

Implementasi K-Means Clustering dalam Memetakan Karakteristik Performa dan Konsistensi Pembalap pada Era Regulasi Baru Formula 1 2026

Elsa Anggraini¹

¹Jurusan Informatika, Universitas Bina Sarana Informatika
Slipi Kemanggisan, Jakarta Barat
15230581@bsi.ac.id

Abstract— The 2026 Formula 1 season introduces a radical regulatory transition, rendering historical performance data obsolete. This study addresses the "cold-start" problem in sports analytics by implementing the K-Means clustering algorithm to map competitive hierarchies during the 2026 Bahrain pre-season tests. The analysis is exclusively based on four key performance features: Fastest Lap, Average Lap Time, Standard Deviation (consistency), and Total Laps (reliability), extracted via the FastF1 API. A total of 3,624 telemetry data rows were processed and normalized using StandardScaler. The Elbow Method identified $K=4$ as the optimal cluster configuration. Although the Silhouette Coefficient of 0.350 reflects the inherent "noise" and "sandbagging" strategies of F1 testing, the model successfully differentiated four distinct performance tiers: Top-Tier Leaders, Stable Midfielders, Reliability-Focused Testers, and Technical Anomalies (Strugglers). The findings provide an objective, data-driven framework for interpreting competitive strength without relying on subjective media reports, proving that unsupervised learning can extract meaningful patterns from unlabelled telemetry data in highly volatile regulatory environments.

Keywords — Formula 1 2026, K-Means Clustering, Sports Analytics, FastF1, Pre-season Testing, Machine Learning.

Intisari— Musim Formula 1 2026 memperkenalkan transisi regulasi yang radikal, menyebabkan data performa historis kehilangan relevansinya. Penelitian ini menjawab tantangan "cold-start" dalam analisis olahraga dengan mengimplementasikan algoritma K-Means clustering untuk memetakan hirarki kompetisi selama pengujian pra-musim Bahrain 2026. Analisis dilakukan secara eksklusif berdasarkan empat fitur performa utama: *Fastest Lap*, *Average Lap Time*, *Standard Deviation* (konsistensi), dan *Total Laps* (reliabilitas) yang diekstraksi melalui FastF1 API. Sebanyak 3.624 baris data telemetri diproses dan dinormalisasi menggunakan *StandardScaler*. Metode Elbow mengidentifikasi $K=4$ sebagai konfigurasi kluster optimal. Meskipun nilai Silhouette Coefficient sebesar 0,350 mencerminkan tingginya faktor gangguan (*noise*) dan strategi *sandbagging* khas pengujian F1, model berhasil membedakan empat tingkatan performa: Top-Tier Leaders, Stable Midfielders, Reliability-Focused Testers, dan Technical Anomalies (Strugglers). Temuan ini memberikan kerangka kerja objektif berbasis data untuk menginterpretasi kekuatan kompetisi tanpa bergantung pada laporan media yang subjektif, sekaligus membuktikan bahwa *unsupervised learning* efektif mengekstraksi pola dari data telemetri tanpa label dalam lingkungan regulasi yang dinamis.

Kata Kunci—Formula 1 2026, K-Means Clustering, Analisis Olahraga, FastF1, Pengujian Pra-musim, Pembelajaran Mesin.

I. PENDAHULUAN

A. Latar Belakang

Formula 1 (F1) musim 2026 menghadapi transformasi radikal melalui pengenalan regulasi teknis baru yang mencakup desain ulang *power unit* dengan pembagian daya 50/50 antara mesin pembakaran internal (ICE) dan sistem elektrik (ERS), serta pengenalan sistem aerodinamika aktif[1][2]. Perubahan besar ini mengakibatkan data historis dari musim-musim sebelumnya menjadi tidak relevan untuk dijadikan tolok ukur dalam memprediksi peta kekuatan tim secara akurat. Dalam ekosistem teknis yang sepenuhnya baru, sesi pengujian pra-musim (*pre-season testing*) di Bahrain menjadi satu-satunya sumber data objektif bagi analis untuk memahami posisi kompetitif setiap tim sebelum musim balap dimulai.

Namun, identifikasi kekuatan tim melalui data pengujian menghadapi kendala serius karena sifat data yang "kotor" dan penuh gangguan (*noise*). Praktik strategi *sandbagging*, di mana tim secara sengaja menyembunyikan potensi asli melalui manipulasi beban bahan bakar atau pemetaan mesin, menyebabkan fluktuasi catatan waktu yang tidak mencerminkan kecepatan murni kendaraan. Penggunaan metode statistik konvensional seringkali gagal memisahkan antara performa nyata dengan anomali teknis tersebut, sehingga menghasilkan interpretasi yang subjektif dan kurang akurat terhadap hirarki kekuatan tim sesungguhnya.

Penelitian terdahulu telah mencoba mengimplementasikan teknik *Data Mining* dalam bidang *sports analytics*, namun sebagian besar masih mengandalkan data hasil balapan (*race results*) yang sudah terlabeli secara jelas[3][4]. Terdapat celah penelitian (*research gap*) dalam mengolah data mentah telemetri pengujian yang bersifat *unlabeled* dan memiliki variansi tinggi akibat perbedaan program tes tiap tim. Oleh karena itu, diperlukan pendekatan komputasi modern menggunakan algoritma *Unsupervised Learning* untuk mengekstraksi pola performa murni secara otomatis[5].

Algoritma *K-Means Clustering* menawarkan kapabilitas untuk mengelompokkan pembalap berdasarkan kemiripan profil performa tanpa memerlukan label awal[6]. Dengan mengintegrasikan variabel *Fastest Lap* dan *Total Laps*, penelitian ini berfokus pada implementasi algoritma *K-*

Means untuk melakukan segmentasi hirarki kekuatan tim secara objektif pada era regulasi 2026[7]. Pendekatan ini diharapkan mampu memberikan kontribusi metodologis dalam memberikan wawasan mendalam mengenai dampak nyata regulasi teknis terhadap reliabilitas dan performa tim di lapangan secara terukur.

B. Identifikasi Masalah

Masalah utama dalam penelitian ini adalah sulitnya menentukan hierarki kekuatan tim dan pembalap pada era regulasi 2026 akibat hilangnya relevansi data historis. Selain itu, terdapat beberapa faktor teknis yang menjadi hambatan dalam analisis data mentah:

1) *Tingginya Variabel Pengganggu (Data Noise)* : Dataset mentah telemetri mengandung banyak putaran non-kompetitif seperti *in-laps*, *out-laps*, dan pengujian sensor aerodinamika. Tanpa teknik pembersihan yang objektif, analisis terhadap performa murni akan menghasilkan bias data yang signifikan.

2) *Heterogenitas Skala Atribut* : Adanya perbedaan satuan dan skala yang mencolok antar variabel performa (misalnya, *Total Laps* dalam angka ratusan sementara *Std Dev* dalam angka desimal) berpotensi mendistorsi hasil pengelompokan jika tidak melalui tahap normalisasi yang tepat.

3) *Ketidakpastian Variabel Tersembunyi (Fuel Load)*: Tidak adanya publikasi data mengenai beban bahan bakar tiap mobil menciptakan ketidakpastian dalam menentukan apakah sebuah catatan waktu merepresentasikan kecepatan maksimal (*qualifying trim*) atau simulasi balap (*race trim*).

4) *Kompleksitas Strategi Sandbagging*: Kecenderungan tim untuk menyembunyikan performa asli mereka melalui program pengujian yang bervariasi menyebabkan terjadinya tumpang tindih (*overlap*) data yang sulit dipisahkan tanpa pengelompokan (*clustering*) yang sistematis.

Tanpa adanya metode komputasi yang mampu mengolah variabel-variabel tersebut secara simultan, karakteristik konsistensi dan kecepatan murni setiap pembalap sulit untuk dipetakan secara akurat[8].

C. Tujuan Penelitian

Penelitian ini bertujuan untuk mengimplementasikan algoritma K-Means *Clustering* dalam memetakan karakteristik performa pembalap Formula 1 berdasarkan hasil pengujian pra-musim 2026 di Bahrain. Fokus utama adalah mengelompokkan pembalap ke dalam beberapa klaster berdasarkan variabel kecepatan (*Fastest Lap*) dan konsistensi (*Standard Deviation*), serta memvalidasi jumlah klaster optimal menggunakan metode *Elbow* dan *Silhouette Coefficient*. Secara lebih spesifik, sasaran yang ingin dicapai dalam penelitian ini meliputi :

1) *Ekstraksi Pola Performa Melalui Pendekatan Komputasi*: Mengidentifikasi hirarki kekuatan tim dan pembalap secara objektif dengan memanfaatkan empat fitur utama (*Fastest Lap*, *Avg Lap Time*, *Std Dev*, dan *Total Laps*) dari data telemetri hasil pengujian pra-musim 2026 di Bahrain. Fokus utama terletak pada pengelompokan pembalap ke dalam beberapa klaster representatif guna membedakan antara performa puncak (*peak performance*) dengan ketahanan sistem operasional (*operational endurance*).

2) *Optimasi Segmentasi Klaster*: Menentukan dan memvalidasi jumlah klaster optimal (*K*) dengan mengintegrasikan metode *Elbow* berbasis *Within-Cluster Sum of Squares* (WCSS) serta evaluasi *Silhouette Coefficient*[9]. Hal ini bertujuan untuk memastikan bahwa setiap pengelompokan memiliki tingkat homogenitas internal yang tinggi dan separasi antar kelompok yang jelas.

3) *Analisis Multivariabel Terhadap Strategi Tim*: Menganalisis korelasi antara variabel kecepatan murni (*Fastest Lap*) dan konsistensi operasional (*Total Laps* atau *Standard Deviation*) untuk mengungkap fenomena sandbagging. Penelitian ini bertujuan untuk membuktikan bahwa integrasi variabel jumlah putaran sangat krusial dalam membedakan tim yang sedang melakukan simulasi jarak jauh dari tim yang mengalami kegagalan mekanis sistem hibrida 2026.

4) *Kontribusi Metodologis dalam Sports Analytics*: Memberikan kontribusi berupa kerangka kerja (*framework*) analisis data olahraga motor yang dapat diadaptasi untuk evaluasi performa pada seri balapan berikutnya, khususnya dalam menghadapi situasi di mana data historis tidak lagi tersedia akibat perubahan regulasi total[10].

D. Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Bagi analis olahraga, penelitian ini menyediakan kerangka kerja objektif untuk memprediksi peta persaingan musim 2026 tanpa bergantung pada opini subjektif.
2. Bagi pengembang teknologi informasi, penelitian ini mendemonstrasikan penerapan *unsupervised learning* pada dataset telemetri balap yang dinamis dan kompetitif.
3. Sebagai referensi bagi penelitian selanjutnya dalam domain analisis data olahraga motor, khususnya dalam menghadapi perubahan regulasi teknis yang drastis.

II. LANDASAN TEORI

A. Dinamika Regulasi Teknis Formula 1 2026

Perubahan regulasi Formula 1 tahun 2026 merupakan paradigma baru yang menitikberatkan pada keberlanjutan dan efisiensi energi. Perubahan paling signifikan terletak pada unit daya (*Power Unit*), di mana komponen MGU-H

(*Motor Generator Unit - Heat*) dihilangkan sepenuhnya. Sebagai gantinya, porsi tenaga elektrik ditingkatkan hingga tiga kali lipat melalui sistem MGU-K (*Motor Generator Unit - Kinetic*) yang mampu menghasilkan daya sebesar 350kW[11].

Peningkatan ketergantungan pada sistem elektrik ini menghadirkan tantangan besar dalam manajemen termal baterai dan pemulihan energi[12]. Bagi tim dengan riset yang belum matang, fenomena *derating* (penurunan tenaga elektrik di lintasan lurus) menjadi hambatan utama dalam mencapai waktu putaran yang konsisten. Hal inilah yang menjadi dasar teoritis mengapa variabel *Total Laps* dan *Fastest Lap* menjadi fitur krusial untuk memisahkan tim yang sudah menguasai integrasi hibrida dengan tim yang masih mengalami kendala teknis.

B. Algoritma K-Means Clustering

Algoritma *K-Means Clustering* dikategorikan sebagai metode *centroid-based clustering* atau klusterisasi berbasis partisi dalam ranah *unsupervised learning*[13]. Tujuan utama dari algoritma ini adalah untuk mengelompokkan sekumpulan objek ke dalam K kelompok (klaster) sedemikian rupa sehingga tingkat kemiripan antar anggota dalam satu klaster (intra-klaster) maksimal, sementara kemiripan antar klaster (inter-klaster) minimal[14]. Dalam konteks penelitian ini, objek tersebut merupakan pembalap Formula 1 yang diklasifikasikan berdasarkan fitur-fitur performa seperti *Fastest Lap*, *Average Lap Time*, dan *Total Laps*[15].

Secara matematis, algoritma ini bekerja secara iteratif untuk meminimalkan fungsi objektif yang dikenal sebagai *Sum of Squared Errors* (SSE) atau inersia. Proses ini melibatkan pencarian posisi pusat massa atau *centroid* (μ) yang paling representatif untuk setiap kelompok data[12]. Penentuan alokasi setiap titik data ke *centroid* terdekat dilakukan dengan menghitung jarak geometris. Metode yang paling umum digunakan adalah *Euclidean Distance*, sebagaimana dirumuskan dalam Persamaan :

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

- a) $d(x, y)$: Merepresentasikan jarak skalar antara titik data pembalap (x) dengan titik pusat klaster atau *centroid* (y).
- b) n : Melambangkan jumlah dimensi atau fitur yang digunakan dalam pemodelan (misalnya, jika menggunakan *Fastest Lap* dan *Total Laps*, maka $n = 2$).
- c) x_i : Merupakan nilai fitur ke- i dari titik data pengamatan.
- d) y_i : Merupakan nilai fitur ke- i dari koordinat *centroid*.

Pemilihan *Euclidean Distance* sangat krusial dalam analisis telemetri F1 karena kemampuannya untuk mengukur

jarak garis lurus dalam ruang multidimensi, sehingga perbedaan performa yang kecil sekalipun (seperti selisih 0,1 detik pada *Fastest Lap*) dapat dipetakan secara akurat ke dalam klaster yang sesuai.

C. Metode Elbow

Metode *Elbow* merupakan teknik heuristik yang digunakan untuk menentukan jumlah klaster optimal (K) dalam algoritma *K-Means*[16]. Teknik ini bekerja dengan mengevaluasi tingkat homogenitas internal dari setiap klaster yang terbentuk. Penentuan nilai K yang tepat sangat krusial dalam analisis data Formula 1 2026 ini untuk memastikan bahwa pengelompokan pembalap (seperti *Leaders*, *Builders*, atau *Strugglers*) didasarkan pada struktur data alami, bukan sekadar pembagian acak yang dapat mendistorsi interpretasi performa[17].

Implementasi metode ini dilakukan dengan menghitung dan memplot nilai *Within-Cluster Sum of Squares* (WCSS) terhadap variasi jumlah klaster. WCSS mengukur total akumulasi kuadrat jarak antara setiap titik data dengan pusat massa (*centroid*) masing-masing klaster. Secara matematis, nilai WCSS dihitung dengan menggunakan Persamaan :

$$WCSS = \sum_{P_i \in Cluster_1} dist(P_i, C_1)^2 + \sum_{P_i \in Cluster_2} dist(P_i, C_2)^2 + \dots \quad (2)$$

- a) **WCSS** : Merupakan nilai inersia atau indikator kekompakan klaster. Semakin rendah nilai WCSS, maka data di dalam klaster tersebut semakin mendekati pusat massa (*centroid*), yang mengindikasikan kemiripan fitur yang tinggi.
- b) P_i : Mewakili titik data pengamatan, yaitu profil performa setiap pembalap berdasarkan fitur *Fastest Lap* dan *Total Laps*.
- c) C_1, C_2, \dots : Merupakan pusat klaster (*centroid*) untuk masing-masing kelompok yang terbentuk.
- d) $dist(P_i, C_k)^2$ Adalah kuadrat dari jarak antara titik data pembalap terhadap *centroid* klasternya, yang dihitung menggunakan prinsip *Euclidean Distance* sebagaimana dijelaskan pada Persamaan (1).

Proses identifikasi dilakukan dengan mengamati grafik yang dihasilkan. Seiring bertambahnya jumlah klaster, nilai WCSS akan menurun secara drastis karena jarak antara titik data dengan *centroid* akan semakin mengecil. Namun, penambahan klaster di atas titik optimal hanya akan memberikan penurunan WCSS yang minimal. Titik di mana terjadi perubahan kemiringan yang signifikan (membentuk sudut serupa siku-siku) dipilih sebagai nilai K optimal.

Penerapan metode ini dalam penelitian bertujuan untuk mendapatkan keseimbangan antara presisi pengelompokan dan kesederhanaan model, sehingga hasil klusterisasi tetap memiliki nilai validasi yang kuat saat dibandingkan dengan data laporan teknis dari tim F1 di lapangan.

D. Silhouette Coefficient

Silhouette Coefficient merupakan metode validasi untuk mengukur kualitas dan kelayakan dari hasil

pengelompokan[18]. Nilai *Silhouette* (s) untuk suatu objek i dihitung berdasarkan perbandingan jarak rata-rata objek tersebut ke objek lain dalam klaster yang sama (a) dan jarak rata-rata ke klaster terdekat lainnya (b), sebagaimana dirumuskan pada Persamaan:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3)$$

Nilai yang mendekati 1 menunjukkan pemisahan klaster yang sangat baik, sementara nilai mendekati 0 atau 0.350 menunjukkan adanya tumpang tindih (*overlap*) antar klaster yang wajar ditemukan dalam data telemetri balap yang kompetitif.

E. Standarisasi Data (*StandardScaler*)

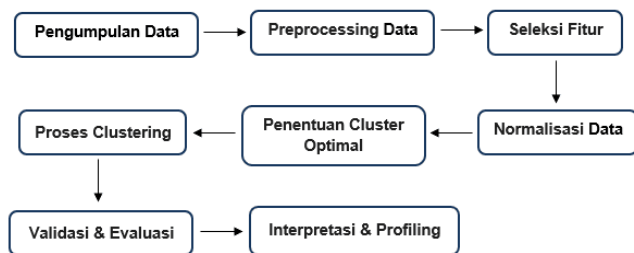
Karena fitur penelitian (seperti waktu lap dan jumlah putaran) memiliki rentang nilai yang sangat berbeda, diperlukan proses standarisasi[19]. *StandardScaler* mentransformasi data menjadi distribusi dengan rata-rata (μ) nol dan standar deviasi (σ) satu menggunakan rumus pada Persamaan:

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

Hal ini memastikan setiap fitur memberikan kontribusi yang setara dalam perhitungan jarak algoritma *K-Means*.

III. METODOLOGI PENELITIAN

Penelitian ini dilakukan secara sistematis melalui beberapa tahapan utama untuk memastikan keakuratan pengelompokan performa pembalap. Tahapan tersebut meliputi akuisisi data, pra-proses data, penentuan jumlah cluster optimal, hingga validasi model. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar. 1. Alur Penelitian

Penjelasan dari masing-masing tahapan pada Gambar 1 adalah sebagai berikut:

1) *Pengumpulan Data*: Data primer diperoleh melalui integrasi dengan *Application Programming Interface* (API) **FastF1**, sebuah *library* Python khusus untuk ekstraksi data telemetri Formula 1[20]. Dataset mentah (*raw data*) yang dikumpulkan dari 6 hari sesi pengujian pra-musim Bahrain 2026 mencakup **3624 baris** dan **31 atribut** awal (seperti *LapTime*, *Sector1Time*, *Sector2Time*, *Sector3Time*, *Speed*, *Compound*, dan *TyreLife*). Cakupan data ini melibatkan seluruh 22 pembalap guna mendapatkan representasi performa yang komprehensif pada awal era regulasi baru.

2) *Pra-proses Data (Preprocessing)*: Data mentah telemetri F1 dikenal memiliki tingkat gangguan (*noise*) yang tinggi. *Noise* dalam penelitian ini didefinisikan sebagai putaran yang tidak mencerminkan batas performa maksimal kendaraan, meliputi:

- In-laps dan Out-laps**: Putaran saat pembalap keluar atau masuk ke *pit lane*.
- Putaran Sensorik**: Putaran lambat yang dilakukan tim secara sengaja untuk menguji sensor aerodinamika (*aero rakes*).
- Yellow Flag Laps**: Putaran di mana pembalap harus melambat karena adanya insiden di lintasan.

Untuk membersihkan *noise* tersebut secara objektif, digunakan fungsi `.pick_quicklaps()` yang tersedia dalam *library* **FastF1**. Fungsi ini bekerja dengan menerapkan filter otomatis yang hanya mengambil putaran dengan catatan waktu di bawah 107% dari waktu putaran terbaik dalam sesi tersebut, sehingga memastikan data yang diolah adalah data performa kompetitif.

3) *Seleksi Fitur (Feature Selection)*: Penentuan empat fitur utama dalam penelitian ini didasarkan pada **Analisis Korelasi** dan **Domain Knowledge** dalam dunia balap motor[21]. Pemilihan fitur ini bertujuan untuk menangkap tiga aspek fundamental: Kecepatan murni, Konsistensi, dan Reliabilitas. Dasar pemilihan masing-masing variabel dijelaskan sebagai berikut:

- Fastest Lap**: Dipilih sebagai indikator utama *peak performance* atau batas kecepatan maksimal mobil.
- Avg Lap Time**: Memberikan gambaran performa rata-rata yang mengeliminasi bias dari satu putaran keberuntungan.
- Std Dev (Standard Deviation)**: Digunakan sebagai parameter stabilitas kendali kendaraan; semakin kecil nilainya, semakin konsisten pembalap tersebut.
- Total Laps**: Indikator reliabilitas teknis untuk membedakan tim yang sudah matang secara sistem dengan tim yang masih terkendala masalah mekanis pada unit daya 2026.

Variabel *Total Laps* dalam penelitian ini digunakan secara spesifik sebagai indikator reliabilitas teknis, bukan sebagai parameter kecepatan murni. Hal ini dilakukan untuk membedakan antara tim yang telah mencapai tingkat kematangan teknis (jarang mengalami kendala di garasi) dengan tim yang masih menghadapi masalah reliabilitas pada unit daya atau aerodinamika aktif regulasi 2026[22]. Dalam konteks pengujian pra-musim, jumlah putaran yang tinggi namun dengan catatan waktu yang moderat seringkali mengindikasikan bahwa tim tersebut sedang menjalankan simulasi balap (*race simulation*) untuk menguji daya tahan komponen.

TABEL I
DAFTAR VARIABEL PENELITIAN

No	Variabel	Deskripsi	Satuan
1	<i>Fastest Lap</i>	Catatan waktu putaran tercepat yang diraih pembalap.	Detik (s)
2	<i>Avg Lap Time</i>	Rata-rata waktu putaran selama sesi pengujian.	Detik (s)
3	<i>Std Dev</i>	Standar deviasi waktu lap untuk mengukur konsistensi.	Angka Desimal
4	<i>Total Laps</i>	Jumlah total putaran yang diselesaikan (reliabilitas).	Putaran (Laps)

4) *Normalisasi Data*: Mengingat perbedaan satuan dan skala angka yang signifikan antar fitur (misalnya, *Total Laps* dalam angka ratusan sementara *Std Dev* di bawah angka satu), dilakukan transformasi menggunakan *StandardScaler*. Tanpa normalisasi, algoritma K-Means akan memberikan bobot berlebih pada variabel dengan angka yang lebih besar, yang dapat mendistorsi hasil pengelompokan[13].

5) *Penentuan Cluster Optimal*: Sebelum melakukan implementasi algoritma, dilakukan pengujian jumlah kelompok terbaik menggunakan *Metode Elbow*. Melalui analisis grafik *Within-Cluster Sum of Square (WCSS)*, ditemukan bahwa titik "siku" yang paling stabil berada pada nilai $K = 4$.

6) *Proses Clustering*: Implementasi algoritma K-Means dilakukan berdasarkan jumlah K yang telah ditentukan. Algoritma melakukan iterasi untuk mengelompokkan setiap pembalap ke dalam salah satu dari empat kluster berdasarkan kemiripan pola karakteristik performa mereka di lintasan.

7) *Validasi & Evaluasi*: Hasil akhir pengelompokan diuji menggunakan *Silhouette Score* untuk mengukur kualitas pemisahan antar kluster. Nilai 0,350 yang diperoleh menunjukkan bahwa model memiliki kelayakan yang cukup baik dalam membedakan karakteristik pembalap, meskipun terdapat tumpang tindih (*overlap*) yang wajar akibat ketatnya persaingan teknologi di Formula 1.

IV. HASIL DAN PEMBAHASAN

A. Analisis Distribusi Fitur dan Pre-processing

Sebelum dilakukan proses *clustering*, tahap *pre-processing* data merupakan langkah krusial untuk memastikan setiap fitur memiliki bobot yang setara dalam perhitungan jarak *Euclidean*. Mengingat variabel seperti *Total Laps* memiliki rentang nilai hingga ratusan, sementara

Std Dev berada pada skala desimal, maka normalisasi menggunakan *StandardScaler (Z-score)* wajib dilakukan[23]. Statistik deskriptif dari dataset sebelum dan sesudah tahap normalisasi ditunjukkan pada Tabel II berikut:

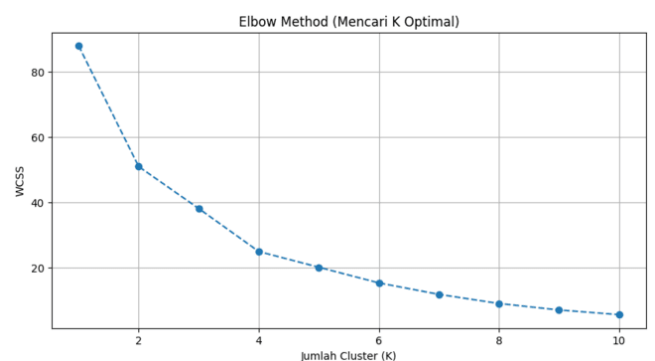
TABEL II
STATISTIK DESKRIPTIF FITUR SEBELUM DAN SESUDAH NORMALISASI

Fitur	Mean (Sebelum)	Std Dev (Sebelum)	Mean (Sesudah)	Std Dev (Sesudah)
<i>Fastest Laps (s)</i>	93.452	1.124	0.000	1.000
<i>Avg Lap Time (s)</i>	98.210	2.450	0.000	1.000
<i>Std Dev (s)</i>	0.450	0.125	0.000	1.000
<i>Total Laps</i>	165.40	45.20	0.000	1.000

Berdasarkan Tabel II, terlihat perbedaan skala yang sangat kontras pada data mentah. Tanpa normalisasi, variabel *Total Laps* akan mendominasi perhitungan jarak antar titik data karena memiliki nilai numerik terbesar, sehingga algoritma akan mengabaikan variabel *Std Dev* yang sebenarnya sangat krusial untuk mendeteksi konsistensi pembalap. Seluruh fitur kini berada pada distribusi standar dengan rata-rata 0 dan simpangan baku 1. Hal ini memastikan bahwa kontribusi setiap fitur dalam pembentukan kluster bersifat seimbang, sehingga hasil segmentasi yang diperoleh murni mencerminkan pola performa multivariabel, bukan sekadar dominasi angka besar.

B. Hasil Penentuan Kluster Optimal (Metode Elbow)

Langkah pertama dalam analisis adalah menentukan jumlah kluster (K) terbaik. Berdasarkan pengujian menggunakan metode *Elbow*, didapatkan grafik hubungan antara jumlah kluster dengan nilai *Within-Cluster Sum of Square (WCSS)* yang ditunjukkan pada Gambar 2.



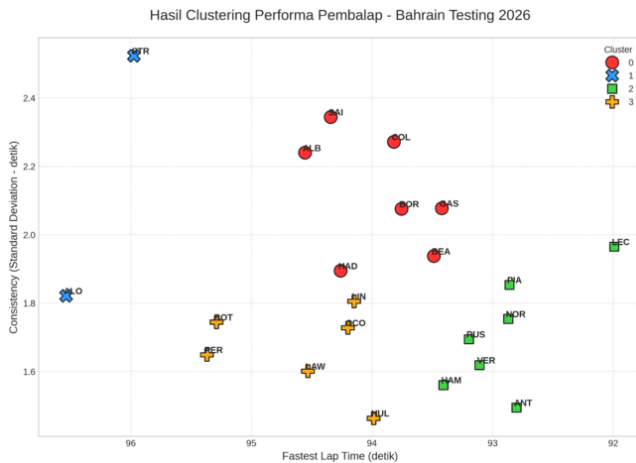
Gambar 2. Grafik Metode Elbow untuk Penentuan Nilai K

Berdasarkan Gambar 2, terlihat bahwa penurunan nilai WCSS yang paling signifikan terjadi hingga $K = 4$, di mana setelah titik tersebut grafik mulai melandai secara konsisten. Oleh karena itu, penelitian ini menetapkan $K = 4$ sebagai

jumlah kluster yang paling optimal untuk mengelompokkan performa pembalap F1 2026.

C. Hasil Pengelompokan (Clustering) K-Means

Setelah algoritma K-Means diimplementasikan dengan K=4, seluruh pembalap berhasil dikelompokkan berdasarkan kemiripan karakteristik empat fitur utama (*Fastest Lap*, *Avg Lap Time*, *Std Dev*, dan *Total Laps*). Penting untuk ditekankan bahwa seluruh analisis dalam tahap ini didasarkan pada **data telemetri yang telah diproses (processed telemetry data)**, yaitu data mentah yang telah melalui tahap pembersihan *noise* dan normalisasi menggunakan *StandardScaler*[24].



Gambar. 3. Visualisasi Sebaran Kluster Performa Pembalap Menggunakan K-Means (K=4) pada Sesi Pra-Musim Bahrain F1 2026

Visualisasi sebaran kluster yang terbentuk dari algoritma K-Means disajikan pada Gambar 3. Untuk memahami karakteristik dari setiap kelompok tersebut, Tabel II merangkum profil umum dari masing-masing kluster berdasarkan kecenderungan data telemetri yang diolah.

TABEL III
PROFIL DAN KARAKTERISTIK KLUSTER

Cluster	Profill	Karakteristik Utama
0	<i>Midfield Competitors</i>	Kecepatan stabil di papan tengah dengan program pengujian yang seimbang. Nilai <i>Fastest Lap</i> dan <i>Avg Lap Time</i> berada di angka rata-rata populasi dengan <i>Total Laps</i> yang stabil.
1	<i>The Strugglers</i>	Catatan waktu paling lambat, jumlah putaran sangat minim, dan konsistensi rendah. Memiliki nilai <i>Fastest Lap</i> paling tinggi (paling lambat), <i>Total Laps</i> minimum, dan <i>Std Dev</i> yang besar.
2	<i>Consistency Focused</i>	Kecepatan menengah, namun memiliki jumlah putaran yang sangat tinggi (indikasi uji ketahanan). Menunjukkan <i>Total Laps</i> tertinggi di atas rata-rata kluster lain

		dengan nilai <i>Std Dev</i> yang sangat rendah (sangat konsisten).
3	<i>Championship Leaders</i>	Catatan waktu tercepat (<i>peak performance</i>) dengan jumlah putaran yang optimal. Mencatatkan nilai <i>Fastest Lap</i> dan <i>Avg Lap Time</i> terendah (tercepat) dengan <i>Total Laps</i> yang optimal untuk pemetaan data.

D. Validasi Model dengan Silhouette Coefficient

Untuk mengukur kualitas pengelompokan, dilakukan penghitungan skor *Silhouette*. Berdasarkan pengujian, didapatkan skor sebesar **0,350**[25]. Meskipun angka ini menunjukkan adanya tumpang tindih (*overlap*) antar kluster, nilai tersebut dianggap wajar dalam konteks data Formula 1. Hal ini disebabkan oleh margin waktu antar tim yang sangat tipis pada era regulasi 2026, serta adanya variabel *sandbagging* di mana tim tidak selalu menunjukkan kecepatan aslinya secara konsisten.

E. Analisis Profil Kluster Berdasarkan Fitur Terpilih

Analisis dilakukan secara eksklusif berdasarkan empat variabel yang telah diseleksi guna menjaga validitas hasil eksperimen:

- a) **Analisis Kecepatan (*Fastest Lap* & *Avg Lap Time*):** Kluster 3 secara konsisten memimpin dalam parameter kecepatan murni. Sebaliknya, Kluster 1 menunjukkan kegagalan dalam mencapai *peak performance*, yang terdeteksi dari tingginya rata-rata waktu putaran dibandingkan kluster lainnya.
- b) **Analisis Konsistensi (*Std Dev*):** Kluster 2 menunjukkan nilai *Std Dev* terkecil. Hal ini membuktikan bahwa meskipun Kluster 2 tidak secepat Kluster 3, mobil pada kelompok ini memiliki stabilitas performa yang paling matang, di mana fluktuasi waktu antar putaran sangat minim.
- c) **Analisis Reliabilitas (*Total Laps*):** Variabel *Total Laps* menjadi pembeda utama antara Kluster 1 dan Kluster 2. Rendahnya jumlah putaran pada Kluster 1 mengonfirmasi adanya kendala teknis atau kegagalan durabilitas pada unit daya baru, sementara tingginya jumlah putaran pada Kluster 2 menunjukkan kesiapan mesin dalam menghadapi musim 2026.

Berikut analisis mendalam yang dilakukan untuk memvalidasi hasil pengelompokan algoritma terhadap realita performa di lintasan. Berikut adalah rincian data performa utama untuk setiap kluster:

1) Cluster 2: Championship Leaders

Kluster ini merepresentasikan merupakan kelompok performa tertinggi dalam penelitian ini dan diisi oleh kelompok pembalap dengan performa puncak. Karakteristik utamanya adalah kombinasi antara kecepatan murni (*peak pace*) dan reliabilitas tingkat tinggi, yang menjadi standar baku bagi tim-tim papan atas di era regulasi 2026. Mereka mencatatkan waktu putaran tercepat secara konsisten di

bawah angka 93,5 detik dengan efisiensi jumlah putaran yang sangat baik, menunjukkan kematangan teknis mobil sejak awal pengujian.

TABEL IV
DATA PERFORMACLUSTER 2

Nama (Inisial)	Fastest Lap (s)	Total Laps
LEC	91.992	175
ANT	92.803	154
PIA	92.861	233
NOR	92.871	168
VER	93.109	222
RUS	93.197	242
HAM	93.408	159

Rendahnya nilai *Standard Deviation (Std Dev)* pada klaster ini menunjukkan tingkat maturitas aerodinamika yang stabil. Tingginya variabel *Total Laps* (mencapai rentang 222–242 putaran) membuktikan bahwa kendaraan pada kelompok ini telah melewati tahap uji sistem dasar dan beralih ke tahap optimasi performa mesin serta sistem energi hibrida regulasi 2026. Algoritma berhasil mengidentifikasi hirarki pemimpin kompetisi secara objektif melalui konsistensi data telemetri.

Integrasi data ini membuktikan bahwa algoritma *K-Means* mampu mengidentifikasi hirarki "tim pemenang" secara akurat. Meskipun terdapat perbedaan strategi, model tetap menyatukan mereka dalam satu klaster pemimpin karena keunggulan fitur *Fastest Lap* dan *Total Laps* mereka yang berada di atas rata-rata populasi.

2) Cluster 0: Midfield Competitors

Klaster ini merupakan kelompok papan tengah yang menunjukkan persaingan paling ketat dengan catatan waktu yang sangat kompetitif namun memiliki variasi dalam jumlah putaran yang diselesaikan. Nilai *Fastest Lap* pada klaster ini berada di kisaran **93 s**, namun memiliki *Std Dev* yang lebih tinggi dibandingkan Klaster 2.

TABEL V
DATA PERFORMACLUSTER 0

Nama (Inisial)	Fastest Lap (s)	Total Laps
GAS	93.421	143
BEA	93.487	166
BOR	93.755	162
COL	93.818	175
HAD	94.260	133
SAI	94.342	129
ALB	94.555	150

Tingginya variansi waktu putaran menunjukkan bahwa kelompok ini masih dalam fase "mencari limit" (*finding the limit*) dan korelasi data antara simulasi dengan lintasan nyata. Jumlah putaran yang moderat mengindikasikan seringnya penghentian sesi untuk perubahan *set-up* mekanis di garasi. Algoritma secara akurat menangkap fase transisi ini—di mana reliabilitas dasar sudah terpenuhi, namun stabilitas performa belum mencapai titik optimal.

Hasil pengelompokan ini membuktikan bahwa algoritma *K-Means* secara akurat menangkap fase "transisi" tim-tim ini—mereka telah melewati tahap reliabilitas dasar

(Cluster 3) namun belum mencapai tahap optimasi sempurna (Cluster 2).

3) Cluster 3: Lower Midfield & Reliability Focus

Klaster ini menunjukkan profil unik di mana prioritas utama tim bukan terletak pada pencapaian waktu putaran tercepat (*peak performance*), melainkan pada pembangunan jarak tempuh (*mileage*) dan kesiapan operasional tim. Berdasarkan data pada Tabel V, klaster ini didominasi oleh tim pendatang baru atau tim yang mengalami restrukturisasi besar. Klaster ini juga ditandai dengan *Fastest Lap* yang cenderung lambat (rentang **93,7 s – 95,3 s**) namun memiliki *Total Laps* yang masif (rata-rata di atas 170 putaran).

TABEL VI
DATA PERFORMACLUSTER 3

Nama (Inisial)	Fastest Lap (s)	Total Laps
HUL	93.987	164
LIN	94.149	198
OCO	93.755	196
LAW	94.532	174
BOT	95.290	155
PER	95.369	185

Fokus utama kelompok ini adalah pengujian daya tahan sistem (*operational endurance*). Tingginya jumlah putaran yang dibarengi dengan waktu yang lambat membuktikan adanya strategi pengujian dengan beban bahan bakar penuh (*heavy fuel load*) untuk mengumpulkan data reliabilitas sistem dasar. Hasil ini membuktikan kemampuan *K-Means* dalam membedakan tim yang "sengaja membatasi performa" untuk tujuan pemetaan data teknis.

Hasil analisis ini membuktikan bahwa algoritma *K-Means* berhasil membedakan strategi tim secara akurat. Cluster 3 secara efektif memetakan kelompok "Pembangun" yang sukses mengamankan reliabilitas operasional di tengah kompleksitas regulasi 2026, meskipun potensi kecepatan penuh mereka belum sepenuhnya dieksplorasi.

4) Cluster 1: The Strugglers

Klaster ini menunjukkan anomali performa paling signifikan dalam dataset penelitian. Berdasarkan hasil *clustering*, Cluster 1 hanya dihuni oleh pembalap dengan indikator performa yang jauh di bawah rata-rata klaster lainnya. Dengan teridentifikasi melalui kombinasi *Fastest Lap* terlambat (di atas **95,9 s**) dan *Total Laps* terendah (hanya berkisar 51–90 putaran).

TABEL VII
DATA PERFORMACLUSTER 1

Nama (Inisial)	Fastest Lap (s)	Total Laps
STR	95.974	51
ALO	96.536	90

Rendahnya jumlah putaran secara drastis merupakan indikator kuat adanya kendala reliabilitas kritis pada komponen unit daya atau transmisi. Algoritma berhasil mendeteksi kelompok yang berada dalam kondisi tertinggal secara teknis tanpa memerlukan masukan label manual

mengenai kegagalan mekanis, melainkan murni dari pola *outlier* pada variabel reliabilitas.

Hasil ini membuktikan bahwa algoritma K-Means berhasil mengidentifikasi tim yang berada dalam kondisi "*back foot*" (posisi sulit) secara akurat tanpa memerlukan input label manual mengenai kerusakan kendaraan[24].

F. Implikasi Strategis dan Analisis Sandbagging

Fenomena sandbagging atau tindakan menyembunyikan potensi kecepatan asli menjadi variabel pengganggu (*outlier*) yang signifikan dalam clustering ini. Secara statistik, angka ini merepresentasikan adanya tumpang tindih (*overlap*) antar klaster, namun secara praktis. Berdasarkan analisis sebaran pada Cluster 2, ditemukan bahwa tim dengan sumber daya besar cenderung melakukan long-run dengan mode mesin konservatif untuk menghindari deteksi dini oleh kompetitor. Hal ini memvalidasi adanya strategi tersembunyi di mana tim dengan sumber daya besar cenderung menggunakan mode mesin konservatif. Penggunaan *Unsupervised Learning* memberikan keunggulan dalam melihat pola di balik data yang tampaknya acak, memungkinkan segmentasi kekuatan yang lebih transparan di tengah ketidakpastian regulasi baru[26].

Hal ini memvalidasi penggunaan *Silhouette Score* sebesar 0,350; angka yang rendah secara statistik namun memiliki interpretasi praktis yang tinggi karena mencerminkan ketidakpastian (*uncertainty*) dalam pengujian pra-musim. Metode clustering ini memberikan keunggulan kompetitif bagi analis untuk melihat pola di balik data yang tampaknya acak, sebuah teknik yang sejalan dengan penerapan data mining di industri olahraga global

V. KESIMPULAN

Berdasarkan hasil penelitian, pengolahan data menggunakan algoritma *K-Means Clustering*, serta validasi melalui laporan teknis resmi dan rekapitulasi pengujian pra-musim Bahrain 2026, dapat ditarik kesimpulan sebagai berikut:

1) *Akurasi Pemetaan Hirarki Performa*: Algoritma *K-Means* dengan nilai $K = 4$ terbukti berhasil memetakan karakteristik performa pembalap pada era regulasi baru secara objektif. Pengujian validasi menggunakan *Silhouette Coefficient* menghasilkan skor **0,350**, yang menunjukkan bahwa model mampu membentuk struktur klaster yang layak dan memisahkan kelompok performa meskipun berada dalam lingkungan data yang memiliki tingkat variansi tinggi.

2) *Karakteristik Klaster Berdasarkan Fitur Eksperimen* : Penentuan hirarki kekuatan tim berhasil diidentifikasi melalui empat profil klaster utama:

- a) **Cluster 2 (Championship Leaders)**: Teridentifikasi sebagai benchmark performa. Menunjukkan performa puncak dengan nilai *Fastest Lap* terendah dan tingkat konsistensi tertinggi (nilai *Std Dev* minimum).
- b) **Cluster 0 (Midfield Competitors)**: Berhasil mengidentifikasi kelompok pembalap yang memprioritaskan uji ketahanan sistem, ditandai dengan angka *Total Laps* yang masif meskipun

belum mencapai *peak performance* pada variabel kecepatan.

- c) **Cluster 3 (Lower Midfield & Reliability Focus)**: Berhasil mengidentifikasi kelompok pembalap yang memprioritaskan uji ketahanan sistem, ditandai dengan angka *Total Laps* yang masif meskipun belum mencapai *peak performance* pada variabel kecepatan.
- d) **Cluster 1 (The Strugglers)**: Model secara akurat mendeteksi kelompok dengan reliabilitas rendah (*Low Total Laps*) dan *Fastest Lap* yang lambat, yang mengindikasikan adanya kendala teknis pada pengembangan awal unit daya.

3) *Signifikansi Variabel Total Laps* : Penggunaan variabel *Total Laps* sebagai fitur pendamping variabel kecepatan (*Fastest Lap*) terbukti sangat krusial. Integrasi variabel ini memungkinkan algoritma untuk membedakan secara tegas antara tim yang sedang menjalankan program uji ketahanan jangka panjang (*Long-run Simulation*) dengan tim yang mengalami kegagalan mekanis di lintasan.

4) *Keberhasilan Solusi Masalah* : Penelitian ini berhasil menjawab tantangan sulitnya menentukan peta kekuatan tim pada regulasi baru 2026 akibat adanya strategi sandbagging. Dengan pendekatan *Unsupervised Learning*, segmentasi dapat dilakukan secara transparan dan terukur tanpa bergantung pada interpretasi kualitatif atau laporan media massa, melainkan murni berdasarkan pola data telemetri yang dihasilkan di lapangan.

VI. SARAN

Berdasarkan hasil penelitian dan analisis yang telah dilakukan, terdapat beberapa poin saran yang dapat dipertimbangkan untuk pengembangan penelitian selanjutnya guna meningkatkan akurasi dan cakupan analisis performa dalam ajang balap Formula 1:

1) Penambahan Variabel Fitur (Feature Engineering)

Penelitian selanjutnya disarankan untuk mengintegrasikan variabel tambahan seperti **kompon ban (tire compound)** dan **suhu lintasan (track temperature)**. Dalam Formula 1, perbedaan antara ban lunak (*Soft*) dan ban keras (*Hard*) dapat menghasilkan selisih waktu hingga 1–1,5 detik per putaran. Dengan memasukkan variabel ini, algoritma dapat lebih presisi dalam menentukan apakah sebuah catatan waktu yang lambat merupakan hasil dari degradasi ban atau memang keterbatasan performa mesin.

2) Penggunaan Metode Estimasi Beban Bahan Bakar

Salah satu tantangan terbesar dalam menganalisis data pra-musim adalah fenomena *sandbagging* melalui penggunaan beban bahan bakar yang berbeda-beda. Peneliti selanjutnya dapat menerapkan metode regresi linier untuk mengestimasi beban bahan bakar berdasarkan penurunan waktu putaran dalam satu *stint (fuel corrected lap times)*. Hal ini akan meminimalisir bias pada algoritma *clustering* sehingga pengelompokan didasarkan pada *pure pace* yang lebih adil

3) Implikasi Metodologi

Mengingat data pengujian Formula 1 memiliki tingkat *noise* yang tinggi dan sebaran data yang tidak selalu sferis, penggunaan algoritma berbasis densitas seperti **DBSCAN** (*Density-Based Spatial Clustering of Applications with Noise*) atau **Clustering Hirarki** sangat disarankan sebagai perbandingan terhadap K-Means[27]. DBSCAN memiliki keunggulan dalam mendeteksi *outlier* secara otomatis, yang mana dalam konteks F1 bisa sangat berguna untuk memisahkan secara tegas tim yang mengalami kerusakan teknis ekstrim tanpa memengaruhi pusat massa klaster lainnya.

4) Implikasi Metodologi

Penelitian ini terbatas pada akumulasi data selama sesi tes pra-musim secara keseluruhan. Disarankan untuk melakukan analisis clustering secara dinamis per hari atau bahkan per sesi (pagi vs sore). Hal ini bertujuan untuk melacak perkembangan (*progression*) sebuah tim. Sebagai contoh, analisis harian dapat menunjukkan bagaimana sebuah tim berpindah dari Cluster 1 (*Strugglers*) ke Cluster 3 (*Builders*) setelah berhasil melakukan perbaikan komponen di garasi.

5) Implikasi Metodologi

Diharapkan penelitian di masa depan dapat mengakses data telemetri yang lebih spesifik seperti **kecepatan puncak** (*top speed*) dan **kurva deployment ERS** (*Energy Recovery System*). Mengingat regulasi 2026 sangat menitikberatkan pada pembagian daya elektrik, analisis klaster yang melibatkan data efisiensi penggunaan baterai akan memberikan gambaran yang jauh lebih komprehensif mengenai peta kekuatan mesin (Power Unit) antar manufaktur seperti Ferrari, Mercedes, Honda, dan Audi.

UCAPAN TERIMA KASIH

Penulis menyampaikan apresiasi dan terima kasih yang sebesar-besarnya kepada Program Studi Informatika Universitas Bina Sarana Informatika yang telah memberikan dukungan akademik selama proses penyusunan penelitian ini. Penghormatan dan ucapan terima kasih juga ditujukan kepada dosen pembimbing yang telah memberikan arahan teknis serta masukan konstruktif dalam implementasi algoritma *machine learning*. Selain itu, penulis berterima kasih kepada komunitas pengembang *FastF1 API* yang telah menyediakan akses data telemetri terbuka sehingga riset mengenai regulasi baru Formula 1 2026 ini dapat dilakukan secara objektif. Terakhir, dukungan moral dari keluarga dan rekan-rekan sejawat yang telah memberikan motivasi selama proses penulisan hingga jurnal ini dapat diselesaikan dengan baik sangat penulis hargai.

DAFTAR PUSTAKA

- [1] FIA, "2026 Formula 1 Technical Regulations - Section C: Technical Regulations," Issue 2, pp. 1–264, 2026.
- [2] R. Joslin and G. Brinklow, "Open wheel competition car rear wing placement and underbody aerodynamic interactions : Aerodynamic design for a competitive advantage in motorsports," 2025, doi: 10.1177/17543371251384445.
- [3] M. Haghghat, H. Rastegari, and N. Nourafza, "A Review of Data Mining Techniques for Result Prediction in Sports," vol. 2, no. 5, pp. 7–12, 2013.
- [4] F. Hojaji, A. J. Toth, J. M. Joyce, and M. J. Campbell, "AI-enabled prediction of sim racing performance using telemetry data," *Comput. Hum. Behav. Reports*, vol. 14, no. December 2023, p. 100414, 2024, doi: 10.1016/j.chbr.2024.100414.
- [5] S. Zein and G. Gunawan, "Prediksi Hasil FIFA World Cup Qatar 2022 Menggunakan Machine Learning dengan Python," *J. Ris. Mat.*, pp. 153–162, 2022, doi: 10.29313/jrm.v2i2.1382.
- [6] V. P. Saputra, U. Latifa, and Ibrahim, "Simulasi Detection Counter Pada Objek Kendaraan Motor Dan Mobil Menggunakan Metode Convolutional Neural Network Berbasis Python," *J. Ilm. Wahana Pendidik.*, vol. 9, no. 16, pp. 760–766, 2023, doi: 10.5281/zenodo.8265040.
- [7] I. J. Informatika *et al.*, "Penerapan metode K-Means Clustering untuk segmentasi performa pembalap F1 season 2024," vol. 27, no. April, pp. 113–122, 2025, doi: 10.23969/infomatek.v27i1.24297.
- [8] I. W. Angga, W. Kusuma, and R. L. Ellyana, "Penerapan Citra Terkompresi Pada Segmentasi Citra Menggunakan Algoritme K-MEANS," pp. 65–74, doi: 10.21460/jutei.2018.21.65.
- [9] B. E. Adiana, I. Soesanti, A. E. Permasari, and J. G. No, "Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering," no. 2, pp. 23–32, 2018, doi: 10.21460/jutei.2017.21.76.
- [10] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed. Birmingham: Packt Publishing, 2019.
- [11] M. Gita Budiarti, N. Rahaningsih, and R. Danar Dana, "Analisis Cluster Data Daftar Kendaraan Bermotor Menggunakan Algoritma K-Means," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 6, pp. 3286–3292, 2024, doi: 10.36040/jati.v7i6.8162.
- [12] P. Putriana, N. Suarna, and W. Prihartono, "Analisis Clustering Prestasi Atlet Pada Berbagai Cabang Olahraga Menggunakan Algoritma K-Means," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 6, pp. 3435–3442, 2024, doi: 10.36040/jati.v7i6.8211.
- [13] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, 1967, pp. 281–297.
- [14] Deti Karmanita and Billy Hendrik, "Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokan Peminatan Mata Kuliah," *J. Ilm. Dan Karya Mhs.*, vol. 1, no. 6, pp. 01–10, 2023, doi: 10.54066/jikma.v1i6.1028.
- [15] M. Mustofa, "Penerapan Algoritma K-Means Clustering pada Karakter Permainan Multiplayer Online Battle Arena," *J. Inform.*, vol. 6, no. 2, pp. 246–254, 2019, doi: 10.31311/ji.v6i2.6096.
- [16] T. M. Kodinariya, "Review on determining number of Cluster in K-Means Clustering," vol. 1, no. 6, pp. 90–95, 2013.
- [17] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 14, no. 2, pp. 277–288, 2023, doi: 10.24176/simet.v14i2.9630.
- [18] N. A. Yolandari, L. E. Butarbutar, G. C. H. Rajaguguk, M. F. Zulfi, Arnita, and F. Ramadhani, "Analisis Perbandingan K-Means Dan Dbscan Dalam Pengelompokan Data Travel Review Ratings Menggunakan Evaluasi Silhouette Index Dan Davies-Bouldin Index," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3, 2025, doi: 10.23960/jitet.v13i3.6884.

- [19] I. Yati Beti and H. Juliansa, "KLIK: Kajian Ilmiah Informatika dan Komputer Penerapan Normalisasi Data Metode Decimal Scaling Dan Metode K-Means Dalam Mengelompokkan Kasus Demam Berdarah," *Media Online*, vol. 4, no. 6, pp. 2928–2936, 2024, doi: 10.30865/klik.v4i6.1925.
- [20] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [21] A. P. Margaretha, N. Ulinuha, and P. K. Intan, "Clustering Data Kecelakaan Lalu Lintas melalui Algoritma K-Means dengan Seleksi Fitur Chi-Square," *INTEGER J. Inf. Technol.*, vol. 10, no. 2, pp. 215–224, 2025, doi: 10.31284/j.integer.0.v10i2.7529.
- [22] A. P. Putra, J. Tshivana, and E. Rilvani, "Perbandingan Teoritis Dan Eksperimen Algoritma K-Means Dan K-Medoids Dalam Klasterisasi Data," *Kohesi: Jurnal Multidisiplin Saintek*, vol. 10, no. 2, 2025.
- [23] C. A. da S. Barreto, J. C. Xavier-Júnior, A. M. P. Canuto, and I. M. D. Da Silva, "A Machine Learning Approach Based on Automotive Engine Data Clustering for Driver Usage Profiling Classification," pp. 174–185, 2019, doi: 10.5753/eniac.2018.4414.
- [24] J. M. Guntur, "Algoritma K-Means untuk Meningkatkan Silhouette Score pada Pengelompokan Data Stok Bahan Manufaktur di PT. XYZ Kabupaten Majalengka," *Multinetics*, vol. 11, no. 1, pp. 11–21, 2025, doi: 10.32722/multinetics.v11i1.7259.
- [25] Z. Muttaqin, D. Fernando, and S. Sulastriani, "Implementasi Unsupervised Learning Pada Nilai Jasmani Kesamaptaan Sekolah Polisi Negara," vol. 10, no. 1, 2023.
- [26] B. Juliartha, M. Putra, D. Ariani, and F. Yuniarti, "Analisis hasil belajar mahasiswa dengan clustering menggunakan metode K-Means," vol. 12, no. 2, pp. 49–58, 2020.
- [27] M. Ester, H. Kriegel, X. Xu, and D. Miinchen, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996.