

PENGARUH ALGORITMA PCA TERHADAP PERHITUNGAN  
NILAI RENTANG PADA *CLUSTERING*  
*SELF ORGANIZING MAP*

Diajukan Sebagai Syarat Untuk Menyelesaikan  
Pendidikan Program Strata-1 Pada  
Jurusan Teknik Informatika



Oleh :

MUHAMMAD ALBAR  
09021281520116

**JURUSAN TEKNIK INFORMATIKA  
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA  
2020**

**LEMBAR PENGESAHAN TUGAS AKHIR**

**PENGARUH ALGORITMA PCA TERHADAP PERHITUNGAN  
NILAI RENTANG PADA *CLUSTERING*  
*SELF ORGANIZING MAP***

Oleh:

MUHAMMAD ALBAR  
NIM : 09021281520116

Palembang, Desember 2020

Pembimbing I,



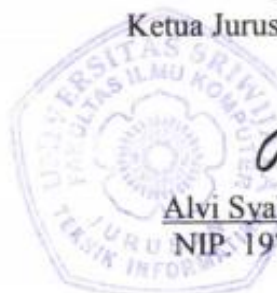
Dian Palupi Rini, M.Kom., Ph.D.  
NIP. 198803302019031007

Pembimbing II,



Kanda Januar Miraswan, M.T  
NIP. 199001092019031012

Mengetahui,  
Ketua Jurusan Teknik Informatika,



Alvi Syahrini Utami, M.Kom  
NIP. 197812222006042003

## TANDA LULUS UJIAN SIDANG TUGAS AKHIR

Pada hari Senin tanggal 31 Desember 2020 telah dilaksanakan ujian sidang tugas akhir oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Muhammad Albar  
NIM : 09021281520116  
Judul : Pengaruh Algoritma PCA Terhadap Perhitungan Nilai Rentang Pada *Clustering Self Organizing Map*

1. Pembimbing I

Dian Palupi Rini, M.Kom., Ph.D.  
NIP. 198803302019031007




2. Pembimbing II

Kanda Januar Miraswan, M.T  
NIP. 199001092019031012




3. Penguji I

Dr. Abdiansah, S.Kom., M.CS.  
NIP. 197102041997021003



4. Penguji II

Desty Rodiah, M.T  
NIP. 1671016112890005



Mengetahui,  
Ketua Jurusan Teknik Informatika



Alvi Syahrini Utami, M.Kom  
NIP. 197812222006042003

## HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Muhammad Albar  
NIM : 09021281520116  
Program Studi : Teknik Informatika  
Judul Skripsi : Pengaruh Algoritma PCA Terhadap  
Perhitungan Nilai Rentang Pada  
*Clustering Self Organizing Map*

Hasil Pengecekan Software *iThenticate/Turnitin* : 5%

Menyatakan bahwa Laporan Projek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan projek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.



Palembang, Desember 2020



Muhammad Albar  
NIM. 09021281520116

## MOTTO DAN PERSEMBAHAN

### MOTTO:

Don't burden yourself with other people's successes

*Kupersembahkan karya tulis ini kepada :*

- *Orang tuaku tercinta*
- *Keluarga besarku*
- *Sahabat dan teman seperjuanganku*
- *Fakultas Ilmu Komputer*
- *Universitas Sriwijaya*

# EFFECT OF PCA ALGORITHM ON THE CALCULATION OF DISTANCE MEASUREMENT VALUES IN CLUSTERING SELF ORGANIZING MAP

By:

Muhammad Albar  
09021281520116

## ABSTRACT

Data mining is a pattern recognition process that aims to find important information in data. Clustering is a data mining method. Clustering works by grouping data based on similarities or dissimilarities. Distance measurement is a very important component in clustering. Different distance measurement will produce different clustering values, but before doing a clustering, the data must go through initial processing or what is called preprocessing. Data preprocessing can improve the results and efficiency of a clustering algorithm. Data reduction is a method in data preprocessing that can be applied before carrying out the clustering process. This study will examine the effect of the data reduction algorithm, Principal Component Analysis, on the results of the Self Organizing Map clustering algorithm which uses 3 distance measurement methods, Chebyshev distance, Minkowski distance, and Cosine distance. The clustering results will be evaluated using the Davies Bouldin Index. The results showed that the Chebyshev distance gave the best clustering results and the application of the Principal Component Analysis algorithm had a negative impact on the results of the Self Organizing Map clustering and the 3 distance measurements used.

Keywords: Data Mining, Clustering, Chebyshev distance, Minkowski distance, Cosine distance, Principal Component Analysis, Self Organizing Map, Davies Bouldin Index

# PENGARUH ALGORITMA PCA TERHADAP PERHITUNGAN NILAI RENTANG PADA *CLUSTERING SELF ORGANIZING MAP*

Oleh:

Muhammad Albar  
09021281520116

## ABSTRAK

Data mining adalah sebuah proses pengenalan pola yang bertujuan untuk menemukan informasi penting pada data. Clustering merupakan salah satu metode data mining. Clustering bekerja dengan cara mengelompokkan data berdasarkan kesamaan atau ketidaksamaannya. Pengukuran rentang merupakan komponen yang sangat penting dalam clustering. Pengukuran rentang yang berbeda akan menghasilkan nilai clustering yang berbeda, akan tetapi sebelum melakukan sebuah clustering, data harus melalui pengolahan awal terlebih dahulu atau yang disebut praproses. Praproses data dapat meningkatkan hasil dan efisiensi dari sebuah algoritma clustering. Reduksi data merupakan salah satu metode pada praproses data yang dapat diterapkan sebelum melakukan proses clustering. Penelitian ini akan meneliti tentang pengaruh algoritma reduksi data, Principal Component Analysis, terhadap hasil clustering algoritma Self Organizing Map yang menggunakan 3 metode pengukuran rentang, yaitu Chebyshev distance, Minkowski distance, dan Cosine distance. Hasil clustering akan dievaluasi menggunakan Davies Bouldin Index. Hasil penelitian menunjukkan bahwa Chebyshev distance memberikan hasil clustering terbaik dan penerapan algoritma Principal Component Analysis memberikan dampak negatif terhadap hasil clustering Self Organizing Map dan 3 pengukuran rentang yang digunakan.

*Keywords: Data Mining, Clustering, Chebyshev distance, Minkowski distance, Cosine distance, Principal Component Analysis, Self Organizing Map, Davies Bouldin Index*

## KATA PENGANTAR

Puji syukur kepada Allah SWT atas berkat dan rahmat-Nya yang telah diberikan kepada Penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan baik. Tugas akhir ini disusun untuk memenuhi salah satu syarat guna menyelesaikan pendidikan program Strata-1 pada Fakultas Ilmu Komputer Program Studi Teknik Informatika di Universitas Sriwijaya.

Dalam menyelesaikan Tugas Akhir ini banyak pihak yang telah memberikan bantuan dan dukungan baik secara langsung maupun secara tidak langsung. Penulis ingin menyampaikan rasa terima kasih kepada:

1. Orang tuaku tercinta, Ayah Edi Thomas dan Lela Khairani, dan seluruh keluarga besarku yang selalu mendoakan serta memberikan dukungan baik moril maupun materil.
2. Bapak Jaidan Jauhari, M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya, Ibu Alvi Syahrini Utami, M.Kom. selaku Ketua Jurusan Teknik Informatika, dan Ibu Mastura Diana Marieska, S.T., M.T. selaku Sekretaris Jurusan Teknik Informatika.
3. Ibu Dian Palupi Rini, M.Kom., Ph.D. selaku dosen pembimbing I dan Bapak Kanda Januar Miraswan, M.T. selaku pembimbing II yang telah membimbing, mengarahkan, dan memberikan motivasi serta kasih sayang kepada penulis selama proses perkuliahan dan pengerjaan Tugas Akhir.
4. Bapak Al Farissi, S.Kom., M.CS. dan Bapak Ali Ibrahim, M.T. selaku dosen pembimbing akademik, yang telah membimbing dan selalu memberikan motivasi kepada penulis dalam proses perkuliahan dan pengerjaan Tugas Akhir.
5. Bapak Dr. Abdiansah, S.Kom., M.CS. selaku dosen penguji I, Bapak Muhammad Ali Buchari, M.T. selaku dosen penguji II Seminar Proposal, dan Ibu Desty Rodiah, M.T. selaku penguji II Komprehensif yang telah memberikan masukan dan dorongan dalam proses pengerjaan Tugas Akhir.
6. Seluruh dosen Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.



7. Seluruh staf tata usaha yang telah membantu dalam kelancaran proses administrasi dan akademik selama masa perkuliahan.
8. Seluruh teman jurusan Teknik Informatika yang telah berbagi keluh kesah, motivasi, semangat, dan canda tawa selama masa perkuliahan.
9. Sri Astuti, S.E.
10. Ade Lismita Zulkarnain

Penulis menyadari dalam penyusunan Tugas Akhir ini masih terdapat banyak kekurangan disebabkan keterbatasan pengetahuan dan pengalaman, oleh karena itu kritik dan saran yang membangun sangat diharapkan untuk kemajuan penelitian selanjutnya. Akhir kata semoga Tugas Akhir ini dapat berguna dan bermanfaat bagi kita semua.

Indralaya, Desember 2020

Muhammad Albar

## DAFTAR ISI

	Halaman
HALAMAN JUDUL .....	i
HALAMAN LEMBAR PENGESAHAN .....	ii
HALAMAN TANDA LULUS UJIAN SIDANG TUGAS AKHIR .....	iii
HALAMAN PERNYATAAN .....	iv
MOTTO DAN PERSEMBAHAN .....	v
ABSTRACT .....	vi
ABSTRAK .....	vii
KATA PENGANTAR .....	viii
DAFTAR ISI .....	x
DAFTAR TABEL .....	xv
DAFTAR GAMBAR .....	xvii
DAFTAR LAMPIRAN .....	xx
BAB I PENDAHULUAN	
1.1 Pendahuluan .....	I-1
1.2 Latar Belakang .....	I-1
1.3 Rumusan Masalah .....	I-4
1.4 Tujuan Penelitian .....	I-4
1.5 Manfaat Penelitian .....	I-5
1.6 Batasan Masalah .....	I-5
1.7 Sistematika Penulisan .....	I-6
1.8 Kesimpulan .....	I-7
BAB II KAJIAN LITERATUR	
2.1 Pendahuluan .....	II-1
2.2 <i>Clustering</i> .....	II-1
2.2.1 <i>Self Organizing Map</i> .....	II-1
2.3 Rentang Data .....	II-5
2.3.1 <i>Chebyshev Distance</i> .....	II-6
2.3.2 <i>Minkowski Distance</i> .....	II-7
2.3.3 <i>Cosine Distance</i> .....	II-8

2.4	Praproses .....	II-9
2.4.1	<i>Principal Component Analysis</i> .....	II-10
2.5	Evaluasi .....	II-13
2.5.1	<i>Davies Bouldin Index (DBI)</i> .....	II-13
2.6	<i>Rational Unified Process (RUP)</i> .....	II-15
2.7	Penelitian Lain yang Relevan .....	II-17
2.7.1	A comparative study of K Means Algorithm by Different <i>Distance Measures</i> .....	II-17
2.7.2	K-means with Three different <i>Distance Metrics</i> .....	II-18
2.7.3	Effect of Different <i>Distance Measures</i> on the Performance of K-Means Algorithm: An Experimental Study in Matlab .....	II-18
2.7.4	Comparison of Dimensional Reduction Using Singular Value Decomposition and Principal Component Analysis for <i>Clustering Results of Indonesian Language Text Documents</i> .....	II-19
2.8	Kesimpulan .....	II-20

### BAB III METODOLOGI PENELITIAN

3.1	Pendahuluan .....	III-1
3.2	Metode Pengumpulan Data .....	III-1
3.2.1	Jenis dan Sumber Data .....	III-1
3.2.2	Metode Pengumpulan Data .....	III-2
3.3	Tahapan Penelitian .....	III-2
3.3.1	Menetapkan Kerangka Kerja .....	III-2
3.3.1.1	Praproses.....	III-2
3.3.1.2	Pengelompokkan Menggunakan Algoritma SOM dengan <i>Chebyshev Distance</i> .....	III-4
3.3.1.3	Pengelompokkan Menggunakan Algoritma SOM dengan <i>Minkowski Distance</i> .....	III-4
3.3.1.4	Pengelompokkan Menggunakan Algoritma SOM dengan <i>Cosine Distance</i> .....	III-5
3.3.1.5	Evaluasi <i>Clustering</i> Menggunakan <i>Davies Bouldin</i>	
3.3.1.6	<i>Index (DBI)</i> .....	III-5

3.3.2	Menetapkan Kriteria Pengujian .....	III-6
3.3.3	Menetapkan Format Data Pengujian .....	III-7
3.3.4	Menentukan Alat yang Digunakan dalam Pelaksanaan Penelitian .	III-9
3.3.5	Melakukan Pengujian Penelitian .....	III-9
3.3.6	Melakukan Analisa Hasil Pengujian dan Membuat Kesimpulan Penelitian .....	III-14
3.4	Metode Pengembangan Perangkat Lunak .....	III-16
3.4.1	Fase Insepsi .....	III-16
3.4.2	Fase Elaborasi .....	III-16
3.4.3	Fase Konstruksi .....	III-17
3.4.4	Fase Transisi.....	III-17
3.5	Manajemen Proyek Penelitian .....	III-18

#### BAB IV PENGEMBANGAN PERANGKAT LUNAK

4.1	Pendahuluan .....	IV-1
4.2	<i>Rational Unified Process</i> (RUP) .....	IV-1
4.2.1	Fase Insepsi .....	IV-1
4.2.1.1	Pemodelan Bisnis .....	IV-1
4.2.1.2	Kebutuhan Sistem .....	IV-2
4.2.1.3	Analisis dan Desain .....	IV-4
4.2.1.3.1	Analisis Kebutuhan Perangkat Lunak .....	IV-4
4.2.1.3.2	Analisis Data .....	IV-5
4.2.1.3.3	Analisis Praproses Data .....	IV-6
4.2.1.3.4	Analisis Perhitungan <i>Chebyshev Distance</i> .....	IV-15
4.2.1.3.5	Analisis Perhitungan <i>Minkowski Distance</i> .....	IV-16
4.2.1.3.6	Analisis Perhitungan <i>Cosine Distance</i> .....	IV-16
4.2.1.3.7	Analisis Algoritma <i>Self Organizing Map</i> .....	IV-17
4.2.1.3.8	Analisis Evaluasi Menggunakan <i>Davies Bouldin Index</i> (DBI).....	IV-21
4.2.1.3.9	Desain Perangkat Lunak .....	IV-24
4.2.2	Fase Elaborasi .....	IV-30
4.2.2.1	Pemodelan Bisnis .....	IV-31

4.2.2.1.1	Perancangan Data .....	IV-31
4.2.2.1.2	Perancangan Antarmuka .....	IV-32
4.2.2.2	Kebutuhan Sistem .....	IV-32
4.2.2.3	Diagram .....	IV-33
4.2.2.3.1	Diagram Aktivitas .....	IV-33
4.2.2.3.2	Diagram <i>Sequence</i> .....	IV-35
4.2.3	Fase Konstruksi .....	IV-40
4.2.3.1	Kebutuhan Sistem .....	IV-40
4.2.3.2	Diagram Kelas .....	IV-40
4.2.3.3	Implementasi .....	IV-41
4.2.3.3.1	Implementasi Kelas .....	IV-41
4.2.3.3.2	Implementasi Antar Muka .....	IV-43
4.2.4	Fase Transisi .....	IV-44
4.2.4.1	Pemodelan Bisnis .....	IV-44
4.2.4.2	Kebutuhan Sistem .....	IV-44
4.2.4.3	Rencana Pengujian .....	IV-45
4.2.4.3.1	Rencana Pengujian <i>Use Case</i> Melakukan <i>Preprocess</i> .....	IV-45
4.2.4.3.2	Rencana Pengujian <i>Use Case</i> Melakukan <i>Set Initial Weight</i> .....	IV-46
4.2.4.3.3	Rencana Pengujian <i>Use Case</i> Melakukan <i>Process Clustering SOM</i> .....	IV-46
4.2.4.3.4	Rencana Pengujian <i>Use Case</i> Melakukan <i>Refresh</i> .....	IV-47
4.2.4.4	Implementasi .....	IV-47
4.2.4.4.1	Pengujian <i>Use Case</i> Melakukan <i>Preprocess</i> .....	IV-48
4.2.4.4.2	Pengujian <i>Use Case</i> Melakukan <i>Set Initial Weight</i> ...	IV-50
4.2.4.4.3	Pengujian <i>Use Case</i> Melakukan <i>Process Clustering SOM</i> .....	IV-51
4.2.4.4.4	Pengujian <i>Use Case</i> Melakukan <i>Refresh</i> .....	IV-52
4.3	Kesimpulan .....	IV-53

## BAB V HASIL DAN ANALISIS PENELITIAN

5.1	Pendahuluan .....	V-1
-----	-------------------	-----

5.2	Data Hasil Percobaan/Penelitian .....	V-1
5.2.1	Konfigurasi Percobaan .....	V-1
5.2.2	Data Hasil Pengujian .....	V-3
5.2.2.1	Pengujian <i>Dataset Iris</i> .....	V-3
5.2.2.2	Pengujian <i>Dataset Wine</i> .....	V-9
5.2.2.3	Pengujian <i>Dataset Car evaluation</i> .....	V-15
5.2.2.4	Pengujian <i>Dataset Abalone</i> .....	V-21
5.2.2.5	Perbandingan Nilai DBI .....	V-28
5.3	Analisis Hasil Penelitian .....	V-32
5.4	Kesimpulan .....	V-54

## BAB VI KESIMPULAN DAN SARAN

6.1	Pendahuluan .....	VI-1
6.2	Kesimpulan .....	VI-1
6.3	Saran.....	VI-2

DAFTAR PUSTAKA .....	xix
----------------------	-----

LAMPIRAN.....	xxii
---------------	------

## DAFTAR TABEL

	Halaman
III-1. Tabel Rancangan Pengujian <i>Dataset Iris</i> .....	III-7
III-2. Tabel Rancangan Pengujian <i>Dataset Wine</i> .....	III-7
III-3. Tabel Rancangan Pengujian <i>Dataset Car evaluation</i> .....	III-8
III-4. Tabel Rancangan Pengujian <i>Dataset Abalone</i> .....	III-8
III-5. Rancangan Tabel Hasil Perbandingan <i>Clustering</i> .....	III-15
IV-1 Tabel Kebutuhan Fungsional .....	IV-3
IV-2 Tabel Kebutuhan Non-Fungsional .....	IV-4
IV-3 Tabel Contoh Data Uji Coba .....	IV-7
IV-4 Tabel Hasil Pembobotan Data Uji Coba .....	IV-7
IV-5 Tabel Hasil Mengganti Nilai yang Hilang .....	IV-8
IV-6 Tabel Hasil Normalisasi Data Uji .....	IV-9
IV-7 Tabel Data Awal.....	IV-9
IV-8 Tabel Hasil Standarisasi Data .....	IV-10
IV-9 Tabel <i>Covariance</i> .....	IV-11
IV-10 Tabel <i>Eigenvalues</i> .....	IV-12
IV-11 Tabel <i>Eigenvectors</i> .....	IV-12
IV-12 Data Hasil PCA .....	IV-14
IV-13 Tabel Contoh Data Bobot Awal .....	IV-20
IV-14 Tabel Hasil <i>Clustering</i> Menggunakan Algoritma SOM dan 3 Metode Perhitungan Rentang Data .....	IV-21
IV-15 Tabel <i>Centeroid</i> Hasil <i>Cluster</i> SOM dan <i>Chebyshev</i> .....	IV-21
IV-16 Tabel Definisi Aktor .....	IV-25
IV-17 Tabel Definisi <i>Use Case</i> .....	IV-25
IV-18 Tabel Skenario <i>Use Case Preprocess</i> .....	IV-27
IV-19 Tabel Skenario <i>Use Case Set Initial Weight</i> .....	IV-28
IV-20 Tabel Skenario <i>Use Case Process</i> .....	IV-29
IV-21 Tabel Skenario <i>Use Case Refresh</i> .....	IV-30
IV-22 Tabel Implementasi Kelas .....	IV-42
IV-23 Tabel Rencana Pengujian <i>Use Case</i> Melakukan <i>Preprocess</i> .....	IV-45
IV-24 Tabel Rencana Pengujian <i>Use Case</i> Melakukan <i>Set Initial Weight</i> ..	IV-46

IV-25	Tabel Rencana Pengujian <i>Use Case</i> Melakukan <i>Process Clustering</i> SOM .....	IV-46
IV-26	Tabel Rencana Pengujian <i>Use Case</i> Melakukan <i>Refresh</i> .....	IV-47
IV-27	Tabel Rencana Pengujian <i>Use Case</i> Melakukan <i>Preprocess</i> .....	IV-48
IV-28	Tabel Rencana Pengujian <i>Use Case</i> Melakukan <i>Set Initial Weight</i> ....	IV-50
IV-29	Tabel Rencana Pengujian <i>Use Case</i> Melakukan <i>Process Clustering</i> SOM .....	IV-51
IV-30	Tabel Rencana Pengujian <i>Use Case</i> Melakukan <i>Refresh</i> .....	IV-52
V-1	Tabel Pengujian <i>Dataset Iris</i> dengan Jumlah <i>Cluster</i> = 3 .....	V-3
V-2	Tabel Pengujian <i>Dataset Iris</i> dengan Jumlah <i>Cluster</i> = 4 .....	V-4
V-3	Tabel Pengujian <i>Dataset Iris</i> dengan Jumlah <i>Cluster</i> = 5 .....	V-6
V-4	Tabel Pengujian <i>Dataset Iris</i> dengan Jumlah <i>Cluster</i> = 6 .....	V-7
V-5	Tabel Pengujian <i>Dataset Wine</i> dengan Jumlah <i>Cluster</i> = 3 .....	V-9
V-6	Tabel Pengujian <i>Dataset Wine</i> dengan Jumlah <i>Cluster</i> = 4 .....	V-10
V-7	Tabel Pengujian <i>Dataset Wine</i> dengan Jumlah <i>Cluster</i> = 5 .....	V-12
V-8	Tabel Pengujian <i>Dataset Wine</i> dengan Jumlah <i>Cluster</i> = 6 .....	V-13
V-9	Tabel Pengujian <i>Dataset Car evaluation</i> dengan Jumlah <i>Cluster</i> = 3 .	V-15
V-10	Tabel Pengujian <i>Dataset Car evaluation</i> dengan Jumlah <i>Cluster</i> = 4 .	V-16
V-11	Tabel Pengujian <i>Dataset Car evaluation</i> dengan Jumlah <i>Cluster</i> = 5 .	V-18
V-12	Tabel Pengujian <i>Dataset Car evaluation</i> dengan Jumlah <i>Cluster</i> = 6 .	V-19
V-13	Tabel Pengujian <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 10 .....	V-21
V-14	Tabel Pengujian <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 20 .....	V-22
V-15	Tabel Pengujian <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 25 .....	V-24
V-16	Tabel Pengujian <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 29 .....	V-25
V-17	Tabel Pengujian <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 35 .....	V-26
V-18	Tabel Perbandingan Nilai DBI pada Iterasi Ke-500 untuk <i>Clustering</i> .....	V-28



## DAFTAR GAMBAR

	Halaman
II-1. Gambar Hubungan Ketetanggaan pada Topologi <i>Linier Array</i> .....	II-4
II-2. Gambar Arsitektur Model Iteratif RUP (Kruchten, 2003) .....	II-16
III-1. Gambar Tahapan Pengujian <i>Clustering</i> Non-PCA .....	III-9
III-2. Gambar Tahapan Pengujian <i>Clustering</i> dengan PCA .....	III-12
III-3. Gambar Perbandingan Nilai DBI antara <i>Clustering</i> Non-PCA dan PCA .....	III-13
IV-1 Gambar Contoh Perhitungan Mencari Nilai <i>Mean</i> .....	IV-8
IV-2 Gambar Contoh Perhitungan <i>Min-Max Normalization</i> .....	IV-8
IV-3 Gambar Contoh Perhitungan <i>Zero-Mean Standardization</i> .....	IV-10
IV-4 Gambar Persamaan <i>Covariance</i> .....	IV-10
IV-5 Gambar Matriks A .....	IV-11
IV-6 Gambar Persamaan Matriks .....	IV-12
IV-7 Gambar Matriks B .....	IV-13
IV-8 Gambar Matriks C .....	IV-13
IV-9 Gambar Matriks D .....	IV-14
IV-10 Gambar Contoh Perhitungan <i>Chebyshev Distance</i> .....	IV-15
IV-11 Gambar Contoh Perhitungan <i>Minkowski Distance</i> .....	IV-16
IV-12 Gambar Contoh Perhitungan <i>Cosine Distance</i> .....	IV-17
IV-13 Gambar Contoh Perhitungan Algoritma SOM dan <i>Chebyshev</i> <i>Distance</i> .....	IV-18
IV-14 Gambar Contoh Perhitungan Algoritma SOM dan <i>Minkowski</i> <i>Distance</i> .....	IV-19
IV-15 Gambar Contoh Perhitungan Algoritma SOM dan <i>Cosine</i> <i>Distance</i> .....	IV-20
IV-16 Gambar Diagram <i>Use Case</i> .....	IV-24
IV-17 Gambar Rancangan Antarmuka .....	IV-32
IV-18 Gambar Diagram Aktivitas <i>Preproces</i> .....	IV-34
IV-19 Gambar Diagram Aktivitas <i>Set Initial Weight</i> .....	IV-34
IV-20 Gambar Diagram Aktivitas <i>Process</i> .....	IV-35
IV-21 Gambar Diagram Aktivitas <i>Refresh</i> .....	IV-35

IV-22	Gambar Diagram <i>Sequence Preprocess</i> .....	IV-36
IV-23	Gambar Diagram <i>Sequence Set Initial Weight</i> .....	IV-37
IV-24	Gambar Diagram <i>Sequence Process</i> .....	IV-38
IV-25	Gambar Diagram <i>Sequence Refresh</i> .....	IV-39
IV-26	Gambar Diagram Kelas .....	IV-41
IV-27	Gambar Antarmuka Perangkat Lunak .....	IV-43
IV-28	Gambar Contoh <i>Output</i> Perangkat Lunak .....	IV-43
V-1	Gambar Perbandingan Nilai DBI pada <i>Dataset Iris</i> dengan Jumlah <i>Cluster</i> = 3 dan Iterasi = 500 .....	V-32
V-2	Gambar Perbandingan Nilai DBI pada <i>Dataset Iris</i> dengan Jumlah <i>Cluster</i> = 4 dan Iterasi = 500 .....	V-33
V-3	Gambar Perbandingan Nilai DBI pada <i>Dataset Iris</i> dengan Jumlah <i>Cluster</i> = 5 dan Iterasi = 500 .....	V-33
V-4	Gambar Perbandingan Nilai DBI pada <i>Dataset Iris</i> dengan Jumlah <i>Cluster</i> = 6 dan Iterasi = 500 .....	V-34
V-5	Gambar Perbandingan Hasil <i>Cluster</i> Berdasarkan Jumlah Nilai DBI Terkecil pada <i>Dataset Iris</i> .....	V-35
V-6	Gambar Perbandingan Hasil <i>Cluster</i> Non-PCA dan PCA Berdasarkan Jumlah Nilai DBI Terkecil pada <i>Dataset Iris</i> .....	V-36
V-7	Gambar Perbandingan Nilai DBI pada <i>Dataset Wine</i> dengan Jumlah <i>Cluster</i> = 3 dan Iterasi = 500 .....	V-37
V-8	Gambar Perbandingan Nilai DBI pada <i>Dataset Wine</i> dengan Jumlah <i>Cluster</i> = 4 dan Iterasi = 500 .....	V-38
V-9	Gambar Perbandingan Nilai DBI pada <i>Dataset Wine</i> dengan Jumlah <i>Cluster</i> = 5 dan Iterasi = 500 .....	V-38
V-10	Gambar Perbandingan Nilai DBI pada <i>Dataset Wine</i> dengan Jumlah <i>Cluster</i> = 6 dan Iterasi = 500 .....	V-39
V-11	Gambar Perbandingan Hasil <i>Cluster</i> Berdasarkan Jumlah Nilai DBI Terkecil pada <i>Dataset Wine</i> .....	V-40
V-12	Gambar Perbandingan Hasil <i>Cluster</i> Non-PCA dan PCA Berdasarkan Jumlah Nilai DBI Terkecil pada <i>Dataset Wine</i> .....	V-41
V-13	Gambar Perbandingan Nilai DBI pada <i>Dataset Car evaluation</i> dengan	

	Jumlah <i>Cluster</i> = 3 dan Iterasi = 500 .....	V-42
V-14	Gambar Perbandingan Nilai DBI pada <i>Dataset Car evaluation</i> dengan Jumlah <i>Cluster</i> = 4 dan Iterasi = 500 .....	V-43
V-15	Gambar Perbandingan Nilai DBI pada <i>Dataset Car evaluation</i> dengan Jumlah <i>Cluster</i> = 5 dan Iterasi = 500 .....	V-43
V-16	Gambar Perbandingan Nilai DBI pada <i>Dataset Car evaluation</i> dengan Jumlah <i>Cluster</i> = 6 dan Iterasi = 500 .....	V-44
V-17	Gambar Perbandingan Hasil <i>Cluster</i> Berdasarkan Jumlah Nilai DBI Terkecil pada <i>Dataset Car evaluation</i> .....	V-45
V-18	Gambar Perbandingan Hasil <i>Cluster</i> Non-PCA dan PCA Berdasarkan Jumlah Nilai DBI Terkecil pada <i>Dataset Car evaluation</i> .....	V-46
V-19	Gambar Perbandingan Nilai DBI pada <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 10 dan Iterasi = 500 .....	V-47
V-20	Gambar Perbandingan Nilai DBI pada <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 20 dan Iterasi = 500 .....	V-48
V-21	Gambar Perbandingan Nilai DBI pada <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 25 dan Iterasi = 500 .....	V-48
V-22	Gambar Perbandingan Nilai DBI pada <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 29 dan Iterasi = 500 .....	V-49
V-23	Gambar Perbandingan Nilai DBI pada <i>Dataset Abalone</i> dengan Jumlah <i>Cluster</i> = 35 dan Iterasi = 500 .....	V-50
V-24	Gambar Perbandingan Hasil <i>Cluster</i> Berdasarkan Jumlah Nilai DBI Terkecil pada <i>Dataset Abalone</i> .....	V-50
V-25	Gambar Perbandingan Hasil <i>Cluster</i> Non-PCA dan PCA Berdasarkan Jumlah Nilai DBI Terkecil pada <i>Dataset Abalone</i> .....	V-51
V-26	Gambar Perbandingan Hasil <i>Cluster</i> Berdasarkan Jumlah Nilai DBI Terkecil pada Semua <i>Dataset</i> .....	V-52
V-27	Gambar Perbandingan Hasil <i>Cluster</i> Non-PCA dan PCA Berdasarkan Jumlah Nilai DBI Terkecil pada Semua <i>Dataset</i> .....	V-53

## DAFTAR LAMPIRAN

	Halaman
L-1. Perbandingan Nilai DBI untuk <i>Clustering</i> .....	L-1
L-2. Perbandingan Nilai DBI pada <i>Dataset Iris</i> .....	L-14
L-3. Perbandingan Nilai DBI pada <i>Dataset Wine</i> .....	L-20
L-4. Perbandingan Nilai DBI pada <i>Dataset Car Evaluation</i> .....	L-26
L-5. Perbandingan Nilai DBI pada <i>Dataset Abalone</i> .....	L-32
L-6. <i>Source Code Program</i> .....	L-39

# **BAB I**

## **PENDAHULUAN**

### **1.1 Pendahuluan**

Bab ini membahas mengenai latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, sistematika penulisan dan kesimpulan. Pada bab ini juga memberikan penjelasan umum mengenai keseluruhan penelitian terkait tentang *clustering* dan pengukuran rentang data. Pendahuluan dimulai dengan penjelasan mengenai apa yang menjadi latar belakang dari penelitian ini.

### **1.2 Latar Belakang Masalah**

*Data mining* didefinisikan sebagai proses menemukan pola pada data. Proses tersebut dapat dilakukan secara otomatis atau semi-otomatis. Pola yang ditemukan harus memiliki makna agar dapat dimanfaatkan (Witten *et al.*, 2017). Salah satu metode yang digunakan pada *data mining* adalah *clustering*. *Clustering* adalah sebuah proses pengelompokan data yang didasarkan pada kemiripan atau ketidakmiripan dari hasil pengukurannya. *Clustering* dapat diaplikasikan ke banyak area seperti *data mining*, pengenalan pola dan kompresi data. Secara umum, teknik *clustering* terbagi menjadi dua kategori utama yaitu *hierarchical* dan *partitional clustering* (Kumar *et al.*, 2014).

Komponen penting pada algoritma *clustering* adalah pengukuran rentang (Kouser & Sunita, 2013). Pengukuran rentang berfungsi menentukan jarak antara

elemen atau objek dari suatu set. Kesamaan antara berbagai objek ditentukan oleh pengukuran rentang. Pemilihan pengukuran rentang yang tepat pada *dataset* yang digunakan bukanlah hal yang sepele, karena hal ini akan sangat mempengaruhi hasil dari algoritma *clustering* (Vimal *et al.*, 2008).

Penelitian mengenai pengukuran rentang telah dilakukan oleh beberapa peneliti. Penelitian yang membandingkan performa dari beberapa metode pengukuran rentang yaitu *Euclidean distance*, *Manhattan distance*, dan *Chebyshev distance* serta algoritma yang digunakan adalah algoritma *clustering K-Means* (Kouser & Sunita, 2013). Hasil dari penelitian ini menyatakan bahwa algoritma *K-Means* dengan menggunakan *Chebyshev distance* memberikan hasil akurasi yang paling besar dibandingkan dengan *Euclidean distance* dan *Manhattan distance*, namun jumlah iterasi pada *Chebyshev distance* jauh lebih besar dibandingkan dengan *Euclidean distance* dan *Manhattan distance*.

Penelitian serupa juga telah dilakukan Kirdat dan Patil. Dalam penelitiannya, metode pengukuran rentang yang digunakan adalah *Chebyshev distance* dan *Minkowski distance* pada teknik *Content Based Image Retrieval (CBIR)* (Kirdat & Patil, 2016). Hasil yang didapat dari penelitian ini adalah *Chebyshev distance* memberikan performa yang lebih tinggi dibandingkan *Minkowski Distance*.

Penelitian pengukuran rentang data menggunakan tiga metode yaitu *Euclidean distance*, *Manhattan distance*, dan *Chebyshev distance* juga telah dilakukan oleh Sari *et al* (2018). Penelitian ini menggunakan algoritma *clustering Self Organizing Map*. Hasil yang didapatkan pada penelitian tersebut adalah *Chebyshev distance* menghasilkan hasil *cluster* yang lebih baik dibandingkan

*Euclidean distance* dan *Manhattan distance*. Namun, waktu komputasinya lebih lama dibandingkan kedua metode perhitungan rentang data tersebut (Sari *et al.*, 2018).

Penelitian perhitungan rentang lain terdahulu menunjukkan bahwa *clustering* metode perhitungan rentang *Chebyshev distance* cenderung memberikan hasil yang lebih baik dibandingkan metode perhitungan rentang lainnya, akan tetapi proses komputasinya membutuhkan waktu yang lebih lama. Mengingat masih banyaknya metode perhitungan rentang lain yang bisa dibandingkan dengan metode perhitungan rentang *Chebyshev distance* untuk mendapatkan metode yang menghasilkan performa terbaik, oleh karena itu pada penelitian ini akan digunakan tiga metode perhitungan rentang yaitu *Chebyshev distance*, *Minkowski distance*, dan *Cosine distance*. Untuk dapat meningkatkan akurasi serta efisiensi dari sebuah algoritma *data mining* maka dibutuhkan praproses data (Han *et al.*, 2012), salah satu tahapan yang dilakukan saat praproses data yaitu reduksi data.

Reduksi dimensi dibagi menjadi dua, yaitu seleksi fitur dan ekstraksi fitur. Pendekatan teknik reduksi dimensi yang baik digunakan untuk algoritma *clustering* adalah ekstraksi fitur karena dapat mereduksi dimensi dengan sedikit kehilangan informasi pada data (Jun *et al.*, 2014). Algoritma *Principal Component Analysis* (PCA) memiliki performa yang baik dalam ekstraksi fitur karena dapat menghilangkan korelasi 100% antar fitur pada (Paukkeri *et al.*, 2011).

Berdasarkan masalah yang diuraikan diatas, penelitian ini akan menguji pengaruh reduksi data menggunakan PCA terhadap metode perhitungan rentang

data dengan tujuan untuk memilih teknik *clustering* yang lebih baik dalam meningkatkan hasil kualitas *Clustering Self Organizing Map* (SOM).

### 1.3 Rumusan Masalah

Masalah yang diangkat pada penelitian ini adalah bagaimana pengaruh Algoritma *Principal Component Analysis* (PCA) terhadap performa pengukuran rentang pada *Clustering Self Organizing Map* (SOM).

Untuk dapat menyelesaikan permasalahan di atas maka diuraikan pertanyaan penelitian sebagai berikut:

1. Bagaimana penggunaan algoritma *Principal Component Analysis* (PCA) sebelum proses *Clustering* ?
2. Bagaimana mengukur rentang data menggunakan metode *Chebyshev distance*, *Minkowski distance*, dan *Cosine distance* ?
3. Bagaimana perbedaan nilai *Clustering Self Organizing Map* (SOM) untuk setiap algoritma pengukuran rentang data antara yang menggunakan algoritma *Principal Component Analysis* (PCA) dengan yang tidak menggunakan algoritma *Principal Component Analysis* (PCA) ?

### 1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mengetahui penggunaan algoritma *Principal Component Analysis* (PCA).



2. Mengetahui cara pengukuran rentang data menggunakan metode metode *Chebyshev distance*, *Minkowski distance*, dan *Cosine distance* serta mengetahui nilai DBI dari setiap metode pengukuran rentang.
3. Mengetahui perbedaan nilai DBI *Clustering Self Organizing Map* (SOM) untuk setiap algoritma pengukuran rentang data antara yang menggunakan algoritma *Principal Component Analysis* (PCA) dengan yang tidak menggunakan algoritma *Principal Component Analysis* (PCA).

### 1.5 Manfaat Penelitian

Manfaat penelitian ini adalah:

1. Mengetahui algoritma pengukuran rentang data mana yang memiliki nilai terbaik untuk algoritma *Clustering Self Organizing Map* (SOM).
2. Mengetahui pengaruh algoritma reduksi data *Principal Component Analysis* (PCA) terhadap nilai *Clustering Self Organizing Map* (SOM).

### 1.6 Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah sebagai berikut:

1. *Dataset* yang digunakan sebagai objek penelitian merupakan *dataset* standar yang diambil dari *UCI Machine Learning Repository*, berupa *Dataset Iris*, *Wine*, *Car evaluation*, dan *Abalone* yang memiliki tipe data dan jumlah atribut yang berbeda-beda.
2. Topologi dari SOM yang digunakan adalah *Linier Array* dan jari-jari 0

3. Nilai  $P$  yang digunakan untuk *Minkowski distance* adalah 3
4. Pengukuran performansi algoritma *clustering Self Organizing Map* (SOM) menggunakan *Davies Bouldin Index* (DBI).

## 1.7 Sistematika Penulisan

Sistematika penulisan ini adalah sebagai berikut:

### **BAB I. PENDAHULUAN**

Bab I menguraikan latar belakang masalah, rumusan masalah, tujuan dan manfaat penelitian, batasan masalah, sistematika penulisan penelitian dan kesimpulan.

### **BAB II. KAJIAN LITERATUR**

Bab II berisi landasan teori yang digunakan pada penelitian. Pembahasan dasar-dasar teori tersebut meliputi penjelasan mengenai *clustering* dan metode pengukuran rentang data, penjelasan mengenai algoritma *clustering Self Organizing Map* (SOM), algoritma *Principal Component Analysis* (PCA), serta penjelasan lainnya yang berkaitan dengan penelitian ini. Pada bab ini juga akan membahas hasil penelitian terdahulu yang relevan dengan penelitian ini.

### **BAB III. METODOLOGI PENELITIAN**

Bab III akan membahas mengenai tahapan serta pengembangan perangkat lunak yang akan dilaksanakan pada penelitian ini. Arsitektur mengenai tahapan penelitian dibangun dengan sistem kerangka kerja yang dideskripsikan secara rinci.

## **BAB IV. PENGEMBANGAN PERANGKAT LUNAK**

Bab IV membahas perancangan perangkat lunak yang akan dibangun pada penelitian ini.

## **BAB V. HASIL DAN ANALISIS PENELITIAN**

Bab V menampilkan hasil pengujian perangkat lunak berdasarkan langkah-langkah yang telah direncanakan. Analisis diberikan sebagai basis dari kesimpulan yang diambil dalam penelitian ini.

## **BAB VI. KESIMPULAN DAN SARAN**

Bab VI berisi tentang kesimpulan dari semua uraian pada semua bab dan juga berisi saran-saran yang diharapkan dapat membantu dalam pengembangan penelitian selanjutnya.

### **1.8 Kesimpulan**

Bab ini telah menguraikan latar belakang masalah penelitian. Dimana pengukuran rentang data memiliki peran penting dalam proses *clustering*, namun dengan melihat beberapa penelitian sebelumnya dapat disimpulkan bahwa *clustering* pada *dataset* yang berbeda akan menghasilkan metode pengukuran rentang terbaik yang mungkin berbeda, serta algoritma reduksi data dapat meningkatkan hasil *clustering* (Hasanah *et al.*, 2018).

Oleh karena itu, penelitian ini akan membahas mengenai pengaruh algoritma reduksi data *Principal Component Analysis* (PCA) terhadap performa perhitungan rentang data *Chebyshev distance*, *Minkowski distance*, dan *Cosine distance* pada algoritma *Clustering Self Organizing Map* (SOM).

## DAFTAR PUSTAKA

- Alasadi, S. A., & Bhaya, W. S. (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, 12(16), 4102–4107. <https://doi.org/10.3923/jeasci.2017.4102.4107>
- Ali, S., & Smith, K. A. (2006). On Learning Algorithm Selection for Classification. *Applied Soft Computing Journal*, 6(2), 119–138. <https://doi.org/10.1016/j.asoc.2004.12.002>
- Anand, S., Padmanabham, P., & Govardhan, A. (2015). Effect of *Distance* Measures on Partitional *Clustering* Algorithms using Transportation Data. *International Journal of Computer Science and Information Technologies*, 6(6), 5308–5312.
- Arechiga, A., Barocio, E., Ayon, J. J., & Garcia-Baleon, H. A. (2017). Comparison of Dimensionality Reduction Techniques for *Clustering* and Visualization of Load Profiles. *2016 IEEE PES Transmission and Distribution Conference and Exposition-Latin America, PES T and D-LA 2016*, 2, 1–6. <https://doi.org/10.1109/TDC-LA.2016.7805661>
- Awodele, O., & Jegede, O. (2009). Neural Networks and its Application in Engineering. *Proceedings of Informing Science & IT Education Conference (InSITE) 2009*, 1–13. <http://proceedings.informingscience.org/InSITE2009/InSITE09p083-095Awodele542.pdf>
- Bora, M. D. J., & Gupta, D. A. K. (2014). *Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab*. 5(2), 2501–2506. <http://arxiv.org/abs/1405.7471>
- Fausett, L. (1994). *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice-Hall, Inc.
- Galluccio, L., Michel, O., Comon, P., & Hero, A. O. (2012). Graph based K-Means *Clustering*. *Signal Processing*, 92(9), 1970–1984. <https://doi.org/10.1016/j.sigpro.2011.12.009>
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On *Clustering* Validation Techniques. *Journal of Intelligent Information Systems*, 17(3), 107–145. <https://doi.org/10.1023/A>
- Han, J., Kamber, M., & Pei, J. (2012). Data Mining. In *Data Mining* (Third). Morgan Kaufmann. <https://doi.org/10.1016/b978-0-12-381479-1.00001-0>
- Hasanah, S. I. R., Jambak, M. I., & Saputra, D. M. (2018). *Comparison of Dimensional Reduction Using Singular Value Decomposition and Principal Component Analysis for Clustering Results of Indonesian Language Text Documents*. Sriwijaya University.

- Jun, S., Park, S. S., & Jang, D. S. (2014). Document *Clustering* Method using Dimension Reduction and Support Vector *Clustering* to Overcome Sparseness. *Expert Systems with Applications*, 41(7), 3204–3212. <https://doi.org/10.1016/j.eswa.2013.11.018>
- Kirdat, T., & Patil, V. V. (2016). *INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY* 28 Application of Chebyshev Distance and Minkowski Distance to CBIR Using Color Histogram. 2(9), 28–31.
- Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed.). Springer International Publishing.
- Kouser, K., & Sunita, S. (2013). A Comparative Study of K Means Algorithm by Different *Distance* Measures. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(9), 2443–2447. [www.ijirce.com](http://www.ijirce.com)
- Kruchten, P. (2003). The Rational Unified Process--An Introduction. *Rational Software*, May.
- Kumar, V., Chhabra, J. K., & Kumar, D. (2014a). Impact of *Distance* Measures on the Performance of *Clustering* Algorithms. *Advances in Intelligent Systems and Computing*, 243, 183–190. [https://doi.org/10.1007/978-81-322-1665-0\\_17](https://doi.org/10.1007/978-81-322-1665-0_17)
- Liu, P., El-Darzi, E., Lei, L., Vasilakis, C., Chountas, P., & Huang, W. (2005). An Analysis of Missing Data Treatment Methods and Their Application to Health Care *Dataset*. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3584 LNAI, 583–590.
- Miljkovic, D. (2017). Brief Review of Self-Organizing Maps. *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings, October*, 1061–1066. <https://doi.org/10.23919/MIPRO.2017.7973581>
- Mythili, S., & Madhiya, E. (2014). An Analysis on *Clustering* Algorithms in Data Mining. *International Journal of Computer Science and Mobile Computing*, 3(1), 334–340.
- Nawrin, S., Rahatur, M., & Akhter, S. (2017). Exploring K-Means with Internal Validity Indexes for Data *Clustering* in Traffic Management System. *International Journal of Advanced Computer Science and Applications*, 8(3). <https://doi.org/10.14569/ijacsa.2017.080337>
- Pachghare, V. K., Patole, V. A., & Kulkarni, D. P. (2010). Self Organizing Maps to Build Intrusion Detection System. *International Journal of Computer Applications*, 1(8), 1–4. <https://doi.org/10.5120/191-328>
- Pande, S. R., Sambare, M. S. S., & Thakre, V. M. (2012). *Data Clustering Using Data Mining Techniques*. 1(8), 494–499.

- Patro, S. G. K., & sahu, K. K. (2015). Normalization: A Preprocessing Stage. *Iarjset*, 20–22. <https://doi.org/10.17148/iarjset.2015.2305>
- Paukkeri, M. S., Kivimäki, I., Tirunagari, S., Oja, E., & Honkela, T. (2011). Effect of dimensionality reduction on different *distance* measures in document clustering. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7064 LNCS(PART 3), 167–176. [https://doi.org/10.1007/978-3-642-24965-5\\_19](https://doi.org/10.1007/978-3-642-24965-5_19)
- Putra, J. W. G. (2020). *Pengenalan Konsep Pembelajaran Mesin dan Deep Learning* (1.4, Vol. 4, p. 227).
- Rani, S., & Sonika, S. (2014). Effectiveness of Data Preprocessing for Data Mining. *International Journal of Current Engineering and Technology*, 4(5), 3480–3483. <http://www.theseus.fi/handle/10024/62626>
- Sari, T., Rini, D. P., & Yusliani, N. (2018). *PENGARUH PERHITUNGAN RENTANG DATA TERHADAP PERFORMANSI ALGORITMA CLUSTERING SELF ORGANIZING MAP*. Sriwijaya University.
- Singh, A., Yadav, A., & Rana, A. (2013). K-Means with Three Different *Distance* Metrics. *International Journal of Computer Applications*, 67(10), 13–17. <https://doi.org/10.5120/11430-6785>
- Tamilselvi, R., Sivasakthi, B., & Kavitha, R. (2015). *An Efficient Preprocessing and Postprocessing Techniques in Data Mining*.
- Vimal, A., Valluri, S. R., & Karlapalem, K. (2008). *An Experiment with Distance Measures for Clustering . An Experiment with Distance Measures for Clustering. June 2014*.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining Practical Machine Learning Tools and Techniques* (Fourth). Elsevier Inc.
- Zaware, P. S. N., Gautam, M. A., Nashte, M. S., & Khanuja, M. P. (2015). *AN EFFECTUAL APPROACH FOR CALCULATING COSINE SIMILARITY*. 13–18.