

Optimisasi Klasterisasi Nilai Ujian Nasional dengan Pendekatan Algoritma K-Means, Elbow, dan Silhouette

Allbila Rahajeng Lashiyanti^{1*}, Ibnu Rasyid Munthe², Fitri Aini Nasution³

^{1,2,3}Manajemen Informatika, Universitas Labuhan Batu, Rantauprapat, Indonesia

Email: ¹ allbilarahajeng7@gmail.com, ² ibnurasyidmunthe@gmail.com, ³ fitriaininasution689@gmail.com

Email Penulis Korespondensi: ¹ allbilarahajeng7@gmail.com

Abstrak– Penelitian ini bertujuan untuk menerapkan algoritma K-Means dalam klasterisasi data nilai Ujian Nasional (UN) dengan pemanfaatan metode optimasi Elbow dan Silhouette. Klasterisasi data nilai UN memiliki potensi untuk mengidentifikasi pola yang ada dalam hasil ujian dan membantu dalam pemahaman lebih lanjut tentang karakteristik kelompok nilai yang berbeda. Dalam penelitian ini, kami menggunakan data nilai UN sebagai input untuk algoritma K-Means. Proses klasterisasi dilakukan dengan mempertimbangkan penggunaan metode optimasi Elbow dan Silhouette. Metode Elbow digunakan untuk menentukan jumlah klaster yang optimal, sementara metode Silhouette digunakan untuk mengevaluasi kualitas klaster yang terbentuk. Hasil penelitian ini menunjukkan bahwa penerapan algoritma K-Means dengan optimasi Elbow dan Silhouette dapat menghasilkan klaster yang relevan dari data nilai UN. Penentuan jumlah klaster menggunakan metode Elbow memberikan indikasi tentang jumlah kelompok nilai yang paling sesuai, sedangkan evaluasi menggunakan metode Silhouette membantu mengukur sejauh mana kelompok-kelompok tersebut terisolasi dan konsisten. Diharapkan bahwa hasil penelitian ini akan memberikan wawasan lebih lanjut tentang penggunaan algoritma K-Means dalam klasterisasi data nilai UN. Penemuan pola dalam klaster nilai UN dapat memberikan informasi berharga bagi lembaga pendidikan dan pengambil keputusan dalam mengembangkan strategi pendidikan yang lebih efektif. Dengan menggabungkan algoritma K-Means dengan metode optimasi Elbow dan Silhouette, penelitian ini memberikan kontribusi pada pemahaman kita tentang bagaimana teknik klasterisasi dapat diterapkan secara efektif dalam analisis data nilai Ujian Nasional. Selain itu, metodologi yang digunakan dalam penelitian ini dapat memiliki implikasi lebih luas dalam analisis data pada berbagai bidang lainnya.

Kata Kunci: Klasterisasi, Data Nilai Ujian Nasional, Algoritma K-Means, Optimasi Elbow, Metode Silhouette

Abstract– This study aims to apply the K-Means algorithm in clustering National Examination (UN) score data by utilizing Elbow and Silhouette optimization methods. The clustering of UN score data has the potential to identify patterns present in exam results and aid in further understanding of the characteristics of different grade groups. In this study, we used UN value data as input for the K-Means algorithm. The clustering process is carried out taking into account the use of Elbow and Silhouette optimization methods. The Elbow method is used to determine the optimal number of clusters, while the Silhouette method is used to evaluate the quality of the clusters formed. The results of this study show that the application of the K-Means algorithm with Elbow and Silhouette optimization can generate relevant clusters from UN value data. Determination of the number of clusters using the Elbow method gives an indication of the number of groups of values that are most suitable, while evaluation using the Silhouette method helps measure the extent to which the groups are isolated and consistent. It is hoped that the results of this study will provide further insight into the use of the K-Means algorithm in clustering UN value data. Finding patterns in UN value clusters can provide valuable information for educational institutions and decision-makers in developing more effective education strategies. By combining the K-Means algorithm with Elbow and Silhouette's optimization methods, this research contributes to our understanding of how clustering techniques can be effectively applied in the analysis of National Exam score data. In addition, the methodology used in this study can have broader implications in data analysis in various other fields.

Keywords: Clustering, National Test Score Data, K-Means Algorithm, Elbow Optimization, Silhouette Method

1. PENDAHULUAN

Ujian Nasional (UN) merupakan sistem evaluasi standar pendidikan dasar yang mendukung kelulusan siswa. Sesuai dengan peraturan Pemerintah Republik Indonesia, evaluasi hasil belajar bertujuan untuk mengevaluasi pencapaian siswa lulusan nasional. Selain itu, UN juga berpengaruh untuk melanjutkan ke jenjang pendidikan berikutnya, serta dapat meningkatkan mutu pendidikan sekolah tersebut [6]. Peneliti memakai data Nilai rata-rata

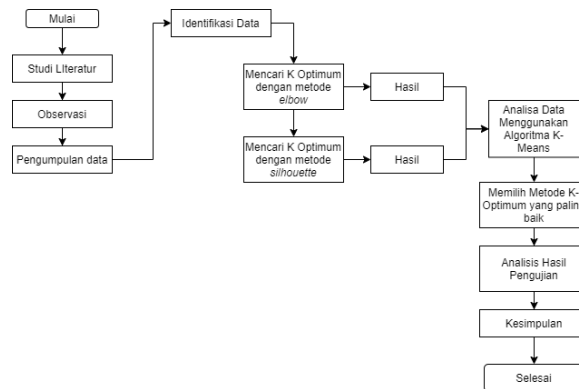
ujian nasional Sekolah Menengah Kejuruan (SMK) Tahun 2019 yang ada di Provinsi Jawa Tengah pada website resmi Kementerian Pendidikan dan Kebudayaan (KEMDIKBUD). Data dipublikasikan dalam bentuk format file Microsoft Excel (.xls).

Permasalahannya data ini masih ditampilkan secara acak dan bukan berdasarkan kategori atau peringkat tertentu, sehingga informasi menjadi kurang jelas. Untuk mengatasi masalah tersebut maka dilakukan pengelompokan dengan metode data mining.

Metode yang dipakai adalah clustering dengan algoritma k-means dengan pencarian K optimum menggunakan metode optimasi elbow dan silhouette dengan jumlah record 1536. Dengan optimasi elbow dan silhouette didapatkan jumlah K optimum yaitu K=3 untuk metode optimasi elbow dan K=2 untuk metode optimasi silhouette.

Pada penelitian ini data yang diteliti yaitu nilai rata-rata UN di Provinsi Jawa Tengah untuk Sekolah Menengah Kejuruan dengan menggunakan atribut Mata Pelajaran Bahasa Indonesia, Bahasa Inggris, Matematika dan Kompetensi. Dengan penelitian ini diharapkan mampu menambah dan memperjelas informasi dalam menunjang keputusan.

2. METODE PENELITIAN



Gambar 1. Flowchart Metodologi Penelitian

Penelitian yang dilakukan [1] menggunakan metodologi *Cross Industry Standard Process for Data Mining* (CRISP-DM). Metodologi ini memiliki enam tahap sebagai berikut.



Gambar 2. Tahapan Metode CRISP-DM [1]

- a. Fase pemahaman bisnis tujuannya untuk mencari metode terbaik yang dapat memecahkan masalah dengan menerapkan algoritma *k-means clustering* dengan metode optimasi *elbow* dan *silhouette*. [4]
- b. Fase pemahaman data Sumber data yang digunakan berasal dari data rata-rata nilai ujian nasional tingkat SMK tahun ajaran 2018-2019 yang didapat dari web <http://hasilun.puspendik.kemendikbud.go.id/>.

- c. Fase persiapan data
Fase ini melakukan persiapan data yang akan diimplementasikan kedalam data mining, seperti menghilangkan data yang kosong atau data yang tidak diperlukan dalam implementasi kedalam data mining. Memilih kolom yang sekiranya dapat dilakuakn proses pengelompokan, karena algoritma yang digunakan adalah *clustering k-means* maka data yang diperlukan hanyalah data yang berbentuk *numeric* saja, kolom yang sekiranya diperlukan bisa ditambahkan seperti kolom NO., KODE, NAMA SATUAN PENDIDIKAN.
- d. Fase pemodelan
Tahap pertama yaitu input dataset. Dataset yang sudah diinput lalu dilakukan *preprocessing data*. Setelah itu menentukan jumlah K optimum dengan menggunakan metode *elbow* dan *silhouette*. Jika K optimum sudah ditentukan maka lakukan metode *clustering* dengan algoritma *k-means*. Kemudian hasil yang didapat dibuat kesimpulan.
- e. Fase evaluasi
Jika sudah didapat hasil dari *clustering k-means* ,data hasil analisa diberi kesimpulan dengan menampilkan profil dari setiap *cluster*.
- f. Fase penyebaran
Hasil yang sudah disimpulkan dalam bentuk laporan dan dokumentasi dapat direkomendasikan kepada pembaca yang ingin mengembangkan kembali penelitian ini ke tahap yang lebih baik.

3. HASIL DAN PEMBAHASAN

Hasil dan implementasi penelitian dari perhitungan komputerisasi dan perhitungan manual, menggunakan *tools* Microsoft Excel dan RStudio.

1. Perhitungan Manual

Perhitungan manual menggunakan *tools* Microsoft Excel 2010. Data dihitung menggunakan perhitungan SSE [7] untuk mencari pusat *cluster* terbaik dengan anggota dalam ietrasi. Dataset yang digunakan berjumlah 1536 data dengan jumlah kluster terbaik adalah $K = 3$ dan iterasi yang dilakukan sebanyak 22 kali. Langkah pertama yang dilakukan adalah menentukan jumlah kluster terbaik yaitu $K = 1$ sampai dengan $K = 3$. Setelah itu menentukan pusat *centroid* [8] acak pada tiap kluster uji coba. Berikutnya menghitung jarak objek ke *centroid* menggunakan rumus *Euclidian*.

Rumus *Euclidian* :

$$D(i, j) = \sqrt{(X1i - X1j)^2 + .. + (Xki - Xkj)^2} \text{ [2]}$$

- 1) K1 (Kluster 1)

$$d(1,1) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + (x3 - y3)^2 + (x4 - y4)^2}$$

$$d(1,1) = \sqrt{(79,23 - 79,23)^2 + (53,26 - 53,26)^2 + (40,64 - 40,64)^2 + (48,4 - 48,4)^2}$$

$$d(1,1) = 0$$

- 2) K2 (Kluster 2)

$$d(1,2) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + (x3 - y3)^2 + (x4 - y4)^2}$$

$$d(1,2) = \sqrt{(79,23 - 86,47)^2 + (53,26 - 64,49)^2 + (40,64 - 64,37)^2 + (48,4 - 63,5)^2}$$

$$d(1,2) = \sqrt{52,4176 + 126,1129 + 563,1129 + 228,01}$$

$$d(1,2) = \sqrt{969,6534}$$

$$d(1,2) = 31,13926$$

3) K3 (Kluster 3)

$$d(1,3) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + (x3 - y3)^2 + (x4 - y4)^2}$$

$$d(1,3) = \sqrt{\begin{matrix} (79,23 - 75,31)^2 + (53,26 - 48,02)^2 \\ (40,64 - 37,64)^2 + (48,4 - 45,73)^2 \end{matrix}}$$

$$d(1,3) = \sqrt{15,3664 + 27,4576 + 9 + 7,1289}$$

$$d(1,3) = \sqrt{58,9529}$$

$$d(1,3) = 7,678079$$

Setelah didapat hasil jaraknya maka dapat dilakukan perhitungan manual. Seperti contoh yang didapat pada iterasi 1.

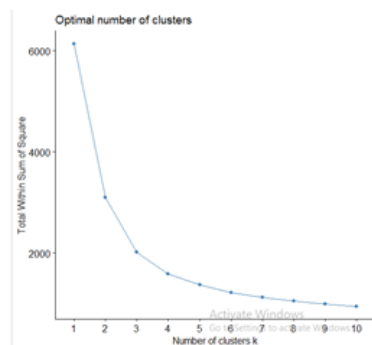
Tabel 1. Iterasi 1 dari perhitungan manual

No.	Nama Satuan Pendidikan	Kluster
1.	SMK NEGERI 1 SEMARANG	1
2.	SMK NEGERI 2 SEMARANG	2
3.	SMK NEGERI 3 SEMARANG	3
4.	SMK NEGERI 4 SEMARANG	1
5.	SMK NEGERI 5 SEMARANG	3
6.	SMK NEGERI 6 SEMARANG	1

Iterasi dilakukan sebanyak 22 kali. Dengan jumlah optimum $K = 3$. Kluster 1 terdapat 675 anggota, kluster 2 terdapat 152 anggota, kluster 3 sebesar 707 anggota.

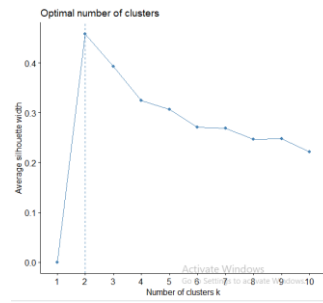
2. Implementasi Algoritma K-Means dengan RStudio

Langkah pertama yang dilakukan adalah menginput dataset kedalam RStudio menggunakan *source code*. Lalu dilakukan data preparation dengan mengambil data yang hanya bertipe numeric. Setelah dilakukan *preparation* data, dicari kluster optimal dengan menggunakan metode optimasi *elbow* dan *silhouette*. Yang didapat hasilnya sebagai berikut.



Gambar 3. Grafik Optimasi *Elbow*

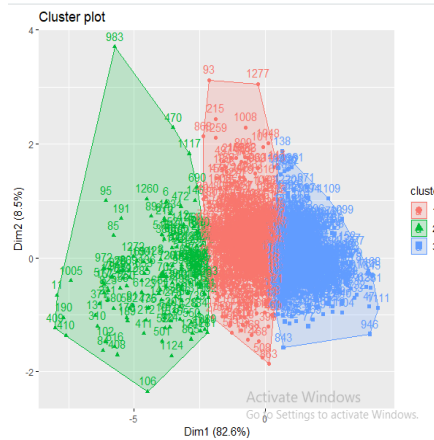
Untuk metode optimasi *elbow* dijelaskan bahwa untuk menentukan K optimum dilihat presentase grafik yang membentuk sudut. Gambar 4.2 menunjukkan presentase garis berbentuk sudut ada pada angka 3. Maka K optimumnya adalah $K = 3$.



Gambar 4. Grafik Optimasi *Silhouette*

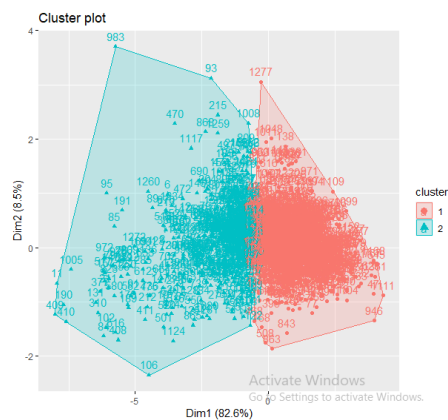
Untuk metode optimasi *silhouette* dilihat bahwa presentasi garis yang paling tinggi adalah nilai yang paling baik. Pada Gambar 4.3 presentasi yang paling tinggi ada pada nilai 2. Jadi nilai K optimum yang didapat adalah $K = 2$.

Kemudian setelah didapat jumlah K optimum terbaik, dilanjutkan dengan perhitungan *clustering* dengan algoritma *k-means*. Untuk hasil dari *clustering* menggunakan metode optimasi *elbow* didapatkan hasil Kluster 1 dengan 675 anggota, kluster 2 152 anggota dan kluster 3 sebesar 707 anggota. Dijelaskan juga hasil *sum of square* dari 3 kluster tersebut yaitu kluster 1 sebesar 776,3203, kluster 2 sebesar 516,4798, kluster 3 sebesar 717, 4831. Dengan nilai total_ss sebesar 67,2%. Perhitungan optimasi *silhouette* didapatkan hasil *clustering* dengan metode yaitu



Gambar 5 Visualisasi K

Optimum 3



Gambar 6. Visualisasi K Optimum 2

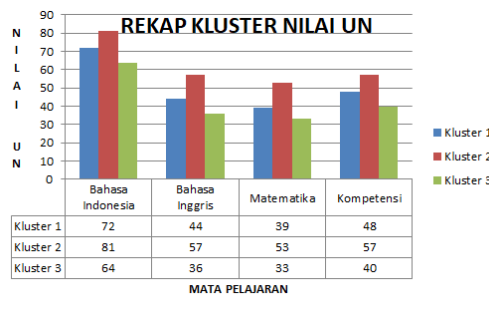
kluster 1 memiliki 1044 anggota dan kluster 2 memiliki 490 anggota. Untuk hasil *sum of square* untuk kluster 1 berjumlah 1416,049 dan kluster 2 berjumlah 1675,441 dengan nilai *between_ss* totalnya 49,6%. Setelah semua proses sudah selesai maka perhitungan akan dimasukkan kedalam variabel baru bernama 'final' lalu ditampilkan hasil visualisasinya. Untuk uraian gambar 4.5 warna merah melambangkan kluster 1 dan warna biru melambangkan kluster 2. Hasil rekap kluster dapat dilihat pada tabel berikut.

Tabel 2 Hasil Rekap Cluster

Nilai	Manual	Aplikasi
Data Uji	1536	1536
K optimasi kluster <i>elbow</i>	Lakukan siku pada K =3	Lakukan siku pada K = 3
K optimasi kluster <i>silhouette</i>	Nilai tertinggi pada K = 2	Nilai tertinggi pada K = 2

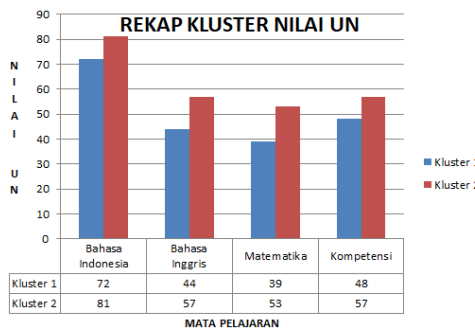
Analisa Rekap Cluster

Analisa rekap cluster ini akan berisi rata-rata dari tiap cluster yang sudah dibulatkan ke bilangan bulat. Diagram rekap cluster dapat dilihat pada gambar 4.



Gambar 7. Rekap kluster metode optimasi *elbow*

Uraian diagram diatas yaitu rekap kluster dengan nilai K optimum K = 3. Untuk nilai Bahasa Indonesia pada kluster 1 rata-rata nilainya 72, kluster 2 rata-ratanya 81, dan kluster 3 rata-ratanya 64. Bahasa Inggris untuk kluster 1 jumlah rata-ratanya 44, kluster 2 rata-ratanya 57 dan kluster 3 jumlah rata-rata 38. Matematika nilai rata-rata untuk kluster 1 yaitu 39, kluster 2 rata-ratanya 53 dan kluster 3 rata-ratanya 33. Kompetensi memiliki nilai rata-rata yaitu kluster 1 rata-rata nilainya 48, kluster 2 rata-rata nilainya 57 dan kluster 3 nilai rata-ratanya adalah 40.



Gambar 8. Rekap kluster metode optimasi *silhouette*

Uraian dari gambar 4.7 adalah rekap kluster dengan nilai K optimum $K = 2$ dengan metode optimasi *silhouette*. Pada mata pelajaran Bahasa Indonesia rata-rata untuk kluster 1 adalah 72 dan kluster berjumlah 81. Mata pelajaran Bahasa Inggris nilai rata-rata untuk kluster 1 adalah 44 dan kluster 2 yaitu 57. Mata pelajaran Matematika nilai rata-ratanya untuk kluster 1 39 dan kluster 2 53. Mata Pelajaran Kompetensi memiliki rata-rata nilai untuk kluster 1 yaitu 48 dan kluster 2 bernilai 57.

4. KESIMPULAN

Berdasarkan analisa yang sudah dilakukan diatas dapat disimpulkan : Rata-rata keseluruhan dari $K = 3$ nilai UN Bahasa Indonesia adalah 72,79906. Rata-rata nilai UN Bahasa Inggris 45,941. Rata-rata nilai UN Matematika 41,324. Rata-rata nilai UN Kompetensi 48,1947. Rata-rata keseluruhan dari $K = 2$ nilai UN Bahasa Indonesia 76,95. Rata-rata nilai UN Bahasa Inggris 33,425. Rata-rata nilai UN Matematika 45,65. Rata-rata nilai UN Kompetensi 52,54. Dari data nilai ujian nasional tingkat SMK diperoleh pengelompokan sebanyak 3 dengan menggunakan *cluster k-means* dengan metode optimasi *elbow*. Pada kluster 1 memiliki 707 anggota, kluster 2 memiliki 152 anggota. Kluster 3 memiliki 675 anggota. Dengan *cluster sum of squares* Kluster 1 717,4831. Kluster 2 516,4798. Kluster 3 776,3202. Dengan nilai *between_ss* 67.2%. Pengelompokan data nilai ujian nasional menggunakan *cluster k-means* dengan metode optimasi *silhouette* sebanyak 2 pengelompokan. Kluster 1 memiliki anggota sebanyak 1044 anggota dan kluster 2 sebanyak 490 anggota. Untuk nilai *sum of squares* kluster 1 sebesar 1416,049 dan kluster 2 sebesar 1675,411. Total *ss* sebesar 49,6%. Optimasi *cluster elbow* dan *silhouette* yang digunakan untuk metode *clustering k-means* hasilnya kedua optimasi ini sama-sama baik untuk mengelompokkan permasalahan yang ada.

REFERENSI

- 1) Budiman, I., Prahasto, T., & Christyono, Y. (2014). Data Clustering Menggunakan Metodologi CRISP-DM Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma. *Jurnal Sistem Informasi Bisnis*, 1(3), 15–16. <https://doi.org/10.21456/vol1iss3pp129-134>
- 2) Ekasetya, V. A. (2020). *OPTIMASI K-MEANS CLUSTERING MENGGUNAKAN METODE ELBOW PADA SEGMENTASI KECELAKAAN SEMARANG*.
- 3) Febrianti, F., Hafiyusholeh, M., & Asyhar, A. H. (2016). Perbandingan Pengklusteran Data Iris Menggunakan Metode K-Means Dan Fuzzy C-Means. *Jurnal Matematika "MANTIK,"* 2(1), 7. <https://doi.org/10.15642/mantik.2016.2.1.7-1>
- 4) Izzadin, F. M. (2020). *OPTIMASI JUMLAH CLUSTER K-MEANS DENGAN METODE ELBOW DAN SILHOUETTE PADA PRODUKTIVITAS TANAMAN PANGAN DI PROVINSI JAWA TENGAH TAHUN 2018*. Yogyakarta : Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.
- 5) Kementerian Pendidikan dan Budaya 2021. "LAPORAN HASIL UJIAN NASIONAL" <http://hasilun.puspendik.kemdikbud.go.id/> diakses pada 21 Juli 2020.
- 6) Muliono, R., & Sembiring, Z. (2019). *Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen*. 4(2), 2502–2714.
- 7) Muningsih, E., & Yogyakarta, A. B. S. I. (2017). Optimasi jumlah cluster k-means dengan metode elbow untuk pemetaan pelanggan. *Prosiding Seminar Nasional ELINVO, September*, 105–114.
- 8) Setyadi .F. 2020. "Metodologi CRISP-DM Beserta Contoh Kasusnya". <https://flinsetyadi.com/metodologi-crisp-dm-beserta-contoh-kasusnya/>. diakses pada 24 Juni 2021.
- 9) Simbolon, C. L., Kusumastuti, N., & Irawan, B. (2013). Clustering lulusan mahasiswa matematika fmipa untan pontianak menggunakan algoritma fuzzy c - means. *Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster)*, 02(1), 21–26.
- 10) Wikipedia 2020. "Ujian Nasional". https://id.wikipedia.org/wiki/Ujian_Nasional. diakses pada 11 Juni 2021.