**Komparasi Algoritma DBSCAN dan K-Means terhadap Hasil Bebras Computational Thinking Challenge Tingkat SMP**

Marta Yulianti1, \*, Victoria Alysha F. S. 2

1,2*Program Studi Informatika Fakultas Sains dan Teknologi,*

*Universitas Sanata Dharma, Yogyakarta, Indonesia*

***Abstract***

*The Bebras Computational Thinking Challenge is an international competition aimed at testing and enhancing students' computational thinking skills. In this study, data from the junior high school level of the competition, consisting of 397 entries with 30 attributes, is used to compare the performance of DBSCAN and K-Means algorithms in data clustering. The data underwent preprocessing to ensure quality and relevance before analysis. The DBSCAN and K-Means algorithms were then used to identify patterns in student performance, and the clustering results were evaluated using the Silhouette Coefficient. The results show that K-Means outperforms DBSCAN, with a Silhouette Coefficient of 0.674 at k = 3, while DBSCAN only achieved 0.538 with two clusters, indicating that K-Means provides better cluster validity in this data analysis.*

*Keyword: Clustering, Computational Thinking, DBSCAN, K-Means, Silhouette Coefficient*

**Abstrak**

‘Bebras Computational Thinking Challenge’ adalah kompetisi internasional yang bertujuan untuk menguji dan meningkatkan keterampilan berpikir komputasional siswa. Dalam penelitian ini, data dari kompetisi tingkat SMP, yang terdiri dari 397 data dengan 30 atribut, digunakan untuk membandingkan kinerja algoritma DBSCAN dan K-Means dalam pengelompokan data. Data tersebut telah melalui proses preprocessing untuk memastikan kualitas dan relevansinya sebelum dianalisis. Algoritma DBSCAN dan K-Means kemudian digunakan untuk mengidentifikasi pola performa siswa, dan hasil clustering dievaluasi menggunakan Silhouette Coefficient. Hasilnya menunjukkan bahwa K-Means lebih unggul dengan nilai Silhouette Coefficient sebesar 0,674 pada k = 3, sementara DBSCAN hanya mencapai 0,538 dengan dua cluster, sehingga K-Means memberikan validitas cluster yang lebih baik dalam analisis data ini.

Kata Kunci: Berpikir Komputasional, DBSCAN, K-Means, Klastering, Koefisien Silhouette

**1. PENDAHULUAN**

Bebras Computational Thinking Challenge adalah sebuah kompetisi internasional yang dirancang untuk menguji dan meningkatkan kemampuan berpikir komputasional siswa dari jenjang SD hingga SMA. Kompetisi ini diselenggarakan secara rutin setiap tahun dan melibatkan ribuan peserta dari berbagai negara, termasuk Indonesia.

Perkembangan teknologi yang pesat di abad ke-21 telah membawa dampak signifikan pada dunia pendidikan di seluruh dunia, termasuk Indonesia. Di era *Society* 5.0, kemampuan untuk beradaptasi dengan teknologi dan memiliki keterampilan pemecahan masalah yang baik menjadi sangat penting. *Computational Thinking*, sebagai metode untuk memahami dan menyelesaikan masalah kompleks menggunakan teknik dan konsep ilmu komputer seperti dekomposisi, pengenalan pola, abstraksi, dan algoritma, dipandang oleh banyak ahli sebagai salah satu keterampilan kunci dalam pendidikan abad ke-21 [1]. Melalui *computational thinking*, siswa didorong untuk mengembangkan keterampilan berpikir kritis, kreatif, komunikatif, dan kolaboratif dalam menyelesaikan masalah.

Penelitian ini memanfaatkan dua dataset yang diperoleh dari hasil Bebras Computational Thinking Challenge, yang berisi data peserta dan hasil dari sejumlah siswa SMP yang mengikuti kompetisi tersebut. Dengan data ini, akan dilakukan komparasi antara dua algoritma *clustering* yang populer, yaitu DBSCAN (Density-Based Spatial Clustering of Applications with Noise) dan K-Means, untuk menganalisis pola-pola dalam hasil Bebras Computational Thinking Challenge. K-Means adalah algoritma clustering yang membagi data ke dalam sejumlah kelompok (cluster) berdasarkan jarak *centroid*. Sementara itu, DBSCAN adalah algoritma *clustering* berbasis kepadatan yang dapat menangani data dengan bentuk yang tidak beraturan dan mampu mendeteksi noise.

Pada penelitian sebelumnya yang membandingkan algoritma DBSCAN dan K-Means untuk pengelompokan kasus Covid-19 di dunia, hasilnya menunjukkan bahwa K-Means lebih unggul dengan nilai *silhouette index* (SI) terbaik sebesar 0,6902 pada k = 8. Sementara itu, nilai SI terbaik untuk DBSCAN adalah 0,3624 dengan Eps 0,2 dan MinPts 3 [2]. Selain itu, penelitian lain yang membandingkan algoritma DBSCAN dan K-Means untuk pengelompokan status desa di Jawa Tengah tahun 2020 menunjukkan bahwa DBSCAN lebih unggul dengan rata-rata nilai *silhouette* sebesar 0,41 dibandingkan K-Means yang hanya sebesar 0,35 [3]. Berdasarkan hasil-hasil penelitian tersebut, dapat disimpulkan bahwa kinerja model clustering terbaik berdasarkan *silhouette coefficient* atau *silhouette index* masih menunjukkan variasi yang signifikan tergantung pada konteks dan karakteristik data yang digunakan. Penelitian ini bertujuan untuk membandingkan kinerja algoritma DBSCAN dan K-Means dalam mengelompokkan data hasil Bebras Computational Thinking Challenge. Melalui komparasi ini, diharapkan dapat memberikan wawasan lebih lanjut tentang pola-pola performa siswa dan efektivitas kedua algoritma dalam mengelompokkan data hasil kompetisi.

**2. METODOLOGI PENELITIAN**

Penelitian ini dimulai dengan tahap pengumpulan data yang akan digunakan sebagai data mentah. Data yang terkumpul kemudian akan diproses untuk memastikan kualitas dan relevansinya. Setelah data diolah, langkah selanjutnya adalah melatih model menggunakan dua algoritma yang berbeda, yaitu DBSCAN dan K-Means. Setelah proses pelatihan selesai, model tersebut akan diuji untuk menilai kinerjanya.

A diagram of a data cleaning process

Description automatically generated

**Gambar 1.** Metodologi Penelitian

1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah dataset dari hasil Bebras Computational Thinking Challenge untuk siswa SMP. Dataset pertama berisi 397 data dengan 20 atribut, sedangkan dataset kedua berisi 396 data dengan 10 atribut. Penjelasan lebih lanjut mengenai atribut-atribut tersebut dapat dilihat pada **Tabel 1**.

**Tabel 1**. Atribut Dataset

|  |  |
| --- | --- |
| Atribut | Keterangan |
| ID\_Siswa | Nomor identitas siswa |
| Sekolah | Nama sekolah asal siswa |
| Jml\_ikut | Berapa kali siswa tersebut mengikuti challenge |
| Persiapan | Cara siswa mempersiapkan diri untuk mengikuti challenge |
| Wkt\_persiapan | Waktu yang dipergunakan siswa untuk belajar mempersiapkan challenge |
| N\_Bindo | Nilai mata pelajaran Bahasa Indonesia milik siswa |
| N\_IPA | Nilai mata pelajaran IPA milik siswa |
| N\_Mat | Nilai mata pelajaran Matematika milik siswa |
| Kesan | Kesan siswa mengenai challenge |
| Minat | Minat siswa untuk mengikuti challenge lagi pada kesempatan mendatang |
| JK | Jenis kelamin siswa |
| Skor\_Total | Nilai total yang diperoleh siswa dalam challenge |
| Waktu | Waktu yang dipergunakan siswa untuk mengerjakan soal |
| Q1 - Q15 | Nilai yang diperoleh siswa untuk soal ke-1 sampai ke-15 |

1. *Preprocessing* Data

*Preprocessing* data adalah tahap penting dalam mempersiapkan data mentah agar siap digunakan dalam proses analisis dengan lebih mudah dan akurat [4]. Tahapan preprocessing mencakup beberapa keputusan krusial, seperti penanganan *missing value* untuk mengisi atau menghapus data yang hilang, menghilangkan duplikasi, seleksi data untuk memilih fitur-fitur yang relevan dan transformasi data.

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

DBSCAN adalah algoritma pengelompokan yang didasarkan pada kepadatan (*density*) data. Algoritma ini memiliki dua parameter utama yang menentukan pembentukan cluster yaitu jumlah sampel minimal (minPts) dan Eps (*ε*). Parameter pertama menentukan jumlah titik minimal yang dapat diklasifikasikan sebagai sampel inti, yang juga menentukan toleransi noise dari algoritma tersebut [2].

Langkah-langkahnya adalah sebagai berikut:

* 1. Tentukan nilai parameter MinPts dan Eps.
  2. Tentukan nilai p atau titik awal secara acak.
  3. Hitung Eps atau semua jarak titik yang *density reachable* terhadap p menggunakan rumus jarak *euclidean* pada **Persamaan (1).**

(1)

Keterangan :

= variabel ke-a dari obyek i (i = 1, …, n; a=1, …, p)

= nilai *euclidean distance*

* 1. Terbentuk sebuah cluster ketika titik yang memenuhi Eps lebih dari MinPts dan titik p sebagai *core point*.
  2. Ulangi langkah 3 – 4 hingga semua titik diproses. Jika p merupakan titik border dan tidak ada titik yang *density reachable* terhadap p, maka proses dilanjutkan ke titik yang lain.

1. K-Means

K-Means adalah salah satu metode pengelompokan data nonhierarki (sekatan) yang berusaha membagi data ke dalam dua atau lebih kelompok [5]. Metode ini mengelompokkan data sedemikian rupa sehingga data dengan karakteristik yang sama dimasukkan ke dalam kelompok yang sama, sementara data dengan karakteristik yang berbeda dimasukkan ke dalam kelompok yang berbeda. Pengelompokan data menggunakan metode K-Means umumnya dilakukan dengan algoritma berikut:

* + - 1. Menentukan banyaknya k kelompok.
      2. Membagi data ke dalam k kelompok.
      3. Menghitung pusat kelompok (*centroid)* dari data yang ada di masing-masing kelompok dan dinyatakan dalam bentuk **Persamaan (2).**

= (2)

di mana *C* adalah *centroid*, *M* adalah banyaknya data, dan *i* adalah banyaknya kelompok.

* + - 1. Masing-masing data dialokasikan ke *centroid* terdekat dengan menghitung jarak data ke setiap *centroid* menggunakan jarak *Euclidean* dan dinyatakan dalam bentuk **Persamaan (3).**

(3)

* + - 1. Kembali ke langkah 3 apabila masih ada data yang berpindah kelompok.

1. Evaluasi

Evaluasi adalah tahap yang bertujuan untuk mengukur kualitas model dengan menilai keakuratan hasil yang dicapai oleh model tersebut. Model *clustering* dapat diukur dengan *silhouette coefficient,* yang menilai seberapa baik setiap objek data sesuai dengan kelompoknya dibandingkan dengan kelompok lainnya. Evaluasi ini menggabungkan konsep kohesi (cohesion) dan pemisahan (separation) dalam pengelompokan data. Nilai *silhouette coefficient* berkisar dari -1 hingga 1, dengan nilai mendekati 1 menunjukkan kualitas pengelompokan yang baik [6]. Nilai *silhouette coefficient* dinyatakan dalam **Persamaan (4)** [3].

(4)

Keterangan:

rata-rata jarak titik inti dengan semua titik pada kelompok yang sama.

rata-rata jarak titik inti dengan semua titik pada kelompok yang berbeda menggunakan jarak *Euclidean.*

**3.** **HASIL DAN ANALISIS**

Data mentah yang sudah digabungkan mencakup 30 atribut dan 397 baris data. Data akan dilakukan *preprocessing,* dimulai dengan pengecekan missing value dan penghapusan baris yang memiliki missing valu. Lalu dilakukan seleksi atribut, dimana atribut yang dianggap tidak relevan seperti ‘Persiapan’, ‘Kesan’, dan ‘ID\_siswa’ dihapus. Transformasi data diterapkan dengan mengubah data ‘Waktu’ menjadi nilai dalam hitungan detik dan atribut kategorikal menjadi numerik.

Penelitian ini mengkomparasikan kinerja model K-Means dan DBSCAN. Hasil pengujian menunjukkan bahwa DBSCAN memperoleh nilai *silhouette coefficient* sebesar 0.538 dengan jumlah *cluster* sebanyak 2, dan K-means mempreoleh *silhouette coefficient* sebesar 0.674 dengan jumlah *cluster* sebanyak 3.

**Figur 1.** Komparasi

Berdasarkan hasil perbandingan yang tertera dalam **Figur 1**, terlihat bahwa K-Means memiliki kemampuan untuk melakukan pengelompokan (clustering) dengan lebih baik, dengan nilai *silhouette coefficient* sebesar 0.674.

**4. KESIMPULAN**

Setelah menerapkan metode K-Means dan DBSCAN untuk analisis data hasil kompetisi Bebras Computational Thinking Challenge, dengan menggunakan dataset berukuran 397 data. Hasil klasterisasi tersebut telah divalidasi menggunakan *silhouette coefficient*. Algoritma K-Means menunjukkan kinerja yang lebih baik dengan nilai *silhouette coefficient* terbaik sebesar 0,674 pada k = 3, sementara DBSCAN memperoleh *silhouette coefficient* sebesar 0.538 dengan jumlah k *=* 2. Dengan demikian, penelitian ini menyimpulkan bahwa algoritma K-Means memiliki validitas *cluster* yang lebih baik dibandingkan dengan DBSCAN, menghasilkan *cluster* paling optimal pada k = 2.

**REFERENSI**

[1]Ansori, M. (2020). Pemikiran Komputasi (Computational Thinking) dalam Pemecahan Masalah. *DIRASAH*, *3*(1). <https://ejournal.iaifa.ac.id/index.php/dirasah>

[2]Adha, R., Nurhaliza, N., & Soleha, U. (2021). Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia. *Jurnal Sains, Teknologi Dan Industri*, *18*(2), 206–211. <https://covid19.who.int>.

[3]Putri, M. M., Dewi, C., Permata Siam, E., Asri Wijayanti, G., Aulia, N., & Nooraeni, R. (2021). *Comparison of DBSCAN and K-Means Clustering for Grouping the Village Status in Central Java 2020 Komparasi DBSCAN dan K-Means Clustering pada Pengelompokan Status Desa di Jawa Tengah Tahun 2020*. *17*(3), 394–404. https://doi.org/10.20956/j.v17i3.11704

[4]Agung, A., Daniswara, A., Kadek, I., & Nuryana, D. (2023). Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru. *Journal of Informatics and Computer Science*, *05*.

[5]Sisca Agustin Diani Budiman, Diah Safitri, & Dwi Ispriyanti. (2016). Perbandingan Metode K-Means dan Metode DBSCAN Pada Pengelompokan Rumah Kost Mahasiswa di Kelurahan Tembalang Semarang. *JURNAL GAUSSIAN*, *5*(4), 757–762. <https://doi.org/10.14710/j.gauss.5.4.757-762>

[6]Paembonan, S., & Abduh, H. (2021). Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat. *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, *2*, 48. https://doi.org/10.51557/pt\_jiit.v6i2.659