Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 6 Iss. 4 October 2026, pp: 1930-1940
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Comparison of K-Means and DBSCAN Algorithms in Segmenting Risk and Profit Profiles of Stocks on the Indonesia Stock Exchange

Perbandingan Algoritma K-Means dan DBSCAN dalam Segmentasi Profil Risiko dan Keuntungan Saham di Bursa Efek Indonesia

Zeydan Fazle Mawla[1], Dwiki Aprilian Aryanda[2], Masna Wati[3*], Joan Angelina Widians[4]

[1,2,3]Program Studi Informatika, Fakultas Teknik, Universitas Mulawarman, Indonesia
[4]Program Studi Magister Informatika, Fakultas Teknik, Universitas Mulawarman, Indonesia

E-Mail: [1]zeydanfm2006@gmail.com, [2]aprilianaryandad@gmail.com, [3]masnawati@fkti.unmul.ac.id, [4]angelwidians@unmul.ac.id

Received May 04th 2026; Revised Jun 25th 2026; Accepted Jul 12th 2026; Available Online Oct 30th 2026
Corresponding Author: Masna Wati
Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

The high dynamics of daily stock price fluctuations require investors to have a precise portfolio strategy. Traditional clustering algorithms like K-Means are often vulnerable and biased toward outlier data. This study compares the performance of K-Means and DBSCAN algorithms in segmenting the risk and profit profiles of IDX30 index stocks. The methodology uses Knowledge Discovery in Databases (KDD) by extracting expected return and daily volatility features. The cluster architecture evaluation was measured comparatively using the Silhouette Score, Davies-Bouldin Index (DBI), and Dunn Index. The results proved that DBSCAN is significantly superior to K-Means. DBSCAN obtained the highest Silhouette Score of 0.518, the lowest DBI of 0.481, and the highest Dunn Index of 0.245, while K-Means only reached 0.391, 0.810, and 0.200. DBSCAN effectively separated five highly speculative anomalous stocks as noise (pump-and-dump stocks), and accurately grouped the remaining constituents into moderate and aggressive clusters without overlapping.

Keywords: Algorithms, DBSCAN, Indonesia Stock Exchange, K-Means, Segmentation

Abstrak

Tingginya dinamika fluktuasi harga saham harian menuntut investor untuk memiliki strategi portofolio yang presisi. Algoritma clustering tradisional seperti K-Means seringkali rentan dan bias terhadap data pencilan (outliers). Penelitian ini membandingkan kinerja algoritma K-Means dan DBSCAN dalam menyegmentasikan profil risiko dan keuntungan saham indeks IDX30. Metodologi yang digunakan berbasis Knowledge Discovery in Databases (KDD) dengan mengekstraksi fitur ekspektasi keuntungan dan volatilitas harian. Evaluasi arsitektur klaster diukur secara komparatif menggunakan Silhouette Score, Davies-Bouldin Index (DBI) dan Dunn Index. Hasil penelitian membuktikan bahwa DBSCAN jauh lebih unggul dibandingkan K-Means. DBSCAN memperoleh Silhouette Score tertinggi sebesar 0.518, DBI terendah sebesar 0.481 dan Dunn Index tertinggi sebesar 0.245, sedangkan K-Means hanya mencapai 0.391, 0.810, dan 0.200. DBSCAN terbukti secara efektif memisahkan lima saham anomali yang sangat spekulatif sebagai noise (saham gorengan), serta mengelompokkan sisa emiten secara akurat ke dalam klaster moderat dan agresif tanpa tumpang tindih.

Kata Kunci: Algoritma, Bursa Efek Indonesia, DBSCAN, K-Means, Segmentasi

1. PENDAHULUAN

Pasar modal merupakan instrumen krusial dalam mendorong stabilitas ekonomi nasional melalui mobilisasi dana masyarakat. Di Bursa Efek Indonesia (BEI), investor dihadapkan pada tantangan untuk menyeimbangkan antara potensi keuntungan (*return*) dan tingkat risiko (*risk*) yang ada. Dinamika pasar modal sangat dipengaruhi oleh kondisi ketidakpastian global atau fase *Volatility, Uncertainty, Complexity, Ambiguity* (VUCA), yang memicu fluktuasi harga saham secara signifikan [1]. Tingginya dinamika fluktuasi harga saham harian, khususnya pada konstituen indeks IDX30, menuntut investor untuk memiliki strategi manajemen portofolio yang presisi guna memitigasi risiko dan terhindar dari saham gorengan [2]. Agar strategi investasi

ini berhasil, pemahaman mendalam terhadap faktor-faktor lingkungan pasar seperti ekspektasi keuntungan dan metrik volatilitas sangatlah penting. Akan tetapi, masih banyak investor yang melakukan analisis pemetaan profil risiko secara parsial atau mengandalkan pengamatan manual, yang memakan banyak waktu, tenaga, serta rawan terhadap kesalahan yang dapat memengaruhi keseimbangan hasil investasi. Kondisi ini menuntut adanya strategi pengelolaan portofolio yang terukur di mana penelitian [3] menegaskan pentingnya pemilihan kombinasi aset yang tepat melalui optimalisasi portofolio untuk memitigasi risiko investasi. Sebagai solusi, diperlukan penerapan teknologi penambangan data (*data mining*), khususnya algoritma *clustering*, sebagai langkah yang tepat untuk mensegmentasi profil saham secara objektif dan akurat. Proses *clustering* memungkinkan pengolahan data riwayat perdagangan yang besar menjadi kelompok-kelompok emiten dengan karakteristik serupa. Penerapan metode analitik ini dapat membuktikan peningkatan ketepatan pemantauan profil pasar serta membantu pengambilan keputusan investasi yang lebih cepat dan meminimalkan risiko kerugian [4].

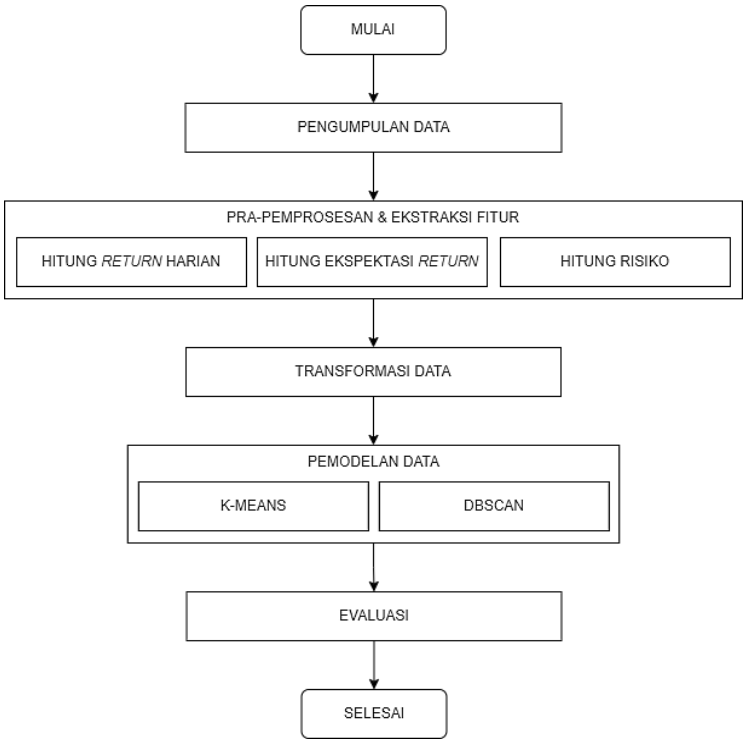
Beberapa penelitian terdahulu menunjukkan bahwa metode *clustering* mampu digunakan secara efektif untuk menganalisis pola pergerakan saham dan mendukung pengambilan keputusan investasi. Penelitian [4] memanfaatkan K-Means dan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) untuk mengelompokkan saham berdasarkan kemiripan informasi tekstual dan pergerakan harga. Namun, penelitian tersebut lebih berfokus pada integrasi *textual embeddings* untuk kebutuhan prediksi harga, sehingga evaluasi kualitas struktur klaster dan analisis data anomali tidak menjadi fokus utama penelitian. Sementara itu, penelitian [5] menunjukkan bahwa K-Means dapat digunakan untuk mengidentifikasi pola harga saham melalui pembentukan kelompok data yang memiliki karakteristik serupa. Meskipun demikian, metode K-Means secara teoritis memiliki sensitivitas terhadap *outlier* karena seluruh data cenderung dipartisi ke dalam klaster tertentu. Penelitian lain seperti [6], [7] juga menegaskan bahwa teknik *clustering* berperan penting dalam segmentasi portofolio, optimasi investasi, dan identifikasi hubungan antar saham. Akan tetapi, kedua penelitian tersebut lebih menitikberatkan pada strategi perdagangan dan optimasi portofolio dibandingkan evaluasi performa algoritma *clustering* dalam pemetaan profil risiko saham. Selain itu, studi [8] memperlihatkan bahwa kombinasi *clustering* dan optimasi mampu meningkatkan efisiensi strategi alokasi aset pada pasar saham, meskipun belum secara spesifik membahas efektivitas metode berbasis densitas seperti DBSCAN dalam mengisolasi data anomali pasar.

Berdasarkan beberapa penelitian tersebut, masih terdapat sejumlah keterbatasan yang belum banyak dibahas secara mendalam. Sebagian besar penelitian lebih berfokus pada optimalisasi portofolio, prediksi harga, atau pemilihan pasangan saham, namun belum secara spesifik mengevaluasi kemampuan algoritma clustering dalam menangani data anomali (*outlier*) pada pasar saham yang memiliki volatilitas tinggi [6]. Penggunaan algoritma K-Means juga cenderung sensitif terhadap keberadaan *noise* karena seluruh data dipaksa masuk ke dalam klaster tertentu, sehingga berpotensi menghasilkan segmentasi yang bias [5]. Di sisi lain, beberapa penelitian belum melakukan analisis komparatif yang komprehensif antara metode berbasis *centroid* seperti K-Means dan metode berbasis densitas seperti DBSCAN menggunakan metrik evaluasi internal, misalnya Silhouette Score dan Davies-Bouldin Index [9]. Selain itu, penerapan *clustering* pada saham indeks IDX30 di Bursa Efek Indonesia masih relatif terbatas, khususnya dalam konteks pemetaan profil risiko dan tingkat pengembalian saham [10].

Untuk mengatasi keterbatasan tersebut, implementasi strategi segmentasi saham yang memetakan parameter *expected return* dan volatilitas risiko secara terintegrasi sangat diperlukan. Oleh karena itu, penelitian ini mengusulkan pendekatan komparatif antara algoritma K-Means dan DBSCAN untuk mensegmentasi saham indeks IDX30. Penelitian ini tidak hanya berfokus pada pembentukan klaster, tetapi juga mengevaluasi kualitas struktur klaster secara objektif menggunakan metrik Silhouette Score, Davies-Bouldin Index dan Dunn Index guna memperoleh model yang lebih optimal [11]. Penggunaan DBSCAN secara khusus diharapkan mampu mengisolasi saham-saham anomali (saham gorengan) secara efektif tanpa memaksakan data pencilan masuk ke dalam klaster utama [12]. Dengan analisis data yang terus dikumpulkan, model evaluasi ini diharapkan dapat memberikan pemetaan profil investasi yang lebih akurat dan adaptif terhadap dinamika pasar, sehingga dapat menjadi landasan perencanaan portofolio yang efisien, terukur, dan relevan bagi pengambilan keputusan investor.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif yang memanfaatkan teknik penambangan data (data mining) untuk mengelompokkan dan mengidentifikasi anomali pada data pasar saham. Metodologi yang digunakan secara umum mengikuti tahapan Knowledge Discovery in Databases (KDD). KDD adalah proses sistematis dan terstruktur untuk mengekstraksi pengetahuan, menemukan pola yang bermakna, serta mengidentifikasi hubungan tersembunyi dari sekumpulan data berukuran besar [13], yang meliputi pemilihan data, pra-pemrosesan, transformasi, pemodelan data mining, dan evaluasi yang ditampilkan pada Gambar 1.



Gambar 1. Metode Penelitian

2.1. Pengumpulan Data

Penelitian ini memanfaatkan data sekunder historis Bursa Efek Indonesia (BEI) yang diakuisisi melalui teknik ekstraksi data (*web scraping*) pada platform Yahoo Finance. Fokus utama observasi bertumpu pada catatan perdagangan harian, secara spesifik menggunakan metrik Harga Penutupan yang Disesuaikan (*Adjusted Closing Price*). Penggunaan *adjusted close* dipilih karena nilainya telah mengakomodasi penyesuaian terhadap aksi korporasi (*corporate action*) di masa lalu, seperti pembagian dividen, pemecahan saham (*stock split*), dan *right issue*, sehingga mencerminkan nilai pergerakan harga dan pengembalian (return) yang jauh lebih akurat secara historis. Adapun sampel penelitian mencakup saham-saham yang terdaftar dalam konstituen indeks IDX30. Untuk merepresentasikan dinamika pasar terkini, periode pengamatan ditetapkan selama satu tahun penuh, yakni rentang 3 Maret 2025 hingga 3 Maret 2026.

Lebih lanjut, untuk memastikan integritas dataset hasil *scraping*, prosedur prapemrosesan awal dilakukan dengan mengevaluasi keberadaan data hilang (*missing values*) yang umumnya disebabkan oleh perbedaan pencatatan hari libur bursa atau anomali sistem Yahoo Finance [14]. Berdasarkan hasil ekstraksi, diperoleh 237 hari perdagangan aktif yang seluruh datanya terisi secara lengkap dan utuh tanpa adanya *missing values*, sehingga tidak memerlukan tindakan imputasi data lebih lanjut. Sebagai langkah verifikasi akhir, proses validasi dilakukan secara acak silang (*cross-check validation*). Sampel pergerakan harga dari hasil *scraping* divalidasi dengan membandingkannya terhadap publikasi data historis resmi dari situs Bursa Efek Indonesia (BEI) guna memastikan tidak ada deviasi harga yang signifikan sebelum data tersebut diekstraksi menjadi fitur pemodelan.

2.2. Pra-pemrosesan Data dan Ekstraksi Fitur

Data mentah tidak dapat dimasukkan secara langsung ke dalam algoritma *clustering*. Data tersebut memerlukan transformasi menjadi metrik keuangan yang relevan: Ekspektasi Keuntungan (*Expected Return*) dan Risiko atau Volatilitas (*Risk*).

1. Perhitungan *Return* Harian

   Langkah pertama adalah menghitung persentase perubahan harian dari harga penutupan untuk setiap saham. Rumus untuk *return* harian (R_t) adalah:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \quad (1)$$

   Di mana P_t adalah harga penutupan pada hari ke-t, dan P_{t-1} adalah harga penutupan pada hari sebelumnya.

2. Ekspektasi Keuntungan (*Expected Return* yang Disetahunkan)

Untuk menstandarisasi pengukuran potensi profitabilitas, rata-rata *return* harian pada penelitian ini disetahunkan (*annualized*). Berdasarkan total 237 hari bursa yang tercatat antara rentang waktu 3 Maret 2025 sampai 3 Maret 2026, formula perhitungan Ekspektasi Keuntungan atau *Expected Return* ($E(R)$) adalah sebagai berikut:

$$E(R) = \underline{R} \times 237 \tag{2}$$

Di mana R adalah rata-rata aritmatika dari return harian.

3. Risiko atau Volatilitas (Disetahunkan)

Risiko direpresentasikan oleh standar deviasi dari *return* harian, yang mengukur penyebaran (dispersi) *return* di sekitar nilai rata-ratanya. Risiko yang disetahunkan (σ_{annual}) dihitung menggunakan rumus:

$$\sigma_{annual} = \sigma_{daily} \times \sqrt{237} \tag{3}$$

Di mana σ_{daily} adalah standar deviasi dari return harian. Kedua fitur yang diekstraksi ini, yaitu $E(R)$ dan σ_{annual}, berfungsi sebagai variabel utama (fitur) untuk proses clustering.

2.3. Transformasi Data

Tahap transformasi data diawali dengan mengintegrasikan dua variabel utama, yakni tingkat pengembalian yang diharapkan (*expected return*) dan tingkat risiko (*risk*), ke dalam sebuah matriks fitur. Setelah dataset terbentuk, proses dilanjutkan dengan tahap normalisasi menggunakan metode *Standard Scaler* (Standardisasi Z-Score) untuk menyeragamkan distribusi skala antar variabel [15]. Secara matematis, transformasi Z-Score dirumuskan sebagai berikut:

$$Z = \frac{(X-\mu)}{\sigma} \tag{4}$$

Dimana Z merupakan nilai data setelah melalui tahap normalisasi (*standardized value*). Nilai dihitung berdasarkan X, yakni nilai data pada fitur, kemudian dikurangi dengan μ sebagai representasi nilai rata-rata (*mean*) dari keseluruhan observasi fitur. Hasil metrik tersebut selanjutnya dibagi dengan σ, yang mendefinisikan nilai standar deviasi (*standard deviation*) dari fitur guna menstandarkan rentang skala distribusinya.

Langkah pra-pemrosesan ini sangat krusial bagi algoritma *clustering* berbasis perhitungan jarak spasial guna mencegah atribut dengan rentang nilai asal yang lebih besar mendominasi atau membiaskan hasil pengelompokan [16].

2.4. Pemodelan Data Mining

Pemodelan memanfaatkan dua teknik klaster yang digunakan untuk mencapai tujuan penelitian, berikut adalah algoritmanya:

2.4.1. Clustering K-Means

Pada algoritma K-Means, proses diawali dengan penentuan jumlah klaster (K) yang optimal melalui Elbow Method. Metode ini dilakukan dengan memplot nilai inersia terhadap berbagai variasi nilai K untuk mengidentifikasi titik belok (*elbow*). Setelah nilai K optimal diperoleh, model K-Means diinisialisasi dengan memilih titik-titik pusat (*centroid*) awal secara acak. Algoritma kemudian dieksekusi untuk mempartisi data, yang diakhiri dengan visualisasi hasil klasterisasi [17], [18]. Rumus umum yang digunakan ditunjukkan pada Persamaan 5.

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|^2 \tag{5}$$

Dimana J merepresentasikan nilai fungsi objektif atau inersia dari model pengelompokan. Nilai ini dihitung dengan mempertimbangkan k, yakni jumlah total klaster yang telah ditentukan sebelumnya, dengan *j* bertindak sebagai indeks iterasi yang bergerak dari klaster ke-1 hingga klaster ke- k. Pada setiap iterasi, algoritma akan mengevaluasi C_j, yaitu himpunan titik data yang dikelompokkan ke dalam klaster ke-j. Secara spesifik, pengukuran dilakukan pada x_i yang mewakili titik data ke-i (merepresentasikan fitur emiten saham pada titik tertentu) terhadap μ_j, yakni titik pusat (*centroid*) dari klaster ke-j tersebut. Hasil akhirnya didapatkan melalui metrik $\left\| x_i - \mu_j \right\|^2$, yang mendefinisikan kuadrat dari jarak spasial antara titik data x_i dengan *centroid* μ_j.

2.4.2. *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN)

Tahap awal dalam implementasi algoritma DBSCAN adalah penetapan parameter Epsilon (ε) dan min_samples. Parameter ε berfungsi sebagai batas radius ketetanggaan, sedangkan min_samples menetapkan jumlah minimum observasi dalam radius ε tersebut agar suatu titik data dapat diklasifikasikan sebagai titik inti (*core point*). Nilai-nilai parameter yang telah didefinisikan tersebut kemudian diintegrasikan untuk menginisialisasi model DBSCAN [12], [19]. Persamaan umum yang digunakan yaitu ditunjukkan pada Persamaan 6.

$$N_\epsilon(p) = \{q \in D \mid d(p,q) \leq \epsilon\} \tag{6}$$

Dimana $N_\epsilon(p)$ mendefinisikan himpunan ketetanggaan Epsilon dari sebuah titik data p, yang bertindak sebagai titik data acuan yang sedang dievaluasi area ketetanggaannya. Evaluasi ini dilakukan pada D, yakni keseluruhan himpunan data atau *dataset* observasi, dengan membandingkannya terhadap q yang merupakan titik data lain. Pengukuran dilakukan menggunakan $\mathrm{d(p, q)}$, yaitu sebuah fungsi metrik jarak untuk menghitung jarak spasial antara titik p dan titik q. Ketetanggaan ini secara langsung dibatasi oleh ϵ, yakni parameter *Epsilon* yang merepresentasikan nilai batas radius maksimal dari area evaluasi tersebut.

2.5. Evaluasi

Fokus utama pada tahap evaluasi adalah melakukan analisis komparatif antara hasil pengelompokan algoritma K-Means dan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). Kualitas arsitektur klaster dari kedua model tersebut dinilai menggunakan metrik *Silhouette Score* dan *Davies-Bouldin Index* untuk mengevaluasi kemampuan algoritma dalam mengisolasi pola data secara optimal tanpa adanya tumpang tindih (*overlapping*) antar klaster [9], [11], [20]. Serta metrik Dunn Index yang ditambahkan untuk mengevaluasi kualitas klaster berdasarkan rasio antara jarak minimum antar-klaster terhadap diameter maksimal di dalam suatu klaster, di mana perolehan nilai yang lebih tinggi merepresentasikan bahwa klaster yang terbentuk sangat padat dan terpisah secara maksimal [21]. Metrik Silhouette Score dihitung untuk mengukur seberapa dekat suatu titik dengan klasternya sendiri dibandingkan dengan klaster tetangganya ditunjukkan pada Persamaan 7.

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{7}$$

Sementara itu, untuk menilai rasio jumlah penyebaran di dalam klaster terhadap pemisahan antar klaster, digunakan Davies-Bouldin Index ditunjukkan pada Persamaan 8.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{s_i+s_j}{d_{ij}}\right) \tag{8}$$

Lalu, untuk mengevaluasi seberapa padat dan terpisahnya klaster yang terbentuk, ditambahkan pengujian menggunakan metrik Dunn Index, ditunjukkan pada Persamaan 9.

$$DI = \frac{\min\limits_{1 \leq i < j \leq k} \delta(C_i, C_j)}{\max\limits_{1 \leq m \leq k} \Delta(C_m)} \tag{9}$$

Selanjutnya, model yang telah tervalidasi keandalannya akan memasuki tahap akhir KDD, yakni proses interpretasi pola matematis menjadi pengetahuan (*knowledge*) bisnis yang bermakna, khususnya mengenai profil risiko dan potensi tingkat pengembalian saham bagi investor [11].

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini bersumber dari platform Yahoo Finance, berupa riwayat Harga Penutupan yang Disesuaikan (*Adjusted Closing Price*) dari 30 perusahaan yang tergabung dalam indeks IDX30. Periode observasi data mencakup rentang waktu satu tahun, yaitu mulai 3 Maret 2025 hingga 3 Maret 2026. Dataset ini memiliki dimensi sebanyak 237 baris yang merepresentasikan hari perdagangan aktif, dan 30 kolom yang mewakili masing-masing saham, dengan tahapan pengolahan data sebagai berikut:

3.1. Pra-pemrosesan Data dan Ekstraksi Fitur

Langkah pertama dalam analisis ini adalah mentransformasi data mentah pergerakan harga saham harian menjadi metrik keuangan yang dapat dikuantifikasi, yaitu *Expected Return* atau $E(R)$ dan Risiko/Volatilitas atau σ_{annual} yang disetahunkan. Hasil perhitungan untuk 30 emiten selama periode observasi disajikan pada Tabel 1.

Tabel 1. Hasil Ekstraksi Fitur

Ticker	Expected Return	Risk
AADI	0.623060	0.424646
ADRO	0.437051	0.427474
AMRT	-0.286910	0.425761
ANTM	1.237528	0.564354
ASII	0.435443	0.340459
…	…	…
PGEO	0.366566	0.518861
PTBA	0.274535	0.246871
TLKM	0.506672	0.363726
UNTR	0.439313	0.384203
UNVR	0.811411	0.515488

3.2. Normalisasi Data

Setelah dataset terbentuk, tahap selanjutnya adalah melakukan normalisasi data menggunakan metode Standard Scaler (*Z-Score Standardization*). Proses ini mengkonversi nilai setiap fitur, yakni *expected return* dan *risk*, agar memiliki nilai rata-rata (mean) sebesar nol dan standar deviasi sebesar satu. Langkah transformasi ini sangat krusial untuk mencegah terjadinya bias akibat perbedaan skala angka saat model menghitung jarak antar titik data. Hasil normalisasi disajikan pada Tabel 2.

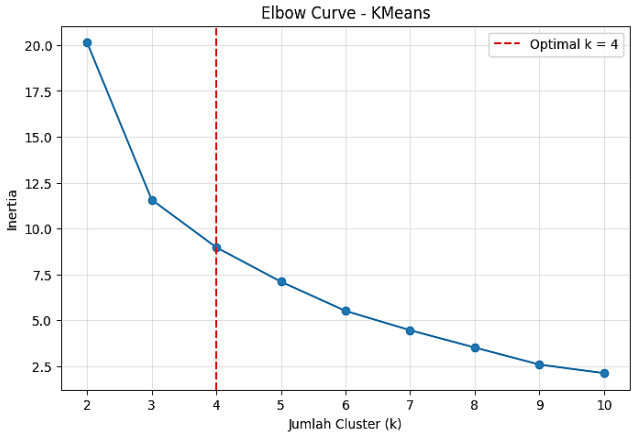
Tabel 2. Hasil Normalisasi dengan Standard Scaler

Ticker	Expected Return	Risk
AADI	0.317799	-0.232357
ADRO	-0.085701	-0.214741
AMRT	-1.656157	-0.225413
ANTM	1.650735	0.637990
ASII	-0.089191	-0.756823
…	…	…
PGEO	-0.219624	0.354575
PTBA	-0.557046	-1.339853
TLKM	0.006243	-0.611879
UNTR	-0.163805	-0.484310
UNVR	0.681530	0.333567

30 Rows × 3 Columns

3.3. Pembentukan Model dengan K-Means

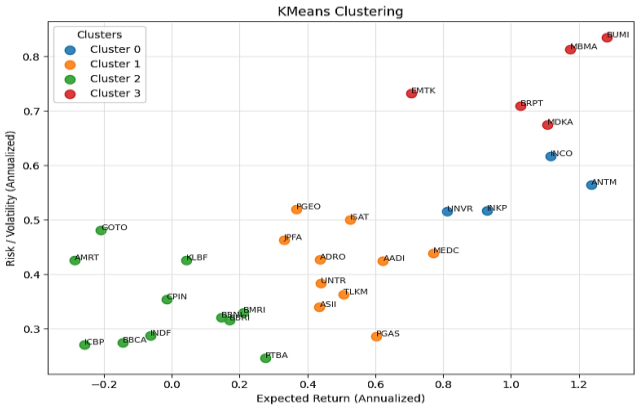
Dalam implementasi algoritma K-Means, tahapan pertama yang dilakukan adalah menentukan jumlah klaster optimal menggunakan *Elbow Method*. Secara komputasional, titik belok (siku) yang paling optimal dapat dicari dengan menghitung jarak maksimum dari setiap titik evaluasi terhadap garis lurus yang ditarik dari titik pertama hingga titik terakhir kurva.



Gambar 2. Grafik Elbow Curve

Mengacu pada Gambar 2, terlihat bahwa penurunan inersia mulai melandai secara signifikan dan membentuk titik belok terjauh pada K=4, yang mengindikasikan jumlah klaster terbaik. Berdasarkan temuan tersebut, pemodelan dilanjutkan dengan menginisialisasi parameter K=4 untuk mempartisi dataset yang

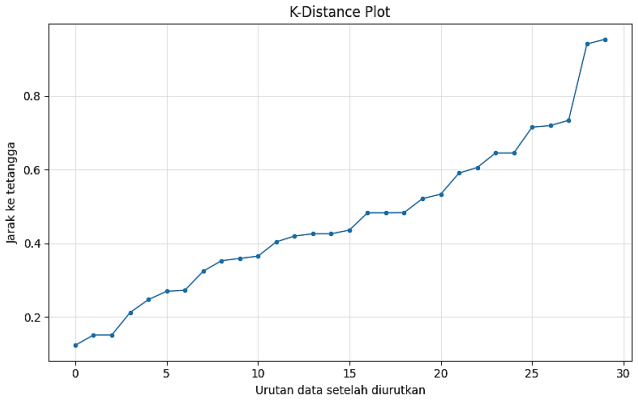
berisikan variabel *Expected Return* dan Tingkat Risiko. Adapun pemetaan visual dari keempat klaster yang terbentuk disajikan pada Gambar 3.



Gambar 3. Visualisasi K-Means

3.4. Pembentukan Model dengan DBSCAN

Berbeda dengan algoritma K-Means, penentuan jumlah klaster pada DBSCAN tidak diinisialisasi di awal, melainkan terbentuk secara otomatis berdasarkan optimasi parameter Epsilon (ε) dan Minimum Samples (min_samples). Dalam penelitian ini, parameter optimal dicari melalui proses iteratif yang menguji rentang nilai ε mulai dari 0.05 hingga 1.05 yang didapatkan berdasarkan observasi titik belok kurva pada grafik K-distance (Gambar 4) serta dikombinasikan dengan variasi nilai min_samples.



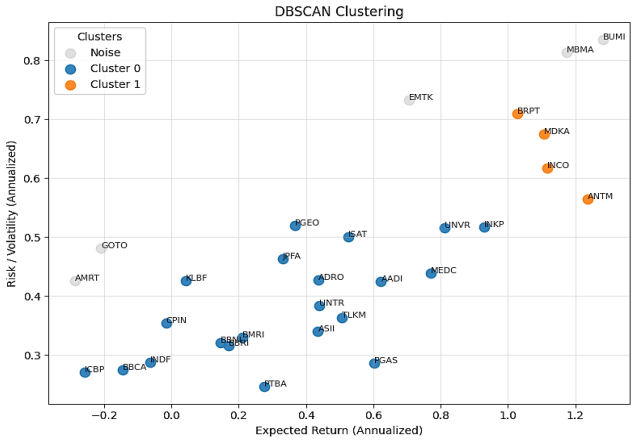
Gambar 4. Grafik K-Distance

Pada setiap iterasi, kinerja model dievaluasi menggunakan metrik *Silhouette Score*, dengan konstrain pembentukan klaster lebih dari satu dan proporsi noise yang terkendali guna menjaga stabilitas serta interpretabilitas struktur klaster yang dihasilkan [22], [23]. Ringkasan hasil komputasi dari proses iterasi tersebut disajikan pada Tabel 3.

Tabel 3. Hasil Iterasi

i	ε	min_samples	silhouette
0	0.75	5	0.539555
1	0.50	3	0.517942
2	0.55	3	0.517942
3	0.60	3	0.517942
4	0.65	4	0.517942
5	0.70	4	0.510124
6	0.65	3	0.505674
7	0.70	3	0.505674
8	1.05	7	0.483435

Berdasarkan hasil pengujian pada Tabel 3, parameter yang menghasilkan *Silhouette Score* tertinggi dan ditetapkan sebagai model terbaik adalah ε = 0.75 dan min_samples = 5. Adapun pemetaan visual dari klaster yang terbentuk menggunakan parameter terbaik tersebut disajikan pada Gambar 5.



Gambar 5. Visualisasi DBSCAN

3.5. Evaluasi

Tahap evaluasi dalam penelitian ini difokuskan pada analisis komparatif terhadap hasil pengelompokan data dari algoritma K-Means dan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN). Penilaian kualitas arsitektur klaster diukur menggunakan tiga parameter pengujian validasi internal, yakni Silhouette Score, Davies-Bouldin Index (DBI), dan Dunn Index. Ketiga metrik ini digunakan untuk menentukan algoritma mana yang mampu mengisolasi pola data secara lebih optimal, memiliki kohesi internal yang kuat, terpisah dengan jarak yang maksimal, serta bebas dari tumpang tindih (*overlapping*) antar klaster.

Evaluasi model dilakukan dengan membandingkan nilai kinerja dari konfigurasi terbaik masing-masing algoritma. Pada metrik Silhouette Score, nilai yang mendekati 1 merepresentasikan bahwa titik data berdekatan dengan klaster yang tepat dan berjarak jauh dari klaster lain. Sebaliknya, pada Davies-Bouldin Index, nilai yang semakin mendekati 0 menunjukkan pembagian klaster yang semakin baik dengan tingkat kesamaan antar-klaster yang rendah. Sementara itu, pada metrik Dunn Index, nilai yang semakin tinggi mengindikasikan bahwa klaster yang terbentuk sangat padat dan terpisah dengan sangat baik (jarak intra-klaster kecil dan jarak antar-klaster besar). Tabel 4 menyajikan hasil komputasi dari ketiga pengujian algoritma tersebut.

Tabel 4. Hasil Evaluasi Model

Model	Silhouette Score	Davies-Bouldin Index	Dunn Index
K-Means	0.391	0.810	0.200
DBSCAN	0.518	0.481	0.245

Tahap evaluasi membuktikan bahwa algoritma DBSCAN memiliki performa yang lebih superior dibandingkan K-Means, yang ditunjukkan oleh perolehan Silhouette Score lebih tinggi (0.518 berbanding 0.391), nilai Davies-Bouldin Index yang lebih rendah (0.481 berbanding 0.810) serta perolehan nilai Dunn Index yang lebih maksimal, yakni 0.245 berbanding 0.200. Hasil tersebut mengindikasikan bahwa DBSCAN mampu membentuk struktur klaster yang lebih kompak dan memiliki separasi antar-klaster yang lebih baik. Temuan ini sejalan dengan penelitian [23] yang menyatakan bahwa metode berbasis densitas seperti DBSCAN lebih efektif dalam menangani data yang mengandung *noise* dan *outlier* dibandingkan metode berbasis c*entroid* seperti K-Means. Tingginya kualitas klasterisasi ini menjadikannya sebagai model final yang dipilih. Setelah kualitas model tervalidasi, tahap akhir dari proses Knowledge Discovery in Databases (KDD) adalah interpretasi matematis menjadi informasi yang bermakna (*knowledge*).

Tabel 5. Statistik Klasterisasi DBSCAN

Klaster	Expected Return			Risk			Jumlah Saham
	Mean	Min	Max	Mean	Min	Max	
-1	0.533309	-0.286910	1.282658	0.657261	0.425761	0.834278	5
0	0.339985	-0.258034	0.928415	0.381247	0.246871	0.518861	21
1	1.122629	1.028191	1.237528	0.641458	0.564354	0.709379	4

Tabel 5 menyajikan statistik hasil klasterisasi menggunakan algoritma DBSCAN berdasarkan parameter *expected return*, tingkat risiko (*risk*), dan jumlah saham pada masing-masing klaster. Secara umum, hasil tersebut menunjukkan adanya perbedaan karakteristik yang cukup jelas antar kelompok saham yang terbentuk. Rata-rata *expected return* seluruh saham berada pada kisaran 0.49 dengan rata-rata risiko sekitar 0.56, sehingga distribusi klaster memperlihatkan variasi profil investasi yang berbeda dari kondisi pasar secara umum.

Berdasarkan hasil pemodelan, ke-30 saham konstituen IDX30 tidak dipetakan secara kaku ke dalam kelompok tertentu, melainkan terbagi ke dalam dua klaster utama dan satu kelompok *noise* yang merepresentasikan saham anomali (saham gorengan).

1. Klaster 0 dikategorikan sebagai kelompok Moderat (*Medium Risk, Stable Return*) karena memiliki jumlah anggota terbanyak, yaitu 21 saham dengan rata-rata *expected return* sebesar 0.339985 dan rata-rata risiko sebesar 0.381247. Nilai risiko tersebut berada di bawah rata-rata risiko keseluruhan saham IDX30, yang menunjukkan bahwa saham dalam kelompok ini memiliki tingkat volatilitas yang relatif lebih rendah dan pergerakan harga yang lebih stabil [24]. Selain itu, rentang risiko pada klaster ini juga cenderung sempit, yaitu berada pada kisaran 0.246871 hingga 0.518861, sehingga mencerminkan tingkat konsistensi yang cukup baik antar saham di dalam kelompok tersebut. Karakteristik ini menunjukkan bahwa saham-saham pada Klaster 0 lebih sesuai bagi investor dengan profil risiko moderat yang mengutamakan kestabilan pertumbuhan investasi jangka menengah hingga panjang [25].
2. Klaster 1 dikategorikan sebagai kelompok Agresif (*High Risk, High Return*) yang terdiri dari 4 saham, yaitu ANTM, INCO, MDKA, dan BRPT. Berdasarkan Tabel 5, klaster ini memiliki rata-rata *expected return* tertinggi sebesar 1.122629 dengan rata-rata risiko sebesar 0.641458, yang berada di atas rata-rata risiko keseluruhan saham IDX30. Selain itu, nilai *expected return* minimum pada kelompok ini tetap berada pada angka tinggi, yaitu 1.028191, yang menunjukkan bahwa seluruh saham dalam klaster tersebut memiliki potensi keuntungan yang relatif besar. Namun, tingginya tingkat pengembalian tersebut diikuti oleh volatilitas yang lebih tinggi dengan rentang risiko mencapai 0.564354 hingga 0.709379. Kondisi ini mengindikasikan bahwa saham dalam kelompok tersebut memiliki fluktuasi harga yang lebih ekstrem dan tingkat ketidakpastian yang lebih besar. Dalam literatur keuangan modern, saham dengan volatilitas tinggi umumnya dikategorikan sebagai instrumen *high risk-high return* yang lebih sesuai bagi investor bertipe *risk-taker* atau *trader* aktif dengan toleransi risiko tinggi [24].
3. Kelompok noise atau Klaster -1 dikategorikan sebagai kelompok Anomali (*Highly Speculative Stocks*) yang terdiri dari 5 saham pencilan, yaitu GOTO, AMRT, EMTK, MBMA, dan BUMI. Berbeda dengan dua klaster utama, saham-saham pada kelompok ini tidak memenuhi kepadatan data yang cukup untuk dimasukkan ke dalam klaster tertentu oleh algoritma DBSCAN. Berdasarkan Tabel 5, kelompok ini memiliki rata-rata risiko tertinggi sebesar 0.657261 dengan rentang *expected return* yang sangat lebar, mulai dari -0.286910 hingga 1.282658. Variasi tersebut menunjukkan bahwa saham pada kelompok *noise* memiliki pola pergerakan yang sangat tidak stabil dan menyimpang dari pola mayoritas saham lainnya. Tingginya volatilitas yang disertai ketidakpastian return mengindikasikan bahwa saham dalam kelompok ini bersifat sangat spekulatif dan memiliki tingkat ketidakpastian yang tinggi, atau yang lazim dikenal sebagai fenomena saham gorengan [2], [26]. Data yang teridentifikasi sebagai *noise* pada DBSCAN umumnya merepresentasikan objek yang memiliki karakteristik berbeda secara signifikan dari pola umum data [23]. Oleh karena itu, investor memerlukan analisis fundamental dan teknikal yang lebih mendalam sebelum mengambil keputusan investasi pada saham-saham dalam kelompok ini.

4. KESIMPULAN

Penelitian ini bertujuan untuk membandingkan dari kedua algoritma yaitu *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) dan K-Means dalam pengelompokan (*clustering*) saham indeks IDX30 berdasarkan tingkat risiko (volatilitas) dan pengembalian yang diharapkan (*expected return*). Berdasarkan evaluasi, hasil pemodelan membuktikan secara empiris bahwa pendekatan *clustering* berbasis densitas (DBSCAN) jauh lebih adaptif dan andal dalam memetakan karakteristik investasi dibandingkan pendekatan berbasis *centroid* (K-Means). Kemampuan DBSCAN dalam mendeteksi dan mengisolasi saham-saham anomali (saham gorengan) secara objektif sebagai kelompok *noise* menjadi karakteristik dalam mencegah terjadinya distorsi atau bias pada struktur klaster utama, sehingga menghasilkan pembagian kelompok investasi (moderat dan agresif) yang memiliki kohesi internal kuat tanpa adanya tumpang tindih (*overlapping*). Secara praktis, hasil pemetaan profil risiko yang akurat ini memberikan landasan strategis yang relevan bagi investor maupun manajer investasi dalam mengoptimalkan keputusan alokasi aset dan mitigasi risiko di tengah volatilitas pasar finansial yang tinggi.

Meskipun penelitian memberikan hasil pemetaan yang komparatif dan optimal, penelitian ini memiliki beberapa keterbatasan metodologis yang perlu diperhatikan seperti periode yang relatif pendek (terbatas pada 1 tahun penuh) belum mampu menangkap atau merepresentasikan perilaku dan tren karakteristik emiten dalam kondisi ekonomi jangka panjang. Fitur untuk proses segmentasi hanya bertumpu pada dua variabel keuangan berbasis harga penutupan harian, yaitu *expected return* dan volatilitas risiko (*risk*) yang disetahunkan, tanpa mengintegrasikan variabel eksternal maupun indikator performa fundamental internal perusahaan. Selain itu, pengimplementasian algoritma DBSCAN memiliki ketergantungan yang tinggi terhadap penentuan parameter Epsilon dan min_samples yang harus dicari secara iteratif, serta memiliki batasan algoritmik.

Beberapa saran untuk kelanjutan penelitian ini yaitu, untuk penelitian selanjutnya, disarankan untuk mengeksplorasi penggunaan algoritma *clustering* lainnya seperti *Ordering Points To Identify the Clustering Structure* (OPTICS) atau *Gaussian Mixture Models* (GMM) untuk segmentasi profil saham IDX30. Selain itu,

penelitian dapat diperluas dengan menggabungkan data tambahan seperti indikator fundamental perusahaan (misalnya rasio PER, PBV), sentimen pasar, dan volume perdagangan harian untuk mendapatkan pemahaman yang lebih komprehensif tentang perilaku dan nilai intrinsik saham. Menggunakan data dengan periode waktu yang lebih panjang (misalnya 3 hingga 5 tahun) juga dapat memberikan wawasan tentang tren dan perubahan karakteristik saham menghadapi berbagai siklus ekonomi dari waktu ke waktu. Terakhir, implementasi metode *machine learning* yang lebih canggih, seperti teknik *deep learning* untuk prediksi deret waktu (*time series*), dapat dieksplorasi untuk meningkatkan akurasi dan efektivitas dalam pembentukan portofolio investasi.

REFERENSI

[1] F. Wulandari, "Analisis Volatilitas Indeks Harga Saham Dan Indikator Ekonomi Makro Di Indonesia Dalam Menghadapi Kondisi Vuca," *Bisnis dan Pendidikan*, vol. 5, no. 5, p. 2025, doi: 10.17977/um066.v5.i5.2025.1.

[2] N. Syahla and M. I. Harahap, "Studi Literatur : Saham Gorengan," *Media Ekonomi*, vol. 22, no. 2, pp. 75–84, Jan. 2022, doi: 10.30595/MEDEK.V22I2.14086.

[3] F. Seru and B. F. Kuddi, "Pendekatan Optimalisasi Portofolio dengan Capital Asset Pricing Model dan Model Markowitz sebagai Strategi Investasi Cerdas bagi Investor Milenial," *Journal of Mathematics: Theory and Applications*, vol. 6, no. 2, pp. 214–224, Oct. 2024, doi: 10.31605/jomta.v6i2.4117.

[4] A. D. C. M. de Oliveira, P. F. A. Pinto, and S. Colcher, "Stocks Clustering Based on Textual Embeddings for Price Forecasting," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12320 LNAI, pp. 665–678, 2020, doi: 10.1007/978-3-030-61380-8_45.

[5] M. Steinbacher, M. Steinbacher, and M. Steinbacher, "An Identification and Estimation of Stock Price Pattern Equations using K-Means," *Comput. Econ.*, Feb. 2025, doi: 10.1007/S10614-025-10879-3.

[6] P. B. Martin and S. S. Suneesh, "Data-Driven Intelligence in Pair Trading: Exploring Advanced Pair Selection Techniques for Long–Short Portfolio," *Lecture Notes on Multidisciplinary Industrial Engineering*, vol. Part F1254, pp. 331–345, 2026, doi: 10.1007/978-981-95-0853-2_31.

[7] H. Dastkhan, A. Mosaffa, and F. J. Fabozzi, "Integrated machine learning framework for dynamic portfolio management: clustering, forecasting, and optimization," *Ann. Oper. Res.*, 2026, doi: 10.1007/S10479-026-07053-9.

[8] J. Chevallier, "Particle Swarm Optimization with Stretching and Clustering for Asset Allocation," *International Journal of Financial Studies*, vol. 14, no. 2, Feb. 2026, doi: 10.3390/IJFS14020038.

[9] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan Dbscan," 2024.

[10] A. Ridwan, … S. S.-I. J. of, and undefined 2021, "IDX30 stocks clustering with K-means algorithm based on expected return and value at risk," *v2.rescollacomm.comAF Ridwan, S Subiyanto, S SupianInternational Journal of Quantitative Research and Modeling, 2021•v2.rescollacomm.com*, vol. 2, no. 4, pp. 201–208, 2021, Accessed: May 21, 2026. [Online]. Available: https://v2.rescollacomm.com/index.php/ijqrm/article/view/157

[11] A. Anfossi and D. Chicco, "An easy guide to the Davies-Bouldin index for unsupervised internal clustering evaluation," *Discover Computing*, vol. 29, no. 1, Dec. 2026, doi: 10.1007/S10791-026-10094-0.

[12] U. Kazemi and S. Soleimani, "A new approach data processing: *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) clustering using game-theory: U. Kazemi, S. Soleimani," *SpringerU Kazemi, S SoleimaniSoft Computing, 2025•Springer*, vol. 29, no. 3, pp. 1331–1346, Feb. 2025, doi: 10.1007/S00500-025-10405-5.

[13] C. Llatas, B. Soust-Verdaguer, L. C. Torres, and D. Cagigas, "Application of Knowledge Discovery in Databases (KDD) to environmental, economic, and social indicators used in BIM workflow to support sustainable design," *Journal of Building Engineering*, vol. 91, Aug. 2024, doi: 10.1016/j.jobe.2024.109546.

[14] R. Swastika, S. Mukodimah, F. Susanto, and M. Muslihudin, "Implementasi Data Mining (Clastering, Association, Prediction, Estimation, Classification)," 2023, Accessed: May 21, 2026.

[15] A. K. Hidayah and R. E. Putra, "Penerapan Metode Long Short Term Memory untuk Memprediksi Harga Beras di Indonesia," *Journal of Informatics and Computer Science*, vol. 06, 2024.

[16] C. Wongoutong, "The impact of neglecting feature scaling in k-means clustering," *PLoS One*, vol. 19, no. 12, Dec. 2024, doi: 10.1371/journal.pone.0310839.

[17] D. A. Setiady and H. Leong, "Implementation Of K-Means Algorithm Elbow Method And Silhouette Coefficient For Rainfall Classification," 2020.

[18] I. Chintia Sarumaha *et al.*, "of Superstore Data using K-Means and K-Medoids for Product Delivery Insights," 2025.

[19] F. B. Wijaya, W. Budiaji, and A. S. Wicaksono, "Applied Machine Learning DBSCAN for Identifying Clusters of Micro and Small Industries," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 2, pp. 380–386, May 2025, doi: 10.31004/riggs.v4i2.515.

[20] Z. Teng, J. Yan, D. Liu, and P. Zhang, "When Does the Silhouette Score Work? A Comprehensive Study in Network Clustering," Dec. 2025, [Online]. Available: http://arxiv.org/abs/2512.24841

[21] P. Hasugian, … D. M.-M. J., and undefined 2025, "Comparative Evaluation of Data Clustering Accuracy through Integration of Dimensionality Reduction and Distance Metric," *journal.universitasbumigora.ac.idPM Hasugian, D Mathelinea, S Simamora, PBN SimangunsongMATRIK: Jurnal Manajemen, Teknik Informatika dan, 2025•journal.universitasbumigora.ac.id*, Accessed: May 23, 2026. [Online]. Available: https://journal.universitasbumigora.ac.id/matrik/article/view/5057

[22] A. Yudono, A. R. R. T. Hidayat, W. P. Wijayanti, F. Afrianto, and A. D. Fitrianto, "Spatial Clustering of The Village Development Index Using K-Means and DBSCAN Approaches in Machine Learning: A Case Study in Malang Regency, Indonesia," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1595, no. 1, p. 012031, Mar. 2026, doi: 10.1088/1755-1315/1595/1/012031.

[23] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems*, vol. 42, no. 3, Jul. 2017, doi: 10.1145/3068335.

[24] P. Sakowski, R. Sieradzki, and R. Ślepaczuk, "Systemic risk indicator based on implied and realized volatility," *Quality & Quantity 2025 60:2*, vol. 60, no. 2, pp. 6389–6427, Dec. 2025, doi: 10.1007/S11135-025-02515-X.

[25] R. M. Leushuis and N. Petkov, "Advances in forecasting realized volatility: a review of methodologies," *Financial Innovation 2026 12:1*, vol. 12, no. 1, pp. 14-, Jan. 2026, doi: 10.1186/S40854-025-00809-5.

[26] B. M. Barber, X. Huang, T. Odean, and C. Schwarz, "Attention-induced trading and returns: Evidence from Robinhood users," *Wiley Online LibraryBM Barber, X Huang, T Odean, C SchwarzThe Journal of Finance, 2022•Wiley Online Library*, vol. 77, no. 6, pp. 3141–3190, Dec. 2022, doi: 10.1111/JOFI.13183.