dianalisis secara visualisasi pesebaran data pada masing-masing *Cluster* yang dibedakan berdasarkan *warna*, pada metode *Cosine Similarity* *Cluster* 1 di gambarkan dengan warna biru, dan *Cluster* 2 digambarkan dengan warna hijau, grafik ini juga dapat dianalisis juga kedekatan data satu dengan data yang lain yang terhimpun pada suatu *Cluster*

h. Hasil pusat Cluster akhir metode Cosine Similarity

Tabel 8. Hasil pusat *Cluster* akhir metode *Cosine Similarity*

Atibut	C_1	C_2
Umur Tanaman	0.355	0.524
Tinggi Tanaman	0.435	0.454
Kerontokan	0.545	0.500
Kadar Amilosa	0.653	0.729
Berat per_1000 butir	0.349	0.376
Rata_rata hasil	0.576	0.231

Sainteks
ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

Tabel 8 menunjukkan hasil pusat *Cluster* akhir yang terbentuk pada metode *Cosine Similarity*. Nilai pusat *Cluster* akhir ini merupakan nilai rata-rata dari data yang terhimpun pada suatu *Cluster*, nilai rata-rata ini dapat dianalisis berdasarkan karakteristik yang dihasilkan, bahwa C1 menunjukkan produktifitas padi yang cukup banyak dibandingkan dengan C2.

i. Hasil jumlah data disetiap Cluster

Hasil *Clustering* menggunakan algoritma K-Means dengan metode perhitungan jarak *Cosine Similarity* menghasilkan 2 *Cluster* dan data yang terkelompok pada masing-masing *Cluster*, hasil dapat dilihat pada Tabel 9.

Tabel 9. Hasil jumlah data disetiap *Cluster*

Klaster	Jumlah data padi
C_1	66
C_2	53

Tabel 9 menunjukkan hasil jumlah data bibit padi yang terhimpun pada suatu *Cluster* ini memiliki jarak kedekatan data satu dengan pusat *Cluster*, sehingga data bibit padi yang dikelompokkan memiliki tingkat kemiripan data karakteristik terhadap pusat *Cluster* yang ditentukan oleh *system* secara acak, kemudian di pisahkan bedasarkan tingkat kesamaan antar data.

5. Evaluasi

Tabel 10 merupakan penyajian hasil dari evaluasi metode perhitungan jarak menggunakan algoritma K-Means dengan data bibit padi sebagai data yang digunakan pada proses penelitian.

Tabel 10. Hasil *performance Davies Bouldin Index*

Euclidean Distance	*Manhattan Distance*	*Cosine Similarity*
0,307	0,318	0,307

Berdasarkan Tabel 10 hasil perhitungan evaluasi *Davies Bouldin Index* menggunakan 119 data bibit padi yang diproses dengan algoritma K-Means menggunakan 3 metode perhitungan jarak yaitu *Euclidean Distance*, *Manhattan Distance*, dan *Cosine Similarity* menghasilkan 3 nilai *Davies Bouldin Index* yaitu 0,307 untuk metode *Euclidean Distance*, 0.318 untuk *Manhattan Distance* dan 0,307 menggunakan metode *Cosine Similarity*. Berdasarkan ketiga hasil tersebut metode perhitungan jarak *Euclidean Distance* dan metode *Cosine Similarity* memperoleh nilai *Davies Bouldin Index* sebesar 0.307 serta memiliki hasil yang cukup baik karena mendekati nilai 0. Perhitungan pada evaluasi *Davies Bouldin Index* (DBI) dapat disimpulkan bahwa semakin kecil nilai *Davies Bouldin Index* (DBI) yang diperoleh (non negatif >= 0) maka hasil *Cluster* dari metode pengukuran jarak tersebut semakin baik.

KESIMPULAN

Berdasarkan hasil perhitungan yang telah dilakukan menggunakan Algoritma K-Means dengan 3 metode perhitungan jarak *Euclidean Distance*, *Manhattan Distance*, dan *Cosine Similarity* dan diuji menggunakan teknik evaluasi *Davies Bouldin Index* menghasilkan nilai DBI yaitu 0.307 (pada metode *Euclidean Distance* dengan *Cosine Similarity*), dan 0.318 (untuk metode *Manhattan Distance*). Hasil dari perhitungan evaluasi *Davies bouldin index* tersebut dapat disimpulkan bahwa metode *Euclidean Distance* dan *Cosine Similarity* merupakan 2 metode yang baik digunakan dalam melakukan

pengelompokkan data bibit padi berdasarkan karakteristik benih dengan hasil nilai DBI mendekati 0 yaitu 0.307.

DAFTAR PUSTAKA

Adeyemo, A., Wimmer, H. and Powell, L. (2019) "Effects of Normalization Techniques on Logistic Regression in Data Science," *Journal of Information Systems Applied Research*, 12(2). Available at: http://conisar.org.

Afiani, F.R.K. (2018) "PENERAPAN K-MEANS CLUSTERING UNTUK MENGETAHUI VARIETAS PADI UNGGUL PRODUKSI BALAI PENGKAJIAN TEKNOLOGI PERTANIAN JAWA TIMUR," *Jurnal Mahasiswa Teknik Informatika*, 2(1).

Azmi, F. *et al.* (2020) "Initial Centroid Optimization of K-Means Algorithm Using Cosine Similarity," *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, 3(2), pp. 224–231. Available at: https://doi.org/10.31289/jite.v3i2.3211.

Hanifa, T.T., Adiwijaya and Al-Faraby, S. (2017) "Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging," *e-Proceeding of Engineering* , 4, pp. 3210–3225.

Ismail, Y. *et al.* (2019) "Pengenalan Wajah Berbasis Perhitungan Jarak Fitur LBP Menggunakan Euclidean, Manhattan, Chi Square Distance," *SEMNASTIK*, pp. 386–393.

Nishom, M. (2019) "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *Jurnal Informatika: Jurnal Pengembangan IT*, 4(1), pp. 20–24. Available at: https://doi.org/10.30591/jpit.v4i1.1253.

Nosra, A., Arifianto, D. and Rahman, M. (2021) "Penerapan Metode Cosine Similarity Untuk Meningkatkan Kinerja K-Means Pada Pengelompokkan Wilayah Penanganan Covid Di Dki Jakarta," *Jurnal Smart Teknologi*, 1(1), pp. 2774–1702. Available at: http://jurnal.unmuhjember.ac.id/index.php/JST.

Rahman, A.T., Wiranto and Anggrainingsih, R. (2017) "Coal Trade Data Clusterung Using K-Means (Case Study PT. Global Bangkit Utama)," *Jurnal Ilmiah Teknologi dan Informasi*, 6(1), pp. 24–31.

Religia, Y. and sunge, A.S. (2019) "COMPARISON OF DISTANCE METHODS IN K-MEANS ALGORITHM FOR DETERMINING VILLAGE STATUS IN BEKASI DISTRICT," *ICAIIT* [Preprint].

Sadewo, M.G., Windarto, A.P. and Wanto, A. (2018) "PENERAPAN ALGORITMA CLUSTERING DALAM MENGELOMPOKKAN BANYAKNYA DESA/KELURAHAN MENURUT UPAYA ANTISIPASI/ MITIGASI BENCANA ALAM MENURUT PROVINSI DENGAN K-MEANS," *KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer)*, 2(1), pp. 311–319. Available at: http://ejurnal.stmik-budidarma.ac.id/index.php/komik.

Sari, R.W., Wanto, A. and Windarto, A.P. (2018) "IMPLEMENTASI RAPIDMINER DENGAN METODE K-MEANS (STUDY KASUS: IMUNISASI CAMPAK PADA BALITA BERDASARKAN PROVINSI)," *KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer)*, 2(1), pp. 224–230. Available at: http://ejurnal.stmik-budidarma.ac.id/index.php/komik.

ISSN: 2686-0546
Volume 19 No 2, Oktober 2022
DOI: 10.30595/sainteks.v19i2.14495 (141 – 155)

Sasmita, P. *et al*. (2019) *Varietas Unggul Baru Padi*. Available at: https://bbpadi.litbang.pertanian.go.id/ (Accessed: July 10, 2022).

Setiawan, S. (2019) "ANALISIS CLUSTER MENGGUNAKAN ALGORITMA K-MEANS UNTUK MENGETAHUI KEMAMPUAN PEGAWAI DIBIDANG IT PADA CV. ROXED LTD," *Jurnal Pelita Informatika*, 7(3), pp. 341–347.