International Journal of Applied Sciences and Smart Technologies

Volume 5, Issue 2, pages 1–12

p-ISSN 2655-8564, e-ISSN 2685-9432

IJASST

Perbandingan Algoritma K-Means dan DBSCAN terhadap Klasterisasi Data Bebras Computational Thinking Challenge

Resiana Kinanti Jati[1, *], Loadtriana Oktavia S[2]

[1,2] *Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Sanata Dharma, Yogyakarta, Indonesia*

[*]*Corresponding Author: 215314159@student.usd.ac.id*

(Received DD-MM-YYYY; Revised DD-MM-YYYY; Accepted DD-MM-YYYY)

Abstrak

Penelitian ini membandingkan dua algoritma klasterisasi data, K-Means dan DBSCAN, dalam mengelompokkan data Bebras Computational Thinking Challenge. Dataset yang digunakan terdiri dari hasil kompetisi Bebras untuk siswa SMP. Algoritma K-Means dan DBSCAN diterapkan pada data tersebut, dan kinerja masing-masing algoritma dievaluasi menggunakan Silhouette Coefficient. Hasil penelitian menunjukkan bahwa algoritma K-Means menghasilkan klaster yang lebih baik dengan Silhouette Coefficient sebesar 0,613 dibandingkan dengan DBSCAN yang memiliki nilai 0,493.

Kata kunci: Bebras, DBSCAN, Klastering, K-Means, Penambangan Data.

Abstract

This study compares two data clustering algorithms, K-Means and DBSCAN, in clustering data from the Bebras Computational Thinking Challenge. The dataset used consists of results from the Bebras competition for junior high school students. Both K-Means and DBSCAN algorithms were applied to the data, and the performance of each algorithm was evaluated using the Silhouette Coefficient. The study found that the K-Means algorithm produced better clusters with a Silhouette Coefficient of 0,613 compared to DBSCAN, which had a value of 0,493.

Keywords: Bebras, Clustering, Data Mining, DBSCAN, K-Means.

1 PENDAHULUAN

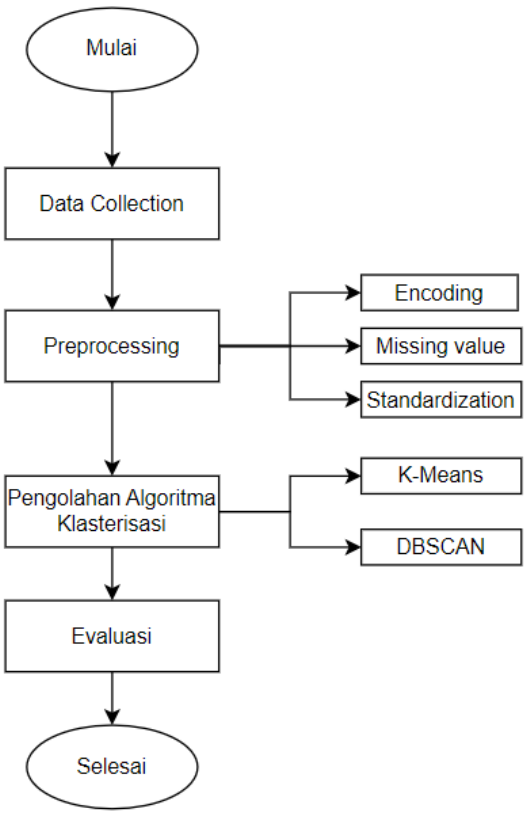
Bebras Computational Thinking Challenge (Bebras CT Challenge) merupakan kompetisi internasional yang bertujuan untuk mengukur kemampuan *Computational Thinking* (CT) di kalangan siswa dan guru mulai dari tingkat SD. CT adalah kemampuan untuk menyelesaikan masalah dengan menggunakan konsep dan metode informatika.

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

Di Indonesia, Bebras CT Challenge diselenggarakan oleh Bebras Indonesia, sebuah organisasi nirlaba yang dibentuk pada tahun 2014. Kompetisi ini menghasilkan data yang berharga tentang kemampuan CT para siswa. Data ini dapat dianalisis untuk berbagai keperluan, salah satunya adalah untuk mengelompokkan (*clustering*) para siswa berdasarkan kemampuan CT mereka.

2 METODE PENELITIAN

Dalam penelitian ini dilakukan dengan beberapa tahapan. Setelah data didapatkan, peneliti merancang langkah-langkah yang bertujuan untuk mendapatkan hasil perbandingan yang maksimal antara dua algoritma terhadap data Bebras Computational Thinking Challenge. Langkah-langkah tersebut tergambar dalam gambar 1.



Gambar 1 Metode Penelitian

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

2.1 Data Collection

Data collection merupakan suatu langkah untuk mengumpulkan atau memperoleh data [1]. Pada penelitian ini, kumpulan data atau *dataset* didapatkan dari dosen pengampu yang mana *dataset* dibagikan melalui laman belajar mahasiswa. *Dataset* yang digunakan dalam penelitian ini adalah data Bebras Computational Thinking Challenge yang mana terdiri dari dua (2) dataset kemudian digabungkan menjadi satu.

2.2 Data Preprocessing

Data preprocessing merupakan langkah untuk mempersiapkan data sebelum digunakan [1] seperti mengatasi data dengan nilai kosong atau *null* (*missing value*), menghapus atribut yang tidak diperlukan, pengecekan duplikasi data, mengatasi nilai *outliers*, dan normalisasi data.

2.3 K-Means

K-Means adalah metode pengelompokan data yang membagi data menjadi dua atau lebih kelompok. Metode ini mengelompokkan data dengan karakteristik yang sama ke dalam satu kelompok, sedangkan data dengan karakteristik yang berbeda dikelompokkan ke dalam kelompok lain [2]. Proses pengelompokan ini dilakukan dengan iteratif, di mana data akan dialokasikan ke pusat *cluster* terdekat, yang diperbaharui setiap iterasi untuk meminimalkan jarak antara data dan pusat cluster tersebut. Untuk menghitung jarak dari titik I (*xi*) ke pusat *cluster* titik K (*ck*), yang dinamakan (*dik*), dapat digunakan rumus Euclidean berikut.

$$d_{ik} = \sqrt{\sum_{j=1}^{m} \left(c_{ij} - x_{ik}\right)^2} \quad (1)$$

2.4 DBSCAN

Algoritma DBSCAN (*Density-Based Spatial Clustering of Aplication with Noise*) dapat menemukan area dengan kepadatan tinggi dan mengembangkan *cluster* dari area tersebut. Ada dua parameter utama yang menentukan pembentukan *cluster*: jumlah sampel minimal dan ε (epsilon). Parameter pertama menetapkan jumlah titik

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

minimal yang dibutuhkan untuk dianggap sebagai inti dari suatu *cluster* [3]. Hitung *Eps* atau semua jarak titik yang bisa dijangkau dari *p* berdasarkan kepadatan menggunakan rumus jarak Euclidean berikut.

$$d_{ij} = \sqrt{\sum_{a}^{p} \left(x_{ia} - x_{ja}\right)^2} \quad (2)$$

2.5 Silhouette Coefficient

Silhouette Coefficient digunakan untuk menemukan klaster terbaik dengan membandingkan seberapa dekat setiap titik data dalam sebuah klaster dengan klaster lainnya. Hasilnya membantu dalam mengevaluasi seberapa baik pengelompokan data tersebut sesuai dengan struktur sebenarnya dari data yang ada [4]. Untuk menghitung nilai Silhouette Index, dapat menggunakan rumus berikut:

$$SI = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{b(i) - a\,(i)}{max\{a(i), b(i)\}}\right) \quad (3)$$

2.6 Principal Component Analysis

Principal Component Analysis (PCA) adalah teknik analisis yang digunakan untuk menyederhanakan data dengan mengubahnya menjadi kombinasi linear yang lebih sedikit namun tetap mempertahankan sebagian besar variasi dari data awal. PCA juga membantu dalam menggambarkan struktur matriks varians-kovarians dari sekumpulan variabel dengan menggunakan kombinasi linier dari variabel-variabel tersebut [5].

3 HASIL DAN ANALISIS

3.1 Dataset

Data yang digunakan berjumlah 317 data. Terdiri dari dua dataset yaitu data berisi hasil Bebras Computational Thinking Challenge dari sejumlah siswa SMP dan data peserta Bebras Computational Thinking Challenge dari sejumlah siswa SMP yang mengikuti challenge. Atribut-atribut data yang digunakan antara lain: Jenis Kelamin (JK), Sekolah, Skor Total, Waktu, Q1 - Q15 (nilai yang diperoleh untuk soal ke-1

International Journal of Applied Sciences and Smart Technologies
Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

sampai dengan soal ke-15), Jumlah Keikutsertaan, Waktu Persiapan, Nilai Bahasa Indonesia, Nilai IPA, Nilai Matematika dan Minat.

ID_Siswa	JK	Sekolah	Skor_Total	Waktu	Q1	Q2	Q3	Q4	Q5	...	Q14	Q15	Jml_ikut	Persiapan	Wkt_persiapan	N_Bindo	N_IPA	N_Mat	Kesan	Minat
1	L	SMP A	61	42 min 43 detik	6,67	6,67	-1,67	6,67	-1,67	...	0,00	6,67	3 kali	Belajar mengerjakan soal Bebras sendiri di rum...	Lebih atau sama dengan 3 jam, tapi kurang dari...	> 90	> 90	> 90	Menantang	Ya
2	P	SMP A	21	31 min 30 detik	6,67	6,67	6,67	-0,83	-1,67	...	0,00	0,00	1 kali	Belajar mengerjakan soal Bebras sendiri di rum...	Lebih atau sama dengan 5 jam, tapi kurang dari...	> 90	81 - 90	> 90	Menyenangkan, Menantang	Ya
3	P	SMP A	84	44 min 45 detik	6,67	6,67	6,67	-0,83	6,67	...	6,67	6,67	3 kali	Belajar mengerjakan soal Bebras sendiri di rum...	Lebih dari 7 jam	> 90	> 90	> 90	Menyenangkan, Menantang, Penting untuk dipelaj...	Ya
4	P	SMP A	38	44 min 55 detik	6,67	-1,67	6,67	-0,83	6,67	...	0,00	0,00	> 3 kali	Belajar mengerjakan soal Bebras sendiri di rum...	Lebih atau sama dengan 1 jam, tapi kurang dari...	81 - 90	81 - 90	81 - 90	Biasa saja	Ya
5	P	SMP A	70	43 min 49 detik	6,67	6,67	6,67	-0,83	6,67	...	6,67	0,00	3 kali	Belajar mengerjakan soal Bebras bersama Ibu/Ba...	Kurang dari 1 jam	81 - 90	81 - 90	81 - 90	Menyenangkan, Biasa saja	Ya

Gambar 2 Dataset Mentah

3.2 Data Preprocessing

Langkah pertama setelah mendapatkan data adalah mempersiapkan data. Gambar 2 menunjukkan tampilan data mentah, di mana masih terdapat nilai kosong atau *null* (*missing value*) dan atribut yang tidak diperlukan. Selain itu juga diperlukan normalisasi atau standardisasi data. Hasil dari langkah preprocessing ditunjukkan pada gambar 3.

JK	Sekolah	Skor_Total	Waktu	Q1	Q2	Q3	Q4	Q5	Q6	...	Q12	Q13	Q14	Q15	Jml_ikut	Wkt_persiapan	N_Bindo	N_IPA	N_Mat	Minat
0	0	61	2563.0	6.67	6.67	-1.67	6.67	-1.67	-1.33	...	6.67	6.67	0.00	6.67	2	2	4	4	4	1
1	0	21	1890.0	6.67	6.67	6.67	-0.83	-1.67	-1.33	...	-1.67	-1.67	0.00	0.00	0	3	4	2	4	1
1	0	84	2685.0	6.67	6.67	6.67	-0.83	6.67	6.67	...	6.67	6.67	6.67	6.67	2	4	4	4	4	1
1	0	38	2695.0	6.67	-1.67	6.67	-0.83	6.67	-1.33	...	-1.67	-1.67	0.00	0.00	3	1	2	2	2	1
1	0	70	2629.0	6.67	6.67	6.67	-0.83	6.67	6.67	...	6.67	6.67	6.67	0.00	2	0	2	2	2	1

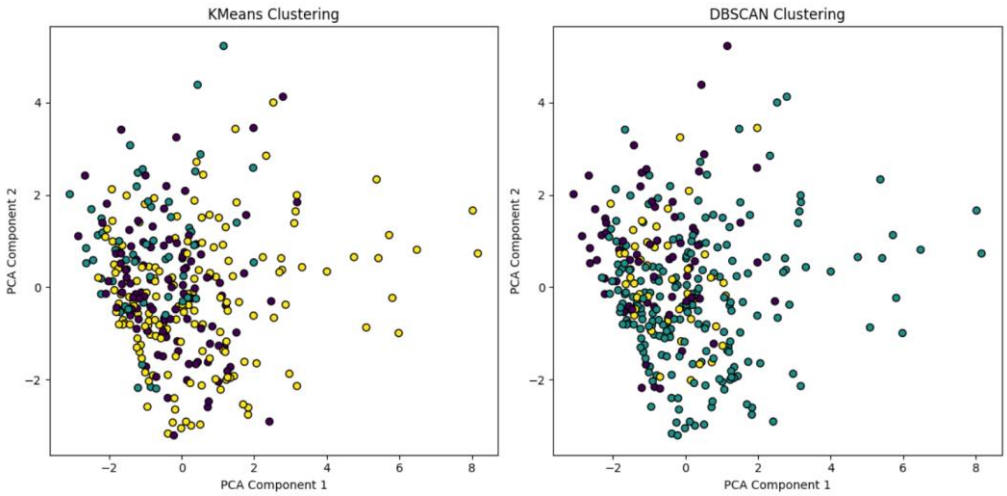
Gambar 3 Hasil Preprocessing Data

3.3 Hasil Pengolahan Data

Pada proses pengolahan kedua algoritma, masing-masing algoritma memiliki parameter yang mempengaruhi kinerja hasil klasterisasi. Pada algoritma KMeans diatur parameter n_cluster yaitu tiga (3). Sedangkan pada algoritma DBSCAN diatur parameter epsilon berjumlah 50 dan minimal samples berjumlah 10. Kemudian kedua algoritma juga akan dihitung silhouette indexnya untuk mengukur jarak antar cluster. Maka didapatkan hasil score SI untuk algoritma KMeans sebesar 0,613 sedangkan untuk algoritma DBSCAN sebesar 0,493.

Dengan menggunakan PCA dilakukan visualisasi hasil dari klasterisasi kedua algoritma seperti ditunjukkan pada gambar 4.

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST



Gambar 4 Visualisasi Hasil Clustering

4 KESIMPULAN

Berdasarkan hasil pengolahan data yang dilakukan untuk melakukan perbandingan algoritma K-Means dan DBSCAN, dalam klasterisasi data Bebras Computational Thinking Challenge dengan menggunakan data sebanyak 317 data, disimpulkan bahwa KMeans memberikan hasil yang lebih baik dibandingkan DBSCAN. Kedua metode clustering telah berhasil mengelompokkan data dengan cukup baik, tetapi hasil yang diperoleh belum dapat dijadikan acuan bahwa algoritma KMeans lebih baik dari DBSCAN. Perubahan parameter pada kedua algoritma dapat dilakukan pada penelitian selanjutnya.

REFERENSI

[1] N. P. Sutramiani, I. M. T. Arthana, P. F. Lampung, S. Aurelia, M. Fauzi, and I. W. A. S. Darma, "The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover," *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 13–24, 2024, doi: 10.20473/jisebi.10.1.13-24.

[2] M. Faisal, "Analisis Cluster untuk Pengelompokan Desa Berdasarkan Indikator Penyakit Diare," *SAINTIFIK*, vol. 5, no. 1, pp. 75–80, 2019.

[3] R. Adha, N. Nurhaliza, and U. Soleha, "Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia," *Jurnal Sains, Teknologi dan Industri*, vol. 18, no. 2, pp. 206–211, 2021, [Online]. Available: https://covid19.who.int.

[4] A. Kristianto, E. Sediyono, and K. D. Hartomo, "Implementation dbscan algorithm to clustering satellite surface temperature data in indonesia,"

Volume 5, Issue 2, pages 1–12
p-ISSN 2655-8564, e-ISSN 2685-9432
IJASST

Register: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 6, no. 2, pp. 109–118, 2020, doi: 10.26594/register.v6i2.1913.

[5] M. Z. Nasution, "PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA) DALAM PENENTUAN FAKTOR DOMINAN YANG MEMPENGARUHI PRESTASI BELAJAR SISWA (Studi Kasus : SMK Raksana 2 Medan)," *Jurnal Teknologi Informasi*, vol. 3, no. 1, 2019.