IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING FROM SCRATCH UNTUK SEGMENTASI RISIKO ASURANSI

OPTIMASI PENENTUAN PREMI ASURANSI KARYAWAN BERBASIS DATA AKTIVITAS FISIK

Alfaizz Dyandra Ardin
25/562922/NPA/19986

LATAR BELAKANG & TUJUAN

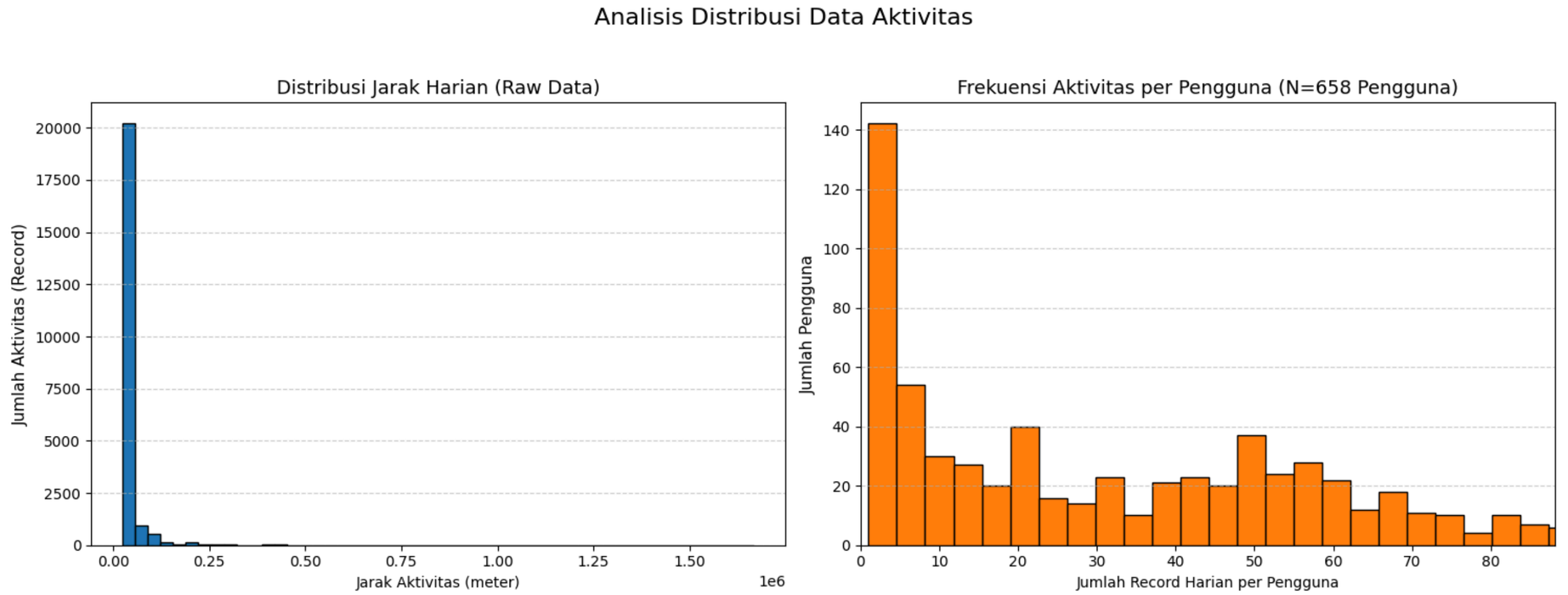
- Masalah Utama : Perusahaan membutuhkan metode objektif untuk mengklasifikasikan risiko kesehatan karyawan, bukan hanya asumsi.

- Solusi : Membangun model Machine Learning (K-Means) untuk mengelompokkan karyawan berdasarkan data aktivitas riil (Jarak, Kecepatan, Detak Jantung).

- Tantangan Teknis : Algoritma dibangun tanpa library menggunakan Python untuk mendemonstrasikan pemahaman mendalam tentang fungsi optimasi matematika.

LEARN MORE

EXPLORATORY DATA ANALYSIS



Data mentah sangat skewed (miring) ke kiri. Sebagian besar karyawan memiliki aktivitas rendah, dengan sedikit outlier yang sangat aktif. Ini menjadi sinyal bahwa data perlu dinormalisasi.

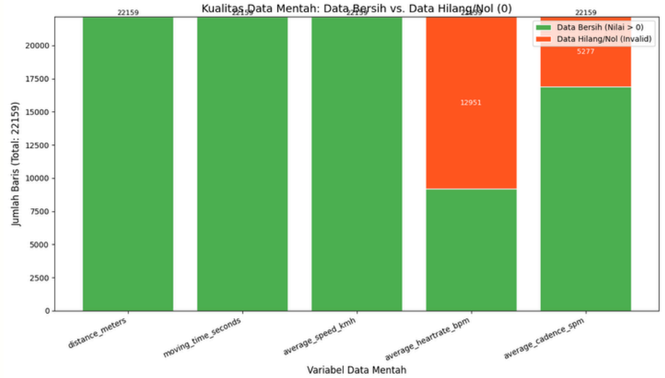
DATA CLEANING

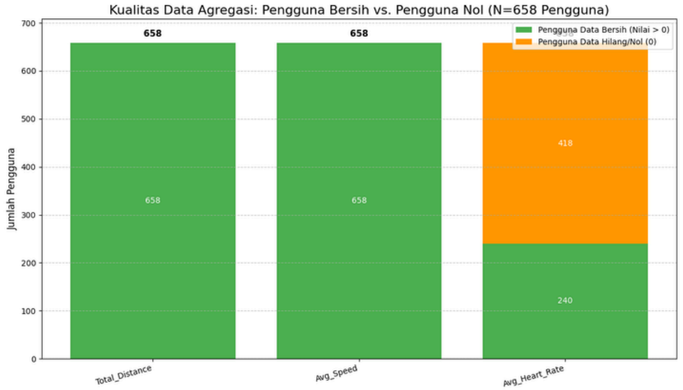
distance_meters	moving_time_seconds	elapsed_time_seconds	type	start_date_utc	end_date_utc	average_speed_kmh	max_speed_kmh	average_heartrate_bpm	average_cadence_spm	badge
31260	2353	2484	Walk	4/12/2025 10:33	4/12/2025 11:15	1329	2773	0	0	13289547
26900	1375	1375	Walk	1/2/2025 18:31	1/2/2025 18:54	1956	0	1226	1022	13267699
30200	1521	1521	Walk	1/3/2025 18:54	1/3/2025 19:20	1986	0	1261	998	13267699
28500	1480	1480	Walk	1/4/2025 9:07	1/4/2025 9:31	1926	0	1292	998	13267699
26300	1286	1286	Walk	1/13/2025 18:46	1/13/2025 19:07	2045	0	1345	1002	13267699

- Data awal kita terdiri dari 22.159 catatan harian (baris data). Awalnya, data Jarak dan Kecepatan terlihat sangat bagus, hampir 100% lengkap. Masalah utamanya ada pada Detak Jantung Rata-rata, di mana 12.951 catatan (lebih dari setengah) tidak ada nilainya atau nol.

- Karena kita perlu menganalisis perilaku pengguna, kita kelompokkan (agregasi) catatan-catatan tadi menjadi 658 orang unik. Setelah dikelompokkan, kita cek lagi. Ternyata, 418 orang dari 658 itu memiliki rata-rata Detak Jantung nol (berarti data aktivitas intensitas mereka tidak terekam).

- Karena data Detak Jantung ini penting, kita putuskan untuk mengeluarkan 418 orang yang datanya tidak valid tersebut. Hasilnya, kita mendapatkan 240 orang dengan data yang benar-benar bersih dan siap untuk diolah ke tahap selanjutnya.





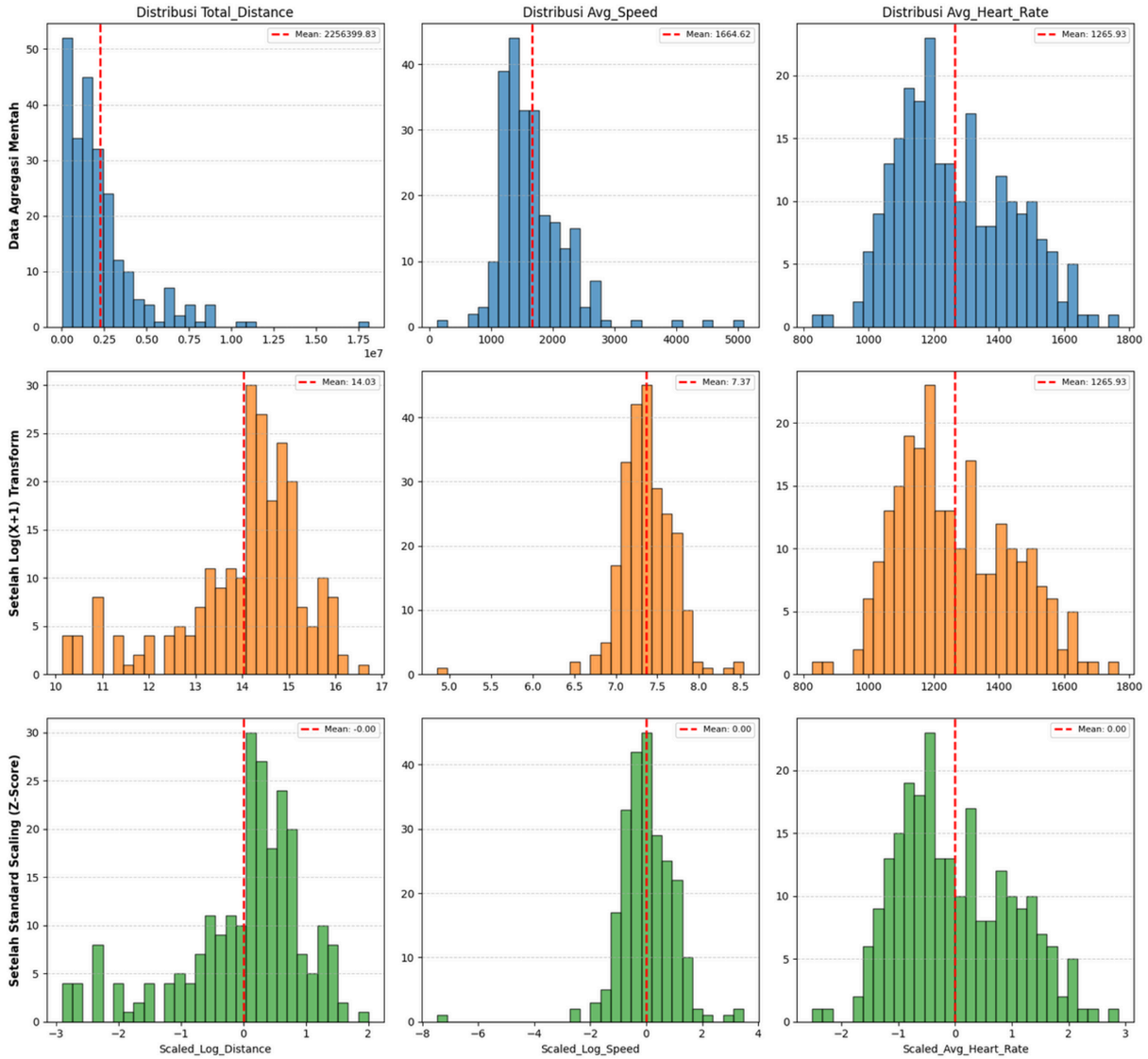
PREPROCESSING

- Masalah Skala : Data Distance memiliki satuan jutaan meter, sedangkan Speed hanya ribuan. Jika dibiarkan, K-Means hanya akan menghitung jarak berdasarkan Distance dan mengabaikan Speed.

Solusi :

- Log Transformation : Mengubah data yang timpang menjadi lebih berdistribusi normal.

- Standard Scaling (Z-Score) : Mengubah semua variabel agar memiliki rata-rata 0 dan standar deviasi 1, sehingga setiap fitur memiliki bobot setara dalam perhitungan jarak Euclidean.



FUNGSI OBJEKTIF

Within-Cluster Sum of Squares (WCSS)

Tujuan algoritma adalah meminimalisir variansi dalam klaster. Semakin kecil nilai WCSS, semakin mirip data di dalam satu kelompok.

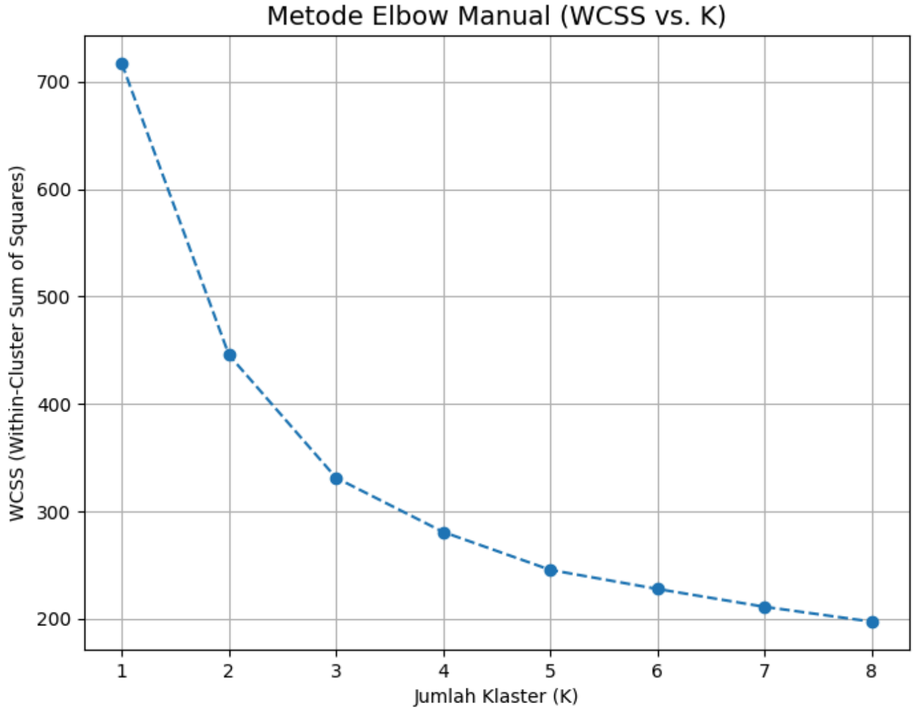
Rumus Utama :

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

- J: Fungsi Objektif (Nilai Error yang ingin diminimalkan).
- k: Jumlah klaster (dalam kasus ini 3).
- x_i: Titik data aktivitas karyawan.
- c_j: Titik pusat (centroid) klaster.
- $||...||^2$: Jarak Euclidean kuadrat (jarak garis lurus).

PENENTUAN JUMLAH KLASTER (ELBOW METHOD)

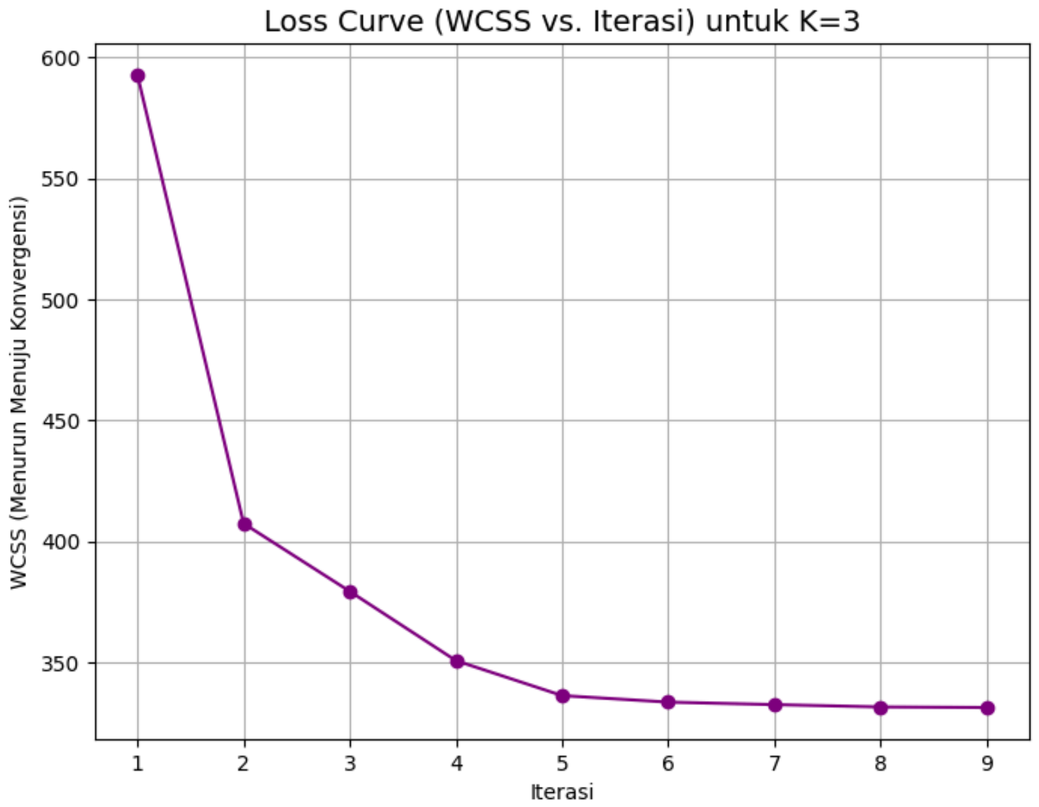
```
Memulai Perhitungan Manual Elbow Method (WCSS)...
K=1: WCSS=717.00
K=2: WCSS=445.72
K=3: WCSS=331.35
K=4: WCSS=280.89
K=5: WCSS=245.62
K=6: WCSS=227.86
K=7: WCSS=211.24
K=8: WCSS=197.45

Menghitung Loss Curve (WCSS vs. Iterasi) untuk K=3...

--- Iterasi 1/20 ---
Cluster 0: Centroid = (0.2693, 0.9119, 1.3601), Ukuran = 62 pengguna
Cluster 1: Centroid = (0.1903, -0.4891, -0.9083), Ukuran = 92 pengguna
Cluster 2: Centroid = (-0.3977, -0.1342, -0.0089), Ukuran = 86 pengguna

--- Iterasi 2/20 ---
Cluster 0: Centroid = (0.4627, 0.9142, 1.1587), Ukuran = 75 pengguna
Cluster 1: Centroid = (0.2206, -0.4993, -0.8985), Ukuran = 90 pengguna
Cluster 2: Centroid = (-0.7274, -0.3150, -0.0805), Ukuran = 75 pengguna

--- Iterasi 3/20 ---
Cluster 0: Centroid = (0.5025, 0.8766, 1.0341), Ukuran = 83 pengguna
Cluster 1: Centroid = (0.2734, -0.4046, -0.8030), Ukuran = 96 pengguna
Cluster 2: Centroid = (-1.1140, -0.5560, -0.1433), Ukuran = 61 pengguna
```



- Dilakukan uji coba dari K=1 hingga K=8.
- Terlihat penurunan nilai WCSS (Error) yang drastis dari K=1 ke K=3, namun mulai melandai (stagnan) setelah K=3.
- Keputusan : K=3 dipilih karena marginal gain (penurunan WCSS) setelah K=3 tidak lagi signifikan dibandingkan kompleksitas model yang bertambah.

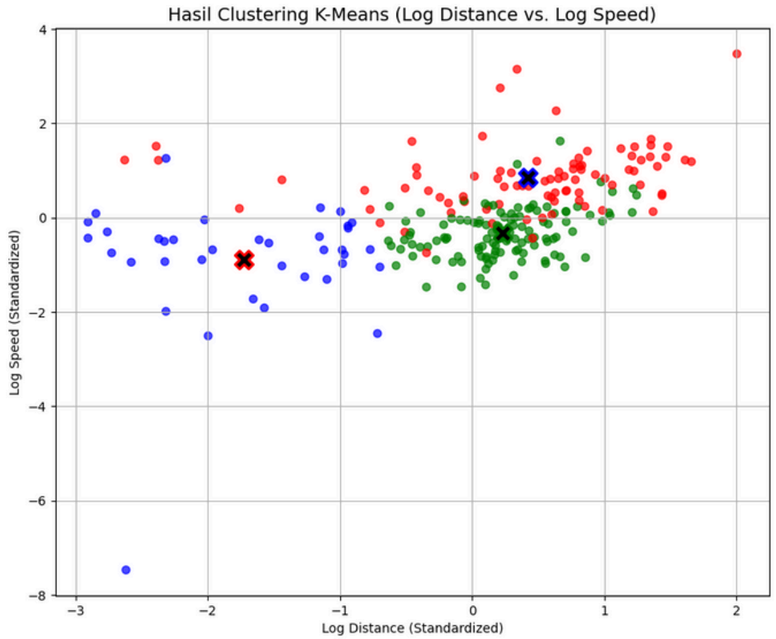
PROSES TRAINING & KONVERGENSI

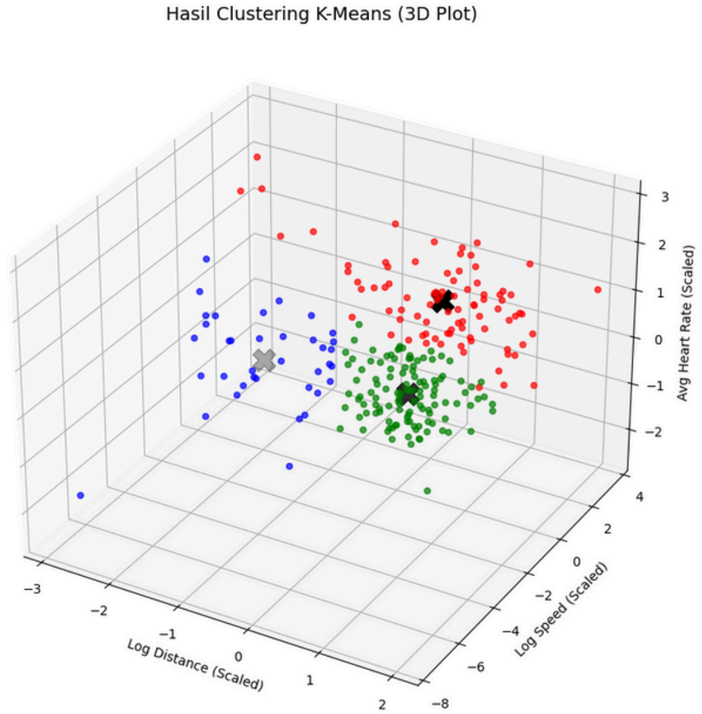
```
--- Iterasi 7/20 ---
Cluster 0: Centroid = (0.4376, 0.8425, 1.0404), Ukuran = 86 pengguna
Cluster 1: Centroid = (0.2406, -0.3399, -0.6184), Ukuran = 116 pengguna
Cluster 2: Centroid = (-1.7248, -0.8690, -0.4669), Ukuran = 38 pengguna

--- Iterasi 8/20 ---
Cluster 0: Centroid = (0.4244, 0.8420, 1.0569), Ukuran = 86 pengguna
Cluster 1: Centroid = (0.2332, -0.3350, -0.6103), Ukuran = 117 pengguna
Cluster 2: Centroid = (-1.7239, -0.8978, -0.5270), Ukuran = 37 pengguna

--- Iterasi 9/20 ---
Cluster 0: Centroid = (0.4244, 0.8420, 1.0569), Ukuran = 86 pengguna
Cluster 1: Centroid = (0.2332, -0.3350, -0.6103), Ukuran = 117 pengguna
Cluster 2: Centroid = (-1.7239, -0.8978, -0.5270), Ukuran = 37 pengguna

Algoritma Konvergen, menghentikan iterasi lebih awal.
```



- Algoritma dijalankan secara iteratif (berulang).
- Grafik ini membuktikan stabilitas kode manual yang dibuat. Algoritma mencapai konvergensi (stabil) pada iterasi ke-9.
- Kriteria penghentian (stopping criteria) tercapai ketika pergeseran posisi centroid berada di bawah ambang batas, menunjukkan solusi lokal optimum telah ditemukan.

VISUALISASI

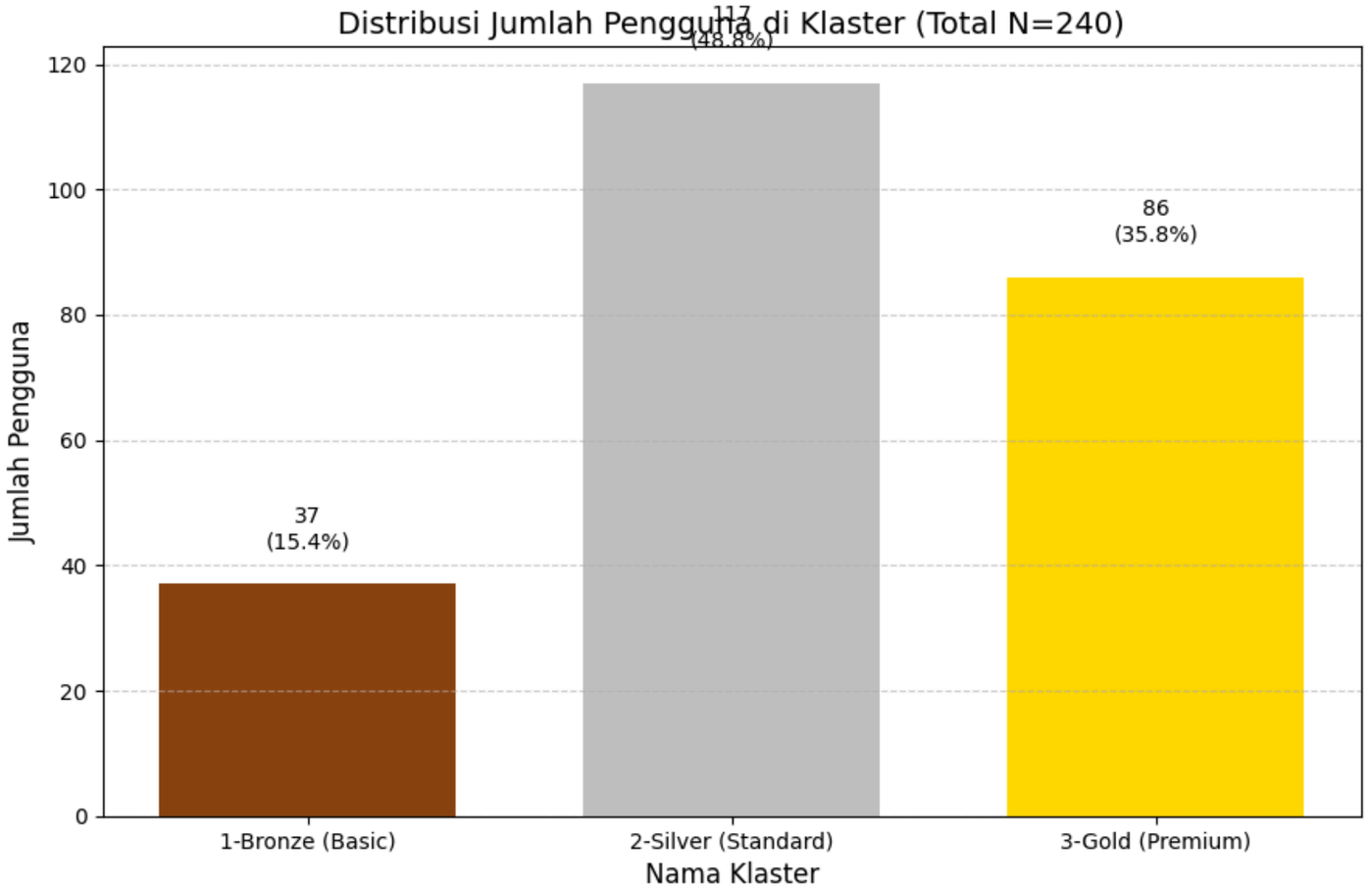




- Plotting dilakukan pada fitur yang sudah direduksi dimensinya.
- Titik-titik data terpisah secara jelas menjadi 3 warna (Merah, Hijau, Biru) dengan titik pusat (Centroid) yang tegas.
- Tidak ada tumpang tindih (overlap) yang signifikan antar kelompok, menandakan kualitas klaster yang baik.

PROFIL KLASTER

- Klaster 1 (Bronze/Basic - 15.4%): Aktivitas fisik rendah. perlu program pemicu kesehatan.
- Klaster 2 (Silver/Standard - 48.8%): Aktivitas sedang. Mayoritas karyawan berada di sini.
- Klaster 3 (Gold/Premium - 35.8%): Sangat aktif (Atletis). Risiko penyakit rendah.



Link Collab : https://colab.research.google.com/drive/10dXnIAWJInT2wRjiGgK6oZtMpQPZ8qDF?usp=sharing

THANK YOU