Tersedia Online : http://e-journals.unmul.ac.id/

ADOPSI TEKNOLOGI DAN SISTEM INFORMASI (ATASI)

Alamat Jurnal : http://e-journals2.unmul.ac.id/index.php/atasi/index

Penerapan *K-Means Clustering* dalam Analisis URL *Phishing* untuk Identifikasi Risiko Keamanan Menggunakan Model PCA

Tua Delima Sitompul [1)*], Davina Putri Ananta [2)], Muhammad Rafif Hanif [3)], Masna Wati [4)], Haviluddin [5)]

[1,2,3,4,5)] Jurusan Informatika, Fakultas Teknik, Universitas Mulawarman, Samarinda

E-Mail : delimasitompul514@gmail.com [1)]; davinaputriananta1912@gmail.com [2)]; rafifdragneel@gmail.com [3)]; masnawati@fkti.unmul.ac.id [4)]; haviluddin@unmul.ac.id [5)]

ARTICLE INFO

Article history:
Received : May 22, 2025
Revised : June 22, 2025
Accepted : June 26, 2025
Available online :
November 30, 2025

Keywords:
Clustering
K-Means
Machine Learning
PCA
Phising

ABSTRACT

Phishing is an evolving cyber threat, and blacklist-based detection methods have significant limitations in identifying new Phishing sites. This study implements K-Means Clustering to group Phishing URLs based on their characteristics, using the PhiUSIIL Phishing URL Dataset with 235,795 samples. Through comprehensive data preprocessing, analysis of optimal cluster numbers using Silhouette Score yielded k = 2 with a score of 0.972 for the hybrid approach utilizing URLLength and IsDomainIP features. Visualization results through PCA and t-SNE show very clear cluster separation, confirming that a simple combination of two features can effectively distinguish Phishing URLs from normal URLs. This research demonstrates that K-Means Clustering offers a more adaptive solution compared to blacklist-based methods for Phishing detection, with the ability to recognize new attack patterns without requiring labeled data.

Kata Kunci :
Clustering
K-Means
Machine Learning
PCA
Phising

ABSTRAK

Phishing merupakan ancaman siber yang terus berkembang, dan metode deteksi berbasis daftar hitam memiliki keterbatasan signifikan dalam mengidentifikasi situs *Phishing* baru. Penelitian ini menerapkan *K-Means Clustering* untuk mengelompokkan URL *Phishing* berdasarkan karakteristiknya, menggunakan dataset PhiUSIIL *Phishing* URL dengan 235.795 sampel. Melalui preprocessing data yang komprehensif, analisis jumlah klaster optimal menggunakan *Silhouette Score* menghasilkan k = 2 dengan skor 0,972 pada pendekatan hibrid yang menggunakan fitur URLLength dan IsDomainIP. Hasil visualisasi melalui PCA dan t-SNE menunjukkan pemisahan klaster yang sangat jelas, mengonfirmasi bahwa kombinasi sederhana dari dua fitur dapat secara efektif membedakan URL *Phishing* dari URL normal. Penelitian ini membuktikan bahwa *K-Means Clustering* menawarkan solusi yang lebih adaptif dibandingkan metode berbasis daftar hitam dalam deteksi *Phishing*, dengan kemampuan mengenali pola serangan baru tanpa memerlukan data berlabel.

1. PENDAHULUAN

Phishing adalah jenis serangan siber yang bertujuan untuk memperoleh informasi pribadi, seperti kredensial login, data kartu kredit, dan informasi sensitif lainnya, dengan cara menipu korban agar mengunjungi situs web palsu yang menyerupai situs resmi. Berdasarkan laporan Kaspersky, lebih dari 36 juta ancaman siber berhasil terdeteksi di Indonesia pada tahun 2024, dengan sebagian besar serangan berasal dari *Phishing* (Kaspersky, 2024). Teknik yang digunakan dalam serangan *Phishing* terus berkembang, sehingga tantangan utama dalam mendeteksi situs *Phishing* adalah mengenali situs yang baru muncul. Metode berbasis daftar hitam (*blacklist*) memiliki keterbatasan dalam hal ini, karena hanya dapat mendeteksi situs *Phishing* yang sudah terdaftar sebelumnya (Fatiha, Setiawan, Ikhsan, & Yunita, 2024).

K-Means Clustering menawarkan fleksibilitas yang lebih tinggi dalam mendeteksi situs *Phishing* baru yang belum terdaftar dalam sistem. Beberapa penelitian juga telah menunjukkan penerapan teknik dimensionality reduction untuk meningkatkan kinerja *K-Means Clustering*, salah satunya adalah penggunaan *Principal Component*

*) Corresponding Author
https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.

Analysis (PCA) untuk mereduksi dimensi data dan memudahkan analisis lebih lanjut (Ady Saputro, Sugiarto, Surya Nugraha, Studi Informatika, & Amikom, 2024). Selain itu, *Silhouette Score* digunakan untuk mengevaluasi kualitas klaster dan membantu memilih jumlah klaster yang optimal (Dewi et al. 2023).

Penelitian ini bertujuan untuk mengimplementasikan *K-Means Clustering* dalam mendeteksi URL *Phishing* menggunakan dataset PhiUSIIL *Phishing* URL Dataset, yang berisi lebih dari 235.795 sampel. Dengan memanfaatkan PCA untuk reduksi dimensi dan *Silhouette Score* untuk menentukan jumlah klaster yang optimal, penelitian ini diharapkan dapat mengembangkan solusi yang lebih efisien dan adaptif dalam mendeteksi situs *Phishing* baru. Hasil dari penelitian ini diharapkan dapat memperkuat sistem deteksi *Phishing* berbasis URL dan memberikan alternatif yang lebih fleksibel dibandingkan dengan metode berbasis daftar hitam.

2. TINJAUAN PUSAKA

A. Machine Learning untuk Deteksi *Phishing*

Machine learning telah menjadi pendekatan yang efektif dalam deteksi *Phishing*. Beberapa penelitian sebelumnya banyak mengaplikasikan metode *supervised learning* seperti Naïve Bayes, Random Forest, SVM, dan *LightGBM* untuk mendeteksi URL *Phishing* dengan akurasi yang tinggi (Tampinongkol, Kamila, Wardhana, Kusuma, & Revaldo, 2024)(Muriithi & Karani, 2024)(Gusthvi, Roza, & Allo, 2023). Penelitian oleh Foozy et al. menunjukkan bahwa *LightGBM* memberikan akurasi tertinggi sebesar 95%, diikuti oleh Random Forest dan Naïve Bayes dalam mendeteksi URL *Phishing*. Pendekatan *supervised learning* meskipun efektif, membutuhkan data berlabel, yang sering kali sulit diperoleh, terutama untuk situs *Phishing* yang baru. Oleh karena itu, pendekatan lain berbasis un*supervised learning* seperti *K-Means Clustering* mulai banyak diterapkan. Metode ini tidak memerlukan data berlabel dan dapat mengelompokkan URL berdasarkan karakteristik tertentu, seperti panjang URL, jumlah simbol khusus, dan penggunaan alamat IP sebagai domain, yang sering ditemukan pada situs *Phishing* (Foozy, Anuar, Maslan, Adam, & Mahdin, 2024; Rahmah, 2024; Wijaya & Subandi, 2024).

B. *Principal Component Analysis (PCA)*

Principal Component Analysis (PCA) adalah teknik reduksi dimensi yang banyak digunakan dalam preprocessing data untuk analisis clustering. PCA bekerja dengan cara mengidentifikasi komponen-komponen utama (principal components) dari data yang memiliki varians terbesar, yang memungkinkan representasi data yang lebih efisien dalam dimensi yang lebih rendah sambil tetap mempertahankan informasi yang signifikan (Saputro, Sugiarto, & Nugraha, 2024).

Dalam konteks deteksi *Phishing*, PCA membantu mengatasi masalah dimensionalitas tinggi dengan mengekstraksi fitur-fitur penting dari URL yang dapat digunakan untuk membedakan URL *Phishing* dan legitimate. Secara matematis, PCA mencari transformasi linier yang memaksimalkan varians data yang diproyeksikan, dengan persamaan:

$$X' = XW \qquad \text{.....(1)}$$

Dimana:
X adalah matriks data asli
W adalah matriks transformasi (berisi eigenvector dari matriks kovarians data)
X' adalah matriks data hasil reduksi dimensi.

C. T-Distributed Stochastic Neighbor Embedding (t-SNE)

T-Distributed Stochastic Neighbor Embedding (t-SNE) adalah teknik reduksi dimensi nonlinier yang fokus pada preservasi struktur lokal data. Berbeda dengan PCA yang mempertahankan varians global, t-SNE mempertahankan hubungan ketetanggaan antara titik-titik data, sehingga sangat cocok untuk visualisasi data berdimensi tinggi (Wang et al. 202(Ghojogh et al., 2022; Wang et al., 2021).

t-SNE bekerja dengan dua langkah utama: pertama, menghitung probabilitas kemiripan antara pasangan titik data dalam ruang dimensi tinggi; kedua, mencari representasi titik-titik dalam ruang dimensi rendah sehingga probabilitas kemiripan dipertahankan semaksimal mungkin. t-SNE menggunakan distribusi *t-Student* dalam ruang dimensi rendah untuk mengatasi masalah "*crowding problem*" yang sering terjadi pada teknik reduksi dimensi lainnya. Dalam konteks deteksi *Phishing*, t-SNE membantu memvisualisasikan bagaimana URL *Phishing* dan legitimate terkelompok dalam ruang fitur, yang memungkinkan analisis visual terhadap pola-pola *Phishing*.

D. *K-Means Clustering*

K-Means Clustering adalah algoritma un*supervised learning* yang populer untuk mengelompokkan data menjadi k klaster berdasarkan kemiripan fitur. Algoritma ini bekerja dengan cara iteratif, dimulai dengan menginisialisasi k centroid secara acak, kemudian mengelompokkan setiap titik data ke centroid terdekat, dan memperbarui posisi centroid berdasarkan rata-rata titik data dalam kelompok tersebut.

Proses iteratif K-Means bertujuan untuk meminimalkan *within-cluster sum of squares (WCSS),* yang merupakan jumlah kuadrat jarak antara setiap titik data dan centroid klasternya. WCSS dihitung dengan persamaan:

https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu i||^2 \quad \text{.....(2)}$$

Dimana:
k adalah jumlah klaster,
C_i adalah himpunan data dalam klaster ke-i,
μ_i adalah centroid dari klaster ke-$_i$,
dan x adalah data point pada klaster tersebut.

Dalam konteks deteksi *Phishing*, *K-Means Clustering* dapat digunakan untuk mengelompokkan URL berdasarkan karakteristik strukturalnya, seperti panjang URL, jumlah simbol khusus, dan penggunaan alamat IP sebagai domain. Keuntungan utama dari pendekatan ini adalah kemampuannya untuk mendeteksi pola *Phishing* baru tanpa memerlukan data berlabel, yang menjadikannya solusi yang lebih adaptif dibandingkan dengan metode berbasis daftar hitam.

E. *Silhouette Score*

Silhouette Score adalah metrik evaluasi yang digunakan untuk menilai kualitas klaster yang terbentuk. Nilai *Silhouette Score* berkisar antara -1 hingga 1, dimana nilai yang mendekati 1 menunjukkan pemisahan klaster yang sangat baik, nilai mendekati 0 menunjukkan klaster yang saling tumpang tindih, dan nilai negatif mengindikasikan kesalahan dalam klasterisasi. *Silhouette Score* untuk setiap titik data dihitung dengan persamaan:

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

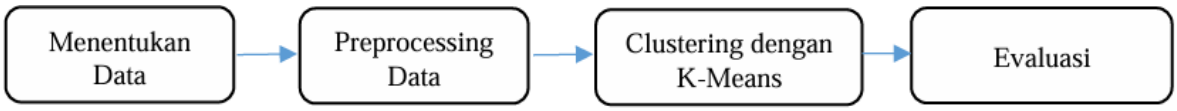
Dimana:
$a(i)$: rata-rata jarak antara data i dan semua titik lain dalam klaster yang sama,
$b(i)$: rata-rata jarak antara data i dan semua titik dalam klaster terdekat berikutnya.

3. METODE PENELITIAN

A. Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif untuk menguji efektivitas *K-Means Clustering* dalam mendeteksi situs *Phishing* berbasis URL. Proses analisis dilakukan dengan menggunakan Google Colab, yang menguji dataset PhiUSIIL *Phishing* URL Dataset yang terdiri dari lebih dari 235,795 sampel.

B. Tahapan Penelitian



Gambar 1. Tahapan Penelitian

Bagan 1 menunjukkan alur tahapan penelitian yang dilakukan dalam studi ini. Tahapan dimulai dengan menentukan data yang akan digunakan, PhiUSIIL *Phishing* URL Dataset. Selanjutnya, data tersebut melalui tahap preprocessing untuk memastikan kualitas dan kesiapannya, termasuk proses normalisasi, pemilihan fitur, dan reduksi dimensi. Setelah data siap, dilakukan penentuan jumlah klaster menggunakan metode evaluasi seperti *Silhouette Score*, yang kemudian dilanjutkan dengan proses clustering menggunakan algoritma K-Means. Hasil klasterisasi tersebut kemudian dianalisis untuk mengidentifikasi pola dalam data, dan pada akhirnya digunakan sebagai dasar dalam menyusun kesimpulan dan saran yang relevan terhadap tujuan penelitian.

C. Dataset dan Sumber Data

Dataset yang digunakan dalam penelitian ini adalah PhiUSIIL *Phishing* URL Dataset, yang diperoleh dari Kaggle. Dataset ini mencakup 56 fitur, dengan empat fitur yang dipilih karena relevansinya dalam mendeteksi *Phishing*. Fitur-fitur yang digunakan adalah URLLength (panjang URL dalam karakter), NoOfOtherSpecialCharsInURL (jumlah simbol khusus dalam URL), NoOfQMarkInURL (jumlah tanda tanya (?) dalam URL), dan IsDomainIP (indikator apakah URL menggunakan alamat IP sebagai domain). Fitur-fitur tersebut memiliki kontribusi signifikan dalam membedakan URL *Phishing* dan URL sah.

D. Proses Processing

Sebelum dilakukan clustering, dilakukan beberapa tahap preprocessing untuk memastikan data yang digunakan siap untuk analisis. Berikut adalah tahap preprocessing:

*) Corresponding Author
https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.

1. Handling Missing Values

Dataset dicek untuk mengetahui keberadaan missing values pada setiap fitur. Data yang memiliki nilai kosong pada fitur yang dipilih untuk clustering akan dihapus untuk menghindari distorsi dalam analisis. Tahap preprocessing mencakup normalisasi fitur menggunakan StandardScaler, teknik yang direkomendasikan untuk dataset tidak seimbang (Windarni et al., 2023).

2. Pemilihan Fitur Untuk Clustering

Dataset asli memiliki 56 fitur, namun tidak semua fitur relevan untuk analisis *Phishing* berbasis URL. Oleh karena itu, dilakukan seleksi fitur dengan mempertimbangkan korelasi terhadap identifikasi *Phishing*. Fitur seperti jumlah simbol khusus dan penggunaan IP terbukti signifikan dalam mengidentifikasi *Phishing*. Dari 56 fitur yang tersedia, hanya 4 fitur utama yang dipilih untuk proses clustering karena dianggap paling berkontribusi dalam membedakan URL *Phishing* dan legitimate:

1. URLLength – Panjang URL dalam karakter, sering kali URL *Phishing* memiliki panjang yang tidak biasa.
2. NoOfOtherSpecialCharsInURL – Jumlah simbol khusus dalam URL, yang sering digunakan dalam teknik manipulasi URL.
3. NoOfQMarkInURL – Jumlah tanda tanya (?) dalam URL, yang dapat digunakan dalam teknik *query manipulation*.
4. IsDomainIP – Indikator apakah URL menggunakan alamat IP sebagai domain, yang sering digunakan dalam *Phishing*.

3. Dimensionality Reduction

Sebelum dilakukan clustering, digunakan teknik PCA untuk mereduksi dimensi data. PCA digunakan untuk mengekstrak informasi penting dari fitur yang tersedia dan meningkatkan efisiensi komputasi. Selanjutnya, t-SNE digunakan untuk memvisualisasikan data dalam ruang dua dimensi sehingga pola clustering dapat lebih mudah diinterpretasikan.

4. Sampling Data

Untuk efisiensi komputasi, dilakukan random sampling sebanyak 235.795 sampel dari total dataset. Sampling ini bertujuan untuk mengurangi waktu pemrosesan tanpa mengurangi representasi pola data secara signifikan.

5. Normalisasi Data

Agar semua fitur memiliki skala yang sebanding, dilakukan normalisasi menggunakan StandardScaler. Teknik ini memastikan bahwa setiap fitur memiliki distribusi dengan rata-rata nol dan varians satu, sehingga hasil clustering lebih optimal.

E. Klasterisasi

Setelah data selesai diproses melalui *tahap preprocessing*, langkah selanjutnya adalah melakukan klasterisasi menggunakan *K-Means Clustering*.

1. Konsep *K-Means Clustering*

K-Means Clustering adalah metode un*supervised learning* yang digunakan untuk mengelompokkan data berdasarkan kemiripan fitur. Algoritma ini bekerja dengan cara menginisialisasi k centroid secara acak, menghitung jarak setiap titik data ke centroid terdekat, dan mengelompokkan data berdasarkan kedekatannya. Setelah setiap klaster terbentuk, posisi centroid diperbarui berdasarkan rata-rata titik data dalam klaster, dan proses ini diulang hingga posisi centroid stabil atau iterasi maksimum tercapai (Pribadi & Sulianta, 2024; Saputra & Nataliani, 2021).

Proses clustering mengoptimalkan pengelompokan dengan meminimalkan WCSS (Persamaan 4). Hasil analisis enam skenario preprocessing (Gambar 2) menunjukkan bahwa F2 (PCA Penuh) mencapai *Silhouette Score* tertinggi (0.345, Persamaan 4), mengindikasikan pemisahan klaster yang jelas. Nilai ini konsisten bahwa PCA meningkatkan separabilitas data URL. Hasil proses clustering menggunakan algoritma K-Means bertujuan meminimalkan within-cluster sum of squares (WCSS), yang dihitung dengan rumus sebagai berikut:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu i||^2 \quad \text{.....(4)}$$

Keterangan:
k adalah jumlah klaster,
C_i adalah himpunan data dalam klaster ke-i,
μ_i adalah centroid dari klaster ke-$_i$,

https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.

dan x adalah data point pada klaster tersebut.

2. Pemilihan Jumlah Klaster (k) dengan *Silhouette Score*

Untuk menentukan jumlah klaster yang optimal, digunakan *Silhouette Score*, yang mengukur kualitas pemisahan antara klaster-klaster yang terbentuk. Nilai *Silhouette Score* berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan pemisahan klaster yang sangat baik, nilai mendekati 0 menunjukkan klaster yang saling tumpang tindih, dan nilai negatif mengindikasikan kesalahan dalam klasterisasi.

Proses dimulai dengan menjalankan K-Means untuk berbagai nilai k, diikuti dengan perhitungan *Silhouette Score* untuk setiap nilai k. Nilai k yang menghasilkan *Silhouette Score* tertinggi dipilih sebagai jumlah klaster optimal, yang menandakan pemisahan klaster terbaik (Fatiha et al., 2024; Guntara & Lutfi, 2023; Mulyani, Setiawan, & Fathi, 2023). Berdasarkan hasil perhitungan *Silhouette Score* pada berbagai nilai *k*, diperoleh bahwa nilai k = 4 memberikan skor tertinggi, sehingga dipilih sebagai jumlah klaster yang optimal untuk proses klasterisasi pada penelitian ini. Untuk menentukan jumlah klaster yang optimal, digunakan metode *Silhouette Score*, yang mengukur seberapa baik data telah terkelompokkan. Nilai ini berkisar antara -1 hingga 1. Rumus yang digunakan untuk menghitung nilai Silhouette adalah:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad \text{.....(5)}$$

Keterangan:
$a(i)$: rata-rata jarak antara data iii dan semua titik lain dalam klaster yang sama,
$b(i)$: rata-rata jarak antara data iii dan semua titik dalam klaster terdekat berikutnya.

4. HASIL DAN PEMBAHASAN

Perbandingan kuantitatif 4 pendekatan berbeda dapat dilihat melalui Tabel 1. Data ini memberikan gambaran tentang efektivitas masing-masing metode. Tabel ini menjadi landasan awal dengan menyajikan perbandingan objektif antar metode.

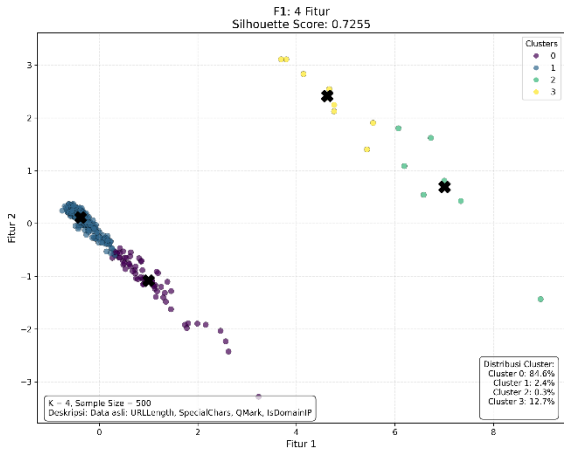
Tabel 1. Perbandingan Pendekatan Clustering dengan k=4

Pendekatan	Deskripsi	*Silhouette Score*	Distribusi Klaster
F1: 4 Fitur	URLLength, SpecialChars, QMark, IsDomainIP	0,7255	8.458, 239, 29, 1.274
F2: PCA Penuh	Reduksi 4 fitur ke 2 komponen PCA	0,7890	8.836, 99, 29, 1.036
F3: Hibrid 1	URLLength + SpecialChars (PCA) + QMark + IsDomainIP	0,7255	8.458, 239, 29, 1.274

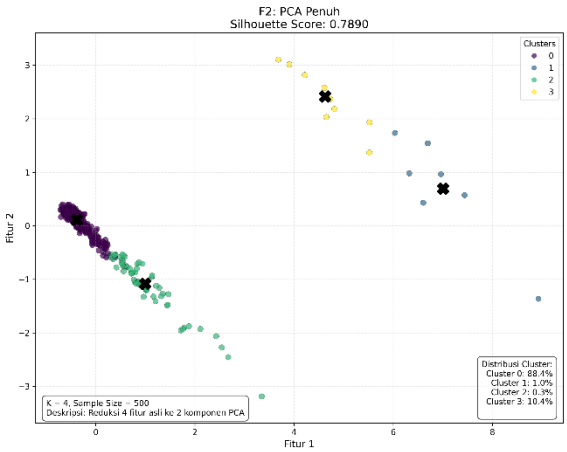
Pada Tabel 1. Mengungkapkan bahwa pendekaan F4 (Hibrid 2) dengan hanya 2 kunci fitur kunci, yaitu URLLength dan IsDomainIP yang mencapai *Silhouette Score* tertinggi (0.8398), mengungguli metode lain yang menggunakan lebih banyak fitur atau teknik reduksi dimensi. Distribusi klasternya pun paling jelas: satu kelompok besar URL normal dan beberapa kelompok kecil *Phishing*. Sementara itu, F1 (4 fitur) dan F3 (hibrid) menunjukkan skor lebih rendah (0.7255) dengan distribusi kurang seimbang, membuktikan bahwa kompleksitas tambahan justru tidak selalu meningkatkan performa.

Hasil kuantitaif dari tabel tersebut, dapat dihasilkan visualisasi yang lebih nyata pada Gambar 2 a hingga d. Gambar tersebut menampilkan secara langsung bagaimana perbedaan skor Silhouette tercermin dalam pola pengelompokan URL, sekaligus memahami alasan mengapa model sederhana seperti F4 justru paling efektif.

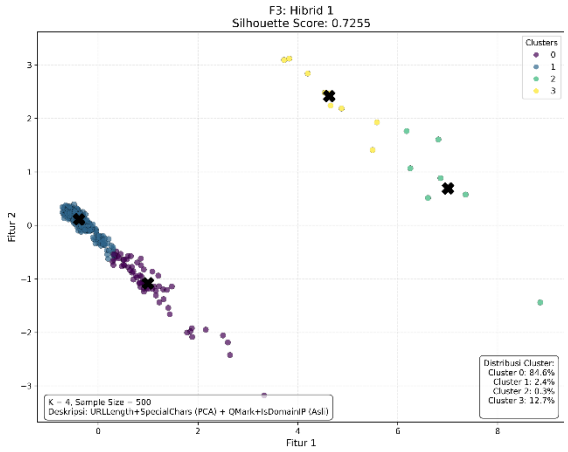
*) Corresponding Author
https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.



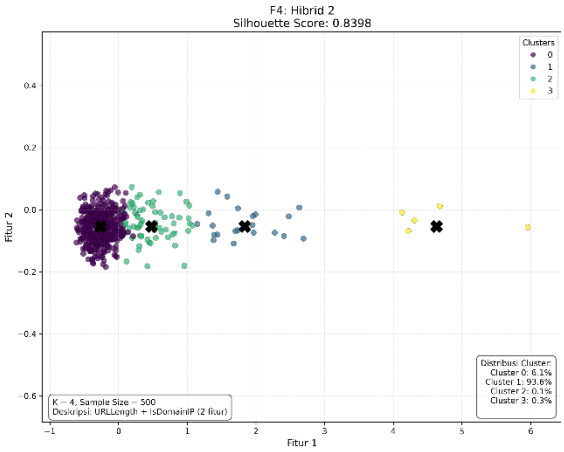
a). F1: 4 Fitur



b). F2: PCA Penuh



c). F3 Hibrid 1



d). Hibrid 2

Gambar 2. Visualisasi hasil Clustering k=4

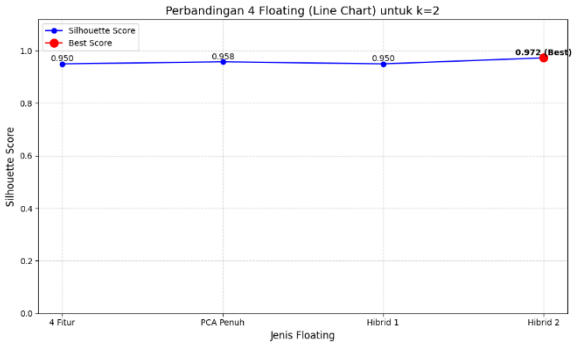
Visualisasi clustering dengan k=4 menunjukkan perbedaan mencolok antar model. Gambar a). F1: 4 Fitur, dan Gambar c). F3: Hibrid 1, memperlihatkan pola yang hampir sama. Keduanya menghasilkan *Silhouette Score* 0.7255 dengan klaster yang tumpang tindih di tengah plot. Fakta bahwa F3 menggunakan PCA namun hasilnya tetap sama menunjukkan bahwa penambahan PCA tidak memberikan kontribusi nyata dalam membedakan URL *Phishing* dari yang normal. Gambar b) F2: PCA Penuh, mencatat peningkatan *Silhouette Score* menjadi 0.7890. Titik-titik data lebih terkonsentrasi, meskipun outlier masih muncul. Ini membuktikan bahwa reduksi dimensi bisa membantu, tapi belum sepenuhnya efektif. Menariknya, struktur dasar pola tetap menyerupai F1 dan F3, mengindikasikan bahwa transformasi PCA masih membawa sifat asli data.

Perubahan signifikan terjadi di Gambar d). F4: Hibrid 2, dengan *Silhouette Score* tertinggi: 0.8398. Klaster *Phishing* tampak jelas terpisah di sisi kanan, mengarah pada fitur-fitur seperti panjang URL dan penggunaan IP sebagai domain. URL normal terkelompok rapi di sisi lain. Ini bukan sekadar peningkatan kecil, tapi perbedaan kualitas yang signifikan. F4 juga mengungkap fakta penting: pelaku *Phishing* cenderung memakai URL panjang dan IP langsung, sedangkan situs sah memakai domain pendek yang mudah diingat. Pendekatan ini membuktikan bahwa solusi efektif sering kali muncul dari eksplorasi fitur sederhana, bukan dari kompleksitas algoritma.

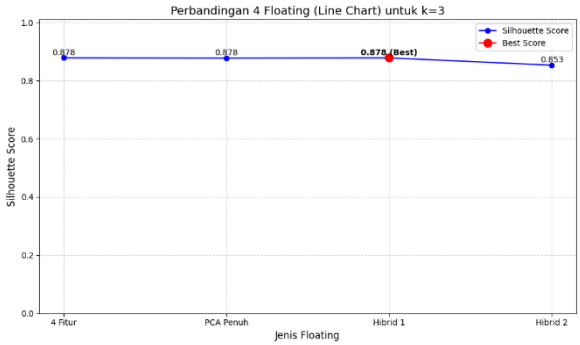
5. EVALUASI

Evaluasi dalam penelitian ini bertujuan untuk mengukur efektivitas masing-masing pendekatan dalam membentuk klaster yang optimal untuk membedakan URL *Phishing* dari URL normal. Fokus evaluasi difokuskan pada variasi jumlah klaster (*k*) dari 2 hingga 9, dengan metrik utama berupa *Silhouette Score*, yang dikombinasikan dengan analisis visual hasil klasterisasi.

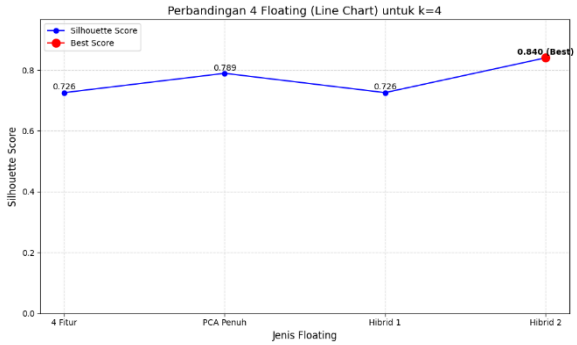
https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.



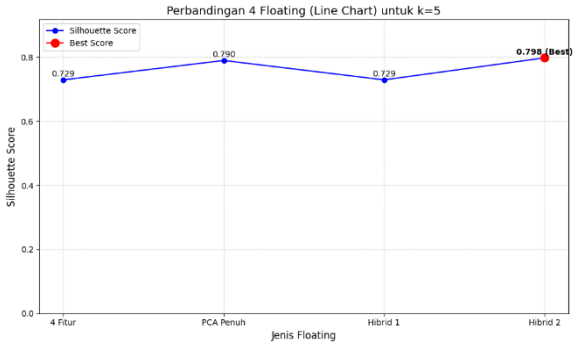
a). K = 2



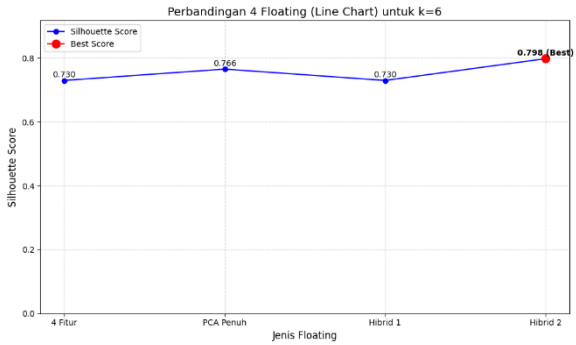
b). K = 3



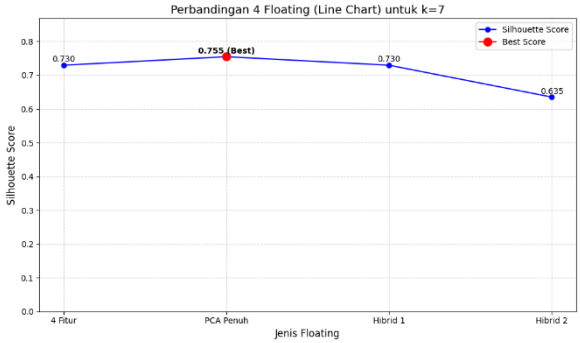
c). K = 4



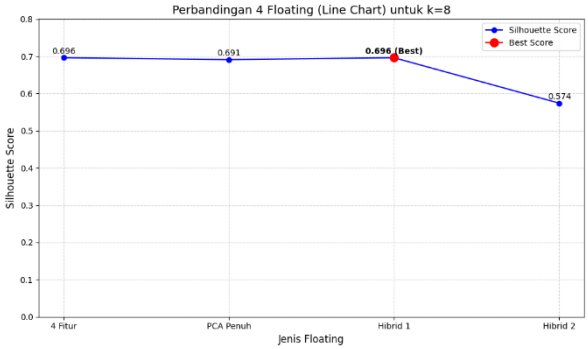
d). K = 5



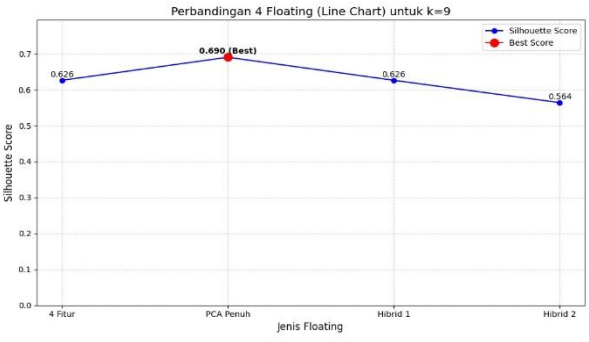
e). K = 6



f). K = 7



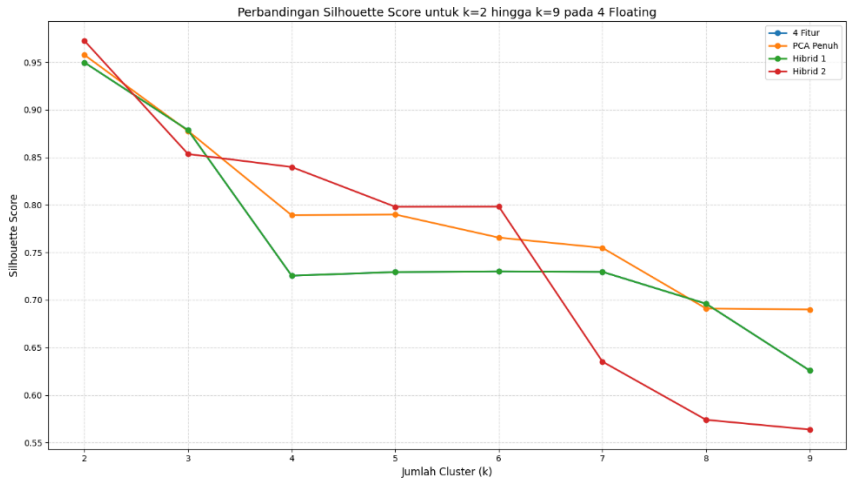
g). K = 8



h). K = 9

Gambar 3. Visualisasi 4 Floating untuk K=2 hingga K=9

*) Corresponding Author
https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.



Gambar 2. Perbandingan *Silhouette Score* K=2 hingga K=9

Rekapitulasi nilai *Silhouette Score* dari keempat pendekatan (F1 hingga F4) pada seluruh rentang *k* disajikan dalam Tabel 2. Hasil menunjukkan bahwa pendekatan F4 (Hibrid 2) mencatat skor tertinggi secara keseluruhan, dengan nilai 0.972 pada saat $k = 2$. Nilai ini merupakan puncak dari seluruh skenario yang diuji dan menjadi indikator kuat bahwa dua klaster merupakan pemisahan yang paling optimal. Skor F4 juga tetap unggul untuk k = 4, 5, dan 6, meskipun mengalami penurunan performa setelah k > 6. Sebaliknya, pendekatan F1 dan F3 mencatat nilai yang relatif stagnan di seluruh variasi k, dan tidak menunjukkan peningkatan signifikan bahkan ketika jumlah klaster ditambah.

Tabel 2 Hasil 4 Floting pada K=2 hingga k=9

K	F1 (4 Fitur)	F2 (PCA Penuh)	F3 (Hibrid 1)	F4 (Hibrid 2)	Skenario Terbaik
2	0.950	0.958	0.950	**0.972**	F4: Hibrid 2
3	0.878	0.878	0.878	0.853	F3: Hibrid 1
4	0.726	0.789	0.726	0.840	F4: Hibrid 2
5	0.729	0.790	0.729	0.798	F4: Hibrid 2
6	0.730	0.766	0.730	0.798	F4: Hibrid 2
7	0.730	0.755	0.730	0.635	F2: PCA Penuh
8	0.696	0.691	0.696	0.574	F3: Hibrid 1
9	0.626	0.690	0.626	0.564	F2: PCA Penuh

Hasil visualisasi klasterisasi untuk berbagai nilai *k* pada masing-masing pendekatan ditampilkan pada Gambar 3. a) Pendekatan F1, yang menggunakan empat fitur, menunjukkan distribusi data yang cenderung tidak stabil, dengan banyak titik yang tumpang tindih di area tengah. Ketika nilai *k* meningkat, tidak tampak adanya pemisahan klaster yang jelas, sehingga mengindikasikan keterbatasan fitur dalam membedakan dua jenis URL. b) Pendekatan F2, yang menggunakan reduksi dimensi penuh melalui PCA, menunjukkan struktur yang relatif lebih terarah pada $k = 2$ dan 3. Namun, visualisasi menjadi semakin tidak teratur ketika *k* meningkat, dan pemisahan klaster melemah secara signifikan. Hal ini mengindikasikan obahwa PCA tidak cukup efektif dalam mempertahankan struktur pemisahan klaster pada kondisi kompleks. c) Pendekatan F3 (Hibrid 1), yang menggabungkan sebagian fitur asli dengan hasil PCA, tidak memberikan peningkatan visual yang berarti. Distribusi data menunjukkan pola serupa dengan F1, dengan klaster yang kurang terdefinisi secara visual. d) Sebaliknya, pendekatan F4 (Hibrid 2) menunjukkan struktur klaster yang paling stabil dan terpisah. Pada nilai $k = 2$ hingga 4, tampak terbentuknya dua atau lebih kelompok data dengan batas yang cukup tegas, mengindikasikan bahwa dua fitur terpilih sudah cukup untuk menghasilkan pemisahan yang efektif. Meskipun pada nilai *k* di atas 6 bentuk klaster mulai melemah, distribusi dari F4 masih terlihat lebih terstruktur dibanding pendekatan lainnya. e) hingga h) menampilkan kelanjutan visualisasi dari pendekatan yang sama untuk nilai *k* berikutnya, disusun dalam grid yang seragam untuk memudahkan analisis perbandingan visual secara menyeluruh. Secara umum, hanya pendekatan F4 yang mampu mempertahankan konsistensi bentuk klaster yang relatif stabil, bahkan ketika jumlah klaster berubah.

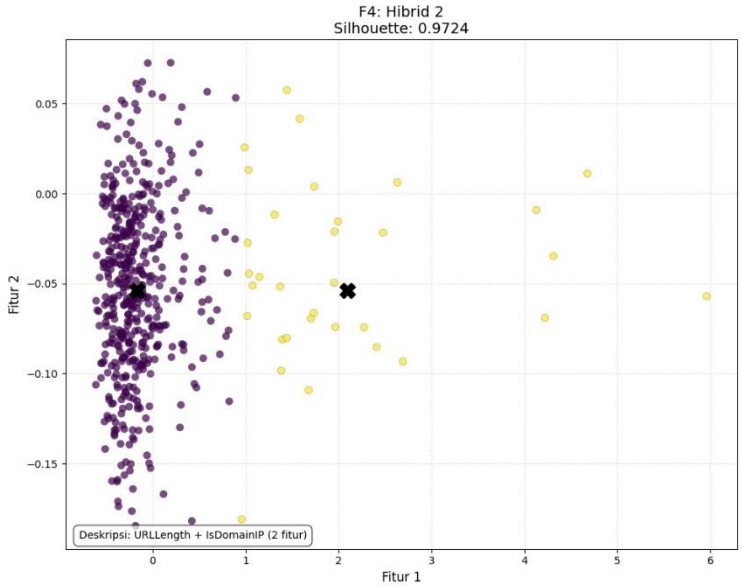
Untuk mendukung temuan visual ini, Gambar 4 menyajikan perbandingan nilai *Silhouette Score* dari masing-masing pendekatan terhadap variasi *k*. Nilai skor dari F4 menempati posisi tertinggi pada $k = 2$ dan tetap kompetitif hingga $k = 6$, sebelum mengalami penurunan. Sementara itu, pendekatan F1, F2, dan F3 menunjukkan tren skor yang relatif datar, menandakan bahwa penambahan fitur maupun penerapan PCA tidak memberikan kontribusi yang signifikan terhadap kualitas pemisahan klaster.

Berdasarkan evaluasi menyeluruh yang merujuk pada Gambar 3 a) hingga h), Gambar 4, serta Tabel 2, dapat disimpulkan bahwa pendekatan F4 menunjukkan performa paling unggul secara visual maupun kuantitatif. Temuan ini menegaskan bahwa pemilihan fitur yang tepat, meskipun dalam jumlah terbatas, dapat menghasilkan struktur

https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.

klaster yang stabil tanpa memerlukan transformasi dimensi yang kompleks. Berdasarkan hasil evaluasi, pendekatan F4 (Hibrid 2) dengan k = 2 memberikan hasil terbaik dengan *Silhouette Score* tertinggi (0.972). Hal ini menunjukkan bahwa:

1. Fitur URLLength dan IsDomainIP memiliki daya diskriminatif yang tinggi dalam membedakan URL normal dan URL *Phishing*, bahkan tanpa memerlukan fitur tambahan seperti NoOfOtherSpecialCharsInURL dan NoOfQMarkInURL.
2. Penggunaan alamat IP sebagai domain merupakan indikator kuat dari URL *Phishing*, seperti yang ditunjukkan oleh karakteristik Cluster 1 dalam pendekatan F4.
3. Jumlah cluster optimal adalah 2, yang sesuai dengan intuisi bahwa URL dapat dikelompokkan menjadi URL normal dan URL *Phishing*.
4. Pendekatan sederhana dengan dua fitur lebih efektif daripada pendekatan yang lebih kompleks dengan empat fitur atau kombinasi PCA, yang menunjukkan bahwa kompleksitas model tidak selalu menghasilkan performa yang lebih baik



Gambar 5. Visualisasi klasterisasi F4 (Hibrid 2) pada K=2

Gambar 5 memperlihatkan hasil klasterisasi F4 (Hibrid 2) dengan k=2, menggunakan fitur *URLLength* dan *IsDomainIP*. Klaster berwarna ungu dengan jumlah data 9,971 merepresentasikan URL normal dengan distribusi padat, sedangkan klaster kuning dengan jumlah data 29 menunjukkan URL *Phishing* yang tersebar atau setara dengan 0.0029% dari 10,000 data. Perbedaan posisi dan kepadatan antar klaster memperkuat hasil evaluasi sebelumnya, dengan *Silhouette Score* tertinggi sebesar 0,9724. Visualisasi ini menegaskan bahwa dua fitur tersebut sudah cukup untuk membedakan URL *Phishing* tanpa tambahan fitur atau reduksi dimensi.

6. KESIMPULAN

Pendekatan K-Means dengan fitur *URLLength* dan *IsDomainIP* pada k=2 memberikan pemisahan klaster terbaik dengan *Silhouette Score* 0.9724. dengan perlehan 9,971 data yang mempresentasikan URL normal, dan 29 URL Phising yang terdeteksi atau sekitar 0.0029% dari 10,000 data URL yang digunakan. Visualisasi memperlihatkan klaster ungu sebagai URL normal yang padat, dan klaster kuning sebagai URL *Phishing* yang lebih tersebar. Hasil ini menunjukkan bahwa dua fitur tersebut sudah cukup untuk membedakan URL *Phishing* secara efektif tanpa perlu fitur tambahan atau teknik reduksi dimensi.

7. DAFTAR PUSTAKA

Saputro, I. A., Sugiarto, L., & Nugraha, F. S. (2024). Analisis Kesadaran Masyarakat Terhadap Bahaya Internet Phishing Menggunakan K-Means Clustering. *STRING (Satuan Tulisan Riset Dan Inovasi Teknologi)*, *9*(2), 139–146.

Windarni, V. A., Nugraha, A. F., Ramadhani, S. T. A., Istiqomah, D. A., Puri, F. M., & Setiawan, A. (2023). Deteksi Website Phishing Menggunakan Teknik Filter Pada Model Machine. *Information System Journal (INFOS)* |, *6*(1), 39–43.

*) Corresponding Author
https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.

Dewi, S., & Pakereng, M. A. I. (2023). Implementasi Principal Component Analysis Pada K-Means Untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *8*(4), 1186–1195. https://doi.org/10.29100/jipi.v8i4.4101

Fatiha, M. R., Setiawan, I., Ikhsan, A. N., & Yunita, I. R. (2024). Optimisasi Sistem Deteksi Phishing Berbasis Web Menggunakan Algoritma Decision Tree. *Jurnal Ilmiah IT CIDA : Diseminasi Teknologi Informasi*, *10*(2). Retrieved from https://www.kaggle.com

Tampinongkol, F. F., Kamila, A. R., Wardhana, A. cahya, Kusuma, A. W. C., & Revaldo, D. (2024). Implementation of Random Forest Classification and Support Vector Machine Algorithms for Phishing Link Detection. *Journal of Informatics, Information System, Software Engineering and Applications (INISTA)*, *7*(1), 127–137. https://doi.org/10.20895/INISTA.V7I1.1588

Foozy, C. F. M., Anuar, M. A. I., Maslan, A., Adam, H. A. M., & Mahdin, H. (2024). Phishing URLs Detection Using Naives Baiyes, Random Forest and LightGBM Algorithms. *International Journal of Data Science*, *5*(1), 56–63.

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2022). Stochastic Neighbor Embedding with Gaussian and Student-t Distributions: Tutorial and Survey. *Stochastic Neighbor Embedding with Gaussian and Student-t Distributions: Tutorial and Survey*, 1–13. Retrieved from http://arxiv.org/abs/2009.10301

Guntara, M., & Lutfi, N. (2023). Optimasi Cacah Klaster pada Klasterisasi dengan Algoritma KMeans Menggunakan Silhouette Coeficient dan Elbow Method. *JuTI "Jurnal Teknologi Informasi," 2*(1), 43. https://doi.org/10.26798/juti.v2i1.944

Gusthvi, W., Roza, A. A., & Allo, C. B. G. (2023). Perbandingan Metode Klasifikasi Decission Tree, Naive Bayes, K-Nearest-Neighbor, dan Logistic Regression pada Dataset Phishing. *CENDERAWASIH Journal of Statistics and Data Science*, *1*. Retrieved from https://ejurnal.fmipa.uncen.ac.id/index.php/CJSDS

Kaspersky. (2024). Laporan Ancaman Siber di Indonesia 2024. Retrieved February 24, 2025, from https://www.antaranews.com/berita/4656245/kaspersky-deteksi-36-juta-ancaman-siber-lokal-di-indonesia-pada-2024

Mulyani, H., Setiawan, R. A., & Fathi, H. (2023). Optimization Of K Value In Clustering Using Silhouette Score (Case Study: Mall Customers Data). *JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION*, *6*, 45–49.

Muriithi, N. M., & Karani, J. (2024). A Systematic Literature Review on Phishing Detection Model. *International Journal of Computer and Information Technology*, *13*(2), 2279–0764. Retrieved from www.ijcit.com62

Pribadi, R. A., & Sulianta, F. (2024). *Metode K-Means Clustering dalam Pengelompokan Penjualan Produk Indofood*. 1–9.

Rahmah, S. A. (2024). Review Terbaru Tentang Klasterisasi Data Mining Menggunakan Metode K-Means: Tantangan Dan Aplikasi. *Jurnal Teknologi Informasi*, *5*(2), 297–303. https://doi.org/10.46576/djtechno

Saputra, E. A., & Nataliani, Y. (2021). Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means. *Journal of Information Systems and Informatics*, *3*(3), 424–439. Retrieved from http://journal-isi.org/index.php/isi

Wijaya, A. T., & Subandi. (2024). Penerapan Metode Clustering Dengan Algoritma K-Means Pada Sistem Pendeteksi Pencucian Uang Perbankan Berbasis Web. *SENAFTI (Seminar Nasional Mahasiswa Fakultas Teknologi Informasi)*, *3*(2), 398–406.

Wang, Y., Huang, H., Rudin, C., & Shaposhnik, Y. (2021). Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data Visualization. *Journal of Machine Learning Research*, 1–73. Retrieved from http://arxiv.org/abs/2012.04456

https://doi.org/10.30872/atasi.v4i2.2887
2025 Adopsi Teknologi dan Sistem Informasi (ATASI) with CC BY SA license.