Informatika sains teknologi
FAKULTAS SAIN & TEKNOLOGI
PROGRAM STUDI TEKNIK INFORMATIKA

https://doi.org/10.34005/insit.v2i2.4119

Vol. 02 No. 02 (2024) 9 - 14

E-ISSN : 3025-3608

Komparasi Algoritma K-Means dan *Hierarchical Clustering* Untuk Mengetahui Data Customer Dalam Layanan Internet

Afri Yudha[1], Shofi Khaerunnisa[2], Timor Setiyaningsih[3]

[1,2,3]Program Studi Teknologi Informasi Universitas Darma Persada

ibnugazali@gmail.com

Abstract

This study aims to compare two popular clustering algorithms, namely K-Means and Hierarchical Clustering, in grouping customer data of internet services. The data used includes information such as date, purchase order number, type of internet, product description, quantity in Mbps, region, city, and rate in IDR. Clustering is a technique in data mining used to group data based on similar characteristics, helping companies understand usage patterns, market segmentation, and improve services to customers. In this study, both algorithms were tested using the Sum of Squared Errors (SSE) metric to evaluate clustering performance. The results showed that the K-Means algorithm outperformed Hierarchical Clustering in terms of efficiency and accuracy when evaluating SSE values. K-Means was able to cluster data more quickly and produce clearer clusters, while Hierarchical Clustering required longer computation time and yielded less optimal clustering results in some cases. In conclusion, the K-Means algorithm is more suitable for analyzing customer data of internet services in this context due to its advantages in speed and accuracy of clustering based on SSE results. This study provides guidance for companies in choosing the appropriate clustering algorithm to group customer data to enhance business strategies and service quality.

Keywords*: K-Means, Hierarchical Clustering, clustering, customer data, internet services, data mining, market segmentation, Sum of Squared Errors (SSE).*

Abstrak

Penelitian ini bertujuan untuk membandingkan dua algoritma clustering yang populer, yaitu K-Means dan Hierarchical Clustering, dalam mengelompokkan data pelanggan layanan internet. Data yang digunakan mencakup informasi seperti tanggal, nomor pesanan, jenis internet, deskripsi produk, kuantitas Mbps, wilayah, kota, dan tarif IDR. Clustering adalah teknik dalam data mining yang digunakan untuk mengelompokkan data berdasarkan karakteristik yang serupa, membantu perusahaan dalam memahami pola penggunaan, segmentasi pasar, dan meningkatkan layanan kepada pelanggan. Pada penelitian ini, kedua algoritma diuji menggunakan metrik *Sum of Squared Errors* (**SSE**) untuk mengevaluasi kinerja clustering. Hasil penelitian menunjukkan bahwa algoritma K-Means lebih unggul dibandingkan dengan Hierarchical Clustering dalam hal efisiensi dan akurasi ketika mengevaluasi nilai SSE. K-Means mampu mengelompokkan data dengan lebih cepat dan menghasilkan cluster yang lebih jelas, sementara Hierarchical Clustering membutuhkan waktu komputasi yang lebih lama dan hasil clustering yang kurang optimal dalam beberapa kasus. Kesimpulannya, algoritma K-Means lebih cocok digunakan untuk analisis data pelanggan layanan internet dalam konteks ini karena keunggulannya dalam kecepatan dan akurasi clustering berdasarkan hasil pengecekan SSE. Penelitian ini memberikan panduan bagi perusahaan dalam memilih algoritma clustering yang tepat untuk mengelompokkan data pelanggan guna meningkatkan strategi bisnis dan kualitas layanan.

Kata Kunci: K-Means, Hierarchical Clustering, clustering, data pelanggan, layanan internet, data mining, segmentasi pasar, *Sum of Squared Errors* (**SSE**).

I. Pendahuluan

PT XYZ adalah perusahaan telekomunikasi terkemuka di Indonesia, yang bergerak dalam sektor telekomunikasi. PT XYZ memiliki jaringan yang luas dan canggih di seluruh Indonesia, termasuk infrastruktur serat optik, jaringan nirkabel, satelit, dan kabel bawah laut. Sebagai BUMN, PT XYZ berkomitmen untuk mendukung pembangunan masyarakat dan perekonomian Indonesia. Mengingat tingginya permintaan pelanggan, PT XYZ perlu mengolah data untuk menyusun strategi bisnis yang efektif guna mempertahankan hubungan baik dengan pelanggan, terutama pelanggan setia. Salah satu cara untuk mencapai hal ini adalah dengan melakukan segmentasi pelanggan, yaitu mengelompokkan pelanggan dengan karakteristik serupa ke dalam satu kelompok. Segmentasi ini dapat dilakukan menggunakan metode clustering, yang merupakan salah satu teknik data mining untuk mengelompokkan data menjadi kelompok-kelompok tertentu. Metode clustering memiliki berbagai algoritma, di antaranya algoritma K-Means dan Hierarchical. Algoritma K-Means adalah teknik partisi yang membagi atau memisahkan objek ke dalam wilayah yang terpisah. Pada algoritma ini, setiap objek harus dimasukkan ke dalam kelompok atau cluster tertentu. Sebaliknya, Hierarchical Clustering mengelompokkan data berdasarkan hierarki atau tingkatan tertentu, memungkinkan analisis yang lebih mendalam terhadap struktur data.

Penelitian ini memilih algoritma K-Means dan Hierarchical Clustering dengan beberapa pertimbangan penting. Pertama, algoritma K-Means dikenal efisien dalam menangani dataset besar dan mampu menghasilkan cluster yang jelas dan terpisah. K-Means juga relatif mudah diimplementasikan dan memberikan hasil clustering yang mudah diinterpretasikan. Di sisi lain, Hierarchical Clustering dipilih karena kemampuannya membangun hierarki atau dendrogram yang memberikan wawasan mendalam tentang struktur data dan hubungan antar cluster. Kombinasi kedua algoritma ini memungkinkan peneliti mendapatkan gambaran yang lebih komprehensif tentang data pelanggan. Dalam penelitian ini, peneliti melakukan studi kasus pada perusahaan distributor yang bergerak di bidang layanan, dengan target pemasaran yang dapat mencakup berbagai segmen pelanggan. Penelitian ini tidak hanya fokus pada pemilihan algoritma clustering yang tepat, tetapi juga bertujuan memberikan rekomendasi strategi bisnis yang lebih efektif berdasarkan hasil clustering tersebut. Dengan demikian, PT XYZ dapat lebih memahami kebutuhan pelanggan, meningkatkan kualitas layanan, dan mempertahankan loyalitas pelanggan yang tinggi.

Selain itu, penelitian ini juga bertujuan untuk menjawab tantangan yang dihadapi oleh banyak perusahaan di era digital saat ini, yaitu bagaimana mengelola dan memanfaatkan data pelanggan yang besar dan kompleks untuk mendukung keputusan bisnis strategis. Dengan menerapkan metode clustering, perusahaan dapat mengidentifikasi pola-pola tersembunyi dalam data pelanggan, mengoptimalkan kampanye pemasaran, dan meningkatkan efisiensi operasional. Penelitian ini diharapkan dapat memberikan kontribusi signifikan bagi PT XYZ dalam upaya mereka tetap menjadi pemimpin di industri telekomunikasi nasional, serta memberikan wawasan berharga bagi perusahaan lain yang menghadapi tantangan serupa dalam pengelolaan data pelanggan.

II. Metodologi penelitian

1) Data Mining

Data mining merupakan “proses pengolahan data untuk menemukan pola-pola baru dan bermanfaat dari kumpulan data besar” (Efori Buulolo, 2020). Menurut Han dan Kamber data mining adalah “bagian dari proses Knowledge Discovery in Databases (KDD) yang terdiri dari beberapa tahap, mulai dari pembersihan data hingga interpretasi hasil” (Han et al., 2012). Teknik-teknik dalam data mining digunakan untuk berbagai tujuan seperti klasifikasi, regresi, clustering, dan asosiasi. *Output* dalam *data mining* dapat dipergunakan sebagai alternatif dalam pengambilan keputusan atau untuk memperbaiki keputusan di masa yang akan datang (Efori Buulolo, 2020).

2) Clustering

Clustering adalah “salah satu teknik dalam data mining yang digunakan untuk mengelompokkan data menjadi beberapa kelompok atau cluster berdasarkan kemiripan karakteristik”. Menurut Jain clustering adalah “proses pengelompokan objek-objek ke dalam kelompok yang anggotanya memiliki kesamaan tertentu”. Clustering membantu dalam memahami struktur dan hubungan dalam data serta mengidentifikasi pola-pola yang tersembunyi”(Jain et al., 1999).

3) K-Means

Metode K-Means adalah “salah satu metode populer dalam analisis data dan merupakan teknik yang umum digunakan untuk melakukan klausterisasi (clustering) data”. Algoritma ini mengelompokkan data

berdasarkan jarak antara setiap data dengan pusat kelompok yang mewakili kelompok tersebut. Rumus penggunaan K- Means melibatkan dua komponen utama, yaitu jarak Euclidean dan pembaruan pusat kelompok(Kaur et al., 2020).

4) Hierarchichal Clustering

Hierarchical Clustering adalah "Metode klausterisasi yang berusaha untuk membentuk hierarki kluster berdasarkan jarak antara data point". Hierarchical Clustering adalah "Metode clustering yang membangun hierarki atau dendrogram dari data. Metode ini bisa bersifat agglomerative (bottom-up) atau divisive (top-down)". Menurut Johnson *Agglomerative Clustering* memulai dengan setiap data point sebagai cluster terpisah dan menggabungkannya secara bertahap, sedangkan *divisive clustering* memulai dengan satu cluster dan memecahnya secara bertahap (Johnson, 1967).

5) Sum Of Squared Eror (SSE)

Sum of Squared Errors (SSE) adalah "metrik yang digunakan untuk mengevaluasi kinerja clustering dengan mengukur jumlah kesalahan kuadrat antara setiap data point dan centroid terdekatnya". SSE yang lebih rendah menunjukkan clustering yang lebih baik. Menurut Hartigan SSE digunakan untuk menentukan kualitas clustering dan membantu dalam memilih jumlah cluster yang optimal (Jollyta et al., 2019).

6) Segmentasi Pelanggan

Segmentasi pelanggan adalah "proses mengelompokkan pelanggan ke dalam kelompok-kelompok berdasarkan karakteristik atau perilaku mereka". Segmentasi ini membantu perusahaan dalam memahami kebutuhan pelanggan, mengidentifikasi peluang pasar, dan merancang strategi pemasaran yang lebih efektif. Menurut Kotler dan Keller, segmentasi pelanggan adalah "langkah kunci dalam manajemen hubungan pelanggan (CRM) yang berfokus pada meningkatkan kepuasan dan loyalitas pelanggan" (Kotler & Keller, 2016)

7) UML

Unified Modeling Language (UML) adalah "bahasa standar yang digunakan untuk spesifikasi, visualisasi, pengembangan, dan pendokumentasian artefak dari sistem perangkat lunak". UML menyediakan serangkaian notasi untuk menggambarkan berbagai aspek dari sistem, termasuk struktur, perilaku, dan interaksi antar komponen (Sudarman, 2019).

8) PHP-MYSQL-Bootstrap

PHP adalah "bahasa skrip server-side yang dirancang khusus untuk pengembangan web. PHP memungkinkan pembuatan halaman web yang dinamis dengan kemampuan untuk berinteraksi dengan basis data dan memproses input pengguna". Bootstrap adalah "framework front-end yang digunakan untuk pengembangan web responsif dan mobile-first". Bootstrap menyediakan komponen HTML, CSS, dan JavaScript yang sudah siap pakai, yang memudahkan pengembang dalam membuat desain web yang konsisten dan modern. MySQL adalah "sistem manajemen basis data relasional (RDBMS) yang menggunakan SQL (*Structured Query Language*) sebagai bahasa pengelolaan data". (Wibowo P, 2018).

III. HASIL PEMBAHASAN

Dalam penelitian ini, langkah-langkah penelitian yang penulis lakukan menggunakan CRISP-DM (Cross-Industry Standard Process for Data Mining).

Gambar 1. CRISP-DM
Sumber: Michael Fuchs Python (2024)

1. *Business Understanding* (Pemahaman Bisnis)
   Pada fase ini penulis ingin memahami tujuan bisnis PT XYZ terkait dengan segmentasi pelanggan layanan internet dan Mengidentifikasi pola penggunaan layanan internet oleh pelanggan.
2. *Data Understanding* (Pemahaman Data)
   Pada fase ini penulis Mengumpulkan data pelanggan yang relevan, menganalisis data untuk memahami distribusi, korelasi, dan anomali. Mengidentifikasi data yang hilang atau tidak konsisten dan memeriksa kelengkapan data dan mengatasi data yang hilang atau anomali
3. *Data Preparation* (Persiapan Data)

Pada fase ini penulis memilih atribut yang relevan untuk analisis clustering, melakukan pembersihan data. Mengatasi data yang hilang atau salah. Menormalkan atau menskalakan data jika diperlukan.
Jika diperlukan, membuat atribut baru berdasarkan atribut yang ada.

4. *Modeling* (Pemodelan)
   Pada fase ini penulis memilih algoritma K-Means dan Hierarchical Clustering untuk diimplementasikan dan Membangun Model. Mengaplikasikan algoritma K-Means pada data pelanggan. Mengaplikasikan algoritma Hierarchical Clustering pada data pelanggan. Menentukan jumlah cluster yang optimal untuk K-Means. Mengatur parameter yang sesuai untuk Hierarchical Clustering. Mengulangi proses pemodelan untuk mengoptimalkan hasil.
5. *Evaluation* (Evaluasi)
   Pada fase ini penulis menggunakan metrik *Sum of Squared Errors* (SSE) untuk mengevaluasi performa clustering dari kedua algoritma. Membandingkan hasil clustering K-Means dan Hierarchical.
   Mengevaluasi hasil clustering terhadap tujuan bisnis yang telah ditetapkan. Memastikan bahwa cluster yang dihasilkan memiliki makna dan dapat digunakan untuk pengambilan keputusan bisnis.
6. *Deployment* (Penerapan)
   Pada fase ini, penulis membuat sebuah produk dari hasil pemodelan yang telah didapatkan dan Mengintegrasikan model clustering yang terpilih ke dalam sistem bisnis PT XYZ.

4. KESIMPULAN DAN SARAN

Dalam implementasi metode k-means pada segmentasi pelanggan layanan internet, hasil dari cluster dapat dijelaskan dengan menggunakan rumus k-means seperti dibawah ini.

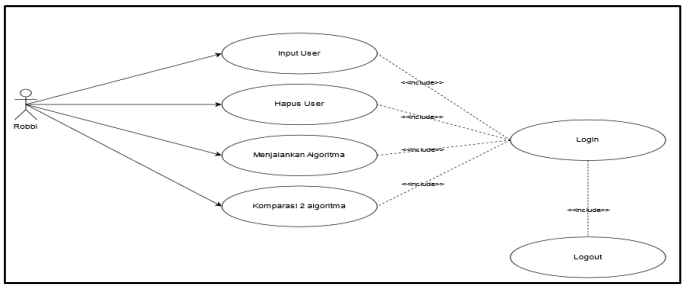
$$C_j = \{x_i \mid \min_i \|x_i - \mu_j\|\}$$

di mana $\|x_i - \mu_j\|$ adalah jarak Euclidean antara data x_i dan centroid μ_j.

Sedangkan rumus *Hierarchichal Clustering* pada segmentasi pelanggan layanan internet, dapat dijelaskan dengan menggunakan rumus seperti dibawah ini.

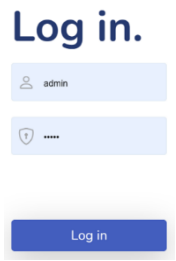
$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

Setelah merumuskan, maka penulis membuat rancangan aplikasi berdasarkan skenario dengan diagram UseCase yang telah penuli buat.



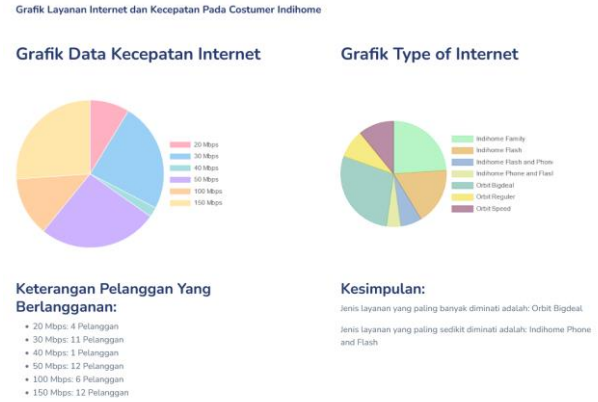
Gambar 2. Diagram Use Case Aplikasi K-Means dan *Hierarchichal Clustering*

Penulis membuat login untuk memverifkasi awal user akan masuk kedalam aplikasi, yang dapat di lihat dari gambar dibawah ini.



Gambar 3 Login Aplikasi

Penulis membuat grafik dalam aplikasi untuk menunjukkan paket dan kecepatan internet yang sering disewa oleh pelanggan



Gambar 4. Grafik Kecepatan Dan Paket Internet

Penulis menampilkan hasil algoritma K-Means setelah melakukan proses perhitungan dari data yang telah di masukkan.

Gambar 5. Hasil Algoritma K-Means dan SSE

Penulis menampilkan hasil algoritma *Hierarchichal Clustering* setelah melakukan proses perhitungan dari data yang telah di masukkan.

Gambar 6. Hasil Algoritma *Hierarchichal Clustering* dan SSE

Penulis melakukan komparasi algoritma k-means dan *Hierarchichal Clustering* dengan menggunakan SSE sebagai standart yang dapat dilihat di bawah ini.

Gambar 7. Hasil Komparasi Algoritma K-Means dan *Hierarchichal Clustering*

Hasil penerapan Algoritma K-Means dan *Hierarchichal Clustering* sudah di lakukan dari tahap pemodelan hingga tahap perhitungan. Metode K-Means dan Hierarchical Clustering sesuai dengan yang diharapkan yaitu mengcluster data customer dalam layanan internet PT. XYZ.

5. Kesimpulan dan Saran

1. Kesimpulan

a) Pengembangan sistem informasi menggunakan analisis clustering berhasil dilakukan dengan menerapkan metode K-Means dan Hierarchical Clustering.

b) Hasil analisis menunjukkan bahwa model K-Means memiliki Sum of Squared Errors (SSE) yang lebih rendah (SSE 98935) dibandingkan dengan model Hierarchical (SSE 155100), menunjukkan prediksi yang lebih baik dalam menclusterisasi layanan internet.

c) Kedua metode clustering ini efektif dalam mengorganisir data pelanggan layanan internet berdasarkan pola penggunaan yang berbeda-beda.

2. Saran

a) Gunakan dataset yang lebih luas, seperti data dari beberapa tahun terakhir, untuk meningkatkan akurasi dan keakuratan clustering.

b) Uji performa model dalam situasi yang lebih kompleks untuk memvalidasi keandalan dan keefektifan prediksi pasar.

c) Eksplorasi metode clustering lain atau variasi dari K-Means dan Hierarchical Clustering untuk memperluas pemahaman terhadap data pelanggan dalam layanan internet.

d) Fokuskan penelitian pada pengembangan strategi penjualan yang lebih efektif berdasarkan hasil clustering untuk meningkatkan kepuasan pelanggan dan pertumbuhan bisnis.

Daftar Pustaka

Efori Buulolo. (2020). *Data Mining Untuk Perguruan Tinggi - Efori Buulolo*. Google Book.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining : Concepts and Techniques : Concepts and Techniques (3rd Edition). In *Data Mining*. Morgan Kaufmann. http://linkinghub.elsevier.com/retrieve/pii/B9780123814791000010

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering ACM Computing Surveys. *Intelligent Multidimensional Data Clustering and Analysis*, *31*(3), 265–323.

Johnson, S. C. (1967). *Hierarchical Clustering Schemes. Psychometrika*.

Jollyta, D., Efendi, S., Zarlis, M., & Mawengkang, H. (2019). Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum of Square Error dan Davies Bouldin Index.

Prosiding Seminar Nasional Riset Information Science (SENARIS), *1*(September), 918. https://doi.org/10.30645/senaris.v1i0.100

Kaur, A., Pal, S. K., & Singh, A. P. (2020). Quantum-inspired ant lion optimized hybrid K-Means for cluster analysis and intrusion detection. *Knowledge-Based Syst*, *97*. https://doi.org/10.1016/j.asoc.2019.105523

Kotler, P., & Keller, K. L. (2016). *Manajemen Pemasaran* (G. A. Pratama (ed.); 13th ed.). Gelora Aksara Pratama.

Sudarman. (2019). *Pemodelan Sistem Informasi Menggunakan Unified Modeling Language (UML)*. Elex Media Komputindo.

Wibowo P. (2018). *Pemrograman Web Dinamis Menggunakan PHP dan MySQL*. Andi.