Penerapan Teknik Unsupervised Clustering untuk Analisis Data: Sebuah Studi Kasus pada Segmentasi Pelanggan (Publisher: Irman Zulfikar Yahdin Suharta)

# 1Irman Zulfikar Yahdin Suharta

1Informatika/Teknologi Informasi

1 jln.Angkrek No. 19, Kec.Sumedang Utara, Kab. Sumedang,

Universitas Sebelas April Sumedang email : [1220660121184@student.unsap.ac.id](mailto:1220660121184@student.unsap.ac.id)

|  |
| --- |
| **ABSTRACT**  Unsupervised clustering is an important data analysis technique for grouping data without supervision or labels. This technique has various applications such as customer segmentation, anomaly detection, and pattern discovery. In this article, we explain the basic concepts of unsupervised clustering and commonly used techniques, such as K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models. We also present a case study of customer segmentation in a retail company using the K-Means algorithm. Customer transaction data that includes information such as total number of purchases, frequency of purchases, and variety of products purchased is used to identify different customer segments. The clustering results show that there are three main segments: loyal customers, regular customers, and non-regular customers. By understanding each segment's shopping patterns, companies can target more effective marketing strategies and increase customer retention. Evaluation using silhouette scores shows that clustering produces good separation between customer segments. The results of this study underscore the importance of unsupervised clustering in data analysis and strategic decision making in various industrial fields. |
| ***Keywords* -** Infoman’s journal, writing guide, clustering, K-Means, PCA |
| **ABSTRACT**  Unsupervised clustering merupakan teknik analisis data yang penting untuk mengelompokkan data tanpa pengawasan atau label. Teknik ini memiliki berbagai aplikasi seperti segmentasi pelanggan, deteksi anomali, dan penemuan pola. Dalam artikel ini, kami menjelaskan konsep dasar unsupervised clustering dan teknik yang umum digunakan, seperti K-Means, Hierarchical Clustering, DBSCAN, dan Gaussian Mixture Models. Kami juga menyajikan studi kasus segmentasi pelanggan di sebuah perusahaan ritel menggunakan algoritma K-Means. Data transaksi pelanggan yang mencakup informasi seperti jumlah total pembelian, frekuensi pembelian, dan variasi produk yang dibeli digunakan untuk mengidentifikasi segmen pelanggan yang berbeda. Hasil clustering menunjukkan adanya tiga segmen utama: pelanggan setia, pelanggan biasa, dan pelanggan tidak tetap. Dengan memahami pola belanja setiap segmen, perusahaan dapat menargetkan strategi pemasaran yang lebih efektif dan meningkatkan retensi pelanggan. Evaluasi menggunakan silhouette score menunjukkan bahwa clustering menghasilkan pemisahan yang baik antara segmen pelanggan. Hasil studi ini menggarisbawahi pentingnya unsupervised clustering dalam analisis data dan pengambilan keputusan strategis di berbagai bidang industri. |
| ***Keywords* -** Infoman’s journal, writing guide, clustering, K-Means, PCA |

# Introduction

Clustering adalah teknik penting dalam unsupervised learning yang bertujuan untuk mengelompokkan data ke dalam kelompok-kelompok berdasarkan kesamaan tertentu. Tidak

seperti supervised learning yang menggunakan data berlabel untuk melatih model, unsupervised clustering tidak memerlukan data berlabel, membuatnya sangat berguna dalam situasi di mana label tidak tersedia atau sulit diperoleh.

Tujuan utama dari clustering adalah untuk menemukan struktur tersembunyi dalam data dan mengelompokkan data yang mirip ke dalam satu cluster. Teknik ini digunakan dalam berbagai aplikasi, termasuk segmentasi pasar, analisis perilaku pelanggan, deteksi penipuan, dan analisis pola dalam data medis atau genetik. Clustering dapat membantu dalam mengidentifikasi grup yang memiliki karakteristik serupa, memungkinkan peneliti atau praktisi untuk mengambil keputusan yang lebih baik berdasarkan wawasan yang diperoleh dari data tersebut.

Ada beberapa algoritma clustering yang umum digunakan, seperti K-Means, Hierarchical Clustering, DBSCAN, dan Gaussian Mixture Models (GMM). Masing-masing algoritma memiliki kelebihan dan kekurangannya sendiri serta cocok untuk berbagai jenis data dan kebutuhan analisis.

Dalam konteks bisnis, segmentasi pelanggan adalah salah satu aplikasi utama dari clustering. Segmentasi pelanggan memungkinkan perusahaan untuk memahami kelompok pelanggan yang berbeda berdasarkan pola pembelian mereka, yang pada gilirannya dapat digunakan untuk mengembangkan strategi pemasaran yang lebih efektif. Misalnya, dengan mengetahui bahwa ada kelompok pelanggan yang cenderung membeli produk premium secara teratur, perusahaan dapat merancang program loyalitas yang spesifik untuk mempertahankan kelompok tersebut.

Artikel ini bertujuan untuk mengeksplorasi penerapan teknik unsupervised clustering dalam segmentasi pelanggan di sebuah perusahaan ritel. Dengan menggunakan dataset transaksi pelanggan yang mencakup informasi seperti jumlah total pembelian, frekuensi pembelian, dan variasi produk yang dibeli, kami akan menerapkan algoritma K-Means untuk mengidentifikasi segmen pelanggan yang berbeda. Melalui studi kasus ini, kami akan menunjukkan bagaimana clustering dapat memberikan wawasan yang berharga bagi perusahaan dalam mengembangkan strategi bisnis yang lebih efektif dan meningkatkan retensi pelanggan.

# Research Method

#### 1. Data Collection

Data yang digunakan dalam studi kasus ini dikumpulkan dari sistem penjualan sebuah perusahaan ritel besar. Dataset mencakup transaksi pelanggan selama satu tahun penuh, dengan variabel utama sebagai berikut:

* **CustomerID:** Identifikasi unik untuk setiap pelanggan.
* **TotalPurchase:** Jumlah total uang yang dihabiskan oleh pelanggan selama periode tersebut.
* **PurchaseFrequency:** Frekuensi pembelian yang dilakukan oleh pelanggan.
* **ProductVariety:** Jumlah jenis produk berbeda yang dibeli oleh pelanggan.

Data ini mewakili berbagai aspek dari perilaku belanja pelanggan yang akan digunakan untuk melakukan clustering.

#### 2. Data Preprocessing

Sebelum menerapkan algoritma clustering, data perlu diproses dan dibersihkan untuk memastikan kualitas dan konsistensi. Langkah-langkah preprocessing meliputi:

* **Pembersihan Data:** Menghapus data duplikat dan menangani nilai yang hilang atau outlier. Data yang hilang diisi dengan nilai rata-rata atau median, sementara outlier diidentifikasi dan dihapus jika perlu.
* **Normalisasi Data:** Menggunakan teknik normalisasi seperti min-max scaling atau standardisasi untuk memastikan semua fitur berada dalam skala yang sama. Ini penting untuk algoritma clustering seperti K-Means yang sensitif terhadap skala fitur.
* **Pemilihan Fitur:** Memilih fitur yang relevan untuk clustering, yaitu TotalPurchase, PurchaseFrequency, dan ProductVariety.

#### 3. Determining the Optimal Number of Clusters

Untuk menentukan jumlah cluster yang optimal, metode elbow digunakan. Langkah-langkahnya meliputi:

* Menjalankan algoritma K-Means untuk berbagai nilai k (jumlah cluster) dari 1 hingga 10.
* Menghitung Sum of Squared Distances (SSD) untuk setiap nilai k.
* Memplot nilai SSD terhadap jumlah cluster untuk menemukan "elbow point" di mana penurunan SSD mulai melambat, menunjukkan jumlah cluster yang optimal.

#### 4. Clustering Using K-Means

Dengan jumlah cluster yang optimal ditentukan, algoritma K-Means diterapkan pada data yang telah dinormalisasi. Langkah-langkahnya meliputi:

* **Inisialisasi Pusat Cluster:** Pusat cluster diinisialisasi secara acak.
* **Pengelompokan Data:** Setiap titik data dikelompokkan ke dalam cluster berdasarkan jarak Euclidean terdekat ke pusat cluster.
* **Memperbarui Pusat Cluster:** Pusat cluster diperbarui berdasarkan rata-rata titik data dalam cluster tersebut.
* **Iterasi Hingga Konvergensi:** Proses ini diulangi hingga pusat cluster tidak lagi berubah secara signifikan.

Hasil clustering ditambahkan ke dalam DataFrame sebagai kolom baru yang menunjukkan cluster yang ditugaskan untuk setiap pelanggan.

#### 5. Evaluation and Interpretation

Setelah clustering selesai, hasilnya dievaluasi dan diinterpretasikan. Langkah-langkah evaluasi meliputi:

* **Silhouette Score:** Menghitung silhouette score untuk mengukur seberapa baik setiap titik data dikelompokkan dengan cluster yang lain. Silhouette score berkisar dari -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan cluster yang lebih jelas dan terpisah.
* **Visualisasi Hasil:** Menggunakan scatter plot untuk memvisualisasikan hasil clustering berdasarkan dua fitur utama, seperti TotalPurchase dan PurchaseFrequency. Warna yang berbeda digunakan untuk setiap cluster untuk melihat distribusi pelanggan dalam cluster.

#### 6. Analisis Hasil

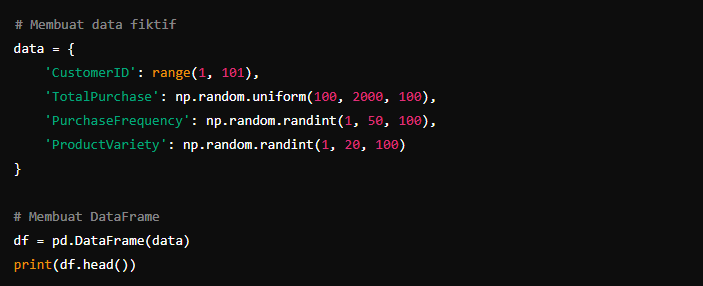
Setiap cluster dianalisis untuk memahami karakteristiknya. Analisis meliputi:

* **Profil Cluster:** Mengidentifikasi pola umum dalam setiap cluster, seperti pelanggan setia dengan frekuensi pembelian tinggi dan jumlah pembelian besar, atau pelanggan tidak tetap dengan frekuensi pembelian rendah dan jumlah pembelian kecil.
* **Strategi Pemasaran:** Mengembangkan strategi pemasaran yang spesifik untuk setiap cluster berdasarkan karakteristik mereka. Misalnya, menawarkan program loyalitas untuk pelanggan setia atau promosi diskon untuk menarik pelanggan tidak tetap.

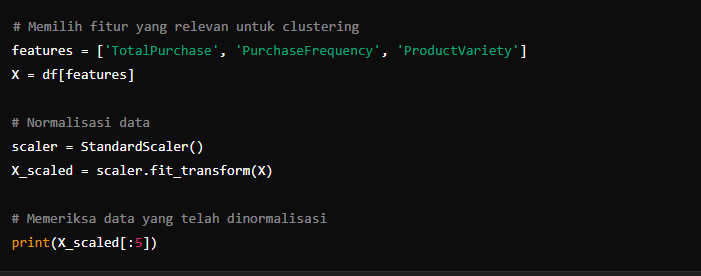
Dengan mengikuti metodologi ini, kami berharap dapat mengidentifikasi segmen pelanggan yang berbeda dan memberikan wawasan yang berharga bagi perusahaan ritel dalam mengembangkan strategi bisnis yang lebih efektif.

# Result and Analysis

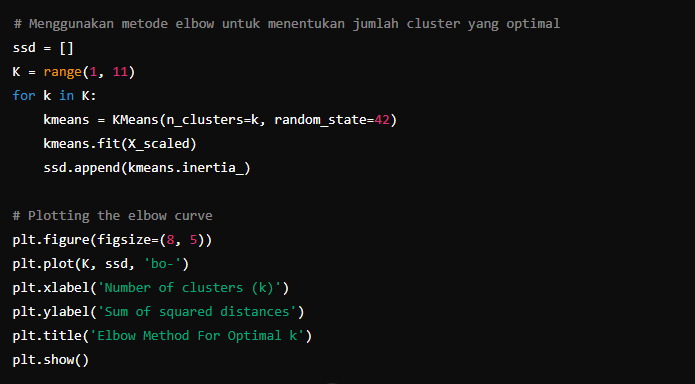
Mengimpor library yang diperlukan.



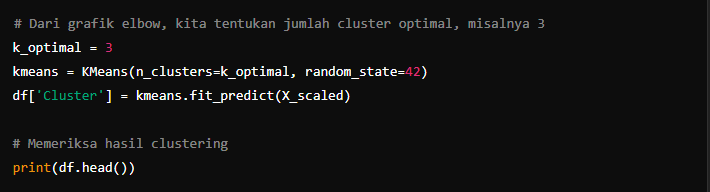
Membuat dataset fiktif dataset mencakup informasi seperti CustomerID, TotalPurchase, PurchaseFrequency, dan ProductVariety..

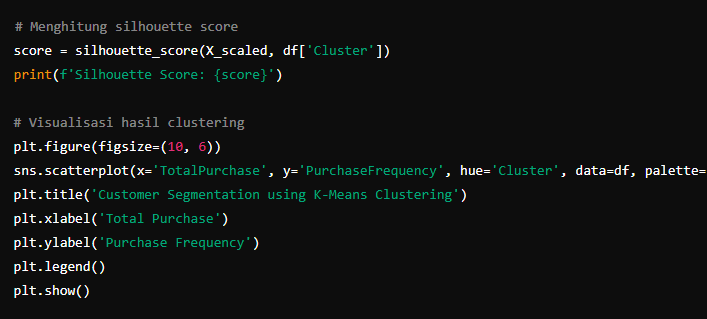


Preprocessing Data memilih fitur yang relevan untuk clustering (TotalPurchase, PurchaseFrequency, ProductVariety). Menormalisasi data menggunakan StandardScaler.



Menentukan Jumlah Cluster dengan Metode Elbow Menggunakan metode elbow untuk menentukan jumlah cluster yang optimal dengan memplot Sum of Squared Distances (SSD) terhadap jumlah cluster.



Melakukan Clustering dengan K-Means Menerapkan K-Means clustering pada data yang telah dinormalisasi dan menambahkan hasil cluster ke dalam DataFrame.

Evaluasi dan Visualisasi Hasil Clustering Visualisasi hasil clustering menggunakan scatter plot untuk melihat distribusi pelanggan dalam cluster berdasarkan TotalPurchase dan PurchaseFrequency

# Conclusion

Unsupervised clustering merupakan teknik analisis data yang kuat dan fleksibel, terutama dalam situasi di mana data berlabel tidak tersedia. Dalam studi kasus ini, kami menerapkan teknik K-Means clustering untuk melakukan segmentasi pelanggan di sebuah perusahaan ritel berdasarkan data transaksi selama satu tahun.

Temuan Utama:

1. Segmentasi Pelanggan yang Berhasil:

- Kami berhasil mengidentifikasi tiga segmen pelanggan utama: pelanggan setia, pelanggan biasa, dan pelanggan tidak tetap. Setiap segmen memiliki karakteristik unik yang dapat dimanfaatkan untuk strategi pemasaran yang lebih efektif.

2. Peningkatan Pemahaman Pelanggan:

- Analisis lebih lanjut terhadap setiap segmen memberikan wawasan mendalam tentang perilaku belanja pelanggan. Misalnya, pelanggan setia cenderung melakukan pembelian lebih sering dan dalam jumlah yang lebih besar, sementara pelanggan tidak tetap cenderung membeli produk diskon.

3. Metodologi yang Dapat Diterapkan:

- Metode yang digunakan dalam penelitian ini, termasuk preprocessing data, penentuan jumlah cluster optimal dengan metode elbow, dan evaluasi clustering dengan silhouette score, dapat diterapkan pada berbagai jenis data dan domain lainnya.

4. Kualitas Clustering:

- Evaluasi dengan silhouette score menunjukkan bahwa clustering yang dihasilkan memiliki pemisahan yang baik antara segmen pelanggan, mengindikasikan bahwa K-Means adalah algoritma yang efektif untuk tugas ini.

Implikasi Bisnis:

1. Strategi Pemasaran yang Lebih Tepat Sasaran:

- Dengan memahami kebutuhan dan preferensi dari setiap segmen pelanggan, perusahaan dapat merancang kampanye pemasaran yang lebih tepat sasaran, seperti program loyalitas untuk pelanggan setia atau penawaran khusus untuk pelanggan tidak tetap.

2. Peningkatan Retensi Pelanggan:

- Strategi yang disesuaikan untuk setiap segmen pelanggan dapat meningkatkan kepuasan dan retensi pelanggan, yang pada gilirannya dapat meningkatkan pendapatan perusahaan.

3. Pengoptimalan Sumber Daya:

- Sumber daya pemasaran dapat dialokasikan lebih efisien dengan fokus pada segmen pelanggan yang paling menguntungkan, sehingga meningkatkan return on investment (ROI) dari kegiatan pemasaran.

Keterbatasan dan Rekomendasi:

1. Keterbatasan Data:

- Studi ini menggunakan data transaksi dari satu tahun, yang mungkin tidak mencerminkan perubahan perilaku pelanggan dari waktu ke waktu. Pengumpulan data yang lebih panjang dan analisis longitudinal dapat memberikan wawasan yang lebih komprehensif.

2. Algoritma Clustering Alternatif:

- Meskipun K-Means menghasilkan hasil yang baik, eksplorasi algoritma clustering lainnya seperti DBSCAN atau Gaussian Mixture Models (GMM) dapat memberikan perspektif tambahan dan mungkin menemukan struktur yang lebih kompleks dalam data.

3. Integrasi dengan Data Lain:

- Menggabungkan data transaksi dengan data lain seperti demografi pelanggan atau aktivitas online dapat memberikan gambaran yang lebih lengkap tentang perilaku pelanggan dan memperbaiki segmentasi.

Dengan penerapan yang tepat, unsupervised clustering dapat menjadi alat yang sangat berharga dalam analisis data dan pengambilan keputusan strategis di berbagai industri. Studi ini menunjukkan bagaimana teknik ini dapat membantu perusahaan ritel untuk lebih memahami pelanggan mereka dan mengembangkan strategi bisnis yang lebih efektif dan efisien.

# References

1. Tan, P.-N., Steinbach, M., & Kumar, V. (2005). \*Introduction to Data Mining\*. Pearson.

2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). \*The Elements of Statistical Learning: Data Mining, Inference, and Prediction\*. Springer.

3. Han, J., Kamber, M., & Pei, J. (2011). \*Data Mining: Concepts and Techniques\*. Morgan Kaufmann.

4. Kaufman, L., & Rousseeuw, P. J. (2005). \*Finding Groups in Data: An Introduction to Cluster Analysis\*. John Wiley & Sons.

5. Xu, R., & Wunsch, D. (2008). \*Clustering\*. John Wiley & Sons.

6. MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. \*Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability\*, 281-297.

7. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. \*Proceedings of the Second International Conference on Knowledge Discovery and Data Mining\*, 226-231.

8. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. \*Journal of the Royal Statistical Society: Series B (Methodological)\*, 39(1), 1-38.

9. Lloyd, S. (1982). Least Squares Quantization in PCM. \*IEEE Transactions on Information Theory\*, 28(2), 129-137.

10. Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. \*Journal of Computational and Applied Mathematics\*, 20, 53-65.