



UNIVERSITAS
GADJAH MADA

Final Project Metode Optimasi

Bagus Cipta Pratama – 23/516539/PA/22097

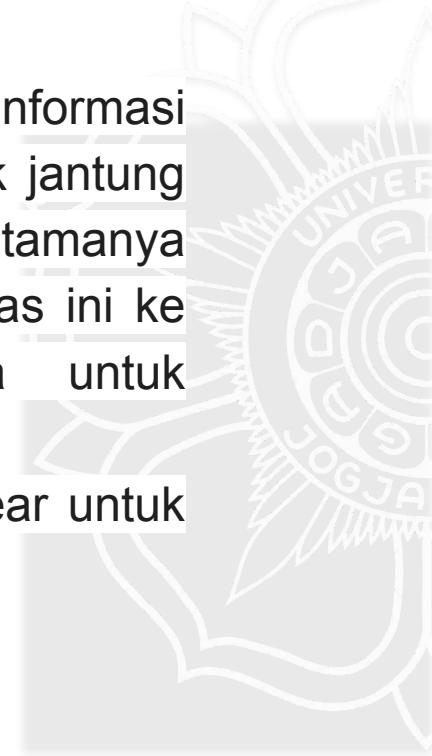
Ravie Arjun Nadhief – 23/522765/PA/22491

Nasya Putri Raudhah Dahlia – 23/513931/PA/21967



Pendahuluan

- Dataset aktivitas harian (*daily activity*) memuat informasi multidimensi seperti jarak, durasi, kecepatan, dan detak jantung yang polanya sulit dipahami secara manual. Masalah utamanya adalah bagaimana mengelompokkan ribuan data aktivitas ini ke dalam segmen-segmen (cluster) yang bermakna untuk memahami karakteristik pengguna.
- Secara matematis, ini adalah masalah optimasi non-linear untuk meminimalkan variansi di dalam setiap kelompok.



Proposed Method

Solusi yang diajukan adalah mengimplementasikan algoritma K-Means Clustering yang dibangun dari nol (from scratch) menggunakan metode optimasi Gradient Descent

Pendekatan ini tidak menggunakan solver bawaan library, melainkan secara eksplisit menghitung turunan gradien dari fungsi objektif Within-Cluster Sum of Squares (WCSS) untuk memperbarui posisi centroid secara iteratif hingga mencapai konvergensi optimal.



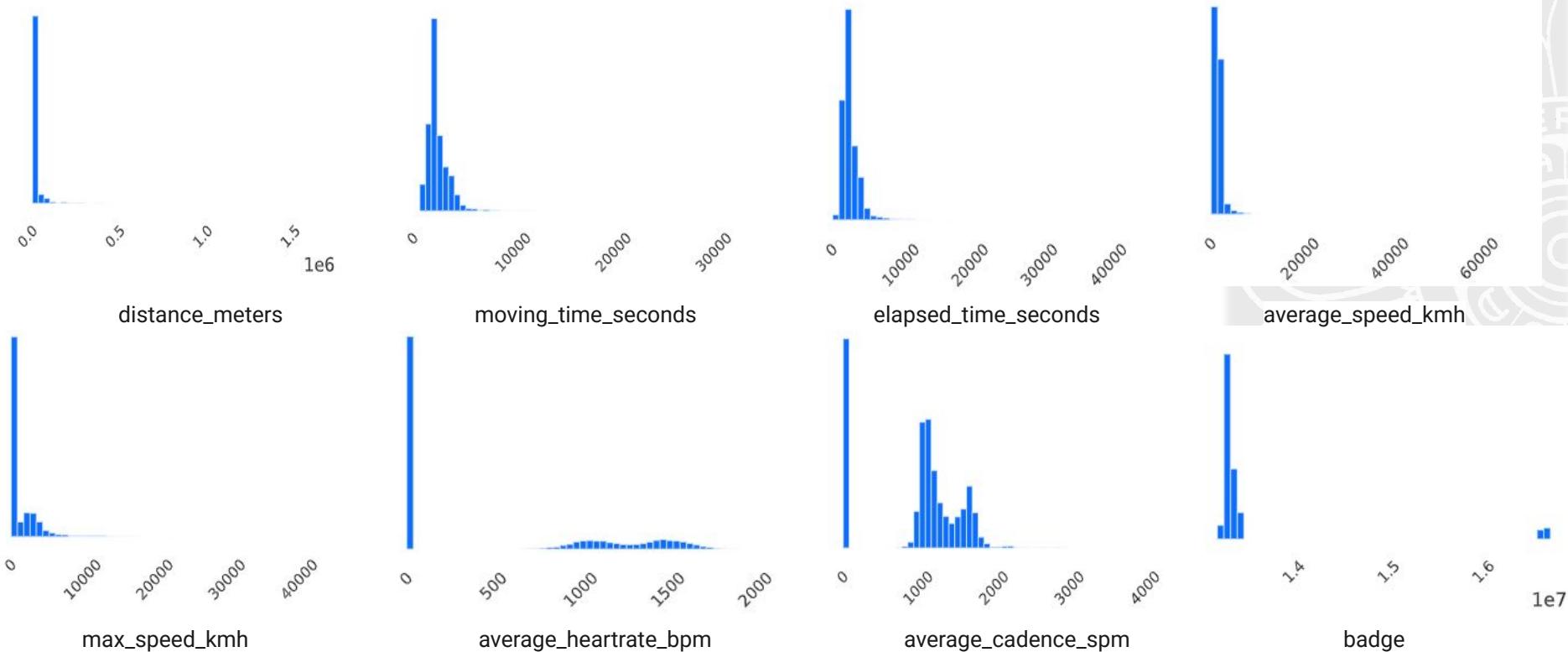


Exploratory Data Analysis

Sample 5 data pertama

distance_meters	moving_time_seconds	elapsed_time_seconds	type	start_date_utc	end_date_utc	average_speed_kmh	max_speed_kmh	average_heartrate_bpm	average_cadence_spm	badge
31260	2353	2484	Walk	4/12/25 10:33	4/12/25 11:15	1329	2773	0	0	13289547
26900	1375	1375	Walk	1/2/25 18:31	1/2/25 18:54	1956	0	1226	1022	13267699
30200	1521	1521	Walk	1/3/25 18:54	1/3/25 19:20	1986	0	1261	998	13267699
28500	1480	1480	Walk	1/4/25 9:07	1/4/25 9:31	1926	0	1292	998	13267699
26300	1286	1286	Walk	1/13/25 18:46	1/13/25 19:07	2045	0	1345	1002	13267699

Exploratory Data Analysis : Distribusi Fitur Numerik

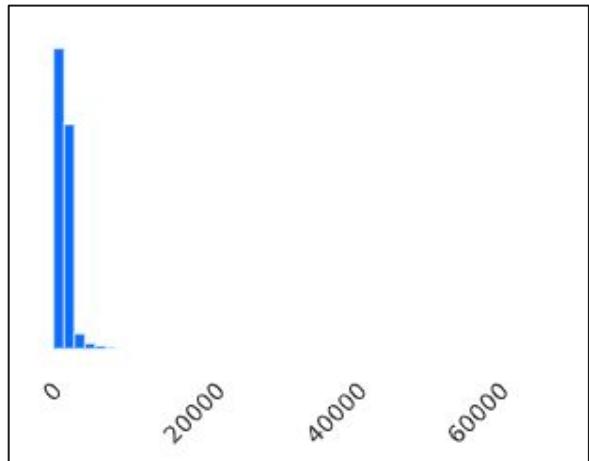


Exploratory Data Analysis

Masalah 1 :

distribusi tidak normal dan **outlier esktrem**

Solusi : Clipping dan normalisasi
menggunakan Yeo-Johnson transformation

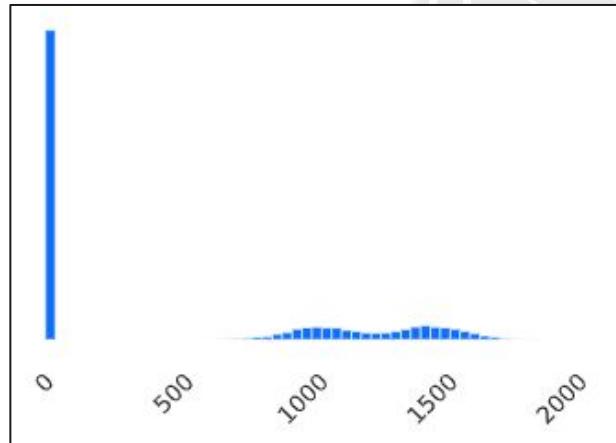


average_speed_kmh

Masalah 2 :

terdapat kolom yang didominasi **nilai 0**

Solusi : drop baris dengan fiitur 0



average_heartrate_bpm



Preprocessing dan feature engineering yang dilakukan

1. Ekstraksi Durasi
2. One-Hot Encoding kolom 'type'
3. Transformasi Fitur Siklis 'start_date', 'end_date'
4. Clipping + Normalisasi Yeo-Johnson

Objective Function

Within-Cluster Sum of Squares (WCSS). Loss Function ini didefinisikan sebagai jumlah kuadrat jarak antara tiap titik data dengan centroid terdekatnya

$$J(\mu) = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

Dimana:

- k : Jumlah cluster.
- C_j : Himpunan titik data yang termasuk dalam cluster ke- j .
- μ_j : Centroid dari cluster ke- j .
- $\|x_i - \mu_j\|^2$: Jarak Euclidean kuadrat.



Gradient Computation

Hitung gradient fungsi J terhadap setiap centroid. Gradien menunjukkan arah kenaikan terbesar dari fungsi error, sehingga untuk meminimalkan error, kita harus bergerak ke arah berlawanan (negative gradient).

Turunan parsial fungsi objektif terhadap μ_j adalah:

$$\nabla_{\mu_j} J = \frac{\partial}{\partial \mu_j} \sum_{x_i \in C_j} (x_i - \mu_j)^2$$

Dengan aturan rantai (*chain rule*), turunannya menjadi:

$$\nabla_{\mu_j} J = \sum_{x_i \in C_j} 2(\mu_j - x_i)$$

Persamaan ini dapat disederhanakan menjadi bentuk vektor yang lebih efisien secara komputasi:

$$\nabla_{\mu_j} J = 2 \left(N_j \mu_j - \sum_{x_i \in C_j} x_i \right)$$



Parameter

Learning Rate (α) : 0.01

Jumlah Inisialisasi Ulang (n_{init}) : 10

Batas Minimal Perubahan Loss (Tolerance / Δ Loss minimum) : 1×10^{-6}

Batas Minimal Nilai Loss untuk Konvergensi (Loss Threshold) : 1×10^{-3}



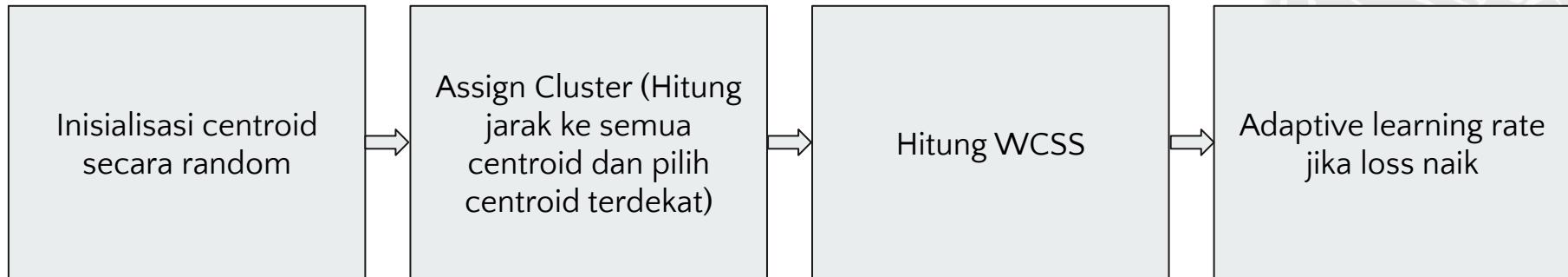
Update Rule

Algoritma Gradient Descent memperbarui posisi *centroid* dengan menggesernya ke arah negatif gradien dikalikan dengan laju pembelajaran (*learning rate* atau α).

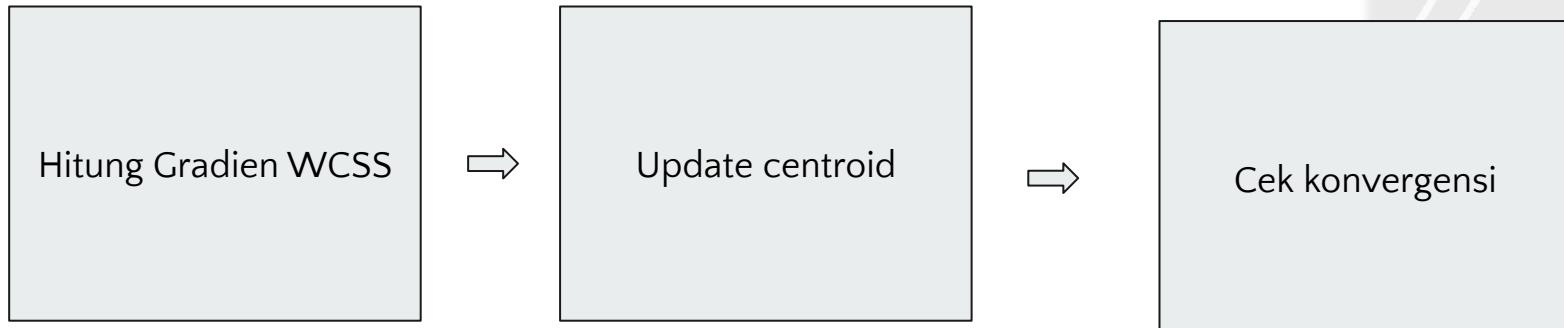
$$\mu_j^{(t+1)} = \mu_j^{(t)} - \alpha \cdot \nabla_{\mu_j} J$$



Steps of the Algorithm



Steps of the Algorithm

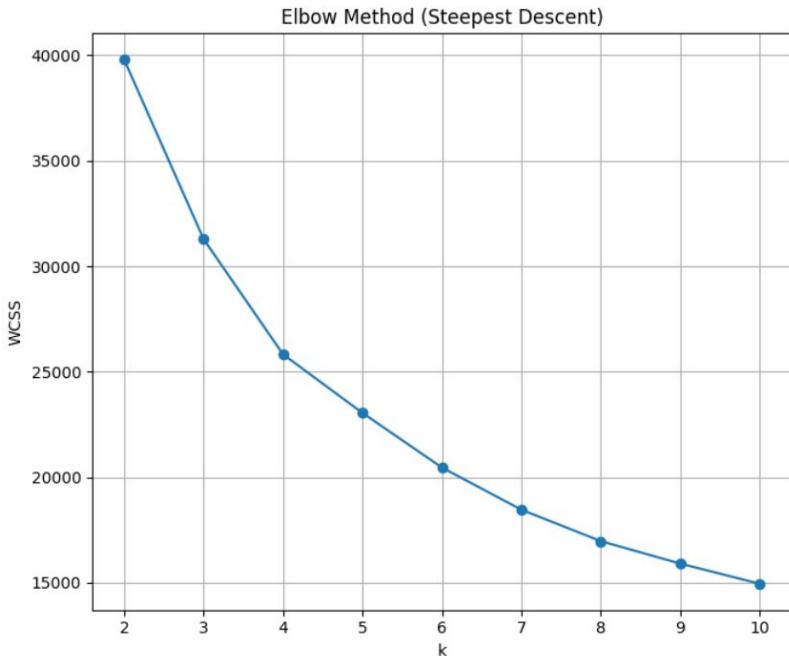


$$\nabla_{\mu_j} J = 2 \left(N_j \mu_j - \sum_{x_i \in C_j} x_i \right)$$

$$\mu_j^{(t+1)} = \mu_j^{(t)} - \alpha \cdot \nabla_{\mu_j} J$$

Clustering

setelah melalui proses sebelumnya didapatkan k optimal dengan nilai 4

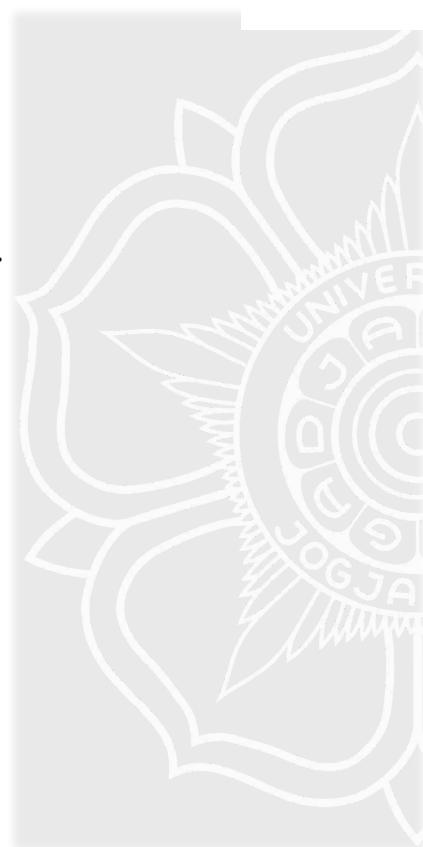
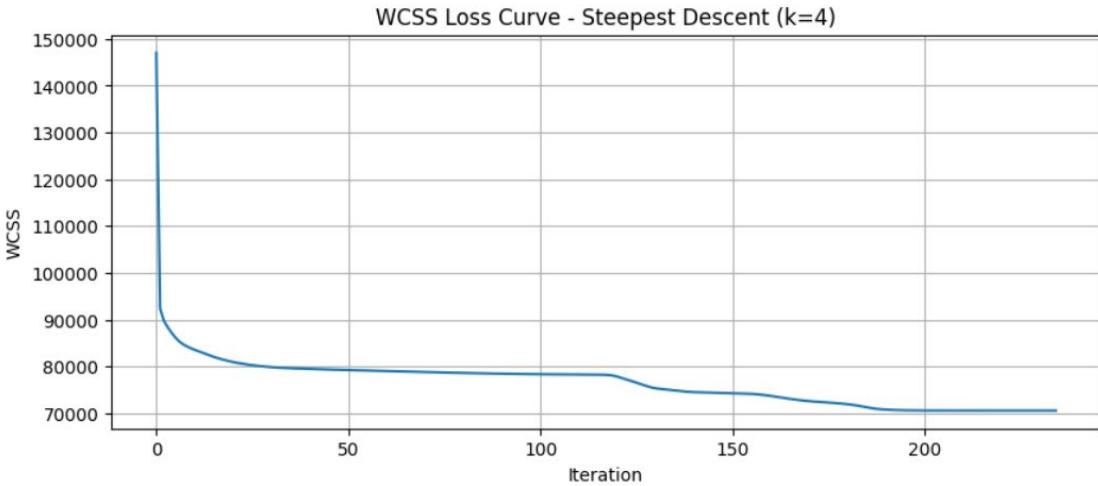


k	WCSS
2	39,795.08
3	31,283.62
4	25,822.96
5	23,044.18
6	20,454.50
7	18,456.09
8	16,966.67
9	15,901.57
10	14,946.77

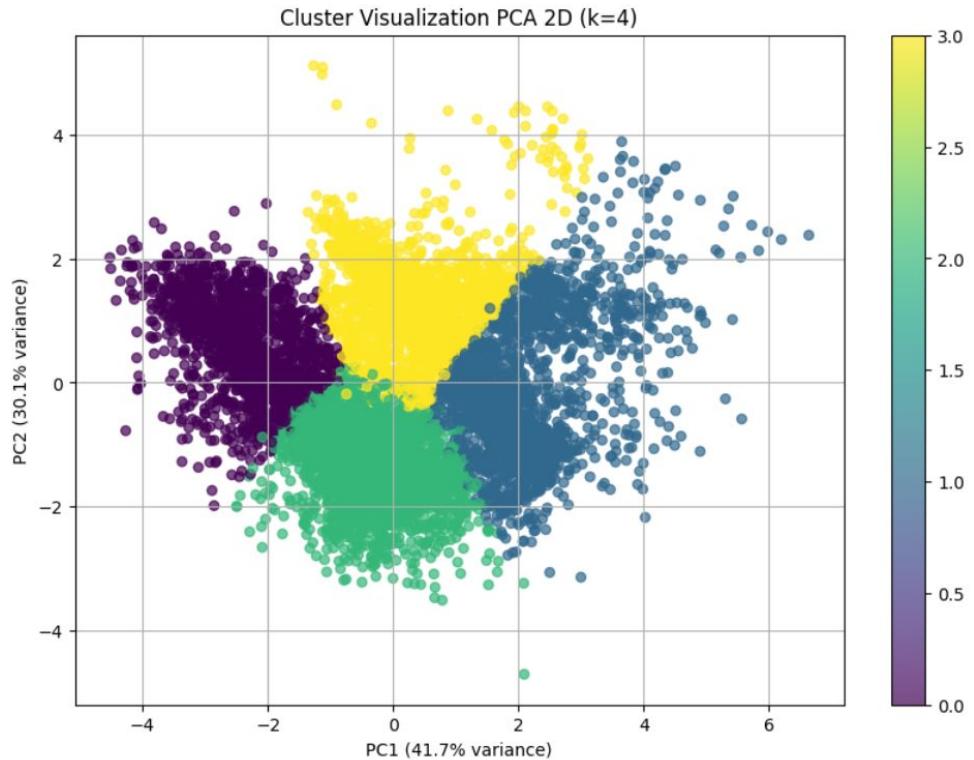


Visualisasi Loss Curve

- Grafik Loss (WCSS) menunjukkan penurunan stabil.
- Algoritma Steepest Descent berhasil mencapai konvergensi.

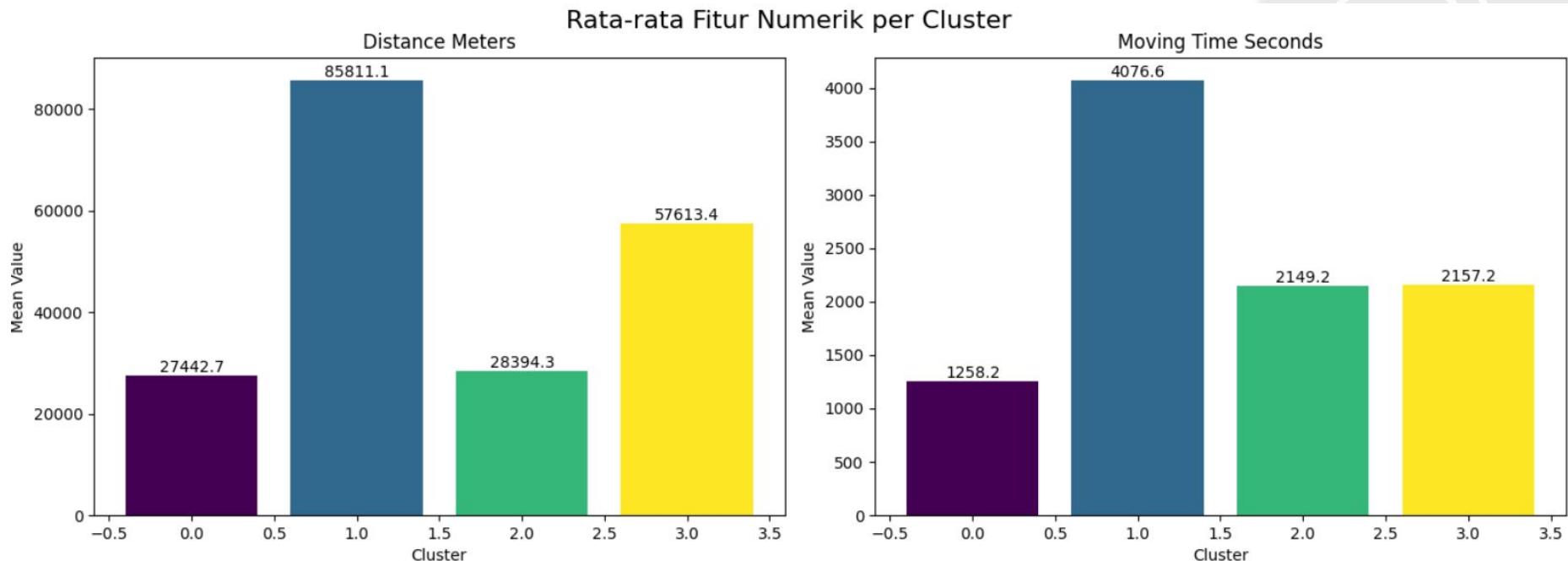


Visualisasi PCA 2D

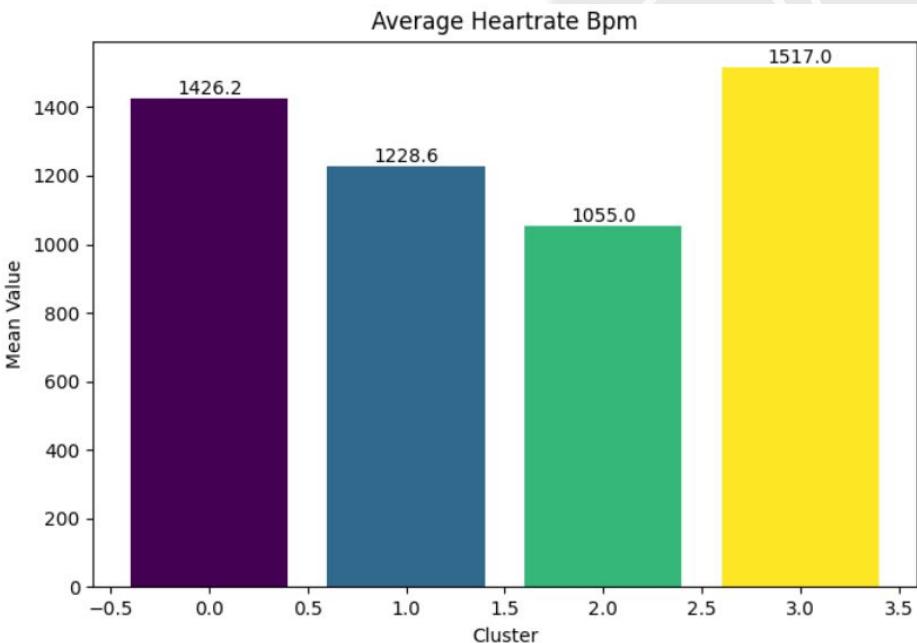
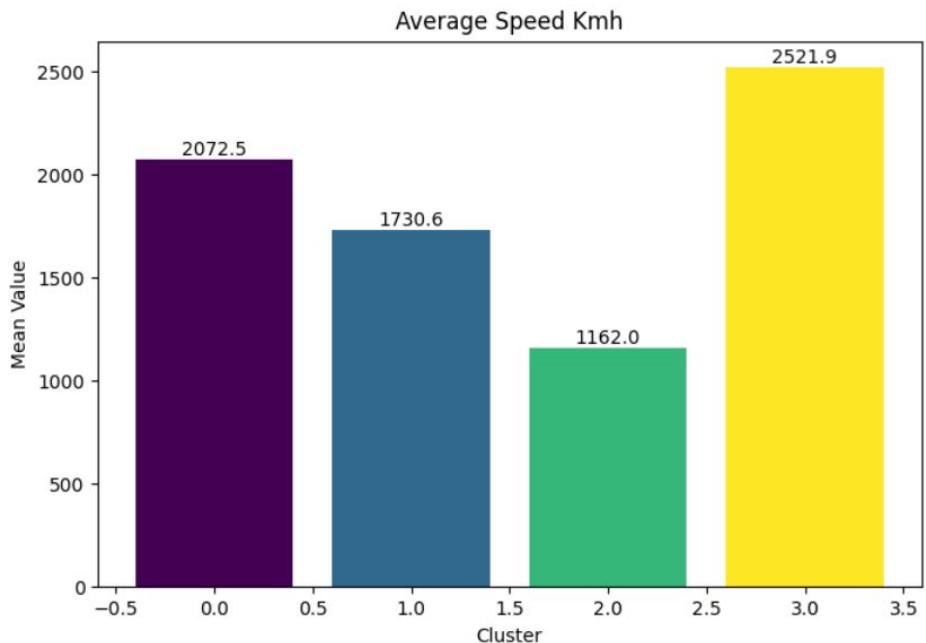


- PCA digunakan untuk memproyeksikan fitur ke 2 dimensi.
- Warna berbeda mewakili cluster.
- Terlihat pemisahan cukup jelas antara cluster intensitas tinggi dan cluster berjalan/aktivitas ringan.

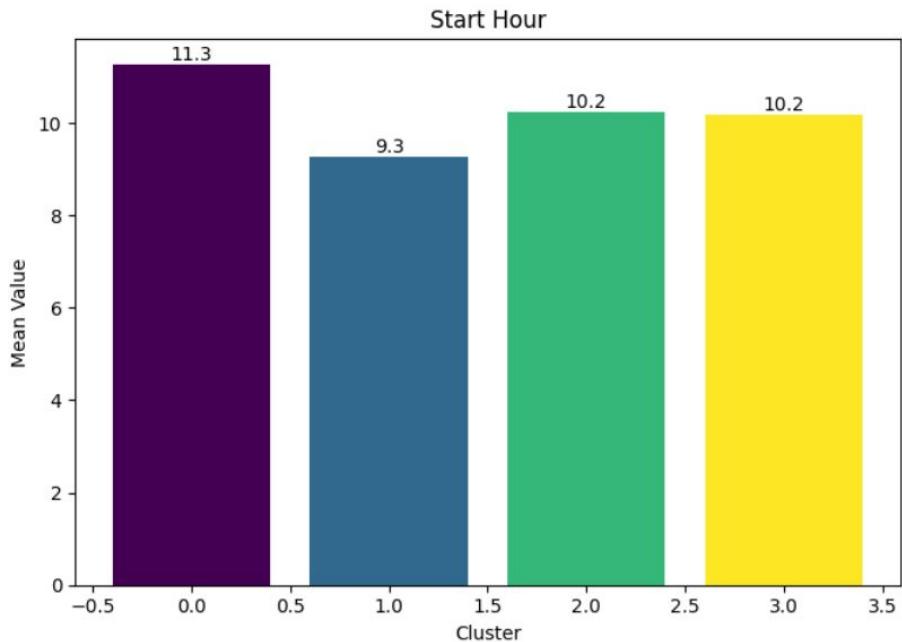
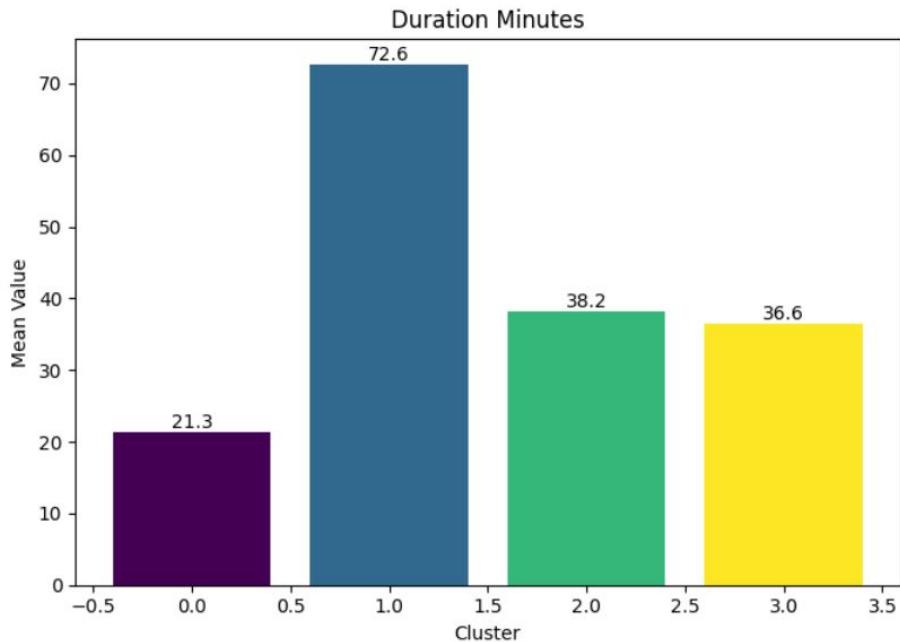
Analisis Per Cluster Berdasarkan Fitur Numerik



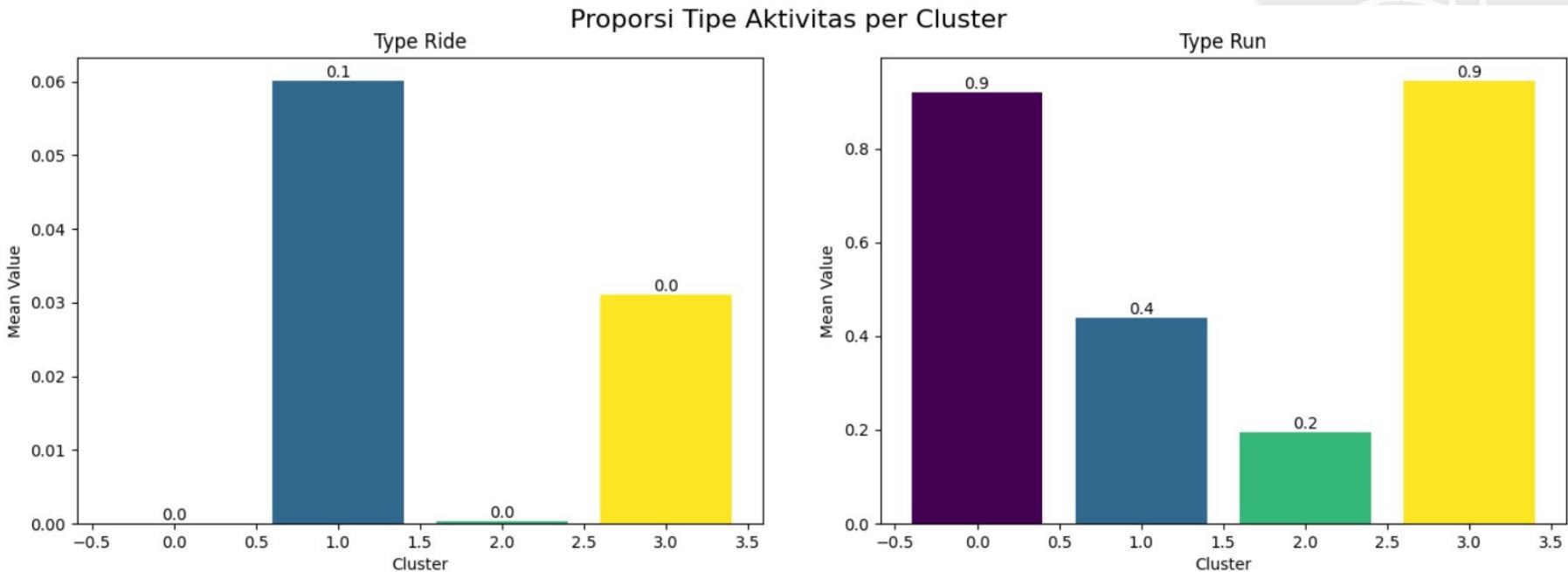
Analisis Per Cluster Berdasarkan Fitur Numerik



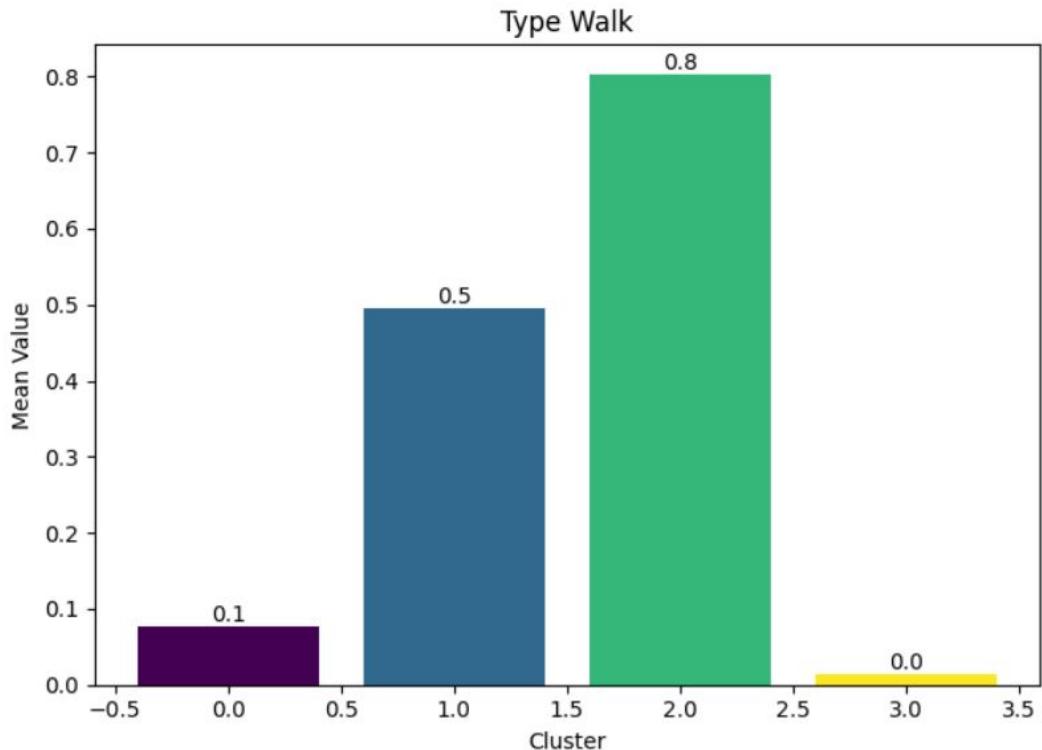
Analisis Per Cluster Berdasarkan Fitur Numerik



Analisis Per Cluster Berdasarkan Aktivitas



Analisis Per Cluster Berdasarkan Aktivitas



Interpretasi Cluster

Aspek	Cluster 0	Cluster 1	Cluster 2	Cluster 3
jumlah aktivitas	1.516	1.944	3.148	2.600
jarak tempuh	Rendah (~27.442 m)	Paling Tinggi (~85.811 m)	Rendah (~28.482 m)	Tinggi (~57.613 m)
Durasi Aktivitas	Singkat (~21 menit)	Paling Lama (~68 menit)	Sedang (~36 menit)	Sedang (~36 menit)
Kecepatan	Tinggi (~2.072)	Sedang (~1.730)	Paling Rendah (~1.162)	Tertinggi (~2.521)
Detak Jantung	Tinggi (~1.426)	Sedang (~1.228)	Terendah (~1.055)	Tertinggi (~1.517)

Interpretasi Cluster

Aspek	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Tipe Aktivitas	Running Dominan (>90%)	Campuran (Ride tertinggi, Walk, Run)	Walking Sangat Dominan (>80%)	Running Sangat Dominan (>90%)
Interpretasi	Short Distance Fast Run (Lari Cepat Jarak Pendek)	Endurance / Long Activity (Aktivitas Ketahanan Jarak Jauh)	Casual Walking (Jalan Santai Intensitas Ringan)	High Performance Runners (Pelari Performa Tinggi)



UNIVERSITAS
GADJAH MADA

Terimakasih

