

**IMPLEMENTASI METODE *FUZZY C-MEANS* BERBASIS
JARAK *MANHATTAN* DENGAN VALIDASI *MODIFIED*
PARTITION COEFFICIENT PADA *CLUSTERING* DATA
GAYA HIDUP MAHASISWA**

SKRIPSI

**OLEH
AYU HABIBATUN NURWAHYU ZAINURI
NIM. 220601110054**



**PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026**

**IMPLEMENTASI METODE *FUZZY C-MEANS* BERBASIS
JARAK *MANHATTAN* DENGAN VALIDASI *MODIFIED
PARTITION COEFFICIENT* PADA *CLUSTERING* DATA
GAYA HIDUP MAHASISWA**

SKRIPSI

**Diajukan Kepada
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
untuk Memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Matematika (S.Mat)**

**Oleh
Ayu Habibatun Nurwahyu Zainuri
NIM. 220601110054**

**PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2026**

**IMPLEMENTASI METODE *FUZZY C-MEANS* BERBASIS
JARAK *MANHATTAN* DENGAN VALIDASI *MODIFIED*
PARTITION COEFFICIENT PADA *CLUSTERING* DATA
GAYA HIDUP MAHASISWA**

SKRIPSI

Oleh
Ayu Habibatul Nurwahyu Zainuri
NIM. 220601110054

Telah Disetujui untuk Diuji

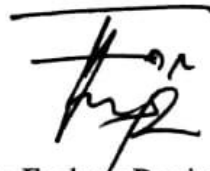
Malang, 18 Mei 2026

Dosen Pembimbing I



Prof. Dr. H. Turmudi, M.Si, Ph.D
NIP. 19571005 198203 1 006

Dosen Pembimbing II



Dr. Fachrur Rozi, M.Si
NIP. 19800527 200801 1 012

Mengetahui,
Ketua Program Studi Matematika



Dr. Fachrur Rozi, M.Si
NIP. 19800527 200801 1 012

**IMPLEMENTASI METODE *FUZZY C-MEANS* BERBASIS
JARAK *MANHATTAN* DENGAN VALIDASI *MODIFIED*
PARTITION COEFFICIENT PADA *CLUSTERING* DATA
GAYA HIDUP MAHASISWA**

SKRIPSI

Oleh
Ayu Habibatul Nurwahyu Zainuri
NIM. 220601110054

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima sebagai Salah Satu Persyaratan
untuk Memperoleh Gelar Sarjana Matematika (S.Mat)

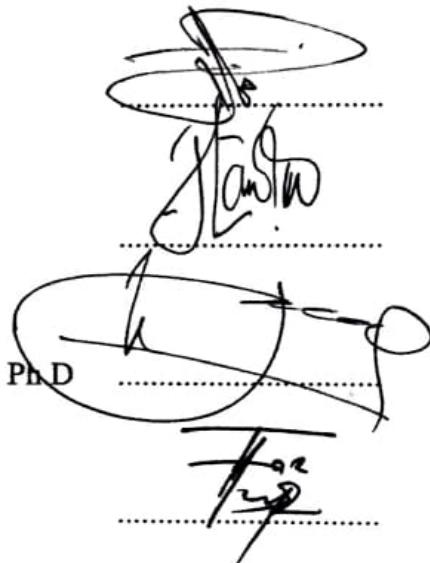
Tanggal 12 Juni 2026

Ketua Penguji : Hisyam Fahmi, M.Kom

Anggota Penguji 1 : Intan Nisfulaila, M.Si

Anggota Penguji 2 : Prof. Dr. H. Turmudi, M.Si, Ph.D

Anggota Penguji 3 : Dr. Fachrur Rozi, M.Si



Mengetahui,
Ketua Program Studi Matematika



Dr. Fachrur Rozi, M.Si
NIP. 19800527 200801 1 012

PERNYATAAN KEASLIAN TULISAN

Saya bertanda tangan di bawah ini

Nama : Ayu Habibatun Nurwahyu Zainuri

NIM : 220601110054

Program Studi : Matematika

Fakultas : Sains dan Teknologi

Judul Skripsi : Implementasi Metode *Fuzzy C-Means* Berbasis Jarak *Manhattan* dengan Validasi *Modified Partition Coefficient* pada *Clustering* Data Gaya Hidup Mahasiswa

Menyatakan bahwa skripsi yang saya susun merupakan hasil karya sendiri dan tidak mengambil atau menyalin tulisan maupun pemikiran orang lain. Seluruh isi skripsi ini adalah hasil pemikiran saya, kecuali bagian yang telah disebutkan sumbernya dalam daftar rujukan pada halaman terakhir. Apabila di kemudian hari terbukti bahwa skripsi ini merupakan hasil plagiasi atau meniru karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 12 Juni 2026



Ayu Habibatun Nurwahyu Zainuri
NIM. 220601110051

MOTO

Allah tidak pernah berkata semuanya mudah, tetapi dua kali Allah menegaskan bahwa, “setiap kesulitan pasti ada kemudahan”

(Q.S. Al-Insyirah:5)

PERSEMBAHAN

Pertama-tama, penulis ucapkan terima kasih kepada Allah SWT yang telah melimpahkan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Skripsi ini penulis persembahkan kepada:

Kedua orang tua tercinta, Ayah Nur dan Ibu Sri, yang selalu menjadi alasan terbesar penulis untuk terus berjuang dan bertahan hingga sampai pada titik ini. Terima kasih atas segala doa, kasih sayang, pengorbanan, dukungan, serta kerja keras yang tidak pernah berhenti diberikan kepada penulis. Selalu menguatkan penulis di setiap proses yang tidak mudah, selalu percaya pada kemampuan penulis, dan selalu memberikan semangat ketika penulis merasa lelah dan ingin menyerah. Tidak ada kata yang cukup untuk menggambarkan rasa syukur penulis memiliki orang tua seperti Ayah dan Ibu.

Adik-adik tercinta, La'Salina dan Levina, yang selalu menjadi penyemangat dan penghibur di tengah berbagai kesibukan dan tekanan selama proses penyusunan skripsi ini.

Semoga segala perjuangan, doa, dan pengorbanan yang telah diberikan mendapatkan balasan terbaik dari Allah SWT. Penulis berharap skripsi ini menjadi langkah awal untuk membanggakan dan membahagiakan keluarga tercinta.

KATA PENGANTAR

Puji syukur ke hadirat Allah SWT atas limpahan rahmat, taufik, dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi ini yang berjudul “Implementasi Metode *Fuzzy C-Means* Berbasis Jarak *Manhattan* dengan Validasi *Modified Partition Coefficient* pada *Clustering* Data Gaya Hidup Mahasiswa” sebagai salah satu syarat untuk memperoleh gelar sarjana pada Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Penulis menyadari bahwa penyusunan skripsi ini tidak terlepas dari bantuan, bimbingan, serta dukungan dari berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si., CAHRM., CRMP selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. Agus Mulyono, M.Kes., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrur Rozi, M.Si., selaku Ketua Program Studi Matematika Universitas Islam Negeri Maulana Malik Ibrahim Malang sekaligus Dosen Pembimbing II yang dengan penuh kesabaran telah memberikan arahan, ilmu, serta bimbingan kepada penulis.
4. Prof. Dr. H. Turmudi, M.Si., Ph. D., selaku Dosen Pembimbing I yang telah banyak memberikan arahan, nasihat, ilmu dan bimbingan kepada penulis dalam penyusunan skripsi ini.
5. Hisyam Fahmi, M.Kom., selaku Ketua Penguji dalam Ujian Skripsi, yang telah berkenan memberikan arahan serta saran yang membangun bagi penyusunan skripsi.
6. Intan Nisfulaila, M.Si., selaku Penguji I dalam Ujian Skripsi, yang telah memberikan arahan dan saran yang bermanfaat kepada penulis dalam penyusunan skripsi.
7. Seluruh Dosen Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim.

8. Dengan penuh rasa hormat dan cinta, penulis ingin mengucapkan terima kasih yang sebesar-sebarnya kepada kedua orang tua tercinta, Ayah Nurwakid dan Ibu Sri. Penulis menyadari bahwa tidak ada kata yang mampu sepenuhnya menggambarkan betapa besar rasa syukur atas segala doa, kasih sayang, serta pengorbanan yang telah diberikan. Terima kasih atas dukungan yang tidak pernah terputus, kesabaran dalam mendampingi setiap proses serta menunggu proses penulis hingga dititik ini. Ayah dan Ibu adalah sumber kekuatan dan alasan utama penulis untuk terus bertahan dan berjuang hingga sampai dititik ini. Setiap waktu, usaha, dan tenaga yang telah dicurahkan untuk penulis menjadi bagian penting dalam perjalanan ini. Semoga skripsi sederhana ini bisa menjadi penghargaan atas segala cinta dan pengorbanan yang telah diberikan. Semoga setiap doa yang dipanjatkan menjadi jalan yang mengantarkan penulis menuju keberhasilan di dunia dan akhirat. Aamiin.
9. Kepada kedua adik tercinta, Lala dan Levina, yang juga menjadi sumber semangat dan kekuatan dalam setiap proses yang dilalui. Terima kasih atas dukungan, keceriaan, serta kehadiran kalian yang memberi energi positif di tengah lelahnya perjalanan menyelesaikan skripsi ini. Meskipun tidak selalu diungkapkan secara langsung, perhatian kecil, canda tawa, dan kebersamaan yang kita miliki memiliki arti yang sangat besar bagi penulis. Semoga kebersamaan dan kasih sayang ini senantiasa terjaga, dan semoga kalian juga dapat meraih impian serta kesuksesan di masa depan.
10. Kepada sepupu tercinta, Puteri, yang juga berjuang menyelesaikan skripsinya. Terima kasih dukungan, perhatian, semangat dan kebersamaan yang kita lalui bersama di tengah proses yang tidak mudah ini. Saling berbagi cerita, keluh kesah, dan motivasi menjadi hal yang sangat berarti bagi penulis untuk tetap kuat melangkah hingga menyelesaikan skripsi ini. Semoga setiap usaha yang kita lakukan membuahkan hasil yang terbaik.
11. Kepada Esha Arya Erdina, sahabat yang telah menemani penulis sejak masa SMA hingga saat ini. Terima kasih karena selalu hadir dalam setiap cerita hidup penulis. Terima kasih atas setiap cerita, dukungan, doa, serta selalu

mendengarkan keluh kesah penulis ketika merasa lelah di tengah kesibukan masing-masing.

12. Kepada Mutiara Alfi, Amira Adelia Putri, Do'aul Isma Mufidah, Nurus Syafi'ah, dan Rossima Eva Yuliana. Terima kasih telah menjadi bagian dari cerita masa perkuliahan penulis. Penulis tidak hanya menemukan teman untuk belajar dan berdiskusi, tetapi juga tempat berbagi cerita, tawa, keluh kesah, serta berbagai pengalaman yang mewarnai kehidupan sebagai mahasiswa. Sejak semester satu hingga di tahap penyelesaian skripsi ini, kehadiran kalian membuat setiap proses yang dijalani terasa lebih ringan dan bermakna. Terima kasih telah menjadi tempat pulang untuk penulis di perantauan ini.
13. Kepada Erma Nur Fatimah, yang telah menjadi salah satu sahabat penulis bertumbuh selama masa perkuliahan. Di tengah berbagai proses yang tidak selalu mudah, selalu ada ruang untuk saling menguatkan dan mengingatkan agar tetap bertahan hingga mencapai tahap ini. Semoga segala kebaikan, dukungan, doa, dan kebersamaan yang telah diberikan menjadi kenangan indah bagi penulis.
14. Seluruh mahasiswa Program Studi Matematika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang angkatan 2022.
15. Terakhir, dengan penuh rasa syukur, saya ingin mengapresiasi dan berterima kasih kepada diri saya sendiri, Ayu Habibatul Nurwahyu Zainuri, seorang anak pertama perempuan yang telah bertahan, berjuang, dan terus melangkah meskipun sering merasa lelah dan hampir menyerah. Terima kasih karena tetap kuat menghadapi setiap proses yang tidak mudah, melewati hari-hari penuh tekanan, serta malam-malam panjang yang diisi dengan usaha dan doa tanpa henti demi menyelesaikan skripsi ini. Hari ini adalah hasil dari perjuangan panjang yang penuh air mata, kesabaran, dan harapan yang tak pernah padam. Berbahagialah dan banggalah atas pencapaian ini, meskipun masih jauh dari kata sempurna tapi kamu telah sampai pada titik yang dulu hanya sebuah impian. Kamu boleh merasa lelah

dan menangis dalam prosesnya, tetapi percayalah bahwa Allah SWT tidak akan membawamu sejauh ini hanya untuk berakhir dengan kegagalan.

Semoga Allah SWT melimpahkan rahmat dan karunia-Nya kepada kita semua. Penulis menyadari bahwa dalam penyusunan skripsi ini masih terdapat berbagai kekurangan yang perlu diperbaiki. Meskipun demikian, penulis berharap agar skripsi ini dapat memberikan manfaat terutama bagi penulis dan pembaca.

Malang, 12 Juni 2026

Penulis

DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PENGAJUAN	ii
HALAMAN PERSETUJUAN	iii
HALAMAN PENGESAHAN	iv
PERNYATAAN KEASLIAN TULISAN	v
MOTO	vi
PERSEMBAHAN	vii
KATA PENGANTAR	viii
DAFTAR ISI	xii
DAFTAR TABEL	xiv
DAFTAR GAMBAR	xv
DAFTAR LAMPIRAN	xvii
ABSTRAK	xviii
ABSTRACT	xix
مستخلص البحث	xx
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	6
1.3 Tujuan Penelitian	6
1.4 Manfaat Penelitian	6
1.5 Batasan Masalah	7
1.6 Definisi Istilah	7
BAB II KAJIAN TEORI	9
2.1 Logika <i>Fuzzy</i>	9
2.1.1 Pengertian Logika <i>Fuzzy</i>	9
2.1.2 Himpunan <i>Fuzzy</i>	10
2.1.3 Derajat Keanggotaan	11
2.2 <i>Clustering</i>	11
2.3 <i>Fuzzy Clustering</i>	13
2.4 <i>Fuzzy C-Means</i>	14
2.5 Normalisasi <i>Min-Max</i>	26
2.6 Jarak <i>Manhattan</i>	27
2.7 <i>Outlier</i>	29
2.8 Validasi <i>Cluster</i>	29
2.8.1 <i>Modified Partition Coefficient</i> (MPC)	29
2.8.2 <i>Purity</i>	32
2.9 Gaya Hidup Mahasiswa	34
2.10 Kajian Integrasi Topik dengan Al-Quran/Hadis	36
BAB III METODE PENELITIAN	38
3.1 Jenis Penelitian	38
3.2 Data dan Sumber Data	38
3.3 Teknik Analisis Data	39
BAB IV HASIL DAN PEMBAHASAN	43
4.1 Analisis Deskriptif	43
4.2 <i>Preprocessing</i> Data	45
4.2.1 Seleksi Variabel	45

4.2.2 Deteksi <i>Outlier</i>	45
4.2.3 Normalisasi Data	50
4.3 Implementasi Algoritma <i>Fuzzy C-Means</i>	52
4.3.1 Perhitungan Menggunakan Jarak <i>Euclidean</i>	53
4.3.2 Perhitungan Menggunakan Jarak <i>Manhattan</i>	65
4.3.3 Perbandingan Hasil <i>Clustering</i> Jarak <i>Euclidean</i> dan Jarak <i>Manhattan</i>	78
4.4 Evaluasi <i>Modified Partition Coefficient</i>	80
4.5 Evaluasi <i>Purity</i> Berdasarkan Label Aktual	83
4.6 Interpretasi Hasil <i>Clustering</i>	90
4.7 Pencarian Solusi Pengelompokan yang Lebih Baik dalam Prespektif Islam	94
BAB V PENUTUP	97
5.1 Kesimpulan	97
5.2 Saran dan Penelitian Lanjutan	98
DAFTAR PUSTAKA	100
LAMPIRAN	103
RIWAYAT HIDUP	114

DAFTAR TABEL

Tabel 3.1	Deskripsi Data	39
Tabel 4.1	Analisis Deskriptif Variabel	43
Tabel 4.2	Normalisasi Data	51
Tabel 4.3	Bilangan <i>Random</i> $\mu_{i,k}$ untuk $c = 2$	55
Tabel 4.4	Kuadrat Bilangan <i>Random</i> $(\mu_{i,k})^2$	56
Tabel 4.5	Pusat <i>Cluster</i> Menggunakan Jarak <i>Euclidean</i>	57
Tabel 4.6	Nilai <i>Cluster</i> 1	59
Tabel 4.7	Nilai <i>Cluster</i> 2	59
Tabel 4.8	Fungsi Objektif Menggunakan Jarak <i>Euclidean</i>	60
Tabel 4.9	Nilai $\mu_{i,k}$ Baru Menggunakan Jarak <i>Euclidean</i>	62
Tabel 4.10	Hasil <i>Clustering</i> Menggunakan Jarak <i>Euclidean</i>	64
Tabel 4.11	Bilangan <i>Random</i> $\mu_{i,k}$ untuk $c = 2$	67
Tabel 4.12	Kuadrat Bilangan <i>Random</i> $(\mu_{i,k})^2$	68
Tabel 4.13	Pusat <i>Cluster</i> Menggunakan Jarak <i>Manhattan</i>	69
Tabel 4.14	Nilai <i>Cluster</i> 1	71
Tabel 4.15	Nilai <i>Cluster</i> 2	71
Tabel 4.16	Hasil Kuadrat Jarak	73
Tabel 4.17	Fungsi Objektif Menggunakan Jarak <i>Manhattan</i>	73
Tabel 4.18	Nilai $\mu_{i,k}$ Baru Menggunakan Jarak <i>Manhattan</i>	75
Tabel 4.19	Hasil <i>Clustering</i> Menggunakan Jarak <i>Manhattan</i>	77
Tabel 4.20	Tabel Perbandingan <i>Outlier</i>	79
Tabel 4.21	Perbandingan Hasil <i>Clustering</i> Jarak <i>Euclidean</i> dan Jarak <i>Manhattan</i>	80
Tabel 4.22	Hasil Validasi $c = 2$ hingga $c = 5$	82
Tabel 4.23	Tabulasi Silang $c = 2$ dan Label Aktual pada Jarak <i>Euclidean</i> ...	84
Tabel 4.24	Tabulasi Silang $c = 2$ dan Label Aktual pada Jarak <i>Manhattan</i> ..	85
Tabel 4.25	Tabulasi Silang $c = 3$ dan Label Aktual pada Jarak <i>Euclidean</i> ...	87
Tabel 4.26	Tabulasi Silang $c = 3$ dan Label Aktual pada Jarak <i>Manhattan</i> ..	88
Tabel 4.27	Hasil <i>Purity</i> $c = 2$ hingga $c = 5$	89
Tabel 4.28	Nilai Rata-rata Variabel Tiap Cluster pada Jarak <i>Euclidean</i>	90
Tabel 4.29	Nilai Rata-rata Variabel Tiap Cluster pada Jarak <i>Manhattan</i>	91

DAFTAR GAMBAR

Gambar 3.1 Diagram Alir Penelitian	42
Gambar 4.1 <i>Boxplot Study Hours Per Day</i>	46
Gambar 4.2 <i>Boxplot Extracurricular Hours Per Day</i>	47
Gambar 4.3 <i>Boxplot Sleep Hours Per Day</i>	48
Gambar 4.4 <i>Boxplot Social Hours Per Day</i>	48
Gambar 4.5 <i>Boxplot Physical Activity Hours Per Day</i>	49
Gambar 4.6 <i>Boxplot GPA</i>	50
Gambar 4.7 Visualisasi Hasil <i>Clustering</i> Menggunakan Jarak <i>Euclidean</i>	65
Gambar 4.8 Visualisasi Hasil <i>Clustering</i> Menggunakan Jarak <i>Manhattan</i>	78

DAFTAR SIMBOL

X_1	: Variabel <i>Study Hours Per Day</i>
X_2	: Variabel <i>Extracurricular Hours Per Day</i>
X_3	: Variabel <i>Sleep Hours Per Day</i>
X_4	: Variabel <i>Social Hours Per Day</i>
X_5	: Variabel <i>Physical Hours Per Day</i>
X_6	: Variabel GPA
X	: Himpunan data
c	: Banyaknya <i>cluster</i>
w	: Pangkat pembobot
$MaxIter$	: Maksimum iterasi
P	: Indeks fungsi objektif
t	: Indeks Iterasi
ε	: <i>Error</i> terkecil yang diharapkan
$\mu_{i,k}$	: Derajat keanggotaan data ke- i pada <i>cluster</i> ke- k
Q_i	: Jumlah derajat keanggotaan pada data ke- i
$\mu_{i,k}^{(0)}$	: Derajat keanggotaan data ke- i pada <i>cluster</i> ke- k sebelum normalisasi
$U^{(0)}$	: Matriks partisi awal
$V_{k,j}$	: Pusat <i>cluster</i> ke- k pada variabel ke- j
$x_{i,j}$	: Data sampel ke- i dan variabel ke- j
P_t	: Fungsi objektif pada iterasi ke- t
n	: Jumlah data
m	: Banyak variabel
x_{min}	: Nilai minimum data
x_{max}	: Nilai maksimum data
x'	: Nilai data normalisasi
$d_{i,k}^2$	: Jarak kuadrat antara data ke- i dan pusat <i>cluster</i> ke- k
$d(x_i, V_k)$	: Jarak <i>Manhattan</i> antara data ke- i dan pusat <i>cluster</i> ke- k
x_i	: Vektor data ke- i dengan elemen x_{ij} untuk variabel ke- j
V_k	: Vektor pusat <i>cluster</i> ke- k dengan elemen $V_{k,j}$
N	: Banyak seluruh data
r	: Banyak label aktual
$n_{k,j}$	: Banyak data pada <i>cluster</i> ke- k yang memiliki label aktual ke- j
$\max_{1 \leq j \leq r} n_{k,j}$	: Banyak label aktual terbanyak pada <i>cluster</i> ke- k

DAFTAR LAMPIRAN

Lampiran 1	<i>Dataset Penelitian</i>	103
Lampiran 2	<i>Analisis Deskriptif</i>	103
Lampiran 3	<i>Dataset setelah Preprocessing</i>	103
Lampiran 4	<i>Hasil Clustering untuk $c = 2$ Menggunakan Jarak Euclidean .</i>	103
Lampiran 5	<i>Hasil Clustering untuk $c = 3$ Menggunakan Jarak Euclidean .</i>	104
Lampiran 6	<i>Hasil Clustering untuk $c = 4$ Menggunakan Jarak Euclidean .</i>	104
Lampiran 7	<i>Hasil Clustering untuk $c = 5$ Menggunakan Jarak Euclidean .</i>	104
Lampiran 8	<i>Hasil Clustering untuk $c = 2$ Menggunakan Jarak Manhattan</i>	104
Lampiran 9	<i>Hasil Clustering untuk $c = 3$ Menggunakan Jarak Manhattan</i>	105
Lampiran 10	<i>Hasil Clustering untuk $c = 4$ Menggunakan Jarak Manhattan</i>	105
Lampiran 11	<i>Hasil Clustering untuk $c = 5$ Menggunakan Jarak Manhattan</i>	105
Lampiran 12	<i>Source Code FCM Menggunakan Jarak Euclidean</i>	105
Lampiran 13	<i>Source Code FCM untuk Menggunakan Jarak Manhattan</i>	108
Lampiran 14	<i>Source Code Visualisasi Hasil Clustering</i>	112

ABSTRAK

Zainuri, Ayu Habibatul Nurwahyu. 2026. **Implementasi Metode Fuzzy C-Means Berbasis Jarak Manhattan dengan Validasi Modified Partition Coefficient pada Clustering Data Gaya Hidup Mahasiswa**. Skripsi. Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Pembimbing: (I) Prof. Dr. H. Turmudi, M.Si, Ph.D. (II) Dr. Fachrur Rozi, M.Si.

Kata Kunci: *Fuzzy C-Means, Euclidean, Manhattan, clustering, Modified Partition Coefficient, Purity.*

Clustering merupakan salah satu metode dalam *data mining* yang digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik tertentu. Salah satu algoritma *clustering* yang dapat digunakan adalah *Fuzzy C-Means*, yaitu metode pengelompokan data yang memungkinkan satu data memiliki derajat keanggotaan pada lebih dari satu *cluster*. Dalam proses *clustering*, ukuran jarak memiliki peran penting karena dapat memengaruhi hasil pengelompokan, terutama pada data yang mengandung *outlier*. Penelitian ini membandingkan hasil *clustering* algoritma *Fuzzy C-Means* menggunakan jarak *Euclidean* dan jarak *Manhattan* pada data gaya hidup mahasiswa. Metode penelitian dilakukan melalui beberapa tahapan, yaitu *preprocessing* data, seleksi variabel, deteksi *outlier*, normalisasi *Min-Max*, penerapan algoritma *Fuzzy C-Means*, serta evaluasi hasil *clustering* menggunakan *Modified Partition Coefficient* dan *Purity*. Variabel yang digunakan dalam proses *clustering* meliputi *Study Hours Per Day*, *Extracurricular Hours Per Day*, *Sleep Hours Per Day*, *Social Hours Per Day*, *Physical Activity Hours Per Day*, dan GPA, sedangkan *Stress Level* digunakan sebagai label aktual pada evaluasi *Purity*. Hasil penelitian menunjukkan bahwa jarak *Manhattan* menghasilkan nilai *Modified Partition Coefficient* dan *Purity* yang lebih tinggi dibandingkan jarak *Euclidean* pada beberapa banyak *cluster* yang diuji. Nilai MPC terbaik diperoleh pada jarak *Manhattan* dengan banyak *cluster* $c = 2$, sedangkan nilai *Purity* tertinggi diperoleh pada jarak *Manhattan* dengan banyak *cluster* $c = 5$. Berdasarkan hasil tersebut, dapat disimpulkan bahwa penggunaan jarak *Manhattan* pada algoritma *Fuzzy C-Means* memberikan hasil *clustering* yang lebih baik dibandingkan jarak *Euclidean* karena lebih stabil terhadap data yang mengandung *outlier* dan menghasilkan derajat keanggotaan yang lebih tegas.

ABSTRACT

Zainuri, Ayu Habibatul Nurwahyu Zainuri. 2026. **Implementation of the Manhattan Distance-Based Fuzzy C-Means Method with Modified Partition Coefficient Validation on Clustering Student Lifestyle Data.** Undergraduate Thesis. Mathematics Study Program, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang.
Advisors: (I) Prof. Dr. H. Turmudi, M.Si, Ph.D. (II) Dr. Fachrur Rozi, M.Si.

Kata Kunci: Fuzzy C-Means, Euclidean, Manhattan, clustering, Modified Partition Coefficient, Purity.

Clustering is a data mining method used to group data based on the similarity of specific characteristics. One clustering algorithm that can be used is Fuzzy C-Means, a data grouping method that allows a single data point to have a degree of membership in more than one cluster. In the clustering process, the choice of distance plays a crucial role because it can influence the clustering results, especially in data containing outliers. This study aims the clustering results of the Fuzzy C-Means algorithm using Euclidean distance and Manhattan distance on student lifestyle data. The research method was conducted through several stages, namely data preprocessing, variable selection, outlier detection, Min-Max normalization, application of the Fuzzy C-Means algorithm, and evaluation of the clustering results using the Modified Partition Coefficient and Purity. The variables used in the clustering process include Study Hours Per Day, Extracurricular Hours Per Day, Sleep Hours Per Day, Social Hours Per Day, Physical Activity Hours Per Day, and GPA, while Stress Level was used as the actual label in the Purity evaluation. The results show that the Manhattan distance yields higher Modified Partition Coefficient and Purity values compared to the Euclidean distance for several tested cluster sizes. The best MPC value was obtained using Manhattan distance with $c = 2$ clusters, while the highest Purity value was obtained using the Manhattan distance with $c = 5$ clusters. Based on these results, it can be concluded that using the Manhattan distance in the Fuzzy C-Means algorithm yields better clustering results than the Euclidean distance because it is more robust to data containing outliers and produces more distinct membership degrees.

مستخلص البحث

العلمي، زين النورايوحبية نوروحي. ٢٠٢٦. تطبيق طريقة *Fuzzy C-Means* القائمة على مسافة مانهاتن مع التحقق من صحة معامل التقسيم المعدل *Modified Partition Coefficient* في تجميع بيانات نمط حياة الطلاب. البحث جامعي. برنامج القسم الرياضيات، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية بمالانج.

المشرف: (١) الأستاذ الدكتور الحاج تورمودي، الماجستير، في العلوم. (٢) الدكتور فخرالرازي، الماجستير في العلوم.

الكلمات الأساسية: *Fuzzy C-Means*، إقليدس *Euclidean*، مانهاتن *Manhattan*، التجميع *Clustering*، معامل التقسيم المعدل MPC، النقاء *Purity*.

التجميع *Clustering* هو أحد الأساليب في تنقيب البيانات المستخدمة لتصنيف البيانات بناءً على تشابه خصائص معينة. ومن خوارزميات التجميع التي يمكن استخدامها هي خوارزمية *Fuzzy C-Means*، وهي طريقة لتجميع البيانات تسمح للبيانات المفردة بامتلاك درجة عضوية في أكثر من مجموعة واحدة. في عملية التجميع، يلعب مقياس المسافة دوراً مهماً لأنه يمكن أن يؤثر على نتائج التجميع، خاصة في البيانات التي تحتوي على قيم متطرفة *outlier*. تهدف هذه الدراسة إلى مقارنة نتائج تجميع خوارزمية *Fuzzy C-Means* باستخدام مسافة إقليدس ومسافة مانهاتن على بيانات نمط حياة الطلاب. تم إجراء منهجية البحث من خلال عدة مراحل، وهي: المعالجة المسبقة للبيانات، واختيار المتغيرات، واكتشاف القيم المتطرفة، والتطبيع بطريقة الحد الأدنى والأقصى *Min-Max*، وتطبيق خوارزمية *Fuzzy C-Means*، بالإضافة إلى تقييم نتائج التجميع باستخدام معامل التقسيم المعدل MPC والنقاء *Purity*. تشمل المتغيرات المستخدمة في عملية التجميع: ساعات الدراسة يوميًا، وساعات الأنشطة اللامنهجية يوميًا، وساعات النوم يوميًا، وساعات التواصل الاجتماعي يوميًا، وساعات النشاط البدني يوميًا، والمعدل التراكمي GPA، بينما تم استخدام مستوى التوتر Stress Level كمصنف فعلي في تقييم النقاء. وأظهرت النتائج أن مسافة مانهاتن أنتجت قيمًا أعلى لمعامل التقسيم المعدل والنقاء مقارنة بمسافة إقليدس في العديد من المجموعات المختبرة. تم الحصول على أفضل قيمة لمعامل التقسيم المعدل MPC باستخدام مسافة مانهاتن مع عدد مجموعات $c = 2$ ، في حين تم الحصول على أعلى قيمة للنقاء باستخدام مسافة مانهاتن مع عدد مجموعات $c = 5$. وبناءً على هذه النتائج، يمكن استنتاج أن استخدام مسافة مانهاتن في خوارزمية *Fuzzy C-Means* يعطي نتائج تجميع أفضل مقارنة بمسافة إقليدس لأنها أكثر استقرارًا تجاه البيانات التي تحتوي على قيم متطرفة وتنتج درجات عضوية أكثر وضوحًا وحسماً.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Mengelola data yang kompleks serta tidak pasti menjadi tantangan yang terus berkembang, karena metode konvensional belum mampu mengatasi data yang tidak lengkap dan samar. Oleh karena itu, diperlukan pendekatan yang mampu menangani ketidakpastian ini, seperti logika *fuzzy*. Logika *fuzzy* merupakan pengembangan dari logika klasik yang hanya memiliki dua nilai kebenaran, yaitu 0 dan 1, dengan memperluasnya menjadi rentang nilai antara 0 hingga 1 (Zadeh, 1965). Setiap nilai memiliki derajat keanggotaan yang menunjukkan seberapa besar nilai tersebut termasuk dalam suatu himpunan. Pendekatan ini memberikan gambaran data yang lebih fleksibel dan akurat, sehingga banyak diterapkan dalam berbagai bidang analisis data yang mengandung ketidakpastian dan ambiguitas.

Konsep logika *fuzzy* kemudian berkembang luas dan menjadi dasar bagi metode analisis data, salah satunya *clustering* atau pengelompokan data. *Clustering* bertujuan untuk membagi sekumpulan data menjadi beberapa kelompok berdasarkan tingkat kemiripan antar data, di mana setiap kelompok memiliki karakteristik yang berbeda satu sama lain (Erick, 2024). Dengan demikian, hasil pengelompokan menjadi lebih realistis dan sesuai dengan kondisi data yang sering kali tidak memiliki batas yang tegas.

Secara umum, metode *clustering* dapat dibedakan menjadi dua, yaitu *hard clustering* dan *fuzzy clustering* (Johra, 2021). *Hard clustering* menetapkan bahwa setiap data hanya menjadi anggota satu *cluster* secara pasti, sedangkan *fuzzy*

clustering memungkinkan suatu data memiliki derajat keanggotaan pada lebih dari satu *cluster* (Erick, 2024). Salah satu metode *hard clustering* yang banyak digunakan adalah *K-Means*, yaitu algoritma yang mengelompokkan data berdasarkan jarak terdekat terhadap pusat *cluster* (Suryadi, 2024). Meskipun sederhana dan efisien, metode ini memiliki keterbatasan karena setiap data hanya dapat dimasukkan ke dalam satu kelompok secara pasti, sehingga kurang efektif dalam menangani data yang memiliki batas antar kelompok yang tidak tegas.

Sebagai pengembangan dari metode *clustering* konvensional, *fuzzy clustering* mengintegrasikan konsep himpunan *fuzzy* ke dalam proses pengelompokan. Salah satu algoritma yang paling banyak digunakan dalam *fuzzy clustering* adalah *Fuzzy C-Means*. Metode ini mengelompokkan data berdasarkan derajat keanggotaan setiap data terhadap masing-masing *cluster*. Berbeda dengan metode *hard clustering* seperti *K-Means*, *Fuzzy C-Means* memungkinkan satu data memiliki tingkat keanggotaan pada beberapa *cluster* sekaligus, sehingga lebih fleksibel dalam menangani data yang tidak pasti atau memiliki batas antar kelompok yang samar (Kusumadewi & Purnomo, 2004).

Fuzzy C-Means (FCM) merupakan metode *clustering* yang banyak digunakan, namun memiliki beberapa kelemahan, salah satunya adalah sensitivitas terhadap pusat *cluster* yang dapat menyebabkan hasil pengelompokan sulit dikontrol (Siringoringo & Jamaluddin, 2019). Selain itu, dalam beberapa penelitian, diperlukan metode tambahan seperti DBSCAN untuk mendeteksi dan menangani *outlier* sebelum proses *clustering* menggunakan FCM dilakukan (Lubis & Hasibuan, 2024). Hal ini menunjukkan bahwa FCM belum sepenuhnya *robust* terhadap data yang mengandung nilai ekstrem.

Pengaruh *outlier* terhadap hasil *clustering* tersebut tidak terlepas dari penggunaan ukuran jarak dalam proses perhitungan, di mana FCM umumnya menggunakan *Euclidean distance*. Jarak *Euclidean* memiliki sifat berbasis kuadrat yang dapat memperbesar pengaruh nilai ekstrem, sehingga keberadaan *outlier* dapat menggeser pusat *cluster* dan menurunkan kualitas hasil *clustering*. Oleh karena itu, diperlukan pendekatan alternatif dengan menggunakan ukuran jarak yang dapat mengurangi pengaruh *outlier* terhadap proses *clustering*, salah satunya adalah *Manhattan distance*. Berbeda dengan *Euclidean distance*, *Manhattan distance* menggunakan selisih *absolut* sehingga tidak memperbesar pengaruh nilai ekstrem (Strauss & Maltitz, 2017). Sehingga, keberadaan *outlier* tetap digunakan dalam data, tetapi pengaruhnya terhadap perhitungan jarak dan pembentukan cluster menjadi lebih kecil. Dengan demikian, penggunaan *Manhattan distance* diharapkan dapat meningkatkan stabilitas dan kualitas hasil *clustering* pada metode FCM.

Berdasarkan permasalahan tersebut, penelitian ini berfokus pada penerapan dan perbandingan metode FCM dengan menggunakan dua ukuran jarak, yaitu *Euclidean distance* dan *Manhattan distance*. Perbandingan ini dilakukan untuk menganalisis perbedaan hasil *clustering* yang dihasilkan oleh kedua ukuran jarak tersebut, sehingga dapat diketahui metode yang lebih optimal dalam menangani data yang memiliki variasi tinggi atau mengandung *outlier*.

Untuk memastikan hasil modifikasi, penelitian ini menerapkan validasi yang berfungsi mengevaluasi kualitas hasil pengelompokan secara objektif. Salah satu teknik validasi yang sering digunakan pada *fuzzy clustering* adalah *Modified Partition Coefficient* (MPC). MPC merupakan salah satu indeks validasi yang digunakan dalam metode FCM untuk menilai kualitas hasil pengelompokan. Indeks

ini pertama kali diperkenalkan oleh Dave 1996 sebagai pengembangan dari *Partition Coefficient* (PC) yang dikemukakan oleh Bezdek (1974). Nilai MPC berada pada rentang antara 0 hingga 1. Semakin tinggi nilai MPC atau yang mendekati 1, maka semakin jelas batas antar *cluster* yang terbentuk. Sebaliknya, nilai MPC yang rendah atau mendekati 0 menunjukkan bahwa hasil pengelompokan masih bersifat kabur, di mana banyak data memiliki tingkat keanggotaan yang hampir sama pada beberapa *cluster* (Dave, 1996).

Proses modifikasi dan validasi metode yang dilakukan dalam penelitian ini tidak hanya menjadi upaya pengembangan metode, tetapi juga mencerminkan penerapan nilai-nilai Islam yang menegaskan pentingnya kerja keras serta tanggung jawab manusia dalam menemukan solusi yang lebih baik dan bermanfaat. Islam menegaskan bahwa setiap manusia memiliki kewajiban untuk berusaha memperbaiki keadaan, serta berinovasi dalam berbagai aspek kehidupan untuk mendapatkan manfaat bagi diri sendiri maupun orang lain. Prinsip ini ditegaskan oleh Allah SWT dalam firman-Nya pada Surah Al-Insyirah ayat 7:

فَإِذَا فَرَغْتَ فَانصَبْ ۝٧

“Apabila engkau telah selesai (dengan suatu kebajikan) teruslah bekerja keras (untuk kebajikan yang lain)” (QS. Al-Insyirah: 7)

Menurut Tafsir Tahlili terhadap Surah Al-Insyirah ayat 7, Allah SWT memberikan perintah kepada manusia agar senantiasa menunjukkan ketekunan dalam beramal serta tidak berhenti pada satu bentuk usaha saja, melainkan melanjutkannya dengan usaha yang lebih baik, bernilai manfaat, dan membawa kebaikan. Prinsip tersebut selaras dengan penelitian ini yang berfokus pada pengembangan metode yang telah ada sebelumnya. Dalam penelitian ini, dilakukan upaya modifikasi algoritma *Fuzzy C-Means* dengan mengganti Jarak *Euclidean*

menjadi jarak *Manhattan*, serta menerapkan proses validasi untuk memperoleh hasil yang akurat. Penelitian ini mencerminkan nilai-nilai keislaman seperti tanggung jawab dan komitmen untuk menemukan solusi yang lebih baik dan memberikan manfaat bagi pengembangan ilmu pengetahuan (Kementerian Agama, 2022).

Gaya hidup mahasiswa merupakan salah satu aspek penting dalam kehidupan perguruan tinggi yang mencerminkan pola aktivitas, kebiasaan, serta cara mahasiswa mengelola waktu dan kesehariannya. Perubahan sosial, kemajuan teknologi, dan perkembangan media digital telah membawa pengaruh signifikan terhadap pola hidup generasi muda, termasuk mahasiswa. Mereka tidak hanya dituntut untuk berprestasi secara akademik, tetapi juga harus mampu menyeimbangkan berbagai aspek kehidupan seperti kegiatan sosial, kesehatan fisik, waktu istirahat, hingga pengelolaan stres.

Dataset gaya hidup mahasiswa digunakan sebagai studi kasus untuk menerapkan metode FCM yang dimodifikasi ini. Data yang digunakan bersumber dari *platform Kaggle*, yang berisi berbagai variabel numerik terkait kebiasaan, aktivitas, serta pola perilaku mahasiswa, terutama dalam konteks keseharian dan keseimbangan hidup. Data ini dimanfaatkan untuk menguji performa metode dalam pengelompokan dengan karakteristik data yang beragam dan tidak pasti. Sehingga, fokus utama penelitian ini terletak pada pengujian efektivitas algoritma FCM yang telah dimodifikasi.

Urgensi penelitian ini untuk meningkatkan akurasi metode FCM dalam proses pengelompokan data yang kompleks dan mengandung ketidakpastian. FCM merupakan salah satu algoritma *fuzzy clustering* yang banyak digunakan untuk

menangani data yang kompleks dan tidak pasti. Namun, penggunaan jarak *Euclidean* pada algoritma ini memiliki keterbatasan, terutama pada data yang mengandung *outlier*, sehingga dapat memengaruhi kualitas hasil *clustering*. Sehingga, penerapan jarak *Manhattan*, yang dinilai lebih tahan terhadap *outlier* dan mampu menghasilkan pengelompokan yang stabil. Selain itu, penelitian ini digunakan validasi MPC untuk menilai kualitas hasil *clustering* secara objektif. Kombinasi antara *Fuzzy C-Means* berbasis jarak *Manhattan* dan validasi menggunakan MPC diharapkan dapat memberikan kontribusi dalam memperkuat metode analisis data berbasis *fuzzy*, khususnya dalam pengelompokan data yang bersifat kompleks dan tidak pasti.

1.2 Rumusan Masalah

Merujuk pada penjelasan yang telah disampaikan dalam bagian latar belakang, maka perumusan masalah dalam penelitian ini adalah apa pengaruh modifikasi algoritma *Fuzzy C-Means* dengan jarak *Manhattan* terhadap kualitas hasil *clustering*, berdasarkan ukuran validasi menggunakan *Modified Partition Coefficient* (MPC)?

1.3 Tujuan Penelitian

Dengan merujuk pada rumusan masalah yang telah dikemukakan sebelumnya, maka tujuan dari penelitian ini adalah untuk menganalisis pengaruh modifikasi algoritma *Fuzzy C-Means* dengan jarak *Manhattan* terhadap kualitas hasil *clustering* berdasarkan validasi *Modified Partition Coefficient* (MPC).

1.4 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah:

1. Bagi Penulis

Menambah pengetahuan dan wawasan peneliti mengenai implementasi metode *Fuzzy C-Means* dengan modifikasi jarak serta validasi menggunakan *Modified Partition Coefficient* (MPC) dalam analisis data gaya hidup mahasiswa.

2. Bagi Instansi

Memberikan kontribusi bagi instansi dalam menghasilkan sistem pengelompokan yang lebih akurat dan stabil, sehingga dapat mendukung pengambilan keputusan berbasis data yang seimbang dan tidak bias.

3. Bagi Pembaca

Menjadi bahan referensi dalam penelitian berikutnya yang dapat dikembangkan menggunakan *dataset* lain, metode *clustering* lain, atau kombinasi metode validasi yang berbeda untuk memperkuat hasil analisis.

1.5 Batasan Masalah

1. Data yang digunakan terbatas pada data gaya hidup mahasiswa, dengan variabel *Study Hours Per Day*, *Extracurricular Hours Per Day*, *Sleep Hours Per Day*, *Social Hours Per Day*, *Physical Activity Hours Per Day*, dan GPA.
2. Rentang banyak *cluster* yang diuji dibatasi antara $c = 2$ hingga $c = 5$.
3. Pengaruh dalam penelitian ini dibatasi pada hasil evaluasi *clustering* yang diukur menggunakan *Modified Partition Coefficient* (MPC).

1.6 Definisi Istilah

Beberapa istilah yang digunakan dalam penelitian ini:

Clustering : Proses mengelompokkan data ke dalam beberapa kelompok (*cluster*) berdasarkan

tingkat kemiripan atau jarak antar data.

- Fuzzy Clustering* : Metode *clustering* yang mengelompokkan data ke dalam beberapa *cluster* dengan tingkat keanggotaan tertentu, di mana setiap data dapat menjadi anggota dari lebih dari satu *cluster* dengan derajat keanggotaan antara 0 dan 1.
- Centroid* : Titik pusat dari setiap *cluster* yang mewakili rata-rata posisi seluruh data dalam *cluster* tersebut.
- Outlier* : Data yang memiliki nilai yang berbeda jauh dari sebagian besar data lainnya dalam suatu *dataset*.
- Jarak *Manhattan* : Ukuran kedekatan antara dua titik dalam ruang data yang dihitung berdasarkan jumlah nilai mutlak selisih tiap variabel.
- Modified Partition Coefficient* : Indeks validasi yang digunakan untuk mengukur kualitas hasil pengelompokan *fuzzy*.

BAB II

KAJIAN TEORI

2.1 Logika *Fuzzy*

2.1.1 Pengertian Logika *Fuzzy*

Prof. Lotfi A. Zadeh adalah seorang ilmuwan yang mengajar dan melakukan penelitian di Universitas California, untuk pertama kalinya mengembangkan logika *fuzzy* pada tahun 1965. Sepanjang abad ke-20, peranan penting dalam penyelesaian masalah ketidakpastian dipegang oleh teori probabilitas (Kusumadewi dkk., 2006). Namun, melalui pengembangan logika *fuzzy* oleh Prof. Lotfi A. Zadeh, pendekatan ini bukan digunakan untuk menggantikan teori probabilitas melainkan untuk pendekatan alternatif dalam menyelesaikan masalah yang bersifat tidak pasti (Puryanto & Sutikno, 2016).

Logika *fuzzy* merupakan metode matematika yang digunakan untuk mengelola data yang bersifat tidak pasti atau kabur. Secara umum logika *fuzzy* sering dianggap sebagai logika kabur, tetapi sebenarnya logika *fuzzy* berfungsi untuk menggambarkan ketidakjelasan. Menurut Kusumadewi dan Purnomo (2004), terdapat sejumlah faktor yang mendasari penggunaan logika *fuzzy*, antara lain:

1. Logika *fuzzy* memiliki konsep dasar yang mudah dipahami karena landasan matematisnya sederhana.
2. Logika *fuzzy* sangat fleksibel dalam penerapannya.
3. Logika *fuzzy* mampu mentoleransi ketidakakuratan pada data.
4. Logika *fuzzy* dapat merepresentasikan fungsi nonlinear yang kompleks.

5. Logika *fuzzy* memungkinkan penerapan pengalaman pakar secara langsung tanpa proses pelatihan.
6. Logika *fuzzy* dapat diintegrasikan dengan metode kendali konvensional.
7. Logika *fuzzy* berbasis pada bahasa alami sehingga memudahkan untuk diaplikasikan.

Logika *fuzzy* menjadi salah satu metode yang banyak digunakan dalam pemodelan dan pengambilan keputusan. Logika *fuzzy* dapat dijadikan sebagai pendekatan yang efektif dalam pengolahan data yang mengandung ketidakpastian. Teori ini menjadi dasar dari berbagai metode *fuzzy* lainnya, seperti *Fuzzy C-Means* (FCM), yang digunakan dalam penelitian ini untuk mengelompokkan data berdasarkan derajat keanggotaan masing-masing objek terhadap *cluster*.

2.1.2 Himpunan *Fuzzy*

Teori himpunan *fuzzy* dijadikan sebagai dasar dari logika *fuzzy*. Berbeda dengan himpunan tegas (*crisp*) yang hanya mengenal dua kondisi, yaitu anggota dan bukan anggota, himpunan *fuzzy* memungkinkan suatu data memiliki tingkat keanggotaan dalam suatu himpunan (Rahakbauw dkk., 2017). Hal ini membuat himpunan *fuzzy* lebih fleksibel dalam merepresentasikan data yang bersifat tidak pasti atau kabur. Suatu data tidak langsung dikategorikan secara mutlak, tetapi dapat berada dalam suatu himpunan dengan tingkat keanggotaan tertentu. Dengan pendekatan ini, kondisi yang tidak pasti dalam data dapat digambarkan dengan lebih realistis.

Dalam himpunan *fuzzy*, keberadaan suatu data dalam himpunan ditentukan oleh derajat keanggotaan, dan nilai tersebut dianggap sebagai ciri utama dalam

penalaran *fuzzy* (Santosa, 2014). Nilai ini menunjukkan seberapa besar suatu data termasuk dalam suatu himpunan. Dengan demikian, suatu data tidak hanya diklasifikasikan sebagai anggota dan bukan anggota, tetapi dapat memiliki tingkat keanggotaan tertentu. Konsep derajat keanggotaan ini menjadi dasar dalam metode berbasis *fuzzy*, termasuk dalam proses *clustering* seperti *Fuzzy C-Means* (Lubis & Hasibuan, 2024).

2.1.3 Derajat Keanggotaan

Derajat keanggotaan merupakan nilai yang menunjukkan tingkat keanggotaan suatu data dalam himpunan *fuzzy*. Nilai ini berada pada rentang 0 hingga 1, di mana nilai 0 menunjukkan bahwa suatu data tidak termasuk dalam himpunan, sedangkan nilai 1 menunjukkan bahwa data tersebut sepenuhnya menjadi anggota (Santosa, 2014). Sedangkan, nilai di antara 0 dan 1 menunjukkan tingkat keanggotaan parsial sesuai dengan kondisi data. Dengan adanya rentang nilai ini, keanggotaan suatu data tidak bersifat mutlak, melainkan memiliki tingkat tertentu.

Berbeda dengan logika klasik yang hanya mengenal dua kondisi, yaitu anggota dan bukan anggota, logika *fuzzy* memungkinkan suatu data memiliki tingkat keanggotaan yang tidak pasti (Klir & Yuan, 1996). Hal ini membuat suatu data memiliki tingkat keanggotaan pada lebih dari satu himpunan secara bersamaan. Pendekatan ini lebih mampu menggambarkan kondisi nyata yang seringkali tidak bersifat pasti.

2.2 *Clustering*

Clustering merupakan salah satu teknik dalam analisis data yang berfungsi untuk mengelompokkan sejumlah objek ke dalam beberapa kelompok atau *cluster*.

Pengelompokan tersebut berdasarkan tingkat kesamaan karakteristik yang dimilikinya. Dalam proses ini, setiap objek yang berada dalam satu *cluster* memiliki kemiripan yang relatif tinggi satu sama lain dibandingkan dengan objek yang tergolong dalam *cluster* berbeda. Prinsip dari metode *clustering* digambarkan oleh definisi ini sebagai upaya untuk menjaga tingkat homogenitas internal yang tinggi di dalam *cluster* serta heterogenitas eksternal yang jelas antar *cluster* (Rahayu dkk., 2024). *Clustering* merupakan salah satu teknik pemrosesan data yang berfungsi untuk mengidentifikasi dan mengungkap pola-pola tersembunyi dalam suatu kumpulan data (Žalik, 2008). *Clustering* tidak hanya digunakan untuk memahami pola tersembunyi dalam data, tetapi juga memungkinkan digunakan untuk mengidentifikasi struktur alami yang terdapat di dalam suatu kumpulan data.

Selain berfungsi untuk mengelompokkan data berdasarkan kesamaan karakteristik, *clustering* juga dapat dipahami sebagai suatu bentuk untuk menyederhanakan data, di mana sekumpulan data besar disederhanakan menjadi sejumlah kecil *cluster*. Melalui proses ini, informasi yang semula tersebar dalam banyak data dapat dirangkum menjadi struktur yang lebih ringkas tanpa menghilangkan karakteristik utama yang mewakili keseluruhan data. Terdapat beberapa jenis ukuran kesamaan yang dapat diterapkan untuk mengidentifikasi suatu kelas atau kelompok data, tetapi tergantung pada karakteristik serta tujuan aplikasi data yang akan digunakan. Ukuran kesamaan ini berperan penting dalam menentukan bagaimana suatu *cluster* terbentuk, karena untuk melihat kemiripan antar objek di dalam *dataset*. Beberapa nilai dan parameter yang umum digunakan sebagai dasar pengukuran kesamaan antara lain yaitu: jarak, konektivitas, dan intensitas (Sianturi, 2015).

Metode pengelompokan (*clustering*) dapat diklasifikasikan menjadi dua kategori utama, yaitu *hard clustering* dan *soft clustering*. *Hard clustering* digunakan pada setiap data yang hanya dapat menjadi anggota dari satu *cluster* tertentu secara tegas. Salah satu metode yang paling banyak digunakan dalam kategori ini adalah *K-Means*, yang berfungsi dengan membagi himpunan data ke dalam beberapa *cluster* berdasarkan jarak terdekat terhadap pusat *cluster*. Berbeda dengan *hard clustering*, *soft clustering* digunakan pada setiap objek yang memiliki tingkat keanggotaan pada lebih dari satu *cluster* sekaligus, sehingga memungkinkan hasil yang lebih fleksibel terhadap data yang kompleks. Salah satu metode yang paling banyak digunakan dalam kategori ini adalah *Fuzzy C-Means* (FCM) (Johra, 2021).

2.3 *Fuzzy Clustering*

Fuzzy clustering adalah salah satu metode pengelompokan data berdasarkan derajat keanggotaan, di mana setiap objek memiliki tingkat keterhubungan yang berbeda terhadap masing-masing *cluster*. Pendekatan ini menggunakan himpunan *fuzzy* sebagai dasar dalam proses pembobotan. Metode ini juga digunakan untuk menentukan *cluster* yang optimal dalam suatu ruang vektor dengan menggunakan jarak *Euclidean* sebagai dasar perhitungan (Kusumadewi & Purnomo, 2004). Selain berfungsi untuk pengelompokan data, *fuzzy clustering* juga dimanfaatkan dalam pemodelan *fuzzy* untuk mengidentifikasi aturan-aturan *fuzzy* yang dibutuhkan dalam proses pemodelan (Rahakbauw dkk., 2017).

Dalam proses pengelompokan ke dalam *cluster* yang memiliki karakteristik yang sama, dibutuhkan langkah sistematis untuk menjamin hasil pengelompokan yang akurat. Pada metode *clustering* klasik, setiap objek dikelompokkan secara tegas

hanya ke dalam satu partisi tertentu. Namun, dalam kenyataannya, sering ditemukan objek yang posisinya berada di antara dua atau lebih partisi, sehingga menyebabkan kesulitan dalam menentukan pengelompokan secara pasti. Untuk mengatasi permasalahan tersebut, konsep logika *fuzzy* memberikan pendekatan yang lebih efisien melalui metode pengelompokan yang dikenal sebagai *fuzzy clustering*. Metode ini dianggap lebih alami dibandingkan dengan metode *clustering* klasik karena kemampuannya dalam merepresentasikan ketidakpastian posisi objek di antara beberapa *cluster*. Secara umum, sebagian besar algoritma *fuzzy clustering* dikembangkan berdasarkan prinsip optimasi fungsi objektif atau hasil modifikasinya, yang bertujuan untuk meminimalkan kesalahan dalam proses pengelompokan serta memperbaiki akurasi dalam menggambarkan struktur data (Sari & Suranti, 2016).

Dengan demikian, *fuzzy clustering* tidak hanya berperan dalam pengelompokan data secara fleksibel, tetapi juga menjadi dasar penting dalam pengembangan berbagai algoritma berbasis logika *fuzzy* seperti *Fuzzy C-Means* (FCM). Pendekatan ini mampu meningkatkan keakuratan dalam pemodelan dan analisis data yang mengandung unsur ketidakpastian.

2.4 Fuzzy C-Means

Fuzzy C-Means ditemukan oleh Bezdek pada tahun 1981. Metode *Fuzzy C-Means* adalah teknik pengklasteran data di mana pada setiap titik data dikelompokkan dalam suatu *cluster* dengan derajat keanggotaan tertentu, sehingga memungkinkan fleksibilitas yang lebih tinggi dalam analisis data. Di mana derajat keanggotaannya dalam rentang 0 hingga 1 (Christian & Hakim, 2019). Ini berarti seberapa besar suatu data dapat dikatakan milik suatu kelompok, ditentukan oleh

derajat keanggotaannya. Proses ini dimulai dengan menemukan pusat *cluster*, yang menunjukkan lokasi rata-rata untuk setiap *cluster*. Namun, pada awalnya, posisi pusat ini mungkin tidak tepat, dan setiap data memiliki tingkat keanggotaan untuk setiap *cluster*. Dengan melakukan perbaikan terus-menerus pada pusat *cluster* dan nilai keanggotaan data, pusat *cluster* akan secara perlahan bergerak menuju lokasi yang lebih tepat.

Dalam FCM, setiap titik data memiliki derajat keanggotaan yang berbeda untuk setiap *cluster* yang ada (Rahakbauw dkk., 2017). Metode ini menggunakan ukuran kekaburan dan menghitung jarak antar titik dengan menggunakan jarak *Euclidean* yang jelas (tegas), sehingga data bisa menjadi bagian dari semua *cluster* dengan tingkat keanggotaan yang berbeda. Oleh karena itu, FCM sangat berguna untuk mengelompokkan data yang rumit di mana batas antar kelompok tidak selalu jelas. Ketika pusat *cluster* diperbaiki dan nilai keanggotaan untuk setiap titik data diperbarui secara berulang, proses ini dapat menghasilkan posisi pusat *cluster* yang lebih akurat sesuai dengan data yang ada.

Hasil akhir dari FCM adalah daftar pusat *cluster* dan derajat keanggotaan untuk setiap titik data. Informasi ini bisa digunakan untuk membuat sistem inferensi *fuzzy* yang lebih baik, karena memberikan gambaran yang lebih jelas tentang bagaimana data saling berinteraksi dalam setiap *cluster*. FCM memberikan lebih banyak fleksibilitas dibandingkan dengan metode pengelompokan tradisional, karena data tidak hanya terikat pada satu *cluster*, tetapi bisa masuk ke beberapa *cluster* sesuai dengan derajat keanggotaannya. Dengan demikian, FCM menjadi solusi yang efektif untuk pengelompokan data, terutama ketika karakteristik data sangat beragam dan rumit.

Algoritma dari FCM adalah sebagai berikut (Kusumadewi & Purnomo, 2004):

1. Menentukan dan *input* data yang digunakan dalam algoritma FCM disusun dalam bentuk matriks data yang dinotasikan dengan X , yaitu berupa matriks berukuran $n \times m$ di mana n menyatakan banyak data dan m menyatakan banyak variabel setiap data. Entri $x_{i,j}$ menunjukkan nilai data sampel ke- i ($i = 1, 2, \dots, n$), variabel ke- j ($j = 1, 2, \dots, m$). Dengan menentukan parameter awal sebagai berikut:
 - a. Banyak *cluster* (c)
 - b. Pangkat pembobot (w)
 - c. Maksimum iterasi ($MaxIter$)
 - d. *Error* terkecil yang diharapkan (ε)
 - e. Fungsi objektif awal $P_0 = 0$
 - f. Iterasi awal $t = 1$

Sebagai contoh proses perhitungan, misalkan terdapat data yang disusun dalam bentuk matriks X yang berukuran 4×3 . Secara umum, bentuk matriks data X dapat dituliskan sebagai berikut:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ x_{3,1} & x_{3,2} & x_{3,3} \\ x_{4,1} & x_{4,2} & x_{4,3} \end{bmatrix}$$

Selanjutnya, sebagai contoh digunakan nilai data matriks X sebagai berikut:

$$X = \begin{bmatrix} 2 & 5 & 4 \\ 3 & 4 & 5 \\ 6 & 7 & 6 \\ 8 & 6 & 7 \end{bmatrix}$$

kemudian, parameter yang digunakan pada tahap *clustering* ditentukan sebagai berikut:

- a. Banyak *cluster* (c) = 2
 - b. Pangkat pembobot (w) = 2
 - c. Maksimum iterasi = 100
 - d. *Error* terkecil yang diharapkan (ε) = 0,001
 - e. Fungsi objektif awal $P_0 = 0$
 - f. Iterasi awal $t = 1$
2. Membangkitkan bilangan *random* $\mu_{i,k}$, $i = 1, 2, \dots, n$; $k = 1, 2, \dots, c$; sebagai entri matriks partisi awal U . Untuk setiap baris akan memiliki jumlah nilai entri baris yang sama yaitu 1. Secara umum, matriks partisi U dapat dituliskan sebagai berikut:

$$U = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \dots & \mu_{1,c} \\ \mu_{2,1} & \mu_{2,2} & \dots & \mu_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n,1} & \mu_{n,2} & \dots & \mu_{n,c} \end{bmatrix}$$

Pada tahap awal, nilai derajat keanggotaan belum diketahui. Sehingga, dibangkitkan nilai awal sebelum normalisasi yang disusun dalam matriks $U^{(0)}$. Entri dari matriks $U^{(0)}$ dinotasikan dengan $\mu_{i,k}^{(0)}$. Secara umum, matriks nilai awal sebelum normalisasi dapat dituliskan sebagai berikut:

$$U^{(0)} = \begin{bmatrix} \mu_{1,1}^{(0)} & \mu_{1,2}^{(0)} & \dots & \mu_{1,c}^{(0)} \\ \mu_{2,1}^{(0)} & \mu_{2,2}^{(0)} & \dots & \mu_{2,c}^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n,1}^{(0)} & \mu_{n,2}^{(0)} & \dots & \mu_{n,c}^{(0)} \end{bmatrix}$$

Hitung jumlah setiap baris menggunakan rumus:

$$Q_i = \sum_{k=1}^c \mu_{i,k}^{(0)} \text{ dengan } i = 1, 2, \dots, n \quad (2.1)$$

Hitung:

$$\mu_{i,k} = \frac{\mu_{i,k}^{(0)}}{Q_i} \quad (2.2)$$

di mana:

Q_i : Jumlah derajat keanggotaan pada data ke- i

$\mu_{i,k}$: Derajat keanggotaan data ke- i pada *cluster* ke- k setelah normalisasi

$\mu_{i,k}^{(0)}$: Derajat keanggotaan data ke- i pada *cluster* ke- k sebelum normalisasi

Pada contoh proses perhitungan ini digunakan $c = 2$, sehingga matriks nilai awal sebelum normalisasi $U^{(0)}$ berordo 4×2 . Misalkan matriks nilai awal sebelum normalisasi yang dibangkitkan adalah:

$$U^{(0)} = \begin{bmatrix} 3 & 2 \\ 4 & 1 \\ 2 & 5 \\ 6 & 3 \end{bmatrix}$$

Matriks $U^{(0)}$ tidak langsung digunakan sebagai matriks partisi, karena hasil penjumlahan entri pada setiap baris belum sama dengan 1, sehingga setiap baris perlu dinormalisasi terlebih dahulu. Jumlah nilai pada setiap baris dihitung sesuai dengan Persamaan (2.1) maka diperoleh:

$$Q_1 = 3 + 2 = 5$$

$$Q_2 = 4 + 1 = 5$$

$$Q_3 = 2 + 5 = 7$$

$$Q_4 = 6 + 3 = 9$$

karena nilai derajat keanggotaan yang dibangkitkan secara acak belum memenuhi syarat, yaitu hasil penjumlahan entri-entri pada setiap baris belum sama dengan 1, maka dilakukan normalisasi terhadap setiap entri matriks $U^{(0)}$ sesuai dengan Persamaan (2.2) dengan membagi setiap entri pada baris ke- i dengan jumlah total baris. Tujuan dari proses ini adalah untuk menghasilkan matriks partisi U dengan entri derajat keanggotaan yang baru $(\mu_{i,k})$. Maka diperoleh:

$$U = \begin{bmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{4}{5} & \frac{1}{5} \\ \frac{5}{7} & \frac{2}{7} \\ \frac{2}{9} & \frac{3}{9} \\ \frac{6}{9} & \frac{3}{9} \end{bmatrix} = \begin{bmatrix} 0,60 & 0,40 \\ 0,80 & 0,20 \\ 0,29 & 0,71 \\ 0,67 & 0,33 \end{bmatrix}$$

Setiap baris pada matriks partisi U telah memenuhi syarat, yaitu hasil penjumlahan derajat keanggotaan sama dengan 1. Dengan demikian, matriks partisi U sudah memenuhi syarat sebagai matriks partisi dan dapat digunakan pada proses perhitungan pusat *cluster*.

3. Hitung pusat *cluster* ke- k , $V_{k,j}$ dengan $k = 1, 2, \dots, c$ dan $j = 1, 2, \dots, m$:

$$V_{k,j} = \frac{\sum_{i=1}^n ((\mu_{i,k})^w \cdot x_{i,j})}{\sum_{i=1}^n (\mu_{i,k})^w} \quad (2.3)$$

di mana:

$V_{k,j}$: Pusat *cluster* ke- k pada variabel ke- j

$\mu_{i,k}$: Derajat keanggotaan data ke- i pada *cluster* ke- k

$x_{i,j}$: Data sampel ke- i dan variabel ke- j

w : Pangkat pembobot

n : Banyak data

Hasil perhitungan $V_{k,j}$, pusat *cluster* dapat ditulis dalam bentuk vektor. Pusat *cluster* ke- k dinotasikan dengan V_k , yaitu vektor yang berisi seluruh nilai pusat *cluster* ke- k pada setiap variabel. Secara umum dapat dituliskan sebagai berikut:

$$V_k = (V_{k,1}, V_{k,2}, \dots, V_{k,m})$$

Dengan demikian, V_k bukan satu nilai tunggal, melainkan vektor pusat *cluster* ke- k . Misalkan, jika terdapat 3 variabel, maka pusat *cluster* pertama dapat dituliskan sebagai berikut:

$$V_1 = (V_{1,1}, V_{1,2}, V_{1,3})$$

Artinya, V_1 adalah vektor pusat *cluster* pertama, yang terdiri dari pusat *cluster* pertama pada variabel pertama, variabel kedua, dan variabel ketiga.

Pada contoh proses perhitungan pusat *cluster* dengan menggunakan $w = 2$, sehingga setiap entri derajat keanggotaan $\mu_{i,k}$ dipangkatkan 2. Sehingga, diperoleh:

$$[(\mu_{i,k})^2] = \begin{bmatrix} 0,36 & 0,16 \\ 0,64 & 0,04 \\ 0,0841 & 0,5041 \\ 0,4489 & 0,1089 \end{bmatrix}$$

Matriks ini adalah hasil pemangkatan setiap entri dari matriks partisi U .

Untuk menghitung pusat *cluster* 1, variabel ke-1 ($j = 1$)

$$V_{1,1} = \frac{(0,36 \times 2) + (0,64 \times 3) + (0,0841 \times 6) + (0,4489 \times 8)}{0,36 + 0,64 + 0,0841 + 0,4489}$$

$$V_{1,1} = \frac{0,72 + 1,92 + 0,5046 + 3,5912}{1,533} = \frac{6,7358}{1,533} \approx 4,3939$$

Untuk menghitung pusat *cluster* 1, variabel ke-2 ($j = 2$)

$$V_{1,2} = \frac{(0,36 \times 5) + (0,64 \times 4) + (0,0841 \times 7) + (0,44489 \times 6)}{0,36 + 0,64 + 0,0841 + 0,4489}$$

$$V_{1,2} = \frac{1,80 + 2,56 + 0,5887 + 2,6934}{1,533} = \frac{7,6421}{1,533} \approx 4,9851$$

Untuk menghitung pusat *cluster* 1, variabel ke-3 ($j = 3$)

$$V_{1,2} = \frac{(0,36 \times 4) + (0,64 \times 5) + (0,0841 \times 6) + (0,44489 \times 7)}{0,36 + 0,64 + 0,0841 + 0,4489}$$

$$V_{1,3} = \frac{1,44 + 3,20 + 0,5046 + 3,1423}{1,533} = \frac{8,2869}{1,533} \approx 5,4057$$

sehingga pusat *cluster* pertama dapat dituliskan dalam bentuk vektor V_1 , yaitu:

$$V_1 = (4,3939, 4,9851, 5,4057)$$

Untuk menghitung *cluster* 2 dilakukan dengan menggunakan metode yang serupa yang digunakan pada *cluster* 1.

4. Hitung fungsi objektif pada iterasi ke- t , P_t :

$$P_t = \sum_{i=1}^n \sum_{k=1}^c \left(\left[\sum_{j=1}^m (x_{i,j} - V_{k,j})^2 \right] (\mu_{i,k})^w \right) \quad (2.4)$$

di mana:

P_t : Fungsi objektif pada iterasi ke- t

$x_{i,j}$: Data sampel ke- i dan variabel ke- j

$V_{k,j}$: Pusat *cluster* ke- k pada variabel ke- j

$\mu_{i,k}$: Derajat keanggotaan data ke- i pada *cluster* ke- k

w : Pangkat pembobot

c : Banyak *cluster*

n : Banyak data

m : Banyak variabel

Sebelum melakukan perhitungan contoh fungsi objektif, terlebih dahulu dihitung jarak antara setiap data dengan masing-masing pusat *cluster*. Jarak kuadrat antara data ke- i dan pusat *cluster* ke- k dinotasikan dengan $d_{i,k}^2$. Notasi ini digunakan karena jarak dihitung dari data ke- i terhadap *cluster* ke- k , sedangkan indeks j digunakan untuk menunjukkan variabel. Jarak kuadrat antara data ke- i dan pusat *cluster* ke- k dihitung dengan rumus:

$$d_{i,k}^2 = \sum_{j=1}^m (x_{i,j} - V_{k,j})^2 \quad (2.5)$$

di mana:

$d_{i,k}^2$: Jarak kuadrat antara data ke- i dan pusat *cluster* ke- k

$x_{i,j}$: Data sampel ke- i dan variabel ke- j

$V_{k,j}$: Pusat *cluster* ke- k pada variabel ke- j

Contoh proses perhitungan untuk data pertama adalah:

$$x_{1,1} = (2,5,4)$$

Untuk menghitung *cluster* 1

jarak antara data pertama dan pusat cluster ke-1 ditulis sebagai $d_{1,1}^2$ dan dihitung sebagai berikut:

$$d_{1,1}^2 = (2 - 4,3939)^2 + (5 - 4,9851)^2 + (4 - 5,4057)^2$$

$$d_{1,1}^2 = (-2,3939)^2 + (0,0149)^2 + (-1,4057)^2$$

$$d_{1,1}^2 = 5,7308 + 0,0002 + 1,9758$$

$$d_{1,1}^2 = 7,7068$$

kemudian nilai tersebut dikalikan dengan derajat keanggotaan yang telah dipangkatkan:

$$7,7068 \times 0,36 = 2,7744$$

Untuk menghitung *cluster* 2

jarak antara data pertama dan pusat *cluster* ke-2 ditulis sebagai $d_{1,2}^2$ dan dihitung sebagai berikut:

$$d_{1,2}^2 = (2 - 5,3331)^2 + (5 - 6,3248)^2 + (4 - 5,6911)^2$$

$$d_{1,2}^2 = (-3,3331)^2 + (-1,3248)^2 + (-1,6911)^2$$

$$d_{1,2}^2 = 11,1095 + 1,7551 + 2,8601$$

$$d_{1,2}^2 = 15,7247$$

kemudian dikalikan dengan derajat keanggotaan:

$$15,7274 \times 0,16 = 2,5159$$

Sehingga nilai fungsi objektif untuk data pertama adalah:

$$2,7744 + 2,5159 = 5,2903$$

Perhitungan yang sama dilakukan untuk data kedua hingga data kelima dengan menggunakan langkah yang sama dengan data pertama. Kemudian diperoleh nilai fungsi objektif untuk seluruh data yang kemudian dijumlahkan untuk mendapatkan nilai fungsi objektif total.

5. Hitung perubahan matriks partisi untuk masing-masing elemen:

$$\mu_{i,k} = \frac{\left[\sum_{j=1}^m (x_{i,j} - V_{k,j})^2 \right]^{\frac{-1}{w-1}}}{\sum_{k=1}^c \left[\sum_{j=1}^m (x_{i,j} - V_{k,j})^2 \right]^{\frac{-1}{w-1}}} \quad (2.6)$$

dengan $i = 1, 2, \dots, n$; dan $k = 1, 2, \dots, c$

di mana:

$\mu_{i,k}$: Derajat keanggotaan data ke- i pada *cluster* ke- k

$x_{i,j}$: Data sampel ke- i dan variabel ke- j

$V_{k,j}$: Pusat *cluster* ke- k pada variabel ke- j

w : Pangkat pembobot

m : Banyak data

c : Banyak *cluster*

Contoh proses perhitungan untuk data pertama adalah:

$$X_1 = (2,5,4)$$

Karena pangkat pembobot yang digunakan adalah:

$$w = 2$$

maka:

$$-\frac{1}{w-1} = -\frac{1}{2-1} = -1$$

sehingga untuk mempermudah perhitungan rumus pembaruan derajat

keanggotaan untuk data pertama dapat ditulis sebagai berikut:

$$\mu_{1,1} = \frac{\frac{1}{d_{1,1}^2}}{\frac{1}{d_{1,1}^2} + \frac{1}{d_{1,2}^2}}$$

$$\mu_{1,2} = \frac{\frac{1}{d_{1,2}^2}}{\frac{1}{d_{1,1}^2} + \frac{1}{d_{1,2}^2}}$$

Hitung jarak ke *cluster* 1

jarak antara data pertama dan pusat cluster ke-1 ditulis sebagai $d_{1,1}^2$ dan

dihitung sebagai berikut:

$$d_{1,1}^2 = (2 - 4,3939)^2 + (5 - 4,9851)^2 + (4 - 5,4057)^2$$

$$d_{1,1}^2 = (-2,3939)^2 + (0,0149)^2 + (-1,4057)^2$$

$$d_{1,1}^2 = 5,7308 + 0,0002 + 1,9758$$

$$d_{1,1}^2 = 7,7068$$

Hitung jarak ke *cluster* 2

jarak antara data pertama dan pusat *cluster* ke-2 ditulis sebagai $d_{1,2}^2$ dan dihitung sebagai berikut:

$$d_{1,2}^2 = (2 - 5,3331)^2 + (5 - 6,3248)^2 + (4 - 5,6911)^2$$

$$d_{1,2}^2 = (-3,3331)^2 + (-1,3248)^2 + (-1,6911)^2$$

$$d_{1,2}^2 = 11,1095 + 1,7551 + 2,8601$$

$$d_{1,2}^2 = 15,7247$$

Menghitung derajat keanggotaan data pertama terhadap *cluster* 1

$$\mu_{1,1} = \frac{\frac{1}{7,7068}}{\left(\frac{1}{7,7068}\right) + \left(\frac{1}{15,7247}\right)}$$

$$\mu_{1,1} = \frac{0,1298}{0,1298 + 0,0636} = 0,6711$$

Menghitung derajat keanggotaan data pertama terhadap *cluster* 2

$$\mu_{1,2} = \frac{\frac{1}{15,7247}}{\left(\frac{1}{7,7068}\right) + \left(\frac{1}{15,7247}\right)}$$

$$\mu_{1,2} = \frac{0,0636}{0,1298 + 0,0636} = 0,3289$$

Sehingga, hasil pembaruan derajat keanggotaan untuk data pertama terhadap masing-masing *cluster* adalah:

$$(\mu_{1,1}, \mu_{1,2}) = (0,6711, 0,3289)$$

Perhitungan yang sama dilakukan untuk data kedua hingga data kelima dengan menggunakan langkah yang sama dengan data pertama.

6. Memeriksa kondisi berhenti:

- a. Jika: $(|P_t - P_{t-1}| < \varepsilon)$ atau $(t > MaxIter)$ maka berhenti;

b. Jika tidak: $t = t + 1$, ulangi langkah ke-4

Misalkan dari perhitungan diperoleh nilai fungsi objektif iterasi ke-1 $P_t = 17,2170$ dan batas toleransinya ($\varepsilon = 0,001$). Untuk menghitung selisihnya, yaitu:

$$|P_1 - P_0| = |17,2170 - 0| = 17,2170$$

maka,

$$17,2170 > 0,001$$

Karena selisih nilai fungsi objektif masih lebih besar dari batas toleransi dan banyaknya iterasi belum mencapai maksimum, maka proses iterasi dilanjutkan ke iterasi berikutnya dengan memperbarui nilai $t = t + 1$.

2.5 Normalisasi *Min-Max*

Menurut Rahayu dkk. (2024), normalisasi termasuk dalam tahapan *preprocessing* data yang berfungsi meningkatkan kualitas data dan mendukung kinerja algoritma analisis data agar lebih akurat. Normalisasi data merupakan suatu proses transformasi nilai-nilai variabel ke dalam rentang tertentu agar variabel tersebut memiliki skala yang sama. Normalisasi dilakukan agar variabel dengan nilai yang besar tidak mendominasi hasil analisis. Salah satu metode normalisasi yang sering digunakan adalah normalisasi *Min-Max*. Teknik ini bekerja dengan cara melakukan transformasi terhadap data asli sehingga seluruh nilai variabel berada pada rentang tertentu, umumnya antara 0 hingga 1. Rumus yang digunakan dalam metode ini dapat dituliskan sebagai berikut:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.7)$$

di mana:

x' : Hasil Normalisasi

x : Nilai asli pada data

x_{min} : Nilai minimum dari seluruh data

x_{max} : Nilai maksimum dari seluruh data

Normalisasi *Min-Max* berperan penting karena jika data belum dinormalisasi, variabel yang memiliki skala lebih besar akan memberikan pengaruh lebih kuat dalam perhitungan, sehingga hasilnya menjadi bias. Sehingga, normalisasi diperlukan untuk memastikan bahwa setiap variabel memiliki nilai yang seimbang dalam proses analisis data.

2.6 Jarak *Manhattan*

Dalam proses analisis data, metrik kesamaan dan jarak memiliki peran penting dalam menentukan tingkat kemiripan atau kedekatan dua objek yang dibandingkan. Metrik ini digunakan dalam berbagai algoritma, terutama dalam *clustering* dan sistem rekomendasi yang membutuhkan pengukuran antar data. Salah satunya jenis metrik jarak yang umum digunakan adalah Jarak *Manhattan*, yang juga dikenal dengan sebutan *City Blok Distance*.

Jarak *manhattan* atau *City Blok Distance* merupakan salah satu ukuran jarak yang digunakan dalam analisis data atau metode *clustering* untuk mengukur jarak antara dua titik ketika pergerakannya hanya dilakukan secara horizontal atau vertikal, atau seperti bergerak di sekitar blok kota (Rahayu dkk., 2024). Istilah “*Manhattan*” sendiri di perkenalkan karena mirip dengan jarak yang ditempuh di jalan perkotaan berbentuk grid, seperti Kota *Manhattan* di *New York*. Sehingga jarak *Manhattan* menggambarkan total perbedaan mutlak antar koordinat dua titik data berdasarkan arah pergerakannya (Kaufman & Rousseeuw, 1989).

Jarak *Manhattan* dikenal sebagai *City Blok Distance* atau *L1 Distance* yang merupakan salah satu ukuran jarak yang banyak digunakan dalam analisis data. Ukuran jarak ini merupakan bentuk khusus dari Minkowski distance dengan parameter $p = 1$ (Kaufman & Rousseeuw, 1989). Secara sistematis, jarak *Manhattan* adalah dua vektor antara x dan y dapat didefinisikan sebagai jumlah nilai mutlak dari selisih masing-masing komponennya (Nishom, 2019):

$$d(x_i, V_k) = \sum_{j=1}^n |x_{i,j} - V_{k,j}| \quad (2.8)$$

di mana:

$d(x_i, V_k)$: Jarak *Manhattan* antara data ke- i dan pusat *cluster* ke- k

x_i : Vektor data ke- i dengan elemen x_{ij} untuk variabel ke- j

V_k : Vektor pusat *cluster* ke- k dengan elemen $V_{k,j}$

n : Banyak data

$x_{i,j}$: Data sampel ke- i dan variabel ke- j

$V_{k,j}$: Nilai pusat *cluster* ke- k pada variabel ke- j

Berbeda dengan jarak *Euclidean* (L2 norm) yang menghitung jarak dengan mengkuadratkan selisih antar komponen, *Manhattan distance* (L1 norm) hanya menjumlahkan selisih nilai mutlak (Kaufman & Rousseeuw, 1990). Perbedaan ini berdampak pada sensitivitas terhadap nilai pencilan atau *outlier*. Jarak *Euclidean* cenderung lebih sensitif terhadap *outlier* karena kuadrat dari selisih memperbesar deviasi besar, sedangkan jarak *Manhattan* lebih tahan karena selisih dihitung apa adanya tanpa dikuadratkan (Suheri dkk., 2021).

Keberadaan *outlier* dalam data dapat memengaruhi bentuk, ukuran, dan arah *cluster*, sehingga pemilihan ukuran jarak yang tepat menjadi sangat penting dalam

metode *clustering*. Sehingga, jarak *Manhattan* sering dipilih karena lebih stabil dalam menghadapi data yang mengandung nilai pencilan dibandingkan jarak *Euclidean* (Suheri dkk., 2021).

Berdasarkan pertimbangan tersebut, penggunaan jarak *Manhattan* dalam penelitian ini diharapkan dapat memberikan hasil *clustering* yang lebih akurat, terutama pada data gaya hidup mahasiswa yang berpotensi memiliki variasi nilai pencilan.

2.7 Outlier

Outlier merupakan data yang memiliki karakteristik atau nilai yang menyimpang secara signifikan dari sebagian besar data lainnya dalam suatu himpunan, sehingga berpotensi memengaruhi hasil analisis serta menurunkan akurasi. Keberadaan *outlier* biasanya muncul akibat beberapa faktor, seperti kesalahan dalam proses pengukuran, atau kesalahan input data. Untuk mengatasi permasalahan ini, dapat mendeteksi dan menghapus data yang menyimpang, atau melakukan normalisasi data untuk meminimalkan pengaruhnya terhadap analisis data. Terdapat beberapa teknik umum yang digunakan untuk mendeteksi *outlier*, seperti pendekatan statistik *Interquartile Range* (IQR) dan *z-score*, yang mampu mengidentifikasi nilai-nilai pencilan berdasarkan distribusi data (Rahayu dkk., 2024).

2.8 Validasi Cluster

2.8.1 Modified Partition Coefficient (MPC)

Dalam analisis *clustering* salah satu tahapan yang penting adalah validasi *cluster*, karena berfungsi untuk menilai sejauh mana kualitas pembentukan *cluster* benar-benar mencerminkan pola yang ada dalam data. Tahap ini tidak

hanya berperan dalam mengevaluasi kualitas hasil pengelompokan data, tetapi juga membantu dalam menentukan banyak *cluster* yang paling sesuai dengan karakteristik data yang dianalisis. Dengan melakukan validasi, dapat mengidentifikasi tingkat keseragaman objek-objek yang berada dalam satu kelompok. Selain itu, proses validasi berperan dalam mengevaluasi tingkat homogenitas data dalam satu *cluster* serta tingkat heterogenitas antar *cluster*, sehingga nilai pengelompokan dapat diinterpretasikan secara lebih bermakna secara statistik maupun kontekstual (Ikotun dkk., 2025).

Dalam kajian literatur, ukuran validasi telah berkembang, diantaranya *Silhouette Index* (SI), *Dunn Index*, *Davies-Bouldin Index*, serta *Partition Coefficient* (PC). Keempat ukuran ini tidak bersifat universal, melainkan memiliki relevansi dan keunggulan yang berbeda sesuai dengan pendekatan *clustering* yang diterapkan. Sebagai ilustrasi, *Silhouette Index* yang digunakan dalam metode *hard clustering* seperti *K-Means*, karena indikator ini menilai tingkat kedekatan suatu objek terhadap *cluster* asalnya dibandingkan dengan jarak ke *cluster* lain. Dengan demikian, SI mampu memberikan gambaran mengenai seberapa baik data terklasifikasi secara tegas ke dalam kelompok tertentu. Berbeda halnya dengan pendekatan *fuzzy clustering*, misalnya *Fuzzy C-Means* (FCM), di mana sifat keanggotaan data tidak bersifat absolut, melainkan berupa derajat keanggotaan. Pada penelitian ini, ukuran validasi yang akan digunakan adalah *Modified Partition Coefficient* (PC).

Modified Partition Coefficient (MPC) merupakan salah satu ukuran validasi yang digunakan dalam metode *Fuzzy C-Means* (FCM) untuk menilai sejauh mana hasil pengelompokan (*clustering*) dapat dikatakan berkualitas.

Indeks ini diperkenalkan oleh Dave (1996) sebagai bentuk penyempurnaan dari *Partition Coefficient* (PC) yang sebelumnya dikemukakan oleh Bezdek (1974). Pengembangan ini bertujuan untuk mengatasi kelemahan PC yang cenderung sensitif terhadap banyak *cluster*, sehingga hasil evaluasi terkadang tidak konsisten ketika banyak *cluster* berubah.

Menurut Dave (1996), *Modified Partition Coefficient* (MPC) merupakan sebuah indeks yang digunakan untuk menilai tingkat kejelasan dari hasil pengelompokan *fuzzy*. Indeks ini dihitung dengan mempertimbangkan derajat keanggotaan setiap data terhadap masing-masing *cluster* yang terbentuk dalam proses FCM. Nilai MPC berada pada rentang 0 hingga 1. Semakin tinggi nilai MPC atau mendekati 1, maka hasil pengelompokan memiliki batas yang lebih tegas antara satu *cluster* dengan *cluster* lainnya. Sebaliknya, nilai MPC yang rendah atau mendekati 0 menunjukkan bahwa hasil pengelompokan masih *fuzzy* (kabur) atau tidak tegas, di mana banyak data memiliki tingkat keanggotaan yang hampir sama terhadap beberapa *cluster*. MPC dijadikan sebagai kriteria untuk menentukan banyak *cluster* yang paling optimal dalam analisis data *fuzzy*. Banyak *cluster* terbaik ditentukan dari nilai MPC tertinggi.

Rumus *Modified Partition Coefficient* (MPC) dapat dinyatakan sebagai berikut (Dave, 1996):

$$MPC(c) = 1 - \frac{c}{c-1} (1 - PC) \quad (2.9)$$

di mana:

c : Banyak *cluster*

PC : Nilai *Partition Coefficient*

dengan persamaan untuk menghitung nilai indeks *Partition Coefficient* (PC) dapat dinyatakan dengan rumus berikut (Ramadhan dkk., 2017):

$$PC = \frac{1}{N} \left(\sum_{i=1}^N \sum_{j=1}^c (\mu_{i,j})^2 \right) \quad (2.10)$$

di mana:

N : Banyak data

c : Banyak *cluster*

$\mu_{i,k}$: Derajat keanggotaan data ke- i pada *cluster* ke- k

Menurut Suleman (2015), *Modified Partition Coefficient* (MPC) secara geometris dapat dipahami sebagai jarak *Euclidean* rata-rata antara derajat keanggotaan setiap data terhadap pusat partisi *fuzzy*. Perhitungan MPC tidak bergantung pada banyak *cluster*, sehingga tetap memberikan evaluasi yang objektif ketika banyak *cluster* berubah. Sehingga indeks ini banyak digunakan dalam penerapan metode *Fuzzy C-Means* (FCM) untuk menilai sejauh mana struktur pengelompokan yang dihasilkan memiliki kejelasan dan ketepatan yang optimal.

2.8.2 Purity

Purity merupakan salah satu ukuran evaluasi hasil *clustering* yang digunakan untuk melihat tingkat kesesuaian antara hasil *cluster* dengan label aktual yang terdapat pada data. Berbeda dengan metode validasi internal yang hanya menilai kualitas *cluster* berdasarkan struktur data dan derajat keanggotaan, *Purity* termasuk ke dalam evaluasi eksternal karena membutuhkan label aktual. Oleh karena itu, *Purity* hanya dapat digunakan apabila *dataset* memiliki label

yang asli yang dapat dijadikan sebagai pembanding (Sikhakolli & Sikhakolli, 2023).

Dalam *clustering*, *Purity* digunakan untuk mengetahui tingkat kemurnian suatu *cluster*. Suatu *cluster* dapat dikatakan memiliki *Purity* yang baik jika sebagian besar anggotanya berasal dari satu label aktual yang sama. Sebaliknya, jika satu *cluster* berisi campuran label aktual yang cukup merata, maka nilai *Purity* akan lebih rendah. Nilai *Purity* berada pada rentang 0 hingga 1. Nilai yang semakin mendekati 1 menunjukkan bahwa hasil *clustering* semakin sesuai dengan label aktual, sedangkan nilai yang lebih rendah menunjukkan bahwa anggota dalam *cluster* masih berasal dari label aktual yang beragam (Sikhakolli & Sikhakolli, 2023).

Purity dihitung dengan melihat label aktual yang paling banyak muncul pada setiap *cluster*. Label dengan banyak data terbesar dalam suatu *cluster* disebut sebagai label mayoritas. Setelah label mayoritas pada setiap *cluster* diperoleh, banyak label mayoritas dari seluruh *cluster* dijumlahkan, kemudian dibagi dengan banyak seluruh data. Prinsip ini sesuai dengan rumus *Purity*, yaitu mengambil nilai maksimum dari setiap baris *cluster* pada tabulasi silang, menjumlahkannya, lalu membaginya dengan total data (Sikhakolli & Sikhakolli, 2023). Rumus *Purity* yang digunakan dalam penelitian ini adalah sebagai berikut:

$$Purity = \frac{1}{N} \sum_{k=1}^c \max_{1 \leq j \leq r} n_{k,j} \quad (2.11)$$

di mana:

N : Banyak seluruh data

c : Banyak *cluster*

r : Banyak label aktual

$n_{k,j}$: Banyak data pada *cluster* ke- k yang memiliki label aktual ke- j

$\max_{1 \leq j \leq r} n_{k,j}$: Banyak label aktual terbanyak pada *cluster* ke- k

Langkah perhitungan *Purity* dalam penelitian ini dilakukan dengan membuat tabulasi silang antara hasil *cluster* dan label aktual *Stress Level*. Tabulasi silang ini digunakan untuk mengetahui banyak data dengan label *Low*, *Moderate*, dan *High* pada masing-masing *cluster*. Kemudian dipilih label aktual yang terbesar pada setiap *cluster* sebagai label mayoritas. Banyak label mayoritas dari seluruh *cluster* kemudian dibagi dengan banyak data untuk memperoleh nilai *Purity*.

2.9 Gaya Hidup Mahasiswa

Gaya hidup (*lifestyle*) merupakan salah satu aspek dari kebutuhan sekunder manusia yang bersifat dinamis dan dapat berubah sesuai dengan perkembangan zaman serta keinginan individu untuk menyesuaikan diri terhadap perubahan tersebut. Dengan demikian, gaya hidup dapat dipandang sebagai urusan pribadi yang mencerminkan pilihan dan kecenderungan seseorang dalam menjalani kehidupan sehari-hari. Namun, meskipun bersifat personal, penerapan gaya hidup tetap perlu mempertimbangkan norma dan nilai sosial yang berlaku di masyarakat agar tidak menimbulkan dampak negatif bagi lingkungan sosialnya. Gaya hidup seseorang dapat tercermin melalui berbagai aspek perilaku, seperti cara berpakaian, kebiasaan, pola konsumsi, aktivitas sosial, dan cara berinteraksi. Penilaian gaya hidup tergantung dari cara seseorang memandang atau relatif, karena apa yang dianggap sebagai gaya hidup ideal bagi seseorang belum tentu diterima dengan cara yang sama oleh orang lain (Laana & Sukri, 2022).

Setiap individu pasti memiliki pola kehidupan yang berbeda-beda, yang terbentuk dari tanggung jawab serta peran sosial yang dijalani dalam kehidupan sehari-hari. Perbedaan tersebut berpengaruh terhadap cara seseorang mengelola waktu, menyusun prioritas aktivitas, serta menyesuaikan diri terhadap lingkungan sosialnya. Dalam dunia pendidikan, seorang mahasiswa memiliki tuntutan dan kewajiban yang harus diselesaikan, yang secara tidak langsung ini dapat mempengaruhi pola hidup dan keseimbangan kehidupannya.

Mahasiswa tidak hanya dituntut untuk menyelesaikan studi tepat waktu, tetapi juga harus mampu mengerjakan berbagai tugas akademik yang menjadi penilaian setiap mata kuliah. Selain itu, mahasiswa sering kali mengikuti kegiatan organisasi, kepanitiaan, ataupun kegiatan lainnya. Di mana kegiatan tersebut bisa memberikan dampak positif untuk mahasiswa seperti mengembangkan kemampuan kepemimpinan, memperluas jaringan sosial, dan menyalurkan minat serta bakat individu.

Gaya hidup mahasiswa pada dasarnya terbentuk dari pola kebiasaan dan aktivitas yang dilakukan secara berulang dalam kehidupan sehari-hari. Secara umum, gaya hidup mahasiswa tercermin dari beberapa aspek utama, antara lain aktivitas akademik, pola tidur, interaksi sosial, serta kegiatan fisik yang dilakukan mahasiswa. Kesehatan mahasiswa dipengaruhi oleh keseimbangan antara aspek-aspek tersebut, baik dari kesehatan fisik maupun mental. Menurut Bhat & Sheikh (2023), *Study-Life Balance* merupakan kemampuan individu dalam menyeimbangkan tanggung jawab akademik dengan aspek kehidupan lainnya sehingga dapat tercapai kehidupan yang lebih baik secara keseluruhan. Misalnya, pola tidur yang tidak teratur atau tingginya tingkat stress dapat memengaruhi daya

konsentrasi, motivasi belajar, dan kemampuan berpikir kritis (Sahri & Hasibuan, 2023). Sebaliknya, keteraturan dalam aktivitas fisik, pengelolaan waktu belajar yang baik, serta interaksi sosial yang positif dapat membantu mahasiswa menjaga stabilitas emosional sekaligus meningkatkan produktivitas pendidikannya (Bhat & Sheikh, 2023).

Gaya hidup mahasiswa kini tidak hanya terbentuk dari aktivitas akademik dan sosial, tetapi juga dipengaruhi oleh perkembangan teknologi dan media sosial. Mahasiswa pada era saat ini termasuk dalam kelompok yang dikenal sebagai generasi Z, yaitu kelompok individu yang tumbuh dari lingkungan digital dan teknologi (Hastini dkk., 2020). Kemudahan dalam mengakses informasi, komunikasi secara digital, serta tingkat keterhubungan dengan dunia maya, dapat mempengaruhi perilaku mahasiswa generasi Z, karena dibentuk oleh budaya digital, di mana penggunaan media sosial, *platform* daring, dan teknologi menjadi bagian yang tidak terpisahkan dari kehidupan sehari-hari (Nadhifah dkk., 2024). Sehingga, dapat disimpulkan bahwa gaya hidup mahasiswa saat ini merupakan hasil dari interaksi antara kebutuhan pendidikan, pengaruh lingkungan sosial, dan perkembangan teknologi digital.

2.10 Kajian Integrasi Topik dengan Al-Quran/Hadis

Dalam penelitian ini, nilai-nilai Islam dapat diimplementasikan melalui *ijtihad*, yakni usaha sungguh-sungguh untuk menemukan solusi atas suatu permasalahan dengan memanfaatkan ilmu pengetahuan yang dimiliki manusia. Penelitian ini melakukan pengembangan metode untuk mencari alternatif solusi yang lebih baik. Prinsip ini ditegaskan oleh Allah SWT dalam firman-Nya pada surah Al-Ankabut ayat 69:

وَالَّذِينَ جَاهَدُوا فِينَا لَنَهْدِيَنَّهُمْ سُبُلَنَا وَإِنَّ اللَّهَ لَمَعَ الْمُحْسِنِينَ ٦٩

“Orang-orang yang berusaha dengan sungguh-sungguh untuk (mencari keridaan) Kami benar-benar akan Kami tunjukkan kepada mereka jalan-jalan Kami. Sesungguhnya Allah benar-benar bersama orang-orang yang berbuat kebaikan” (QS. Al-Ankabut: 69)

Menurut Tafsir Tahlili terhadap Al-Ankabut ayat 69, istilah *jihad* tidak selalu diartikan sebagai peperangan fisik, tetapi juga mencakup seluruh bentuk kesungguhan, perjuangan, dan pengorbanan dalam menegakkan kebenaran. Konsep *jihad* dalam konteks ini mencakup usaha yang dilakukan secara sungguh-gungguh untuk mencari solusi atas berbagai persoalan yang dihadapi. Penelitian ini, menanamkan semangat *jihad* melalui modifikasi serta memvalidasi metode *Fuzzy C-Means* untuk menghasilkan pengelompokan yang akurat. Sehingga, penelitian ini tidak hanya merepresentasikan penerapan prinsip ilmiah dalam pengembangan metode analisis data, tetapi juga mencerminkan nilai-nilai Islam yang menekankan pentingnya kesungguhan dalam berupaya mencapai kebaikan (Kementerian Agama, 2022).

BAB III

METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif eksperimental yang berfokus pada pengujian algoritma *Fuzzy C-Means* (FCM) yang telah dimodifikasi untuk mengelompokkan data gaya hidup mahasiswa sebagai studi kasus. Pendekatan kuantitatif dipilih karena penelitian ini melibatkan proses pengumpulan, pengolahan, dan analisis data numerik untuk menguji kinerja algoritma FCM dalam menghasilkan pengelompokan data. Penelitian ini tidak hanya bersifat deskriptif, tetapi juga evaluatif, dengan tujuan menilai kualitas hasil klasterisasi menggunakan indeks *Modified Partition Coefficient* (MPC).

Dalam penelitian ini, eksperimen dilakukan melalui beberapa tahapan, seperti pra-pemrosesan data, penerapan algoritma FCM klasik yang menggunakan jarak *Euclidean* dan algoritma FCM yang telah dimodifikasi dengan jarak *Manhattan*, serta evaluasi kualitas hasil pengelompokan menggunakan MPC. Melalui rancangan eksperimen ini, penelitian diharapkan dapat menunjukkan sejauh mana modifikasi penggunaan jarak *Manhattan* berpengaruh terhadap performa algoritma FCM dalam menghasilkan pembagian *cluster* yang lebih baik dan lebih tahan terhadap data *outlier*.

3.2 Data dan Sumber Data

Data yang digunakan dalam penelitian ini adalah *dataset Student Lifestyle*, sebuah kumpulan data publik yang tersedia di *platform Kaggle* dan telah banyak digunakan dalam berbagai penelitian terkait analisis dampak kebiasaan sehari-hari

terhadap prestasi akademik dan kesejahteraan mahasiswa. Secara teknik, *dataset Student Lifestyle* terdiri dari 2000 baris data dan 8 variabel atau kolom, yang mewakili berbagai informasi penting. Variabel data yang digunakan pada penelitian seperti pada Tabel 3.1.

Tabel 3.1 Deskripsi Data

Variabel	Deskripsi
<i>Student ID</i>	Identitas Mahasiswa
<i>Study Hours Per Day</i>	Jumlah rata-rata jam yang digunakan mahasiswa untuk belajar mandiri atau mengerjakan tugas akademik setiap hari.
<i>Extracurricular Hours Per Day</i>	Jumlah rata-rata yang dihabiskan mahasiswa untuk kegiatan ekstrakurikuler, seperti organisasi, ukm, atau kegiatan non-akademik lainnya setiap hari.
<i>Sleep Hours Per Day</i>	Jumlah rata-rata jam tidur mahasiswa dalam satu hari, yang mencerminkan pola istirahat dan kualitas tidur.
<i>Social Hours Per Day</i>	Jumlah rata-rata jam yang digunakan mahasiswa untuk bersosialisasi dengan teman, keluarga, atau kegiatan sosial lainnya setiap hari.
<i>Physical Activity Hours Per Day</i>	Jumlah rata-rata jam yang digunakan mahasiswa untuk aktivitas fisik atau olahraga setiap hari.
GPA	Indeks Prestasi Kumulatif mahasiswa sebagai ukuran pencapaian akademik secara keseluruhan.
<i>Stress Level</i>	Tingkat stress Mahasiswa

3.3 Teknik Analisis Data

Analisis data dalam penelitian ini dilakukan dengan pendekatan kuantitatif menggunakan algoritma *clustering Fuzzy C-Means*. Tahapan analisis data yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Melakukan analisis deskriptif.
2. Melakukan *preprocessing* data
 - a. Seleksi Variabel

Seleksi variabel dilakukan seleksi variabel untuk menentukan variabel yang digunakan dalam proses *clustering*.

b. Deteksi *Outlier*

Deteksi *outlier* dilakukan menggunakan metode *Interquartile Range* (IQR).

c. Normalisasi Data

Melakukan normalisasi data menggunakan metode *Min-Max* pada variabel numerik agar memiliki skala yang sama.

3. Implementasi Algoritma *Fuzzy C-Means* (FCM)

a. Memasukkan parameter awal yaitu:

- 1) Banyak *cluster* (c) = Ditetapkan dalam rentang 2 hingga 5 untuk memperoleh banyak *cluster* yang optimal.
- 2) Pangkat pembobot (w) = 2 dipilih berdasarkan penelitian Pal & Bezdek (1995), karena nilai pangkat pembobot ini merupakan nilai standar yang banyak digunakan dan dinilai mampu memberikan keseimbangan stabilitas *cluster* dan tingkat *fuzziness*.
- 3) Maksimum iterasi ($MaxIter$) = 1000 ditetapkan nilai ini karena pada percobaan awal dengan $MaxIter = 100$, proses *clustering* belum mencapai batas *error* yang ditentukan (ϵ) = 10^{-5} . Oleh karena itu, jumlah iterasi maksimum diperbesar untuk memberikan kesempatan hingga algoritma mencapai kondisi konvergen.
- 4) *Error* terkecil yang diharapkan (ϵ) = 10^{-5}
- 5) Fungsi objektif awal (P_0) = 0 nilai ini berperan sebagai inisialisasi pada tahap awal proses *clustering*. Setelah itu, nilai fungsi objektif akan dihitung dan diperbarui pada setiap iterasi sampai tercapai kondisi konvergen.

- 6) Iterasi awal $(t) = 1$
- b. Membangkitkan bilangan *random* $\mu_{i,k}$ menggunakan Persamaan (2.1) dan (2.2).
- c. Menghitung pusat *cluster* $P_{k,j}$ menggunakan Persamaan (2.3).
- d. Menghitung fungsi objektif pada iterasi ke- t .

Pada penelitian ini, jarak *Euclidean* dimodifikasi menjadi jarak *Manhattan*, merujuk pada Persamaan (2.4), sehingga fungsi objektif dituliskan sebagai berikut:

$$P_t = \sum_{i=1}^n \sum_{k=1}^c \left(\left[\sum_{j=1}^m |x_{i,j} - V_{k,j}| \right]^2 \cdot (\mu_{i,k})^w \right) \quad (3.1)$$

- e. Menghitung perubahan matriks partisi yang merujuk pada Persamaan (2.5), sehingga perubahan matriks partisi dituliskan sebagai berikut:

$$\mu_{i,k} = \frac{\left[\left(\sum_{j=1}^m |x_{i,j} - V_{k,j}| \right)^2 \right]^{\frac{-1}{w-1}}}{\sum_{k=1}^c \left[\left(\sum_{j=1}^m |x_{i,j} - V_{k,j}| \right)^2 \right]^{\frac{-1}{w-1}}}$$

$$\mu_{i,k} = \frac{\left[\sum_{j=1}^m |x_{i,j} - V_{k,j}| \right]^{\frac{-2}{w-1}}}{\sum_{k=1}^c \left[\sum_{j=1}^m |x_{i,j} - V_{k,j}| \right]^{\frac{-2}{w-1}}} \quad (3.2)$$

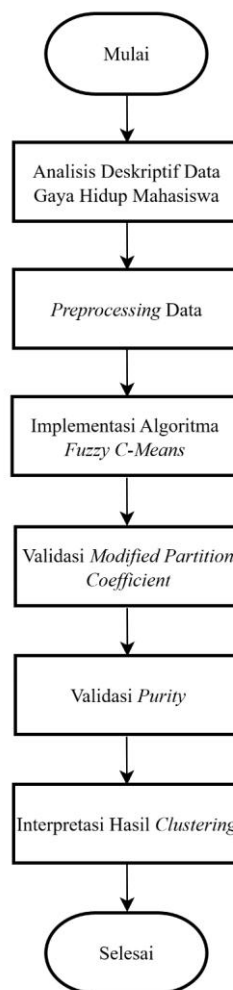
- f. Memeriksa kondisi berhenti
- 1) Jika: $(|P_t - P_{t-1}| < \varepsilon)$ atau $(t > MaxIter)$ maka berhenti
 - 2) Jika tidak: $t = t + 1$, ulangi langkah ke-b
4. Menghitung nilai validasi *Modified Partition Coefficient* (MPC) menggunakan Persamaan (2.8) pada hasil *clustering* yang diperoleh dari penggunaan jarak *Euclidean* dan jarak *Manhattan*. Nilai MPC dari kedua ukuran jarak dibandingkan untuk melihat apakah terdapat perbedaan kualitas

clustering yang dihasilkan. Perbedaan nilai tersebut digunakan untuk melihat pengaruh dari penggunaan jarak yang berbeda dalam proses *clustering*.

5. Menghitung nilai validasi eksternal *Purity* menggunakan Persamaan 2.10 pada hasil *clustering* yang diperoleh dari penggunaan jarak *Euclidean* dan jarak *Manhattan* berdasarkan label aktual.
6. Interpretasi Hasil *Clustering*

Interpretasi dilakukan dengan menganalisis nilai rata-rata variabel pada setiap *cluster* untuk menentukan karakteristik data.

Alur tahapan penelitian yang akan dilakukan ditunjukkan pada Gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian

BAB IV

HASIL DAN PEMBAHASAN

4.1 Analisis Deskriptif

Analisis deskriptif ini bertujuan untuk memahami pola distribusi data pada masing-masing variabel penelitian. Penelitian ini menggunakan 2000 data yang menggambarkan gaya hidup mahasiswa. Dalam penelitian ini, terdapat 6 variabel, yaitu *Study Hours Per Day* (X_1), *Extracurricular Hours Per Day* (X_2), *Sleep Hours Per Day* (X_3), *Social Hours Per Day* (X_4), *Physical Activity Hours Per Day* (X_5), dan GPA (X_6). Analisis untuk masing-masing variabel terdapat pada Tabel 4.1.

Tabel 4.1 Analisis Deskriptif Variabel

Variabel	Mean	Median	Standar Deviasi	Minimum	Maximum
X_1	7,4758	7,4	1,42388837	5	10
X_2	1,9901	2	1,155854658	0	4
X_3	7,50125	7,5	1,460948535	5	10
X_4	2,70455	2,6	1,688514092	0	6
X_5	4,3283	4,1	2,514110078	0	13
X_6	3,11596	3,11	0,298673536	2,24	4

Variabel X_1 menunjukkan waktu belajar mahasiswa per hari (*study hours per day*). Berdasarkan Tabel 4.1, diketahui bahwa variabel X_1 memiliki nilai *mean* atau rata-rata sebesar 7,48 jam dengan nilai standar deviasi sebesar 1,42. Nilai *minimum* yang diperoleh adalah 5 jam dan nilai *maksimum* sebesar 10 jam. Hal ini menunjukkan bahwa waktu belajar mahasiswa cenderung konsisten dan tidak memiliki perbedaan yang terlalu jauh antar data.

Variabel X_2 merupakan waktu yang digunakan mahasiswa untuk kegiatan ekstrakurikuler (*extracurricular hours per day*). Berdasarkan Tabel 4.1, variabel X_2 memiliki nilai *mean* atau rata-rata sebesar 1,99 jam dengan standar deviasi sebesar 1,16. Nilai *minimum* sebesar 0 jam menunjukkan bahwa terdapat mahasiswa tidak mengikuti kegiatan ekstrakurikuler, sedangkan nilai *maksimum* sebesar 4 jam menunjukkan adanya mahasiswa yang cukup aktif dalam kegiatan tersebut.

Variabel X_3 merupakan waktu tidur mahasiswa per hari (*sleep hours per day*). Berdasarkan Tabel 4.1, variabel X_3 memiliki nilai *mean* atau rata-rata sebesar 7,50 jam dengan standar deviasi sebesar 1,46. Nilai *minimum* sebesar 5 jam dan *maksimum* sebesar 10 jam menunjukkan bahwa sebagian besar mahasiswa memiliki pola tidur yang cukup baik dan relatif stabil.

Variabel X_4 merupakan waktu aktivitas sosial mahasiswa (*social hours per day*). Berdasarkan Tabel 4.1, variabel X_4 memiliki nilai *mean* atau rata-rata sebesar 2,70 jam dengan standar deviasi sebesar 1,69. Nilai *minimum* sebesar 0 jam dan *maksimum* sebesar 6 jam menunjukkan bahwa aktivitas sosial mahasiswa cukup beragam, ada yang hampir tidak melakukan aktivitas sosial, tetapi ada juga yang cukup aktif.

Variabel X_5 merupakan waktu aktivitas fisik mahasiswa (*physical activity hours per day*). Berdasarkan Tabel 4.1, variabel X_5 memiliki nilai *mean* atau rata-rata sebesar 4,33 jam dengan standar deviasi sebesar 2,51. Nilai *minimum* sebesar 0 jam dan *maksimum* sebesar 13 jam menunjukkan bahwa aktivitas fisik mahasiswa sangat bervariasi dan memiliki penyebaran data yang cukup luas dibandingkan variabel lainnya.

Variabel X_6 merupakan nilai indeks prestasi kumulatif mahasiswa (*GPA*). Berdasarkan Tabel 4.1, variabel X_6 memiliki nilai *mean* atau rata-rata sebesar 3,12 dengan standar deviasi sebesar 0,30. Nilai *minimum* sebesar 2,24 dan *maksimum* sebesar 4,00 menunjukkan bahwa performa akademik mahasiswa cenderung baik dan nilainya relatif berdekatan satu sama lain.

4.2 *Preprocessing Data*

Pada tahap ini akan dilakukan *preprocessing* yang merupakan salah satu tahap paling penting untuk melakukan analisis data dari data mentah menjadi data yang siap digunakan. Tahap *preprocessing* yang diterapkan dalam penelitian ini meliputi deteksi *outlier* menggunakan metode *Interquartile Range* (IQR), serta normalisasi data untuk menyamakan skala antar variabel.

4.2.1 *Seleksi Variabel*

Pada tahap awal, dilakukan seleksi variabel untuk menentukan variabel yang digunakan dalam proses *clustering*. Variabel *stress level* tidak digunakan dalam proses *clustering* karena berperan sebagai label. Selain itu, variabel *Student ID* juga tidak diolah karena hanya berfungsi sebagai identitas data dan tidak memiliki pengaruh terhadap proses *clustering*.

Dengan demikian, variabel yang digunakan dalam proses *clustering* adalah *Study Hours Per Day*, *Extracurricular Hours Per Day*, *Sleep Hours Per Day*, *Social Hours Per Day*, *Physical Activity Hours Per Day*, dan *GPA*.

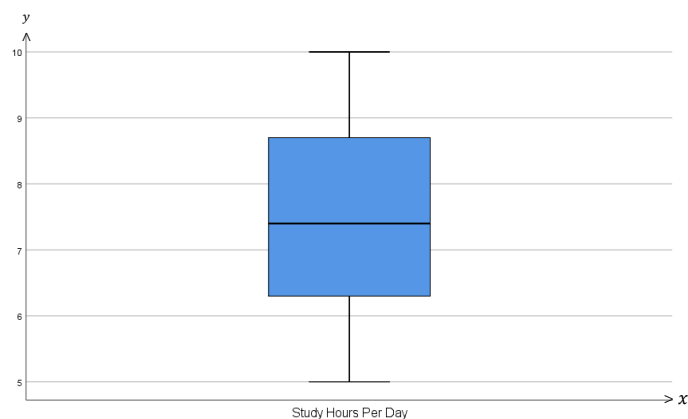
4.2.2 *Deteksi Outlier*

Deteksi *outlier* pada penelitian ini dilakukan untuk melihat apakah terdapat data yang nilainya menyimpang cukup jauh dibandingkan dengan sebagian besar data lainnya. Untuk itu, digunakan metode *Interquartile Range*

(IQR) dengan bantuan visualisasi *boxplot*. Berdasarkan hasil *boxplot*, sebagian besar variabel tidak menunjukkan adanya data yang berada di luar batas *whisker*, yaitu batas nilai yang masih dianggap normal berdasarkan metode IQR. Variabel tersebut meliputi *Study Hours Per Day*, *Extracurricular Hours Per Day*, *Sleep Hours Per Day*, dan *Social Hours Per Day*. Sehingga menunjukkan bahwa sebaran data pada variabel-variabel tersebut relatif stabil dan tidak terdapat nilai yang menyimpang secara signifikan.

Namun, pada variabel *Physical Activity Hours Per Day* ditemukan beberapa data yang berada di atas batas *whisker*, yang menunjukkan adanya nilai yang cukup jauh dibandingkan data yang lainnya. Selain itu, pada variabel GPA juga ditemukan beberapa data yang berada di luar batas bawah maupun batas atas *whisker*, sehingga dapat dikatakan terdapat nilai ekstrem baik pada sisi rendah maupun tinggi.

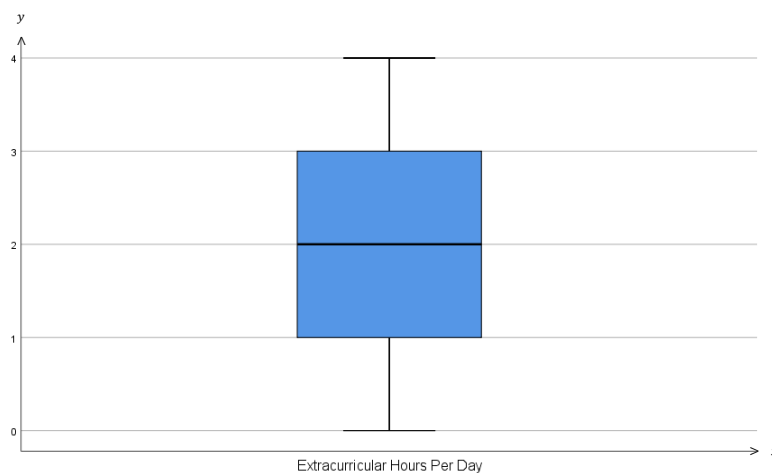
Untuk memberikan gambaran yang lebih jelas mengenai sebaran data pada setiap variabel, dilakukan visualisasi menggunakan *boxplot* secara terpisah. Dengan cara ini, pola data pada masing-masing variabel bisa dilihat dengan lebih detail. Visualisasi *boxplot* untuk variabel *Study Hours Per Day* ditampilkan pada Gambar 4.1.



Gambar 4.1 *Boxplot Study Hours Per Day*

Berdasarkan Gambar 4.1, tidak ditemukan data yang berada di luar batas *whisker*, sehingga variabel *Study Hours Per Day* tidak memiliki *outlier*. Batas *whisker* terlihat berimpit dengan nilai *minimum* dan *maksimum* data, yaitu antara 5 hingga 10 jam per hari. Hal ini menunjukkan bahwa seluruh data masih berada dalam rentang yang dianggap normal. Selain itu, sebaran data juga terlihat cukup merata di sekitar nilai tengah, sehingga dapat dikatakan bahwa variabel ini memiliki pola yang relatif stabil.

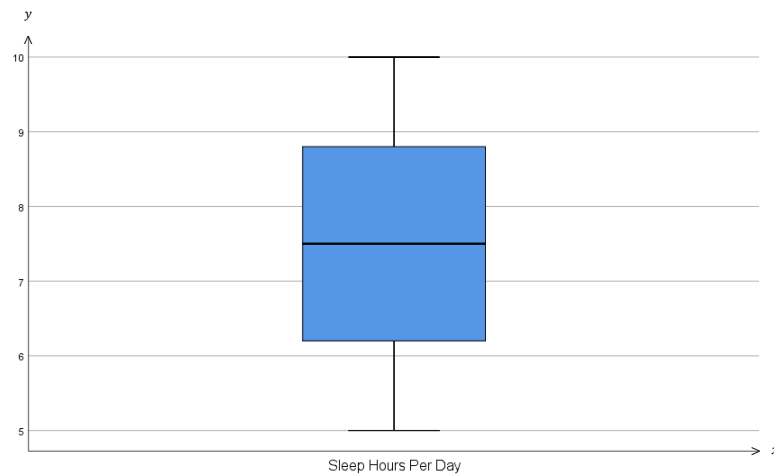
Selanjutnya, visualisasi boxplot pada variabel *Extracurricular Hours Per Day* ditunjukkan pada Gambar 4.2.



Gambar 4.2 Boxplot *Extracurricular Hours Per Day*

Berdasarkan Gambar 4.2, tidak terdapat data yang berada di luar batas *whisker*, sehingga variabel *Extracurricular Hours Per Day* tidak memiliki *outlier*. Batas *whisker* berada pada rentang 0 hingga 4 jam per hari, yang menunjukkan bahwa seluruh data masih berada dalam batas yang dianggap normal. Sebaran data terlihat cukup merata dengan nilai tengah berada di sekitar 2 jam per hari, sehingga dapat dikatakan bahwa variasi data pada variabel ini berada pada kondisi stabil.

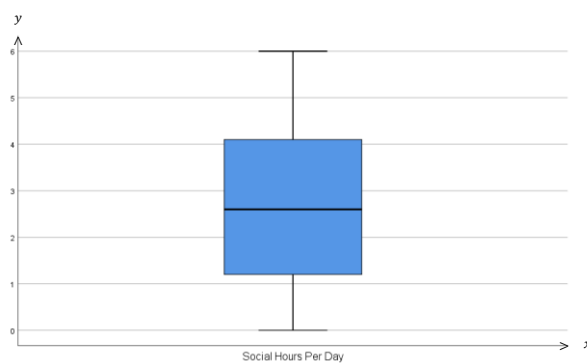
Visualisasi *boxplot* pada variabel *Sleep Hours Per Day* ditunjukkan pada Gambar 4.3.



Gambar 4.3 Boxplot *Sleep Hours Per Day*

Berdasarkan Gambar 4.3, tidak ditemukan data yang berada di luar batas *whisker*, sehingga variabel *Sleep Hours Per Day* tidak memiliki *outlier*. Batas *whisker* berada pada rentang 5 hingga 10 jam per hari, yang menunjukkan bahwa seluruh data masih berada dalam batas yang dianggap normal. Sebaran data juga terlihat cukup merata di sekitar nilai tengah, sehingga dapat dikatakan bahwa variabel ini memiliki pola yang relatif stabil.

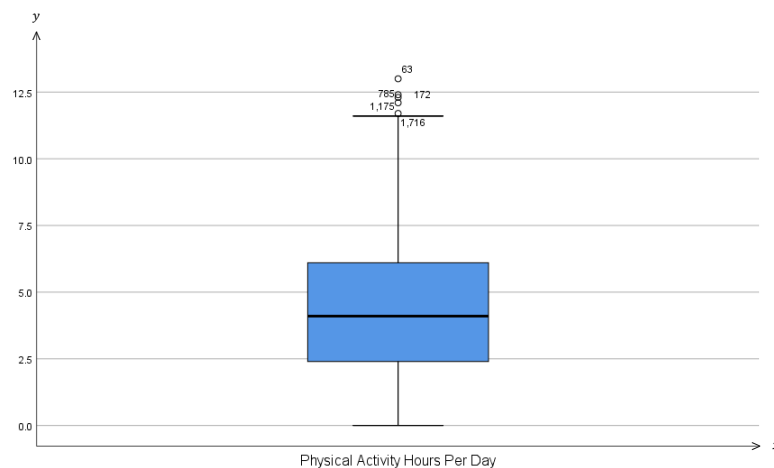
Visualisasi *boxplot* pada variabel *Social Hours Per Day* ditunjukkan pada Gambar 4.4.



Gambar 4.4 Boxplot *Social Hours Per Day*

Berdasarkan Gambar 4.4, tidak ditemukan data yang berada di luar batas *whisker*, sehingga variabel *Social Hours Per Day* tidak memiliki *outlier*. Batas *whisker* berada pada rentang 0 hingga 6 jam per hari, yang menunjukkan bahwa seluruh data masih berada dalam batas yang dianggap normal. Sebaran data juga terlihat cukup merata di sekitar nilai tengah, sehingga dapat dikatakan bahwa variabel ini memiliki pola yang relatif stabil.

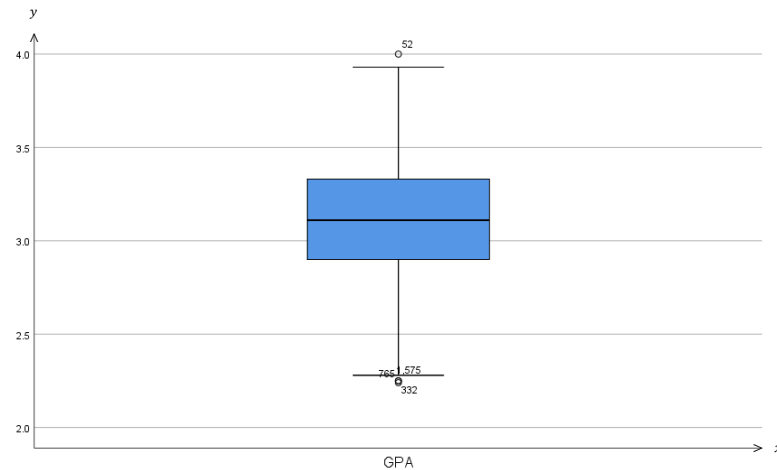
Visualisasi *boxplot* pada variabel *Physical Activity Hours Per Day* ditunjukkan pada Gambar 4.5.



Gambar 4.5 *Boxplot Physical Activity Hours Per Day*

Berdasarkan Gambar 4.5, terlihat adanya beberapa data yang berada di atas batas *whisker*, yang menunjukkan keberadaan *outlier* pada variabel *Physical Activity Hours Per Day*. Data ini memiliki nilai yang lebih tinggi dibandingkan dengan sebagian besar data lainnya. Pada variabel *Physical Activity Hours Per Day*, terlihat beberapa data yang berada di atas batas *whisker*, yaitu pada data ke-63, 172, 785, 1.175, dan 1.716. Nilai-nilai tersebut menunjukkan durasi aktivitas fisik yang lebih tinggi dibandingkan dengan sebagian besar data lainnya.

Visualisasi *boxplot* pada variabel GPA ditunjukkan pada Gambar 4.6.



Gambar 4. 6 *Boxplot* GPA

Berdasarkan Gambar 4.6, terlihat adanya data yang berada di luar batas *whisker* baik pada sisi atas maupun bawah, sehingga menunjukkan adanya *outlier* pada variabel GPA. Keberadaan *outlier* pada kedua sisi ini menunjukkan bahwa variasi nilai akademik antar data cukup beragam. Sebagian memiliki nilai akademik yang sangat tinggi, sementara sebagian lainnya memiliki nilai yang relatif rendah dibandingkan lainnya. *Outlier* tersebut pada data ke-52 untuk nilai yang tinggi, serta pada data ke-332, 765, dan 1.575 untuk nilai yang relatif rendah.

Meskipun sebagian besar data menunjukkan pola yang relatif konsisten, masih terdapat beberapa nilai yang menyimpang dari pola tersebut. Dalam penelitian ini, data *outlier* tidak dihapus dan keberadaan *outlier* juga digunakan sebagai pembanding dalam analisis selanjutnya.

4.2.3 Normalisasi Data

Perhitungan normalisasi data gaya hidup mahasiswa pada setiap variabel dihitung menggunakan metode normalisasi *Min-Max*, yang merupakan teknik untuk mengubah nilai setiap variabel berada pada rentang yang sama yaitu 0

hingga 1. Agar perbedaan skala antar variabel bisa diminimalkan sebelum proses analisis. Nilai normalisasi dihitung menggunakan

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

di mana:

x' : Hasil Normalisasi

x : Nilai asli pada data

x_{min} : Nilai minimum dari seluruh data

x_{max} : Nilai maksimum dari seluruh data

Proses normalisasi tersebut dilakukan dengan bantuan software *Excel*.

Hasil normalisasi data penelitian ini dapat dilihat pada Tabel 4.2. Tabel ini menyajikan nilai-nilai yang telah dinormalisasi untuk setiap variabel dari semua data.

Tabel 4.2 Normalisasi Data

Data	X_1	X_2	X_3	X_4	X_5	X_6
1	0,380000	0,950000	0,740000	0,466667	0,138462	0,426136
2	0,060000	0,875000	0,600000	0,700000	0,230769	0,289773
3	0,020000	0,975000	0,840000	0,200000	0,353846	0,244318
4	0,300000	0,525000	0,440000	0,283333	0,500000	0,363636
5	0,620000	0,150000	0,300000	0,366667	0,507692	0,721591
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1996	0,300000	0,050000	0,480000	0,350000	0,600000	0,613636
1997	0,260000	0,700000	0,760000	0,250000	0,353846	0,232955
1998	0,240000	0,000000	0,240000	0,133333	0,830769	0,511364
1999	0,620000	0,175000	0,520000	0,583333	0,315385	0,454545
2000	0,800000	0,425000	0,460000	0,516667	0,223077	0,761364

Berikut merupakan contoh perhitungan manual pada data pertama untuk semua variabel menggunakan normalisasi *Min-Max*. Variabel X_1 , diketahui nilai minimum 5 dan nilai maksimum sebesar 10. Sehingga perhitungan normalisasi data untuk variabel X_1 sebagai berikut:

$$x_{1,1} = \frac{6,9 - 5}{10 - 5} = 0,380000$$

Variabel X_2 , diketahui nilai minimum 0 dan nilai maksimum sebesar 4.

Sehingga perhitungan normalisasi data untuk variabel X_2 sebagai berikut:

$$x_{2,1} = \frac{3,8 - 0}{4 - 0} = 0,950000$$

Variabel X_3 , diketahui nilai minimum 5 dan nilai maksimum sebesar 10.

Sehingga perhitungan normalisasi data untuk variabel X_3 sebagai berikut:

$$x_{3,1} = \frac{8,7 - 5}{10 - 5} = 0,740000$$

Variabel X_4 , diketahui nilai minimum 0 dan nilai maksimum sebesar 6.

Sehingga perhitungan normalisasi data untuk variabel X_4 sebagai berikut:

$$x_{4,1} = \frac{2,8 - 0}{6 - 0} = 0,466667$$

Variabel X_5 , diketahui nilai minimum 0 dan nilai maksimum sebesar 13.

Sehingga perhitungan normalisasi data untuk variabel X_5 sebagai berikut:

$$x_{5,1} = \frac{1,8 - 0}{13 - 0} = 0,138462$$

Variabel X_6 , diketahui nilai minimum 2,24 dan nilai maksimum sebesar

4. Sehingga perhitungan normalisasi data untuk variabel X_6 sebagai berikut:

$$x_{6,1} = \frac{2,99 - 2,24}{4 - 2,24} = 0,426136$$

4.3 Implementasi Algoritma *Fuzzy C-Means*

Setelah melalui tahap *preprocessing* data, langkah berikutnya adalah menerapkan algoritma *Fuzzy C-Means* untuk melakukan proses pengelompokan. Pada penelitian ini, proses *clustering* tidak hanya menggunakan satu jenis ukuran jarak, tetapi dua, yaitu jarak *Euclidean* dan jarak *Manhattan*. Penggunaan kedua

pendekatan ini bertujuan untuk melihat apakah ada perbedaan hasil *clustering* yang dihasilkan, sekaligus membandingkan bagaimana masing-masing metode jarak memengaruhi proses pengelompokan data.

4.3.1 Perhitungan Menggunakan Jarak *Euclidean*

Algoritma *Fuzzy C-Means* menggunakan jarak *Euclidean* untuk mengukur kedekatan antar data. Data yang digunakan dalam penelitian ini merupakan data gaya hidup mahasiswa. Sebelum dilakukan proses *clustering*, data tersebut sudah melalui tahap *preprocessing*, salah satunya normalisasi. Setelah itu, data yang sudah siap digunakan direpresentasikan ke dalam bentuk matriks X dengan ukuran $n \times m$ (n = banyaknya sampel data, m = banyaknya variabel setiap data), pada penelitian ini, terdapat 2000 data dengan 6 variabel. Sehingga matriks X dapat direpresentasikan sebagai berikut:

$$X = \begin{bmatrix} 0,380000 & 0,950000 & 0,740000 & \dots & 0,426136 \\ 0,060000 & 0,875000 & 0,600000 & \dots & 0,289773 \\ 0,020000 & 0,975000 & 0,840000 & \dots & 0,244318 \\ 0,300000 & 0,525000 & 0,440000 & \dots & 0,363636 \\ 0,620000 & 0,150000 & 0,300000 & \dots & 0,721591 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0,300000 & 0,050000 & 0,480000 & \dots & 0,613636 \\ 0,260000 & 0,700000 & 0,760000 & \dots & 0,232955 \\ 0,240000 & 0 & 0,240000 & \dots & 0,511364 \\ 0,620000 & 0,175000 & 0,520000 & \dots & 0,454545 \\ 0,800000 & 0,425000 & 0,460000 & \dots & 0,761364 \end{bmatrix}$$

Selanjutnya proses *clustering* dengan menentukan parameter yang digunakan, antara lain:

1. Banyak *cluster* tidak ditetapkan secara langsung, tetapi diuji dalam beberapa kemungkinan. Rentang yang digunakan adalah $c = 2$ hingga $c = 5$, sehingga hasil *clustering* dari masing masing *cluster* bisa dibandingkan untuk menentukan banyak *cluster* yang paling sesuai.

2. Pangkat pembobot (w) = 2.
3. Maksimum iterasi ($MaxIter = 1000$) digunakan untuk mengetahui apakah proses iterasi sudah mencapai kondisi konvergen.
4. *Error* terkecil yang diharapkan ($\varepsilon = 10^{-5}$), sebagai syarat untuk menghentikan iterasi
5. Fungsi objektif awal ($P_0 = 0$) dimulai dari nilai awal 0, yang kemudian digunakan sebagai acuan dalam proses iterasi selanjutnya.
6. Iterasi awal dimulai dari $t = 1$.

Proses perhitungan algoritma dilakukan setelah menentukan parameter, kemudian langkah selanjutnya adalah membangkitkan bilangan *random* $\mu_{i,k}$ sebagai elemen matriks partisi U . Nilai ini digunakan untuk menunjukkan derajat keanggotaan setiap data terhadap *cluster*. Dengan syarat untuk setiap baris akan memiliki jumlah nilai elemen kolom yang sama yaitu 1. Proses pembangkitan bilangan *random* dibangkitkan menggunakan *tools Python* dengan mengacu pada Persamaan (2.1) dan (2.2). Karena banyak *cluster* yang digunakan ada pada rentang $c = 2$ hingga $c = 5$, maka matriks partisi awal U yang dibangkitkan juga akan berbeda untuk setiap banyak *cluster*.

Perhitungan algoritma *Fuzzy C-Means* tidak ditampilkan secara rinci untuk semua banyak *cluster*. Sebagai gambaran proses, perhitungan difokuskan pada $c = 2$. Seluruh tahapan ditunjukkan secara bertahap, mulai dari pembentukan matriks partisi awal, kemudian dilanjutkan dengan perhitungan pusat *cluster*, hingga fungsi objektif dan proses pembaruan derajat keanggotaan. Untuk banyak *cluster* lainnya, yaitu $c = 3$, $c = 4$, dan $c = 5$, langkah yang dilakukan pada dasarnya sama. Perbedaannya hanya terletak pada banyak *cluster*

yang digunakan. Sebagai gambaran, hasil pembangkitan bilangan *random* $\mu_{i,k}$ untuk $c = 2$ ditunjukkan pada Tabel 4.3.

Tabel 4.3 Bilangan *Random* $\mu_{i,k}$ untuk $c = 2$

Data	$\mu_{i,1}$	$\mu_{i,2}$
1	0,282618	0,717382
2	0,550102	0,449898
3	0,500039	0,499961
4	0,062843	0,937157
5	0,459151	0,540849
$\vdots$	$\vdots$	$\vdots$
1996	0,283471	0,716529
1997	0,936183	0,063817
1998	0,526977	0,473023
1999	0,458729	0,541271
2000	0,766577	0,233423

Berdasarkan Tabel 4.3 yang menunjukkan nilai bilangan random $\mu_{i,k}$ yang dibangkitkan secara acak. Nilai ini digunakan sebagai awal sebelum proses iterasi dilakukan pada algoritma *Fuzzy C-Means*. Setiap baris memiliki jumlah nilai sebesar 1, sehingga sudah memenuhi syarat sebagai matriks partisi.

Tahap selanjutnya setelah matriks partisi awal U terbentuk, proses kemudian dilanjutkan ke tahap perhitungan pusat *cluster* ($V_{k,j}$). Nilai pusat *cluster* tersebut dihitung menggunakan Persamaan (2.3).

Langkah pertama dalam menghitung nilai pusat *cluster* dimulai dengan mengolah bilangan *random* $\mu_{i,k}$ pada Tabel 4.3 dipangkatkan dengan pangkat pembobot (w) yang telah ditentukan, yaitu 2. Beberapa hasil dari $(\mu_{i,k})^2$ ditunjukkan pada Tabel 4.4.

Tabel 4.4 Kuadrat Bilangan *Random* $(\mu_{i,k})^2$

Data	$(\mu_{i,1})^2$	$(\mu_{i,2})^2$
1	0,079873	0,514635
2	0,302612	0,202408
3	0,250039	0,249961
4	0,003949	0,878263
5	0,210819	0,292518
$\vdots$	$\vdots$	$\vdots$
1996	0,080355	0,513415
1997	0,876439	0,004073
1998	0,277705	0,223750
1999	0,210432	0,292974
2000	0,587642	0,054486

Setelah diperoleh nilai $(\mu_{i,k})^2$, nilai dari setiap $(\mu_{i,k})^2$ dikalikan dengan data setiap variabel. Selanjutnya, hasil tersebut dibagi dengan jumlah $(\mu_{i,k})^2$ untuk mendapatkan nilai pusat *cluster*. Proses perhitungan untuk pusat *cluster* ditunjukkan sebagai berikut:

Cluster 1

$$V_{1,1} = \frac{(0,079873 \times 0,380000) + (0,302612 \times 0,060000) + \dots + (0,587642 \times 0,800000)}{0,079873 + 0,302612 + \dots + 0,587642}$$

$$= 0,492824$$

$$V_{1,2} = \frac{(0,079873 \times 0,950000) + (0,302612 \times 0,875000) + \dots + (0,587642 \times 0,425000)}{0,079873 + 0,302612 + \dots + 0,587642}$$

$$= 0,492033$$

$\vdots$

$$V_{1,6} = \frac{(0,079873 \times 0,426136) + (0,302612 \times 0,289773) + \dots + (0,587642 \times 0,761364)}{0,079873 + 0,302612 + \dots + 0,587642}$$

$$= 0,499058$$

Cluster 2

$$V_{2,1} = \frac{(0,514635 \times 0,380000) + (0,202408 \times 0,060000) + \dots + (0,054486 \times 0,800000)}{0,514635 + 0,202408 + \dots + 0,054486}$$

$$\begin{aligned}
&= 0,498198 \\
V_{2,2} &= \frac{(0,514635 \times 0,950000) + (0,202408 \times 0,875000) + \dots + (0,054486 \times 0,425000)}{0,514635 + 0,202408 + \dots + 0,054486} \\
&= 0,499636 \\
&\vdots \\
V_{2,6} &= \frac{(0,514635 \times 0,426136) + (0,202408 \times 0,289773) + \dots + (0,054486 \times 0,761364)}{0,514635 + 0,202408 + \dots + 0,054486} \\
&= 0,496400
\end{aligned}$$

Sehingga diperoleh pusat *cluster* ($V_{k,j}$) ditunjukkan pada Tabel 4.5.

Tabel 4.5 Pusat *Cluster* Menggunakan Jarak *Euclidean*

Variabel	Cluster 1	Cluster 2
X_1	0,492824	0,498198
X_2	0,492033	0,499636
X_3	0,499592	0,500889
X_4	0,458473	0,445858
X_5	0,332227	0,333144
X_6	0,499058	0,496400

Berdasarkan Tabel 4.5, diperoleh nilai pusat *cluster* untuk masing-masing *cluster* pada setiap variabel. Nilai ini selanjutnya digunakan untuk perhitungan fungsi objektif pada iterasi selanjutnya. Setelah nilai pusat *cluster* diperoleh, tahap selanjutnya adalah mencari fungsi objektif pada iterasi- t . Perhitungan ini dilakukan untuk mengetahui nilai fungsi objektif pada setiap iterasi yang digunakan sebagai acuan dalam menentukan kondisi konvergen. Nilai fungsi objektif dihitung menggunakan Persamaan (2.4).

Jarak antara setiap objek data ke pusat *cluster* dihitung lalu dikuadratkan, kemudian hasil kuadrat tersebut dijumlahkan untuk masing-masing *cluster*. Selanjutnya, jumlah kuadrat jarak pada setiap *cluster* dikalikan dengan nilai $(\mu_{i,k})^2$ untuk memperoleh nilai fungsi objektif pada masing-masing *cluster*. Nilai fungsi objektif yang diperoleh dari setiap *cluster* kemudian dijumlahkan untuk

mendapatkan nilai fungsi objektif total (P). $\sum_{j=1}^m (x_{i,j} - V_{k,j})^2$ dinyatakan sebagai jarak antara data ke- i terhadap pusat *cluster* ke- k , yang dinotasikan dengan $d_{i,k}^2$. Contoh perhitungan untuk data ke-1 sebagai berikut:

Cluster 1

$$\begin{aligned} d_{1,1}^2 &= (0,380000 - 0,492824)^2 + (0,950000 - 0,492033)^2 + \dots + (0,426136 - 0,499058)^2 \\ &= 0,323189 \end{aligned}$$

Cluster 2

$$\begin{aligned} d_{1,2}^2 &= (0,380000 - 0,498198)^2 + (0,950000 - 0,499636)^2 + \dots + (0,426136 - 0,496400)^2 \\ &= 0,317243 \end{aligned}$$

Setelah diperoleh nilai total jarak kuadrat, langkah berikutnya adalah mengalikan nilai tersebut dengan derajat keanggotaan yang sudah dipangkatkan.

Contoh proses perhitungan pada data ke-1 ditunjukkan sebagai berikut:

Cluster 1

$$\begin{aligned} d_{1,1}^2 \times (\mu_{i,k})^2 &= 0,323187 \times 0,079873 \\ &= 0,025814 \end{aligned}$$

Cluster 2

$$\begin{aligned} d_{1,2}^2 \times (\mu_{i,k})^2 &= 0,317243 \times 0,514635 \\ &= 0,163265 \end{aligned}$$

Sehingga total untuk data ke-1 sebagai berikut:

$$0,025814 + 0,163265 = 0,189079$$

Perhitungan yang sama dilakukan untuk data ke-2 hingga data ke-2000 dengan menggunakan langkah yang sama dengan data ke-1. Kemudian diperoleh nilai fungsi objektif untuk seluruh data yang kemudian dijumlahkan untuk

mendapatkan nilai fungsi objektif total. Beberapa hasil perhitungan untuk *cluster* 1 dan *cluster* 2 ditunjukkan pada Tabel 4.6 dan Tabel 4.7.

Tabel 4.6 Nilai *Cluster* 1

Cluster 1					
Data	$(X_1 - V_1)^2$	$(X_2 - V_1)^2$	...	$(X_6 - V_1)^2$	Σ
1	0,012729	0,209734	...	0,005318	0,323189
2	0,187337	0,146664	...	0,043800	0,456512
3	0,223562	0,233257	...	0,064893	0,704866
4	0,037181	0,001087	...	0,018339	0,118980
5	0,016174	0,116986	...	0,049521	0,261735
⋮	⋮	⋮	⋮	⋮	⋮
1996	0,037181	0,195393	...	0,013128	0,329555
1997	0,054207	0,043250	...	0,070811	0,280009
1998	0,063920	0,242096	...	0,000151	0,727816
1999	0,016174	0,100510	...	0,001981	0,134955
2000	0,094357	0,004493	...	0,068804	0,184522

Tabel 4.7 Nilai *Cluster* 2

Cluster 2					
Data	$(X_1 - V_2)^2$	$(X_2 - V_2)^2$	...	$(X_6 - V_2)^2$	Σ
1	0,013971	0,202827	...	0,004937	0,317243
2	0,192017	0,140898	...	0,042695	0,460501
3	0,228673	0,225970	...	0,063545	0,694060
4	0,039282	0,000643	...	0,017626	0,115515
5	0,014836	0,122246	...	0,050711	0,264888
⋮	⋮	⋮	⋮	⋮	⋮
1996	0,039282	0,202173	...	0,013744	0,336037
1997	0,056738	0,040146	...	0,069403	0,272214
1998	0,066666	0,249637	...	0,000224	0,729893
1999	0,014836	0,105389	...	0,001752	0,141556
2000	0,091085	0,005571	...	0,070206	0,185662

Berdasarkan Tabel 4.6 dan Tabel 4.7, diperoleh hasil perhitungan jarak kuadrat untuk setiap data terhadap masing-masing *cluster*. Nilai ini kemudian digunakan dalam perhitungan fungsi objektif pada tahap selanjutnya, yaitu dengan mengalikan total nilai jarak kuadrat dengan kuadrat derajat keanggotaan $(\mu_{i,k})^2$ pada Tabel 4.4. Beberapa hasil nilai fungsi objektif untuk setiap data pada

masing- masing *cluster*, hingga fungsi objektif total akan ditunjukkan pada Tabel 4.8.

Berdasarkan Tabel 4.8, diperoleh nilai fungsi objektif untuk setiap data pada masing-masing *cluster*, serta nilai fungsi objektif total sebesar 488,034251. Nilai tersebut merupakan hasil penjumlahan dari seluruh data terhadap masing-masing *cluster* berdasarkan derajat keanggotaannya dan digunakan pada perhitungan iterasi selanjutnya.

Tabel 4.8 Fungsi Objektif Menggunakan Jarak *Euclidean*

Data	Cluster 1	Cluster 2	Fungsi Objektif
	$d_{i,k}^2 \times (\mu_{i,k})^2$	$d_{i,k}^2 \times (\mu_{i,k})^2$	
1	0,025814	0,163265	0,189079
2	0,138146	0,093209	0,231355
3	0,176244	0,173488	0,349732
4	0,000470	0,101452	0,101922
5	0,055179	0,077484	0,132663
⋮	⋮	⋮	⋮
1996	0,026482	0,172526	0,199008
1997	0,245411	0,001109	0,246519
1998	0,202118	0,163314	0,365432
1999	0,028399	0,041472	0,069871
2000	0,108433	0,010116	0,118549
Fungsi Objektif Total (P)			488,034251

Untuk memperoleh nilai fungsi objektif P pada iterasi selanjutnya, nilai derajat keanggotaan $\mu_{i,k}$ perlu diperbarui terlebih dahulu. Perhitungannya dimulai dengan mencari jarak masing-masing data terhadap setiap pusat *cluster*. Perhitungan pembaruan nilai μ_{ik} dihitung menggunakan Persamaan (2.5).

Sebagai gambaran proses perhitungan, berikut ditunjukkan cara menentukan derajat keanggotaan untuk data ke-1 terhadap *cluster* 1 dan *cluster* 2. Langkah perhitungannya dapat dituliskan sebagai berikut:

Dengan menggunakan pangkat pembobot $w = 2$, nilai pangkat dapat dituliskan sebagai berikut:

$$\frac{-1}{w-1} = \frac{-1}{2-1} = -1$$

untuk mempermudah proses perhitungan derajat keanggotaan $\mu_{i,k}$, bentuk

$\left[\sum_{j=1}^m (x_{i,j} - V_{k,j})^2 \right]^{-1}$ dapat dituliskan sebagai berikut:

$$\frac{1}{\left[\sum_{j=1}^m (x_{i,j} - V_{k,j})^2 \right]}$$

$\sum_{j=1}^m (x_{i,j} - V_{k,j})^2$ dinyatakan sebagai jarak antara data ke- i terhadap pusat *cluster* ke- k , yang dinotasikan dengan $d_{i,k}^2$.

hitung jarak ke *cluster* 1

$$\begin{aligned} d_{1,1}^2 &= (0,380000 - 0,492824)^2 + (0,950000 - 0,492033)^2 + \dots + (0,426136 - 0,499058)^2 \\ &= 0,323189 \end{aligned}$$

hitung jarak ke *cluster* 2

$$\begin{aligned} d_{1,2}^2 &= (0,380000 - 0,498198)^2 + (0,950000 - 0,499636)^2 + \dots + (0,426136 - 0,496400)^2 \\ &= 0,317243 \end{aligned}$$

sehingga derajat keanggotaan untuk *cluster* 1

$$\begin{aligned} \mu_{1,1} &= \frac{\frac{1}{0,323189}}{\left(\frac{1}{0,323189} \right) + \left(\frac{1}{0,317243} \right)} \\ \mu_{1,1} &= \frac{3,09417}{3,09417 + 3,15216} = 0,495358 \end{aligned}$$

derajat keanggotaan untuk *cluster* 2

$$\begin{aligned} \mu_{1,2} &= \frac{\frac{1}{0,317243}}{\left(\frac{1}{0,323189} \right) + \left(\frac{1}{0,317243} \right)} \\ \mu_{1,2} &= \frac{3,15216}{3,09417 + 3,15216} = 0,504642 \end{aligned}$$

Dari proses perhitungan yang telah dilakukan, maka diperoleh nilai derajat keanggotaan untuk setiap data pada masing-masing *cluster*. Beberapa hasil dari derajat keanggotaan tersebut ditunjukkan pada Tabel 4.9

Tabel 4.9 Nilai $\mu_{i,k}$ Baru Menggunakan Jarak *Euclidean*

Data	Cluster 1	Cluster 2
1	0,495358	0,504642
2	0,502175	0,497825
3	0,496138	0,503862
4	0,492611	0,507389
5	0,502994	0,497006
$\vdots$	$\vdots$	$\vdots$
1996	0,504870	0,495130
1997	0,492942	0,507058
1998	0,500712	0,499288
1999	0,511937	0,488063
2000	0,501539	0,498461

Dari proses perhitungan yang sudah dilakukan, diperoleh nilai derajat keanggotaan untuk setiap data pada masing-masing *cluster*, yang beberapa nilainya ditunjukkan pada Tabel 4.9. Nilai ini akan terus diperbarui hingga mencapai kondisi konvergen.

Langkah berikutnya adalah mengecek apakah proses iterasi sudah bisa dihentikan. Pengecekan ini dilakukan dengan membandingkan hasil nilai fungsi objektif pada iterasi ke- t dengan iterasi sebelumnya, lalu melihat apakah selisih tersebut sudah lebih kecil dari batas toleransi yang telah ditetapkan, yaitu $\varepsilon = 10^{-5}$.

Pada iterasi pertama, diperoleh nilai fungsi objektif sebesar 488,034251 dengan batas toleransi yang digunakan yaitu $\varepsilon = 10^{-5}$. Karena ini merupakan iterasi awal, belum ada nilai fungsi objektif dari iterasi sebelumnya. Oleh karena itu, nilai awal diasumsikan $P_0 = 0$ sebagai pembanding. Selanjutnya, dilakukan

perhitungan selisih antara nilai fungsi objektif saat ini dengan nilai awal. Hasilnya diperoleh:

$$|P_1 - P_0| = |488,034251 - 0| = 488,034251$$

maka,

$$488,034251 > 10^{-5}$$

Dari hasil perhitungan yang didapat, selisih nilainya masih lebih besar dari batas tersebut. Sehingga, proses tersebut belum mencapai kondisi konvergen. Sehingga, iterasi tidak berhenti di sini. Perhitungan dilanjutkan kembali ke tahap berikutnya dengan memperbarui nilai pusat *cluster*, fungsi objektif, dan derajat keanggotaan. proses ini terus diulang dengan langkah yang sama sampai nilai fungsi objektif memenuhi kondisi konvergen.

Setelah proses perhitungan dilakukan berulang dengan langkah yang sama, diperoleh nilai fungsi objektif yang sudah tidak terlalu besar lagi. Pada penelitian ini, kondisi tersebut tercapai pada iterasi ke-47, sehingga proses dihentikan pada iterasi itu. Nilai fungsi objektif pada iterasi ke-47 adalah 393,349220, sedangkan pada iterasi sebelumnya, yaitu iterasi ke-46 bernilai 393,349229. Selisih antara keduanya kemudian dihitung untuk memastikan apakah sudah memenuhi batas toleransi yang ditentukan

$$|P_{47} - P_{46}| = |393,349220 - 393,349229| = 0,000009$$

maka,

$$0,000009 < 10^{-5}$$

dari hasil perhitungan, selisih yang diperoleh sudah lebih kecil dari nilai ε , sehingga proses iterasi telah mencapai kondisi konvergen.

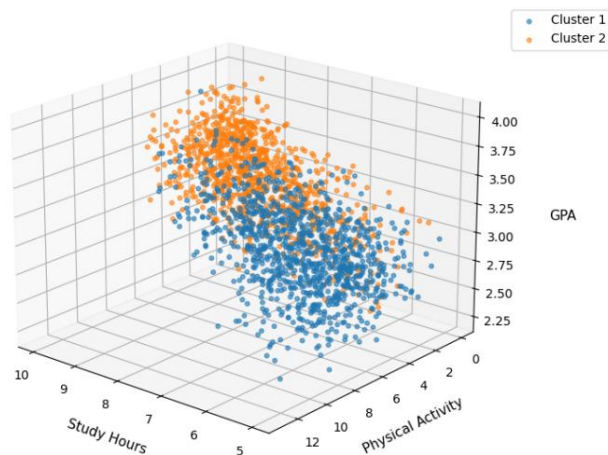
Dengan berhentinya proses iterasi ini, diperoleh nilai derajat keanggotaan akhir untuk setiap data terhadap masing-masing *cluster*. Nilai ini digunakan sebagai hasil akhir dari proses *clustering*, dan akan disajikan dalam bentuk Tabel 4.10.

Tabel 4.10 Hasil *Clustering* Menggunakan Jarak *Euclidean*

Data	$\mu_{i,1}$	$\mu_{i,2}$	Cluster
1	0,400094	0,599906	Cluster 2
2	0,569387	0,430613	Cluster 1
3	0,498739	0,501261	Cluster 2
4	0,708827	0,291173	Cluster 1
5	0,551825	0,448175	Cluster 1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1996	0,629527	0,370473	Cluster 1
1997	0,491279	0,508721	Cluster 2
1998	0,605038	0,394962	Cluster 1
1999	0,555446	0,444554	Cluster 1
2000	0,320883	0,679117	Cluster 2

Berdasarkan hasil pada Tabel 4.10, setiap data memiliki derajat keanggotaan pada masing-masing *cluster*. Nilai ini menunjukkan seberapa dekat suatu data terhadap *cluster* tertentu. Semakin besar nilainya, berarti data tersebut cenderung lebih sesuai dengan *cluster* tersebut. Sehingga setiap data akan dikelompokkan ke dalam *cluster* yang paling sesuai berdasarkan hasil perhitungan tersebut.

Setelah diperoleh hasil *clustering* dengan banyak *cluster* $c = 2$, kemudian dilakukan visualisasi menggunakan plot tiga dimensi (3D). Visualisasi ini bertujuan untuk memberikan gambaran mengenai penyebaran data pada masing-masing *cluster*. Plot dibuat menggunakan tiga variabel yang dipilih berdasarkan hasil analisis korelasi, yaitu *Study Hours Per Day*, *Physical Activity Per Day*, dan GPA. Hasil visualisasi *cluster* ditunjukkan pada Gambar 4.7.



Gambar 4.7 Visualisasi Hasil *Clustering* Menggunakan Jarak *Euclidean*

Berdasarkan Gambar 4.7, terlihat bahwa data pada *Cluster 1* dan *Cluster 2* membentuk pola penyebaran yang berbeda. Perbedaan tersebut dapat dilihat dari posisi titik-titik pada masing-masing *cluster* yang cenderung membentuk kelompok tersendiri. Namun, masih terdapat beberapa titik yang saling berdekatan bahkan tumpang tindih, sebagian besar anggota masing-masing *cluster* cenderung berkumpul pada area tertentu sehingga menunjukkan adanya perbedaan karakteristik antar *cluster*. Terlihat bahwa anggota pada *Cluster 2* cenderung memiliki nilai *Study Hours* dan GPA yang lebih tinggi dibandingkan anggota pada *Cluster 1*.

4.3.2 Perhitungan Menggunakan Jarak *Manhattan*

Algoritma *Fuzzy C-Means* menggunakan jarak *Manhattan* untuk mengukur kedekatan antardata. Data yang digunakan dalam penelitian ini merupakan data gaya hidup mahasiswa. Sebelum dilakukan proses *clustering*, data tersebut sudah melalui tahap *preprocessing*, salah satunya normalisasi. Setelah itu, data yang sudah siap digunakan direpresentasikan ke dalam bentuk matriks X dengan ukuran $n \times m$ (n = banyaknya sampel data, m = banyaknya

variabel setiap data), pada penelitian ini, terdapat 2000 data dengan 6 variabel.

Sehingga matriks X dapat direpresentasikan sebagai berikut:

$$X = \begin{bmatrix} 0,380000 & 0,950000 & 0,740000 & \cdots & 0,426136 \\ 0,060000 & 0,875000 & 0,600000 & \cdots & 0,289773 \\ 0,020000 & 0,975000 & 0,840000 & \cdots & 0,244318 \\ 0,300000 & 0,525000 & 0,440000 & \cdots & 0,363636 \\ 0,620000 & 0,150000 & 0,300000 & \cdots & 0,721591 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0,300000 & 0,050000 & 0,480000 & \cdots & 0,613636 \\ 0,260000 & 0,700000 & 0,760000 & \cdots & 0,232955 \\ 0,240000 & 0 & 0,240000 & \cdots & 0,511364 \\ 0,620000 & 0,175000 & 0,520000 & \cdots & 0,454545 \\ 0,800000 & 0,425000 & 0,460000 & \cdots & 0,761364 \end{bmatrix}$$

Selanjutnya proses *clustering* dengan menentukan parameter yang digunakan, antara lain:

1. Banyak *cluster* tidak ditetapkan secara langsung, tetapi diuji dalam beberapa kemungkinan. Rentang yang digunakan adalah $c = 2$ hingga $c = 5$, sehingga hasil *clustering* dari masing masing *cluster* bisa dibandingkan untuk menentukan banyak *cluster* yang paling sesuai.
2. Pangkat pembobot (w) = 2.
3. Maksimum iterasi ($Maxlter = 1000$) digunakan untuk mengetahui apakah proses iterasi sudah mencapai kondisi konvergen.
4. *Error* terkecil yang diharapkan ($\varepsilon = 10^{-5}$), sebagai syarat untuk menghentikan iterasi
5. Fungsi objektif awal ($P_0 = 0$) dimulai dari nilai awal 0, yang kemudian digunakan sebagai acuan dalam proses iterasi selanjutnya.
6. Iterasi awal dimulai dari $t = 1$.

Proses perhitungan algoritma dilakukan setelah menentukan parameter, kemudian langkah selanjutnya adalah membangkitkan bilangan *random* $\mu_{i,k}$

sebagai elemen matriks partisi U . Nilai ini digunakan untuk menunjukkan derajat keanggotaan setiap data terhadap $cluster$. Dengan syarat untuk setiap baris akan memiliki jumlah nilai elemen kolom yang sama yaitu 1. Proses pembangkitan bilangan *random* dibangkitkan menggunakan *tools Python* dengan mengacu pada Persamaan (2.1) dan (2.2). Karena banyak $cluster$ yang digunakan ada pada rentang $c = 2$ hingga $c = 5$, maka matriks partisi awal U yang dibangkitkan juga akan berbeda untuk setiap banyak $cluster$.

Perhitungan algoritma *Fuzzy C-Means* tidak ditampilkan secara rinci untuk semua banyak $cluster$. Sebagai gambaran proses, perhitungan difokuskan pada $c = 2$. Seluruh tahapan ditunjukkan secara bertahap, mulai dari pembentukan matriks partisi awal, kemudian dilanjutkan dengan perhitungan pusat $cluster$, hingga fungsi objektif dan proses pembaruan derajat keanggotaan. Untuk banyak $cluster$ lainnya, yaitu $c = 3$, $c = 4$, dan $c = 5$, langkah yang dilakukan pada dasarnya sama. Perbedaannya hanya terletak pada banyak $cluster$ yang digunakan. Sebagai gambaran, hasil pembangkitan bilangan *random* $\mu_{i,k}$ untuk $c = 2$ ditunjukkan pada Tabel 4.11.

Tabel 4.11 Bilangan *Random* $\mu_{i,k}$ untuk $c = 2$

Data	$\mu_{i,1}$	$\mu_{i,2}$
1	0,282618	0,717382
2	0,550102	0,449898
3	0,500039	0,499961
4	0,062843	0,937157
5	0,459151	0,540849
$\vdots$	$\vdots$	$\vdots$
1996	0,283471	0,716529
1997	0,936183	0,063817
1998	0,526977	0,473023
1999	0,458729	0,541271
2000	0,766577	0,233423

Tahap selanjutnya setelah matriks partisi awal U terbentuk, kemudian proses dilanjutkan ke tahap perhitungan pusat *cluster* ($V_{k,j}$). Nilai pusat *cluster* tersebut dihitung menggunakan Persamaan (2.3).

Langkah pertama dalam menghitung nilai pusat *cluster* dimulai dengan mengolah bilangan *random* $\mu_{i,k}$ dipangkatkan dengan pangkat pembobot (w) yang telah ditentukan, yaitu 2. Sehingga diperoleh hasil dari $(\mu_{i,k})^2$ yang ditunjukkan pada Tabel 4.12.

Tabel 4.12 Kuadrat Bilangan *Random* $(\mu_{i,k})^2$

Data	$(\mu_{i,1})^2$	$(\mu_{i,2})^2$
1	0,079873	0,514635
2	0,302612	0,202408
3	0,250039	0,249961
4	0,003949	0,878263
5	0,210819	0,292518
$\vdots$	$\vdots$	$\vdots$
1996	0,080355	0,513415
1997	0,876439	0,004073
1998	0,277705	0,223750
1999	0,210432	0,292974
2000	0,587642	0,054486

Berdasarkan Tabel 4.12, diperoleh nilai $(\mu_{i,k})^2$ untuk setiap data pada masing-masing *cluster*. Setelah diperoleh nilai $(\mu_{i,k})^2$, nilai dari setiap $(\mu_{i,k})^2$ dikalikan dengan data setiap variabel. Selanjutnya, hasil tersebut dibagi dengan jumlah $(\mu_{i,k})^2$ untuk mendapatkan nilai pusat *cluster*. Proses perhitungan untuk pusat *cluster* ditunjukkan sebagai berikut:

Cluster 1

$$V_{1,1} = \frac{(0,079873 \times 0,380000) + (0,302612 \times 0,060000) + \dots + (0,587642 \times 0,800000)}{0,079873 + 0,302612 + \dots + 0,587642}$$

$$= 0,492824$$

$$V_{1,2} = \frac{(0,079873 \times 0,950000) + (0,302612 \times 0,875000) + \dots + (0,587642 \times 0,425000)}{0,079873 + 0,302612 + \dots + 0,587642}$$

$$= 0,492033$$

$$\vdots$$

$$V_{1,6} = \frac{(0,079873 \times 0,426136) + (0,302612 \times 0,289773) + \dots + (0,587642 \times 0,761364)}{0,079873 + 0,302612 + \dots + 0,587642}$$

$$= 0,499058$$

Cluster 2

$$V_{2,1} = \frac{(0,514635 \times 0,380000) + (0,202408 \times 0,060000) + \dots + (0,054486 \times 0,800000)}{0,514635 + 0,202408 + \dots + 0,054486}$$

$$= 0,498198$$

$$V_{2,2} = \frac{(0,514635 \times 0,950000) + (0,202408 \times 0,875000) + \dots + (0,054486 \times 0,425000)}{0,514635 + 0,202408 + \dots + 0,054486}$$

$$= 0,499636$$

$$\vdots$$

$$V_{2,6} = \frac{(0,514635 \times 0,426136) + (0,202408 \times 0,289773) + \dots + (0,054486 \times 0,761364)}{0,514635 + 0,202408 + \dots + 0,054486}$$

$$= 0,496400$$

Sehingga diperoleh nilai pusat *cluster* ($V_{k,j}$) yang ditunjukkan pada Tabel 4.13.

Tabel 4.13 Pusat *Cluster* Menggunakan Jarak *Manhattan*

Variabel	<i>Cluster 1</i>	<i>Cluster 2</i>
X_1	0,492824	0,498198
X_2	0,492033	0,499636
X_3	0,499592	0,500889
X_4	0,458473	0,445858
X_5	0,332227	0,333144
X_6	0,499058	0,496400

Berdasarkan Tabel 4.13, diperoleh nilai pusat *cluster* untuk masing-masing *cluster* pada setiap variabel, yang akan digunakan pada perhitungan selanjutnya. Setelah nilai pusat *cluster* diperoleh, tahap selanjutnya adalah

mencari fungsi objektif pada iterasi- t . Perhitungan ini dilakukan untuk mengetahui nilai fungsi objektif pada setiap iterasi yang digunakan sebagai acuan dalam menentukan kondisi konvergen. Nilai fungsi objektif dihitung menggunakan Persamaan (3.1).

Jarak antara setiap objek data ke pusat *cluster* dihitung menggunakan selisih absolut pada masing-masing variabel. Nilai absolut kemudian dijumlahkan untuk setiap variabel sehingga diperoleh total jarak antara data dan pusat *cluster*. Selanjutnya, nilai jarak total dikuadratkan menggunakan rumus di atas. Kemudian, hasilnya dikalikan dengan nilai $(\mu_{i,k})^2$ untuk memperoleh nilai fungsi objektif pada masing-masing *cluster*. Nilai fungsi objektif yang diperoleh dari setiap *cluster* kemudian dijumlahkan untuk mendapatkan nilai fungsi objektif total (P). $\sum_{j=1}^m |x_{i,j} - V_{k,j}|$ dinyatakan sebagai jarak *Manhattan* antara data ke- i terhadap pusat *cluster* ke- k , yang dinotasikan dengan $d_{i,k}$. Contoh proses perhitungan pada data ke-1 ditunjukkan sebagai berikut:

Cluster 1

$$\begin{aligned} d_{1,1} &= |0,380000 - 0,492824| + |0,950000 - 0,492033| + \dots + |0,426136 - 0,499058| \\ &= 1,086080 \end{aligned}$$

Cluster 2

$$\begin{aligned} d_{1,2} &= |0,380000 - 0,498198| + |0,950000 - 0,499636| + \dots + |0,426136 - 0,496400| \\ &= 1,093426 \end{aligned}$$

Berdasarkan contoh perhitungan tersebut, jarak antara setiap data terhadap masing-masing pusat *cluster* dapat dihitung dengan cara yang sama. Perhitungan ini dilakukan untuk seluruh data pada setiap cluster sehingga menghasilkan nilai jarak yang menggambarkan tingkat kedekatan data terhadap pusat cluster. Nilai

jarak yang diperoleh kemudian digunakan pada tahap perhitungan berikutnya dalam proses clustering. Beberapa hasil perhitungan jarak untuk seluruh data ditunjukkan pada Tabel 4.14 dan Tabel 4.15.

Tabel 4.14 Nilai *Cluster 1*

Data	<i>Cluster 1</i>				
	$d_{1,1}$	$d_{2,1}$	...	$d_{6,1}$	Σ
1	0,112824	0,457967	...	0,072922	1,086080
2	0,432824	0,382967	...	0,209286	1,468469
3	0,472824	0,482967	...	0,254740	1,831031
4	0,192824	0,032967	...	0,135422	0,763718
5	0,127176	0,342033	...	0,222533	1,158606
⋮	⋮	⋮	⋮	⋮	⋮
1996	0,192824	0,442033	...	0,114578	1,145273
1997	0,232824	0,207967	...	0,266104	1,197395
1998	0,252824	0,492033	...	0,012305	1,840436
1999	0,127176	0,317033	...	0,044513	0,650832
2000	0,307176	0,067033	...	0,262305	0,843450

Tabel 4.15 Nilai *Cluster 2*

Data	<i>Cluster 2</i>				
	$d_{1,2}$	$d_{2,2}$	...	$d_{6,2}$	Σ
1	0,118198	0,450364	...	0,070263	1,093426
2	0,438198	0,375364	...	0,206627	1,475815
3	0,478198	0,475364	...	0,252082	1,811314
4	0,198198	0,025364	...	0,132763	0,746595
5	0,121802	0,349636	...	0,225191	1,151259
⋮	⋮	⋮	⋮	⋮	⋮
1996	0,198198	0,449636	...	0,117237	1,148675
1997	0,238198	0,200364	...	0,263445	1,177678
1998	0,258198	0,499636	...	0,014964	1,843838
1999	0,121802	0,324636	...	0,041854	0,662638
2000	0,301802	0,074636	...	0,264964	0,863167

Dari Tabel 4.14 dan Tabel 4.15, diperoleh nilai total jarak untuk setiap data terhadap masing-masing *cluster*. Nilai tersebut kemudian dipangkatkan oleh pangkat pembobot $w = 2$, sehingga menghasilkan nilai jarak kuadrat yang akan digunakan untuk perhitungan selanjutnya. Contoh proses perhitungan pada data ke-1 ditunjukkan sebagai berikut:

Cluster 1

$$d_{1,1}^2 = (1,086080)^2 = 1,179570$$

Cluster 2

$$d_{1,2}^2 = (1,093426)^2 = 1,195580$$

Setelah diperoleh nilai total jarak kuadrat, langkah berikutnya adalah mengalikan nilai tersebut dengan derajat keanggotaan yang sudah dipangkatkan. Contoh proses perhitungan pada data ke-1 ditunjukkan sebagai berikut:

Cluster 1

$$\begin{aligned} d_{1,1}^2 \times (\mu_{i,k})^2 &= 1,179570 \times 0,079873 \\ &= 0,094215 \end{aligned}$$

Cluster 2

$$\begin{aligned} d_{1,2}^2 \times (\mu_{i,k})^2 &= 1,195580 \times 0,514635 \\ &= 0,615290 \end{aligned}$$

sehingga total untuk data ke-1 sebagai berikut:

$$0,094215 + 0,615290 = 0,709505$$

Perhitungan yang sama dilakukan untuk data ke-2 hingga data ke-2000 dengan menggunakan langkah yang sama dengan data ke-1. Kemudian diperoleh nilai fungsi objektif untuk seluruh data yang kemudian dijumlahkan untuk mendapatkan nilai fungsi objektif total.

Beberapa hasil dari nilai total jarak yang telah dikuadratkan ditampilkan pada Tabel 4.16.

Tabel 4.16 Hasil Kuadrat Jarak

Data	Cluster 1	Cluster 2
	$d_{i,1}^2$	$d_{i,2}^2$
1	1,179569	1,195580
2	2,156401	2,178030
3	3,352676	3,280859
4	0,583266	0,557404
5	1,342367	1,325398
$\vdots$	$\vdots$	$\vdots$
1996	1,311651	1,319454
1997	1,433755	1,386925
1998	3,387206	3,399739
1999	0,423582	0,439089
2000	0,711408	0,745058

Berdasarkan Tabel 4.16, diperoleh hasil total jarak yang telah dikuadratkan untuk setiap data pada masing-masing *cluster*, nilai tersebut akan digunakan dalam tahap perhitungan selanjutnya. Beberapa hasil nilai fungsi objektif hingga fungsi objektif total akan ditunjukkan pada Tabel 4.17.

Tabel 4.17 Fungsi Objektif Menggunakan Jarak *Manhattan*

Data	Cluster 1	Cluster 2	Fungsi Objektif
	$d_{i,1}^2 \times (\mu_{i,1})^2$	$d_{i,2}^2 \times (\mu_{i,2})^2$	
1	0,094215	0,615290	0,709506
2	0,652552	0,440852	1,093404
3	0,838299	0,820088	1,658387
4	0,002303	0,489547	0,491851
5	0,282997	0,387702	0,670700
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1996	0,105399	0,677426	0,782825
1997	1,256598	0,005648	1,262246
1998	0,940645	0,760693	1,701338
1999	0,089135	0,128642	0,217777
2000	0,418052	0,040596	0,458647
Fungsi Objektif Total (<i>P</i>)			2214,592379

Berdasarkan Tabel 4.17, diperoleh nilai fungsi objektif untuk setiap data pada masing-masing *cluster*, serta nilai fungsi objektif total sebesar 2214,592379.

Untuk memperoleh nilai fungsi objektif *P* pada iterasi selanjutnya, nilai derajat keanggotaan $\mu_{i,k}$ perlu diperbarui terlebih dahulu. Perhitungannya

dimulai dengan mencari jarak masing-masing data terhadap setiap pusat *cluster*. Perhitungan pembaruan nilai $\mu_{i,k}$ dihitung menggunakan Persamaan (3.2).

Sebagai gambaran proses perhitungan, berikut ditunjukkan cara menentukan derajat keanggotaan untuk data pertama terhadap *cluster* 1 dan *cluster* 2. Langkah perhitungannya dapat dituliskan sebagai berikut:

Dengan menggunakan pangkat pembobot $w = 2$, nilai pangkat dapat dituliskan sebagai berikut:

$$\frac{-2}{w-1} = \frac{-2}{2-1} = -2$$

untuk mempermudah proses perhitungan derajat keanggotaan $\mu_{i,k}$, bentuk

$[\sum_{j=1}^m |x_{i,j} - V_{k,j}|]^{-2}$ dapat dituliskan sebagai berikut:

$$\mu_{i,k} = \frac{d_{i,k}^{-2}}{\sum_{k=1}^c d_{i,k}^{-2}} = \frac{\frac{1}{d_{i,k}^2}}{\sum_{k=1}^c \frac{1}{d_{i,k}^2}}$$

contoh proses perhitungan pada data ke-1 ditunjukkan sebagai berikut:

hitung jarak ke *cluster* 1

$$\begin{aligned} d_{1,1} &= |0,380000 - 0,492824| + |0,950000 - 0,492033| + \dots + |0,426136 - 0,499058| \\ &= 1,086080 \end{aligned}$$

$$d_{1,1}^2 = (1,086080)^2 = 1,179570$$

hitung jarak ke *cluster* 2

$$\begin{aligned} d_{1,2} &= |0,380000 - 0,498198| + |0,950000 - 0,499636| + \dots + |0,426136 - 0,496400| \\ &= 1,093426 \end{aligned}$$

$$d_{1,2}^2 = (1,093426)^2 = 1,195580$$

Menghitung derajat keanggotaan untuk *cluster* 1

$$\begin{aligned}\mu_{1,1} &= \frac{\frac{1}{1,179570}}{\frac{1}{1,179570} + \frac{1}{1,195580}} \\ &= \frac{0,847767}{0,847767 + 0,836414} = 0,503370\end{aligned}$$

Menghitung derajat keanggotaan untuk *cluster* 2

$$\begin{aligned}\mu_{1,2} &= \frac{\frac{1}{1,195580}}{\frac{1}{1,179570} + \frac{1}{1,195580}} \\ &= \frac{0,836414}{0,847767 + 0,836414} = 0,496630\end{aligned}$$

Dari proses perhitungan yang telah dilakukan, maka diperoleh nilai derajat keanggotaan untuk setiap data pada masing-masing *cluster*. Beberapa hasil dari derajat keanggotaan ($\mu_{i,k}$) tersebut ditunjukkan pada Tabel 4.18.

Tabel 4.18 Nilai $\mu_{i,k}$ Baru Menggunakan Jarak *Manhattan*

Data	Cluster 1	Cluster 2
1	0,503370	0,496630
2	0,502495	0,497505
3	0,494587	0,505413
4	0,488664	0,511336
5	0,496820	0,503180
⋮	⋮	⋮
1996	0,501483	0,498517
1997	0,491699	0,508301
1998	0,500923	0,499077
1999	0,508987	0,491013
2000	0,511552	0,488448

Dari proses perhitungan yang sudah dilakukan, diperoleh nilai derajat keanggotaan untuk setiap data pada masing-masing *cluster*, yang beberapa nilainya ditunjukkan pada Tabel 4.18. Nilai ini akan terus diperbarui hingga mencapai kondisi konvergen.

Langkah berikutnya adalah mengecek apakah proses iterasi sudah bisa dihentikan. Pengecekan ini dilakukan dengan membandingkan hasil nilai fungsi

objektif pada iterasi ke- t dengan iterasi sebelumnya, lalu melihat apakah selisih tersebut sudah lebih kecil dari batas toleransi yang telah ditetapkan, yaitu $\varepsilon = 10^{-5}$.

Pada iterasi pertama, diperoleh nilai fungsi objektif sebesar 2214,592379 dengan batas toleransi yang digunakan yaitu $\varepsilon = 10^{-5}$. Karena ini merupakan iterasi awal, belum ada nilai fungsi objektif dari iterasi sebelumnya. Oleh karena itu, nilai awal diasumsikan $P_0 = 0$ sebagai pembanding. Selanjutnya, dilakukan perhitungan selisih antara nilai fungsi objektif saat ini dengan nilai awal. Hasilnya diperoleh:

$$|P_1 - P_0| = |2214,592379 - 0| = 2214,592379$$

maka,

$$2214,592379 > 10^{-5}$$

Dari hasil perhitungan yang didapat, selisih nilainya masih lebih besar dari batas tersebut. Sehingga, proses tersebut belum mencapai kondisi konvergen. Sehingga, iterasi tidak berhenti di sini. Perhitungan dilanjutkan kembali ke tahap berikutnya dengan memperbarui nilai pusat *cluster*, fungsi objektif, dan derajat keanggotaan. proses ini terus diulang dengan langkah yang sama sampai nilai fungsi objektif memenuhi kondisi konvergen.

Setelah proses perhitungan dilakukan berulang dengan langkah yang sama, diperoleh nilai fungsi objektif yang sudah tidak terlalu besar lagi. Pada penelitian ini, kondisi tersebut tercapai pada iterasi ke-58, sehingga proses dihentikan pada iterasi itu. Nilai fungsi objektif pada iterasi ke-58 adalah 1744,536444, sedangkan pada iterasi sebelumnya, yaitu iterasi ke-57 bernilai 1744,536451. Selisih antara

keduanya kemudian dihitung untuk memastikan apakah sudah memenuhi batas toleransi yang ditentukan.

$$|P_{58} - P_{57}| = |1744,536444 - 1744,536451| = 0,000007$$

maka,

$$0,000007 < 10^{-5}$$

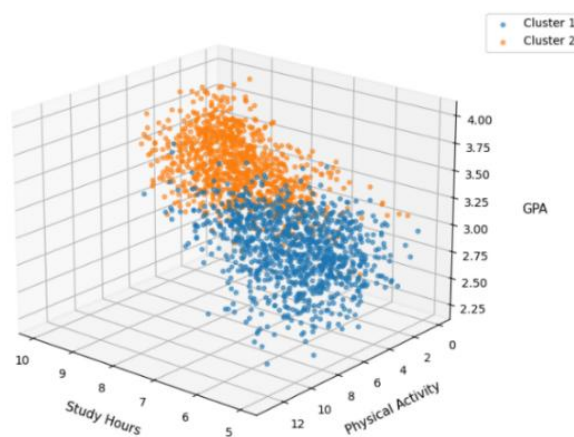
dari hasil perhitungan, selisih yang diperoleh sudah lebih kecil dari nilai ε , sehingga proses iterasi telah mencapai kondisi konvergen. Sehingga, diperoleh nilai derajat keanggotaan akhir untuk setiap data terhadap masing-masing *cluster*. Nilai ini digunakan sebagai hasil akhir dari proses *clustering*. Setiap data memiliki derajat keanggotaan pada masing-masing *cluster*, yang menunjukkan seberapa dekat suatu data terhadap *cluster* tertentu. Semakin besar nilainya, berarti data tersebut cenderung lebih sesuai dengan *cluster* tersebut. Sehingga setiap data akan dikelompokkan ke dalam *cluster* yang paling sesuai berdasarkan hasil perhitungan tersebut. Beberapa hasil derajat keanggotaan $\mu_{i,k}$ ditunjukkan pada Tabel 4.19.

Tabel 4.19 Hasil *Clustering* Menggunakan Jarak *Manhattan*

Data	$\mu_{i,1}$	$\mu_{i,2}$	Cluster
1	0,514468	0,485532	Cluster 1
2	0,522200	0,477800	Cluster 1
3	0,546393	0,453607	Cluster 1
4	0,824037	0,175963	Cluster 1
5	0,477105	0,522895	Cluster 2
⋮	⋮	⋮	⋮
1996	0,650963	0,349037	Cluster 1
1997	0,569937	0,430063	Cluster 1
1998	0,643590	0,356410	Cluster 1
1999	0,430115	0,569885	Cluster 2
2000	0,316763	0,683237	Cluster 2

Hasil pada Tabel 4.19, menyajikan beberapa hasil akhir *clustering* data yang diperoleh setelah proses iterasi berhenti mencapai kondisi konvergen.

Setelah diperoleh hasil *clustering* dengan banyak *cluster* $c = 2$, kemudian dilakukan visualisasi menggunakan plot tiga dimensi (3D). Visualisasi ini bertujuan untuk memberikan gambaran mengenai penyebaran data pada masing-masing *cluster*. Plot dibuat menggunakan tiga variabel yang dipilih berdasarkan hasil analisis korelasi, yaitu *Study Hours Per Day*, *Physical Activity Per Day*, dan GPA. Hasil visualisasi *cluster* ditunjukkan pada Gambar 4.8.



Gambar 4.8 Visualisasi Hasil *Clustering* Menggunakan Jarak *Manhattan*

Berdasarkan Gambar 4.8, terlihat bahwa data pada *Cluster 1* dan *Cluster 2* memiliki pola penyebaran yang berbeda. Namun demikian, masih terdapat beberapa titik yang saling tumpang tindih sehingga batas antar *cluster* tidak terlihat secara tegas. Meskipun demikian, visualisasi ini tetap dapat memberikan gambaran umum mengenai hasil pengelompokan data yang terbentuk.

4.3.3 Perbandingan Hasil *Clustering* Jarak *Euclidean* dan Jarak

Manhattan

Dalam tahap deteksi *outlier*, ditemukan beberapa data yang cukup berbeda dibandingkan sebagian besar data lain, tepatnya pada variabel *Physical Activity Hours Per Day* dan GPA. Data *outlier* ini tidak dihapus karena digunakan untuk

membandingkan hasil *clustering* menggunakan jarak *Euclidean* dan jarak *Manhattan* pada data yang mengandung nilai ekstrem.

Untuk melihat perbedaannya, dibandingkan posisi *outlier* pada hasil *clustering* dengan jarak *Euclidean* dan jarak *Manhattan*. Tabel 4.20 menampilkan *cluster* yang ditempati oleh setiap *outlier* dan nilai derajat keanggotaan tertinggi pada hasil *clustering* menggunakan jarak *Euclidean* dan jarak *Manhattan*.

Tabel 4.20 Tabel Perbandingan *Outlier*

Data	Variabel	$\mu_{i,1}$ <i>Euclidean</i>	$\mu_{i,2}$ <i>Euclidean</i>	Cluster <i>Euclidean</i>	$\mu_{i,1}$ <i>Manhattan</i>	$\mu_{i,2}$ <i>Manhattan</i>	Cluster <i>Manhattan</i>
52	GPA	0,332284	0,667716	Cluster 2	0,276488	0,723512	Cluster 2
63	Physical Activity	0,597700	0,402300	Cluster 1	0,636409	0,363591	Cluster 1
	Hours Per Day						
172	Physical Activity	0,599609	0,400391	Cluster 1	0,640170	0,359830	Cluster 1
	Hours Per Day						
332	GPA	0,575800	0,424200	Cluster 1	0,606026	0,393974	Cluster 1
765	GPA	0,653893	0,346107	Cluster 1	0,711692	0,288308	Cluster 1
785	Physical Activity	0,598975	0,401025	Cluster 1	0,640053	0,359947	Cluster 1
	Hours Per Day						
1175	Physical Activity	0,598317	0,401683	Cluster 1	0,650757	0,349243	Cluster 1
	Hours Per Day						
1575	GPA	0,629602	0,370398	Cluster 1	0,597146	0,402854	Cluster 1
1716	Physical Activity	0,603155	0,396845	Cluster 1	0,644986	0,355014	Cluster 1
	Hours Per Day						

Berdasarkan hasil perbandingan pada data *outlier* Tabel 4.20, ukuran jarak *Manhattan* memberikan hasil *clustering* yang lebih baik. Hal ini dapat dilihat dari nilai derajat keanggotaan pada ukuran jarak *Manhattan* yang cenderung lebih besar pada salah satu *cluster* dibandingkan dengan jarak *Euclidean*. Karena perhitungan jaraknya tidak menggunakan kuadrat selisih seperti jarak *Euclidean*, pengaruh data yang memiliki nilai ekstrem tidak terlalu besar. Oleh karena itu, berdasarkan hasil yang diperoleh, jarak *Manhattan* dapat dikatakan lebih sesuai karena mampu menghasilkan derajat keanggotaan yang lebih tegas dan memudahkan proses penentuan *cluster*.

Perbandingan juga dilakukan pada seluruh data penelitian untuk mengetahui penggunaan ukuran jarak *Euclidean* dan jarak *Manhattan*

memengaruhi hasil *clustering*. Hasil perbandingan dilakukan dengan mencocokkan *cluster* yang diperoleh setiap data pada kedua ukuran jarak, sehingga dapat diketahui banyak data yang memiliki hasil *cluster* yang sama maupun yang mengalami perubahan *cluster*. Hasil perbandingan tersebut ditunjukkan pada Tabel 4.21.

Tabel 4.21 Perbandingan Hasil *Clustering* Jarak *Euclidean* dan Jarak *Manhattan*

Keterangan	Jumlah Data
<i>Cluster</i> sama	1731
<i>Cluster</i> berbeda	269

Berdasarkan Tabel 4.21, dari 2000 data yang digunakan dalam penelitian, sebanyak 1731 data memiliki hasil *cluster* yang sama, sedangkan 269 data mengalami perubahan *cluster*. Hal ini menunjukkan bahwa penggunaan ukuran jarak memberikan pengaruh terhadap hasil *clustering*, yang ditunjukkan adanya perubahan *cluster* pada sebagian data.

Pada Tabel 4.20, data yang mengandung *outlier* tetap berada pada *cluster* yang sama pada kedua ukuran jarak. Sehingga, perubahan hasil *clustering* yang terjadi pada 269 data tidak mengandung data *outlier*. Oleh karena itu, berdasarkan hasil yang diperoleh, jarak *Manhattan* menggunakan selisih absolut dalam perhitungan jarak sehingga pengaruh nilai ekstrem terhadap perhitungan jarak tidak diperbesar.

4.4 Evaluasi *Modified Partition Coefficient*

Pada tahap ini, hasil *clustering* yang sudah diperoleh sebelumnya akan dilakukan evaluasi menggunakan ukuran validasi *Modified Partition Coefficient* (MPC). Akan dilakukan perhitungan untuk setiap banyak *cluster* yang telah diuji, yaitu $c = 2$ hingga $c = 5$. Perhitungan ini menggunakan nilai derajat keanggotaan yang diperoleh dari hasil *clustering* sebelumnya. Sebelum menghitung MPC,

terlebih dahulu dihitung nilai *Partition Coefficient* (PC). Nilai PC diperoleh dari penjumlahan kuadrat derajat keanggotaan setiap data terhadap seluruh *cluster*, lalu dirata-ratakan terhadap jumlah data. Sebagai gambaran proses perhitungan, pada bagian ini ditampilkan contoh perhitungan untuk banyak *cluster* $c = 2$ menggunakan jarak *Euclidean* dan jarak *Manhattan*.

Perhitungan menggunakan Jarak *Euclidean*

$$PC = \frac{1}{2000} \left(\sum_{i=1}^{2000} \sum_{j=1}^2 (0,400094359)^2 + (0,599905641)^2 + \dots + (0,679117006)^2 \right)$$

$$PC = \frac{1}{2000} (1073,531148) = 0,536765574$$

Setelah diperoleh nilai PC, langkah selanjutnya adalah menghitung nilai MPC.

Nilai MPC dihitung menggunakan Persamaan (2.9) sebagai berikut:

$$MPC = 1 - \frac{2}{2-1} (1 - 0,536765574) = 0,073531148$$

Perhitungan menggunakan Jarak *Manhattan*

$$PC = \frac{1}{2000} \left(\sum_{i=1}^{2000} \sum_{j=1}^2 (0,514467561)^2 + (0,485532439)^2 + \dots + (0,683236656)^2 \right)$$

$$PC = \frac{1}{2000} (1111,511606) = 0,555755803$$

Setelah diperoleh nilai PC, langkah selanjutnya adalah menghitung nilai MPC.

Nilai MPC dihitung menggunakan Persamaan (2.9) sebagai berikut:

$$MPC = 1 - \frac{2}{2-1} (1 - 0,555755803) = 0,111511606$$

Perhitungan yang sama dilakukan untuk banyak *cluster* lainnya, yaitu $c = 3$, $c = 4$, dan $c = 5$, baik menggunakan jarak *Euclidean* maupun jarak *Manhattan*.

Hasil perhitungan nilai MPC untuk masing-masing akan ditunjukkan pada Tabel 4.22.

Tabel 4.22 Hasil Validasi $c = 2$ hingga $c = 5$

Banyak Cluster (c)	MPC Euclidean	MPC Manhattan
2	0,073531148	0,111511606
3	0,054729472	0,082398573
4	0,044844737	0,072722283
5	0,037008041	0,058761035

Berdasarkan hasil pada Tabel 4.22, dapat dilihat bahwa nilai MPC untuk masing-masing *cluster* menunjukkan kecenderungan yang berbeda pada kedua jarak yang digunakan. Pada jarak *Euclidean*, nilai MPC tertinggi diperoleh pada $c = 2$ sebesar 0,073531148, kemudian nilainya menurun seiring bertambahnya banyak *cluster*. Pada jarak *Manhattan*, di mana nilai MPC tertinggi diperoleh pada $c = 2$ sebesar 0,111511606, dan juga mengalami penurunan pada banyak *cluster* berikutnya.

Jika dibandingkan antara kedua jarak, nilai MPC pada metode *Manhattan* cenderung lebih tinggi dibandingkan metode *Euclidean* untuk setiap banyak *cluster*. Sehingga menunjukkan bahwa metode *Manhattan* memberikan hasil *clustering* yang lebih baik dalam memisahkan data pada penelitian ini. Dengan demikian, penggunaan jarak *Manhattan* memberikan pengaruh terhadap hasil *clustering* yang dihasilkan oleh algoritma *Fuzzy C-Means*, yang ditunjukkan melalui peningkatan hasil MPC. Berdasarkan hasil di atas, banyak *cluster* yang paling optimal adalah $c = 2$, karena memiliki nilai MPC tertinggi pada kedua jarak.

Nilai MPC sebesar 0,111511606 memang paling tinggi di antara seluruh percobaan, tetapi belum mendekati 1. Sehingga, hasil *clustering* dengan $c = 2$ dan jarak *Manhattan* dapat dikatakan sebagai hasil terbaik dalam penelitian ini, tetapi belum dapat disebut sebagai pemisahan *cluster* yang sangat kuat. Hal ini dapat

terjadi karena distribusi data yang digunakan tidak terpisah secara jelas antar *cluster*, sehingga masih terdapat beberapa data yang berada di antara dua *cluster*.

Sehingga, hasil validasi menunjukkan bahwa $c = 2$ dengan jarak *Manhattan* merupakan hasil *clustering* terbaik. Hasil ini didukung oleh nilai MPC yang paling tinggi dibandingkan percobaan lainnya.

4.5 Evaluasi *Purity* Berdasarkan Label Aktual

Purity digunakan sebagai evaluasi tambahan karena data memiliki label aktual, yaitu *Stress Level*, yaitu dapat dibandingkan dengan hasil *clustering*. Label ini tidak digunakan dalam proses *clustering*, karena proses *clustering* hanya menggunakan variabel numerik yang telah ditentukan sebelumnya. Namun, setelah hasil *clustering* diperoleh, label aktual dapat digunakan sebagai pembanding untuk melihat sejauh mana hasil *cluster* memiliki kesesuaian dengan label asli pada data.

Evaluasi *Purity* digunakan untuk melihat label aktual yang paling banyak muncul pada setiap *cluster*. Jika dalam suatu *cluster* sebagian besar data memiliki label aktual yang sama, maka *cluster* tersebut dapat dikatakan memiliki tingkat kemurnian yang lebih baik. Sebaliknya, apabila dalam satu *cluster* terdapat campuran label aktual yang cukup merata, maka nilai *Purity* akan menjadi lebih rendah.

Secara umum, nilai *Purity* berada pada rentang 0 hingga 1. Semakin mendekati 1, maka hasil *clustering* semakin sesuai dengan label aktual. Perhitungan *Purity* dilakukan dengan mengambil jumlah label aktual terbanyak pada setiap *cluster*, kemudian menjumlahkannya dan membaginya dengan total data. Nilai *Purity* dihitung menggunakan Persamaan (2.11).

Evaluasi *Purity* dilakukan pada hasil *clustering* dengan $c = 2$. Karena pada label aktual *Stress Level* terdiri dari 3 kategori, yaitu *Low*, *Moderate*, dan *High*. Evaluasi *Purity* pada hasil *clustering Euclidean* dilakukan dengan membandingkan hasil *cluster* terhadap label aktual *Stress Level*. Untuk mengetahui kesesuaian antara *cluster* dan label aktual, terlebih dahulu dibuat tabulasi silang. Tabulasi ini digunakan untuk menentukan label mayoritas pada setiap *cluster* sebagai dasar perhitungan nilai *Purity*. Hasil tabulasi silang tersebut disajikan pada Tabel 4.23.

Tabel 4.23 Tabulasi Silang $c = 2$ dan Label Aktual pada Jarak *Euclidean*

<i>Cluster</i>	<i>Low</i>	<i>Moderate</i>	<i>High</i>	Label Mayoritas	Jumlah Mayoritas	Jumlah Data
<i>Cluster 1</i>	270	388	379	<i>Moderate</i>	388	1037
<i>Cluster 2</i>	27	286	650	<i>High</i>	650	963
Total	297	674	1029	-	1038	2000

Berdasarkan hasil tabulasi silang pada Tabel 4.23, diperoleh data sebagai berikut:

$$n_{1,1} = 270, n_{1,2} = 388, n_{1,3} = 379$$

$$n_{2,1} = 27, n_{2,2} = 286, n_{2,3} = 650$$

Nilai maksimum pada setiap *cluster* adalah:

$$\max\{270, 388, 379\} = 388$$

$$\max\{27, 286, 650\} = 650$$

Sehingga, nilai *Purity* untuk hasil *clustering* menggunakan Jarak *Euclidean* adalah:

$$Purity = \frac{1}{2000} [\max\{n_{1,1}, n_{1,2}, n_{1,3}\} + \max\{n_{2,1}, n_{2,2}, n_{2,3}\}]$$

$$Purity = \frac{1}{2000} [388 + 650]$$

$$Purity = \frac{1038}{2000}$$

$$Purity = 0,519$$

Berdasarkan hasil perhitungan nilai *Purity* pada hasil *clustering* menggunakan jarak *Euclidean* adalah 0,519. Nilai ini menunjukkan bahwa hasil *clustering* menggunakan jarak *Euclidean* memiliki tingkat kesesuaian sebesar 0,519 terhadap label aktual *Stess Level*.

Evaluasi *Purity* pada Hasil *Clustering Manhattan* dilakukan dengan membandingkan hasil *cluster* terhadap label aktual *Stress Level*. Label aktual yang digunakan terdiri dari *Low*, *Moderate*, dan *High*. Untuk mengetahui kesesuaian antara *cluster* dan label aktual, terlebih dahulu dibuat tabulasi silang. Tabulasi ini digunakan untuk menentukan label mayoritas pada setiap *cluster* sebagai dasar perhitungan nilai *Purity*. Hasil tabulasi silang tersebut disajikan pada Tabel 4.24.

Tabel 4.24 Tabulasi Silang $c = 2$ dan Label Aktual pada Jarak *Manhattan*

<i>Cluster</i>	<i>Low</i>	<i>Moderate</i>	<i>High</i>	Label Mayoritas	Jumlah Mayoritas	Jumlah Data
<i>Cluster 1</i>	290	458	268	<i>Moderate</i>	458	1016
<i>Cluster 2</i>	7	216	761	<i>High</i>	761	984
Total	297	674	1029	-	1219	2000

Berdasarkan hasil tabulasi silang pada Tabel 4.24, diperoleh data sebagai berikut:

$$n_{1,1} = 290, n_{1,2} = 458, n_{1,3} = 268$$

$$n_{2,1} = 7, n_{2,2} = 216, n_{2,3} = 761$$

Nilai maksimum pada setiap *cluster* adalah:

$$\max\{290, 458, 268\} = 458$$

$$\max\{7, 216, 761\} = 761$$

Sehingga, nilai *Purity* untuk hasil *clustering* menggunakan Jarak *Manhattan* adalah:

$$Purity = \frac{1}{2000} [\max\{n_{1,1}, n_{1,2}, n_{1,3}\} + \max\{n_{2,1}, n_{2,2}, n_{2,3}\}]$$

$$Purity = \frac{1}{2000} [268 + 761]$$

$$Purity = \frac{1219}{2000}$$

$$Purity = 0,610$$

Berdasarkan hasil perhitungan nilai *Purity* pada hasil *clustering* menggunakan jarak *Manhattan* adalah 0,610. Nilai ini menunjukkan bahwa hasil *clustering* menggunakan jarak *Manhattan* memiliki tingkat kesesuaian sebesar 0,610 terhadap label aktual *Stess Level*.

Evaluasi *Purity* dilakukan pada hasil *clustering* dengan $c = 3$. Karena pada label aktual *Stress Level* terdiri dari 3 kategori, yaitu *Low*, *Moderate*, dan *High*. Dengan menggunakan $c = 3$, hasil *cluster* dapat dibandingkan secara langsung dengan banyak kategori label aktual yang tersedia pada *dataset*. Evaluasi *Purity* dilakukan melalui tabulasi silang antara hasil *cluster* dan label aktual *Stress Level*. Tabulasi silang digunakan untuk mengetahui berapa banyak data dengan label *Low*, *Moderate*, dan *High* yang masuk ke masing-masing *cluster*. Setelah itu, label aktual yang jumlahnya paling besar pada setiap *cluster* diambil sebagai label mayoritas.

Evaluasi *Purity* pada hasil *clustering Euclidean* dilakukan dengan membandingkan hasil *cluster* terhadap label aktual *Stress Level*. Label aktual yang digunakan terdiri dari *Low*, *Moderate*, dan *High*. Untuk mengetahui kesesuaian antara *cluster* dan label aktual, terlebih dahulu dibuat tabulasi silang. Tabulasi ini digunakan untuk menentukan label mayoritas pada setiap *cluster* sebagai dasar perhitungan nilai *Purity*. Hasil tabulasi silang tersebut disajikan pada Tabel 4.25.

Berdasarkan hasil tabulasi silang pada Tabel 4.25, diperoleh data sebagai berikut:

$$n_{1,1} = 106, n_{1,2} = 258, n_{1,3} = 268$$

$$n_{2,1} = 183, n_{2,2} = 238, n_{2,3} = 213$$

$$n_{3,1} = 8, n_{3,2} = 183, n_{3,3} = 548$$

Nilai maksimum pada setiap *cluster* adalah:

$$\max\{106, 253, 268\} = 268$$

$$\max\{183, 238, 213\} = 238$$

$$\max\{8, 183, 548\} = 548$$

Sehingga, nilai *Purity* untuk hasil *clustering* menggunakan Jarak *Euclidean* adalah:

$$Purity = \frac{1}{2000} [\max\{n_{1,1}, n_{1,2}, n_{1,3}\} + \max\{n_{2,1}, n_{2,2}, n_{2,3}\} + \max\{n_{3,1}, n_{3,2}, n_{3,3}\}]$$

$$Purity = \frac{1}{2000} [268 + 238 + 548]$$

$$Purity = \frac{1054}{2000}$$

$$Purity = 0,527$$

Tabel 4.25 Tabulasi Silang $c = 3$ dan Label Aktual pada Jarak *Euclidean*

<i>Cluster</i>	<i>Low</i>	<i>Moderate</i>	<i>High</i>	Label Mayoritas	Jumlah Mayoritas	Jumlah Data
<i>Cluster 1</i>	106	253	268	<i>High</i>	268	627
<i>Cluster 2</i>	183	238	213	<i>Moderate</i>	238	634
<i>Cluster 3</i>	8	183	548	<i>High</i>	548	739
Total	297	674	1029	-	1054	2000

Berdasarkan hasil perhitungan nilai *Purity* pada hasil *clustering* menggunakan jarak *Euclidean* adalah 0,527. Nilai ini menunjukkan bahwa hasil *clustering* menggunakan jarak *Euclidean* memiliki tingkat kesesuaian sebesar 0,527 terhadap label aktual *Stess Level*.

Evaluasi *Purity* pada Hasil *Clustering Manhattan* dilakukan dengan membandingkan hasil *cluster* terhadap label aktual *Stress Level*. Label aktual yang digunakan terdiri dari *Low*, *Moderate*, dan *High*. Untuk mengetahui kesesuaian antara *cluster* dan label aktual, terlebih dahulu dibuat tabulasi silang. Tabulasi ini digunakan untuk menentukan label mayoritas pada setiap *cluster* sebagai dasar perhitungan nilai *Purity*. Hasil tabulasi silang tersebut disajikan pada Tabel 4.26.

Tabel 4.26 Tabulasi Silang $c = 3$ dan Label Aktual pada Jarak *Manhattan*

<i>Cluster</i>	<i>Low</i>	<i>Moderate</i>	<i>High</i>	Label Mayoritas	Jumlah Mayoritas	Jumlah Data
<i>Cluster 1</i>	176	264	257	<i>Moderate</i>	264	697
<i>Cluster 2</i>	2	208	630	<i>High</i>	630	740
<i>Cluster 3</i>	119	302	142	<i>Moderate</i>	302	563
Total	297	674	1029	-	1054	2000

Berdasarkan hasil tabulasi silang pada Tabel 4.26, diperoleh data sebagai berikut:

$$n_{1,1} = 176, n_{1,2} = 264, n_{1,3} = 257$$

$$n_{2,1} = 2, n_{2,2} = 108, n_{2,3} = 630$$

$$n_{3,1} = 119, n_{3,2} = 302, n_{3,3} = 142$$

Nilai maksimum pada setiap *cluster* adalah:

$$\max\{176, 264, 257\} = 264$$

$$\max\{2, 108, 630\} = 630$$

$$\max\{119, 302, 142\} = 302$$

Sehingga, nilai *Purity* untuk hasil *clustering* menggunakan Jarak *Manhattan* adalah:

$$Purity = \frac{1}{2000} [\max\{n_{1,1}, n_{1,2}, n_{1,3}\} + \max\{n_{2,1}, n_{2,2}, n_{2,3}\} + \max\{n_{3,1}, n_{3,2}, n_{3,3}\}]$$

$$Purity = \frac{1}{2000} [264 + 630 + 302]$$

$$Purity = \frac{1196}{2000}$$

$$Purity = 0,598$$

Berdasarkan hasil perhitungan nilai *Purity* pada hasil *clustering* menggunakan jarak *Manhattan* adalah 0,598. Nilai ini menunjukkan bahwa hasil *clustering* menggunakan jarak *Manhattan* memiliki tingkat kesesuaian sebesar 0,598 terhadap label aktual *Stess Level*. Untuk melihat perbandingan nilai *Purity* pada beberapa banyak *cluster*, yaitu $c = 2$, $c = 3$, $c = 4$, dan $c = 5$. Perbandingan ini dilakukan

untuk mengetahui banyak *cluster* yang menghasilkan tingkat kesesuaian paling baik terhadap label aktual. Hasil perbandingan nilai *Purity* tersebut disajikan pada Tabel 4.27.

Tabel 4.27 Hasil *Purity* $c = 2$ hingga $c = 5$

Banyak Cluster (c)	<i>Purity Euclidean</i>	<i>Purity Manhattan</i>
2	0,519	0,610
3	0,527	0,598
4	0,616	0,627
5	0,617	0,642

Berdasarkan hasil perhitungan *Purity*, jarak *Manhattan* menunjukkan nilai *Purity* yang lebih tinggi dibandingkan jarak *Euclidean* pada seluruh banyak *cluster* yang diuji. Pada $c = 2$, nilai *Purity* jarak *Manhattan* sebesar 0,610, sedangkan jarak *Euclidean* sebesar 0,519. Pada $c = 3$, nilai *Purity* jarak *Euclidean* sedikit meningkat menjadi 0,527, tetapi nilai jarak *Manhattan* menurun menjadi 0,598.

Pada *clustering* dengan banyak *cluster* $c = 2$ diperoleh hasil *cluster* yang didominasi oleh label aktual *Moderate* dan *High*. Ketika banyak *cluster* $c = 3$, ternyata masih didominasi pada label aktual *Moderate* dan *High*, dan tidak secara dominan merepresentasikan label aktual *Low*. Pemisahan yang terjadi lebih banyak terjadi pada label aktual *Moderate* dan *High*, sehingga penambahan banyak *cluster* tidak meningkatkan kesesuaian hasil *clustering* terhadap label aktual. Hal ini menyebabkan tingkat kesesuaian antara hasil *clustering* dan label aktual pada $c = 3$ lebih rendah dibandingkan $c = 2$.

Selanjutnya, pada $c = 4$ dan $c = 5$, kedua ukuran jarak menunjukkan peningkatan nilai *Purity*. Nilai *Purity* tertinggi diperoleh pada jarak *Manhattan* dengan $c = 5$, yaitu sebesar 0,642. Hal ini menunjukkan bahwa kombinasi jarak *Manhattan* dengan $c = 5$ menghasilkan pengelompokan yang paling sesuai dengan

label mayoritas dibandingkan banyak *cluster* lainnya. Penambahan banyak *cluster* mampu memisahkan data ke dalam kelompok yang lebih sesuai antara hasil *clustering* dan label aktual menjadi lebih baik.

4.6 Interpretasi Hasil *Clustering*

Berdasarkan hasil *clustering Fuzzy C-Means* dengan parameter $c = 2$, data gaya hidup mahasiswa terbagi ke dalam dua *cluster*. Pembagian *cluster* dilakukan berdasarkan nilai derajat keanggotaan tertinggi pada setiap data. Sehingga, apabila suatu data memiliki nilai keanggotaan terbesar pada *Cluster 1*, maka data tersebut dimasukkan ke dalam *Cluster 1*. Sebaliknya, apabila nilai keanggotaan terbesar berada pada *Cluster 2*, maka data tersebut dimasukkan ke dalam *Cluster 2*.

Berdasarkan hasil *clustering* menggunakan jarak *Euclidean* pada iterasi terakhir, yaitu iterasi ke-47, data gaya hidup mahasiswa terbagi ke dalam dua *cluster*. *Cluster 1* terdiri dari 1.037 data, sedangkan *Cluster 2* terdiri dari 963 data. Untuk mengetahui karakteristik masing-masing *cluster*, dilakukan perhitungan nilai rata-rata setiap variabel pada tiap *cluster*. Hasil rata-rata variabel tiap *cluster* menggunakan jarak *Euclidean* ditampilkan pada Tabel 4.28.

Tabel 4.28 Nilai Rata-rata Variabel Tiap *Cluster* pada Jarak *Euclidean*

Cluster	Variabel					
	X_1	X_2	X_3	X_4	X_5	X_6
Cluster 1	6,593828351	1,686403086	6,891996143	3,16509161	5,66268081	2,973741562
Cluster 2	8,425545171	2,317133956	8,157320872	2,208618899	2,891381101	3,269106957

Berdasarkan Tabel 4.28, menunjukkan adanya perbedaan nilai rata-rata setiap variabel pada masing-masing *cluster*. Perbedaan ini digunakan untuk menginterpretasikan karakteristik gaya hidup mahasiswa pada setiap *cluster*. *Cluster 1* memiliki rata-rata waktu belajar sebesar 6,593828351 jam per hari, kegiatan ekstrakurikuler sebesar 1,686403086 jam per hari, waktu tidur sebesar 6,891996143 jam per hari, aktivitas sosial sebesar 3,16509161 jam per hari,

aktivitas fisik sebesar 5,66268081 jam per hari, dan GPA sebesar 2,973741562. Jika dibandingkan dengan *Cluster 2*, *Cluster 1* memiliki nilai yang lebih tinggi pada variabel aktivitas sosial dan aktivitas fisik. Hal ini menunjukkan bahwa mahasiswa dalam *Cluster 1* cenderung lebih banyak menghabiskan waktu untuk bersosialisasi dan melakukan aktivitas fisik.

Namun, *Cluster 1* memiliki rata-rata waktu belajar, waktu tidur, kegiatan ekstrakurikuler, dan GPA yang lebih rendah dibandingkan *Cluster 2*. Sehingga, mahasiswa pada *Cluster 1* lebih menonjol pada aktivitas non-akademik, terutama aktivitas sosial dan fisik. Oleh karena itu, *Cluster 1* dapat diinterpretasikan sebagai kelompok mahasiswa dengan kecenderungan gaya hidup yang lebih aktif pada aktivitas sosial dan fisik.

Pada *Cluster 2* memiliki rata-rata waktu tidur dan kegiatan ekstrakurikuler yang lebih tinggi. Namun, aktivitas sosial dan aktivitas fisik pada *cluster* ini lebih rendah. Dengan demikian, *Cluster 2* dapat diinterpretasikan sebagai kelompok mahasiswa yang cenderung pada kegiatan akademik.

Selanjutnya, berdasarkan hasil *clustering* menggunakan jarak *Manhattan* pada iterasi terakhir, yaitu iterasi ke-58, data gaya hidup mahasiswa juga terbagi ke dalam dua *cluster*. *Cluster 1* terdiri dari 1.016 data, sedangkan *Cluster 2* terdiri dari 984 data. Nilai rata-rata setiap variabel pada masing-masing *cluster* menggunakan jarak *Manhattan* ditampilkan pada Tabel 4.29.

Tabel 4.29 Nilai Rata-rata Variabel Tiap *Cluster* pada Jarak *Manhattan*

Cluster	Variabel					
	X_1	X_2	X_3	X_4	X_5	X_6
Cluster 1	6,374311024	1,835433071	7,20019685	2,864271654	5,725787402	2,928297244
Cluster 2	8,613109756	2,149796748	7,812093496	2,539634146	2,885365854	3,30972561

Berdasarkan Tabel 4.29, menunjukkan adanya perbedaan nilai rata-rata setiap variabel pada masing-masing *cluster*. Perbedaan ini digunakan untuk

menginterpretasikan karakteristik gaya hidup mahasiswa pada setiap *cluster*. *Cluster 1* memiliki rata-rata waktu belajar sebesar 6,374311024 jam per hari, kegiatan ekstrakurikuler sebesar 1,835433071 jam per hari, waktu tidur sebesar 7,20019685 jam per hari, aktivitas sosial sebesar 2,864271654 jam per hari, aktivitas fisik sebesar 5,725787402 jam per hari, dan GPA sebesar 2,928297244. Jika dibandingkan dengan *Cluster 2*, *Cluster 1* memiliki rata-rata aktivitas fisik yang lebih tinggi. Aktivitas sosial pada *Cluster 1* juga sedikit lebih tinggi dibandingkan *Cluster 2*.

Namun, *Cluster 1* memiliki waktu belajar dan GPA yang lebih rendah. Sehingga, mahasiswa pada *Cluster 1* lebih banyak menonjol pada aktivitas fisik, tetapi tidak terlalu kuat pada aspek akademik. Dengan demikian, *Cluster 1* pada hasil *clustering* jarak *Manhattan* dapat diinterpretasikan sebagai kelompok mahasiswa dengan kecenderungan gaya hidup yang lebih aktif pada aktivitas sosial dan fisik.

Pada *Cluster 2* memiliki rata-rata waktu belajar sebesar 8,613109756 jam per hari, kegiatan ekstrakurikuler sebesar 2,149796748 jam per hari, waktu tidur sebesar 7,812093496 jam per hari, aktivitas sosial sebesar 2,539634146 jam per hari, aktivitas fisik sebesar 2,885365854 jam per hari, dan GPA sebesar 3,30972561. Nilai rata-rata waktu belajar dan GPA pada *Cluster 2* lebih tinggi dibandingkan *Cluster 1*. Sehingga, mahasiswa pada *Cluster 2* memiliki kecenderungan akademik yang lebih kuat. *Cluster 2* juga memiliki rata-rata waktu tidur yang lebih tinggi. Sebaliknya, aktivitas fisik pada *cluster* ini lebih rendah. Dengan demikian, *Cluster 2* dapat diinterpretasikan sebagai kelompok mahasiswa yang cenderung pada kegiatan akademik.

Jika dibandingkan antara hasil jarak *Euclidean* dan jarak *Manhattan*, kedua ukuran jarak ini menghasilkan pola karakteristik *cluster* yang relatif konsisten. Pada kedua ukuran jarak, *Cluster 1* cenderung menunjukkan aktivitas fisik dan aktivitas sosial yang lebih tinggi, sedangkan *Cluster 2* cenderung menunjukkan waktu belajar dan GPA yang lebih tinggi. Sehingga, hasil *clustering* yang terbentuk tidak hanya membagi data menjadi dua kelompok, tetapi juga menunjukkan pola gaya hidup mahasiswa yang dapat dijelaskan secara jelas.

Variabel yang digunakan pada visualisasi hasil *clustering*, yaitu *Study Hours Per Day* (X_1), *Physical Activity Per Day* (X_5), dan GPA (X_6), terlihat bahwa selisih nilai rata-rata antara *Cluster 1* dan *Cluster 2* pada jarak *Manhattan* lebih besar dibandingkan jarak *Euclidean*. Selisih rata-rata *Study Hours Per Day* (X_1) pada jarak *Manhattan* sebesar 2,239 jam, sedangkan pada jarak *Euclidean* sebesar 1,832 jam. Pada variabel *Physical Activity Per Day* (X_5), selisih rata-rata antar *cluster* pada jarak *Manhattan* sebesar 2,840 jam, lebih besar dibandingkan jarak *Euclidean* sebesar 2,771 jam. Pada variabel GPA (X_6), selisih rata-rata antar *cluster* pada jarak *Manhattan* 0,381, sedangkan pada jarak *Euclidean* sebesar 0,295. Selisih rata-rata yang lebih besar pada ketiga variabel tersebut menunjukkan bahwa karakteristik *Cluster 1* dan *Cluster 2* lebih berbeda ketika menggunakan jarak *Manhattan*.

Berdasarkan hasil visualisasi *clustering* menggunakan variabel X_1 , X_5 , dan X_6 , kedua *cluster* lebih terpisah pada penggunaan jarak *Manhattan* dibandingkan jarak *Euclidean*. Selisih rata-rata yang lebih besar menunjukkan bahwa karakteristik rata-rata antara *Cluster 1* dan *Cluster 2* berbeda, sehingga pemisahan antar *cluster* terlihat lebih jelas pada visualisasi. Dengan demikian, penggunaan

jarak *Manhattan* menghasilkan pemisahan *cluster* yang lebih jelas dibandingkan jarak *Euclidean* pada variabel yang digunakan dalam visualisasi.

4.7 Pencarian Solusi Pengelompokan yang Lebih Baik dalam Prespektif

Islam

Dalam perspektif Islam, *ijtihad* dapat dipahami sebagai usaha sungguh-sungguh yang dilakukan untuk mencari solusi atas suatu permasalahan melalui proses berpikir dan analisis yang mendalam. Bukan hanya dilihat dari hasil akhirnya saja, tetapi juga proses pencarian solusi tersebut dilakukan secara maksimal dan penuh tanggung jawab. Hal ini selaras dengan ajaran Rasulullah SAW yang menekankan pentingnya usaha dalam mencapai kebenaran. Selain itu, merujuk pada Bab II makna dari Surah Al-Ankabut ayat 69 yang menunjukkan bahwa setiap kesungguhan dalam berusaha akan mengantarkan pada jalan yang lebih baik.

Dalam penelitian ini, konsep *ijtihad* terlihat dari proses yang tidak langsung menggunakan satu pendekatan sebagai solusi. Pada tahap awal, metode *Fuzzy C-Means* diterapkan menggunakan jarak *Euclidean* sebagai pendekatan yang umum digunakan. Namun, dari hasil awal tersebut muncul pertimbangan bahwa kualitas hasil *clustering* masih bisa ditingkatkan, terutama karena data yang digunakan memiliki *outlier*. Sehingga, dilakukan usaha, yaitu menggunakan jarak *Manhattan*.

Pemilihan ini didasarkan pada perbedaan karakteristik kedua ukuran jarak, di mana jarak *Euclidean* menghitung jarak menggunakan kuadrat selisih dari setiap variabel, sehingga perbedaan nilai yang besar akan memberikan pengaruh yang lebih besar dalam proses perhitungan. Sedangkan, pada jarak *Manhattan* perhitungan dilakukan menggunakan nilai absolut dari selisih data terlebih dahulu

sebelum hasil akhirnya dikuadratkan. Perbedaan cara perhitungan tersebut menyebabkan kedua ukuran jarak memberikan hasil yang berbeda terhadap data yang memiliki *outlier*. Perbedaan ini menjadi dasar untuk melihat kemungkinan kualitas hasil *clustering* yang lebih baik.

Proses ini tidak berhenti pada satu tahap. Banyak *cluster* juga tidak langsung ditentukan secara pasti, tetapi dicoba pada beberapa kemungkinan, yaitu dari $c = 2$ sampai $c = 5$. Setiap hasil yang diperoleh kemudian dievaluasi menggunakan nilai *Modified Partition Coefficient* (MPC) untuk melihat kualitas hasil *clustering* yang dihasilkan. Penggunaan MPC dalam penelitian ini menjadi bagian dari proses evaluasi untuk melihat apakah ukuran jarak dan banyak *cluster* yang digunakan sudah menghasilkan kualitas hasil *clustering* yang lebih baik. Dari proses ini, penggunaan jarak *Manhattan* secara konsisten menghasilkan nilai MPC yang lebih tinggi dibandingkan penggunaan jarak *Euclidean* pada setiap banyak *cluster* yang diuji. Hal ini menunjukkan bahwa kualitas hasil *clustering* yang diperoleh menjadi lebih baik.

Proses yang dilakukan tidak bersifat instan, tetapi melalui percobaan, membandingkan, dan mengevaluasi untuk memperoleh hasil yang lebih baik. Proses ini juga tidak berhenti pada satu pendekatan, tetapi terus mencari alternatif lain dan menilai kembali hasil yang diperoleh.

Hal ini sejalan dengan firman Allah SWT dalam Surah Al-Insyirah ayat 7:

فَإِذَا فَرَغْتَ فَانصَبْ ۖ

"Apabila engkau telah selesai (dengan suatu kebajikan), teruslah bekerja keras (untuk kebajikan yang lain)" (Q.S. Al-Insyirah: 7)

Ayat tersebut menekankan pentingnya terus berusaha dan tidak berhenti setelah menyelesaikan satu tahap pekerjaan. Dalam konteks penelitian ini,

perpindahan dari satu pendekatan ke pendekatan yang lain serta adanya proses evaluasi menunjukkan adanya usaha yang dilakukan secara berkelanjutan.

Dengan demikian, penelitian ini mencerminkan nilai *ijtihad* melalui proses yang dijalani, yaitu usaha yang dilakukan secara bertahap, tidak langsung mengambil keputusan, serta adanya evaluasi untuk mendapatkan solusi yang lebih baik.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan seluruh proses penelitian yang sudah dilakukan, terlihat bahwa penggunaan jarak dalam algoritma *Fuzzy C-Means* memberikan pengaruh terhadap hasil *clustering* yang diperoleh. Perbedaan ukuran jarak menyebabkan nilai jarak antara data dan pusat *cluster* menjadi berbeda, sehingga nilai derajat keanggotaan yang dihasilkan juga akan berubah. Perubahan pada nilai derajat keanggotaan memberikan pengaruh terhadap perhitungan pusat *cluster* pada iterasi berikutnya. Proses ini berlangsung secara berulang hingga mencapai kondisi konvergen, sehingga menghasilkan hasil *clustering* yang berbeda. Penggunaan ukuran jarak tidak hanya memengaruhi satu tahap perhitungan, tetapi pada keseluruhan proses hingga hasil akhir *clustering*.

Dari hasil yang diperoleh, nilai MPC pada metode *Manhattan* cenderung lebih tinggi dibandingkan *Euclidean* untuk setiap banyak *cluster* yang diuji. Hal ini menunjukkan bahwa penggunaan jarak *Manhattan* membuat pengelompokan data ke dalam *cluster* menjadi lebih jelas, sehingga kualitas hasil *clustering* menjadi lebih baik. Dari beberapa percobaan banyak *cluster* yang dilakukan, yaitu $c = 2$ hingga $c = 5$, diperoleh bahwa nilai MPC tertinggi berada pada $c = 2$. Artinya, banyak *cluster* tersebut merupakan yang paling sesuai untuk data yang digunakan dalam penelitian ini. Namun, nilai MPC belum mendekati 1, sehingga hasil *clustering* yang diperoleh dapat dikatakan sebagai hasil terbaik secara relatif, tetapi belum menunjukkan pemisahan *cluster* yang sangat kuat.

Selain menggunakan MPC, evaluasi juga dilakukan dengan *Purity* berdasarkan label aktual *Stress Level*. Hasil evaluasi *Purity* menunjukkan bahwa jarak *Manhattan* memiliki nilai *Purity* yang lebih tinggi dibandingkan jarak *Euclidean* pada seluruh banyak *cluster* yang diuji. Nilai *Purity* tertinggi diperoleh pada jarak *Manhattan* dengan $c = 5$, yaitu sebesar 0,642. Hasil ini menunjukkan bahwa jarak *Manhattan* juga lebih baik dalam menghasilkan *cluster* yang sesuai dengan label aktual. Namun, hasil *Purity* dan MPC memberikan penekanan yang berbeda. MPC menunjukkan kualitas struktur *fuzzy clustering* terbaik pada $c = 2$, sedangkan *Purity* menunjukkan kesesuaian tertinggi terhadap label aktual pada $c = 5$.

Dengan demikian, dapat disimpulkan bahwa penggunaan jarak *Manhattan* pada algoritma *Fuzzy C-Means* memberikan pengaruh terhadap proses dan hasil *clustering*, serta mampu menghasilkan *clustering* yang lebih baik dibandingkan dengan jarak *Euclidean*. Hal ini terlihat dari nilai MPC yang lebih tinggi, nilai *Purity* yang lebih baik, serta kemampuan jarak *Manhattan* dalam menghasilkan derajat keanggotaan yang lebih tegas pada data yang mengandung *outlier*.

5.2 Saran dan Penelitian Lanjutan

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa hal yang bisa menjadi saran untuk pengembangan selanjutnya, yaitu:

1. Penelitian ini menggunakan metode *Fuzzy C-Means* dengan dua jenis jarak, yaitu *Euclidean* dan *Manhattan*. Untuk penelitian selanjutnya, bisa menggunakan ukuran jarak yang berbeda atau metode *clustering* yang berbeda agar hasil yang diperoleh dapat dibandingkan dan dianalisis lebih luas.

2. Penelitian selanjutnya, disarankan untuk menggunakan metode validasi lain sebagai pembanding, sehingga hasil penentuan banyak *cluster* dapat lebih kuat dan tidak hanya bergantung pada satu indeks validasi.
3. Penelitian selanjutnya, disarankan menggunakan *dataset* yang memiliki *outlier* lebih banyak atau karakteristik *outlier* yang lebih jelas. Sehingga, pengaruh penggunaan jarak *Manhattan* terhadap kualitas hasil *clustering* dapat diamati dengan lebih jelas.

DAFTAR PUSTAKA

- Bezdek, J. C. (1974). Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, 1(1), 57–71. <https://doi.org/10.1007/BF02339490>
- Bhat, S. A., & Sheikh, B. A. (2023). Study Life Balance of College Students : Evidence from Higher Education Sector of UT of Jammu and Kashmir. *Journal of Learning and Educational Policy*, 03(03), 37–41. <https://doi.org/https://doi.org/10.55529/jlep.33.37.41>
- Christian, B., & Hakim, L. (2019). Penerapan Algoritma Fuzzy C-Means Pada Penentuan Lokasi Gudang Pendukung PT. XYZ. *Jurnal Teknologi Informasi AITI*, 16(1), 31–48. <https://doi.org/10.24246/aiti.v16i1.31-48>
- Dave, R. N. (1996). Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters*, 17(6), 613–623. [https://doi.org/10.1016/0167-8655\(96\)00026-8](https://doi.org/10.1016/0167-8655(96)00026-8)
- Erick. (2024). Analisis Perbandingan Algoritma K-Means Dan Fuzzy C-Means Dalam Pengelompokan Tipe Inflasi Kota Di Indonesia. *Jurnal Komputer Dan Informatika*, 19(1), 52–63.
- Hastini, L. Y., Fahmi, R., & Lukito, H. (2020). Apakah Pembelajaran Menggunakan Teknologi dapat Meningkatkan Literasi Manusia pada Generasi Z di Indonesia? *Jurnal Manajemen Informatika (JAMIKA)*, 10(1), 12–28. <https://doi.org/https://doi.org/10.34010/jamika.v10i1.2678>
- Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Cluster validity indices for automatic clustering: A comprehensive review. *Heliyon*, 11(2), e41953. <https://doi.org/10.1016/j.heliyon.2025.e41953>
- Johra, M. B. (2021). Soft Clustering Dengan Algoritma Fuzzy K-Means (Studi Kasus : Pengelompokan Desa Di Kota Tidore Kepulauan). *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 15(2), 385–392. <https://doi.org/10.30598/barekengvol15iss2pp385-392>
- Kaufman, L., & Rousseeuw, P. J. (1989). Finding Groups in Data: An Introduction to Cluster Analysis. In *Biometrics* (Vol. 47, Issue 2). <https://doi.org/10.2307/2532178>
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley. <https://doi.org/10.1002/9780470316801>
- Kementerian Agama. (2022). *Al-Qur'an dan Terjemahannya. Lajnah Pentashihan Mushaf Al-Qur'an Badan Litbang Kementerian Agama RI*.
- Klir, G. J., & Yuan, B. (1996). Fuzzy sets and fuzzy logic: Theory and applications. In *Endeavour* (Vol. 20, Issue 1). [https://doi.org/10.1016/s0160-9327\(96\)90083-6](https://doi.org/10.1016/s0160-9327(96)90083-6)
- Kusumadewi, S., Hartati, S., Harjoko, A., & Wardoyo, R. (2006). *Fuzzy Multi-Attribute Decision Making (Fuzzy MADM)*. Graha Ilmu.

- Kusumadewi, S., & Purnomo, H. (2004). *Aplikasi Logika Fuzzy Untuk Pendukung Keputusan*. Graha Ilmu.
- Laana, D. L., & Sukri, U. (2022). Life Style: Perilaku Mahasiswa Masa Kini Dan Pengaruh Media Sosial. *Inculco Journal of Christian Education*, 2(1), 67–80. <https://doi.org/10.59404/ijce.v2i1.27>
- Lubis, P. N., & Hasibuan, M. S. (2024). Pendekatan Metode DBSCAN dan Fuzzy C-Means untuk Klasterisasi Skala Prioritas Stunting. *Jurnal Pendidikan Dan Teknologi Indonesia*, 5(2), 301–320. <https://doi.org/https://doi.org/10.52436/1.jpti.655>
- Nadhifah, H. A., Sucipto, M. A. B., & Sudibyo, H. (2024). Tingkat Perilaku Konsumtif Generasi Z Pada Mahasiswa Program Studi Bimbingan Dan Konseling Universitas Pancasakti. *Jurnal Fokus Konseling*, 10(1), 17–27. <https://doi.org/10.52657/jfk.v10i1.2195>
- 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), 20–24. <https://doi.org/10.30591/jpit.v4i1.1253>
- Pal, N. R., & Bezdek, J. C. (1995). *On Cluster Validity for the Fuzzy c-Means Model*. 3(3).
- Puryanto, T., & Sutikno. (2016). Sistem Perencanaan Penambahan Stok Barang menggunakan Metode Fuzzy C-Means dan Fuzzy Tsukamoto (Studi Kasus di Distributor Alfamart Semarang). *Jurnal Nasional Teknologi Dan Sistem Informasi*, 2(2), 43–52. <https://doi.org/10.25077/teknosi.v2i2.2016.43-52>
- Rahakbauw, D. L., Ilwaru, V. Y. I., & M.H, H. (2017). *Implementasi Fuzzy C-Means Clustering Dalam Penentuan Beasiswa* (Vol. 11).
- Rahayu, P. W., Sudipa, I. G. I., Suryani, Surachman, A., Ridwan, A., Darmawiguna, I. G. M., Sutoyo, I. M. N., Slamet, D. I., Harlina, S., & Maysanjaya, I. M. D. (2024). *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia.
- Ramadhan, A., Efendi, Z., & Mustakim. (2017). Perbandingan K-Means dan Fuzzy C-Means untuk Pengelompokan Data User Knowledge Modeling. *Seminar Nasional Teknologi Informasi, Komunikasi Dan Industri*, 9(1), 219–226.
- Sahri, R. P., & Hasibuan, M. T. D. (2023). Hubungan Kualitas Tidur Dengan Prestasi Belajar Pada Mahasiswa. *Jurnal Penelitian Perawat Profesional*, 5, 473–480.
- Santosa, H. (2014). Aplikasi Penentuan Tarif Listrik Menggunakan Metode Fuzzy Sugeno. *Jurnal Sistem Informasi Bisnis*, 4(1), 28–39. <https://doi.org/10.21456/vol4iss1pp28-39>
- Sari, H. L., & Suranti, D. (2016). Perbandingan Algoritma Fuzzy C-Means (FCM) Dan Algoritma Mixture Dalam Penclustoran Data Curah Hujan Kota Bengkulu. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*, 7–15.
- Sianturi, C. J. M. (2015). Analisis Segmentasi Citra USG Hati Menggunakan

- Metode Fuzzy C- Mean. *Creative Information Technology Journal (CITEC Journal)*, 2(3), 256–264.
<https://doi.org/https://doi.org/10.24076/CITEC.2015V2I3.53>
- Sikhakolli, S. S., & Sikhakolli, A. K. (2023). *Effective Purity Method for Measuring the Clustering Accuracy and its Illustration*. 185(9), 28–33.
- Siringoringo, R., & Jamaluddin. (2019). Peningkatan Performa Cluster Fuzzy C-Means Pada Pengklasteran Sentimen Menggunakan Particleswarm Optimization. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 6(4), 349–354.
<https://doi.org/10.25126/jtiik.2018561090>
- Strauss, T., & Maltitz, M. J. von. (2017). Generalising ward’s method for use with manhattan distances. *PLoS ONE*, 12(1), 1–21.
<https://doi.org/10.1371/journal.pone.0168288>
- Suheri, M. E., Alimudin, Javaid, A., Ismail, M. T., & Ali, M. K. M. (2021). *Evaluation of Clustering Approach with Euclidean and Manhattan Distance for Outlier Detection*. 2423.
- Suleman, A. (2015). A new perspective of modified partition coefficient. *Pattern Recognition Letters*, 56, 1–6. <https://doi.org/10.1016/j.patrec.2015.01.008>
- Suryadi, J. (2024). Clustering Ketidakkukupan Konsumsi Pangan Per Kabupaten/Kota Menggunakan Algoritma K-Means Dan Fuzzy C-Means. *Jurnal Komputer Dan Informatika*, 19(1), 30–35.
- Zadeh, L. . (1965). Fuzzy Sets. *Informtion and Control*, 8, 4525–4534.
<https://doi.org/10.1016/j.procs.2022.09.516>
- Žalik, K. R. (2008). An efficient k'-means clustering algorithm. *Pattern Recognition Letters*, 29(9), 1385–1391.
<https://doi.org/10.1016/j.patrec.2008.02.014>

LAMPIRAN

Lampiran 1 *Dataset* Penelitian

<https://docs.google.com/spreadsheets/d/1Mtk-64ctJ87zgI3lxU3ebA4334xtVXzF/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true>



Lampiran 2 Analisis Deskriptif

https://docs.google.com/spreadsheets/d/1Jv_9DZBt0WqysdYZxwN4Ecl2kxgwxAzr/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true



Lampiran 3 *Dataset* setelah *Preprocessing*

<https://docs.google.com/spreadsheets/d/1CcO7FvN5sxK0XlFvp5FcTYWWpa38FdHb/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true>



Lampiran 4 Hasil *Clustering* untuk $c = 2$ Menggunakan Jarak *Euclidean*

<https://docs.google.com/spreadsheets/d/1TkL5rAtjp7X-E-ruh1hMO7vilJwALE-d/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true>



Lampiran 5 Hasil *Clustering* untuk $c = 3$ Menggunakan Jarak *Euclidean*

https://docs.google.com/spreadsheets/d/176OW6Rfa4_nKJwCvSEOTqXdf25aReEHC/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true

**Lampiran 6** Hasil *Clustering* untuk $c = 4$ Menggunakan Jarak *Euclidean*

https://docs.google.com/spreadsheets/d/17XKuLhjoLTUieROtt_pZeNeUKDHi48MP/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true

**Lampiran 7** Hasil *Clustering* untuk $c = 5$ Menggunakan Jarak *Euclidean*

<https://docs.google.com/spreadsheets/d/1CY0pqeRcOQSOGRqgElzlrPaTMXFSVwF5/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true>

**Lampiran 8** Hasil *Clustering* untuk $c = 2$ Menggunakan Jarak *Manhattan*

<https://docs.google.com/spreadsheets/d/1WwpUT7CBIsVB-ck4psKsXhUsir7K-Pez/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true>



Lampiran 9 Hasil *Clustering* untuk $c = 3$ Menggunakan Jarak *Manhattan*

<https://docs.google.com/spreadsheets/d/13YW2PjpS6-03uC4HtWQqxZgJky0VCNoh/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true>

**Lampiran 10** Hasil *Clustering* untuk $c = 4$ Menggunakan Jarak *Manhattan*

<https://docs.google.com/spreadsheets/d/12PNGOOnCgir-ZGivYbiU8uUjMenCQy7x/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true>

**Lampiran 11** Hasil *Clustering* untuk $c = 5$ Menggunakan Jarak *Manhattan*

<https://docs.google.com/spreadsheets/d/1TaAREz6ApaZCkKHyM7XDYG1zf8nJ5ps/edit?usp=sharing&ouid=104539981431764612973&rtpof=true&sd=true>

**Lampiran 12** Source Code *FCM* Menggunakan Jarak Euclidean

```
import numpy as np
import pandas as pd
from IPython.display import display
import matplotlib.pyplot as plt

data = pd.read_csv("/content/Data Normalisasi.csv", usecols=[0, 1, 2, 3, 4, 5]).to_numpy().astype(float)

c = 2 # Menyesuaikan dengan c yang di uji
w = 2
max_iter = 1000
epsilon = 0.00001

n = data.shape[0]
n_features = data.shape[1]
```

```

np.random.seed(42)

U = np.random.rand(n, c)
U = U / np.sum(U, axis=1, keepdims=True)

n_clusters = c

U_initial = U.copy()

def update_centers(U, data_subset, w):
    um = U ** w
    denominator = np.sum(um.T, axis=1)[:, None] + 1e-10
    centers = (um.T @ data_subset) / denominator
    return centers

def compute_objective(U, data_subset, centers, w):
    obj = 0
    for i in range(data_subset.shape[0]):
        for k in range(centers.shape[0]):
            dist_sq = np.sum((data_subset[i] - centers[k]) ** 2)
            obj += (U[i][k] ** w) * dist_sq
    return obj

def update_membership_from_formula(data, centers, w):
    N = data.shape[0]
    C = centers.shape[0]
    new_U = np.zeros((N, C))

    for i in range(N):
        dist_sq = np.array([np.sum((data[i] - centers[k]) ** 2) for k in range(C)]) + 1e-10
        numerator = dist_sq ** (-1 / (w - 1))
        denominator = np.sum(numerator)
        new_U[i] = numerator / denominator

    return new_U

centers_history = []
objective_history = []
U_history = []

print("=== PROSES ITERASI FCM DIMULAI ===\n")

for iteration in range(max_iter):

```

```

centers = update_centers(U, data, w)
centers_history.append(centers)

obj = compute_objective(U, data, centers, w)
objective_history.append((iteration + 1, obj))

U_new = update_membership_from_formula(data, centers, w)
U_history.append(U_new)

if iteration > 0 and abs(obj - objective_history[-2][1]) < epsilon:
    print(f"Konvergen pada iterasi ke-{iteration + 1}\n")
    U = U_new.copy()
    Break

U = U_new.copy()

print("\nMatriks Partisi Awal (Random):")
df_U_initial = pd.DataFrame(U_initial, columns=[f"μik{i+1}" for i in
range(n_clusters)])
df_U_initial.index = [f"Data {i+1}" for i in range(U_initial.shape[0])]
display(df_U_initial)

df_objective = pd.DataFrame(objective_history, columns=["Iterasi", "Fungsi
Objektif"])

U_final = pd.DataFrame(U, columns=[f"μik{i+1}" for i in range(n_clusters)])
U_final["Cluster Dominan"] = U_final.values.argmax(axis=1) + 1

print("\nFungsi Objektif:")
display(df_objective)

print("\nMatriks Partisi Akhir:")
display(U_final)

print("\nPusat Cluster Akhir:")
print(pd.DataFrame(centers, columns=[f"Fitur_{i+1}" for i in
range(n_features)]))

print("\nFungsi Objektif Terakhir:", df_objective.iloc[-1, 1])

output_file = "Hasil FCM EUCLIDEAN CLUSTER 2.xlsx"

with pd.ExcelWriter(output_file, engine='openpyxl') as writer:

```

```

df_U_initial = pd.DataFrame(U_initial, columns=[f"μ_ik{i+1}" for i in
range(n_clusters)])
df_U_initial.index = [f"Data {i+1}" for i in range(U_initial.shape[0])]
df_U_initial.to_excel(writer, sheet_name="U_Awal")

for i, (U_iter, centers_iter, obj_iter) in enumerate(zip(U_history,
centers_history, objective_history), start=1):

    df_U = pd.DataFrame(U_iter, columns=[f"μ_ik{j+1}" for j in
range(n_clusters)])
    df_U.index = [f"Data {idx+1}" for idx in range(U_iter.shape[0])]

    df_cent = pd.DataFrame(centers_iter, columns=[f"Fitur_{k+1}" for k in
range(n_features)])
    df_cent.index = [f"Cluster {j+1}" for j in range(n_clusters)]

    df_obj = pd.DataFrame({"Iterasi": [obj_iter[0]], "Fungsi Objektif":
[obj_iter[1]]})

    startrow = 0
    df_obj.to_excel(writer, sheet_name=f"Iterasi_{i}", startrow=startrow,
index=False)
    startrow += len(df_obj) + 2

    df_cent.to_excel(writer, sheet_name=f"Iterasi_{i}", startrow=startrow)
    startrow += len(df_cent) + 2

    df_U.to_excel(writer, sheet_name=f"Iterasi_{i}", startrow=startrow)

print(f"\nFile Excel tersimpan sebagai: {output_file}")

from google.colab import files
files.download(output_file)

```

Lampiran 13 Source Code FCM untuk Menggunakan Jarak Manhattan

```

import numpy as np
import pandas as pd
from IPython.display import display
import matplotlib.pyplot as plt

data = pd.read_csv("/content/Data Normalisasi.csv", usecols=[0, 1, 2, 3, 4,
5]).to_numpy().astype(float)

c = 2 # Menyesuaikan dengan c yang di uji

```

```

w = 2
max_iter = 1000
epsilon = 0.00001
n_clusters = c
n_features = data.shape[1]

np.random.seed(42)

n = data.shape[0]
U = np.random.rand(n, c)
U = U / np.sum(U, axis=1, keepdims=True)

U_initial = U.copy()
def update_centers(U, data_subset, w):
    um = U ** w
    denominator = np.sum(um.T, axis=1)[:, None] + 1e-10
    centers = (um.T @ data_subset) / denominator
    return centers

def compute_objective_manhattan(U, X, P, w):
    n = X.shape[0]
    c = P.shape[0]
    obj = 0.0

    for i in range(n):
        for k in range(c):
            dist = np.sum(np.abs(X[i] - P[k]))
            obj += (dist ** 2) * (U[i, k] ** w)

    return obj

def update_membership_manhattan(data, centers, w):
    N = data.shape[0]
    C = centers.shape[0]
    new_U = np.zeros((N, C))

    for i in range(N):
        dist = np.array([np.sum(np.abs(data[i] - centers[k])) for k in range(C)]) +
1e-10

        exponent = -2.0 / (w - 1.0)
        numerator = dist ** exponent
        denom = np.sum(numerator)

        new_U[i] = numerator / (denom + 1e-10)

    row_sums = new_U.sum(axis=1, keepdims=True) + 1e-12

```

```

new_U = new_U / row_sums
return new_U

centers_history = []
objective_history = []
U_history = []

for iteration in range(max_iter):
    centers = update_centers(U, data, w)
    centers_history.append(centers)

    obj = compute_objective_manhattan(U, data, centers, w)
    objective_history.append((iteration + 1, obj))
    U_new = update_membership_manhattan(data, centers, w)
    U_history.append(U_new)

    if iteration > 0:
        if abs(objective_history[-1][1] - objective_history[-2][1]) < epsilon:
            print(f"Konvergen pada iterasi ke-{iteration + 1} (Objektif Stabil)")
            U = U_new.copy()
            Break

    U = U_new.copy()

print("\nMatriks Partisi Awal (Random):")
df_U_initial = pd.DataFrame(U_initial, columns=[f"μik{i+1}" for i in range(n_clusters)])
df_U_initial.index = [f"Data {i+1}" for i in range(U_initial.shape[0])]
display(df_U_initial)

iter_data = []
for i, center in enumerate(centers_history, 1):
    for j, c_ in enumerate(center, 1):
        iter_data.append([i, j, *c_])

df_centers = pd.DataFrame(iter_data, columns=["Iterasi", "Cluster",
*[f"Fitur_{i+1}" for i in range(n_features)]])
df_objective = pd.DataFrame(objective_history, columns=["Iterasi", "Fungsi Objektif"])

U_final = pd.DataFrame(U, columns=[f"μik{i+1}" for i in range(n_clusters)])
U_final.index = [f"Data {i+1}" for i in range(U.shape[0])]
U_final["Cluster Dominan"] = U_final.values.argmax(axis=1) + 1

print("\nPusat Cluster per Iterasi:")
display(df_centers)

```

```

print("\nFungsi Objektif per Iterasi:")
display(df_objective)

print("\nMatriks Partisi Akhir:")
display(U_final)

print("\nPusat Cluster Akhir:")
print(pd.DataFrame(centers, columns=[f"Fitur_{i+1}" for i in
range(n_features)]))

print("\nFungsi Objektif Terakhir:", df_objective.iloc[-1, 1])

output_file = "Hasil FCM MANHATTAN CLUSTER 2.xlsx"

with pd.ExcelWriter(output_file, engine='openpyxl') as writer:

    df_U_initial = pd.DataFrame(U_initial, columns=[f"μ_ik{i+1}" for i in
range(n_clusters)])
    df_U_initial.index = [f"Data {i+1}" for i in range(U_initial.shape[0])]
    df_U_initial.to_excel(writer, sheet_name="U_Awal")

    for i, (U_iter, centers_iter, obj_iter) in enumerate(zip(U_history,
centers_history, objective_history), start=1):

        df_U = pd.DataFrame(U_iter, columns=[f"μ_ik{j+1}" for j in
range(n_clusters)])
        df_U.index = [f"Data {idx+1}" for idx in range(U_iter.shape[0])]

        df_centers_iter = pd.DataFrame(centers_iter, columns=[f"Fitur_{k+1}" for
k in range(n_features)])
        df_centers_iter.index = [f"Cluster {j+1}" for j in range(n_clusters)]

        df_obj = pd.DataFrame({"Iterasi": [obj_iter[0]], "Fungsi Objektif":
[obj_iter[1]]})

        startrow = 0
        df_obj.to_excel(writer, sheet_name=f"Iterasi_{i}", startrow=startrow,
index=False)
        startrow += len(df_obj) + 2
        df_centers_iter.to_excel(writer, sheet_name=f"Iterasi_{i}",
startrow=startrow)
        startrow += len(df_centers_iter) + 2
        df_U.to_excel(writer, sheet_name=f"Iterasi_{i}", startrow=startrow)

    print(f"\nFile Excel tersimpan sebagai: {output_file}")

from google.colab import files

```

```
files.download(output_file)
```

Lampiran 14 Source Code Visualisasi Hasil Clustering

```
import pandas as pd
import matplotlib.pyplot as plt
# Membaca data
df = pd.read_excel('PLOT MANHATTAN C2.xlsx')

# Memisahkan cluster
cluster1 = df[df['Cluster'] == 'Cluster 1']
cluster2 = df[df['Cluster'] == 'Cluster 2']

# Membuat figure
fig = plt.figure(figsize=(12,8))
ax = fig.add_subplot(111, projection='3d')

# Mengecilkan area plot agar label tidak terpotong
ax.set_position([0.08, 0.08, 0.72, 0.72])

# Plot Cluster 1
ax.scatter(
    cluster1['Study Hours Per Day'],
    cluster1['Physical Activity Hours Per Day'],
    cluster1['GPA'],
    s=12,
    alpha=0.6,
    label='Cluster 1'
)

# Plot Cluster 2
ax.scatter(
    cluster2['Study Hours Per Day'],
    cluster2['Physical Activity Hours Per Day'],
    cluster2['GPA'],
    s=12,
    alpha=0.6,
    label='Cluster 2'
)

# Label sumbu
ax.set_xlabel('Study Hours', fontsize=11, labelpad=10)
ax.set_ylabel('Physical Activity', fontsize=11, labelpad=10)
ax.set_zlabel('GPA', fontsize=11, labelpad=20)

# Sudut pandang
ax.view_init(elev=20, azim=130)
```

```
# Legend
ax.legend(
    loc='upper left',
    bbox_to_anchor=(1.02, 0.95),
    fontsize=10
)

# Simpan gambar kualitas skripsi
plt.savefig(
    'Plot_FCM_C2.png',
    dpi=600
)

plt.show()
```

RIWAYAT HIDUP



Ayu Habibatun Nurwahyu Zainuri, lahir di Nganjuk pada tanggal 15 Juni 2004. Penulis merupakan anak pertama dari Ayah Nurwakid Zainuri dan Ibu Sri Wahyuningsih serta kakak dari Adik La'Salina Zahro Zainuri dan Adik Levina Salsabila Zainuri. Penulis telah menempuh pendidikan mulai dari TK Pertiwi Kacangan yang lulus pada tahun 2010, kemudian melanjutkan pendidikan sekolah dasar di SDN Jatirejo 1 dan lulus pada tahun 2016. Kemudian penulis melanjutkan pendidikan sekolah menengah pertama di SMPN 2 Nganjuk dan lulus pada tahun 2019. Setelah itu, melanjutkan pendidikan sekolah menengah atas di SMAN 1 Berbek dengan peminatan MIPA dan lulus pada tahun 2022. Pada tahun yang sama, penulis melanjutkan pendidikan di Universitas Islam Negeri Maulana Malik Ibrahim Malang pada Program Studi Matematika, Fakultas Sains dan Teknologi. Selama menempuh pendidikan di perguruan tinggi, pernah mengikuti kepanitiaan menjadi panitia SIGMA, yaitu kegiatan orientasi mahasiswa baru Program Studi Matematika untuk Angkatan 2023. Selain itu, penulis juga melaksanakan Praktik Kerja Lapangan (PKL) secara mandiri di PT PLN (Persero) UID Jawa Timur UP3 Malang pada bidang logistik. Penulis juga beberapa kali mengikuti kegiatan volunteer PeduliLindungi sebagai bentuk partisipasi dalam kegiatan sosial dan pelayanan masyarakat.



**KEMENTERIAN AGAMA RI
UNIVERSITAS ISLAM NEGERI
MAULANA MALIK IBRAHIM MALANG
FAKULTAS SAINS DAN TEKNOLOGI**

Jl. Gajayana No.50 Dinoyo Malang Telp. / Fax. (0341)558933

BUKTI KONSULTASI SKRIPSI

Nama : Ayu Habibatul Nurwahyu Zainuri
NIM : 220601110054
Fakultas / Jurusan : Sains dan Teknologi / Matematika
Judul Skripsi : Implementasi Metode *Fuzzy C-Means* Berbasis Jarak
Manhattan dengan Validasi *Modified Partition Coefficient*
pada *Clustering* Data Gaya Hidup Mahasiswa
Pembimbing I : Prof. Dr. H. Turmudi, M.Si, Ph.D
Pembimbing II : Dr. Fachrur Rozi, M.Si

No	Tanggal	Hal	Tanda Tangan
1.	9 September 2025	Konsultasi Bab I, II, dan III	1.
2.	16 September 2025	Konsultasi Bab I, II, dan III	2.
3.	23 September 2025	Konsultasi Bab I, II, dan III	3.
4.	30 September 2025	Konsultasi Bab I, II, dan III	4.
5.	7 Oktober 2025	Konsultasi Bab I, II, dan III	5.
6.	8 Oktober 2025	ACC Bab I, II, dan III	6.
7.	9 Oktober 2025	Konsultasi Kajian Agama Bab I dan II	7.
8.	15 Oktober 2025	Konsultasi Kajian Agama Bab I dan II	8.
9.	24 Oktober 2025	Konsultasi Kajian Agama Bab I dan II	9.
10.	3 November 2025	ACC Kajian Agama Bab I dan II	10.
11.	18 November 2025	ACC Seminar Proposal	11.
12.	21 April 2026	Konsultasi Revisi Seminar Proposal	12.
13.	24 April 2026	Konsultasi Revisi Seminar Proposal	13.
14.	30 April 2026	Konsultasi Bab IV dan V	14.
15.	30 April 2026	Konsultasi Kajian Agama Bab IV	15.



**KEMENTERIAN AGAMA RI
UNIVERSITAS ISLAM NEGERI
MAULANA MALIK IBRAHIM MALANG
FAKULTAS SAINS DAN TEKNOLOGI**

Jl. Gajayana No.50 Dinoyo Malang Telp. / Fax. (0341)558933

16.	4 Mei 2026	Konsultasi Bab IV dan V	16.
17.	5 Mei 2026	Konsultasi Kajian Agama Bab IV	17.
18.	6 Mei 2026	ACC Bab IV dan V	18.
19.	8 Mei 2026	ACC Kajian Agama Bab IV	19.
20.	8 Mei 2026	ACC Seminar Hasil	20.
21.	18 Mei 2026	Seminar Hasil	21.
22.	25 Mei 2026	Konsultasi Revisi Seminar Hasil	22.
23.	26 Mei 2026	Konsultasi Revisi Seminar Hasil	23.
24.	4 Juni 2026	ACC Sidang Skripsi	24.
25.	12 Juni 2026	ACC Keseluruhan	25.

Malang, 12 Juni 2026

Mengetahui,

Ketua Program Studi Matematika



Dr. Fachrud Rozi, M.Si

NIP. 19800527 200801 1 012