Journal of Artificial Intelligence and Software Engineering
Vol. 6, No. 2, Juni 2026, pp. 233~250
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v6i2.9781

Analysis of Chess Opening Patterns Using Data Mining Methods to Determine Players' Strategic Tendencies

Femasta Sembiring[1], Rahmadani[2], Muhammad Fauzi[3]
[1,3] Teknik Informatika, STMIK Kaputama, Binjai, 20714, Indonesia
[2] Universitas Pembangunan Panca Budi, Medan, 20122, Indonesia

Informasi Artikel

Diterima : 01 Mei 2026
Revisi : 05 Juni 2026
Publikasi : 30 Juni 2026

Kata Kunci:

Data Mining
K-Means Clustering
Chess Opening
Lichess
Davies-Bouldin Index

ABSTRAK

Permainan catur online menghasilkan data pertandingan berskala besar yang dapat dimanfaatkan untuk menganalisis pola pembukaan secara objektif. Penelitian ini bertujuan mengelompokkan opening catur berdasarkan karakteristik statistik permainan menggunakan algoritma K-Means Clustering. Dataset berasal dari Lichess dalam format PGN pada kategori Rapid dengan pemain berating di atas 1200. Data diseleksi, dipreprocessing, ditransformasikan berdasarkan kode ECO, kemudian dipilih 300 kode ECO dengan frekuensi tertinggi. Variabel yang digunakan adalah Win Rate, Draw Rate, dan Average Moves yang distandarisasi menggunakan Z-Score. Pengujian dilakukan pada skenario 3, 4, dan 5 cluster serta dievaluasi menggunakan Davies-Bouldin Index. Nilai DBI masing-masing skenario adalah 0,9538, 0,8715, dan 0,7312. Hasil terbaik diperoleh pada 5 cluster karena memiliki nilai DBI terkecil. Hasil clustering menunjukkan bahwa opening catur dapat dikelompokkan menjadi kecenderungan agresif, solid, seimbang, dan karakteristik khusus.

ABSTRACT

Online chess games generate large-scale match data that can be used to objectively analyze opening patterns. This study aims to cluster chess openings based on game statistical characteristics using the K-Means Clustering algorithm. The dataset was obtained from Lichess in PGN format, focusing on Rapid games played by players with ratings above 1200. The data were selected, preprocessed, transformed based on ECO codes, and reduced to the 300 most frequent ECO codes. The variables used were Win Rate, Draw Rate, and Average Moves, which were standardized using Z-Score. Testing was conducted using 3-, 4-, and 5-cluster scenarios and evaluated using the Davies-Bouldin Index. The DBI values for each scenario were 0.9538, 0.8715, and 0.7312. The best result was obtained by the 5-cluster scenario because it produced the smallest DBI value. The clustering results show that chess openings can be grouped into aggressive, solid, balanced, and special-characteristic tendencies.

***Penulis Koresponden**

Email: rahm4dani@gmail.com

F. Sembiring, R. Rahmadani, dan M. Fauzi, "*Analysis of Chess Opening Patterns Using Data Mining Methods to Determine Players' Strategic Tendencies*," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 6, no. 1, pp. 233-250, Juni 2026. doi:10.30811/ jaise.v6i2.9781

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong meningkatnya aktivitas permainan digital, termasuk permainan catur yang saat ini banyak dimainkan melalui platform daring. Salah satu platform catur online yang banyak digunakan adalah Lichess, karena menyediakan data pertandingan dalam jumlah besar dan dapat dimanfaatkan untuk kebutuhan analisis komputasional. Data permainan catur online umumnya memuat informasi seperti hasil pertandingan, rating pemain, langkah permainan, waktu permainan, serta notasi pertandingan dalam format Portable

Game Notation (PGN). Ketersediaan data yang besar dan terstruktur tersebut menjadikan catur sebagai objek yang relevan untuk dianalisis menggunakan pendekatan data mining dan machine learning ([1], [2]).

Catur merupakan permainan strategi yang memiliki kompleksitas tinggi karena setiap langkah dapat memengaruhi arah permainan berikutnya. Dalam penelitian berbasis data, permainan catur sering digunakan untuk mempelajari perilaku pemain, performa permainan, pola pengambilan keputusan, serta karakteristik strategi yang muncul dari data pertandingan nyata. [2] menunjukkan bahwa data permainan catur dalam skala besar dapat digunakan untuk mengukur performa pemain secara kuantitatif. Selain itu, [3] menjelaskan bahwa data catur dapat digunakan untuk menganalisis hubungan strategi dan pola permainan secara komputasional. Hal ini menunjukkan bahwa data permainan catur tidak hanya berfungsi sebagai arsip pertandingan, tetapi juga sebagai sumber informasi untuk menemukan pola strategi pemain.

Salah satu bagian penting dalam permainan catur adalah opening atau pembukaan. Opening merupakan fase awal permainan yang berperan dalam membangun posisi, mengontrol pusat papan, mengembangkan buah catur, dan menentukan arah strategi pada fase berikutnya. Setiap opening memiliki karakteristik yang berbeda, baik dari sisi tingkat kemenangan, peluang remis, maupun rata-rata jumlah langkah permainan. [4] menjelaskan bahwa opening catur dapat dianalisis berdasarkan kompleksitas dan kemiripan karakteristiknya menggunakan data komunitas catur online. Dengan demikian, opening dapat dipandang sebagai objek analisis yang dapat diukur secara statistik berdasarkan data pertandingan nyata.

Pada praktiknya, pemain sering memilih opening berdasarkan teori permainan, pengalaman pribadi, atau referensi pemain profesional. Namun, banyak opening yang berbeda secara teori dapat menunjukkan karakteristik statistik yang serupa ketika digunakan dalam pertandingan online. Kondisi tersebut menunjukkan perlunya pendekatan berbasis data untuk mengelompokkan opening berdasarkan kemiripan karakteristik permainan. Penelitian sebelumnya telah membahas analisis catur menggunakan pendekatan komputasional, seperti pemodelan perilaku individu pemain [1], analisis performa manusia dalam catur [2], analisis kemiripan opening[4], [5], serta clustering permainan catur berbasis Forsyth-Edwards Notation [6]. Selain itu, [7] juga melakukan studi clustering pada variasi French Defense untuk memetakan karakteristik jalur permainan dalam opening tertentu.

Meskipun berbagai penelitian tersebut telah menunjukkan bahwa data catur dapat dianalisis menggunakan pendekatan machine learning, statistik, dan clustering, penelitian yang secara khusus mengelompokkan opening catur berdasarkan karakteristik statistik permainan masih relatif terbatas. Beberapa penelitian lebih banyak berfokus pada prediksi langkah, klasifikasi hasil pertandingan, analisis perilaku pemain, atau representasi posisi permainan. [8], misalnya menggunakan pendekatan Long Short Term Memory untuk prediksi langkah catur, sedangkan [9] membahas klasifikasi hasil catur blitz pada dataset berskala besar. Berbeda dari penelitian tersebut, penelitian ini berfokus pada pengelompokan opening catur berdasarkan variabel statistik permainan, yaitu Win Rate, Draw Rate, dan Average Moves.

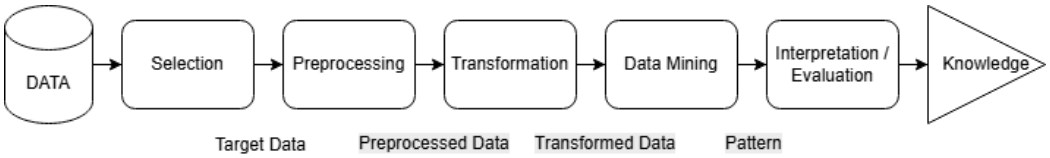
Berdasarkan permasalahan tersebut, penelitian ini menerapkan algoritma K-Means Clustering untuk mengelompokkan pola pembukaan catur berdasarkan kemiripan karakteristik statistik permainan. Data yang digunakan berasal dari dataset pertandingan catur online Lichess pada kategori Rapid dengan pemain yang memiliki rating di atas 1200. Data diolah berdasarkan kode Encyclopedia of Chess Openings (ECO), kemudian dipilih 300 kode ECO dengan frekuensi kemunculan tertinggi. Tahapan penelitian meliputi seleksi dataset, preprocessing, transformasi data, standarisasi menggunakan Z-Score, penerapan K-Means pada skenario 3 cluster, 4 cluster, dan 5 cluster, serta evaluasi menggunakan Davies-Bouldin Index. Hasil pengelompokan diharapkan dapat memberikan gambaran mengenai kecenderungan strategi pemain berdasarkan karakteristik opening yang digunakan.

2. METODE

Penelitian ini menggunakan pendekatan data mining dengan algoritma K-Means Clustering untuk mengelompokkan pola pembukaan catur berdasarkan karakteristik statistik permainan. Metode penelitian disusun secara sistematis mulai dari penyiapan dataset, seleksi data, preprocessing, transformasi data, standarisasi, penerapan algoritma K-Means, hingga evaluasi hasil clustering. Pendekatan data mining digunakan karena mampu mengekstraksi pola dari data berukuran besar dan menghasilkan informasi yang dapat digunakan untuk mendukung analisis karakteristik permainan[10], [8]

Dalam penelitian berbasis data mining, tahapan pengolahan data perlu dilakukan secara terstruktur agar data mentah dapat diubah menjadi informasi yang siap dianalisis. [11] menjelaskan bahwa proses Knowledge Discovery in Databases (KDD) mencakup Data Selection, Data Preprocessing, Data Transformation, Data Mining, dan Knowledge Interpretation/Evaluation. Sejalan dengan hal tersebut, penelitian ini menyusun tahapan metode mulai dari

seleksi dataset Lichess, preprocessing, pembentukan variabel statistik, standarisasi data, penerapan K-Means, hingga evaluasi hasil clustering.



Gambar 1. Proses KDD [11]

Gambar 1 menunjukkan tahapan KDD yang dimulai dari data mentah hingga menghasilkan pengetahuan yang dapat digunakan untuk menjawab tujuan penelitian. Tahap pertama adalah data, yaitu dataset pertandingan catur online Lichess dalam format Portable Game Notation (PGN). Dataset tersebut memuat informasi pertandingan, seperti kategori permainan, rating pemain, kode ECO, nama opening, hasil pertandingan, dan urutan langkah permainan.

Tahap selection dilakukan untuk memilih data pertandingan yang sesuai dengan batasan penelitian. Seleksi dilakukan terhadap pertandingan kategori Rapid, pemain dengan rating di atas 1200, serta data yang memiliki kode ECO, hasil pertandingan, dan jumlah langkah yang dapat diidentifikasi. Data yang memenuhi kriteria tersebut menjadi *target data* yang digunakan pada tahap berikutnya.

Tahap preprocessing dilakukan untuk membersihkan dan menyiapkan data hasil seleksi. Pada tahap ini dilakukan pemeriksaan kelengkapan data, penghapusan data yang tidak valid, ekstraksi atribut dari file PGN, serta perhitungan jumlah langkah permainan. Hasil tahap ini berupa *preprocessed data* yang telah bersih dan dapat diolah lebih lanjut.

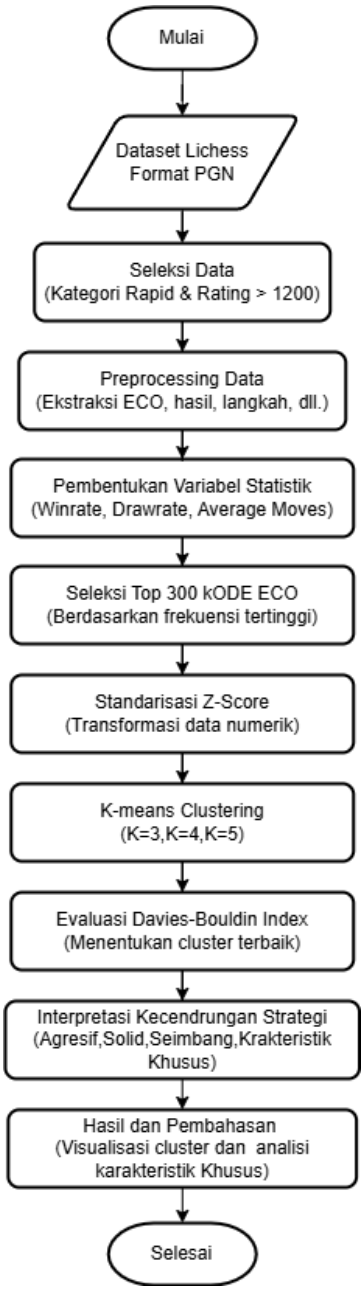
Tahap transformation dilakukan dengan mengelompokkan data pertandingan berdasarkan kode ECO dan mengubahnya menjadi data statistik. Setiap kode ECO direpresentasikan menggunakan tiga variabel, yaitu Win Rate, Draw Rate, dan Average Moves. Selanjutnya, dipilih 300 kode ECO dengan frekuensi kemunculan tertinggi dan dilakukan standarisasi menggunakan Z-Score. Hasil tahap ini berupa *transformed data* yang siap digunakan dalam proses clustering.

Tahap data mining dilakukan menggunakan algoritma K-Means Clustering. Pengujian dilakukan pada tiga skenario jumlah cluster, yaitu K=3, K=4, dan K=5. Pada tahap ini, algoritma mengelompokkan kode ECO berdasarkan kemiripan nilai Win Rate, Draw Rate, dan Average Moves sehingga menghasilkan pola pengelompokan opening catur.

Tahap interpretation/evaluation dilakukan dengan mengevaluasi hasil clustering menggunakan Davies-Bouldin Index. Skenario dengan nilai DBI paling kecil dipilih sebagai konfigurasi cluster terbaik. Selanjutnya, karakteristik setiap cluster dianalisis berdasarkan nilai centroid dan distribusi anggota cluster.

Tahap terakhir adalah knowledge, yaitu pengetahuan yang diperoleh dari hasil pengelompokan. Pengetahuan tersebut berupa kecenderungan karakteristik opening catur, seperti opening yang cenderung agresif, solid, seimbang, atau memiliki karakteristik khusus berdasarkan statistik permainan.

Alur penelitian pada penelitian ini ditunjukkan pada Gambar 2



Gambar 2. Alur penelitian pengelompokan opening catur.

Gambar 2 menunjukkan alur utama penelitian yang dimulai dari penggunaan dataset pertandingan catur Lichess dalam format PGN. Data kemudian diseleksi berdasarkan kategori Rapid, rating pemain di atas 1200, ketersediaan kode ECO, hasil pertandingan, dan jumlah langkah. Data yang lolos seleksi diproses melalui tahap preprocessing untuk mengekstraksi informasi pertandingan dan membentuk variabel statistik berupa Win Rate, Draw Rate, dan Average Moves. Selanjutnya, dipilih 300 kode ECO dengan frekuensi tertinggi, kemudian data distandarisasi menggunakan Z-Score. Proses clustering dilakukan menggunakan algoritma K-Means pada skenario K=3, K=4, dan K=5. Hasil setiap skenario dievaluasi menggunakan Davies-Bouldin Index untuk menentukan jumlah cluster terbaik, kemudian diinterpretasikan berdasarkan kecenderungan strategi opening catur.

2.1 Data Penelitian

Dataset yang digunakan dalam penelitian ini berasal dari data pertandingan catur online pada platform Lichess. Lichess dipilih karena menyediakan data pertandingan catur dalam jumlah besar dan dapat digunakan untuk penelitian berbasis komputasional. Data permainan catur online umumnya tersedia dalam format Portable Game

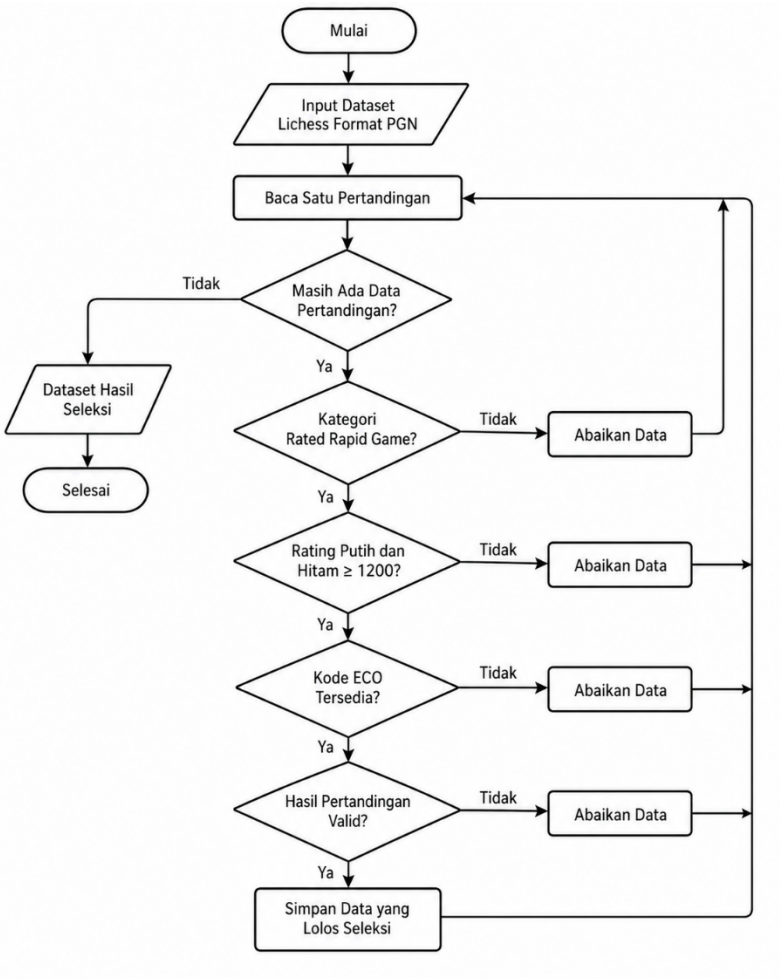
Notation (PGN), yaitu format teks yang menyimpan informasi pertandingan seperti identitas permainan, hasil pertandingan, rating pemain, kode opening, nama opening, serta urutan langkah permainan. Dataset berbasis PGN juga banyak digunakan dalam penelitian catur karena mampu merepresentasikan informasi permainan secara terstruktur ([1], [5], [12]).

Data penelitian difokuskan pada kategori permainan Rapid dengan pemain yang memiliki rating di atas 1200. Pemilihan kategori Rapid dilakukan karena permainan Rapid memiliki waktu berpikir yang relatif lebih stabil dibandingkan kategori Bullet dan Blitz, sehingga karakteristik permainan yang terbentuk dapat dianalisis secara lebih representatif. Batas rating di atas 1200 digunakan untuk menyaring permainan dari pemain yang sudah memiliki pemahaman dasar terhadap opening dan strategi permainan. Dengan demikian, data yang digunakan diharapkan dapat mencerminkan pola permainan yang lebih konsisten.

Atribut awal yang digunakan dari dataset PGN meliputi kode ECO, nama opening, hasil pertandingan, rating pemain, dan jumlah langkah permainan. Kode ECO digunakan sebagai identitas utama opening karena setiap kode merepresentasikan jenis pembukaan tertentu. Hasil pertandingan digunakan untuk menghitung Win Rate dan Draw Rate, sedangkan jumlah langkah permainan digunakan untuk menghitung Average Moves. Penggunaan kode ECO sebagai dasar analisis opening didukung oleh penelitian [4] yang menunjukkan bahwa opening catur dapat dianalisis berdasarkan kompleksitas dan kemiripan karakteristiknya menggunakan data komunitas catur online.

2.2 Seleksi Data

Proses seleksi dilakukan terhadap setiap pertandingan untuk memastikan bahwa data yang digunakan sesuai dengan batasan penelitian. Alur proses seleksi data ditunjukkan pada Gambar 3.



Gambar 3. Alur Seleksi Data Pertandingan Catur

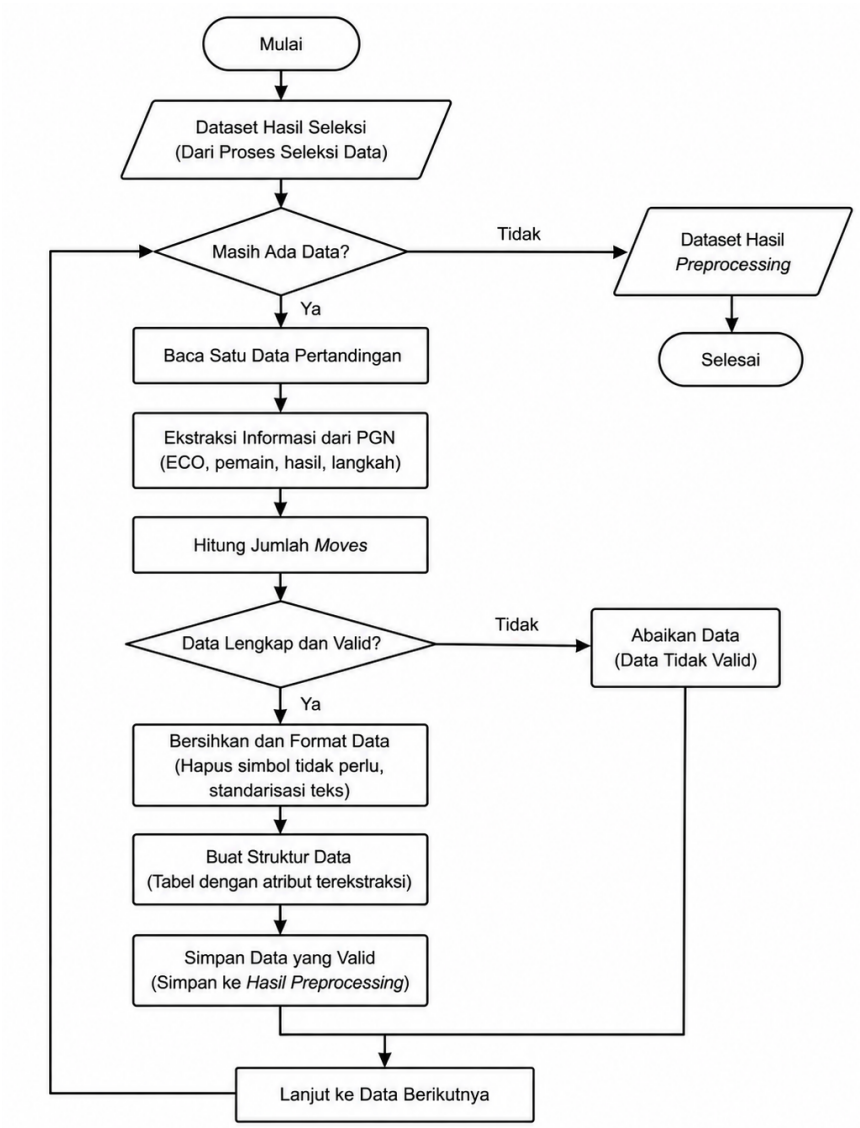
Gambar 3 menunjukkan proses seleksi setiap pertandingan pada dataset Lichess. Data dipilih berdasarkan kategori Rated Rapid Game, rating pemain putih dan hitam minimal 1200, ketersediaan kode ECO, serta hasil

pertandingan yang valid. Data yang tidak memenuhi salah satu kriteria diabaikan, sedangkan data yang memenuhi seluruh kriteria disimpan sebagai dataset hasil seleksi. Proses tersebut dilakukan secara berulang hingga seluruh pertandingan selesai diperiksa.

Seleksi data dilakukan untuk memastikan bahwa data yang masuk ke tahap analisis merupakan data yang relevan dan dapat diolah secara numerik. Data yang tidak memenuhi kriteria, seperti data tanpa kode ECO, data dengan hasil pertandingan tidak valid, atau data yang tidak memiliki jumlah langkah, tidak digunakan dalam proses berikutnya. Proses seleksi ini penting karena kualitas data awal sangat memengaruhi hasil clustering yang diperoleh. Dalam proses data mining, tahap seleksi data menjadi bagian penting untuk menentukan data yang benar-benar sesuai dengan tujuan analisis ([10], [13], [14]).

2.3 Preprocessing Data

Preprocessing dilakukan untuk membersihkan dan menyiapkan data sebelum masuk ke tahap transformasi. Pada tahap ini, data PGN dibaca dan diekstraksi untuk mengambil atribut yang dibutuhkan. Atribut yang dipertahankan adalah kode ECO, nama opening, hasil pertandingan, dan jumlah langkah permainan. Data yang tidak lengkap atau tidak sesuai dengan format yang dibutuhkan dihapus agar tidak mengganggu proses perhitungan statistik.



Gambar 4. Alur Preprocessing Data

Gambar 4 menunjukkan alur preprocessing data setelah tahap seleksi. Proses ini meliputi ekstraksi informasi dari file PGN, perhitungan jumlah langkah, pemeriksaan kelengkapan dan validitas data, pembersihan serta format data, hingga penyusunan data ke dalam tabel terstruktur. Hasil dari tahap ini adalah dataset hasil preprocessing yang siap digunakan pada proses clustering.

Setelah data bersih diperoleh, setiap pertandingan dikelompokkan berdasarkan kode ECO. Pengelompokan awal ini bertujuan untuk menghitung jumlah kemunculan setiap opening, jumlah kemenangan, jumlah remis, dan total langkah permainan. Dengan demikian, data yang awalnya berbentuk data pertandingan individual diubah menjadi data ringkasan statistik berdasarkan opening. Proses preprocessing dan transformasi seperti ini diperlukan agar data mentah dapat menjadi dataset numerik yang sesuai untuk algoritma clustering ([15], [16], [17]).

2.4 Pembentukan Variabel Dataset

Data hasil preprocessing kemudian ditransformasikan menjadi variabel statistik berdasarkan setiap kode ECO. Pada penelitian ini terdapat tiga variabel utama yang digunakan dalam proses clustering, yaitu Win Rate, Draw Rate, dan Average Moves. Ketiga variabel tersebut dipilih karena dapat merepresentasikan karakteristik permainan dari setiap opening.

Win Rate digunakan untuk menunjukkan persentase kemenangan pada setiap kode ECO. Draw Rate digunakan untuk menunjukkan persentase hasil remis pada setiap kode ECO. Average Moves digunakan untuk menunjukkan rata-rata jumlah langkah permainan pada setiap kode ECO. Dengan menggunakan ketiga variabel tersebut, setiap opening dapat direpresentasikan sebagai data numerik yang menggambarkan kecenderungan hasil dan durasi permainan.

Tabel 1. Variabel statistik yang digunakan dalam clustering.

No.	Variabel	Keterangan
1	Win Rate	Persentase kemenangan pada setiap kode ECO.
2	Draw Rate	Persentase hasil remis pada setiap kode ECO.
3	Average Moves	Rata-rata jumlah langkah permainan pada setiap kode ECO.

Secara umum, Win Rate dihitung dari jumlah kemenangan dibandingkan dengan total permainan pada suatu kode ECO. Draw Rate dihitung dari jumlah hasil remis dibandingkan dengan total permainan pada suatu kode ECO. Average Moves dihitung dari total jumlah langkah pada suatu kode ECO dibagi dengan jumlah permainan pada kode ECO tersebut. Hasil dari proses ini berupa dataset statistik opening yang siap digunakan pada tahap seleksi Top 300 ECO.

Pembentukan variabel numerik diperlukan agar objek penelitian dapat dianalisis menggunakan algoritma clustering. [18] menggunakan beberapa atribut numerik untuk membentuk kelompok data menggunakan K-Means dan mengevaluasi hasilnya menggunakan Davies-Bouldin Index. Dalam penelitian ini, setiap kode ECO direpresentasikan dalam bentuk tiga variabel numerik, yaitu Win Rate, Draw Rate, dan Average Moves, sehingga setiap opening dapat dibandingkan berdasarkan kemiripan karakteristik statistik permainan.

2.5 Seleksi Top 300 ECO

Setelah seluruh data pertandingan dikelompokkan berdasarkan kode ECO, dilakukan seleksi terhadap 300 kode ECO dengan frekuensi kemunculan tertinggi. Seleksi ini dilakukan agar data yang digunakan dalam proses clustering berasal dari opening yang memiliki jumlah permainan cukup besar dan representatif. Opening yang terlalu jarang muncul tidak digunakan karena dapat menghasilkan nilai statistik yang kurang stabil dan berpotensi menjadi data pencilan.

Pemilihan 300 kode ECO bertujuan untuk menjaga keseimbangan antara jumlah data yang cukup besar dan kualitas data yang dapat dianalisis. Dengan menggunakan 300 opening yang paling sering muncul, penelitian ini dapat memfokuskan analisis pada opening yang benar-benar banyak digunakan oleh pemain dalam pertandingan online. Data hasil seleksi ini kemudian menjadi dataset utama dalam proses standarisasi dan clustering.

2.6 Standarisasi Data

Sebelum diterapkan ke algoritma K-Means, data terlebih dahulu distandarisasi menggunakan Z-Score. Standarisasi dilakukan karena setiap variabel memiliki skala yang berbeda. Standarisasi data dilakukan agar setiap atribut memiliki pengaruh yang seimbang dalam perhitungan jarak. [19] menjelaskan bahwa normalisasi Z-Score membuat setiap fitur memiliki rata-rata nol dan standar deviasi satu, sehingga pengaruh antarvariabel menjadi lebih seimbang pada proses perhitungan jarak. Oleh karena itu, penelitian ini menggunakan Z-Score sebelum penerapan K-Means agar variabel Win Rate, Draw Rate, dan Average Moves berada pada skala yang sebanding. Win Rate dan Draw Rate berbentuk persentase, sedangkan Average Moves berbentuk rata-rata jumlah langkah. Jika data tidak distandarisasi, variabel dengan rentang nilai lebih besar dapat mendominasi perhitungan jarak pada algoritma K-Means. [15] menjelaskan bahwa feature scaling memiliki pengaruh penting terhadap hasil K-Means karena algoritma ini sangat bergantung pada perhitungan jarak antar data. [20] juga menjelaskan bahwa standarisasi dapat membantu menyeimbangkan atribut sebelum proses clustering.

Standarisasi dilakukan menggunakan persamaan Z-Score sebagai berikut:

$$Z = \frac{x-\mu}{\sigma} \quad (1)$$

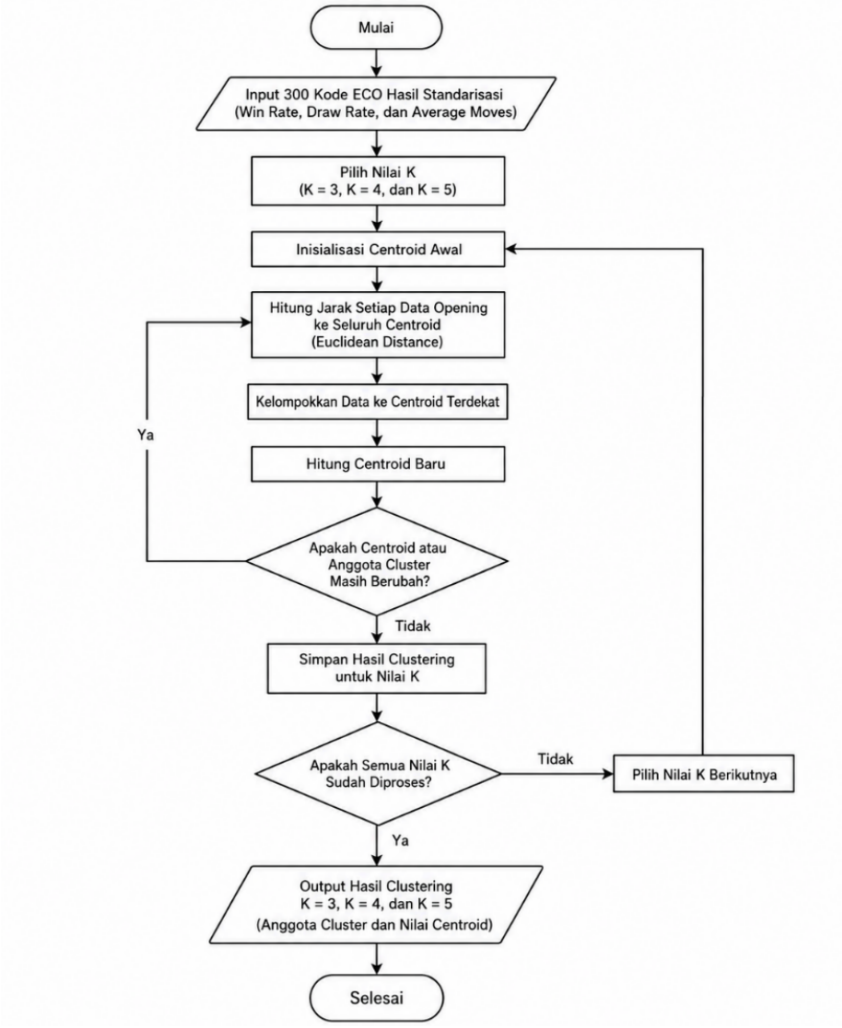
Pada persamaan (1), Z merupakan nilai hasil standarisasi, x merupakan nilai data asli, μ merupakan nilai rata-rata, dan σ merupakan standar deviasi. Setelah proses standarisasi, setiap variabel memiliki skala yang lebih seimbang sehingga proses perhitungan jarak pada K-Means dapat dilakukan secara lebih adil terhadap seluruh atribut.

2.7 Penerapan Algoritma K-Means Clustering

Algoritma K-Means digunakan untuk mengelompokkan 300 data opening catur berdasarkan kemiripan nilai Win Rate, Draw Rate, dan Average Moves yang telah distandarisasi. K-Means termasuk ke dalam metode unsupervised learning karena proses pengelompokan dilakukan tanpa label kelas awal. Setiap data akan ditempatkan ke dalam cluster berdasarkan jarak terdekat terhadap centroid. Proses ini dilakukan secara berulang sampai posisi centroid stabil atau tidak terjadi perubahan anggota cluster secara signifikan ([21], [22]).

Pada penelitian ini, pengujian dilakukan menggunakan tiga skenario jumlah cluster, yaitu 3 cluster, 4 cluster, dan 5 cluster. Pengujian dengan beberapa jumlah cluster dilakukan untuk mengetahui konfigurasi cluster yang menghasilkan kualitas pengelompokan terbaik. Setiap skenario dijalankan menggunakan data yang sama, yaitu 300 kode ECO hasil seleksi dan standarisasi. Hasil dari setiap skenario berupa anggota cluster, nilai centroid, grafik pengelompokan, serta nilai evaluasi Davies-Bouldin Index.

Tahapan penerapan K-Means pada penelitian ini ditunjukan pada Gambar 5.



Gambar 5. Penerapan Algoritma K-Means

Gambar 5 menunjukkan tahapan penerapan algoritma K-Means pada 300 kode ECO hasil standarisasi. Proses dimulai dengan menentukan nilai K, yaitu 3, 4, dan 5, kemudian menginisialisasi centroid awal, menghitung jarak setiap data ke centroid menggunakan Euclidean Distance, mengelompokkan data ke cluster terdekat, dan menghitung centroid baru. Tahapan tersebut diulang sampai hasil clustering konvergen. Hasil akhir dari proses ini adalah clustering untuk K=3, K=4, dan K=5 yang terdiri atas anggota cluster dan nilai centroid.

Perhitungan jarak antara data dan centroid dilakukan menggunakan Euclidean Distance sebagai berikut:

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

Pada persamaan (2), d(x,c) merupakan jarak antara data x dan centroid c, x_i merupakan nilai variabel pada data, c_i merupakan nilai variabel pada centroid, dan n merupakan jumlah variabel yang digunakan. Dalam penelitian ini, variabel yang digunakan adalah Win Rate, Draw Rate, dan Average Moves.

Dalam penelitian ini, setiap kode ECO direpresentasikan menggunakan tiga variabel hasil standarisasi, yaitu Win Rate, Draw Rate, dan Average Moves. Implementasi Persamaan (2) dilakukan dengan menghitung jarak antara setiap data opening dan centroid masing-masing cluster. Meskipun proses clustering dijalankan terhadap keseluruhan 300 kode ECO, sepuluh data ditampilkan sebagai sampel untuk memperjelas proses perhitungan.

Tabel 2. Sampel data hasil standarisasi dari 300 kode ECO

No.	Kode ECO	Win Rate	Draw Rate	Average Moves
1	C41	0,1166	-0,1166	0,0257
2	A00	0,3460	-0,3460	0,0063
3	D00	0,0723	-0,0723	0,0378
4	C00	0,2816	-0,2816	0,0802
5	B01	0,1367	-0,1367	-0,2733
6	C20	0,2454	-0,2454	-0,5620
7	B00	0,4627	-0,4627	-0,6062
8	B20	0,3098	-0,3098	-0,0530
9	A40	0,3299	-0,3299	-0,1474
10	C50	0,2132	-0,2132	-0,5826

Setiap kode ECO direpresentasikan sebagai vektor tiga dimensi. Untuk ECO C41, vektor hasil standarisasinya adalah:

$x_{C41} = (0{,}1166;\ -0{,}1166;\ 0{,}0257)$

Pada skenario K=3, centroid hasil clustering yang digunakan dalam contoh perhitungan adalah:
$C_1 = (-2{,}0354;\ 2{,}0354;\ 0{,}3894)$
$C_2 = (1{,}1274;\ -1{,}1274;\ -0{,}6288)$
$C_3 = (-0{,}0102;\ 0{,}0102;\ 0{,}1362)$

Karena penelitian menggunakan tiga variabel, maka n=3 dan persamaan diimplementasikan menjadi:

$$d(x_j, C_k) = \sqrt{(x_{j1} - C_{k1})^2 + (x_{j2} - C_{k2})^2 + (x_{j3} - C_{k3})^2}$$

Keterangan :
x_j: data ke-j
C_k: centroid cluster ke-k
x_{j1}, x_{j2}, x_{j3}: nilai tiga variabel pada data ke-j
C_{k1}, C_{k2}, C_{k3}: nilai tiga variabel pada centroid ke-k

Perhitungan Jarak ECO C41 terhadap Centroid 1 :

$$d(C41, C_1) = \sqrt{(0{,}1166 - (-2{,}0354))^2 + (-0{,}1166 - 2{,}0354)^2 + (0{,}0257 - 0{,}3894)^2}$$

$$d(\text{C41}, C_1) = \sqrt{(2{,}1520)^2 + (-2{,}1520)^2 + (-0{,}3637)^2}$$
$$d(\text{C41}, C_1) = \sqrt{4{,}6311 + 4{,}6311 + 0{,}1323}$$
$$d(\text{C41}, C_1) = \sqrt{9{,}3945}$$
$$d(\text{C41}, C_1) = 3{,}0650$$

Perhitungan Jarak ECO C41 terhadap Centroid 2 :

$$d(\text{C41}, C_2) = \sqrt{(0{,}1166 - 1{,}1274)^2 + \big(-0{,}1166 - (-1{,}1274)\big)^2 + \big(0{,}0257 - (-0{,}6288)\big)^2}$$
$$d(\text{C41}, C_2) = \sqrt{(-1{,}0108)^2 + (1{,}0108)^2 + (0{,}6545)^2}$$
$$d(\text{C41}, C_2) = \sqrt{1{,}0217 + 1{,}0217 + 0{,}4284}$$
$$d(\text{C41}, C_2) = \sqrt{2{,}4718}$$
$$d(\text{C41}, C_2) = 1{,}5722$$

Perhitungan Jarak ECO C41 terhadap Centroid 3 :

$$d(\text{C41}, C_3) = \sqrt{\big(0{,}1166 - (-0{,}0102)\big)^2 + (-0{,}1166 - 0{,}0102)^2 + (0{,}0257 - 0{,}1362)^2}$$
$$d(\text{C41}, C_3) = \sqrt{(0{,}1268)^2 + (-0{,}1268)^2 + (-0{,}1105)^2}$$
$$d(\text{C41}, C_3) = \sqrt{0{,}0161 + 0{,}0161 + 0{,}0122}$$
$$d(\text{C41}, C_3) = \sqrt{0{,}0444}$$
$$d(\text{C41}, C_3) = 0{,}2106$$

Hasil perhitungan ketiga jarak diringkas pada tabel 3.

Tabel 3. Hasil perhitungan jarak ECO C41 terhadap centroid

Data	Jarak ke C_1	Jarak ke C_2	Jarak ke C_3	Cluster
C41	3,0650	1,5722	**0,2106**	3

Nilai jarak terkecil adalah jarak ECO C41 terhadap centroid C_3, yaitu 0,2106. Oleh karena itu, ECO C41 ditempatkan sebagai anggota Cluster 3. Prosedur yang sama diterapkan terhadap seluruh 300 kode ECO, yaitu menghitung jarak setiap data terhadap seluruh centroid dan menempatkan data ke cluster dengan jarak paling kecil.

2.8 Evaluasi Clustering

Evaluasi clustering dilakukan untuk menilai kualitas hasil pengelompokan yang dihasilkan oleh algoritma K-Means. Metode evaluasi yang digunakan dalam penelitian ini adalah Davies-Bouldin Index (DBI). DBI digunakan karena dapat mengukur kualitas cluster berdasarkan tingkat kekompakan data di dalam cluster dan tingkat pemisahan antar cluster. Nilai DBI yang lebih kecil menunjukkan kualitas clustering yang lebih baik karena cluster yang terbentuk memiliki jarak antar cluster yang lebih jelas dan anggota cluster yang lebih kompak ([23], [24]).

Pada penelitian ini, DBI dihitung untuk tiga skenario jumlah cluster, yaitu 3 cluster, 4 cluster, dan 5 cluster. Nilai DBI dari setiap skenario kemudian dibandingkan untuk menentukan jumlah cluster terbaik. Skenario dengan nilai DBI terkecil dipilih sebagai hasil pengelompokan paling optimal. Selain nilai DBI, hasil clustering juga dianalisis berdasarkan distribusi jumlah anggota cluster, posisi centroid, dan grafik visualisasi hasil pengelompokan.

Selain DBI, beberapa penelitian clustering juga menggunakan metrik lain seperti Silhouette Score untuk memperkuat evaluasi kualitas cluster. [25] membandingkan K-Means dan DBSCAN menggunakan Silhouette Score dan Davies-Bouldin Index, kemudian menunjukkan bahwa K-Means memperoleh performa yang lebih baik pada kasus segmentasi pelanggan Transjakarta. Hal ini menunjukkan bahwa evaluasi internal cluster dapat digunakan untuk membandingkan kualitas pengelompokan dan mendukung pemilihan metode yang sesuai dengan karakteristik data.

2.9 Interpretasi Hasil Clustering

Setelah proses evaluasi dilakukan, hasil cluster diinterpretasikan untuk menentukan kecenderungan strategi opening catur. Interpretasi dilakukan berdasarkan nilai centroid pada setiap cluster. Cluster dengan nilai Win Rate tinggi dan Draw Rate rendah diinterpretasikan sebagai kelompok opening yang cenderung agresif atau taktikal. Cluster dengan nilai Draw Rate tinggi dan Win Rate rendah diinterpretasikan sebagai kelompok opening yang cenderung solid atau posisional. Cluster dengan nilai centroid mendekati rata-rata diinterpretasikan sebagai kelompok opening yang memiliki karakteristik seimbang. Sementara itu, cluster dengan jumlah anggota sangat kecil dapat dianalisis sebagai kelompok opening dengan karakteristik khusus atau data pencilan.

Melalui tahapan tersebut, penelitian ini tidak hanya menghasilkan kelompok opening berdasarkan perhitungan algoritma, tetapi juga memberikan penjelasan mengenai karakteristik strategi dari setiap kelompok. Dengan demikian, hasil clustering dapat digunakan untuk memahami kemiripan karakteristik opening catur berdasarkan data pertandingan nyata.

3. HASIL DAN PEMBAHASAN

Bagian ini menjelaskan hasil pengolahan data, hasil pengelompokan menggunakan algoritma K-Means Clustering, visualisasi hasil cluster, evaluasi menggunakan Davies-Bouldin Index (DBI), serta interpretasi kecenderungan strategi opening catur. Data yang digunakan pada tahap hasil merupakan data yang telah melalui proses seleksi, preprocessing, transformasi, dan standarisasi. Dataset akhir terdiri dari 300 kode ECO yang direpresentasikan menggunakan tiga variabel statistik, yaitu Win Rate, Draw Rate, dan Average Moves.

3.1 Hasil Transformasi Data

Setelah proses preprocessing dilakukan terhadap dataset Lichess dalam format PGN, data pertandingan catur diubah menjadi data statistik berdasarkan kode ECO. Setiap kode ECO merepresentasikan satu jenis opening catur yang memiliki karakteristik permainan tertentu. Data yang awalnya berbentuk data pertandingan individual kemudian dikelompokkan berdasarkan kode ECO untuk memperoleh nilai Win Rate, Draw Rate, dan Average Moves.

Win Rate digunakan untuk menunjukkan tingkat kemenangan pada setiap opening, Draw Rate digunakan untuk menunjukkan tingkat hasil remis, sedangkan Average Moves digunakan untuk menunjukkan rata-rata jumlah langkah permainan. Ketiga variabel tersebut dipilih karena dapat menggambarkan karakteristik hasil dan durasi permainan dari masing-masing opening. Setelah seluruh variabel terbentuk, dilakukan seleksi terhadap 300 kode ECO dengan frekuensi kemunculan tertinggi agar data yang digunakan lebih representatif dan tidak didominasi oleh opening yang sangat jarang dimainkan.

Sebelum masuk ke proses clustering, data distandarisasi menggunakan Z-Score. Standarisasi dilakukan karena variabel Win Rate, Draw Rate, dan Average Moves memiliki skala nilai yang berbeda. Proses ini penting agar setiap variabel memiliki kontribusi yang seimbang dalam perhitungan jarak pada algoritma K-Means. [15] menjelaskan bahwa feature scaling berpengaruh terhadap hasil K-Means karena algoritma tersebut bekerja berdasarkan jarak antar data. Dengan demikian, dataset yang telah distandarisasi menjadi data utama yang digunakan dalam proses pengelompokan opening catur.

3.2 Hasil Pengelompokan K-Means

Pengelompokan data dilakukan menggunakan algoritma K-Means Clustering dengan tiga skenario jumlah cluster, yaitu 3 cluster, 4 cluster, dan 5 cluster. Pengujian dengan beberapa skenario dilakukan untuk mengetahui konfigurasi jumlah cluster yang menghasilkan kualitas pengelompokan terbaik. Setiap skenario menghasilkan anggota cluster, pusat cluster atau centroid, dan visualisasi distribusi data berdasarkan karakteristik statistik opening.

Hasil centroid dan jumlah anggota cluster pada setiap skenario ditunjukkan pada Tabel 4.

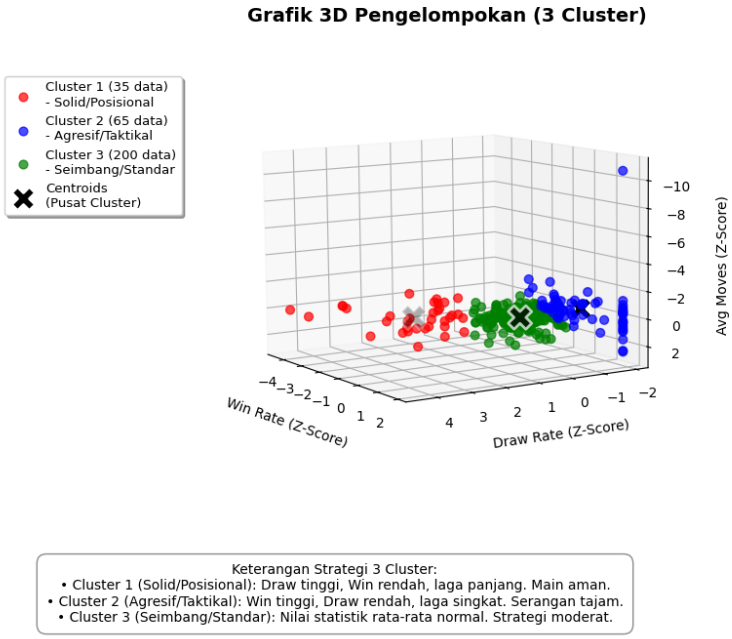
Tabel 4. Ringkasan pusat cluster pada skenario 3, 4, dan 5 cluster.

Jumlah Cluster	Pusat Cluster	Win Rate	Draw Rate	Avg Moves	Jumlah Data
3 Cluster	Centroid 1	-2,0354	2,0354	0,3894	35
3 Cluster	Centroid 2	1,1274	-1,1274	-0,6288	65
3 Cluster	Centroid 3	-0,0102	0,0102	0,1362	200
4 Cluster	Centroid 1	-0,0679	0,0679	0,3041	161
4 Cluster	Centroid 2	0,4980	-0,4980	-0,8871	79
4 Cluster	Centroid 3	1,7135	-1,7135	0,2994	25

Jumlah Cluster	Pusat Cluster	Win Rate	Draw Rate	Avg Moves	Jumlah Data
4 Cluster	Centroid 4	-2,0354	2,0354	0,3894	35
5 Cluster	Centroid 1	-0,1690	0,1690	0,3793	133
5 Cluster	Centroid 2	1,9115	-1,9115	-10,6780	1
5 Cluster	Centroid 3	1,6600	-1,6600	0,2099	28
5 Cluster	Centroid 4	-2,0933	2,0933	0,3627	33
5 Cluster	Centroid 5	0,4166	-0,4166	-0,5576	105

Berdasarkan Tabel 4, variabel Win Rate dan Draw Rate menjadi faktor yang paling menonjol dalam membedakan karakteristik cluster. Nilai Win Rate yang tinggi biasanya diikuti oleh Draw Rate yang rendah, sehingga menunjukkan pola permainan yang lebih desisif. Sebaliknya, cluster dengan Draw Rate tinggi dan Win Rate rendah cenderung menunjukkan pembukaan yang lebih solid, tertutup, dan sering menghasilkan remis. Variabel Average Moves membantu membedakan durasi permainan, terutama pada cluster dengan permainan cepat atau sangat ekstrem.

3.3 Visualisasi Hasil 3 Cluster

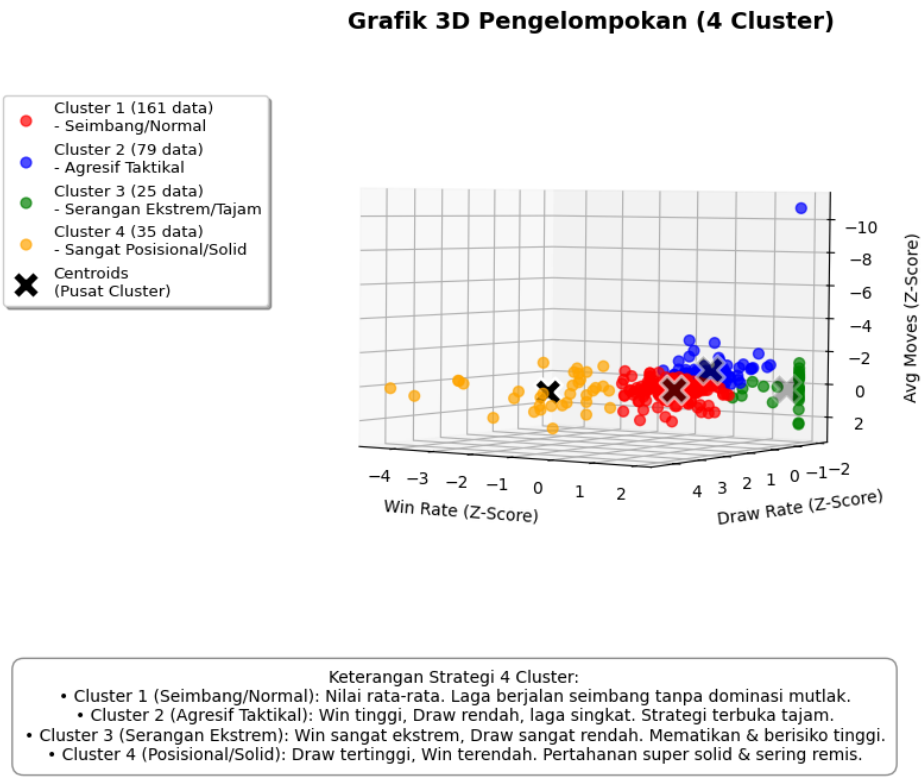


Gambar 6. Grafik 3 Cluster

Gambar 6 menunjukkan hasil visualisasi pengelompokan opening catur pada skenario 3 cluster. Pada skenario ini, data terbagi menjadi tiga kelompok utama. Cluster pertama memiliki nilai Win Rate rendah dan Draw Rate tinggi, sehingga dapat diinterpretasikan sebagai kelompok opening yang cenderung solid atau posisional. Cluster kedua memiliki nilai Win Rate tinggi dan Draw Rate rendah, sehingga menunjukkan kecenderungan opening yang lebih agresif atau taktikal. Cluster ketiga memiliki nilai centroid yang mendekati rata-rata, sehingga menggambarkan kelompok opening dengan karakteristik permainan yang lebih seimbang.

Pada skenario 3 cluster, jumlah anggota cluster terdiri dari C1 sebanyak 35 data, C2 sebanyak 65 data, dan C3 sebanyak 200 data. Jumlah anggota yang cukup besar pada C3 menunjukkan bahwa sebagian besar opening berada pada karakteristik yang relatif seimbang. Namun, karena hanya menggunakan tiga kelompok, pemisahan karakteristik permainan masih bersifat umum. Hal ini menyebabkan beberapa opening dengan karakteristik yang berbeda masih berada dalam kelompok yang sama.

3.4 Visualisasi Hasil 4 Cluster

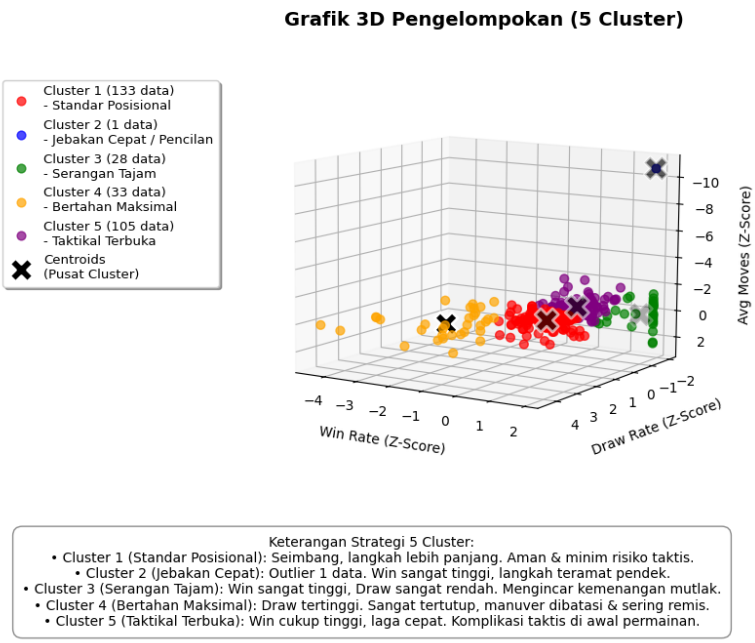


Gambar 7. Grafik 4 Cluster

Gambar 7 menunjukkan hasil visualisasi pengelompokan pada skenario 4 cluster. Pada skenario ini, data opening catur terbagi menjadi empat kelompok, sehingga pemisahan karakteristik menjadi lebih rinci dibandingkan skenario 3 cluster. Cluster pertama terdiri dari 161 data dengan nilai centroid yang relatif mendekati rata-rata, sehingga dapat dikategorikan sebagai kelompok opening seimbang dengan kecenderungan sedikit lebih solid. Cluster kedua terdiri dari 79 data dengan Win Rate lebih tinggi, Draw Rate lebih rendah, dan Average Moves lebih rendah, sehingga menunjukkan kecenderungan permainan yang lebih praktis dan cepat.

Cluster ketiga terdiri dari 25 data dengan nilai Win Rate tinggi dan Draw Rate rendah. Kelompok ini dapat diinterpretasikan sebagai opening yang cenderung desisif atau agresif. Sementara itu, cluster keempat terdiri dari 35 data dengan nilai Win Rate rendah dan Draw Rate tinggi, sehingga menunjukkan karakteristik opening yang lebih solid, tertutup, dan cenderung menghasilkan remis. Dibandingkan skenario 3 cluster, skenario 4 cluster memberikan pemisahan yang lebih jelas antara opening seimbang, agresif, dan solid.

3.5 Visualisasi Hasil 5 Cluster



Gambar 8. Grafik 5 Cluster

Gambar 8 menunjukkan hasil visualisasi pengelompokan opening catur pada skenario 5 cluster. Pada skenario ini, pola karakteristik opening dapat dipisahkan secara lebih detail. Cluster pertama terdiri dari 133 data dengan nilai Win Rate sedikit rendah, Draw Rate sedikit tinggi, dan Average Moves positif. Kelompok ini dapat diinterpretasikan sebagai opening yang cenderung seimbang tetapi sedikit mengarah ke permainan solid atau posisional.

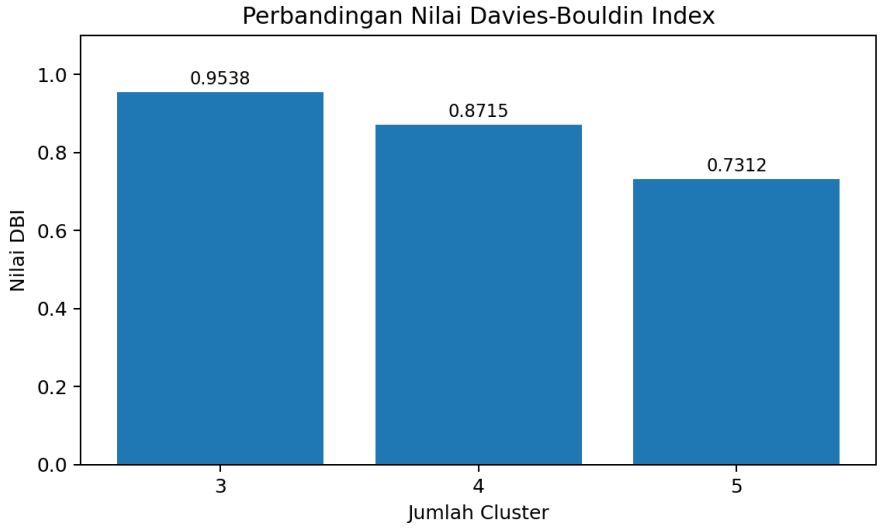
Cluster kedua hanya terdiri dari 1 data dengan nilai Win Rate tinggi, Draw Rate rendah, dan Average Moves sangat rendah. Kondisi ini menunjukkan adanya opening dengan karakteristik yang sangat berbeda dibandingkan data lainnya. Cluster tersebut dapat dianggap sebagai kelompok khusus atau kemungkinan data pencilan karena jumlah anggotanya hanya satu. Meskipun demikian, keberadaan cluster ini tetap penting untuk diperhatikan karena menunjukkan bahwa terdapat opening tertentu yang memiliki pola statistik ekstrem.

Cluster ketiga terdiri dari 28 data dengan nilai Win Rate tinggi dan Draw Rate rendah. Kelompok ini menunjukkan karakteristik opening yang cenderung agresif atau taktikal. Cluster keempat terdiri dari 33 data dengan nilai Win Rate rendah dan Draw Rate tinggi, sehingga dapat dikategorikan sebagai kelompok opening yang cenderung solid atau sering menghasilkan remis. Cluster kelima terdiri dari 105 data dengan nilai Win Rate cukup tinggi, Draw Rate rendah, dan Average Moves negatif, sehingga dapat diinterpretasikan sebagai kelompok opening yang cenderung praktis, lebih cepat, dan cukup desisif.

Hasil skenario 5 cluster menunjukkan bahwa penambahan jumlah cluster mampu memisahkan karakteristik opening secara lebih detail. Hal ini sejalan dengan tujuan clustering, yaitu menempatkan objek yang memiliki kemiripan karakteristik ke dalam kelompok yang sama dan memisahkan objek yang memiliki karakteristik berbeda ke dalam kelompok lain [26] Dengan demikian, skenario 5 cluster memberikan gambaran yang lebih rinci terhadap kecenderungan strategi opening dibandingkan skenario 3 dan 4 cluster.

3.6 Evaluasi Davies-Bouldin Index

Evaluasi hasil clustering dilakukan menggunakan Davies-Bouldin Index (DBI). DBI digunakan untuk mengukur kualitas cluster berdasarkan kekompakan data di dalam cluster dan pemisahan antar cluster. Nilai DBI yang lebih kecil menunjukkan kualitas pengelompokan yang lebih baik karena cluster yang terbentuk memiliki pemisahan yang lebih jelas dan anggota cluster yang lebih kompak ([23], [24]).



Gambar 9. Perbandingan nilai DBI pada skenario jumlah cluster.

Gambar 9 menunjukkan perbandingan nilai DBI pada skenario 3 cluster, 4 cluster, dan 5 cluster. Nilai DBI dari setiap skenario ditampilkan pada Tabel 5.

Tabel 5. Hasil pengujian Davies-Bouldin Index.

No.	Jumlah Cluster	Jumlah Anggota Cluster	Nilai DBI
1	3 Cluster	C1 = 35, C2 = 65, C3 = 200	0,9538
2	4 Cluster	C1 = 161, C2 = 79, C3 = 25, C4 = 35	0,8715
3	5 Cluster	C1 = 133, C2 = 1, C3 = 28, C4 = 33, C5 = 105	0,7312

Berdasarkan Tabel 5, skenario 3 cluster memperoleh nilai DBI sebesar 0,9538. Nilai tersebut merupakan nilai terbesar dibandingkan skenario lainnya, sehingga kualitas pengelompokan pada 3 cluster belum menjadi hasil terbaik. Pada skenario ini, data memang sudah dapat dipisahkan menjadi tiga kelompok utama, tetapi pemisahan antar cluster belum sebaik skenario 4 dan 5 cluster.

Skenario 4 cluster memperoleh nilai DBI sebesar 0,8715. Nilai ini lebih kecil dibandingkan skenario 3 cluster, sehingga dapat dikatakan bahwa kualitas pengelompokan mengalami peningkatan. Pembagian menjadi 4 cluster mampu memisahkan karakteristik opening secara lebih rinci. Namun, nilai DBI pada skenario ini masih lebih besar dibandingkan skenario 5 cluster, sehingga 4 cluster belum menjadi konfigurasi terbaik.

Skenario 5 cluster memperoleh nilai DBI sebesar 0,7312. Nilai ini merupakan nilai terkecil dari seluruh skenario pengujian. Berdasarkan prinsip DBI, nilai yang lebih kecil menunjukkan bahwa cluster memiliki kekompakan internal yang lebih baik dan pemisahan antar cluster yang lebih jelas. Dengan demikian, skenario 5 cluster dipilih sebagai hasil pengelompokan terbaik dalam penelitian ini. Meskipun terdapat satu cluster yang hanya berisi 1 data, skenario 5 cluster tetap menunjukkan nilai evaluasi terbaik berdasarkan DBI.

Hasil evaluasi DBI pada penelitian ini menunjukkan bahwa skenario 5 cluster memiliki nilai terkecil dibandingkan skenario 3 dan 4 cluster. Temuan ini sejalan dengan penelitian [27] yang menggunakan Davies-Bouldin Index untuk menentukan jumlah cluster optimal pada proses K-Means. Prinsip yang digunakan adalah semakin kecil nilai DBI, maka semakin baik kualitas pengelompokan karena data dalam cluster semakin kompak dan jarak antar cluster semakin terpisah.

Penggunaan beberapa skenario jumlah cluster penting dilakukan karena nilai K yang berbeda dapat menghasilkan kualitas pengelompokan yang berbeda. [14] melakukan pengujian K-Means dengan beberapa variasi jumlah cluster dan menggunakan Davies-Bouldin Index untuk menentukan konfigurasi terbaik. Dengan pendekatan yang sama, penelitian ini membandingkan skenario 3 cluster, 4 cluster, dan 5 cluster, kemudian memilih 5 cluster sebagai konfigurasi terbaik karena memiliki nilai DBI paling kecil.

3.7 Pembahasan Kecenderungan Strategi Pemain

Hasil clustering menunjukkan bahwa pola opening catur dapat dikelompokkan berdasarkan kemiripan karakteristik statistik permainan. Hasil clustering perlu diinterpretasikan berdasarkan karakteristik setiap kelompok agar tidak berhenti pada hasil numerik semata. [17] menunjukkan bahwa hasil K-Means dapat dianalisis melalui karakteristik centroid dan anggota cluster untuk memahami perbedaan antar kelompok. Dalam penelitian ini, interpretasi dilakukan dengan melihat nilai centroid pada variabel Win Rate, Draw Rate, dan Average Moves sehingga setiap cluster dapat dikaitkan dengan kecenderungan strategi seperti agresif, solid, seimbang, atau karakteristik khusus. Variabel Win Rate dan Draw Rate menjadi pembeda utama antar cluster karena keduanya menunjukkan hubungan yang berlawanan. Opening dengan Win Rate tinggi dan Draw Rate rendah cenderung menghasilkan permainan yang lebih desisif. Opening dengan Draw Rate tinggi dan Win Rate rendah cenderung menunjukkan karakteristik permainan yang lebih solid atau posisional. Sementara itu, Average Moves membantu melihat kecenderungan panjang atau pendeknya permainan pada setiap kelompok.

Dalam konteks strategi permainan, cluster dengan Win Rate tinggi dapat diinterpretasikan sebagai kelompok opening yang lebih agresif atau taktikal karena cenderung menghasilkan kemenangan lebih tinggi dan tingkat remis lebih rendah. Cluster dengan Draw Rate tinggi dapat diinterpretasikan sebagai kelompok opening yang lebih solid, tertutup, atau posisional karena cenderung menghasilkan permainan yang lebih aman dan lebih sering berakhir remis. Cluster dengan nilai centroid mendekati rata-rata menunjukkan opening dengan karakteristik seimbang, yaitu tidak terlalu agresif dan tidak terlalu solid.

Hasil ini mendukung pandangan bahwa opening catur dapat dianalisis secara kuantitatif berdasarkan data pertandingan online. [4] menjelaskan bahwa opening catur memiliki tingkat kompleksitas dan kemiripan yang dapat dianalisis menggunakan data komunitas catur online. Selain itu, [5] menunjukkan bahwa data catur dalam format PGN dapat digunakan untuk analisis clustering. Dengan demikian, hasil penelitian ini memperkuat bahwa data statistik opening dapat digunakan untuk mengelompokkan karakteristik permainan dan membantu memahami kecenderungan strategi pemain.

Secara keseluruhan, skenario 5 cluster menjadi konfigurasi terbaik karena menghasilkan nilai DBI paling kecil, yaitu 0,7312. Pengelompokan ini memberikan pemisahan karakteristik yang lebih rinci dibandingkan skenario 3 dan 4 cluster. Hasil tersebut menunjukkan bahwa algoritma K-Means Clustering dapat digunakan untuk mengelompokkan opening catur berdasarkan karakteristik statistik permainan. Dengan adanya hasil clustering ini, pemain dapat melihat kemiripan karakteristik antar opening tanpa harus membandingkan setiap variasi pembukaan secara manual.

Visualisasi hasil clustering berperan penting untuk membantu pembaca memahami distribusi data dan perbedaan karakteristik antar cluster. [28] menjelaskan bahwa visualisasi seperti peta interaktif dan grafik dapat memperjelas distribusi hasil clustering serta membantu proses pengambilan keputusan. Dalam penelitian ini, visualisasi 3D digunakan untuk menunjukkan sebaran opening berdasarkan Win Rate, Draw Rate, dan Average Moves, sehingga perbedaan karakteristik antar cluster dapat dilihat secara lebih jelas.

4. KESIMPULAN

Penelitian ini berhasil mengolah data pertandingan catur online dari platform Lichess menjadi dataset statistik pembukaan catur berdasarkan kode Encyclopedia of Chess Openings (ECO). Data yang digunakan difokuskan pada kategori permainan Rapid dengan pemain yang memiliki rating di atas 1200. Dataset akhir terdiri dari 300 kode ECO dengan frekuensi kemunculan tertinggi yang direpresentasikan menggunakan tiga variabel statistik, yaitu Win Rate, Draw Rate, dan Average Moves. Sebelum dilakukan proses clustering, data terlebih dahulu melalui tahapan seleksi, preprocessing, transformasi, dan standarisasi menggunakan Z-Score.

Algoritma K-Means Clustering dapat diterapkan untuk mengelompokkan pola pembukaan catur berdasarkan kemiripan karakteristik statistik permainan. Pengujian dilakukan pada tiga skenario jumlah cluster, yaitu 3 cluster, 4 cluster, dan 5 cluster. Hasil pengujian menunjukkan bahwa skenario 3 cluster memperoleh nilai Davies-Bouldin Index sebesar 0,9538, skenario 4 cluster memperoleh nilai 0,8715, dan skenario 5 cluster memperoleh nilai 0,7312. Berdasarkan nilai Davies-Bouldin Index terkecil, skenario 5 cluster menjadi hasil pengelompokan terbaik dalam penelitian ini.

Hasil clustering menunjukkan bahwa opening catur dapat dikelompokkan ke dalam beberapa kecenderungan strategi berdasarkan karakteristik Win Rate, Draw Rate, dan Average Moves. Cluster dengan Win Rate tinggi dan Draw Rate rendah menggambarkan opening yang cenderung agresif atau taktikal. Cluster dengan Draw Rate tinggi dan Win Rate rendah menggambarkan opening yang cenderung solid atau posisional. Sementara itu, cluster dengan

nilai centroid mendekati rata-rata menunjukkan opening dengan karakteristik permainan yang lebih seimbang. Dengan demikian, hasil penelitian ini dapat memberikan gambaran objektif mengenai kemiripan karakteristik opening catur berdasarkan data pertandingan nyata.

Penelitian selanjutnya dapat dikembangkan dengan menambahkan variabel lain, seperti rating rata-rata pemain, durasi waktu permainan, akurasi langkah, atau perbedaan performa antara pemain putih dan hitam. [29] membandingkan K-Means dan DBSCAN menggunakan Silhouette Index dan Davies-Bouldin Index, serta menunjukkan bahwa karakteristik dataset dapat memengaruhi performa algoritma yang digunakan. Oleh karena itu, penelitian berikutnya dapat membandingkan K-Means dengan DBSCAN, K-Medoids, atau Hierarchical Clustering untuk mengetahui metode yang paling sesuai dalam pengelompokan opening catur. [30] menunjukkan bahwa K-Medoids dapat menghasilkan nilai Davies-Bouldin Index yang lebih kecil dibandingkan K-Means pada data yang memiliki outlier atau distribusi tidak merata. Oleh karena itu, penggunaan algoritma pembanding pada penelitian lanjutan dapat memberikan evaluasi yang lebih komprehensif terhadap pengelompokan opening catur. Hasil penelitian ini juga dapat dikembangkan menjadi dashboard analisis atau sistem rekomendasi opening catur berbasis karakteristik permainan.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada STMIK Kaputama, Program Studi Teknik Informatika, dosen pembimbing, serta seluruh pihak yang telah memberikan dukungan dalam penyusunan penelitian ini.

REFERENSI

[1] R. McIlroy-Young, R. Wang, S. Sen, J. Kleinberg, and A. Anderson, “Learning Models of Individual Behavior in Chess,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, no. August 2022, pp. 1253–1263, 2022, doi: 10.1145/3534678.3539367.
[2] S. Chowdhary, I. Iacopini, and F. Battiston, “Quantifying human performance in chess,” *Sci. Rep.*, vol. 13, no. 1, pp. 1–8, 2023, doi: 10.1038/s41598-023-27735-9.
[3] R. Sanjaya, J. Wang, and Y. Yang, “Measuring the Non-Transitivity in Chess,” pp. 1–22, 2022.
[4] G. De Marzo and V. D. P. Servedio, “Quantifying the complexity and similarity of chess openings using online chess community data,” *Sci. Rep.*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-31658-w.
[5] F. Wijayanto, “Clustering Analysis of Chess Portable Game Notation Text,” *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 3, no. 3, pp. 137–142, 2024, doi: 10.20885/snati.v3.i3.42.
[6] F. Wijayanto, “Forsyth-Edwards Notation in Chess Game Clustering: A Depth-Based Evaluation,” *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 4, no. 1, pp. 18–25, 2025, doi: 10.20885/snati.v4.i1.3.
[7] F. Wijayanto, “Jurnal Sains , Nalar , dan Aplikasi Teknologi Informasi Mapping the Safest Routes : A Clustering Study of the French Defense,” vol. 4, no. 2, pp. 54–61, 2025, doi: 10.20885/snati.v4.i2.40910.
[8] M. A. Aldi *et al.*, “IMPLEMENTASI K-MEANS CLUSTER ING DALAM PENGELOMPOKAN DATA KUNJUNGAN,” vol. 2, no. 1, pp. 13–19, 2025.
[9] B. Aji, K. Batam, and K. Riau, “BLITZ PADA DATASET TIDAK SEIMBANG SKALA BESAR MENGGUNAKAN,” vol. 13, no. 1, pp. 46–52, 2026.
[10] H. L. Siregar, R. Hidayanthi, and A. L. D. Sakti, “Implementation of K-Means clustering on student learning achievements based on social economic and social related,” *Res. Dev. Educ.*, vol. 4, no. 2, pp. 1447–1459, 2024, doi: 10.22219/raden.v4i2.36742.
[11] S. R. Agustin, I. Purnamasari, and B. N. Sari, “Implementasi K-Means Untuk Pengelompokan Kategori Penjualan Barang Berbasis Web,” vol. 5, no. 3, pp. 167–176, 2025, doi: 10.47065/jimat.v5i3.610.
[12] M. D. Adrian and F. Wijayanto, “Chesstify : Chess Game Database Management System using Portable Game Notation,” vol. 8, no. 1, 2026, doi: 10.35842/ijicom.
[13] M. Rochmawati *et al.*, “Implementasi Algoritma K-Means dalam Klasterisasi Penjualan pada Sebuah Perusahaan menggunakan Metodologi KDD Implementation of the K-Means Algorithm in Sales Clustering at a Company using the KDD Methodology,” vol. 13, pp. 54–62, 2024.
[14] D. A. Tarigan, “Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification,” vol. 4, no. 1, pp. 545–552, 2023, doi: 10.30865/klik.v4i1.964.
[15] C. W. Id, “The impact of neglecting feature scaling in k-means clustering,” pp. 1–19, 2024, doi: 10.1371/journal.pone.0310839.
[16] J. Reiner, B. Stilwell, and A. Wahbeh, “Leveraging K-Means Clustering and Z-Score for Anomaly Detection in Bitcoin Transactions,” 2025.
[17] F. Anggraeny, “Penerapan K-Means dengan Evaluasi Davies-Bouldin Index untuk Pengelompokan Kelas Unggulan SMP Wijaya Sukodono,” vol. 15, no. 2, pp. 372–382, 2025.
[18] T. Ikhsan, E. Haerani, F. Wulandari, and F. Syafria, “Clustering Data Penduduk Menggunakan Algoritma K-Means TIN : Terapan Informatika Nusantara,” vol. 5, no. 12, pp. 679–687, 2025, doi: 10.47065/tin.v5i12.7328.
[19] N. Amalia, “Z-Score Based Initialization for K-Medoids Clustering : Application on QSAR Toxicity Data,” vol. 9, no. 5, pp. 2410–2417, 2025.
[20] S. Syaqila and M. Fakhriza, “K-Means Clustering Untuk Mengukur Pengaruh Kompetensi Terhadap Kinerja Pegawai,” vol. 6, no. 2, pp. 1420–1431, 2025, doi: 10.47065/josh.v6i2.6758.
[21] I. D. Setiawan and A. Triayudi, “Penerapan Algoritma Clustering K-Means Data Mining dalam Pengelompokan Mahasiswa Penerima Beasiswa,” vol. 5, no. 2, pp. 430–441, 2024, doi: 10.47065/josyc.v5i2.4971.
[22] U. W. Latifah, S. Bahri, and M. Satriandhini, “Implementasi Algoritma K-Means Clustering untuk Strategi Promosi Kampus IBISA Implementation of K-Means Clustering Algorithm for IBISA Campus Promotion Strategy,” no. 2, pp. 292–300, 2024, doi: 10.26798/jiko.v8i2.1307.
[23] M. Sholeh and K. Aeni, “Perbandingan Evaluasi Metode Davies Bouldin, Elbow dan Silhouette pada Model Clustering dengan

Menggunakan Algoritma K-Means," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 8, no. 1, p. 56, 2023, doi: 10.30998/string.v8i1.16388.

[24] I. F. Ashari, E. D. Nugroho, R. Baraku, I. N. Yanda, and R. Liwardana, "Analysis of Elbow , Silhouette , Davies-Bouldin , Calinski-Harabasz , and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood- Affected Areas in Jakarta," vol. 7, no. 1, pp. 95–103, 2023.

[25] A. Saputra and R. Yusuf, "Comparison of the DBSCAN and K-MEANS Algorithms in Segmenting Customers Using Public Transportation of Transjakarta Using the RFM Method Perbandingan Algoritma DBSCAN dan K-MEANS dalam Segmentasi Pelanggan Pengguna Transportasi Publik Transjakarta Menggunakan Metode RFM," vol. 4, no. October, pp. 1346–1361, 2024.

[26] L. Bai and J. Liang, "A categorical data clustering framework on graph representation," *Pattern Recognit.*, vol. 128, p. 108694, 2022, doi: 10.1016/j.patcog.2022.108694.

[27] I. F. Ashari, R. Banjarnahor, and D. R. Farida, "Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies," vol. 6, no. 1, pp. 7–15, 2022.

[28] B. Ariansah *et al.*, "Penerapan K-Means Clustering untuk Pengelompokan Data Industri Kecil Menengah di Provinsi Jambi," vol. 6, no. September, pp. 359–371, 2025, doi: 10.35957/jtsi.v6i2.13553.

[29] N. A. Yolandari, L. E. Butarbutar, G. Citra, and H. Rajagukguk, "ANALISIS PERBANDINGAN K-MEANS DAN DBSCAN DALAM PENGELOMPOKAN DATA TRAVEL REVIEW RATINGS MENGGUNAKAN EVALUASI SILHOUETTE INDEX DAN DAVIES-BOULDIN INDEX," vol. 13, no. 3.

[30] M. D. Salman, N. R. Pratama, and M. N. F. A, "Comparison of K-Means and K-Medoids Clustering Algorithm Performance in Grouping Schools in Riau Province Based on Availability of Facilities and Infrastructure Perbandingan Kinerja Algoritma Clustering K-Means dan K-Medoids dalam Pengelompokan Sekolah di Provinsi Riau Berdasarkan Ketersediaan Sarana dan Prasarana," vol. 5, no. July, pp. 797–806, 2025.