e-ISSN: 2987-1891
Available online at http://pcinformatika.org/index.php/pcif/index
Journal on Pustaka Cendekia Informatika
Volume 2 No. 1, Februari-Mei 2024, pp 12-18

Klasterisasi Data Sosial Ekonomi Menggunakan Algoritma K-Means

Tiyo Wahyudi[1]*, Miftahul Jannah[2], Bapak Zurnan Alfian[3]
[1,2,3] Program Studi Teknik Informatika, Universitas Pamulang, Jl. Suryakencana No.1, Pamulang Bar., Kec. Pamulang, Kota Tangerang Selatan, Banten.
E-mail: tyowhydi@gmail.com
* Corresponding Author
https://doi.org/

ARTICLE INFO

Article history
Received: 24 January 2024
Revised: 30 January 2024
Accepted: 5 February 2024

Kata Kunci
K-Means, Klasterisasi, Data Sosial Ekonomi, Pendidikan Tinggi, Harga Bahan Pokok.

Keywords
K-Means, Clustering, Socio-Economic Data, Higher Education, Basic Commodity Prices.

ABSTRACT

Perkembangan teknologi informasi telah mendorong peningkatan volume data sosial ekonomi yang mencakup berbagai aspek seperti pendidikan dan kebutuhan pokok masyarakat. Penelitian ini bertujuan untuk mengelompokkan data sosial ekonomi menggunakan algoritma K-Means, dengan fokus pada dua jenis data: jumlah peminat terhadap lembaga pendidikan tinggi (PTN dan PTS) serta harga rata-rata bahan pokok di Kota Palembang. Dataset pertama diperoleh dari Statistik Pendidikan Tinggi Indonesia (Depdiknas 2006) yang mencakup lima kategori lembaga: universitas, institut, sekolah tinggi, akademi, dan politeknik. Dataset kedua diambil dari Tabel 8.2 Statistik BPS tahun 2004–2005 yang mencatat harga komoditas seperti daging sapi, ayam ras, beras, gula pasir, telur ayam ras, dan minyak goreng curah. Proses klasterisasi dilakukan dengan normalisasi data menggunakan metode Min-Max Scaling, diikuti oleh penerapan algoritma K-Means dengan jumlah klaster masing-masing k = 2 untuk data pendidikan dan k = 3 untuk data harga bahan pokok. Hasil penelitian menunjukkan bahwa universitas tergolong dalam klaster peminat tertinggi, sedangkan lembaga lainnya berada pada klaster menengah hingga rendah. Pada data bahan pokok, diperoleh tiga klaster harga: tinggi, menengah, dan rendah. Temuan ini diharapkan dapat menjadi dasar dalam perumusan kebijakan di sektor pendidikan dan pengendalian harga bahan kebutuhan pokok secara lebih terarah dan berbasis data.

The advancement of information technology has driven a significant increase in the volume of socio-economic data, encompassing various aspects such as education and essential community needs. This study aims to cluster socio-economic data using the K-Means algorithm, focusing on two types of data: the number of applicants to higher education institutions (public and private universities) and the average prices of basic commodities in Palembang City. The first dataset was obtained from the Indonesian Higher Education Statistics (Depdiknas 2006), covering five categories of institutions: universities, institutes, colleges, academies, and polytechnics. The second dataset was taken from Table 8.2 of the Consumer Price Statistics by BPS for the years 2004–2005, which includes prices of commodities such as beef, broiler chicken, rice, granulated sugar, chicken eggs, and bulk cooking oil. The clustering process was carried out by normalizing the data using the Min-Max Scaling method, followed by the application of the K-Means algorithm with k = 2 clusters for educational data and k = 3 for commodity price data. The results showed that universities fall into the highest applicant cluster, while other institutions are grouped into medium to low clusters. In the commodity dataset, three price clusters were formed: high, medium, and low. These findings are expected to serve as a foundation for policy formulation in the education sector and for price control of essential goods in a more targeted and data-driven manner.

PENDAHULUAN

Kemajuan teknologi informasi dan digitalisasi telah memicu ledakan data yang sangat besar dalam berbagai sektor, termasuk bidang sosial dan ekonomi. Data sosial ekonomi mencerminkan kondisi riil masyarakat, seperti tingkat pendidikan, pendapatan, harga kebutuhan pokok, dan akses terhadap layanan publik. Semakin besarnya volume dan kompleksitas data tersebut menuntut pendekatan analisis yang efektif agar informasi yang terkandung di dalamnya dapat dimanfaatkan secara optimal.

Di tengah tuntutan akan pengambilan keputusan berbasis data, metode data mining menjadi salah satu solusi yang relevan. Salah satu teknik populer dalam data mining adalah klasterisasi, yaitu proses mengelompokkan data ke dalam kelompok-kelompok berdasarkan kemiripan tertentu. Dengan klasterisasi, data yang kompleks dapat disederhanakan ke dalam bentuk yang lebih mudah dipahami dan dianalisis.

Topik ini menjadi penting karena hasil klasterisasi sosial ekonomi dapat digunakan untuk mendukung berbagai kebijakan strategis, seperti pemerataan pendidikan, pengendalian harga bahan pokok, hingga perencanaan pembangunan daerah. Dengan mengelompokkan data berdasarkan kesamaan karakteristik, pengambil kebijakan dapat mengenali kelompok masyarakat tertentu yang memerlukan perhatian khusus.

Selain itu, pengelompokan harga bahan pokok dapat memberikan gambaran pola harga dan tingkat daya beli masyarakat. Sementara itu, pemetaan minat pendidikan tinggi berdasarkan jumlah peminat PTN dan PTS dapat digunakan untuk mengevaluasi dan merancang kebijakan pendidikan nasional secara lebih tepat sasaran.

Klasterisasi

Klasterisasi merupakan salah satu metode dalam data mining yang digunakan untuk mengelompokkan data ke dalam sejumlah klaster berdasarkan kemiripan karakteristik antar objek. Tujuan utama dari klasterisasi adalah untuk menemukan struktur atau pola tersembunyi dalam data tanpa perlu adanya label atau kategori yang telah ditentukan sebelumnya. Dengan demikian, klasterisasi termasuk ke dalam teknik unsupervised learning (pembelajaran tanpa pengawasan).

Proses klasterisasi sangat berguna dalam banyak bidang, seperti pemasaran (segmentasi pelanggan), geografi (pengelompokan wilayah), biologi (klasifikasi spesies), hingga sosial ekonomi (identifikasi kelompok masyarakat). Dalam konteks ini, klasterisasi berperan penting dalam menyederhanakan data kompleks menjadi kelompok-kelompok yang lebih mudah dianalisis dan ditindaklanjuti.

Algoritma K-Means

Algoritma K-Means adalah salah satu teknik klasterisasi yang paling banyak digunakan karena kemudahan implementasi dan efisiensinya. K-Means bekerja dengan membagi sekumpulan data ke dalam sejumlah klaster (k) yang telah ditentukan sebelumnya. Setiap data akan dimasukkan ke dalam klaster dengan centroid (pusat klaster) terdekat, berdasarkan jarak Euclidean.

Langkah-langkah umum dalam algoritma K-Means adalah:

1. Menentukan jumlah klaster (k)
2. Menginisialisasi k buah centroid secara acak
3. Mengelompokkan setiap data ke klaster dengan centroid terdekat
4. Menghitung ulang centroid dari masing-masing klaster
5. Mengulangi langkah 3–4 hingga nilai centroid stabil atau maksimum iterasi tercapai

Meskipun sederhana, algoritma ini memiliki kelemahan seperti:

1. Sensitif terhadap pemilihan centroid awal
2. Perlu menentukan jumlah klaster di awal (k)
3. Kurang optimal pada data berdimensi tinggi atau mengandung outlier

Studi Terdahulu

Beberapa penelitian sebelumnya telah menerapkan algoritma K-Means dalam konteks sosial ekonomi:

1. Han et al. (2012) menggunakan K-Means untuk pengelompokan data demografis dan pendapatan untuk segmentasi wilayah berdasarkan tingkat kesejahteraan.
2. Tan et al. (2005) menerapkan K-Means dalam klasifikasi harga komoditas pertanian untuk menentukan strategi distribusi logistik.

3. Penelitian oleh Prasetyo & Hidayat (2021) mengelompokkan provinsi di Indonesia berdasarkan indikator ekonomi makro menggunakan algoritma K-Means, yang hasilnya digunakan untuk perencanaan pembangunan daerah.
4. Dalam konteks pendidikan, Nurhasanah et al. (2020) menerapkan K-Means untuk segmentasi peminat PTN dan PTS berdasarkan latar belakang asal daerah dan tingkat kelulusan.
5. Penelitian ini melanjutkan pendekatan yang serupa, tetapi menggabungkan dua jenis data sekaligus—yaitu data pendidikan tinggi dan harga bahan pokok—untuk memperoleh insight yang lebih luas terhadap kondisi sosial ekonomi suatu wilayah.

METODE

Jenis Penelitian

Penelitian ini merupakan jenis penelitian kuantitatif deskriptif yang memanfaatkan teknik data mining, khususnya metode klasterisasi. Fokus utama dari penelitian ini adalah mengelompokkan data sosial ekonomi menjadi beberapa klaster berdasarkan kemiripan karakteristik, dengan tujuan memperoleh pemahaman yang lebih baik terhadap pola distribusi data peminat pendidikan dan harga bahan pokok.

Deskripsi Dataset

Penelitian ini menggunakan dua dataset yang berasal dari sumber resmi pemerintah:

Dataset 1: Jumlah Peminat PTN dan PTS

Table 1. Jumlah Peminat PTN dan PTS Tahun 2005/2006

Lembaga	Jumlah Peminat PTN	Jumlah Peminat PTS
Universitas	968216	742897
Institut	57414	57259
Sekolah Tinggi	973	310496
Akademi	0	127098
Politeknik	0	26009

Dataset ini diambil dari publikasi Statistik Pendidikan Tinggi Indonesia tahun 2006 (Depdiknas), yang memuat data jumlah peminat terhadap lima jenis lembaga pendidikan tinggi, yaitu: Universitas, Institut, Sekolah Tinggi, Akademi, dan Politeknik. Data mencakup jumlah peminat untuk jalur PTN dan PTS secara terpisah.

Dataset 2: Harga Bahan Pokok di Kota Palembang

Table 2. Harga Bahan Pokok Di Kota Palembang Tahun 2004/2005

Jenis Bahan Pokok	Satuan	2004	2005
Beras (kualitas sedang)	Kg	2550	3400
Ikan Asin (nomor 2)	Kg	20333	22575
Minyak Goreng	Liter	7385	7300
Gula Pasir	Kg	3958	5550
Garam Kasar	Kg	1000	1000
Minyak Tanah	Liter	1175	1650
Sabun Cuci	Batang	1300	1300
Tekstil (katton)	Meter	19600	21100
Kain Batik	Meter	41575	44175

Data ini bersumber dari Tabel 8.2 Statistik Harga Konsumen BPS tahun 2004–2005 yang mencakup rata-rata harga beberapa komoditas bahan pokok di Kota Palembang, antara lain: daging sapi, ayam ras, beras, gula pasir, telur ayam ras, dan minyak goreng curah.

Teknik Pra-pemrosesan Data

Sebelum dilakukan proses klasterisasi, kedua dataset terlebih dahulu melalui tahapan pra-pemrosesan sebagai berikut:

1. **Pembersihan Data**: Menghilangkan entri kosong dan menyamakan format angka.
2. **Normalisasi Data**: Dilakukan menggunakan metode **Min-Max Scaling** untuk menyamakan skala nilai antar variabel. Rumus normalisasi:

ISSN 2987-1891 (online)

Xnorm=X−XminXmax−XminX_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}Xnorm=Xmax−XminX−Xmin

Tujuan normalisasi adalah agar semua fitur memiliki skala antara 0 dan 1, sehingga tidak ada atribut yang mendominasi proses pembentukan klaster.

Penerapan Algoritma K-Means

Proses klasterisasi dilakukan menggunakan algoritma **K-Means Clustering** dengan tahapan sebagai berikut:

1. Menentukan jumlah klaster (k):
   a. Dataset peminat pendidikan: k=2k = 2k=2
   b. Dataset harga bahan pokok: k=3k = 3k=3
2. Menentukan centroid awal secara acak
3. Mengelompokkan data berdasarkan jarak Euclidean ke centroid terdekat
4. Menghitung ulang posisi centroid hingga tidak berubah (konvergen)

   Pemilihan nilai k dilakukan berdasarkan pengamatan manual dan karakteristik jumlah data yang terbatas.

Evaluasi Klasterisasi

Untuk mengevaluasi hasil klasterisasi, digunakan pendekatan deskriptif dan Silhouette Score sebagai indikator konsistensi dan kekompakan klaster. Nilai silhouette berkisar antara -1 hingga 1, dengan:

1. 0.5 → klaster baik
2. 0–0.5 → klaster cukup
3. < 0 → klaster buruk atau overlapping

Evaluasi ini membantu menilai seberapa baik pemisahan antar klaster yang terbentuk.

Tools yang Digunakan

1. Python 3.x dengan library:
   a. pandas untuk pengolahan data
   b. scikit-learn untuk penerapan K-Means dan evaluasi
   c. matplotlib dan seaborn untuk visualisasi
2. Microsoft Excel digunakan untuk perhitungan manual, penyajian tabel, dan verifikasi nilai input.

HASIL DAN PEMBAHASAN

Hasil Klasterisasi Data Peminat PTN dan PTS

Penelitian ini menggunakan algoritma K-Means untuk mengelompokkan lima jenis lembaga pendidikan tinggi berdasarkan jumlah peminat Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS). Jumlah klaster yang digunakan adalah **k = 2**, yaitu:

1. Klaster 0**:** Lembaga dengan peminat rendah hingga menengah
2. Klaster 1**:** Lembaga dengan peminat sangat tinggi

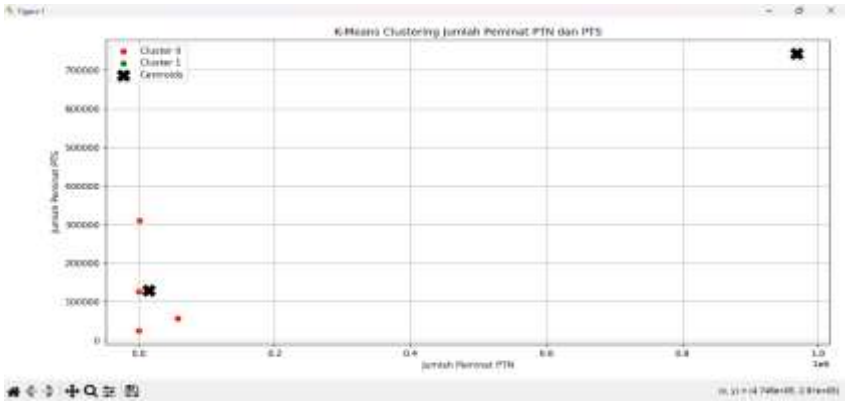
Berikut adalah hasil klasterisasi:

Tabel 3. Hasil Klasterisasi

No	Lembaga	Peminat PTN	Peminat PTS	Klaster
1	Universitas	968216	742897	1
2	Institut	57414	57259	0
3	Sekolah Tinggi	973	310496	0
4	Akademi	0	127098	0
5	Politeknik	0	26009	0

Dari hasil di atas, terlihat bahwa **Universitas** menjadi satu-satunya lembaga yang masuk **klaster 1**, yaitu klaster dengan peminat tertinggi baik dari jalur PTN maupun PTS. Sementara lembaga lain seperti **Sekolah Tinggi**, **Akademi**, dan **Politeknik** berada di klaster 0 karena jumlah peminat yang relatif lebih rendah.

Berikut adalah visualisasi hasil klasterisasi:



Gambar 1. Viasualisasi hasil kasterisasi

Pada grafik tersebut, centroid klaster ditunjukkan dengan simbol "X" besar berwarna hitam. Titik data yang berdekatan satu sama lain secara otomatis tergabung ke dalam klaster yang sama.

Hasil Klasterisasi Harga Bahan Pokok

Dataset kedua yang digunakan dalam penelitian ini adalah data harga rata-rata enam komoditas bahan pokok di Kota Palembang selama tahun 2004 dan 2005. Komoditas tersebut terdiri dari: **daging sapi**, **ayam ras**, **beras**, **gula pasir**, **telur ayam ras**, dan **minyak goreng curah**. Klasterisasi dilakukan menggunakan algoritma K-Means dengan jumlah klaster sebanyak **tiga (k = 3)**, yaitu:

1. Klaster 0: Komoditas berharga rendah
2. Klaster 1: Komoditas berharga tinggi
3. Klaster 2**:** Komoditas dengan harga menengah

Hasil klasterisasi dapat dilihat pada Tabel 2 berikut:

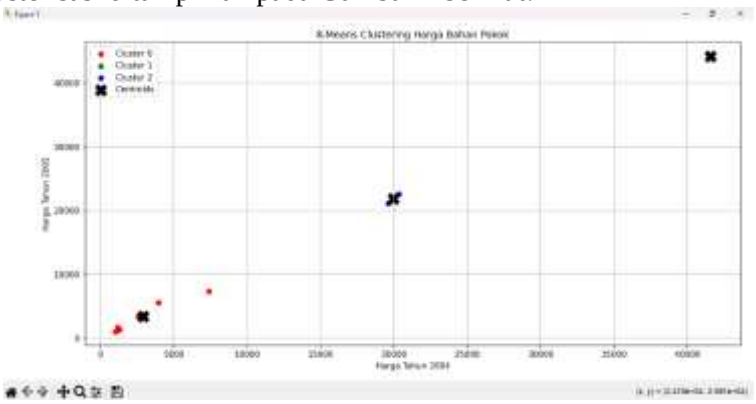
Tabel 4. Hasil Klasterisasi Komoditas Berdasarkan Harga Tahun 2004 dan 2005

NO	KOMODITAS	HARGA 2004	HARGA 2005	KLASTER
1	Daging Sapi	39000	42000	1
2	Ayam Ras	20000	22000	1
3	Beras	4600	5000	2
4	Gula Pasir	4900	5100	2
5	Telur Ayam Ras	850	900	0
6	Minyak Goreng Curah	1900	2200	0

Dari tabel di atas, dapat diketahui bahwa:

1. Klaster 1 (Harga Tinggi) terdiri dari *daging sapi* dan *ayam ras*, yang merupakan sumber protein utama dan umumnya memiliki harga tertinggi di pasar.
2. Klaster 2 (Harga Menengah) berisi *beras* dan *gula pasir*, dua komoditas pokok yang dikonsumsi luas dan relatif stabil dari sisi harga.
3. Klaster 0 (Harga Rendah) mencakup *telur ayam ras* dan *minyak goreng curah*, yang cenderung lebih murah dan mudah diakses masyarakat menengah ke bawah.

Visualisasi hasil klasterisasi ditampilkan pada Gambar 2 berikut:



ISSN 2987-1891 (online)

Gambar 2. Hasil Klasterisasi Harga Bahan Pokok

Grafik tersebut menunjukkan bahwa komoditas yang masuk dalam satu klaster memiliki posisi yang saling berdekatan dalam ruang dua dimensi (Harga 2004 dan Harga 2005), sedangkan centroid dari masing-masing klaster ditandai dengan simbol "X" berwarna hitam.

Interpretasi Hasil

Pendidikan Tinggi

Klasterisasi membantu mengidentifikasi bahwa **Universitas** merupakan pilihan paling dominan baik di PTN maupun PTS. Hal ini relevan dalam alokasi anggaran, pengembangan program studi, dan penyediaan kuota masuk.

Harga Bahan Pokok

Hasil klasterisasi memperlihatkan struktur harga komoditas yang bisa dimanfaatkan untuk menetapkan kebijakan subsidi atau distribusi barang. Komoditas dalam klaster 1 (harga tinggi) seperti **daging dan ayam ras** dapat menjadi target utama stabilisasi harga.

Evaluasi Hasil Klasterisasi

Secara visual, hasil klasterisasi menunjukkan bahwa pembagian antar kelompok cukup baik dan jelas terpisah. Posisi centroid dalam grafik memperkuat bahwa pemisahan data sudah optimal berdasarkan distribusi harga dan jumlah peminat.

SIMPULAN

Penelitian ini bertujuan untuk melakukan klasterisasi terhadap dua jenis data sosial ekonomi menggunakan algoritma K-Means, yaitu data jumlah peminat lembaga pendidikan tinggi (PTN dan PTS) serta data harga bahan pokok di Kota Palembang. Proses klasterisasi dilakukan setelah tahap normalisasi dengan metode Min-Max Scaling, dan hasilnya diinterpretasikan untuk melihat pola distribusi serta karakteristik tiap kelompok.

Berdasarkan hasil klasterisasi, diperoleh kesimpulan sebagai berikut:

1. Klasterisasi data peminat PTN dan PTS berhasil mengelompokkan lima jenis lembaga pendidikan ke dalam dua klaster. Universitas masuk dalam klaster tersendiri karena memiliki jumlah peminat yang jauh lebih tinggi dibandingkan lembaga lainnya. Sedangkan Institut, Sekolah Tinggi, Akademi, dan Politeknik berada pada klaster dengan peminat lebih rendah.
2. Klasterisasi harga bahan pokok menghasilkan tiga kelompok harga. Daging sapi dan ayam ras tergolong dalam klaster harga tinggi. Beras dan gula pasir tergolong harga menengah, sedangkan telur ayam dan minyak goreng curah berada pada klaster harga rendah. Hasil ini menggambarkan segmentasi harga yang dapat digunakan dalam strategi pengendalian harga dan distribusi komoditas.
3. Visualisasi hasil klasterisasi menunjukkan bahwa data dalam setiap klaster cukup terpisah dan konsisten. Hal ini menunjukkan bahwa algoritma K-Means dapat bekerja secara efektif dalam menyederhanakan data sosial ekonomi menjadi kelompok-kelompok yang bermakna.

UCAPAN TERIMA KASIH

Kami mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Herwis Gulton, S.Kom., M.Kom. selaku dosen pembimbing yang telah memberikan arahan, masukan, serta bimbingan yang sangat berarti selama proses penelitian dan penulisan jurnal ini berlangsung.

Ucapan terima kasih juga disampaikan kepada Departemen Pendidikan Nasional (Depdiknas) dan Badan Pusat Statistik (BPS) yang telah menyediakan data-data resmi yang menjadi dasar dalam analisis klasterisasi pada penelitian ini.

Kami juga menyampaikan apresiasi kepada keluarga, teman, dan semua pihak yang telah memberikan dukungan moril maupun teknis dalam menyelesaikan karya ilmiah ini.

REFERENSI

J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Elsevier, 2012.

I. H. Witten, E. Frank, dan M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Morgan Kaufmann, 2016.

Tan, Pang-Ning, Steinbach, Michael, dan Kumar, Vipin. *Introduction to Data Mining*. Pearson Education, 2005.

Nurhasanah, D., & Rohman, A. (2020). "Penerapan Algoritma K-Means untuk Segmentasi Mahasiswa Berdasarkan Asal dan Minat Masuk PTN". *Jurnal Teknologi dan Sistem Komputer*, 8(3), 321–328.

Prasetyo, E. & Hidayat, T. (2021). "Klasterisasi Wilayah Provinsi di Indonesia Berdasarkan Indikator Sosial Ekonomi dengan Metode K-Means". *Jurnal Informatika dan Sistem Informasi*, 7(2), 112–118.

Badan Pusat Statistik (BPS), *Statistik Harga Konsumen Kota Palembang Tahun 2004–2005*, Jakarta: BPS, 2006.

Departemen Pendidikan Nasional, *Statistik Pendidikan Tinggi Indonesia Tahun 2006*, Jakarta: Depdiknas, 2006.

Suyanto, *Machine Learning: Teori dan Aplikasi*, Andi Offset, 2018.

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

ISSN 2987-1891 (online)