Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 9, No. 3, September 2024 (126-135) **https://doi.org/10.32493/informatika.v9i3.44640**

Analisis Visual dan Karakteristik Klub Sepakbola Liga Inggris Berdasarkan Pola Permainan Menggunakan *K-Means Clustering*

Rachmat Bintang Yudhianto[1*], Fajar Athallah Yusuf[2], Anwar Fitrianto[3], L.M. Risman Dwi Jumansyah[4]

[1, 2,3,4]Statistics and Data Science, IPB University, Jl. Raya Dramaga, Babakan, Kec. Dramaga, Kabupaten Bogor, Indonesia, 16680
e-mail: [1]ydth_2000_rachmat@apps.ipb.ac.id, [2]fajarathallah@apps.ipb.ac.id,
[3]anwarstat@gmail.com, [4]rismandwijumansyah@apps.ipb.ac.id
*Corresponding author

Submitted Date: September 20th, 2024 Reviewed Date: September 25th, 2024
Revised Date: November 6th, 2024 Accepted Date: November 29th, 2024

Abstract

This research aimed to analyze and cluster football teams in the English Premier League (EPL) for the 2023/2024 season based on their playing characteristics using K-Means clustering. Understanding the playing styles is essential for optimizing strategies and enhancing team performance. Preprocessing steps included data cleaning, feature engineering, and visualization of key features such as goals, shots, and attacking attempts. Four clusters were identified using the Elbow method, representing teams with varying levels of attacking and defensive capabilities. Evaluation of the clustering results was conducted using Davies-Bouldin (score: 0.47), Calinski-Harabasz (score: 275.89), and Silhouette (score: 0.53) metrics, indicating moderate clustering quality. The findings suggest that EPL teams tend to be attack-oriented, while defensive strength varies across clusters. Limitations in the dataset, such as the number of observations and features, impacted the analysis, and future studies may benefit from incorporating additional features and advanced dimensionality reduction techniques.

Keywords: English Premier League, K-Means, Playing Characteristics, Football Analytics, Cluster Evaluation, Feature Engineering

Abstrak

Penelitian ini bertujuan untuk menganalisis dan mengelompokkan tim-tim sepak bola di Liga Premier Inggris (EPL) musim 2023/2024 berdasarkan karakteristik permainan mereka menggunakan metode klastering K-Means. Memahami gaya permainan sangat penting untuk mengoptimalkan strategi dan meningkatkan kinerja tim. Langkah-langkah prapemrosesan meliputi pembersihan data, rekayasa fitur, dan visualisasi fitur-fitur kunci seperti gol, tembakan, dan upaya menyerang. Empat kluster diidentifikasi menggunakan metode Elbow, mewakili tim dengan tingkat kemampuan menyerang dan bertahan yang bervariasi. Evaluasi hasil klastering dilakukan menggunakan metrik Davies-Bouldin (skor: 0.47), Calinski-Harabasz (skor: 275.89), dan Silhouette (skor: 0.53), menunjukkan kualitas klastering yang moderat. Temuan menunjukkan bahwa tim-tim EPL cenderung berorientasi menyerang, sementara kekuatan bertahan bervariasi antar kluster. Keterbatasan dalam dataset, seperti jumlah observasi dan fitur, berdampak pada analisis, dan penelitian selanjutnya dapat memperoleh manfaat dari pengintegrasian fitur tambahan dan teknik reduksi dimensi yang lebih lanjut.

Kata Kunci: English Premier League, K-Means, Karakteristik Permainan, Analisis Sepakbola, Evaluasi Model, *Feature Engineering*

1. Pendahuluan

Sepakbola merupakan salah satu bidang olahraga yang diminati masyarakat secara luas serta memiliki basis penggemar yang masif di seluruh dunia. Menurut laporan dari Nielsen Sports "World Football 2018," lebih dari 40% populasi di 18 pasar utama di dunia tertarik pada sepakbola, menjadikannya olahraga terpopuler secara global.

Sepakbola juga mendominasi media sosial dengan angka pengikut yang sangat besar. Perkembangan ini diikuti dengan kemajuan dalam pengumpulan data besar yang memuat informasi penting mengenai performa dan karakteristik pemain dalam suatu tim di liga tertentu. Liga Inggris atau *English Premier League* (EPL) merupakan salah satu liga terbesar di dunia yang memuat data mengenai profil pemain, jumlah gol, jumlah umpan, hingga tim yang dihadapi dalam laga kandang maupun tandang (Foo *et al.* 2024). Seluruh informasi ini dihimpun menjadi dataset besar sehingga dapat dilakukan pengolahan untuk memahami karakteristik tiap tim selama beberapa musim terakhir (Herold *et al.* 2019).

Analisis gerombol merupakan salah satu teknik pengolahan data dengan pembelajaran tanpa pengawasan yang digunakan untuk melakukan pengelompokkan data ke dalam grup atau kluster berdasarkan kemiripan antar objek dalam dataset (Millati *et al.* 2021). Metode analisis gerombol telah diterapkan dalam berbagai bidang, termasuk salah satunya adalah di bidang pemasaran, biologi, dan pemrosesan data besar untuk menemukan pola tersembunyi dalam data. Dalam olahraga, khususnya sepakbola, analisis gerombol dapat digunakan untuk mengenali karakteristik tim dalam bermain, baik di laga kandang maupun tandang (Baboota dan Kaur 2019).

Analisis gerombol memiliki beberapa metode salah satunya *K-Means*. Metode *clustering* merupakan metode yang bekerja dengan membagi data ke dalam k *cluster*, setiap variabel dikelompokkan ke dalam kluster yang memiliki pusat atau centroid yang paling dekat berdasarkan kemiripan atribut tertentu (Wu 2024). Penerapan *clustering* dalam sepakbola dapat memperhatikan beberapa variabel seperti jumlah gol, *assist*, tekel, maupun stamina. Penerapan *K-Means* membutuhkan nilai k yang optimal untuk mendapatkan klasterisasi yang tepat sehingga beberapa metode optimasi seperti *Elbow Method* dapat digunakan untuk mencari k optimal dari dataset yang akan digunakan (Shi *et al.* 2021). Namun begitu, metode *K-Means* memiliki beberapa keterbatasan seperti kesensitifan terhadap titik awal *centroid* dan asumsi bentuk kluster yang bulat.

Metode lain sebagai alternatif, seperti *DBSCAN* dan *Hierarchical Clustering* dapat digunakan. Metode tersebut menawarkan keunggulan tertentu, antara lain seperti *DBSCAN* lebih fleksibel dalam menangani data dengan distribusi tidak seragam dan mampu mendeteksi *noise*, sementara *Hierarchical Clustering* dapat memberikan struktur hierarki antar kluster (Murtagh dan Contreras 2012). Metode yang telah disebut disisi lain memiliki keterbatasan. Metode *DBSCAN* sensitif terhadap parameter, sedangkan *Hierarchical Clustering* kurang efisien untuk dataset besar (Xu dan Tian 2015). Oleh karena itu, pemilihan metode *clustering* memerlukan justifikasi yang mendalam untuk memastikan relevansi dan efektivitasnya dalam konteks penelitian yang akan dilakukan.

Penelitian sebelumnya telah menjelaskan mengenai pengaplikasian visualisasi dan analisis sepakbola menggunakan teknik pengolahan data namun lebih banyak berfokus pada klasifikasi karakteristik pemain serta analisis kinerja pemain sepakbola di liga top Eropa. Penelitian yang dilakukan oleh Rommers (Rommers *et al.* 2020) menjelaskan mengenai variabel-variabel yang berpengaruh pada kerentanan cedera pemain sepakbola menggunakan metode *XGBoost* dengan hasil akurasi sebesar 84% dari total 800 amatan pada data yang digunakan, penelitian lainnya dilakukan oleh Al-Asadi (Al-Asadi dan Tasdemir 2022) dan Hewitt (Hewitt dan Karakuş 2023) pada tahun 2022 dan 2023 secara berturut-turut menerapkan teknik pengolahan data menggunakan pembelajaran mesin untuk menilai kualitas pemain dalam video game FIFA dan pengaruh posisi pemain dalam mendukung tercapainya eksptasi terjadinya gol (*Xg*). Kedua penelitian tersebut menggunakan model klasifikasi seperti *regresi linear* dan *random forest*, masing-masing penelitian mampu memvisualisasikan serta memprediksi beberapa peubah penting yang mempengaruhi kinerja pemain sepakbola sesuai dengan permasalahannya. Namun, pengelompokkan tim sepakbola berdasarkan pola permainan masih terbatas dalam penelitian sebelumnya sehingga perlu dilakukan pula penelitian ini dengan tujuan untuk mengenal lebih dalam karakteristik permainan tim yang bermain di EPL.

Pemaparan sebelumnya menjelaskan bahwa penelitian-penelitian sebelumnya lebih berfokus pada klasifikasi karakteristik pemain serta visualisasi faktor-faktor yang berpengaruh pada kinerja pemain sepakbola di liga top eropa, salah satu penelitian juga menggunakan dataset yang berasal dari Liga Primer Inggris. Penelitian sebelumnya secara khusus hanya berfokus pada klasifikasi pemain, visualisasi, atau analisis kinerja

pemain sedang penelitian mengenai pengelompokkan tim berdasarkan pola permainan menggunakan analisis gerombol masih terbatas. Berdasarkan beberapa hal tersebut, tujuan penelitian ini adalah untuk melakukan visualisasi dan klasterisasi tim-tim di Liga Inggris berdasarkan cara dan tipe permainan yang digunakan. Penelitian ini menggunakan analisis gerombol dengan metode *K-Means*, metode tersebut akan diuji nilai kebaikan modelnya menggunakan *Davies-Bouldin*, *Calinski-Harabasz* dan *Silhouette Score*.

2. Metode Penelitian

Langkah penelitian merupakan acuan terkait langkah-langkah yang akan dilakukan oleh penulis. Pada penelitian ini akan dilakukan analisis dan eksplorasi data pertandingan Liga Inggris musim 2014 - 2023, data ini didapatkan melalui *platform Football* Data API dengan jumlah amatan sebesar 3352. Data yang digunakan adalah sebagai berikut.

Tabel 1. Keterangan Data

Variabel	Keterangan
Home	Tim tuan rumah
Away	Tim tandang
Home_Goal	Jumlah Gol tuan rumah
Away_Goal	Jumlah Gol tandang
Home_Corner	Jumlah tendangan pojok tuan rumah
Away_Corner	Jumlah tendangan pojok tandang
Home_Attack	Jumlah serangan tuan rumah
Away_Attack	Jumlah serangan tandang
Home_Shots	Jumlah tembakan tuan rumah
Away_Shots	Jumlah tembakan tandang

Penelitian ini akan menggunakan metode *K-Means* untuk memetakan dan mengklasterisasi tim-tim yang bermain di liga inggris berdasarkan karekteristik permainannya. Evaluasi kebaikan model yang digunakan pada penelitian ini menggunakan *Davies-Bouldin*, *Calinski-Harabasz* dan *Silhouette Score*.

2.1. Sepak bola

Sepak bola merupakan olahraga tim yang dimainkan oleh dua tim, masing-masing tim berisikan 11 pemain dengan tujuan untuk mencetak gol ke gawang lawan. Permainan tim ini dilakukan dalam lapangan berbentuk persegi panjang dengan dua gawang di setiap ujungnya dan dalam permainan sepak bola hanya dapat menggunakan kaki dan kepala serta tidak diperbolehkan untuk menggunakan tangan (Andreff 2011). Sepak bola hingga saat ini telah berkembang menjadi fenomena global dan telah dianggap sebagai olahraga terpopuler saat ini, hal ini dilihat dari minat orang di seluruh dunia atas bidang olahraga ini dan berdasarkan survei terpopuler, Liga Inggris menjadi liga yang paling banyak menarik perhatian di seluruh dunia (Bond *et al.* 2021).

2.2. K-Means Clustering

K-Means Clustering merupakan suatu teknik pengelompokkan yang diaplikasikan berdasarkan *Partioned Clustering* dan memiliki prinsip kerja dari pengelompokkan *Hierarcial Clustering* serta dilakukan secara bertahap. Hasil kluster yang didapat dari metode *K-Means* tergantung pada inisiasi nilai pusat awal kluster yang diberikan. Teknik pengklasteran *K-Means* ini menggunakan ukuran ketidakmiripan untuk pengelompokkan obyeknya. Apabila setelah menggunakan metode *K-Means Clustering* objek yang dikelompokkan memiliki jarak yang dekat maka objek tersebut dapat dikatakan mirip (Kumar *et al.* 2020).Teknik pengelompokkan menggunakan metode *K-Means Clustering* ini adalah dengan mengumpulkan data-data yang memiliki karakteristik sama kedalam satu kluster serta data yang memiliki karakteristik yang berbeda kedalam satu kluster sehingga didapati data yang berada di dalam satu kluster memiliki tingkat variasi yang lebih kecil. Fungsi dari metode *K-Means Clustering* dijelaskan pada rumus berikut.

$$E = \sum_{i=1}^{k} \sum_{p \in ci} \|p - mi\|^2 \qquad (1)$$

Keterangan :

1) E: Total kesalahan kuadrat atau fungsi objektif yang ingin diminimalkan oleh algoritma *K-Means*.
2) k: Jumlah klaster.
3) i: Indeks untuk klaster.
4) $p \in ci$: Menunjukkan bahwa titik data ppp adalah anggota dari klaster ci.
5) p: Titik data dalam dataset.
6) m_i: Titik pusat atau centroid dari klaster i.

$\| p - mi \|^2$: Jarak kuadrat antara titik data ppp dan pusat klaster mim_imi.

2.3. Davies-Bouldin

Metode *Davies Bouldien Index (DBI)* merupakan teknik yang digunakan untuk menilai kualitas dari kluster dengan melihat kedekatan antar hasil kluster yang dihasilkan, apabila nilai DBI yang dihasilkan dari proses *clustering* cenderung lebih dekat dengan nilai 0 maka hasil kluster tersebut akan semakin baik (Vergani dan Binaghi 2018).

2.4. Calinski-Harabasz

Indeks Calinski Harabasz merupakan ukuran yang digunakan untuk mengevaluasi hasil *clustering* dengan menentukan nilai kluster yang telah ditentukan. Proses dilanjutkan dengan memprediksi tiap nilai k *cluster* yang digunakan dimana setiap k memiliki nilai *Calinski Harabasz* masing-masing. Nilai *Calinski-Harabasz* terkecil yang didapat merupakan nilai k terbaik dari proses akhir pengevaluasian (Firman Ashari *et al.* 2023).

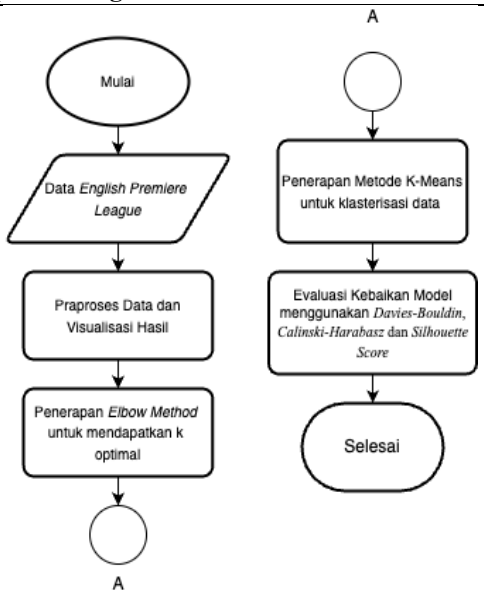
2.5. Silhouette

*Silhouette Coefficien*t merupakan salah satu metode evaluasi untuk menguji kualitas *cluster* yang akan digunakan. Menggabungkan *cohesion* dan *separation*, metode ini bekerja dengan mengukur seberapa dekat hubungan antara objek dalam sebuah kluster serta melihat seberapa jauh kluster yang terbentuk terpisah dari kluster lainnya (Pratama Simanjuntak dan Khaira 2021). Persamaan yang digunakan adalah sebagai berikut.

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \qquad (2)$$

Keterangan :

1) a(i) adalah jarak rata-rata antara objek *i* dengan semua objek lain dalam kluster yang sama.
2) b(i) adalah jarak rata-rata antara objek *i* dan semua objek lain dalam kluster terdekat yang tidak termasuk objek *i.*



Gambar 1. Kerangka Penelitian Penerapan Data *EPL* Menggunakan Metode *K-Means*

Gambar 1 merupakan tahapan-tahapan atau kerangka penelitian yang digunakan dalam penelitian ini. Adapun tahapan-tahapan dalam penelitian ini sebagai berikut:

1. Praproses dan Visualisasi Data

Tahapan ini merupakan langkah awal setelah dilakukannya pemanggilan data *English Premiere League* (*EPL*). Berikut contoh data dalam table berikut;

Table 2. Variabel dalam dataset

…	home	…	home_attack	…
…	0.0	…	NaN	…
…	0.0	…	NaN	…
…	0.0	…	NaN	…

Proses yang dilakukan pada data antara lain pembersihan data yang mencakup penanganan *missing values* melalui *imputasi* data, dan normalisasi data. *Outlier* ditemukan setelah *imputasi* data tetapi tidak dilakukan penanganan karena tidak merusak analisis. Visualisasi kemudian dilakukan setelah praproses selesai dilakukan, dengan tujuan gambaran awal tentang pola yang ada di data, seperti distribusi *variabel* atau hubungan antar fitur. Selanjutnya, akan dilakukan juga mengenai rekayasa fitur berdasarkan *variabel* yang digunakan pada data *EPL*.

ISSN: 2541-1004
e-ISSN: 2622-4615
https://doi.org/10.32493/informatika.v9i3.44640

2. Pencarian k Optimal menggunakan *Elbow Method*

Teknik *Elbow* merupakan metode yang digunakan dengan tujuan untuk menentukan jumlah klaster optimal (k) dalam algoritma *K-Means*. Metode ini memplot nilai inertia (*sum of squared distances* dari setiap titik ke *centroid* terdekat) terhadap berbagai nilai k. Titik siku atau *elbow* yang ditampilkan melalui grafik nantinya menandakan nilai k yang optimal. Titik k yang dituju merupakan k optimal dimana penambahan jumlah klaster tidak lagi memberikan pengurangan signifikan dalam *inertia*.

3. Klasterisasi menggunakan metode *K-Means Clustering*

Setelah menentukan nilai k yang optimal melalui *Elbow Method*, langkah selanjutnya adalah mengaplikasikan algoritma *K-Means Clustering* untuk membagi data ke dalam k klaster. *K-Means* bekerja dengan mengelompokkan data berdasarkan kedekatan atau kesamaan antar titik data, di mana setiap klaster memiliki *centroid* (pusat klaster). Data *EPL* dalam hal ini akan diaplikasikan metode *K-Means Clustering* untuk mengklasterisasi tim-tim di Liga Inggris berdasarkan cara permainan mereka ketika bertanding. Algoritma menetukan k *centroid* awal secara acak sebagai pusat klaster, lalu data dihitung jaraknya ke centroid menggunakan *Eclidean Distance. Centroid* akan diperbarui sebagai rata-rata posisi pada klaster masing-masing dan terus diulang hingga posisi *centroid* stabil. Proses klasterisasi akan melibatkan variabel-variabel seperti jumlah gol dari tiap tim, tendangan pojok yang dimiliki dari tiap tim, dan variabel lainnya sehingga menghasilkan plot klaster dari berbagai variabel.

4. Evaluasi model

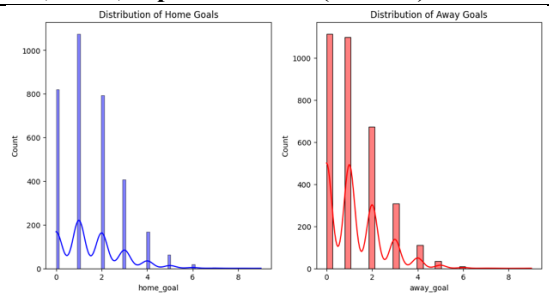
Setelah proses klasterisasi selesai, model perlu dievaluasi untuk menilai seberapa baik data telah dikelompokkan. Evaluasi ini menggunakan tiga metrik utama: *Davies-Bouldin Index, Calinski-Harabasz Index, dan Silhouette Scor*e. *Davies-Bouldin* Index mengukur rasio antara jarak antar klaster dengan ukuran klaster, di mana nilai yang lebih rendah menunjukkan pemisahan antar klaster yang lebih baik. *Calinski-Harabasz Index*, atau *Variance Ratio Criterion*, menghitung rasio antara penyebaran antar klaster dan penyebaran dalam klaster, dengan nilai yang lebih tinggi menunjukkan hasil klasterisasi yang lebih baik. Sementara itu, *Silhouette Score* mengukur seberapa dekat data dalam sebuah klaster dengan klaster lain, dengan nilai berkisar antara -1 hingga 1, di mana skor mendekati 1 menunjukkan klasterisasi yang optimal. Dengan menggunakan ketiga metrik ini, kualitas pengelompokan data dapat dinilai secara menyeluruh, sehingga dapat diketahui apakah model *K-Means* yang digunakan sudah berhasil memisahkan data dengan baik atau tidak.

3. Hasil dan Pembahasan

Hasil dan pembahasan diperoleh melalui beberapa tahapan metode praproses dan analisis data *EPL* menggunakan metode *K-Means Clustering*. Dalam beberapa tahapan meliputi hasil praproses dan visualisasi beberapa fitur dalam data, proses feature engineering untuk menghasilkan fitur atau amatan baru pada data, optimalisasi k serta hasil klasterisasi menggunakan metode *K-Means* dan dievaluasi berdasarkan *Davies-Bouldin*, *Calinski-Harabasz* dan *Silhouette Score*. Analisis klaster menggunakan metode *K-Means* dengan dataset pertandingan *EPL* bertujuan untuk mengenali karakteristik tim serta mengetahui informasi mengenai jumlah gol, total percobaan serangan, serta hasil akhir pertandingan.

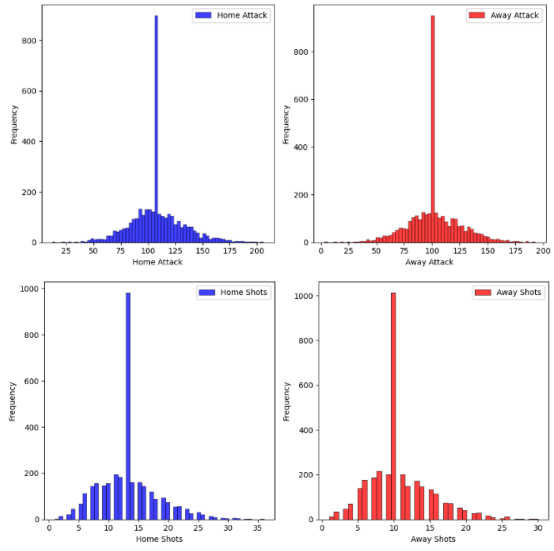
3.1. Praproses dan Visualisasi Data

Dataset *EPL* yang digunakan pada penelitian ini memiliki amatan sebesar 3352 dengan peubah sebanyak 18. Tipe-tipe data untuk peubah yakni berisi *object* dan *float64*. Tipe data *object* merepresentasikan data kategorikal dan tipe data *float64* merepresentasikan data kontinu. Terdapat beberapa data yang hilang pada fitur *home_attack*, *away_attack*, *home_shots*, dan *away_shots* secara berurutan sebesar 802, 801, 803, dan 803. Penanganan data hilang dilakukan dengan metode imputasi, Imputasi dengan median untuk kolom numerik, yang membantu menangani outlier dengan lebih baik dibandingkan menggunakan rata-rata sehingga tidak terdapat lagi amatan hilang.



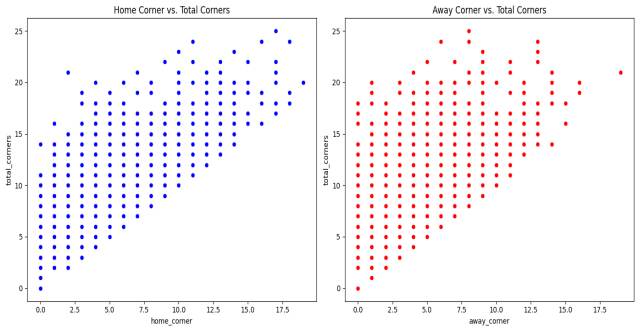
Gambar 2. Distribusi Gol Tim Tuan Rumah dan Tandang

Gambar 2 merupakan hasil visualisasi distribusi gol yang dicetak tim tuan rumah dan tandang. Berdasarkan gambar tersebut terlihat bahwa rata-rata tim di Liga Primer Inggris terutama tim tuan rumah cenderung lebih banyak mencetak gol apabila dibandingkan dengan tim tandang. Kedua distribusi menunjukkan kecenderungan untuk mencetak gol lebih rendah dari keseluruhan tim yang bermain di Liga Primer Inggris, sebagian besar pertandingan yang dilakukan memiliki jumlah rerata 0 hingga 2 gol per laga.



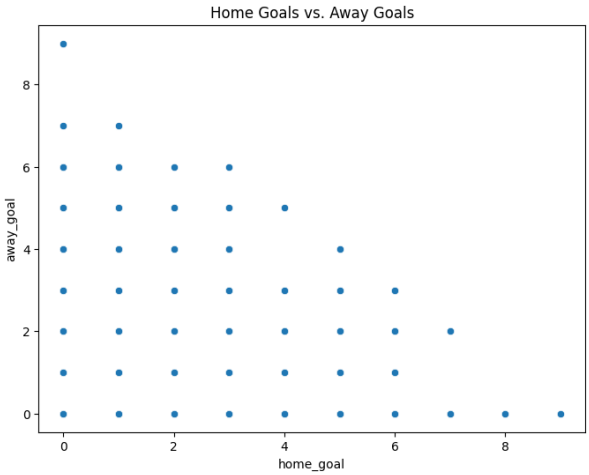
Gambar 3. Distribusi Tembakan dan Serangan Tim Tuan Rumah dan Tandang

Gambar 3 berisi visualisasi mengenai distribusi jumlah tembakan dan serangan ke arah gawang lawan baik dari tim tuan rumah dan tim tandang, terdapat kemiripan antara tim yang bermain sebagai tuan rumah maupun tandang. Tim tuan rumah lebih mendominasi baik dalam jumlah tembakan serta serangan dalam rata-rataan per pertandingan secara berurutan 15 tembakan dan 120 lebih jumlah serangan sedangkan untuk tim tandang secara berurutan hanya memiliki rerata 10 tembakan dan 100 hingga 110 jumlah serangan. Hal ini menunjukkan cara bermain tim tuan rumah yang lebih berani menyerang sedangkan tim tandang cenderung bermain lebih aman dengan tidak banyak melakukan serangan.



Gambar 4. Perbandingan Total tendangan Pojok pada seluruh pertandingan antara tim tuan rumah dan tandang

Gambar 4 menunjukkan visualisasi perbandingan jumlah tendangan pojok yang diambil tuan rumah dan tim tandang dari total tendangan pojok dalam satu pertandingan. Hubungan positif yang kuat terlihat antara jumlah tendangan sudut tim tuan rumah dan tim tandang dengan total jumlah tendangan sudut dari satu pertandingan secara keseluruhan. Sebagian besar tim tuan rumah yang bermain di Liga Primer Inggris lebih mendominasi jumlah tendangan sudut yang diambil apabila dibandingkan dengan jumlah tendangan sudut yang dimiliki tim tandang, hal ini juga didukung dengan visualisasi jumlah serangan dan tembakan yang dimiliki tuan rumah pada visualisasi Gambar 3 sebelumnya.



Gambar 5. Perbandingan Jumlah Gol yang Dicetak Tim Tuan Rumah dan Tandang

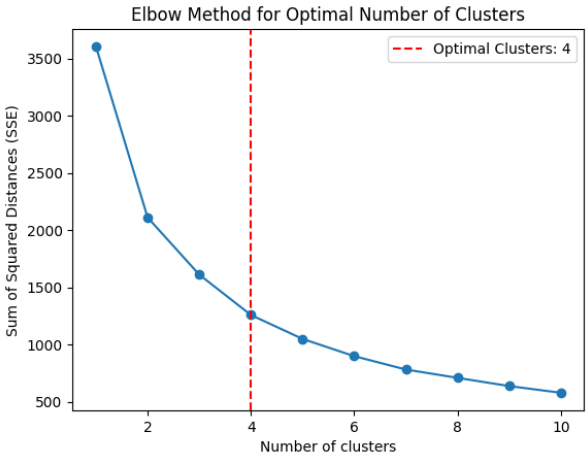
Gambar 5 merupakan visualisasi perbandingan jumlah gol yang dicetak tim tuan rumah dan tim tandang dari keseluruhan laga yang mereka mainkan. Terlihat bahwa dari gambar 6 tersebut skor rendah dengan skala 0-3 lebih banyak terjadi apabila dibandingkan dengan skor imbang dan gol dalam jumlah tinggi (lebih dari 3). Hanya terdapat beberapa tim di Liga Primer Inggris yang mampu mencetak gol dengan rerata diatas 6 sampai dengan 8 keatas.

3.2. *Feature Engineering*

Setelah dilakukan visualisasi terhadap beberapa peubah dalam dataset, kemudian dilakukan proses *Feature Engineering*. Proses ini bertujuan untuk menambah peubah baru berdasarkan peubah peubah yang telah ada pada data sehingga menciptakan dataset dengan penambahan peubah baru (Nargesian *et al.* 2017). Peubah baru yang akan dihasilkan antara lain rata-rata membobol dan rata-rata kebobolan, adanya peubah baru ini bertujuan untuk melihat seberapa kuat tim tuan rumah dan tandang dalam mencetak gol dan seberapa kuat pertahanannya berdasarkan 10 pertandingan terakhir masing-masing tim. Tim yang memiliki rata-rata membobol gawang lawan tinggi menunjukkan seberapa produktif tim dalam mencetak gol sedangkan rata-rata kebobolan yang rendah dapat menunjukkan seberapa baik tim dalam fase bertahan. Peubah baru yang dihasilkan menggunakan beberapa peubah seperti *home_attack* dan *away_attack* sehingga tercipta peubah seperti *rolling_avg_goals_home*, *rolling_avg_goals_away* secara berurutan menggambarkan rata-rata gol yang dicetak tim tuan rumah dan tandang dalam 10 pertandingan terakhirnya. Peubah baru *home_defense*, dan *away_defense* disisi lain menggambarkan kekuatan pertahanan tim tuan rumah dan tandang berdasarkan jumlah gol yang masuk ke gawang mereka. Peubah baru lainnya adalah *home_win_percent*, *away_win_percent* berisi informasi presentase jumlah kemanangan yang diraih tim tuan rumah dan tandang berdasarkan perhitungan dari peubah baru sebelumnya.

3.3. Optimalisasi k terbaik dengan Metode *Elbow*

Proses persiapan dataset yang telah dilakukan pembersihan, visualisasi, serta penambahan fitur baru menggunakan *feature engineering* kemudian dilanjutkan prosesnya untuk mencari k optimal dengan menerapkan metode *Elbow*. Penerapan metode *Elbow* pada dataset yang akan digunakan bertujuan untuk menemukan k optimal dari klasterisasi yang dihasilkan menggunakan metode *K-Means*.

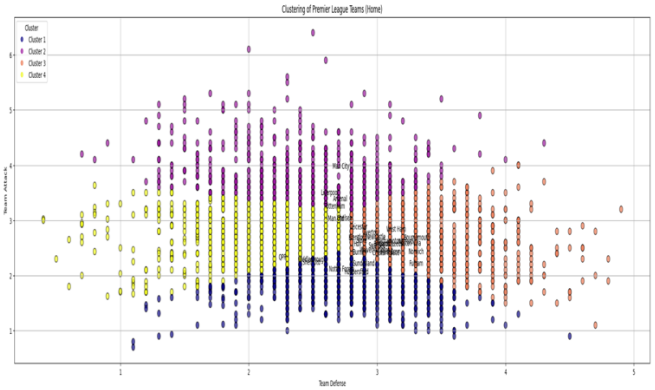


Gambar 6. Hasil Pengoptimalan k terbaik Menggunakan Metode Elbow

Gambar 6 menunjukkan hasil pencarian k optimal berdasarkan jumlah amatan dan peubah yang digunakan dalam dataset Liga Primer Inggris. Klaster k optimal yang terbentuk dengan analisis metode *Elbow* adalah sebanyak 4 klaster dan akan diterapkan pada pemodelan selanjutnya menggunakan metode *K-Means*.

3.4. Hasil Pemodelan dan Evaluasi

Proses ini merupakan visualisasi serta intepretasi hasil pengklasterisasian dataset Liga Primer Inggris dengan 4 klaster yang terbentuk, 4 klaster tersebut didapat dari pengaplikasian metode *Elbow* sebelumnya dan akan menjadi ketentuan klaster yang akan digunakan pada pemodelan menggunakan *K-Means*.



Gambar 7. Visualisasi Klasterisasi menggunakan Metode K-Means

Gambar 7 merupakan visualisasi hasil klasterisasi berdasarkan tim yang memiliki karakteristik bertahan dan menyerang. Distribusi tim menunjukkan tim tersebar dalam berbagai level kombinasi antara kekuatan serangan dan pertahanan, tim dengan nilai *defense* yang tinggi cenderung memiliki nilai serangan yang bervariasi, hal ini menunjukkan bahwa beberapa tim Liga Primer Inggris memiliki pertahanan yang baik namun tidak selalu memiliki daya serang yang kuat. Penjelasan mengenai hasil klasterisasi dapat dijelaskan sebagai berikut :

1) Tim yang tersebar dalam klaster berwarna biru merupakan tim-tim di Liga Primer Inggris yang cenderung memiliki pertahanan yang solid namun daya serang yang mereka miliki kurang efektif. Tim-tim ini cenderung bermain bertahan dan hanya mengandalkan serangan balik untuk mencetak gol. Beberapa tim yang termasuk ke dalam klaster ini antara lain Sheffield United, Nottingham Forest, dan Middlesbrough
2) Tim yang tersebar dalam klaster berwarna ungu merupakan tim-tim yang memiliki daya serang kuat namun pertahanan mereka cenderung sedikit lemah. Tim-tim ini cenderung mendominasi jalannya pertandingan dengan kemampuan daya serang mereka serta mampu mencetak banyak gol. Tim-tim seperti Macnhester City, Liverpool, dan Arsenal terlihat sebagai tim menghuni klaster ini.
3) Tim yang tersebar dalam klaster berwarna kuning dan hijau merupakan tim-tim yang cenderung stabil baik dalam segi menyerang dan bertahan selama pertandingan berlangsung. Tim-tim ini bermain dengan karakteristik efektif dalam menciptakan serangan serta kuat dalam bertahan. Machester United, dan Chelsea menjadi tim dengan nilai serangan dan bertahan yang cukup seimbang sedangkan tim lain seperti West Ham, Aston Villa, dan Fulham terlihat memiliki nilai bertahan yang cukup tinggi apabila dibandingkan dengan nilai serangannya.
4) Tim-tim di Liga Primer Inggris apabila dilihat dari visualisasi analisis gerombol dapat digambarkan sebagai tim yang lebih banyak melakukan serangan secara keseluruhan sehingga pertandingan dengan jual beli serangan lebih banyak terjadi dalam hampir sebagian besar pertandingan.

Pemodelan dengan menggunakan metode *K-Means* kemudian dilakukan uji kebaikan modelnya dengan menggunakan *Davies-Bouldin*, *Calinski-Harabasz* dan *Silhouette Score*. Hasil uji kebaikan model dengan menggunakan beberapa metode uji kebaikal model dapat dilihat dari tabel 1 berikut :

Table 3. Hasil Uji Kebaikan Model K-Means

Davies-Bouldin	*Calinski-Harabasz*	*Silhouette Score*
0.968524	2080.860212	0.325812

Berdasarkan hasil di atas, skor Silhouette sebesar 0,325812 dapat disebabkan oleh jumlah fitur data yang relatif sedikit sehingga tidak dilakukan reduksi dimensi, serta adanya overlap antar-kluster sebagaimana terlihat pada visualisasi pada Gambar 7. Untuk meningkatkan kualitas klasterisasi, diperlukan metode validasi tambahan seperti *Davies-Bouldin Index* atau *Calinski-Harabasz Index* untuk mengevaluasi hasil klasterisasi lebih lanjut. Selain itu, penggunaan algoritme klasterisasi lain yang lebih sesuai dengan struktur data, seperti *DBSCAN* atau *agglomerative clustering*, juga dapat dipertimbangkan sebagai langkah perbaikan.

Perbedaan hasil antara *Silhouette Score* sebesar 0,325812 dan *Davies-Bouldin Index (DBI)* sebesar 0,96 dapat dijelaskan berdasarkan karakteristik matematis dari kedua metrik tersebut. *Silhouette Score* mengevaluasi kualitas klasterisasi dengan membandingkan kedekatan data terhadap klusternya sendiri (kompaksi) dan jarak terhadap kluster terdekat (separasi) untuk setiap titik. Skor rendah mengindikasikan adanya overlap antar-kluster, ketidakseimbangan jumlah data dalam kluster, atau distribusi data yang tidak sesuai dengan asumsi berbentuk *spheris* (Rousseeuw, 1987).

Sebaliknya, DBI menghitung rasio rata-rata jarak intra-kluster terhadap separasi antar-kluster secara global, dengan nilai yang lebih kecil menunjukkan hasil klasterisasi yang lebih baik (Davies & Bouldin, 1979). Nilai DBI yang rendah, meskipun diiringi *Silhouette Score* yang rendah, dapat mengindikasikan bahwa kluster memiliki separasi global yang cukup baik, tetapi distribusi lokal data dalam kluster kurang optimal atau terdapat overlap antar-kluster.

Perbedaan ini mencerminkan sensitivitas masing-masing metrik terhadap struktur data, di mana *Silhouette Score* lebih sensitif terhadap jarak antar-titik individual dan distribusi lokal, sedangkan DBI lebih fokus pada jarak antar pusat

kluster. Penggunaa metrik tambahan seperti *Calinski-Harabasz* dapat membantu interpretasi yang lebih komprehensif.

Penggunaan reduksi dimensi tidak diperlukan dalam visualisasi kluster karena dataset hanya terdiri dari 8 fitur sehingga masih cukup sederhana untuk diolah tanpa mengorbankan *interpretabilitas*.

4. Kesimpulan

Kesimpulan yang didapatkan dari hasil visualisasi dan klasterisasi yang telah dilakukan terhadap data pertandingan Liga Primer Inggris menggunakan metode *K-Means* adalah (1) Visualisasi dan analisis pada tahapan praproses dan visualisasi data menunjukkan bahwa tim-tim di Liga Primer Inggris cukup kompetitif baik tim tuan rumah maupun tandang dengan jumlah tembakan yang cukup besar serta jumlah serangan yang cukup tinggi frekuensinya secara berurutan rerata 10 hingga 12 tembakan dan 120 percobaan menyerang dari masing-masing tim. (2) Hasil pemodelan menggunakan metode K-Means menciptakan 4 klasterisasi berdasarkan karakteristik permainan dari tiap tim-tim Liga Primer Inggris yang bermain. Terlihat bahwa tim-tim tersebut hampir keseluruhan memiliki profil menyerang dan beberapa dari tim tersebut terlihat cukup stabil dalam segi bertahan. (3) Hasil klasterisasi berdasarkan metode *K-Means* dengan uji *Davies-Bouldin*, *Calinski-Harabasz* dan *Silhouette Score* menunjukkan skor yang cukup baik namun masih dapat dioptimalkan hasilnya apabila dilakukan pemodelan menggunakan metode analisis gerombol lainnya.

Secara keseluruhan, penelitian ini menghasilkan klasterisasi tiap tim yang bermain di Liga Inggris berdasarkan pola permainan, serangan, serta pertahanan. Penelitian ini dapat diimplementasikan bagi manajer-manajer serta direktur olahraga dari tim-tim tertentu untuk membaca pola permainan serta bagaimana penyusunan strategi yang tepat sehingga mendapatkan hasil terbaik dari setiap pertandingan. Hal yang perlu diperhatikan yaitu terkait variabel-variabel yang digunakan masih perlu ditambahkan seperti jumlah umpan, progresi umpan kedepan, eksptasi gol (*XG*) dapat ditambahkan di penelitian selanjutnya dengan tujuan menambah wawasan dari klasterisasi tim yang dihasilkan.

5. Saran

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa dataset yang digunakan telah mampu untuk menggerombolkan karakteristik gaya permainan tiap tim yang ada di EPL. Penelitian selanjutnya disarankan dapat menggunakan dataset dengan fitur serta amatan yang lebih banyak sehingga dapat dilakukan rekayasa fitur untuk menambah variasi amatan pada data serta menerapkan penggunaan metode analisis gerombol lainnya seperti *Ward* ataupun *K-Medoid*.

Referensi

Al-Asadi MA, Tasdemir S. 2022. Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. *IEEE Access*. 10:22631–22645.doi:10.1109/ACCESS.2022.3154767.

Andreff W. 2011. Some comparative economics of the organization of sports: competition and regulation in north American vs. European professional team sports leagues. *The European Journal of Comparative Economics*. 8(1):3–27.

Baboota R, Kaur H. 2019. Predictive analysis and modelling football results using machine learning approach for English Premier League. *Int J Forecast*. 35(2):741–755.doi:10.1016/j.ijforecast.2018.01.003.

Bond AJ, Widdop P, Cockayne D, Parnell D. 2021. Prosumption, Networks and Value during a Global Pandemic: Lockdown Leisure and COVID-19. *Leis Sci*. 43(1–2):70–77.doi:10.1080/01490400.2020.1773985.

Firman Ashari I, Dwi Nugroho E, Baraku R, Yanda IN, Liwardana R. 2023. Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. Volume ke-7.

Foo WL, Tester E, Close GL, Cronin CJ, Morton JP. 2024. Professional Male Soccer Players' Perspectives of the Nutrition Culture Within an English Premier League Football Club: A Qualitative Exploration Using Bourdieu's Concepts of Habitus, Capital and Field. *Sports Medicine*..doi:10.1007/s40279-024-02134-w.

Herold M, Goes F, Nopp S, Bauer P, Thompson C, Meyer T. 2019. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *Int J Sports Sci Coach*. 14(6):798–817.doi:10.1177/1747954119879350.

Hewitt JH, Karakuş O. 2023. A machine learning approach for player and position adjusted expected goals in football (soccer). *Franklin Open*. 4:100034.doi:10.1016/j.fraope.2023.100034.

Kumar S, Solanki VK, Choudhary SK, Selamat A, Crespo RG. 2020. Comparative study on ant colony optimization (ACO) and k-means clustering approaches for jobs scheduling and energy optimization model in internet of things (IoT). *International Journal of Interactive Multimedia and Artificial Intelligence*. 6(1):107–116.doi:10.9781/ijimai.2020.01.003.

Millati K, Suhaeni C, Susetyo B. 2021. Penggerombolan Daerah 3T di Indonesia Berdasarkan Rasio Tenaga Kesehatan dengan Metode Penggerombolan Berhierarki dan Cluster Ensemble. *Xplore: Journal of Statistics*. 10(2):197–213.doi:10.29244/xplore.v10i2.744.

Murtagh F, Contreras P. 2012. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2(1):86–97.doi:10.1002/widm.53.

Nargesian F, Samulowitz H, Khurana U, Khalil EB, Turaga D. 2017. Learning feature engineering for classification. Di dalam: *IJCAI International Joint Conference on Artificial Intelligence*. Vol. 0. International Joint Conferences on Artificial Intelligence. hlm. 2529–2535.

Pratama Simanjuntak K, Khaira U. 2021. MALCOM: Indonesian Journal of Machine Learning and Computer Science Hotspot Clustering in Jambi Province Using Agglomerative Hierarchical Clustering Algorithm Pengelompokkan Titik Api di Provinsi Jambi dengan Algoritma Agglomerative Hierarchical Clustering. 1:7–16.

Rommers N, Rössler R, Verhagen E, Vandecasteele F, Verstockt S, Vaeyens R, Lenoir M, D'Hondt E, Witvrouw E. 2020. A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players. *Med Sci Sports Exerc*. 52(8):1745–1751.doi:10.1249/MSS.0000000000002305.

Shi C, Wei B, Wei S, Wang W, Liu H, Liu J. 2021. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP J Wirel Commun Netw*. 2021(1).doi:10.1186/s13638-021-01910-w.

Vergani AA, Binaghi E. 2018. A soft davies-bouldin separation measure. Di dalam: *IEEE International Conference on Fuzzy Systems*. Vol. 2018-July. Institute of Electrical and Electronics Engineers Inc.

Wu R. 2024. Behavioral analysis of electricity consumption characteristics for customer groups using the k-means algorithm. *Systems and Soft Computing*. 6.doi:10.1016/j.sasc.2024.200143.

Xu D, Tian Y. 2015. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*. 2(2):165–193.doi:10.1007/s40745-015-0040-1.