IMPLEMENTASI METODE K-MEANS DALAM ANALISIS PENGELOMPOKKAN PENERIMAAN PROGRAM INDONESIA PINTAR BERDASARKAN PROFIL SOSIAL EKONOMI SISWA

diajukan untuk memenuhi sebagian persyaratan memperoleh gelar Sarjana pada program studi Sistem Informasi.

SKRIPSI

Oleh

Zafira Dea Natasari
222410101003

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS JEMBER
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI SISTEM INFORMASI
JEMBER
2026

PERSEMBAHAN

Alhamdulillah hirobbil alamin, dengan penuh rasa syukur, skripsi ini saya persembahkan untuk:

1. Allah SWT yang senantiasa memberikan rahmat dan hidayah-Nya untuk menjalani hidup dan memberikan kelancaran serta kemudahan dalam menyelesaikan skripsi ini.
2. Kedua orang tua tercinta Bapak dan Ibu, serta keluarga besar yang selalu memberikan doa, semangat, dukungan, dan pengorbanan yang tak ternilai.
3. Saya sendiri, Zafira Dea Natasari yang selalu berdoa dan berusaha keras dalam menyelesaikan penelitian ini.
4. Teman-teman dan seluruh pihak yang tidak bisa disebutkan satu per satu yang telah memberikan do'a dan bantuan dalam mendukung penyelesaian penelitian ini.
5. Almamater Fakultas Ilmu Komputer Universitas Jember.

MOTTO

"Maka, sesungguhnya beserta kesulitan ada kemudahan. Sesungguhnya beserta kesulitan ada kemudahan"
- فَاِنَّ مَعَ الْعُسْرِ يُسْرًاۙ, اِنَّ مَعَ الْعُسْرِ يُسْرًاۗ -
(Q.S Al-Insyirah: 5-6)

"Maka nikmat Tuhan kamu yang manakah yang kamu dustakan?"
- فَبِاَيِّ اٰلَاۤءِ رَبِّكُمَا تُكَذِّبٰنِ -
(Q.S Ar-Rahman: 13)

"Tiada daya dan upaya kecuali dengan kekuatan Allah yang Maha Tinggi lagi Maha Agung"
-لَا حَوْلَ وَلَا قُوَّةَ إِلَّا بِاللهِ الْعَلِيِّ الْعَظِيْمِ -
(H.R Abu Dawud dan Ahmad)

PERNYATAAN ORISINALITAS

Saya yang bertanda tangan di bawah ini :

Nama : Zafira Dea Natasari

NIM : 222410101003

Menyatakan dengan sesungguhnya bahwa skripsi yang berjudul: *Implementasi Metode K-Means Dalam Analisis Pengelompokkan Penerimaan Program Indonesia Pintar Berdasarkan Profil Sosial Ekonomi Siswa* adalah benar-benar hasil karya sendiri, kecuali jika dalam pengutipan substansi disebutkan sumbernya, dan belum pernah diajukan pada institusi manapun, serta bukan karya jiplakan. Saya bertanggung jawab atas keabsahan dan kebenaran isinya sesuai dengan sikap ilmiah yang harus dijunjung tinggi.

Demikian pernyataan ini saya buat dengan sebenarnya, tanpa adanya tekanan dan paksaan dari pihak manapun serta bersedia mendapat sanksi akademik jika ternyata di kemudian hari pernyataan ini tidak benar.

Jember, 18 Juni 2026

Yang menyatakan,

Zafira Dea Natasari

NIM 222410101003

HALAMAN PERSETUJUAN

Skripsi berjudul *Implementasi Metode K-Means Dalam Analisis Pengelompokkan Penerimaan Program Indonesia Pintar Berdasarkan Profil Sosial Ekonomi Siswa* telah diuji dan disetujui oleh Fakultas Ilmu Komputer Universitas Jember pada:

Hari : Senin
Tanggal : 13 Juli 2026
Tempat : Fakultas Ilmu Komputer Universitas Jember

Pembimbing — Tanda Tangan

1. Pembimbing Utama
   Nama : Fajrin Nurman Arifin, S.T., M.Eng (.....)
   NIP : 198511282015041002

Penguji

1. Penguji Utama
   Nama : Nelly Oktavia Adiwijaya, S.Si., MT. (.....)
   NIP : 198410242009122008
2. Penguji Anggota 1
   Nama : Gayatri Dwi Santika, S.SI., M.Kom. (.....)
   NIP : 199201162023212044

ABSTRAK

Dapodik data contains various socioeconomic details about students that can be used to understand their circumstances, such as parental/guardian income, number of siblings, KIP (Student Assistance Card) ownership, KPS (Social Protection Program) enrollment, PIP (Student Assistance Program) eligibility status, and the reasons for PIP eligibility. This data is generally still used for administrative purposes and has not yet been widely utilized to analyze and identify student groups based on socioeconomic characteristics. This study aims to identify student groups and analyze the distribution of PIP proposals at SDN Mangunharjo 7 Probolinggo using the K-Means algorithm. The data used consists of Dapodik data for the 2025/2026 academic year, covering 313 students. The attributes used in the clustering process include family income, the number of siblings receiving KPS, and KIP recipients. This study involved several stages, namely data pre-processing, attribute transformation, data normalization, determining the number of clusters using the Silhouette coefficient, implementing the K-Means algorithm, and testing the association and correlation between attributes and PIP proposal status. The results yielded three clusters: Cluster 0 with 252 students, Cluster 1 with 17 students, and Cluster 2 with 42 students. The results of the correlation and association tests showed that all clustering attributes were significant in relation to PIP proposal status. This study demonstrates that K-Means can be used to describe patterns in students' socioeconomic profiles, but not to determine eligibility for PIP.

Keywords: Dapodik, K-Means, PIP, Clustering, Sosioeconomic

RINGKASAN

Program Indonesia Pintar (PIP) merupakan salah satu program bantuan pendidikan yang bertujuan membantu peserta didik dari keluarga miskin atau rentan miskin agar tetap memperoleh akses pendidikan. Dalam pelaksanaannya, sekolah memiliki data administratif yang tercatat dalam Data Pokok Pendidikan (Dapodik). Data tersebut memuat berbagai informasi siswa, termasuk data sosial ekonomi seperti penghasilan orang tua atau wali, jumlah saudara kandung, kepemilikan Kartu Indonesia Pintar (KIP), kepesertaan Kartu Perlindungan Sosial (KPS), status usulan Layak PIP, dan Alasan Layak PIP. Pada umunya data tersebut masih digunakan sebagai kebutuhan administratif dan belum banyak dimanfaatkan untuk melihat pola kelompok siswa berdasarkan kemiripan profil sosial ekonominya.

Penelitian ini bertujuan untuk mengelompokkan siswa SDN Mangunharjo 7 Kota Probolinggo berdasarkan profil sosial ekonomi menggunakan algoritma K-Means. Penelitian ini tidak digunakan untuk menentukan kelayakan penerimaan PIP, melainkan untuk mendeskripsikan pola kelompok siswa, mengidentifikasi karakteristik tiap *cluster*, menganalisis distribusi status dan alasan usulan Layak PIP, serta mengetahui atribut sosial ekonomi yang cenderung berasosiasi dengan status tersebut.

Penelitian ini menggunakan data Dapodik siswa SDN Mangunharjo 7 Kota Probolinggo tahun ajaran 2025/2026 dengan atribut penghasilan keluarga, jumlah saudara, penerima KPS, dan penerima KIP. Tahapan penelitian meliputi pembersihan data, transformasi, feature engineering, transformasi biner, dan normalisasi. Jumlah *cluster* ditentukan menggunakan metode Silhouette, kemudian dilakukan pengelompokan dengan algoritma K-Means, serta visualisasi dan interpretasi hasil. Analisis tambahan dilakukan terhadap status dan alasan usulan Layak PIP, serta uji asosiasi (Chi-Square) dan korelasi (Spearman).

Hasil penelitian menunjukkan terbentuk tiga *cluster*, yaitu Cluster 0 (81,03%) dengan profil sosial ekonomi relatif lebih baik, Cluster 1 (5,47%) sebagai kelompok paling rentan dengan penghasilan rendah dan jumlah saudara tinggi, serta Cluster 2 (13,50%) dengan kondisi sosial ekonomi sedang dan keterkaitan kuat

terhadap KPS. Visualisasi PCA menunjukkan sebaran *cluster* yang masih berdekatan, sehingga didukung dengan visualisasi Chernoff Face untuk memperjelas karakteristik.

Distribusi status dan alasan Layak PIP berbeda pada tiap *cluster*. Cluster 0 dan 2 didominasi tidak layak, namun masih terdapat alasan tertentu, sedangkan Cluster 1 didominasi layak meskipun alasan tidak tercatat. Hasil uji menunjukkan bahwa penerima KPS, KIP, penghasilan keluarga, dan jumlah saudara memiliki hubungan signifikan dengan status usulan Layak PIP.

Secara keseluruhan, K-Means mampu mendeskripsikan pola pengelompokan sosial ekonomi siswa. Hasil *clustering* tidak digunakan sebagai penentu kelayakan PIP, melainkan sebagai informasi pendukung dalam memahami profil siswa.

PRAKATA

Puji syukur kehadirat Allah Yang Maha Esa atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Implementasi Metode K-Means dalam Pengelompokkan Program Indonesia Pintar Berdasarkan Profil Sosial Ekonomi". Tidak lupa kepada junjungan besar Nabi Muhammad SAW. yang telah membawa manusia dari zaman *jahiliyyah* menuju zaman yang terang benderang. Skripsi ini disusun untuk memenuhi salah satu syarat menyelesaikan pendidikan Strata Satu (S1) pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Jember.

Penyusunan skripsi ini tidak lepas dari doa, bantuan, dan dukungan dari berbagai pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

1. Allah SWT yang senantiasa memberikan rahmat dan hidayah-Nya dalam menyelesaikan penelitian ini;
2. Bapak Prof. Drs. Antonius Cahya Prihandoko, M.App.Sc., Ph.D. selaku Dekan Fakultas Ilmu Komputer;
3. Ibu Dr. Oktalia Juwita, S.Kom., M.MT. selaku ketua Program Studi Sistem Informasi;
4. Bapak Fajrin Nurman Arifin, S.T., M.Eng. selaku Dosen Pembimbing yang telah memberikan arahan, saran, dan waktu yang sangat berarti dalam penyelesaian penelitian ini;
5. Seluruh Bapak dan Ibu dosen beserta staf karyawan di Fakultas Ilmu Komputer Universitas Jember;
6. Kedua orang tua saya Bapak Sujoko, dan Ibu Farida, yang senantiasa memberikan doa dan dukungan, beserta kedua adik saya Egsi dan Al.
7. Saya sendiri, Zafira Dea Natasari yang telah berjuang, berdoa, dan berusaha hingga sampai pada titik ini;
8. Sahabat saya sedari SMP, Yanti Yuliatur Rahma yang selalu memberikan doa dan semangat serta mendengar curhatan saya sehingga saya dapat menyelesaikan tugas akhir ini;

9. Teman-teman terdekat saya sedari SMA, Putri Eka Natasya dan Putri Souffi, serta teman terdekat saya di kuliah Annisa Rofiffah Rochma dan Bhisma Haris Alfitrah yang juga membantu memberikan doa dan semangat;
10. Seluruh pihak yang mendukung dan membantu penulis dalam menyelesaikan tugas akhir ini baik secara materiil dan non-materiil, yang tidak dapat disebutkan satu per satu.

Jember, 18 Juni 2026

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR NOTASI

k	: Jumlah *cluster*
x_i	: Data ke-*i*
C_j	: *Cluster* ke-*j*
min (x)	: Nilai minimum variabel
max (x)	: Nilai maksimum variabel
x'	: Nilai hasil normalisasi
s(i)	: Nilai sillhouette untuk data ke-i
a(i)	: Rata-rata jarak data ke-*i* terhadap data lain yang didalam klaster
b(i)	: Rata-rata jarak data ke-*i* terhadap data dalam klaster dekat lainnya.
$d(x_i, x_j)$	: Jarak Euclidean antara data ke-*i* dan ke-*j*
x_{ip}	: Nilai atribut ke-*p* pada data ke-*i*
c_i	: Centroid dari *cluster* ke-*i*
$\|C_i\|$	: Jumlah data dalam *cluster* ke-*i*
C_i	: Himpunan data yang menjadi anggota *cluster* ke-*i*
$\sum_{x_j \epsilon C_i} \quad x_j$	: Penjumlahan seluruh objek data yang berada di *cluster* C_i
$\left\|x_j - c_i\right\|^2$	: Kuadrat jarak Euclidean antara data dan centroid
O	: *Objected frequency*
E	: *Expected frequency*
p	: Nilai signifikasi

DAFTAR ISTILAH DAN SINGKATAN

Singkatan/Istikal	Arti dan keterangan
PIP	Program Indonesia Pintar
SD	Sekolah Dasar
Dapodik	Data Pokok Pendidikan
BPS	Badan Pusat Statistik
KPS	Kartu Perlindungan Sosial
KIP	Kartu Indonesia Pintar
KKS	Kartu Keluarga Sejahtera
PKH	Program Keluarga Harapan
SSE	*Sum of Squared Error*
NaN	*Not a Number*
PCA	*Principal Component Analysis*
Rp	Rupiah
Kemendikbudristek	Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi
DBSCAN	*Density Based Spatial Clustering of Application with Noise*
SOM	*Self-Organizing Map*

BAB 1. PENDAHULUAN

1.1 Latar Belakang

Pendidikan berperan penting dalam meningkatkan kualitas Sumber Daya Manusia (SDM) dalam melawan arus globalisasi untuk mempertahankan kepribadian yang menjunjung tinggi nilai etika, toleransi, dan tolong menolong (Kholillah et al., 2022). Pendidikan juga bertujuan mengembangkan potensi peserta didik yang mencakup aspek spiritual, kepribadian, kecerdasan, akhlak, dan keterampilan bagi diri, masyarakat, dan negara sesuai UU No. 20 Tahun 2003. Akses pendidikan berkualitas masih menghadapi tantangan dan hambatan, terutama biaya bagi masyarakat yang rentan di bidang ekonomi (Eldita & Umanto, 2025)

Bantuan pendidikan dapat berasal dari suatu lembaga atau organisasi tertentu (Stiawan & Nugroho, 2023). Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 10 Tahun 2020 tentang Program Indonesia Pintar menyatakan bahwa Program Indonesia Pintar (PIP) merupakan program bantuan pemerintah untuk mendukung wajib belajar 12 tahun. Program ini ditujukan untuk memperluas akses pendidikan hingga perguruan tinggi, khususnya bagi peserta didik dari keluarga miskin atau rentan miskin.

Data indikator Badan Pusat Statistik (BPS) tentang Angka Partisipasi Kasar/ Murni (APK/APM) tahun 2015-2021, menunjukkan sekitar 38,35% anak usia 16-18 tahun tidak melanjutkan jenjang SMA, menandakan bahwa adanya ketimpangan akses dan masalah ketepatan sasaran (Ivan, 2024). Partisipasi pendidikan pada setiap daerah berbeda-beda yang disebabkan oleh karakteristik sosial ekonomi rumah tangga, demografi, pendidikan, serta karakteristik daerah (Mulyani et al., 2023). Penerapan PIP dalam perspektif *good governance* berkaitan dengan aspek partisipasi, transparansi, tanggung jawab, keadilan, efektivitas, efisiensi, dan akuntabilitas (Septiawati et al., 2022). Berdasarkan temuan tersebut, diperlukan analisis berbasis data di tingkat sekolah untuk menilai pola penerimaan PIP berdasarkan profil sosial ekonomi siswa.

Teknik data mining dalam hal ini digunakan untuk menggali pola tersembunyi dari data profil siswa yang selama ini hanya digunakan untuk keperluan administratif. *Clustering* merupakan salah satu teknik data mining untuk mengelompokkan data ke dalam beberapa kelompok berdasarkan tingkat kemiripan (Dsouza et al., n.d.). Siswa dapat dikelompokkan berdasarkan kemiripan karakteristiknya dalam kondisi sosial ekonomi sehingga diperoleh gambaran kelompok-kelompok siswa yang memiliki rasio kerentanan yang berbeda, yang nantinya akan dibandingkan dengan status penerimaan PIP.

Sejumlah penelitian telah menerapkan teknik data mining untuk penentuan beasiswa dan bantuan pendidikan (Setiawan & Triayudi, 2024). Penerapan K-means pada teknik *clustering* digunakan untuk merekomendasikan calon penerima beasiswa berdasarkan kriteria tertentu (Rahmah & Antares, 2021). Penelitian oleh (Mardison et al., 2021) melakukan penggabungan algoritma K-Means dan C4.5 dimana algoritma K-Means digunakan untuk mengelompokkan penerima beasiswa dan C4.5 digunakan untuk menghasilkan akurasi model sehingga dapat digunakan untuk penyusunan aturan keputusan bagi pengelola beasiswa. Pada umumnya penelitian tersebut menggunakan data mining untuk penentuan kelayakan, perangkingan, atau prediksi penerima bantuan pendidikan. Penelitian yang dilakukan peneliti akan memposisikan *clustering* sebagai alat analitik utama untuk menganalisis pola kelompok profil sosial ekonomi siswa dan pemerataan PIP di satuan Pendidikan.

SD Negeri Mangunharjo 7 Kota Probolinggo pada tahun pelajaran 2025-2026 merupakan salah satu sekolah dasar yang penerima PIP dengan jumlah yang cukup banyak. Pada praktiknya, penetapan calon penerima PIP didasarkan pada usulan sekolah dengan pendataan administratif tanpa disertai kajian kuantitatif mengenai pola penerimaan dan pemerataan berdasarkan profil sosial ekonomi siswa. Kondisi inilah yang mendorong peneliti untuk memanfaatkan teknik *clustering* sebagai alat analitik guna mengungkap pola kelompok penerimaan dan pemerataannya berdasarkan profil sosial ekonomi SD Negeri Mangunharjo 7 Probolinggo. Penelitian ini tidak hanya menghasilkan pengelompokan data, melainkan juga

memberikan gambaran pola pemerataan penerimaan PIP berbasis data yang dapat dimanfaatkan sebagai bahan evaluasi atau rekomendasi bagi pihak sekolah.

1.2 Rumusan Masalah

Berdasarkan uraian dari latar belakang tersebut, adapun rumusan masalah yang didapat sebagai berikut:

1. Bagaimana analisis pengelompokan siswa berdasarkan profil sosial ekonomi menggunakan teknik *clustering* pada data Dapodik di sekolah SDN Mangunharjo 7 Probolinggo?
2. Bagaimana karakteristik setiap *cluster* siswa yang terbentuk berdasarkan indikator sosial ekonomi?
3. Bagaimana pola distribusi status usulan Layak PIP dan tidak Layak PIP serta Alasan Layak PIP pada masing-masing *cluster*?
4. Apa saja atribut sosial ekonomi yang menujukkan kecenderungan berasosiasi dengan status usulan Layak PIP pada masing-masing *cluster*?

1.3 Batasan Penelitian

Dalam penelitian ini, terdapat beberapa batasan masalah sebagai berikut:

1. Objek penelitian ini hanya pada siswa SDN Mangunharjo 7 Kota Probolinggo yang tercatat sebagai penerima maupun non penerima Program Indonesia Pintar pada tahun ajaran 2025/2026 yang menjadi fokus penelitian.
2. Penelitian ini hanya mengamati karakteristik penerima PIP dan menganalisis pola distribusi penerima PIP.
3. Batas kelayakan yang diperoleh dari hasil analisis merupakan batas empiris berdasarkan pola data penelitian, bukan batas normatif yang menggantikan aturan resmi PIP.

1.4 Tujuan Penelitian

Adapun tujuan yang ingin dicapai dari penelitian ini antara lain yaitu:

1. Mengetahui pengelompokan siswa SDN Mangunharjo 7 Probolinggo menggunakan Teknik *clustering* berdasarkan profil sosial ekonomi yang diperoleh dari data Dapodik.
2. Mengetahui karakteristik setiap *cluster* siswa yang terbentuk berdasarkan indicator sosial ekonomi.
3. Menganalisis pola distribusi status usulan Layak PIP dan Tidak Layak PIP serta Alasan Layak PIP pada masing-masing *cluster*.
4. Mengetahui atribut sosial ekonomi pada siswa yang berasosiasi dengan status usulan Layak PIP pada masing-masing *cluster*.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini sebagai berikut:

1. Bagi peneliti. Memperluas wawasan dan pengalaman dalam menerapkan Teknik *clustering* pada data Dapodik untuk menemukan pattern dan menginterpretasikan hasil pengelompokan dalam konteks pendidikan.
2. Bagi Praktisi Pengelola PIP di SDN Mangunharjo 7 Probolinggo. Membantu dalam memahami kelompok-kelompok profil sosial ekonomi yang dilihat dari karakteristik setiap *cluster*, sekaligus menjadi bahan evaluasi kualitas pendataan sebelum data dikirim ke Dapodik.
3. Bagi Dinas Pendidikan. Menjadi referensi dalam melakukan evaluasi distribusi karakteristik sosial ekonomi siswa berdasarkan data Dapodik sehingga dapat menjadi bahan pertimbangan dalam monitoring pemerataan program bantuan pendidikan.
4. Bagi akademisi. Menjadi referensi dan kontribusi untuk penelitian lebih lanjut dalam penerapan pada jenjang yang berbeda serta pengembangan sistem pendukung Keputusan berbasis hasil *clustering*.

BAB 2. TINJAUAN PUSTAKA

2.1 Kajian Literatur

Penelitian terdahulu yang relevan dengan topik ini umumnya membahas penerapan data mining dalam penentuan penerimaan beasiswa atau bantuan pendidikan. Penelitian-penelitian terdahulu memberikan gambaran penting untuk penelitian ini yang menggunakan metode serupa. Hasil telaah ini akan menjadi perbandingan antara penelitian ini dan penelitian sebelumnya. Pada tabel 2.1 yang menyajikan penelitian terdahulu menggunakan teknik *clustering* dalam berbagai konteks untuk mengelompokkan data dan analisis pemerataan. Gap penelitian pada penelitian terdahulu berfokus pada klasifikasi atau penentuan penerima bantuan. Selain itu data yang digunakan dari perguruan tinggi yang mana datasetnya lebih banyak, dibanding penelitian ini. Pada penelitian ini memanfaatkan *clustering* untuk mengeksplorasi profil sosial ekonomi siswa berdasarkan data Dapodik serta mengkaji karakteristik *cluster* terhadap status usulan Layak PIP.

2.2 Program Indonesia Pintar (PIP)

Program Indonesia Pintar merupakan bantuan pendidikan dari pemerintah yang diberikan kepada anak usia 6-21 tahun dan berasal dari keluarga miskin atau rentan miskin. Indikator prioritas kelayakan penerima PIP adalah peserta didik yang memiliki kartu KIP dan peserta didik dari keluarga miskin atau rentan miskin dengan pertimbangan khusus, yaitu (Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 10 Tahun 2020 tentang Program Indonesia Pintar):

a. Peserta didik dari keluarga peserta Program Keluarga Harapan
b. Peserta Didik dari keluarga pemegang Kartu Keluarga Sejahtera
c. Peserta Didik yang berstatus yatim piatu/yatim/piatu yang dari sekolah/panti sosial/panti asuhan
d. Peserta Didik yang terkena dampak bencana alam

e. Peserta Didik yang tidak bersekolah (drop out) yang diharapkan kembali bersekolah
f. Peserta Didik yang mengalami kelainan fisik, korban musibah, dari orang tua yang mengalami pemutusan hubungan kerja, di daerah konflik, dari keluarga terpidana, berada di Lembaga Pemasyarakatan, memiliki lebih dari 3 (tiga) saudara yang tinggal serumah
g. Peserta pada lembaga kursus atau satuan pendidikan nonformal lainnya.

Indikator tersebut merupakan dasar aturan dan penelitian ini menggunakan beberapa indikator tersebut sebagai proxy untuk mempresentasikan kondisi pengusulan PIP.

2.3 Indikator Kerentanan Sosial Ekonomi

Badan Pusat Statistik (BPS) tahun 2023-2024 di Kota Probolinggo terkait indikator kesejahteraan masyarakat dimana penduduk yang dikategorikan miskin diukur berdasarkan pengeluaran per kapita yang berada di bawah garis kemiskinan. Garis kemiskinan sendiri adalah batas minimum pengeluaran uang yang dibutuhkan untuk memenuhi kebutuhan dasar seperti pangan, perumahan, sandang, pendidikan, dan kesehatan (BPS Probolinggo, 2024).

Indikator kerentanan sosial ekonomi dalam penelitian ini mengacu pada sasaran PIP sebagaimana dijelaskan dalam Permendikbud Nomor 10 Tahun 2020, yaitu peserta didik dari keluarga miskin atau rentan miskin serta peserta didik dengan pertimbangan khusus. Kriteria tersebut kemudian disesuaikan dengan atribut yang tersedia dalam data Dapodik yaitu menggunakan penghasilan keluarga, jumlah saudara kandung, penerima KPS, dan penerima KIP sebagai atribut utama dalam proses *clustering*. Penghasilan keluarga merepresentasikan kondisi ekonomi rumah tangga, jumlah saudara kandung menggambarkan beban tanggungan keluarga, sedangkan penerima KPS dan penerima KIP menunjukkan keterkaitan siswa dengan bantuan sosial maupun bantuan pendidikan. Sementara itu, informasi seperti yatim/piatu dan Alasan Layak PIP dari usulan sekolah digunakan sebagai pendukung dalam interpretasi hasil, bukan sebagai atribut utama pembentuk *cluster*.

Tabel 2.1 Penelitian Terdahulu

No	Judul	Metode	Dataset	Hasil	Gap
1	Integration of fuzzy C-Means and SAW methods on education fee assistance recipients (Fadlil et al., 2023)	Fuzzy C-Means (FCM) untuk mengelompokkan calon penerima bantuan (*clustering*) dan Simple Additive Weight (SAW) untuk meranking prioritas penerima bantuan	Data pendaftar KIP-Kuliah di UMTAS tahun 2022 berisi 191 pendaftar. Atribut yang digunakan meliputi status DTKS, kepemilikan KIP/KKS, rata-rata pendapatan orang tua perbulan, jumlah tanggungan, prestasi siswa.	Hasil *clustering* menghasilkan dua *cluster*, yaitu 72 data pada *cluster* mendekati kriteria penerima bantuandan 119 data pada *cluster* tidak layak. SAW memilih 30 pendaftar dengan skor tertinggi sebagai penerima KIP-Kuliah dan sisanya menjadi cadangan	Objek penelitian adalah KIP-Kuliah di perguruan tinggi dengan menggunakan metode FCM +SAW yang berfokus pada penentuan kelayakan dan prioritas penerima sesuai kuota yang ada
2	A hybrid data mining for predicting scholarship recipient students by combining K-means and C4.5 methods (Hendri et al., 2024)	K-Means untuk mengelompokkan mahasiswa, kemudian hasil *cluster* digunakan dalam metode C4.5 untuk memprediksi penerima beasiswa.	Data mahasiswa Fakultas Ilmu Komputer UPI YPTK Padang sebanyak 200 mahasiswa. Atribut yang digunakan yaitu pendapatan orang tua, IPK, status orang tua, status penerima beasiswa	K-Means menghasilkan 3 *cluster* berdasarkan pendapatan orang tua dan IPK. Model C4.5 setelah hasil *clustering* mencapai akurasi sekitar 85% dalam memprediksi mahasiswa penerima beasiswa	Objek penelitian beasiswa mahasiswa yang menggunakan clustering untuk pengelompokkan dan C4.5 untuk prediksi penerimaan beasiswa.
3.	*Clustering* Analysis for Classifying Student Academic	Membandingkan beberapa metode *clustering* yaitu K-Means, BIRCH, dan DBSCAN. Evaluasi *cluster* menggunakan Davies-Bouldien, Silhouette,	Dataset mahasiswa B40 di perguruan tinggi negeri Malaysia. Setelah proses preprocessing, digunakan 16 atribut yang berkaitan	K-Means Model B (KmoB) memberikan hasil terbaik dibandingkan BIRCH dan DBSCAN. KmoB menghasilkan 5 *cluster*	Objek penelitian adalah mahasiswa B40 dimana berfokus pada peneglompokkan performa akademik

No	Judul	Metode	Dataset	Hasil	Gap
	Performance in Higher Education (Mohamed Nafuri et al., 2022)	dan Calinski-Harabasz. Hasil *cluster* digunakan sebagai label untuk membangun model klasifikasi menggunakan Decission Tree, Random Forest, dan Artificial Neural Network (ANN)	dengan karakteristik dan performa mahasiswa.	performa mahasiswa yaitu, Highest, High, Medium, Low, dan Lowest. Selanjutnya, model klasifikasi ANN menghasilkan akurasi tertinggi sebesar 99,92% dalam mengklasifikasikan label *cluster*.	mahasiswa.

2.4 Data Preprocessing

Preprocessing data merupakan tahap awal yang dilakukan dan sangat penting untuk penemuan hasil akhir (Bhargavi Konda, 2022). Tahap ini berfungsi mengubah data mentah yang masih mengandung noise, inkonsistensi, kesalahan, dan nilai hilang menjadi data yang bersih dan siap dianalisis (Bhargavi Konda, 2022).

Pada data mentah biasanya masih terdapat kolom kosong atau nilai yang kosong (*missing value*) sehingga menyebabkan kehilangan fitur pada model dan hasil tidak akurat. Metode yang digunakan untuk mengatasi nilai hilang yang cukup banyak yaitu dengan menggunakan nilai *mean, median,* atau *modus* yang berkaitan dengan fitur model (Maharana et al., 2022). Setelah itu dilakukan transformasi data yang merupakan hal penting atau inti dari persiapan data (Fernandes et al., 2023). Data akan ditransformasi dari bentuk awalnya menjadi bentuk yang lebih sesuai untuk dianalisis. Selanjutnya data dinormalisasi menggunakan Min-Max untuk menskalakan data [0-1]. Hal ini dihitung dengan mengurangi nilai minimum dari setiap nilai fitur dan kemudian membaginya dengan rentang fitur tersebut (Wongoutong, 2024).

$$x' = \frac{x - min\,(x)}{max\,(x) - min\,(x)} \quad (1)$$

Keterangan :

$min\,(x)$: Nilai minimum variabel
$max\,(x)$: Nilai maksimum variabel
x' : Nilai hasil normalisasi
x : Nilai awal

2.5 Silhouette Coefficient

Penentuan jumlah *k* ditentukan menggunakan koefisien siluet (*silhouette coefficient*). Rumus nilai siluet untuk satu data ke-*i* yaitu (Yuan & Yang, 2019) :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

Keterangan :

$s(i)$: Nilai *sillhouette* untuk data ke-*i*

$a(i)$: Rata-rata jarak data ke-i terhadap seluruh data lain dalam *cluster* yang sama.
$b(i)$: Rata-rata jarak data ke-i terhadap data dalam klaster terdekat lainnya (selain klaster asal).

Nilai $s(i)$ berada pada rentang [-1, 1]. Nilai Silhouette yang mendekati 1 menunjukkan bahwa objek memiliki hubungan yang dekat dengan cluster-nya, sehingga hasil pengelompokan dapat dikatakan sesuai. Semakin dekat nilai s(i) ke 1, maka semakin baik atau semakin masuk akal hasil pengelompokan suatu objek ke dalam cluster tertentu (Yuan & Yang, 2019)

2.6 Metode *Clustering*

Teknik *clustering* berfokus pada pembagian kumpulan data ke dalam *cluster* yang terdiri atas elemen-elemen yang mirip dan elemen ini berbeda dengan data diluar *cluster*. *Clustering* dianggap sebagai *unspervised learning* untuk menemukan *pattern* atau pola dalam sekelompok data yang tidak diketahui (Dsouza et al., 2017). Dalam implementasinya, metode *clustering* memerlukan algoritma untuk mendukung kinerjanya, diantaranya yaitu K-Means, DBSCAN (*Density Based Spatial Clustering of Application with Noise*), SOM (*Self-Organizing Map*), dan sebagainya (Marisa et al., 2019).

2.7 Algoritma K-Means

Algoritma K-Means merupakan algoritma *clustering* yang paling populer di data mining, yang bertujuan meminimalkan jumlah kuadrat jarak dalam *cluster* (*within-cluster sum of squares*) (Ahmed et al., 2020). K-Means dapat dimanfaatkan untuk mengelompokkan berbagai konteks dan kriteria yang diinginkan untuk menemukan pola dan sebagai dasar rekomendasi penerimaan PIP.

Pemilihan titik pusat dalam pengelompokkan nilai *k* tidak ada aturan standar karena sebagian besar diberikan secara acak (Liu, 2022). Dalam perkembangannya, terdapat metode inisialisasi K-Means++ yang digunakan untuk memilih centroid awal secara lebih terarah agar titik pusat awal lebih representatif (Nuraeni et al.,

2023). Algoritma K-Means dikelompokkan dalam sekumpulan data, dimana dinotasikan sebagai berikut (Liu, 2022):

$$X = \{x_1, x_2, x_3, \ldots, x_n\} \quad (3)$$

Notasi tersebut adalah jumlah sekumpulan data atau himpunan data yang terdiri atas n data. Setiap x_i merupakan satu objek data ke-i. Notasi tersebut akan dipartisi kedalam kelompok k (*cluster*) yang dinotasikan sebagai berikut:

$$P_k = \{C_1, C_2, C_3, \ldots, C_k\} \quad (4)$$

P_k merupakan partisi data yang membagi himpunan data X ke dalam k *cluster*, dan setiap C_i menyatakan kluster ke-i dimana berisi sekumpulan objek data. Langkah-langkah algoritma K-Means setelah *centroid* awal ditentukan, selanjutnya dilakukan dengan menghitung jarak antara setiap data dengan masing-masing centroid menggunakan jarak Euclidean. Jarak Euclidean antara dua titik data p-dimensi x_i dan x_j ditetapkan, seperti pada rumus:

$$x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ip}),$$

$$x_j = (x_{j1}, x_{j2}, x_{j3}, \ldots, x_{jp}),$$

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \quad (5)$$

Keterangan:

$d(x_i, x_j)$	: Jarak Euclidean antara data ke-i dan ke-j
x_i	: Data ke-i
x_{ip}	: Nilai atribut ke-p pada data ke-i
x_j	: Data ke-j
x_{jp}	: Nilai atribut ke-p pada data ke-j

Hasil jarak setiap data yang terdekat, dikelompokkan dan akan dimasukkan kedalam *cluster*. Selanjutnya centroid diperbarui dengan menghitung rata-rata nilai seluruh data yang menjadi anggota pada setiap *cluster*. Pembaruan centroid dilakukan menggunakan rumus berikut:

$$c_i = \frac{1}{|C_i|} \sum_{x_j \epsilon C_i} \quad x_j \quad (6)$$

Keterangan:

c_i	: Centroid dari *cluster* ke-i
$\|C_i\|$	: Jumlah data dalam *cluster* ke-i
C_i	: Himpunan data yang menjadi anggota *cluster* ke-i
x_j	: Objek data ke j yang terdapat pada *cluster* C_i

$\sum_{x_j \epsilon C_i}$ x_j : Penjumlahan seluruh objek data yang berada di *cluster* C_i

Algoritma K-Means bertujuan membagi data ke dalam sejumlah (k) *cluster* berdasarkan kedekatan data terhadap centroid (Yuan & Yang, 2019). Dalam proses *clustering*, fungsi objektif yang umum digunakan adalah *Sum of Squared Error* (SSE), yaitu jumlah kuadrat jarak antara setiap data dengan centroid pada *cluster* masing-masing (Gul & Rehman, 2023). Rumusan matematis untuk SSE diberikan di bawah ini (7). Semakin kecil nilai SSE yang diperoleh, maka semakin baik kualitas pengelompokan yang dihasilkan oleh algoritma K-Means (Qi et al., 2017)

$$SSE = \sum_{i=1}^{k} \sum_{x_j \epsilon C_i} \left|\left|x_j - c_i\right|\right|^2 \quad (7)$$

Keterangan:

SSE	: *Sum of Squared Errors*
k	: Jumlah *cluster*
C_i	: Himpunan data pada *cluster* ke-i
x_j	: Data ke-j
c_i	: Centroid *cluster* ke-i
$\left\|\left\|x_j - c_i\right\|\right\|^2$	: Kuadrat jarak Euclidean antara data dan centroid

Proses perhitungan jarak Euclidean dan pembaruan centroid dilakukan secara iteratif hingga tidak terjadi perubahan signifikan pada keanggotaan *cluster* atau nilai centroid, yang menandakan bahwa algoritma telah mencapai kondisi konvergen. Pada kondisi tersebut, nilai SSE menjadi minimum dan solusi yang diperoleh merupakan yang terbaik berdasarkan proses iterasi yang dilakukan.

2.8 Uji Korelasi dan Asosiasi

Uji korelasi dan asosiasi digunakan untuk mengetahui hubungan dan keterkaitan antara dua variabel. Metode uji Chi-Square merupakan uji nonparametrik yang digunakan untuk melihat distribusi frekuensi pada variabel kategorik atau mengetahui hubungan (asosiasi) antara dua variabel kategorik (Cote et al., 2021.). Rumus utama Chi-Square yaitu sebagai berikut:

$$x^2 = \sum \frac{(O-E^2)}{E} \quad (8)$$

Keterangan :

O : *Objected frequency* (frekuensi yang diamati)

E : *Expected frequenc*y (frekuensi harapan)

Rumus tersebut digunakan untuk mengetahui seberapa besar frekuensi yang diamati dan frekuensi yang diharapkan. Semakin besar perbedaannya, maka akan semakin besar pula kemungkinan kedua variabel kategorik memiliki hubungan asosiasi.

Metode korelasi Spearman juga digunakan dalam uji korelasi antar dua variabel. Metode ini digunakan apabila data berbentuk ordinal atau tidak memenuhi asumsi parametik (Goss-Sampson, 2022). Koefisien Spearman berada di rentang -1 sampai +1, dimana nilai positif menunjukkan hubungan searah, nilai negatif menunjukkan berlawanan arah, dan nilai 0 menunjukkan hubungan yang lemah atau tidak berhubungan. Apabila nilai *p-value* lebih kecil, misalnya $p < 0,05$ maka hubungan dikatakan signifikan (Goss-Sampson, 2022).

BAB 3. METODOLOGI PENELITIAN

3.1 Lokasi dan Waktu Penelitian

Penelitian ini dilaksanakan di SD Negeri Mangunharjo 7 Kota Probolinggo. Waktu penelitian direncanakan pada bulan Januari sampai Juni 2026.

3.2 Populasi dan Sampel/Subyek Penelitian

Populasi dalam penelitian ini adalah seluruh siswa SDN Mangunharjo 7 Kota Probolinggo tahun ajaran 2025-2026 yang tercatat dalam data Dapodik. Jumlah data awal berisi 313 baris data siswa.

Sampel penelitian ini menggunakan teknik total sampling, yaitu seluruh populasi dijadikan sampel penelitian. Kriteria data yang dianalisis adalah data siswa yang memiliki informasi terkait profil sosial ekonomi, meliputi penghasilan orang tua, kepemilikan KIP, kepesertaan bantuan sosial seperti PKH/KPS/KKS, dan jumlah saudara. Analisis lebih lanjut menggunakan atribut status usulan layak PIP dan alasan layak PIP.

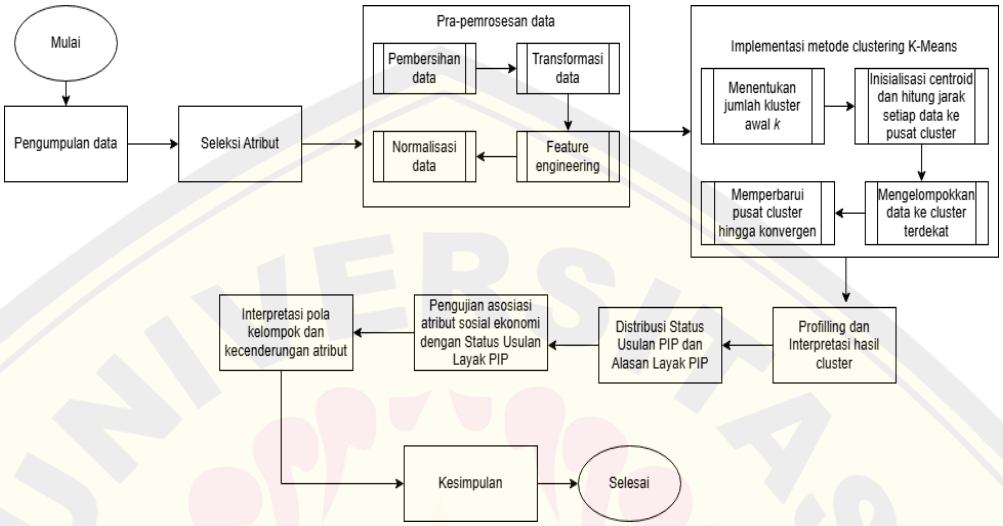
3.3 Data Penelitian

Data penelitian terdiri atas data primer dan sekunder. Data sekunder menggunakan data dari Dapodik siswa SDN Mangunharjo 7 Kota Probolinggo tahun ajaran 2025-2026. Data Dapodik yang diperoleh berisi 313 baris dan 66 atribut. Dalam proses *clustering* tidak semua atribut digunakan, karena atribut disesuaikan dengan tujuan penelitian, kesediaan data, serta indikator sosial ekonomi yang berkaitan dengan PIP.

Data primer sebagai data pendukung yaitu dilakukan wawancara dengan pihak sekolah. Wawancara dilakukan untuk mengonfirmasi terkait indikator kerentanan sosial ekonomi pada siswa yang dipertimbangkan dalam pengusulan PIP.

3.4 Prosedur Penelitian

Prosedur atau tahapan penelitian menggambarkan alur penelitian yang dirancang agar penelitian berjalan sistematis. Tahapan ini akan dilaksanakan oleh peneliti hingga penelitian selesai.



Gambar 3.1 Prosedur Penelitian

3.4.1 Pengumpulan Data

Pengumpulan data diperoleh dari data Dapodik siswa SDN Mangunharjo 7 Probolinggo dan wawancara singkat kepada pihak sekolah untuk mengonfirmasi kriteria yang selama ini dipertimbangkan dalam pengusulan PIP.

3.4.2 Seleksi Atribut

Seleksi atribut bertujuan menentukan atribut yang sesuai dengan tujuan penelitian. Tidak semua atribut pada data Dapodik digunakan. Atribut yang digunakan dalam penelitian ini yaitu penghasilan keluarga, penerima KPS, penerima KIP, dan jumlah saudara.

3.4.3 Pra-Pemrosesan Data

Pada tahap pra-pemrosesan data diawali dengan melakukan pembersihan data. Pembersihan data dilakukan dengan mengecek dan menangani *missing value*, dan menghapus data duplikat. Data kategorikal kemudian ditransformasi menjadi data numerik serta melakukan normalisasi data untuk menyamakan skala antar atribut.

3.4.4 Implementasi Metode *Clustering* K-Means

Metode *clustering* yang dipilih menggunakan algoritma K-Means. Diawali dengan menentukan jumlah *k* menggunakan metode *silhouette coefficient*. Pengujian dilakukan dengan menjalankan algoritma K-Means pada beberapa nilai *k*, yaitu *k*=2 sampai *k*=10. Jumlah *k* yang paling optimal dipilih berdasarkan nilai *k* yang memiliki *silhouette score* tertinggi.

Setelah nilai *k* ditetapkan, selanjutnya algoritma K-Means dijalankan. Inisialisasi *centroid* awal menggunakan K-Means++ agar lebih terarah dan dilakukan secara iteratif. Pada setiap iterasinya akan dihitung jarak setiap data ke masing-masing *centroid*. Proses *clustering* dilakukan dengan menghitung jarak setiap data ke *centroid*, lalu data ditempatkan pada *cluster* dengan jarak terdekat atau terkecil. Setelah terbentuk kelompok, algoritma akan memperbarui *centroid* dengan menghitung rata-rata nilai setiap variabel pada masing-masing *cluster*. Proses ini akan diulang hingga tidak terjadi lagi perubahan anggota *cluster* yang signifikan (konvergen).

3.4.5 Profilling dan Interpretasi Hasil *Cluster*

Profilling hasil *cluster* yaitu menampilkan ciri-ciri dari setiap *cluster.* Interpretasi hasil *cluster* dilakukan dengan menghitung statistik setiap *cluster*. Berdasarkan karakteristik tersebut, setiap *cluster* diberi makna, misalnya *cluster* dengan kerentanan tinggi, sedang, atau rendah. Setelah label *cluster* diperoleh, data hasil *clustering* digabungkan kembali dengan atribut Layak PIP dan Alasan Layak PIP. Penggabungan ini bertujuan untuk mengetahui bagaimana distribusi status kelayakan PIP pada setiap kelompok sosial ekonomi yang terbentuk.

3.4.6 Distribusi Status Usulan PIP dan Alasan Layak PIP

Data hasil *clustering* akan digunakan sebagai dasar untuk mengamati distribusi bantuan pada setiap kelompok profil sosial ekonomi. Setelah label *cluster* diperoleh, label tersebut digabungkan kembali dengan atribut status usulan Layak PIP dan Alasan Layak PIP. Selanjutnya dilakukan perhitungan jumlah dan persentase siswa usulan Layak PIP dan Tidak Layak PIP pada masing-masing *cluster*. Perhitungan distribusi status usulan PIP digunakan

untuk melihat bagaimana sebaran status usulan PIP pada kelompok siswa yang terbentuk berdasarkan profil sosial ekonomi. Hasil perhitungan ini tidak digunakan untuk menetapkan kelayakan siswa sebagai penerima PIP, melainkan untuk membaca pola status usulan PIP pada setiap *cluster*.

3.4.7 Pengujian Asosiasi dan Korelasi

Pengujian asosiasi atribut dilakukan untuk mengetahui atribut sosial ekonomi yang memiliki kecenderungan hubungan dengan status usulan Layak PIP. Pengujian asosiasi dilakukan setelah proses *clustering* untuk mengetahui atribut sosial ekonomi yang memiliki kecenderungan hubungan dengan status usulan Layak PIP. Pengujian ini tidak digunakan untuk menetapkan kelayakan PIP, melainkan untuk mendukung interpretasi pola distribusi status usulan PIP pada masing-masing *cluster*.

3.4.8 Interpretasi Pola Kelompok dan Kecenderungan Atribut

Pada tahap ini, setiap *cluster* dianalisis berdasarkan karakteristik sosial ekonomi yang dominan. Interpretasi juga dilakukan dengan mengaitkan karakteristik tersebut dengan distribusi status usulan Layak PIP dan Alasan Layak PIP pada masing-masing *cluster.* Tahap ini bertujuan untuk menjelaskan makna dari setiap *cluster*, misalnya kelompok siswa dengan keterkaitan bantuan sosial, kelompok penerima PIP, atau kelompok dengan penghasilan orang tua tertentu. Dengan demikian, hasil *clustering* tidak hanya menunjukkan pembagian kelompok siswa, melainkan memberikan gambaran mengenai kecenderungan atribut sosial ekonomi yang membedakan setiap *cluster*.

3.5 Alat/Instrumen Penelitian

Alat atau instrumen yang digunakan dalam penelitian ini mencakup, Data Dapodik, Laptop, Google Colaboratory, Python, dan Library Python. Data Dapodik merupakan data sekunder yang merupakan data utama untuk penelitian ini dan dikonfirmasi dengan wawancara sebagai informasi pendukung. Laptop sebagai perangkat keras digunakan untuk mengolah, menganalisis, dan menyimpan data penelitian. Perangkat keras ini digunakan mulai dari tahap awal penelitian sampai akhir penelitian.

Selain itu, Google Colaboratory digunakan sebagai alat untuk menjalankan kode program Python yang merupakan perangkat lunak utama dalam proses pengolahan dan analisis data. Dalam prosesnya digunakan library python seperti Pandas, NumPy, Scikit-learn, Matplotlib, Matplotlib patches, Seaborn, dan SciPy.

3.6 Metode Analisis

Metode analisis yang digunakan yaitu analisis kuantitatif dengan pendekatan data mining menggunakan algoritma K-Means. Atribut yang digunakan dalam *clustering* yaitu penghasilan keluarga, penerimaan KPS, penerima KIP, dan jumlah saudara.

Sebelum dilakukan proses *clustering* dilakukan pra-pemrosesan data, yaitu pengecekan *missing value,* data duplikat, dan transformasi data. Pada atribut penghasilan dilakukan *feature engineering* sehingga diperoleh atribut penghasilan keluarga dari data penghasilan ayah, ibu, dan wali. Apabila data ayah dan ibu tidak tersedia, maka menggunakan data penghasilan wali sebagai data pengganti. Pembentukan atribut ini dilakukan agar kondisi ekonomi keluarga siswa dapat direpresentasikan dalam satu variabel yang lebih sesuai untuk proses *clustering*.

Data yang telah ditransformasi, kemudian dinormalisasi menggunakan *Min-Max normalization*. Normalisasi ini dilakukan pada seluruh variabel numerik agar berada di rentang [0,1] dan bertujuan untuk mencegah variabel dengan skala yang besar, misalnya penghasilan keluarga yang jumlahnya besar akan mendominasi jarak Euclidean pada saat proses *clustering*. Oleh karena itu, dengan menggunakan *Min-Max normalization*, maka setiap variabel yang digunakan dalam pembentukan *cluster* akan lebih seimbang.

Proses *clustering* dilakukan dengan algoritma K-Means untuk mengelompokkan siswa berdasarkan kemiripan sosial ekonomi berdasarkan jarak Euclidean. Dalam *clustering,* jarak Euclidean digunakan untuk menentukan kemiripan data numerik dan merupakan metrik standar yang digunakan pada algoritma K-Means. Dengan jarak Euclidean, setiap data akan ditempatkan pada *cluster* dengan jarak terdekat terhadap *centroid* yang artinya semakin mirip profilnya.

Jumlah *cluster* optimal ditentukan menggunakan *silhouette coefficient* dengan menguji beberapa nilai *k* yaitu, *k*=2 sampai *k*=10. Hasil dari setiap *clustering* kemudian dievaluasi menggunakan nilai rata-rata *silhouette score* yang nilainya berada pada rentang -1 sampai 1. Nilai yang mendekati 1 menunjukkan bahwa data memiliki pendekatan yang baik dengan *cluster*-nya sendiri dan memiliki keterpisahan yang baik dengan *cluster* yang lain, sedangkan nilai mendekati 0 menunjukkan bahwa data berada dibata antar-*cluster* serta nilai negatif menunjukkan bahwa kemungkinan data lebih sesuai dengan *cluster* lain.

Distribusi status usulan PIP dianalisis menggunakan perhitungan persentase usulan siswa layak PIP dan tidak layak PIP di setiap *cluster* hasil pemodelan. Uji korelasi dan asosiasi menggunakan metode Chi-Square dan Spearman. Teknik Chi-Square untuk atribut kategorik seperti atribut penerima KIP dan penerima KPS. Masing-masing variabel diuji dengan variabel usulan layak PIP.

Atribut yang numerik seperti jumlah saudara dan penghasilan keluarga diuji menggunakan metode Spearman. Dikatakan signifikan secara statistik apabila nilai *p-value<0,05*, sehingga dapat disimpulkan adanya hubungan yang bermakna antara kedua variabel.

BAB 4. HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan hasil dan pembahasan dari penelitian yang sudah dilakukan berdasarkan tahapan penelitian yang sudah disusun sebelumnya. Hasil penelitian akan disajikan melalui tabel.

4.1 Pengumpulan Data

Data yang digunakan dalam penelitian berupa data sekunder yang diperoleh dari data Dapodik siswa SDN Mangunharjo 7 Kota Probolinggo tahun ajaran 2025-2026. Data tersebut memuat bebrapa informasi dari siswa SDN Mangunharjo 7 Probolinggo yang diantaranya terdapat beberapa atribut sosial ekonomi. Atribut yang digunakan disesuaikan dengan tujuan penelitian, ketersediaan data, hasil wawancara dengan pihak sekolah, serta indikator sosial ekonomi yang berkaitan dengan Program Indonesia Pintar.

Tabel 4.1 Keterangan Atribut

Keterangan Atribut	Jumlah
Data siswa	313
Atribut awal	66
Atribut *clustering*	4
Atribut analisis setelah *clustering*	2

Berdasarkan data awal, terdapat 313 data siswa dengan 66 atribut. Pada seleksi atribut, digunakan 4 atribut untuk proses *clustering* yaitu atribut sosial ekonomi siswa. Atribut tersebut diantaranya adalah penghasilan orang tua, penerima KIP, penerima KPS, dan jumlah saudara. Atribut tambahan digunakan setelah *clustering* untuk analisis karakteristik pola penerimaan PIP, yaitu status usulan PIP, dan alasan layak PIP. Pada tabel 4.2 terlampir dataset awal yang memuat atribut sosial ekonomi dan akan digunakan untuk proses *clustering*.

Tabel 4.2 Dataset Penelitian

No	Penghasilan Ayah	Penghasilan Ibu	Penghasilan Wali	Penerima KPS	Penerima KIP	Jumlah Saudara Kandung
0	Rp. 1,000,000-Rp. 1,999,999	Tidak Berpenghasilan	NaN	Tidak	Tidak	1
1	Rp. 2,000,000-Rp. 4,999,999	Rp. 500,000-Rp. 999,999	NaN	Tidak	Tidak	1

No	Penghasilan Ayah	Penghasilan Ibu	Penghasilan Wali	Penerima KPS	Penerima KIP	Jumlah Saudara Kandung
2	Rp. 2,000,000-Rp. 4,999,999	Rp. 1,000,000-Rp. 1,999,999	NaN	Tidak	Tidak	1
3	Rp. 2,000,000-Rp. 4,999,999	Tidak Berpenghasilan	NaN	Tidak	Tidak	2
4	Rp. 2,000,000-Rp. 4,999,999	Tidak Berpenghasilan	NaN	Tidak	Tidak	3

Pada tabel 4.2 atribut penghasilan masih terdiri atas penghasilan ayah, penghasilan ibu, dan penghasilan wali. Ketiga atribut tersebut tidak langsung digunakan dalam proses *clustering* karena perlu dilakukan pembentukan atribut baru berupa penghasilan keluarga. Selain itu, masih terdapat beberapa *missing value* sehingga perlu penanganan.

4.2 Pra-Pemrosesan Data

4.2.1 Pembersihan Data

Pembersihan data dilakukan dengan pengecekan nilai kosong, data duplikat dan data yang tidak sesuai format. Berdasarkan hasil pengecekan, Data Dapodik yang digunakan dalam penelitian ini tidak terdapat data duplikat dan data yang tidak sesuai format, melainkan memiliki nilai kosong pada beberapa atribut yang akan digunakan dalam penelitian ini.

Tabel 4.3 Nilai Kosong

Nama	0
Penghasilan Ayah	5
Penghasilan Ibu	0
Penghasilan Wali	306
Penerima KPS	0
Penerima KIP	0
Layak PIP (usulan dari sekolah)	0
Alasan Layak PIP	256
Jumlah Saudara Kandung	0

Pada tabel 4.3 atribut Penghasilan Ayah, Penghasilan Wali, dan Alasan Layak PIP masih terdapat nilai kosong atau nilai hilang. Nilai pada atribut penghasilan perlu ditangani karena merupakan atribut yang berkaitan dengan

pembentukan atribut penghasilan keluarga, dimana atribut ini yang akan digunakan dalam proses *clustering*.

Pada atribut Alasan Layak PIP terdapat nilai kosong sebanyak 256. Nilai kosong tersebut tidak dihapus melainkan ditangani untuk dianalisis setelah *cluster* terbentuk.

Tabel 4.4 Status Alasan Layak PIP

Alasan Layak PIP	Jumlah
Tidak Layak PIP	238
Siswa Miskin/Rentan Miskin	28
Alasan tidak tercatat	17
Pemegang PKH/KPS/KKS	18
Yatim Piatu/Panti Asuhan/Panti Sosial	6
Menolak	4

Pada tabel 4.4 atribut Alasan Layak PIP, penanganan nilai kosong pada atribut ini dilakukan berdasarkan status pada kolom Layak PIP (usulan dari sekolah). Apabila siswa berstatus Tidak Layak PIP dan kolom alasan kosong, maka alasan diisi dengan kategori Tidak Layak PIP. Apabila siswa berstatus Layak PIP tetapi kolom alasan kosong, maka alasan diisi dengan kategori Layak PIP tetapi alasan tidak tercatat. Penanganan ini dilakukan agar distribusi alasan Layak PIP dapat dianalisis pada masing-masing *cluster*.

4.2.2 Transformasi Penghasilan

Atribut Penghasilan Ayah, Penghasilan Ibu, dan Penghasilan Wali masih berbentuk kategorik rentang penghasilan. Oleh karena itu, dilakukan transformasi data dari bentuk kategorik menjadi bentuk numerik.

Transformasi penghasilan dilakukan dengan memberikan nilai representatif pada setiap kategori penghasilan. Kategori ‘Tidak Berpenghasilan’ diberi nilai 0 karena kategori tersebut tidak memiliki rentang nominal penghasilan yang tercatat pada data. Hal ini bukan berarti penghasilan riil Rp0, atau sebagai kepastian kalau keluarga memiliki penghasilan paling rendah, melainkan sebagai kode numerik untuk kategori tanpa penghasilan

tetap pada data. Kategori penghasilan yang berbentuk rentang penghasilan dikonversi menggunakan nilai tengah dari masing-masing rentang. Hasil transformasi penghasilan dalam penelitian ini diinterpretasikan sebagai representasi data administratif, bukan sebagai nilai penghasilan riil secara mutlak.

Tabel 4.5 Konversi Penghasilan

Kategori Penghasilan	Numerik
Tidak Berpenghasilan	0
Kurang dari Rp. 500,000	250000
Rp. 500,000 – Rp. 999,999	750000
Rp. 1,000,000 – Rp. 1,999,999	1500000
Rp. 2,000,000 – Rp. 4,999,999	3500000
Rp. 5,000,000 – Rp. 20,000,000	12500000
Lebih dari Rp. 20,000,000	20000000

Setelah proses transformasi penghasilan dilakukan, atribut Penghasilan Ayah, Penghasilan Ibu, dan Penghasilan Wali telah berubah menjadi numerik, yaitu ‘penghasilan_ayah_nilai’, ‘penghasilan_ibu_nilai’, dan ‘penghasilan_wali_nilai’ seperti pada tabel 4.6 berikut ini.

Tabel 4.6 Transformasi Penghasilan

No	Penghasilan Ayah	penghasilan_ayah_nilai	Penghasilan Ibu	penghasilan_ibu_nilai	Penghasilan Wali	penghasilan_wali_nilai
0	Rp. 1,000,000 – Rp. 1,999,999	1500000	Tidak Berpenghasilan	0	NaN	NaN
1	Rp. 2,000,000 – Rp. 4,999,999	3500000	Rp. 500,000 – Rp. 999,999	750000	NaN	NaN
2	Rp. 1,000,000 – Rp. 1,999,999	1500000	Rp. 1,000,000 – Rp. 1,999,999	1500000	NaN	NaN
3	Rp. 2,000,000 – Rp. 4,999,999	3500000	Rp. 2,000,000 – Rp. 4,999,999	3500000	NaN	NaN

4.2.3 *Feature Engineering* Penghasilan Keluarga

Feature engineering adalah teknik membentuk atribut baru. Dalam penelitian ini, atribut Penghasilan Ayah, Penghasilan Ibu, dan Penghasilan Wali yang sudah ditransformasikan menjadi numerik, dibentuk atribut baru

yaitu ‘penghasilan_keluarga’. Atribut ini yang nantinya akan digunakan dalam proses *clustering*.

Pembentukan atribut tersebut dilakukan berdasarkan ketersediaan data penghasilan ayah, ibu, dan wali. Jika data ayah dan ibu memiliki penghasilan, maka penghasilan dari keduanya dijumlahkan sebagai nilai penghasilan keluarga. Jika hanya salah satu dari penghasilan orang tua yang terisi dan bernilai lebih dari 0, maka penghasilan yang tersedia tersebut digunakan sebagai penghasilan keluarga, tetapi diberi penanda sebagai data ‘missing_penghasilan_parsial’. Hal ini dilakukan karena data kosong tidak dapat langsung disamakan dengan kategori ‘Tidak Berpenghasilan’.

Data dengan kondisi penghasilan orang tua parsial dan nilai yang tersedia sebesar 0 dikeluarkan dari proses analisis. Hal ini dilakukan karena kondisi tersebut tidak cukup merepresentasikan penghasilan keluarga. Jika nilai 0 dipertahankan, data kosong dapat salah dimaknai sebagai tidak berpenghasilan. Sebaliknya, jika diisi menggunakan median, nilai tersebut dapat menimbulkan bias karena bukan merupakan penghasilan asli. Oleh karena itu, data tersebut dihapus agar atribut penghasilan_keluarga yang digunakan dalam proses *clustering* lebih representatif.

Tabel 4.7 Aturan Pembentukan Atribut

Kondisi Data Penghasilan	**Penanganan**
Penghasilan ayah dan ibu tersedia	Penghasilan ayah dan ibu dijumlahkan
Hanya penghasilan ayah atau ibu yang tersedia dan bernilai lebih dari 0	Penghasilan yang tersedia digunakan dan diberi penanda parsial
Hanya penghasilan ayah atau ibu yang tersedia tetapi bernilai 0, sedangkan penghasilan orang tua lainnya kosong	Di *drop*
Penghasilan ayah dan ibu kosong, tetapi penghasilan wali tersedia	Penghasilan wali digunakan
Penghasilan ayah dan ibu tidak berpenghasilan, tetapi wali memiliki penghasilan	Penghasilan wali digunakan
Penghasilan ayah, ibu, dan wali kosong	Diberi penanda *missing* penghasilan total

Berdasarkan tabel 4.7, pembentukan atribut penghasilan keluarga tidak hanya dilakukan penjumlahan penghasilan ayah dan ibu saja, melainkan juga

mempertimbangkan penghasilan wali. Hal ini dilakukan karena wali dapat menjadi pihak yang menanggung kebutuhan ekonomi siswa. Apabila penghasilan ayah dan ibu tidak tersedia, sedangkan penghasilan wali tersedia, maka penghasilan wali digunakan sebagai representasi penghasilan keluarga. Penggunaan penghasilan wali tidak dimaksudkan untuk menyimpulkan bahwa ayah dan ibu tidak berpenghasilan, tetapi karena informasi penghasilan yang tersedia pada data Dapodik hanya terdapat pada kolom wali. Pada hal ini data tersebut diberi penanda 'penghasilan_wali_digunakan' untuk menunjukkan bahwa nilai penghasilan keluarga berasal dari wali.

Selain itu, apabila penghasilan ayah dan ibu sama-sama tercatat 'Tidak Berpenghasilan', tetapi penghasilan wali tersedia dan memiliki nilai lebih dari 0, maka penghasilan wali digunakan sebagai representasi penghasilan keluarga. Sementara itu, jika semuanya tidak tersedia, maka data tersebut diberi penanda sebagai 'missing_penghasilan_total'. Penanda ini digunakan untuk membedakan data yang benar-benar kosong dengan data yang memang tercatat 'Tidak Berpenghasilan'.

Tabel 4.8 Keterangan Atribut

Atribut	Fungsi
penghasilan_keluarga	Atribut yang digunakan dalam proses *clustering.*
penghasilan_wali_digunakan	Penanda bahwa penghasilan keluarga berasal dari penghasilan wali
missing_penghasilan_parsial	Penanda bahwa data penghasilan orang tua hanya terisi sebagian
missing_penghasilan_total	Penanda bahwa penghasilan ayah, ibu, dan wali kosong semua.

Berdasarkan tabel 4.8, atribut utama yang digunakan dalam proses *clustering* adalah 'penghasilan_keluarga'. Atribut lainnya yang ada di tabel 4.8 merupakan atribut penanda pada tahap pra-pemrosesan data.

```
Missing penghasilan keluarga setelah penanganan:
2

Status penghasilan:
status_penghasilan
Data penghasilan tersedia            307
Data penghasilan orang tua parsial     5
Menggunakan penghasilan wali           1
Name: count, dtype: int64
```

Gambar 4.1 Informasi Atribut Penghasilan

Berdasarkan Gambar 4.1, ditemukan missing penghasilan keluarga setelah penanganan ada 2 data siswa, yang artinya salah satu data orang tua kosong, sedangkan data orang tua yang tersedia bernilai 0. Dalam hal ini data akan di drop.

Tabel 4.9 Kondisi Awal Data

Kondisi Data	Jumlah Data
Data penghasilan tersedia	307
Data penghasilan orang tua parsial	5
Menggunakan penghasilan wali	1
Penghasilan, ayah, ibu, dan wali kosong	0

Berdasarkan Gambar 4.1 dan Tabel 4.9, sebanyak 307 data memiliki penghasilan yang tersedia, 5 data memiliki penghasilan orang tua parsial. Hal ini dapat dilihat pada tabel dibawah ini.

Tabel 4.10 Penghasilan Parsial

	Penghasilan Ayah	Penghasilan Ibu	Penghasilan Wali	penghasilan_ayah_nilai	penghasilan_ibu_nilai	penghasilan_keluarga	missing_penghasilan_parsial	missing_penghasilan_parsial_nol
41	NaN	Rp. 2,000,000-Rp.4,999,999	NaN	NaN	3500000	3500000	1	0
54	NaN	Rp. 2,000,000-Rp.4,999,999	NaN	NaN	3500000	3500000	1	0
75	NaN	Tidak Berpenghasilan	NaN	NaN	0	NaN	1	1
165	NaN	Rp. 500,000-Rp.999,999	NaN	NaN	750000	750000	1	0
234	NaN	Tidak Berpenghasilan	NaN	NaN	0	NaN	1	1

Berdasarkan Tabel 4.10 terlihat bahwa penghasilan pada data index 75 dan 234 merupakan *missing* parsial berjumlah nol, yang dimana penghasilan keluarga tercatat NaN. Kondisi ini memungkinkan interpretasi penghasilan keluarga menjadi kurang tepat, karena data penghasilan ayah sebenarnya tidak tersedia. Oleh karena itu, data dengan kondisi salah satu penghasilan orang tua kosong dan penghasilan orang tua lainnya bernilai 0 dipertimbangkan untuk dikeluarkan dari proses analisis. Hal ini dilakukan karena data kosong tidak dapat langsung disamakan dengan kategori "Tidak Berpenghasilan".

```
Jumlah data setelah penghapusan: 311
```

Gambar 4.2 Jumlah Data

Jumlah data yang awalnya 313 baris sekarang menjadi 311 baris. Hal ini dikarenakan 2 baris dihapus karena bernilai kosong dan bernilai 0.

Pada tabel 4.9, terdapat 1 data menggunakan penghasilan wali sebagai representasi penghasilan keluarga. Tidak ditemukan data dengan penghasilan ayah, ibu, dan wali kosong (NaN). Hasil ini menunjukkan bahwa mayoritas data dengan penghasilan keluarga dapat dibentuk dari data penghasilan ayah dan/atau ibu, sedangkan sebagian kecil data memerlukan penanda tambahan karena bersifat parsial atau menggunakan wali.

Adapun contoh hasil pembentukan atribut penghasilan_keluarga ditunjukkan pada tabel 4.11 dibawah ini.

Tabel 4.11 Hasil Feature Engineering

Index	Penghasilan Ayah	Penghasilan Ibu	Penghasilan Wali	penghasilan_keluarga	missing_penghasilan_parsial	missing_penghasilan_total	penghasilan_wali_digunakan
0	Rp. 1,000,000 – Rp. 1,999,999	Tidak Berpenghasilan	NaN	1500000	0	0	0
1	Rp. 2,000,000 – Rp. 4,999,999	Rp. 500,000 – Rp. 999,999	NaN	4250000	0	0	0
41	NaN	Rp. 2,000,000 – Rp. 4,999,999	NaN	3500000	1	0	0

Index	Penghasilan Ayah	Penghasilan Ibu	Penghasilan Wali	penghasilan_keluarga	missing_penghasilan_parsial	missing_penghasilan_total	penghasilan_wali_digunakan
195	Tidak Berpenghasilan	Tidak Berpenghasilan	Rp. 1,000,000 – 1,999,999	1500000	0	0	1

4.2.4 Transformasi Binner

Transformasi biner dilakukan pada atribut yang memiliki dua kategori jawaban, yaitu Ya dan Tidak. Dalam penelitian ini, atribut yang ditransformasikan menjadi bentuk biner adalah atribut Penerima KPS, Penerima KIP dan atribut Layak PIP (usulan dari sekolah). Transformasi ini dilakukan karena algoritma K-Means memerlukan data dalam bentuk numerik untuk menghitung jarak antar data.

Kategori Ya diberi nilai 1, sedangkan kategori Tidak diberi nilai 0. Nilai 1 menunjukkan bahwa siswa tercatat sebagai penerima atau pemilik atribut tersebut, sedangkan nilai 0 menunjukkan bahwa siswa tidak tercatat sebagai penerima atau pemilik atribut tersebut.

Tabel 4.12 Hasil Transformasi Biner

	Penerima KPS	penerima_kps	Penerima KIP	penerima_kip	Layak PIP (usulan dari sekolah)	status_layak_pip
0	Tidak	0	Tidak	0	Tidak	0
1	Tidak	0	Tidak	0	Tidak	0
...	...	...	...	...	...	...
309	Tidak	0	Tidak	0	Ya	1
310	Tidak	0	Tidak	0	Ya	1

Pada tabel 4.12 hasil transformasi biner untuk atribut Penerima KPS, Penerima KIP, dan atribut Layak PIP. Hasil transformasi ini kemudian digunakan dalam proses *clustering*, kecuali atribut Layak PIP yang akan digunakan sebagai analisis setalah terbentuk *cluster*.

4.2.5 Normalisasi Data

Normalisasi data dilakukan setalah seluruh atribut telah menjadi numerik. Tahap ini bertujuan untuk menyamakan skala antar atribut agar tidak

ada atribut tertentu yang mendominasi proses perhitungan jarak. Hal ini penting karena algoritma K-Means menggunakan jarak antar data dalam proses pembentukan *cluster*.

Normalisasi dilakukan menggunakan metode *Min-Max Normalization* yang mengubah nilai setiap atribut kedalam rentang 0 sampai 1. Normalisasi diperlukan karena atribut penghasilan keluarga memiliki skala nilai yang jauh lebih besar dibandingkan atribut lain seperti jumlah saudara kandung, penerima KPS, dan penerima KIP. Tanpa normalisasi, atribut penghasilan dapat lebih dominan dalam perhitungan jarak dibandingkan atribut lainnya.

Atribut yang digunakan dalam proses *clustering* adalah penghasilan_keluarga, jumlah_saudara, penerima_kps, dan penerima_kip. Sementara itu, atribut Layak PIP (usulan dari sekolah) dan Alasan Layak PIP tidak digunakan dalam proses *clustering*, melainkan digunakan setelah *cluster* terbentuk untuk analisis distribusi status dan alasan Layak PIP pada masing-masing *cluster*.

Tabel 4.13 Hasil Normalisasi Min-Max

	penghasilan_keluarga	jumlah_saudara	penerima_kps	penerima_kip
0	0.214286	0.125	0.0	0.0
1	0.607143	0.125	0.0	0.0
2	0.428571	0.125	0.0	0.0
3	0.500000	0.250	0.0	0.0
4	0.500000	0.375	0.0	0.0

Berdasarkan Tabel 4.13, seluruh atribut yang digunakan dalam proses *clustering* telah berada pada rentang 0 sampai 1. Hasil normalisasi ini kemudian digunakan sebagai data masukan dalam proses *clustering* menggunakan algoritma K-Means. Dengan demikian, setiap atribut memiliki skala yang sebanding dalam perhitungan jarak antar data.

4.3 Penentuan Jumlah *Cluster*

Penentuan jumlah *k* menggunakan metode *silhouette coefficient.* Metode ini menghitung koefisien siluet pada beberapa jumlah *cluster* untuk melihat kualitas pemisahan antar *cluster*. Pada tahap ini, algoritma K-Means dijalankan pada beberapa nilai *k*, kemudian setiap hasil *cluster* yang terbentuk dihitung menggunakan koefisien Silhouette.

```
Jumlah cluster terbaik: 3
Silhouette Score terbaik: 0.695
```

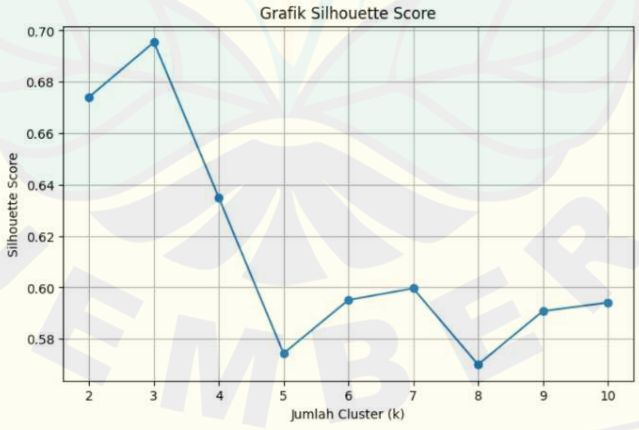
Gambar 4.3 Hasil Silhouette Terbaik

Berdasarkan Gambar 4.3 diperoleh hasil *k* terbaik yaitu 3. Oleh karena itu, jumlah *cluster* yang digunakan dalam implementasi algoritma K-Means pada penelitian ini adalah 3 *cluster*.

Tabel 4.14 Evaluasi *Cluster*

	Jumlah *Cluster*	Silhouette Score
0	2	0.674
1	3	0.695
...	...	...
7	9	0.591
8	10	0.594

9 rows × 3 columns



Gambar 4.4 Grafik Silhouette

Setelah jumlah *k* dipilih, dilakukan evaluasi *k*. Berdasarkan Gambar 4.3 dan tabel 4.14 menunjukkan jumlah *cluster* = 3 memiliki nilai silhouette score tertinggi

yang mendekati 1, dengan nilai 0,695. Hal ini menunjukkan bahwa pembentukan tiga *cluster* menghasilkan pemisahan antar-*cluster* yang lebih baik dibandingkan jumlah *cluster* lainnya.

4.4 Implementasi K-Means

Atribut *clustering* dijalankan oleh algoritma K-Means untuk menghasilkan kelompok dengan profil sosial ekonomi yang mirip. Algoritma K-Means melakukan inisialisasi *centroid* sebanyak jumlah *cluster* yang telah dipilih. Setelah *centroid* awal terbentuk, setiap data siswa dihitung jaraknya terhadap masing-masing *centroid*. Jarak *centroid* terdekat akan dijadikan satu kelompok. *Centroid* terus diperbarui dengan menghitung rata-rata anggota pada masing-masing *cluster*. Proses ini dilakukan secara iteratif sampai hasil konvergen.

```
Jumlah iterasi sampai konvergen: 3
SSE akhir: 25.428809098043946
```

Gambar 4.5 Iterasi dan SSE Akhir

Berdasarkan hasil implementasi, proses K-Means mencapai konvergensi pada iterasi ke-3 dengan nilai SSE sebesar 25.5. Nilai SSE menunjukkan total jarak kuadrat data terhadap centroid pada masing-masing *cluster*.

Tabel 4.15 Centroid Akhir Skala Normalisasi

	penghasilan_keluarga	jumlah_saudara	penerima_kps	penerima_kip
0	0.374150	0.167163	-7.771561e-16	-8.326673e-17
1	0.205882	0.323529	6.470588e-01	1.000000e+00
2	0.305272	0.187500	1.000000e+00	3.469447e-17

Tabel 4.16 Centroid Akhir Skala Asli

	penghasilan_keluarga	jumlah_saudara	penerima_kps	penerima_kip
0	2619048.0	1.34	0.00	0.0
1	1441176.0	2.59	0.65	1.0
2	2136905.0	1.50	1.00	0.0

Berdasarkan tabel 4.16 tersebut setiap *cluster* memiliki nilai *centroid* yang berbeda-beda pada masing-masing atribut. Pada atribut numerik seperti penghasilan_keluarga dan jumlah_saudara, *centroid* menunjukkan nilai rata-rata anggota *cluster*. Pada atribut biner seperti penerima_kps dan penerima_kip, nilai *centroid* menunjukkan kecenderungan atau proporsi anggota *cluster*. Nilai

mendekati 1 menunjukkan sebagian besar anggota *cluster* memiliki kategori "Ya", sedangkan nilai mendekati 0 menunjukkan sebagian besar anggota *cluster* memiliki kategori "Tidak".

Hasil implementasi K-Means menghasilkan tiga *cluster* dengan jumlah anggota yang berbeda-beda. Berikut ini tabel hasil implementasi *cluster* yang menunjukkan jumlah siswa per-*cluster* dan persentase setiap *cluster*.

Tabel 4.17 Hasil *Cluster*

Cluster	**Jumlah Siswa**	**Persentase**
0	252	81.03
1	17	5.47
2	42	13.30

Tabel 4.17 menunjukkan bahwa Cluster 0 memiliki jumlah data terbanyak yaitu 253 siswa dengan persentase 81.03% dan Cluster 1 memiliki jumlah data paling kecil yaitu 17 siswa dengan persentase 5.47%. Perbedaan jumlah anggota pada setiap *cluster* menunjukkan bahwa pola profil sosial ekonomi siswa dalam data Dapodik memiliki kecenderungan yang berbeda-beda. Cluster 0 dengan jumlah anggota terbesar menunjukkan profil sosial ekonomi yang paling banyak muncul pada data, sedangkan Cluster 1 dengan jumlah anggota lebih kecil menunjukkan kelompok siswa dengan karakteristik yang lebih spesifik. Pada Cluster 2 berada di tengah-tengah atau sedang.

4.5 Profiling Karakteristik Hasil *Cluster*

Profiling dilakukan dengan melihat pada nilai centroid akhir dalam skala asli pada tabel 4.16 Centroid Akhir Skala Asli. Nilai centroid digunakan untuk melihat kecenderungan karakteristik sosial ekonomi pada masing-masing *cluster*.

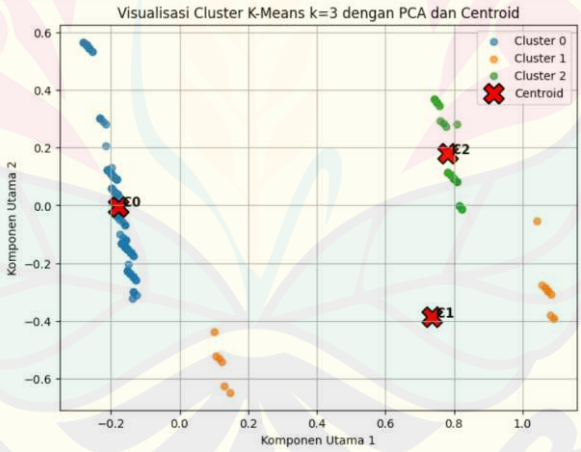
Berdasarkan tabel 4.16, Cluster 0 memiliki rata-rata penghasilan keluarga paling tinggi sebesar Rp. 2.619.048 dengan rata-rata jumlah saudara sebesar 1,34. Nilai centroid pada atribut penerima_kip dan penerima_kps cenderung mendekati 0 yang artinya pada *cluster* ini cenderung tidak tercatat sebagai penerima KPS, dan KIP.

Cluster 1 memiliki rata-rata penghasilan kelurga paling rendah, yaitu sebesar Rp. 1.441.176 dengan rata-rata jumlah saudara tertinggi yaitu 2,59. Nilai centroid pada 'penerima_kps' sebesar 0,65 dan 'penerima_kip' sebesar 1,00 yang artinya sebagian besar anggota *cluster* memiliki keterkaitan dengan KPS dan menunjukkan bahwa seluruh anggota *cluster* memiliki KIP.

Cluster 2 memiliki rata-rata penghasilan keluarga sebesar Rp. 2.136.905 dengan rata-rata jumlah saudara sekitar 1,50. Nilai centroid 'penerima_kps' mencapai 1,00 menunjukkan bahwa seluruh anggota *cluster* tercatat sebagai penerima KPS. Sedangkan nilai centroid pada 'penerima_kip' mendekati 0,00 yang artinya menunjukkan bahwa anggota Cluster 2 tidak tercatat sebagai penerima KIP.

4.6 Visualisasi dan Interpretasi Karakteristik *Cluster*

Visualisasi data digunakan untuk melihat perbedaan karakteristik serta memahami hasil pengelompokkan. Hal ini bertujuan dalam proses interpretasi hasil *clustering* sehingga kesimpulan yang diperoleh menjadi lebih mudah dipahami.



Gambar 4.6 Visualisasi Menggunakan Scatter Plot PCA

Hasil visualisasi menggunakan scatter plot PCA digunakan untuk melihat sebaran data hasil *clustering* dalam bentuk dua dimensi. Berdasarkan visualisasi PCA, centroid Cluster 1 tampak tidak berada tepat di tengah sebaran titik data. Hal ini disebabkan karena proses *clustering* dilakukan pada ruang empat dimensi, sedangkan visualisasi PCA hanya merupakan proyeksi dua dimensi yang

mempertahankan sebagian besar variasi data. Posisi centroid pada PCA tidak selalu sama dengan posisi centroid sebenarnya pada ruang fitur asli.

Berdasarkan scatter plot PCA terlihat bahwa sebagian besar data terkonsentrasi pada satu kelompok besar dengan beberapa kelompok yang lebih kecil. Sebaran data tersebut memengaruhi proses pembentukan *cluster* sehingga jumlah *cluster* optimal yang diperoleh adalah tiga *cluster*. Jumlah anggota Cluster 1 yang relatif sedikit menyebabkan centroid lebih sensitif terhadap karakteristik anggotanya dibandingkan *cluster* lain. Berdasarkan visualisasi tersebut, hasil scatter plot menggunakan PCA kurang kuat dalam mendefinisikan atau interpretasi karakteristik antarcluster. Oleh karena itu, penelitian ini menggunakan visualisasi tambahan yaitu Chernoff Face.

Visualisasi Chernoff Face digunakan sebagai visualisasi pendukung karena mampu merepresentasikan beberapa atribut centroid secara bersamaan dalam bentuk karakteristik wajah. Visualisasi ini mempermudah interpretasi perbedaan karakteristik antarcluster yang tidak selalu terlihat jelas pada scatter plot PCA.

Tabel 4.18 Pemetaan Atribut Chernoff

Atribut	Elemen Wajah	Makna Visual
Penghasilan keluarga	Ukuran wajah	Semakin besar wajah, semakin tinggi penghasilan keluarga
Jumlah saudara	Ukuran telinga	Semakin besar telinga, jumlah saudara semakin banyak
Penerima KIP	Ukuran mata	Semakin besar mata, semakin tinggi jumlah kecenderungan penerima KIP
Penerima KPS	Bentuk mulut	Semakin melengkung mulut, semakin tinggi kecenderungan penerima KPS

Gambar 4.7 Visualisasi Chernoff

Berdasarkan gambar 4.7 dan tabel 4.18, Cluster 0 ditampilkan dengan ukuran wajah terbesar yang artinya penghasilan pada Cluster 0 merupakan yang paling tinggi dibandingkan *cluster* lain. Cluster 1 menunjukkan ukuran telinga yang besar sehingga menunjukkan bahwa Cluster 1 memiliki jumlah saudara terbanyak. Cluster 1 juga memiliki ukuran mata yang besar yang artinya jumlah penerima KIP terbanyak. Bentuk mulut yang sangat melengkung terlihat pada Cluster 2 yang menandakan bahwa jumlah penerima KPS terbanyak.

Kondisi pada Cluster 2 dapat dipahami sebagai pola khusus. Rata-rata penghasilan keluarga pada Cluster 2 cukup tinggi tetapi bukan yang tertinggi, diantara ketiga Cluster. Rata-rata jumlah saudara pada Cluster 2 relatif rendah yang dilihat dari ukuran telinga yang relatif kecil, sehingga beban tanggungan keluarga tidak sebesar Cluster 1. Meskipun seluruh anggota Cluster 2 tercatat sebagai penerima KPS, atribut penghasilan keluarga dan jumlah saudara menunjukkan bahwa *cluster* ini tidak dapat dikategorikan sebagai kelompok dengan kerentanan paling tinggi. Hal ini didukung oleh distribusi status usulan PIP yang menunjukkan bahwa sebagian besar anggota Cluster 2 tidak berstatus Layak PIP. Titik berat visualisasi Chernoff dalam penelitian ini adalah kepesertaan bantuan sosial, terutama penerima KPS, dengan dukungan atribut penerima KIP, penghasilan keluarga, dan jumlah saudara. Oleh karena itu, Cluster 2 lebih tepat diinterpretasikan sebagai kelompok sosial ekonomi sedang dengan keterkaitan kuat terhadap KPS, bukan sebagai kelompok yang sepenuhnya tidak rentan.

Visualisasi Chernoff Face ini memperkuat hasil profiling *cluster*. Cluster 0 menunjukkan profil sosial ekonomi relatif lebih baik, Cluster 1 menunjukkan profil sosial ekonomi rentan, sedangkan Cluster 2 menunjukkan profil sosial ekonomi sedang menuju mampu dengan keterkaitan kuat terhadap KPS. Dengan demikian, visualisasi Chernoff Face membantu menjelaskan bahwa masing-masing *cluster* memiliki karakteristik sosial ekonomi yang berbeda berdasarkan kombinasi atribut penghasilan keluarga, jumlah saudara, penerima KIP, dan penerima KPS.

Berdasarkan hasil *profiling* dan visualisasi, setiap *cluster* diberi label interpretatif. Pelabelan ini tidak dimaksudkan sebagai status ekonomi resmi siswa, melainkan sebagai deskripsi pola sosial ekonomi yang terbentuk dari hasil *clustering*.

Tabel 4.19 Label *Cluster*

Cluster	**Karakteristik Utama**	**Label Interpretatif**
0	Penghasilan keluarga relatif tinggi, jumlah saudara rendah, penerima KPS dan KIP rendah	Sosial ekonomi relatif baik (relatif mampu).
1	Penghasilan keluarga relatif kecil, jumlah saudara tinggi, sebagian besar penerima KPS dan penerima KIP tinggi.	Sosial ekonomi rentan.
2	Penghasilan keluarga sedang, jumlah saudara relatif sedikit, penerima KPS tinggi, dan penerima KIP rendah	Sosial ekonomi sedang menuju mampu

Berdasarkan tabel 4.19 diperoleh 3 *cluster*, yaitu Cluster 0, Cluster 1, dan Cluster 2. Cluster 1 merupakan *cluster* paling rentan yang dilihat berdasarkan karakteristik utama, dan *cluster* relatif mampu berada pada Cluster 0, sedangkan Cluster 2 diberi label *cluster* dengan sosial ekonomi sedang. Dengan demikian, KPS dapat menjadi indikator kerentanan sosial, tetapi tidak selalu sejalan langsung dengan penghasilan keluarga paling rendah. Kondisi ini menunjukkan bahwa status bantuan sosial seperti KPS dapat dipengaruhi oleh faktor administratif atau kondisi sosial ekonomi lain yang tidak seluruhnya tercermin dalam atribut penghasilan keluarga.

4.7 Distribusi Status Usulan Layak PIP dan Alasan Layak PIP

Setelah diperoleh label dari hasil *clustering,* label tersebut digabungkan kembali dengan atribut Layak PIP (usulan dari sekolah) dan Alasan Layak PIP. Analisis ini bertujuan untuk melihat bagaimana pola siswa berstatus layak PIP dan tidak layak PIP tersebar pada *cluster* yang telah terbentuk berdasarkan profil sosial ekonomi.

4.7.1 Distribusi Status Usulan PIP

Distribusi Status Usulan PIP untuk melihat persebaran pola kelompok siswa per *cluster*.

Tabel 4.20 Distribusi Status Usulan Layak PIP

Layak PIP (usulan dari sekolah)	Tidak	Ya
cluster		
0	217 (86.11%)	35 (13.89%)
1	0 (0.0%)	17 (100.0%)
2	29 (69.05%)	13 (30.95%)

Dari hasil tabel 4.20 tersebut, Cluster 0 memiliki jumlah Tidak Layak PIP tertinggi dibanding *cluster* lain. Hal ini sejalan dengan pembahasan sebelumnya pada 4.6 bahwa Cluster 0 menunjukkan karakteristik yang relatif mampu. Pada Cluster 0 juga beberapa diusulkan layak, yaitu sebesar 13.89% sebanyak 35 siswa.

Pada Cluster 1 menunjukkan persentase penerima PIP paling tinggi, yaitu 100% anggota *cluster* layak diusulkan PIP. Hal ini juga sejalan dengan karakteristik Cluster 1 yang rentan dalam sosial ekonomi. Sedangkan pada Cluster 2 menunjukkan sebanyak 13 orang layak diusulkan PIP, dan yang 30 orang tidak. Pada Cluster 2 memiliki pola yang berbeda karena pada Cluster 2 memiliki keterkaitan dengan penerimaan KPS (bantuan sosial), tetapi tidak tercatat sebagai penerima KIP. Kondisi ini menunjukkan bahwa *clustering* tidak bersifat mutlak dalam menentukan status PIP, karena algoritma K-Means hanya mengelompokkan siswa berdasarkan kemiripan atribut sosial ekonomi yang digunakan.

4.7.2 Distribusi Alasan Layak PIP

Distribusi Alasan Layak PIP digunakan untuk melihat kategori alasan apa yang menjadi dominan pada siswa berstatus Layak PIP dalam masing-masing *cluster*, misalnya siswa miskin/rentan miskin, pemegang PKH/KPS/KKS, yatim/piatu/panti sosial, atau alasan lain yang tercatat pada data.

Tabel 4.21 Distribusi Alasan Usulan Layak PIP

alasan_layak_pip	Alasan tidak tercatat	Menolak	Pemegang PKH/KPS/ KKS	Siswa Miskin/Rentan Miskin	Tidak Layak PIP	Yatim piatu/panti asuhan/panti sosial
cluster						
0	0 (0.0)	3 (1.19)	5 (1.98)	28 (11.11)	210 (83.33)	6 (2.38)
1	17 (100.0)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
2	0 (0.00)	1 (2.38)	13 (30.95)	0 (0.00)	28 (66.67)	0 (0.00)

Berdasarkan tabel 4.21 dan pembahasan pada 4.6, Cluster 0 didominasi oleh kategori Tidak Layak PIP. Terdapat total siswa 210 dengan persentase 83% yang tidak layak PIP. Meskipun didominasi oleh kategori tidak layak PIP, tetapi pada Cluster 0 masih terdapat siswa dengan alasan layak PIP, yaitu 5 orang siswa pemegang PKH/KPS/KKS, 28 orang siswa misikin/rentan miskin, dan 6 orang yatim piatu/panti asuhan/ pasnti sosial. Selain itu, ada 3 orang siswa dengan alasan menolak. Hal ini menunjukkan meskipun Cluster 0 memiliki karakteristik sosial ekonomi relatif baik, masih terdapat beberapa siswa yang mempunyai alasan tertentu dalam usulan PIP.

Cluster 1 memiliki jumlah anggota 17 siswa, dengan karakteristik paling rentan. Siswa pada Cluster 1 memiliki usulan Layak PIP namun alasannya tidak tercatat. Hal ini dikarenakan *cluster* mengelompokkan berdasarkan atribut profil sosial ekonomi.

Cluster 2 memiliki 28 siswa pada kategori Tidak Layak PIP, 13 siswa dengan alasan Pemegang PKH/KPS/KKS, dan 1 siswa dengan kategori Menolak. Hal ini menunjukkan bahwa Cluster 2 memiliki keterkaitan dengan alasan bantuan sosial, terutama karena pada hasil profiling sebelumnya Cluster 2 memiliki nilai penerima KPS yang tinggi.

Distribusi alasan Layak PIP tidak digunakan dalam proses *clustering*, melainkan digunakan untuk interpretasi dan menemukan pola distribusi alasan

Layak PIP pada masing-masing kelompok siswa. Dengan demikian, hasil ini berfungsi sebagai informasi tambahan untuk memahami karakteristik siswa pada setiap *cluster*, bukan sebagai dasar penentuan kelayakan PIP.

4.8 Pengujian Asosiasi dan Korelasi Atribut

Pengujian ini dilakukan dengan menguji atribut sosial ekonomi dengan status usulan Layak PIP. Pengujian bertujuan untuk mengetahui atribut manakah yang berasosiasi dan berkorelasi dengan status usulan Layak PIP. Pengujian ini tidak untuk menentukan kelayakan PIP melainkan sebagai analisis dalam mendukung pola hubungan antara atribut sosial ekonomi siswa dengan status usulan layak PIP.

4.8.1 Uji Asosiasi Atribut Kategorik

Uji asosiasi bertujuan melihat apakah 2 variabel memiliki keterkaitan atau hubungan. Uji asosiasi dilakukan pada atribut kategorik yaitu penerima KPS dan penerima KIP. Dalam penelitian ini menggunakan Chi-Square untuk melihat hubungan antar variabel.

Tabel 4.22 Kontingensi Penerima KIP dengan Status Usulan PIP

Layak PIP (usulan dari sekolah)	Tidak	Persentase	Ya	Persentase
Penerima KIP				
Tidak	246	83.67%	48	16.33%
Ya	0	0.00%	17	100.0%

Berdasarkan tabel 4.22 siswa yang tercatat tidak menerima KIP sebanyak 246 siswa berstatus tidak layak diusulkan PIP dan 48 siswa diusulkan layak PIP, sedangkan 17 siswa yang tercatat penerima KIP seluruhnya berstatus usulan Layak PIP. Hal ini menunjukkan masih terdapat siswa yang tidak memiliki KIP tetapi tetap berstatus Layak PIP, sehingga status Layak PIP tidak hanya berkaitan dengan kepemilikan KIP, tetapi juga dapat dipengaruhi oleh alasan atau kondisi sosial ekonomi lain yang tercatat pada data.

Tabel 4.23 Kontingensi Penerima KPS dengan Status Usulan PIP

Layak PIP (usulan dari sekolah)	Tidak	Persentase	Ya	Persentase
Penerima KPS				
Tidak	217	84.11%	41	15.89%
Ya	29	54.72%	24	45.28%

Berdasarkan tabel 4.23, siwa yang tidak menerima KPS berjumlah 258, yang diantaranya sebanyak 217 tidak diusulkan layak PIP, sedangkan 41 berstatus layak PIP. Sementara itu, siswa yang tercatat memiliki KPS berjumlah 53, yang diantaranya 29 berstatus tidak layak usulan PIP, dan 24 berstatus layak PIP. Berdasarkan hasil tersebut, secara proporsi, siswa penerima KPS memiliki persentase Layak PIP yang lebih tinggi dibandingkan siswa yang tidak menerima KPS. Hal ini menunjukkan adanya kecenderungan bahwa kepesertaan KPS berkaitan dengan status usulan Layak PIP.

Berikut ini merupakan hasil Chi-Square yang diperoleh untuk atribut penerima KPS dan KIP.

```
Chi-Square KPS: 21.231378018411938
p-value: 4.070458388153338e-06
df: 1
```

Gambar 4.8 Hasil Chi-Square KPS

```
Chi-Square KIP: 63.09153304536324
p-value: 1.9732048796914324e-15
df: 1
```

Gambar 4.9 Hasil Chi-Square KIP

Berdasarkan hasil tersebut, menunjukkan bahwa p-value lebih kecil dari 0,05 yang artinya terdapat asosiasi signifikan. Hasil ini tidak menunjukkan hubungan sebab-akibat, melainkan hanya menunjukkan adanya kecenderungan hubungan antara kepesertaan KPS dan penerima KIP dengan status usulan Layak PIP pada data penelitian.

4.8.2 Uji Korelasi Atribut Numerik

Uji korelasi atribut numerik bertujuan untuk melihat hubungan dan arah hubungannya (positif atau negatif). Uji korelasi dilakukan pada atribut numerik

seperti penghasilan keluarga dan jumlah saudara. Dalam penelitian, untuk menguji korelasi menggunakan metode Spearman.

```
Spearman penghasilan: -0.3764994934473093
p-value: 6.545641817017468e-12
```

Gambar 4.10 Uji Korelasi Penghasilan

Berdasarkan Gambar 4.10, koefisien Spearman pada penghasilan keluarga bernilai negatif, dan p-value lebih kecil dari 0,05. Hal ini menunjukkan hubungan signifikan antara penghasilan keluarga dengan status usulan layak PIP. Selain itu, nilai koefisien -0,37 menunjukkan adanya arah hubungan negatif. Hal ini maksudnya adalah semakin tinggi penghasilan keluarga, maka terdapat kecenderungan semakin rendah status usulan Layak PIP dan begitupun sebaliknya.

```
Spearman jumlah saudara: 0.2132069953927706
p-value: 0.0001515864073026811
```

Gambar 4.11 Uji Korelasi Jumlah Saudara

Berdasarkan Gambar 4.11, atribut jumlah saudara memperoleh koefisien Spearman sebesar 0,213 dengan p-value sebesar 0,00015. Nilai p-value yang lebih kecil dari 0,05 menunjukkan bahwa jumlah saudara memiliki hubungan yang signifikan dengan status usulan Layak PIP. Koefisien positif menunjukkan bahwa semakin banyak jumlah saudara, maka terdapat kecenderungan semakin tinggi status usulan Layak PIP. Namun, nilai koefisien yang relatif kecil menunjukkan bahwa hubungan tersebut bersifat lemah. Berikut ini ditampilkan tabel hasil asosiasi dan korelasi.

Tabel 4.24 Hasil Uji Korelasi dan Asosiasi

Atribut	**Jenis Uji**	**Nilai Statistik**	**p-value**	**Keterangan**
Penerima KPS	Chi-Square	21.231378	4.070458e-06	Signifikan
Penerima KIP	Chi-Square	63.091533	1.973205e-15	Signifikan
Penghasilan keluarga	Spearman	-0.376499	6.545642e-12	Signifikan
Jumlah saudara kandung	Spearman	0.213207	1.515864e-04	Signifikan

Hasil ini tidak dikaitkan sebagai hubungan sebab-akibat, melainkan sebagai kecenderungan antarvariabel. Berdasarkan hasil profiling *cluster*, distribusi status usulan Layak PIP, serta uji asosiasi dan korelasi, diketahui bahwa atribut penerima KPS, penerima KIP, penghasilan keluarga, dan jumlah saudara menunjukkan kecenderungan berasosiasi dengan status usulan Layak PIP pada masing-masing *cluster* dan dapat digunakan sebagai informasi pendukung dalam mendeskripsikan pola sosial ekonomi siswa terhadap status usulan Layak PIP.

Cluster 1 menunjukkan dominasi atribut penghasilan keluarga yang rendah, jumlah saudara yang tertinggi, dan penerima KIP cukup tinggi sehingga mencerminkan kelompok yang paling rentan, sebaliknya Cluster 0 didominasi oleh penghasilan keluarga tertinggi dan penerima KIP dan KPS tidak dominan, sehingga mencerminkan kelompok yang mampu. Cluster 2 ditandai oleh dominasi atribut penerima KPS dengan tingkat penghasilan keluarga tertinggi dan jumlah saudara terendah setelah Cluster 1. Hasil ini didukung oleh hasil uji Chi-Square dan Spearman yang menunjukkan bahwa keempat atribut tersebut saling berasosiasi signifikan terhadap usulan Layak PIP. Hasil tersebut digunakan sebagai pendukung dalam mendeskripsikan pola sosial ekonomi siswa, bukan sebagai dasar penentuan kelayakan PIP.

BAB 5. KESIMPULAN, KETERBATASAN, DAN SARAN

5.1 Kesimpulan

1. Pengelompokan profil sosial ekonomi siswa menggunakan algoritma K-Means melalui beberapa tahapan, yaitu dilakukan pembersihan data untuk mengetahui dan menangani nilai kosong, selanjutnya dilakukan transformasi penghasilan karena atribut penghasilan masih berbentuk rentang kategori. Hasil transformasi dari atribut penghasilan kemudian dilakukan *feature engineering* dan menghasilkan atribut baru yaitu penghasilan_keluarga. Setelah dilakukan transformasi penghasilan, selanjutnya dilakukan transformasi biner, dan normalisasi data. Penentuan jumlah *cluster* dilakukan menggunakan metode *silhouette coefficient* dengan nilai optimal k=3. Atribut yang digunakan dalam proses *clustering* adalah penghasilan keluarga, penerima KIP, penerima KPS, dan jumlah saudara. Setelah algoritma K-Means dijalankan, diperoleh hasil yaitu Cluster 0 sebanyak 252 siswa, Cluster 1 sebanyak sebanyak 17 siswa, dan Cluster 2 sebanyak 42 siswa.
2. Berdasarkan *cluster* yang terbentuk, pada Cluster 0 memiliki rata-rata penghasilan paling tinggi, jumlah saudara relatif rendah, serta cenderung tidak tercatat sebagai penerima KPS dan KIP. Pada Cluster 1 memiliki karakteristik penghasilan keluarga relatif paling rendah, jumlah saudara paling tinggi, seluruh anggota tercatat sebagai penerima KIP dan sebagian besar anggota tercatat sebagai penerima KPS. Pada Cluster 2 memiliki rata-rata penghasilan keluarga sedang, jumlah saudara sedang, seluruh anggota tercatat sebagai penerima KPS, dan sebagian tercatat sebagai penerima KIP. Dari karakteristik tersebut disimpulkan bahwa Cluster 0 merupakan kategori relatif mampu, Cluster 1 rentan, dan Cluster 2 sedang.
3. Berdasarkan analisis yang telah dilakukan, Cluster 1 memiliki proporsi usulan Layak PIP tinggi, namun alasan kelayakannya tidak tercatat. Sedangkan Cluster 0 didominasi oleh siswa yang tidak layak usulan PIP. Meskipun demikia, pada Cluster 0 masih terdapat siswa dengan alasan Layak

PIP, seperti siswa miskin/rentan miskin, pemegang PKH/KPS/KKS, dan yatim piatu/panti asuhan/panti sosial. Kondisi ini dapat terjadi karena proses *clustering* menggunakan atribut sosial ekonomi, bukan atribut Alasan Layak PIP. Oleh karena itu, siswa dengan alasan Layak PIP tertentu masih dapat berada pada Cluster 0 apabila pola atribut sosial ekonominya lebih mirip dengan karakteristik di Cluster 0. Sementara itu, pada Cluster 2 didominasi oleh siswa tidak layak usulan PIP, tetapi masih terdapat yang layak usulan PIP berupa pemegang PKH/KPS/KKS.

4. Berdasarkan uji asosiasi dan korelasi, semua atribut sosial ekonomi menunjukkan hubungan yang signifikan dengan status usulan Layak PIP. Atribut penerima KIP memiliki asosiasi paling kuat dibanding penerima KPS berdasarkan nilai Crame's V, sedangkan pada atribut numerik, penghasilan keluarga memiliki hubungan paling kuat yang berdasarkan pada nilai koefisien Spearman. Korelasi menunjukkan hubungan negatif, yang artinya semakin tinggi penghasilan keluarga, maka status usulan PIP rendah, begitupun sebaliknya. Atribut jumlah saudara memiliki hubungan positif, yang artinya semakin banyak jumlah saudara maka status usulan PIP cenderung tinggi.

5.2 Keterbatasan Penelitian

1. Data penghasilan orang tua/wali ditemukan beberapa yang kosong, sehingga diperlakukan proses *feature engineering*.
2. Data penghasilan berbentuk kategorikal, sehingga diperlukan transformasi untuk mempresentasikan penghasilan.
3. Bebrapa siswa yang tercatat layak PIP (usulan dari sekolah) memiliki alasan PIP yang tidak tercatat, sehingga interpretasi pada alasan kelayakan dilakukan secara hati-hati.

5.3 Saran

Penelitian selanjutnya disarankan dapat mengeksplorasi algoritma *clustering* lain seperti K-Medoids, DBSCAN, BIRCH, atau Hiearchial *Cluster* terutama

dalam melihat pola pengelompokan siswa berdasarkan profil sosial ekonomi. Selain itu, disarankan untuk menambahkan atribut penelitian seperti kepemilikan aset, kondisi tempat tinggal, status DTKS dan atribut lain yang dapat memperkaya hasil *clustering* serta memberikan gambaran profil sosial ekonomi siswa yang lebih komprehensif.

Penelitian selanjutnya disarankan dapat mengembangkan hasil *clustering* kedalam bentuk dashboard sederhana. Hal ini bertujuan untuk mempermudah pihak sekolah dalam melihat profil kelompok siswa, distribusi status usulan PIP, serta karakteristik sosial ekonomi setiap *cluster*.

Bagi pihak sekolah disarankan agar data Dapodik dapat terus diperbarui kelengkapannya secara berkala, terutama pada penghasilan orang tua/wali. Kelengkapan data tersebut dapat memberikan gambaran yang lebih akurat terhadap kondisi sosial ekonomi siswa.

Selain itu, hasil penelitian ini sebaiknya tidak digunakan sebagai dasar tunggal dalam penentuan PIP karena keputusan kelayakan PIP tetap harus mengacu pada aturan resmi dan verifikasi pihak berwenang. Hasil penelitian ini digunakan sebagai informasi pendukung dalam memahami pola profil sosial ekonomi siswa terutama siswa yang memiliki ekonomi rentan.

DAFTAR PUSTAKA

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. In *Electronics (Switzerland)* (Vol. 9, Number 8, pp. 1–12). MDPI AG. https://doi.org/10.3390/electronics9081295

Badan Pusat Statistik Kota Probolinggo. (2024). *Indikator kesejahteraan rakyat Kota Probolinggo 2023/2024* (Vol. 7). Badan Pusat Statistik Kota Probolinggo.

Bhargavi Konda. (2022). The impact of data preprocessing on data mining outcomes. *World Journal of Advanced Research and Reviews*, *15*(3), 540–544. https://doi.org/10.30574/wjarr.2022.15.3.0931

Cote, L. R., Gordon, R., Randell, C. E., Schmitt, J., & Marvin, H. (n.d.). *IRL @ UMSL IRL @ UMSL Open Educational Resources Collection Open Educational Resources*. Retrieved https://irl.umsl.edu/oer/4/

Dsouza, S., Dsouza, J. D., & Vanitha, &. (n.d.). *ANALYSIS OF DATA USING K-MEANS AND K-MEDOIDS ALGORITHMS*.

Eldita, S., & Umanto, U. (2025). Efektivitas Program Indonesia Pintar dalam mengatasi kesenjangan biaya pendidikan. *JAMP: Jurnal Administrasi dan Manajemen Pendidikan, 8*(1), 29–51. http://journal2.um.ac.id/index.php/jamp/

Fadlil, A., Riadi, I., & Mulyana, Y. (2023). Integration of Fuzzy C-Means and SAW Methods on Education Fee Assistance Recipients. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*. https://doi.org/10.22219/kinetik.v8i2.1636

Fernandes, A. A. A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., & Sakellariou, R. (2023). Data Preparation: A Technological Perspective and Review. *SN Computer Science*, *4*(4). https://doi.org/10.1007/s42979-023-01828-8

Goss-Sampson, M. A. (2022). *Statistical analysis in JASP: A guide for students* (5th ed., JASP v0.16.1). University of Greenwich.

Gul, M., & Rehman, M. A. (2023). Big data: an optimized approach for cluster initialization. *Journal of Big Data*, *10*(1), 120. https://doi.org/10.1186/s40537-023-00798-1

Hendri, H., Andrianof, H., Robianto, R., Awal, H., Putra, O. A., Wijaya, R., Gusman, A. P., Hafizh, M., & Pondrinal, M. (2024). A hybrid data mining for predicting scholarship recipient students by combining K-means and C4.5 methods. *Indonesian Journal of Electrical Engineering and Computer Science*, *33*(3), 1726–1735. https://doi.org/10.11591/ijeecs.v33.i3.pp1726-1735

Ivan, M. (2024). Evaluasi Kebijakan Bantuan Pendidikan (Program Indonesia Pintar/Bantuan Operasional Sekolah) Dalam Mengatasi Anak Tidak Sekolah (Ats) Dan Peningkatan Angka Partisipasi Kasar/Angka Partisipasi Murni (Apk/Apm) Di Indonesia. *Jurnal Transformasi Administrasi,* 14(1), 80-92.

Kholillah, M. K., Furnamasari, Y. F., & Dewi, D. A. (2022). Peran pendidikan dalam menghadapi arus globalisasi. *Edumaspul: Jurnal Pendidikan,* 6(1), 515–518.

Liu, R. (2022). Data Analysis of Educational Evaluation Using K-Means Clustering Method. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/3762431

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Mardison, M., Defit, S., & Alturky, S. (2021). Prediction of Scholarship Recipients Using Hybrid Data Mining Method with Combination of K-Means and C4.5 Algorithms. *International Journal of Artificial Intelligence Research*, *5*(2). https://doi.org/10.29099/ijair.v5i2.224

Marisa, F., Ahmad, S. S. S., Yusof, Z. I. M., Fachrudin, & Aziz, T. M. A. (2019). Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model. *International Journal of Integrated Engineering*, *11*(3), 169–180. https://doi.org/10.30880/ijie.2019.11.03.018

Menteri Pendidikan dan Kebudayaan Republik Indonesia. (2020). *Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 10 Tahun 2020 tentang Program Indonesia Pintar*.

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering Analysis for Classifying Student Academic Performance in Higher Education. *Applied Sciences (Switzerland)*, *12*(19). https://doi.org/10.3390/app12199467

Mulyani, F., Ridwan, E., & Nazer, M. (2023). Efektivitas Program Indonesia Pintar terhadap Partisipasi Sekolah di Kawasan Barat dan Timur Indonesia. *Jurnal Informatika Ekonomi Bisnis*, 1328–1332. https://doi.org/10.37034/infeb.v5i4.652

Nuraeni, F., Kurniadi, D., & Fauzian Dermawan, G. (2023). Pemetaan Karakteristik Mahasiswa Penerima Kartu Indonesia Pintar Kuliah (KIP-K) menggunakan Algoritma K-Means++. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *11*(3), 437–443. https://doi.org/10.32736/sisfokom.v11i3.1439

Qi, J., Yu, Y., Wang, L., Liu, J., & Wang, Y. (2017). An effective and efficient hierarchical K-means clustering algorithm. *International Journal of Distributed Sensor Networks*, *13*(8), 1–17. https://doi.org/10.1177/1550147717728627

Rahmah, S. A., & Antares, J. (2021). I N F O R M A T I K A KLASTERISASI SELEKSI MAHASISWA CALON PENERIMA BEASISWA YAYASAN MENGGUNAKAN K-MEANS CLUSTERING. *Jurnal Informatika, Manajemen Dan Komputer*, *13*(2).

Septiawati, S. E., Prasetijowati, T., Nurany, F., Publik, P. A., & Surabaya, B. (2022). *Analisis Penerapan Program Indonesia Pintar (PIP) Prespektif Good Governance di Lingkungan Madrasah* (Vol. 11, Number 3). www.publikasi.unitri.ac.id

Setiawan, I. D., & Triayudi, A. (2024). Penerapan Algoritma Clustering K-Means Data Mining dalam Pengelompokan Mahasiswa Penerima Beasiswa. *Journal of Computer System and Informatics (JoSYC)*, *5*(2), 430–441. https://doi.org/10.47065/josyc.v5i2.4971

Stiawan, D. B., & Nugroho, Y. S. (2023). Perbandingan Performa Algoritma Decision Tree untuk Klasifikasi Penerima Beasiswa Bank Indonesia. *Indonesian Journal of Computer Science Attribution*, *12*(4), 2108-2121. https://doi.org/10.33022/ijcs.v12i4.3339

Wongoutong, C. (2024). The impact of neglecting feature scaling in k-means clustering. *PLoS ONE*, *19*(12). https://doi.org/10.1371/journal.pone.0310839

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016

LAMPIRAN-LAMPIRAN

Lampiran 1 Pedoman dan Hasil Wawancara

A. Identitas Wawancara

Narasumber	Jabatan	Hari/Tanggal	Tujuan
Tini Susiyowati, S.E.	Operator Dapodik	13 Maret 2026	Memperoleh informasi data Dapodik dan proses pendataan sosial ekonomi siswa terkait pengusulan PIP.
Lailatul Fitria, S.Pd.SD.	Guru dan Operator Dapodik	8 April 2026	Memperoleh informasi pendukung mengenai proses usulan PIP.

B. Pedoman dan Hasil Wawancara

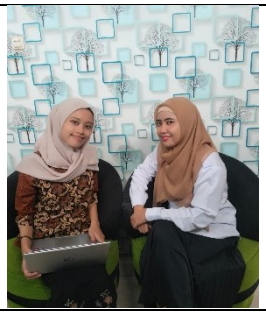
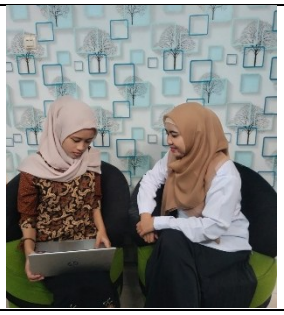
1. Wawancara dengan operator Dapodik

No	Pertanyaan	Ringkasan Jawaban
1.	Data apa saja dalam Dapodik yang berkaitan dengan usulan PIP?	Usulan PIP berdasarkan sosial ekonomi siswa, dilihat dari penghasilan orang tua, jumlah saudara yang menjadi tanggungan, lalu dilihat apakah memiliki bantuan sosial seperti KPS/PKH, dan siswa yang memiliki KIP berpeluangjuga dalam pengusulan PIP.
2.	Bagaimana proses pencatatan penghasilan orang tua/wali?	Dicatat berdasarkan orang tua/wali siswa saat pendataan. Hasilnya berupa rentang rata-rata penghasilan.
3.	Bagaimana jika di data Dapodik tidak tercatat penghasilan orang tuanya atau tidak lengkap?	Data yang kosong artinya informasi belum tercatat atau belum diperbarui, terkadang siswa ada yang lupa untuk menyuruh orang tua/wali dalam pengisian, jadi pihak sekolah akan memperbaruinya.
4.	Apakah terdapat kendala dalam kelengkapan data Dapodik?	Kendalanya ada, yaitu data dari orang tua dan wali belum lengkap, belum diperbarui, atau terdapat informasi yang belum rinci. Jadi, kelengkapan datanya masih perlu diperbaiki secara berkala.

No	Pertanyaan	Ringkasan Jawaban
5.	Apakah diperbarui secara berkala?	Iya, sesuai periode pembaruan dan kebutuhan sekolah, terutama jika terdapat perubahan data siswa atau keluarga.
Dokumentasi		

2. Wawancara dengan Guru

No.	Pertanyaan	Hasil Wawancara
1.	Bagaimana pihak sekolah dalam melihat kondisi apa siswa diusulkan PIP?	Dilihat dari kondisi sosial ekonominya, seperti penghasilan keluarganya berapa, kepemilikan kartu bantuan sosial seperti KPS/KKS/PKH, kepemilikan KIP. Dari sekolah hanya mengusulkan, karena yang menentukan kelayakan dari pusatnya.
2.	Apakah dalam mengusulkan PIP berdasarkan dengan peraturan menteri?	Iya, sebagai aturan dan gambaran dasarnya, pada praktiknya, sekolah menggunakan data Dapodik yang melihat kondisi sosial ekonomi siswa. Terkadang siswa mengusulkan kalau mempunyai kepemilikan kartu bantuan, jadi sekolah mencatatnya dan bisa diusulkan, namun keputusan akhir dari pusat, sekolah hanya mengusulkan.
3.	Bagaimana sekolah memastikan data siswa yang digunakan untuk usulan PIP sesuai dengan kondisi siswa?	Sekolah melakukan pengecekan melalui data administratif dan bantuan informasi dari wali kelas ataupun pihak sekolah yang mengetahui kondisi siswa. Jika diperlukan, data dapat dikonfirmaasi kembali kepada orang tua atau wali siswa.
4.	Mengapa pada data Dapodik terdapat status usulan layak, namun pada alasannya beberapa tidak tercatat?	Hal ini memang terjadi karena data belum lengkap, atau alasan belum tercatat secara detail pada sistem. Pada praktiknya tetap mengacu pada kepemilikan bantuan sosial, kepemilikan KIP, jumlah saudara dan penghasilan dari orang tua atau wali.

No.	Pertanyaan	Hasil Wawancara
Dokumentasi		 

Lampiran 2 Dataset Penelitian

Dataset Penelitian

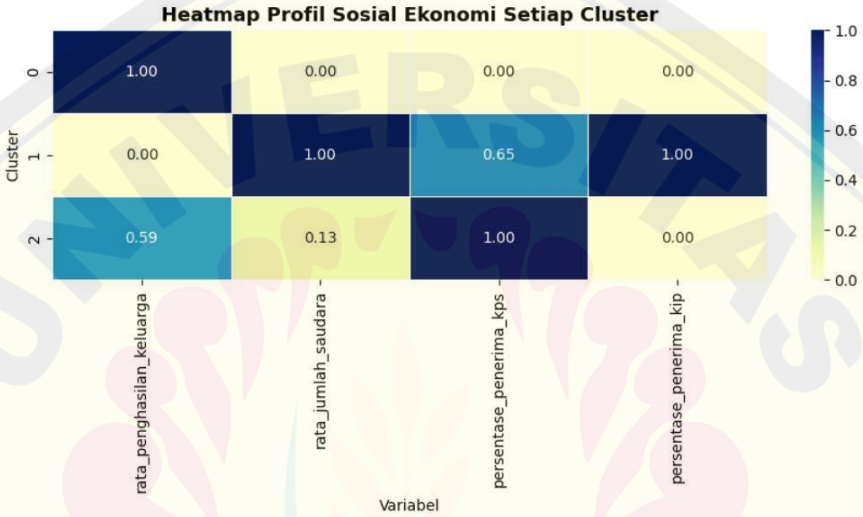
Dataset Setelah Preprocessing

Lampiran 3 Kode Program Penelitian

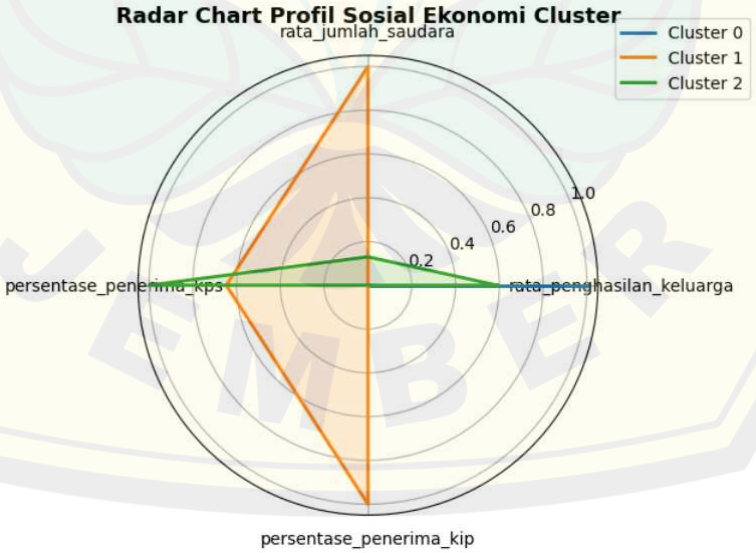
Kode Program

Lampiran 4 Hasil *Cluster*

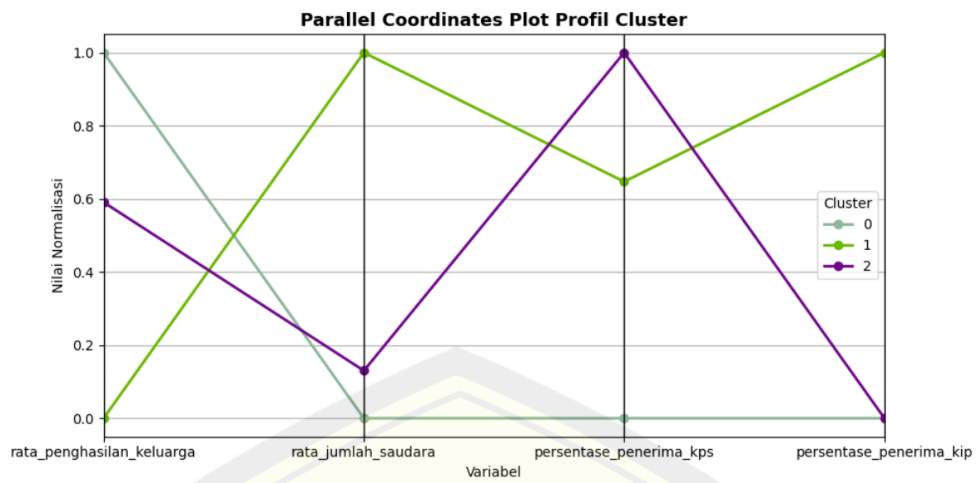
Tabel Hasil Clustering



Visualisasi Heatmap



Radar Chart



Parallel Coordinat Plot