JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 4, Desember 2023, Pp. 1186-1195

IMPLEMENTASI PRINCIPAL COMPONENT ANALYSIS PADA K-MEANS UNTUK KLASTERISASI TINGKAT PENDIDIKAN PENDUDUK KABUPATEN SEMARANG

Syarafina Dewi*[1)], Magdalena A. Ineke Pakereng [2)]
1. Universitas Kristen Satya Wacana, Indonesia
2. Universitas Kristen Satya Wacana, Indonesia

Article Info
Kata Kunci: K-Means; *Clustering*; Reduksi dimensi; PCA; Tingkat pendidikan

Keywords: *K-Means; Clustering; Dimension reduction; PCA; Level of Education*

Article history:
Received 9 May 2023
Revised 23 May 2023
Accepted 6 June 2023
Available online 1 December 2023

DOI :
https://doi.org/10.29100/jipi.v8i4.4101

*Corresponding author.
Syarafina Dewi
E-mail address:
672019063@student.uksw.edu

ABSTRAK
Indonesia diperkirakan akan menghadapi bonus demografi pada tahun 2035. Sebagai upaya untuk menghadapi hal tersebut, negara Indonesia membutuhkan sumber daya manusia yang berkualitas salah satunya melalui pendidikan. Namun ditinjau dari Angka Partisipasi Kasar (APK) untuk tingkat pendidikan tinggi di Kabupaten Semarang masih tergolong rendah, yaitu 19,25%. Oleh karena itu, penelitian ini bertujuan untuk mengelompokkan tingkat pendidikan berdasarkan jenis kelamin, umur, dan status individu dalam keluarga dengan menggunakan algoritma K-Means dan reduksi dimensi melalui metode *Principal Component Analysis* (PCA). Penelitian ini dilakukan dengan mengolah data hasil Survei Sosial Ekonomi Nasional 2021 Maret (KOR) Wilayah Kabupaten Semarang yang bersumber dari data Badan Pusat Statistik (BPS). Setelah dilakukan analisis, dihasilkan dua komponen utama dengan proporsi kumulatif mencapai 70%. Kemudian, setelah dilakukan reduksi dimensi dengan menggunakan metode PCA, analisis klaster dilakukan dengan menggunakan algoritma K-means yang menghasilkan empat kelompok dengan karakteristik yang berbeda untuk masing-masing klaster.

ABSTRACT
Indonesia is expected to face a demographic bonus in 2035. As an effort to deal with this, the Indonesian state needs quality human resources, one of which is through education. However, in terms of the gross enrollment rate (APK) for the higher education level in Semarang Regency, it is still relatively low, namely 19.25%. Therefore, this research was conducted to classify the level of education of the people in Semarang Regency based on gender, age, and individual status in the family using the K-Means algorithm and dimension reduction through the Principal Component Analysis (PCA) method. This research was conducted by processing data from the results of the March 2021 National Socioeconomic Survey for the Semarang Regency area which was sourced from data from the Central Statistics Agency (BPS). After analysis, two main components were produced with a cumulative proportion of 70%. Then, after dimension reduction using the PCA method, cluster analysis is performed using the K-means algorithm which produces four groups with different characteristics for each cluster.

I. PENDAHULUAN

PADA tahun 2035, diperkirakan negara Indonesia akan menghadapi bonus demografi berupa peningkatan jumlah masyarakat usia produktif [1]. Peningkatan masyarakat usia produktif bisa berdampak baik pada pertumbuhan ekonomi sebuah negara karena tersedianya banyak sumber daya manusia. Namun fenomena seperti ini membutuhkan berbagai perubahan dan penyesuaian terhadap pola kehidupan masyarakat itu sendiri.

Sejalan dengan hal tersebut, dibutuhkan persiapan dan peningkatan kualitas sumber daya manusia dalam berbagai bidang. Di Indonesia beberapa upaya sudah dilakukan, salah satunya kebijakan wajib belajar 12 tahun. Namun jika dilihat dari Angka Partisipasi Kasar (APK) tahun 2021 di wilayah Kabupaten Semarang, terdapat 106,16% masyarakat tingkat sekolah dasar, 98% pada tingkat sekolah menengah pertama, 80,91% pada tingkat sekolah menengah atas, dan 19,25% pada tingkat perguruan tinggi. Terdapat lebih dari 100% masyarakat pada tingkat sekolah dasar dikarenakan terdapat siswa pada usia yang terlalu dini, terlambat masuk sekolah, atau yang mengulang kelas. Di sisi lain, hanya terdapat 19,25% masyarakat yang mencapai tingkat perguruan tinggi [2], [3].

Berdasarkan permasalahan tersebut, diperlukan pengolahan data untuk mengetahui bagaimana kondisi tingkat pendidikan masyarakat berdasarkan kriteria tertentu. Untuk mendapatkan hasil yang sesuai, penelitian ini dilakukan dengan mengolah data hasil Survei Sosial Ekonomi Nasional 2021 Maret (KOR) Wilayah Kabupaten Semarang yang bersumber dari data Badan Pusat Statistik (BPS). Untuk melakukan penelitian tersebut, diperlukan data dengan atribut pendidikan terakhir, jenis kelamin, umur, dan status individu dalam keluarga. Pengolahan data dilakukan menggunakan teknik *clustering* dan reduksi dimensi. *Clustering* adalah sebuah metode *data mining* yang bertujuan untuk membagi objek dalam kumpulan data menjadi beberapa kelompok. Tujuan dari *clustering* adalah mengelompokkan objek-objek yang memiliki kemiripan ke dalam kategori yang sama, sedangkan objek-objek yang memiliki perbedaan akan dikelompokkan ke dalam kategori yang berbeda [4]. Lalu reduksi dimensi dilakukan untuk mengurangi jumlah fitur atau dimensi dari suatu kumpulan data tanpa menghilangkan informasi penting yang terkandung di dalamnya [5].

Pada penelitian ini, teknik *clustering* yang digunakan adalah algoritma K-Means. Algoritma ini populer digunakan dalam melakukan *clustering* karena sederhana dan mudah diterapkan. Namun beberapa algoritma *clustering* seperti K-Means sering mengalami kendala ketika diaplikasikan pada data dengan dimensi atau fitur yang tinggi. Terdapat beberapa kendala yang muncul seperti penurunan akurasi klasifikasi, kualitas klaster yang buruk, dan membutuhkan waktu yang lama dalam komputasi. Untuk menjaga kinerja algoritma tetap optimal, salah satu cara yang dapat dilakukan adalah dengan melakukan reduksi dimensi [6].

Reduksi dimensi dapat dilakukan menggunakan dua metode, yaitu metode pemilihan fitur dan metode ekstraksi fitur [7]. Metode yang digunakan dalam penelitian ini adalah metode ekstraksi fitur, yaitu melakukan ekstraksi fitur baru dari kumpulan data awal. Salah satu teknik ekstraksi fitur yang dapat digunakan adalah *Principal Component Analysis* (PCA). PCA adalah metode yang digunakan untuk menyederhanakan kumpulan data dengan cara mentransformasi linier sehingga terbentuk sistem koordinat baru dengan varians minimum. PCA dapat digunakan untuk mereduksi dimensi data yang lebih rendah dengan risiko kehilangan informasi yang sangat kecil [8].

Pengolahan data dilakukan dengan menggunakan bahasa pemrograman Python dan beberapa *library* seperti numpy dan pandas untuk memproses data. Lalu *library* sklearn digunakan dalam proses *clustering*, standarisasi data, PCA, dan evaluasi klaster. Sedangkan untuk melakukan visualisasi data dapat menggunakan *library* matplotlib dan seaborn.

(Revisi 1) Penelitian yang dilakukan oleh Kurniawan tahun 2021 berjudul "Klasterisasi Tingkat Pendidikan di DKI Jakarta pada Tingkat Kecamatan Menggunakan Algoritma K-Means" menggunakan algoritma K-Means untuk melakukan clustering data tingkat pendidikan di wilayah DKI Jakarta. Hasil dari penelitian tersebut dapat mengelompokkan wilayah kecamatan ke dalam beberapa klaster, selain itu data dapat digunakan untuk mengidentifikasi wilayah yang memerlukan perhatian khusus dalam peningkatan kualitas pendidikan [9]. Serupa dengan penelitian sebelumnya, Wulandari dan Sumarah tahun 2021 yang berjudul "Klaster Rata-rata Lama Sekolah (RLS) Menurut Jenis Kelamin di Provinsi Jawa Tengah dengan K-Means" juga menggunakan algoritma K-Means untuk melakukan pengelompokan data. Hasil dari penelitian tersebut terbagi menjadi dua klaster yaitu klaster tinggi dan klaster rendah [10].

Studi lain yang berjudul "Penerapan Principal Component Analysis (PCA) untuk Reduksi Dimensi pada Proses Clustering Data Produksi Pertanian di Kabupaten Bojonegoro" menggunakan kombinasi antara algoritma K-Means dan teknik PCA. Hasil penelitian ini menunjukkan bahwa penggunaan metode PCA dalam reduksi dimensi data produksi pertanian di Kabupaten Bojonegoro dapat meningkatkan efisiensi analisis dan hasil clustering. Dengan mengurangi dimensi data, analisis dapat dilakukan dengan lebih efisien dan hasil clustering dapat lebih akurat dan mudah diinterpretasikan [12]. Lalu penelitian oleh Harahap et al. tahun 2022 yang berjudul "Analisis Pemasaran Bisnis dengan Data Science : Segmentasi Kepribadian Pelanggan berdasarkan Algoritma K-Means Clustering" memanfaatkan kombinasi yang sama antara K-Means dan PCA pada data pelanggan untuk melakukan segmentasi terhadap karakteristik pelanggan [13].

Metode PCA dapat membantu dalam mengekstraksi informasi penting dari dataset yang kompleks dengan mengurangi dimensi data. Umumnya PCA dan K-Means digunakan bersama untuk mengelompokkan data dengan dimensi yang tinggi. PCA berguna untuk mengurangi dimensi data yang kompleks menjadi lebih kecil sehingga K-Means dapat bekerja secara efektif pada dataset yang baru. Oleh sebab itu, akan dilakukan penelitian yang berjudul "Implementasi Principal Component Analysis pada K-Means untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang".

(Revisi 2) Meninjau dari beberapa penelitian sebelumnya, dimana PCA tidak diterapkan pada K-Means membuat karakteristik setiap variabel dalam dimensi data yang tinggi belum terlihat dengan jelas. Hal ini dapat menyulitkan dalam memahami pola dan perbedaan yang signifikan antara variabel-variabel pendidikan. Namun, dengan menerapkan PCA, penelitian ini dapat mengatasi masalah tersebut dengan memproyeksikan variabel-variabel

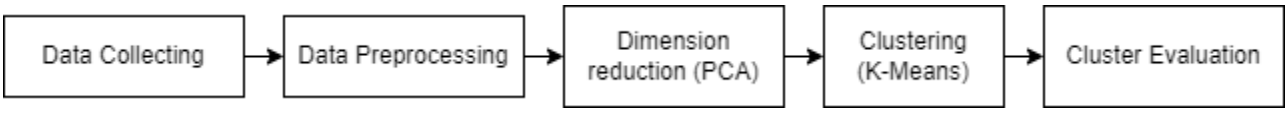
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 4, Desember 2023, Pp. 1186-1195

tersebut ke ruang dimensi yang lebih rendah. Dengan demikian, karakteristik dari setiap variabel dapat dibedakan dengan jelas, memberikan pemahaman yang lebih baik tentang hubungan dan pola tingkat pendidikan penduduk di Kabupaten Semarang

(Revisi 3) Dalam konteks kebaruan, penelitian ini memberikan kontribusi dengan mengaplikasikan teknik *Principal Component Analysis* (PCA) pada metode klasterisasi K-Means untuk menganalisis dan mengklasterisasi tingkat pendidikan penduduk di Kabupaten Semarang. Dengan menggunakan PCA sebagai metode reduksi dimensi, penelitian ini dapat mengidentifikasi kombinasi linear dari variabel pendidikan yang paling berpengaruh dalam membentuk kelompok-kelompok penduduk dengan tingkat pendidikan yang serupa. Pendekatan ini dapat membantu dalam pemahaman yang lebih baik tentang struktur dan pola yang mendasari tingkat pendidikan, serta memfasilitasi pengambilan keputusan berbasis data untuk pengembangan program dan kebijakan pendidikan di Kabupaten Semarang. Dengan demikian, penelitian ini memberikan kontribusi baru dalam memanfaatkan PCA dalam konteks klasterisasi tingkat pendidikan penduduk.

II. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini dijelaskan pada Gambar 1.



Gambar. 1. Metode Penelitian

Gambar 1 menjelaskan tahap yang dilakukan yaitu mulai dari (1) tahap pengumpulan data, (2) tahap prapemrosesan data, (3) tahap reduksi dimensi, (4) tahap *clustering*, dan (5) tahap evaluasi.

A. Pengumpulan Data

Tahap pengumpulan data merupakan langkah pertama dalam penelitian yang dilakukan dengan cara mengumpulkan data sesuai dengan tujuan penelitian. Data yang digunakan bersumber dari Survei Sosial Ekonomi Nasional 2021 Maret (KOR) Wilayah Kabupaten Semarang. Data tersebut didapat dari *website* Badan Pusat Statistik (BPS) yaitu https://pst.bps.go.id/. Jumlah observasi dari *dataset* sebanyak 3059 observasi dan jumlah atribut dari *dataset* sebanyak 328 atribut.

(Revisi 4)

TABEL I
SAMPEL DATA YANG DIGUNAKAN

Pendidikan Tertinggi	Status dalam Keluarga	Jenis Kelamin	Umur
3	1	1	62
4	2	2	61
11	3	1	18
3	1	1	57
8	2	2	51

B. Prapemrosesan Data

Prapemrosesan data dilakukan sebelum *dataset* diterapkan ke dalam model. Tujuannya untuk mengurangi kebisingan data agar menjadi lebih bersih sehingga didapat hasil yang maksimal. Proses prapemrosesan data meliputi pembersihan data, transformasi data, integrasi data, dan pemilihan fitur [9]. Proses pembersihan data dilakukan dengan mengambil 4 atribut, beberapa data yang tidak memenuhi syarat akan dihapus. Atribut yang digunakan yaitu pendidikan terakhir, jenis kelamin, umur, dan status individu dalam keluarga. Pengelompokan tingkat pendidikan juga perlu dilakukan agar memudahkan proses analisis. Pengelompokan dimulai dari pendidikan dasar yaitu SD/sederajat dan SMP/sederajat, lalu pendidikan menengah yaitu SMA/sederajat, dan pendidikan tinggi yang mencakup program pendidikan diploma, sarjana, magister, spesialis, dan doktor [10].

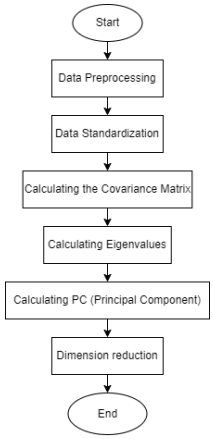
(Revisi 5)

TABEL II
SAMPEL DATA SETELAH PRAPEMROSESAN DATA

Pendidikan Tertinggi	Status dalam Keluarga	Jenis Kelamin	Umur
1	2	1	62
1	2	2	61
2	1	1	18
1	2	1	57
1	2	2	51

C. *Reduksi Dimensi*

Setelah data diproses, dapat terjadi masalah pada dimensi data yaitu terdapat banyak dimensi atau fitur. Oleh karena itu, tahap dimensi reduksi dilakukan untuk mengurangi jumlah dimensi data dengan teknik *Principal Component Analysis* (PCA).

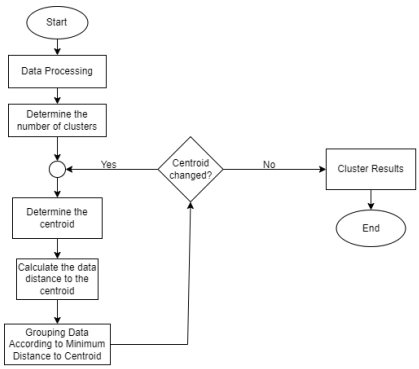


Gambar. 2. *Principal Component Analysis Flowchart*

Gambar 2 merupakan diagram alur tahapan dalam melakukan PCA. Langkah pertama adalah dengan melakukan standarisasi data. Standarisasi data diperlukan untuk menyeragamkan satuan variabel satu dengan yang lainnya. Selanjutnya, menghitung *matriks covariance* yang berfungsi sebagai nilai masukan untuk mendapatkan nilai *eigen* dan vektor *eigen*. Langkah berikutnya adalah menghitung nilai *eigen* untuk menyatakan seberapa besar keragaman yang mampu dijelaskan oleh suatu variabel *Principal Component Analysis* (PCA). Pada tahapan reduksi dimensi tidak semua variabel digunakan, hanya variabel yang mempunyai nilai *eigen* lebih dari 1 yang akan dipilih [11].

D. *Clustering*

Setelah tahap dimensi reduksi selesai, tahap selanjutnya adalah melakukan *clustering* menggunakan algoritma K-Means.



Gambar. 3. K-Means *Flowchart*

Gambar 3 adalah diagram alur dalam menentukan klaster menggunakan K-Means. K-Means adalah suatu metode analisis klaster yang berfungsi untuk membagi objek ke dalam beberapa klaster. Langkah pertama dalam penerapannya adalah menentukan jumlah kelompok atau klaster yang diinginkan. Setelah itu, titik awal secara acak dipilih sebagai pusat kelompok, dan jarak setiap titik data ke pusat kelompok terdekat dihitung. Setiap titik data kemudian dimasukkan ke dalam kelompok dengan pusat terdekat. Pusat baru kemudian dihitung berdasarkan rata-rata jarak titik-titik dalam kelompok, dan proses ini diulang sampai pusat kelompok tidak lagi berubah secara signifikan atau telah mencapai batas iterasi. Hasil akhir dari algoritma ini adalah kelompok-kelompok yang terbentuk berdasarkan jarak data ke pusat kelompok [12]. Untuk menggunakan algoritma K-Means dapat mengikuti langkah-langkah berikut: [13], [14]

(Revisi 6)

a. Menentukan nilai k sebagai jumlah *cluster* yang akan dibentuk
b. Menemukan *centroid* (titik pusat *cluster*) pada tahap awal secara acak, sedangkan pada tahap iterasi menggunakan rumus sebagai berikut:

$$v_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} Xkj \quad (1)$$

Penjelasan:
v_{ij} = *centroid* dari rata-rata *cluster* ke-i untuk variabel ke-j
Ni = jumlah anggota *cluster* ke-i
i, k = indeks dari *cluster*
j = indeks dari variabel
Xkj = nilai data ke-k variabel ke-j untuk *cluster* tersebut

c. Hitung jarak setiap objek ke masing-masing *centroid* dari setiap *cluster* dengan menggunakan rumus *Euclidean Distance*. Rumus *Euclidean Distance* merupakan jarak garis lurus biasa antara dua titik dalam ruang *Euclidean*, dengan rumus sebagai berikut:

$$De = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (2)$$

Penjelasan:
De = *Euclidean Distance*
i = Banyaknya objek
(x, y) = Koordinat objek
$\prod_{k=1}^{n} A_k$ = Koordinat *centroid*

d. Kelompokkan objek berdasarkan jarak ke *centroid* terdekat
e. Ulangi langkah ke-2 hingga ke-4, lakukan iterasi hingga *centroid* bernilai optimal

E. Evaluasi Klaster

Hasil dari *clustering* harus dievaluasi untuk memastikan kualitasnya. Metode evaluasi klaster yang digunakan dalam penelitian ini adalah sebagai berikut: [15]

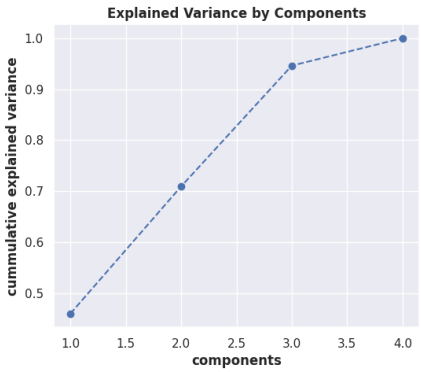
1) Metode Elbow
Metode yang digunakan untuk mengukur penurunan variansi dalam setiap kelompok saat jumlah klaster bertambah. Jika plot menunjukkan penurunan yang signifikan pada suatu titik, titik tersebut kemudian dapat dipilih sebagai jumlah optimal dari kelompok-kelompok.

2) Metode Silhouette
Mengukur seberapa baik setiap objek data yang cocok dengan kelompoknya sendiri dibandingkan dengan kelompok lain. Semakin besar skor *silhouette* yang dihasilkan, semakin baik objek data tersebut cocok dengan kelompoknya.

3) Davies-Bouldin Index
Metode yang dilakukan dengan mengukur kualitas klaster berdasarkan perbandingan antara jarak antar kelompok dan jarak dalam kelompok. Semakin kecil nilai indeks, semakin baik kualitas klaster.

III. HASIL DAN PEMBAHASAN

Hasil pengelompokan dengan menggunakan algoritma K-Means menghasilkan data yang tidak berlabel, polanya dapat ditemukan dengan cara mengelompokkan titik data yang memiliki atribut serupa dalam kumpulan data. Langkah awal yang dilakukan adalah melakukan standarisasi data untuk memastikan bahwa semua fitur memiliki standar deviasi 1 dan rata-rata 0. Selanjutnya, digunakan metode PCA untuk menentukan variabel yang penting dalam data dengan mengambil jumlah komponen yang menyumbang 70% sampai dengan 80%. Gambar 4 menunjukkan hasil dari penerapan metode PCA.



Kumulatif Varians Variabel dengan PCA

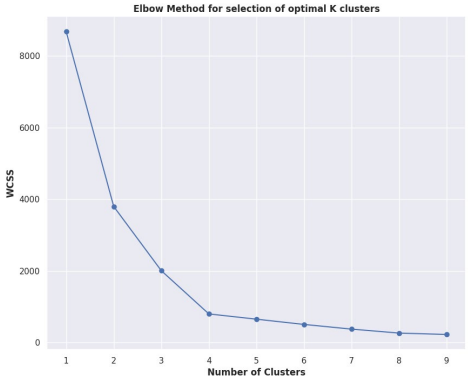
JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 4, Desember 2023, Pp. 1186-1195

Seperti yang terlihat pada Gambar 4, komponen pertama memiliki varian sebesar 0,45. Kemudian diikuti komponen kedua yang menjelaskan persentase varian sebesar 0,7. Sebanyak 2 komponen diperlukan untuk mempertahankan 70% dari total varian data. Berdasarkan hasil analisis, terdapat 2 komponen variabel yang dianggap penting dan akan digunakan dalam algoritma K-Means.

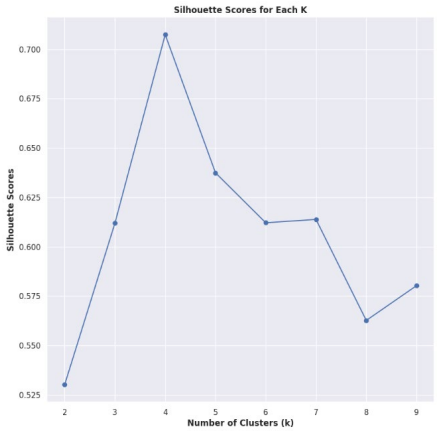
TABEL III
SAMPEL DATA HASIL REDUKSI DIMENSI

Component 1	Component 2
1.300303	1.033570
1.419263	-0.946898
-1.179237	1.088140
1.140623	1.019076
1.099903	-0.975887

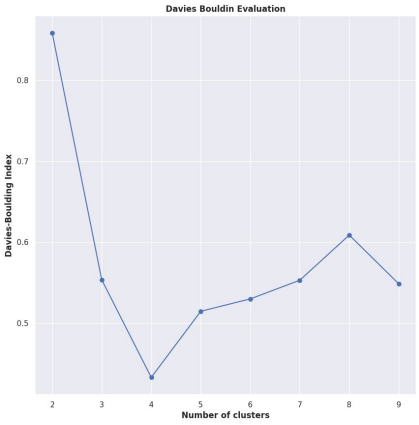
Tiga metode digunakan untuk menentukan pusat *cluster* yang optimal dalam K-Means, yaitu *Elbow, Silhouette*, dan *Davies Bouldin Index*. Hasil dari ketiga metode tersebut terlihat pada Gambar 5, Gambar 6, dan Gambar 7.



Gambar. 5. Evaluasi dengan Metode *Elbow*



Gambar. 6. Evaluasi dengan Metode *Silhouette Score*



Gambar. 7. Evaluasi dengan Metode *Davies Bouldin*

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 4, Desember 2023, Pp. 1186-1195

Berdasarkan ketiga hasil visualisasi metode tersebut, didapatkan bahwa metode *Elbow,* metode *Silhouette*, dan metode *Davies Bouldin* mencapai hasil yang sama dengan nilai $k = 4$. Ketiga metode evaluasi klaster yang digunakan menunjukkan bahwa klaster dengan jumlah 4 adalah yang paling optimal. (Revisi 7) Langkah berikutnya adalah mencari titik *centroid* atau titik pusat masing-masing klaster. Pemilihan *centroid* dalam penelitian ini dilakukan dengan coding menggunakan bahasa pemrograman Python.

TABEL IV
TITIK PUSAT KLASTER

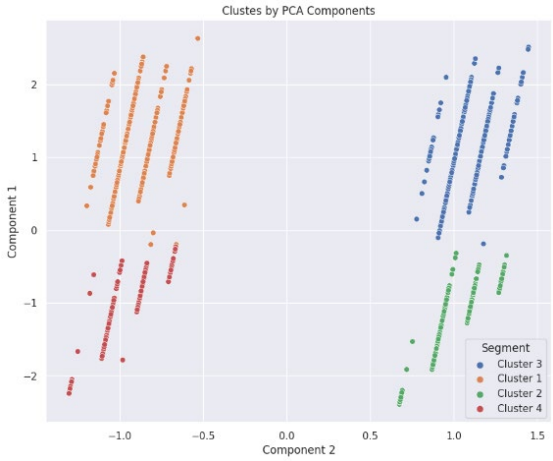
	Component 1	Component 2
C1	1.08444644	-0.92689143
C2	-1.57697179	0.91363876
C3	1.00297294	1.05870647
C4	-1.46355269	-1.06529566

Selanjutnya menghitung jarak dari setiap data ke pusat klaster menggunakan persamaan *Euclidean Distance*. Jarak yang dipilih adalah jarak yang paling dekat antara setiap data dengan *centroid*.

TABEL V
SAMPEL JARAK SETIAP DATA DENGAN TITIK PUSAT KLASTER

Distance 1	Distance 2	Distance 3	Distance 4	Jarak Terdekat
1.972309185	2.879773281	0.298390756	3.470466275	C3
0.335413844	3.526899571	2.048352249	2.885246048	C1
3.03061336	0.434330998	2.182408843	2.17212373	C2
1.946777757	2.719639274	0.143241445	3.335616011	C3
0.051375673	3.27657819	2.036901043	2.565014143	C1

Dengan demikian, hasil dari analisis K-Means dapat divisualisasikan menggunakan *scatter plot* agar dapat dengan mudah melihat pola atau struktur yang muncul dari pengelompokan tersebut. Gambar 8 adalah visualisasi hasil *clustering* menggunakan algoritma K-Means.

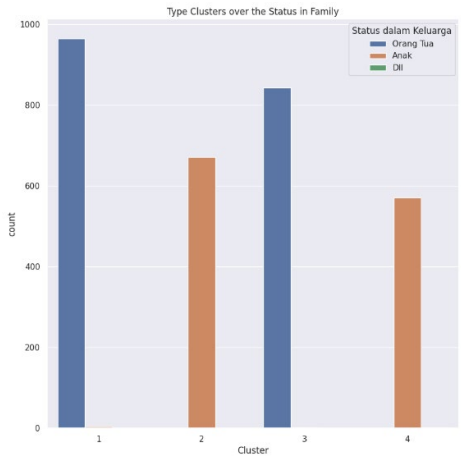


Gambar. 8. Hasil *Clustering* Menggunakan K-Means

Gambar 8 merupakan hasil *clustering* menggunakan K-Means yang terbagi menjadi 4 klaster, menunjukkan bahwa data telah berhasil dikelompokkan menjadi empat kelompok yang berbeda berdasarkan kesamaan fitur. Setiap titik pada *scatter plot* mewakili sebuah data dalam *dataset* yang telah dikelompokkan ke dalam salah satu klaster. Warna atau simbol yang berbeda dapat digunakan untuk membedakan setiap klaster. (Revisi 8) Hal ini dapat dilakukan karena telah dilakukan reduksi dimensi pada data dengan menggunakan PCA. Tanpa menggunakan PCA, karakteristik dari setiap klaster akan sulit dibedakan karena setiap klasternya memiliki dimensi data yang tinggi.

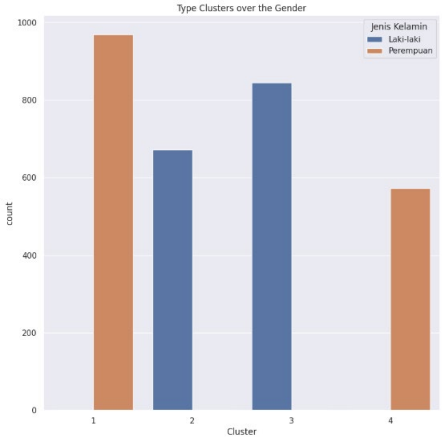
Setiap klaster memiliki karakteristik unik yang membedakannya dari klaster lainnya. *Exploratory Data Analysis* (EDA) pada masing-masing klaster dijelaskan sebagai berikut.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 4, Desember 2023, Pp. 1186-1195



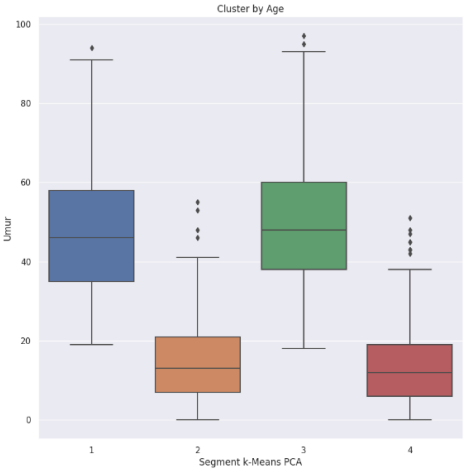
Gambar. 9. Klaster Berdasarkan Status dalam Keluarga

Gambar 9 memberikan informasi mengenai setiap klaster berdasarkan status dalam keluarga. Klaster satu dan klaster tiga merupakan orang tua, sedangkan klaster dua dan klaster empat merupakan anak.



Gambar. 10. Klaster Berdasarkan Jenis Kelamin

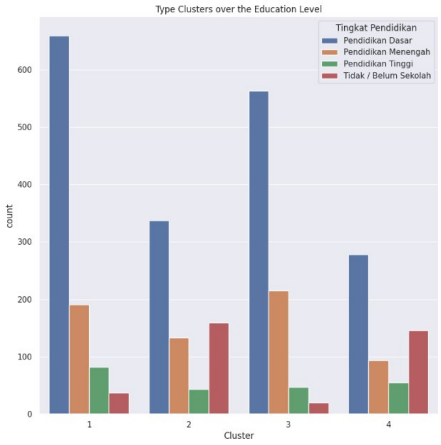
Gambar 10 memberikan informasi mengenai setiap klaster berdasarkan jenis kelamin. Berdasarkan visualisasi data tersebut, klaster satu dan klaster empat memiliki jenis kelamin perempuan, sedangkan klaster dua dan klaster tiga memiliki jenis kelamin laki-laki.



Gambar. 11. Klaster Berdasarkan Umur

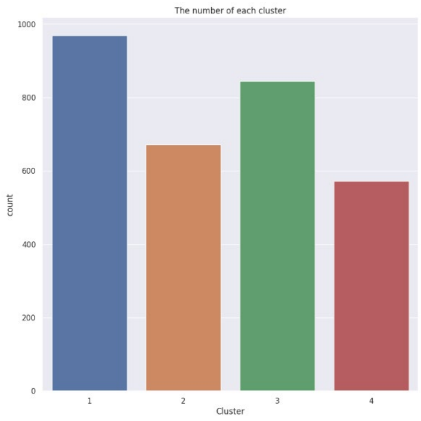
Gambar 11 memberikan informasi mengenai setiap klaster berdasarkan umur. Terlihat bahwa pusat data di klaster satu berada pada rentang usia 35 tahun sampai 57 tahun. Selanjutnya pusat data di klaster dua berada pada rentang usia 7 tahun sampai 21 tahun. Pusat data di klaster tiga berada pada rentang usia 38 tahun sampai 60 tahun, dan yang terakhir klaster empat di mana pusat data berada pada rentang usia 5 tahun sampai dengan 19 tahun.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 4, Desember 2023, Pp. 1186-1195



Gambar. 12. Klaster Berdasarkan Tingkat Pendidikan

Gambar 12 memberikan informasi mengenai setiap klaster berdasarkan tingkat pendidikan. Klaster satu dengan tingkat Pendidikan yang didominasi oleh pendidikan dasar, diikuti pendidikan menengah, pendidikan tinggi, dan yang terakhir tidak atau belum sekolah. Klaster dua tingkat pendidikan didominasi oleh pendidikan dasar, diikuti tidak atau belum sekolah, pendidikan menengah, dan yang terakhir pendidikan tinggi. Untuk klaster tiga, tingkat pendidikannya didominasi oleh pendidikan dasar, diikuti pendidikan menengah, sedikit pendidikan tinggi dan tidak atau belum sekolah. Untuk klaster empat, tingkat pendidikan didominasi oleh pendidikan dasar, diikuti tidak atau belum sekolah, pendidikan menengah, dan yang terakhir pendidikan tinggi.



Gambar. 13. Jumlah Observasi di Setiap Klaster

Gambar 13 memberikan informasi jumlah observasi di setiap klaster. Klaster 1 terdapat 969 observasi, klaster 2 terdapat 672 observasi, klaster 3 terdapat 845 observasi, dan klaster 4 terdapat 573 observasi. Informasi yang didapat dari hasil *clustering* menggunakan K-Means adalah sebagai berikut:

Klaster 1:
- Jumlah observasi yang terdapat di klaster 1 sebanyak 969
- Memiliki status sebagai orang tua dalam keluarga
- Memiliki jenis kelamin perempuan
- Pusat data berada di rentang usia 35 sampai dengan 57 tahun
- 68% menempuh pendidikan dasar, 20% menempuh pendidikan menengah, 8% menempuh pendidikan tinggi, dan 4% tidak atau belum bersekolah

Klaster 2:
- Jumlah observasi yang terdapat di klaster 2 sebanyak 672
- Memiliki status sebagai anak dalam keluarga
- Memiliki jenis kelamin laki-laki
- Pusat data berada di rentang usia 7 sampai dengan 21 tahun
- 50% menempuh pendidikan dasar, 24% tidak atau belum bersekolah, 20% menempuh pendidikan menengah, dan 6% menempuh pendidikan tinggi

Klaster 3:
- Jumlah observasi yang terdapat di klaster 3 sebanyak 845
- Memiliki status sebagai orang tua dalam keluarga

- Memiliki jenis kelamin laki-laki
- Pusat data berada di rentang usia 38 sampai dengan 60 tahun
- 67% menempuh pendidikan dasar, 25% menempuh pendidikan menengah, 6% menempuh pendidikan tinggi, dan 2% tidak atau belum bersekolah

Klaster 4:

- Jumlah observasi yang terdapat di klaster 4 sebanyak 573
- Memiliki status sebagai anak dalam keluarga
- Memiliki jenis kelamin perempuan
- Pusat data berada di rentang usia 5 sampai dengan 19 tahun
- 49% menempuh pendidikan dasar, 25% tidak atau belum bersekolah, 16% menempuh pendidikan menengah, dan 10% menempuh pendidikan tinggi

IV. KESIMPULAN

Dalam penelitian ini, teknik *clustering* K-Means dan PCA digunakan untuk mengelompokkan tingkat pendidikan masyarakat di Kabupaten Semarang berdasarkan jenis kelamin, umur, dan status individu dalam keluarga. Hasil dari penelitian ini menunjukkan bahwa penerapan PCA pada algoritma K-Means dapat mengurangi dimensi pada data tanpa menghilangkan informasi yang penting. Hasil dari penerapan PCA didapatkan dua komponen utama dengan proporsi kumulatif dari kedua komponen tersebut mencapai 70%. Setelah dilakukan reduksi dimensi dengan menggunakan metode PCA, analisis cluster dilakukan dengan menggunakan algoritma K-means yang menghasilkan empat kelompok dengan karakteristik yang berbeda untuk masing-masing cluster. Dengan demikian, penggunaan metode K-Means yang diimplementasikan dengan PCA dapat menjadi alternatif untuk melakukan segmentasi data dan memberikan pemahaman yang lebih baik dalam pengelompokan data dengan dimensi yang tinggi. Saran pengembangan untuk penelitian ke depan adalah dapat ditambahkan beberapa variabel baru yang terkait dengan faktor ekonomi seperti jumlah pendapatan, jenis pekerjaan, dan teknologi yang digunakan. Penambahan variabel tersebut, diharapkan dapat menambah validitas dan reliabilitas penelitian berikutnya terhadap tingkat pendidikan penduduk.

DAFTAR PUSTAKA

[1] O. Achmad and N. Sutikno, "BONUS DEMOGRAFI DI INDONESIA," *Visioner: Jurnal Pemerintahan Daerah di Indonesia*, vol. 12, no. 2, 2020, doi: https://doi.org/10.54783/jv.v12i2.285.

[2] BPS Provinsi Jawa Tengah, "Angka Partisipasi Kasar (APK) (Persen), 2019-2021." https://jateng.bps.go.id/indicator/28/70/1/angka-partisipasi-kasar-apk-.html (accessed Dec. 12, 2022).

[3] BPS, "Angka Partisipasi Kasar (APK)," 2014. https://sirusa.bps.go.id/sirusa/index.php/indikator/565 (accessed Dec. 12, 2022).

[4] W. Sheng *et al.*, "A Differential Evolution Algorithm With Adaptive Niching and K-Means Operation for Data Clustering," *IEEE Trans Cybern*, vol. 52, no. 7, pp. 6181–6195, 2020, doi: 10.1109/TCYB.

[5] C. H. Yu, F. Gao, S. Lin, and J. Wang, "Quantum data compression by principal component analysis," *Quantum Inf Process*, vol. 18, no. 8, Aug. 2019, doi: 10.1007/s11128-019-2364-9.

[6] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar Joseph, "A Review of Dimensionality Reduction Techniques for Efficient Computation," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 104–111. doi: 10.1016/j.procs.2020.01.079.

[7] B. Venkatesh and J. Anuradha, "A review of Feature Selection and its methods," *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3–26, 2019, doi: 10.2478/CAIT-2019-0001.

[8] K. Anwar, R. Goejantoro, and S. Prangga, "Pengelompokan Kabupaten/Kota Di Pulau Kalimantan Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2020 Menggunakan Optimasi K-Means Cluster Dengan Principle Component Analysis (PCA)," *Jurnal EKSPONENSIAL*, vol. 13, no. 2, pp. 131–140, 2022.

[9] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC - Trends in Analytical Chemistry*, vol. 132. Elsevier B.V., Nov. 01, 2020. doi: 10.1016/j.trac.2020.116045.

[10] "UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 20 TAHUN 2003 TENTANG SISTEM PENDIDIKAN NASIONAL." [Online]. Available: www.hukumonline.com

[11] D. A. Ihsani, A. Arifin, and M. H. Fatoni, "Klasifikasi DNA Microarray Menggunakan Principal Component Analysis (PCA) dan Artificial Neural Network (ANN)," *JURNAL TEKNIK ITS* , vol. 9, no. 1, pp. A124–A129, 2020.

[12] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," *Atmospheric Pollution Research*, vol. 11, no. 1. Elsevier B.V., pp. 40–56, Jan. 01, 2020. doi: 10.1016/j.apr.2019.09.009.

[13] Z. Nabila, A. Rahman Isnain, and Z. Abidin, "ANALISIS DATA MINING UNTUK CLUSTERING KASUS COVID-19 DI PROVINSI LAMPUNG DENGAN ALGORITMA K-MEANS," *Jurnal Teknologi dan Sistem Informasi (JTSI)*, vol. 2, no. 2, p. 100, 2021, [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[14] D. Transaksi Bongkar Muat di Provinsi Riau, I. Kamila, U. Khairunnisa, P. Studi Sistem Informasi, and F. Sains dan Teknologi UIN Sultan Syarif Kasim Riau, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan," *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, vol. 5, no. 1, pp. 119–125, 2019.

[15] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J (Basel)*, vol. 2, no. 2, pp. 226–235, Jun. 2019, doi: 10.3390/j2020016.