Segmentasi Mahasiswa Berdasarkan Performa Akademik dan Karakter Pribadi Menggunakan K-Means Clustering

**Juliyana Herman**

Universitas Haluoleo/Teknik Informatika

e-mail: [julianaoppo87](mailto:julianaoppo87)[@gmail.](mailto:1xxxx@xxxx.xxx)com

**Abstrak**

Pendidikan tinggi memainkan peran penting dalam pengembangan sumber daya manusia, namun ketimpangan dalam pencapaian akademik dan akses masih terjadi di kalangan mahasiswa. Ketimpangan ini dipengaruhi oleh faktor akademik seperti nilai dan kehadiran, serta aspek sosial seperti tingkat stres, pendidikan orang tua, dan keterlibatan dalam kegiatan ekstrakurikuler. Penelitian ini bertujuan untuk mengidentifikasi ketimpangan pendidikan dengan mengelompokkan mahasiswa berdasarkan faktor akademik dan sosial menggunakan algoritma clustering K-Means. Data yang digunakan adalah dataset berjumlah 5.000 mahasiswa dari sumber terbuka. Tahapan pra-pemrosesan mencakup penanganan data hilang, pengkodean variabel kategorik, dan penskalaan data numerik. Jumlah cluster optimal ditentukan menggunakan metode Elbow dan Silhouette Score, kemudian dilakukan proses clustering dan visualisasi menggunakan PCA. Model menghasilkan nilai Silhouette Score sebesar 0,9137 yang menunjukkan performa pengelompokan yang sangat baik. Segmentasi ini menghasilkan enam kelompok mahasiswa dengan karakteristik akademik dan sosial yang berbeda, yang dapat membantu pemahaman lebih baik terhadap kesenjangan pendidikan. Hasil penelitian ini penting sebagai dasar penyusunan kebijakan pendidikan yang lebih inklusif dan tepat sasaran. Penelitian ini juga menunjukkan potensi pembelajaran tak terawasi (unsupervised learning) dalam mengungkap pola ketimpangan di pendidikan tinggi.

**Kata kunci**—Clustering, Segmentasi Mahasiswa, Ketimpangan Pendidikan, Faktor Akademik, Faktor Sosial, K-Means, Metode Elbow, Silhouette Score

***Abstract***

*Higher education plays a vital role in human resource development; however, disparities in academic achievement and access still occur among students. These disparities are influenced by academic factors such as grades and attendance, as well as social aspects such as stress levels, parental education, and involvement in extracurricular activities. This study aims to identify educational inequality by clustering students based on academic and social factors using the K-Means clustering algorithm. The data used consists of a dataset of 5,000 students from an open-source platform. The preprocessing stages include handling missing values, encoding categorical variables, and scaling numerical data. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, followed by clustering and visualization using PCA. The model achieved a Silhouette Score of 0.9137, indicating excellent clustering performance. This segmentation resulted in six student groups with distinct academic and social characteristics, which can enhance understanding of educational inequality. The findings of this study are important as a basis for formulating more inclusive and targeted educational policies. The study also demonstrates the potential of unsupervised learning in uncovering patterns of inequality in higher education.*

***Keywords—*** *Clustering, Student Segmentation, Educational Inequality, Academic Factors, Social Factors, K-Means, Elbow Method, Silhouette Score*

1. PENDAHULUAN

Pendidikan mempengaruhi secara penuh pertumbuhan ekonomi suatu bangsa. Dalam peningkatan kualitas manusia Indonesia, pemerintah tidak merupakan satu sistem yang lepas dengan pihak swasta dan masyarakat. Hubungan pemerintah, masyarakat dan swasta merupakan hubungan yang tidak terpisahkan dalam peranannya dalam meningkatkan pemerataan dan mutu pendidikan ndidikan tinggi merupakan salah satu pilar utama dalam pembangunan sumber daya manusia yang berkualitas [1]. Namun, ketimpangan dalam akses, partisipasi, dan pencapaian pendidikan masih menjadi isu yang signifikan di berbagai perguruan tinggi. Ketimpangan ini tidak hanya dipengaruhi oleh faktor akademik seperti nilai ujian dan kehadiran, tetapi juga oleh faktor sosial seperti latar belakang keluarga, tingkat stres, dan aktivitas ekstrakurikuler. Oleh karena itu, penting untuk memahami bagaimana berbagai faktor ini saling berinteraksi dalam membentuk profil mahasiswa yang beragam.

Segmentasi mahasiswa berdasarkan faktor akademik dan sosial dapat memberikan gambaran yang lebih menyeluruh mengenai pola ketimpangan pendidikan. Dengan melakukan segmentasi, pihak kampus dapat mengidentifikasi kelompok mahasiswa yang rentan atau membutuhkan intervensi khusus, serta merancang kebijakan yang lebih tepat sasaran. Salah satu pendekatan yang dapat digunakan dalam segmentasi ini adalah algoritma K-Means, yaitu metode unsupervised learning yang efektif dalam mengelompokkan data berdasarkan kemiripan karakteristik.

Beberapa penelitian sebelumnya telah mengkaji segmentasi mahasiswa untuk berbagai tujuan. Di Indonesia, Christian dan Jimmy (2021) menggunakan algoritma K-Means untuk memprediksi performa akademik mahasiswa berdasarkan IPK, program studi, dan lama studi. Hasilnya digunakan sebagai model pemantauan kinerja mahasiswa guna mendukung perencanaan akademik universitas [2]. Suhanda et al. (2020) menerapkan algoritma K-Means dengan pendekatan CRISP-DM untuk mengelompokkan mahasiswa berdasarkan IPK, program studi, dan jenis kelamin guna mendukung pengambilan keputusan akademik [3].

Penelitian ini bertujuan untuk mengidentifikasi ketimpangan pendidikan dengan cara mengelompokkan mahasiswa berdasarkan faktor-faktor akademik seperti nilai ujian, kehadiran, dan jam belajar mingguan, serta faktor pribadi seperti tingkat stres, aktivitas ekstrakurikuler, akses internet di rumah, dan tingkat pendidikan orang tua. Dengan menerapkan algoritma K-Means pada dataset mahasiswa, diharapkan dapat diperoleh segmen-segmen mahasiswa yang mencerminkan kondisi pendidikan yang tidak merata, sehingga menjadi dasar untuk perumusan kebijakan pendidikan yang lebih inklusif.

2. METODE PENELITIAN

## 2.1 Tahapan Penelitian

Gambar 1 Tahapan Penelitian

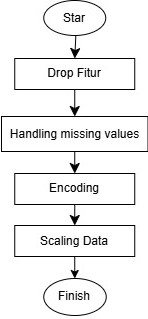
Berikut penjelasan masing-masing tahapan dari metode penelitian, yaitu:

## Data Collection

Dataset yang digunakan dalam penelitian ini adalah dataset mahasiswa yang berisi informasi akademik dan sosial yang diambil dari Kaggle dan berjumlah 5000 data.

b. *Exploratory* *Data Analysis* (EDA)

Tahap ini dilakukan untuk memahami karakteristik data, termasuk distribusi nilai, deteksi outlier, dan pencarian pola melalui visualisasi seperti histogram dan korelasi[4].

c. *Pra-Processing*

Gambar 2 Flow Chart *Pra-Processing*

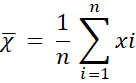
Berikut penjelasan ringkas setiap alur dalam *pra-processing*.

1) Drop Fitur

Langkah ini bertujuan untuk menghapus kolom atau fitur yang tidak relevan, redundan, atau tidak berguna dari dataset. Misalnya, kolom seperti ID unik, data duplikat, atau fitur yang tidak berkaitan dengan tujuan analisis dihilangkan untuk menyederhanakan data dan mengurangi noise.

2) Handling Missing Values

Data yang hilang (kosong atau tidak tercatat) perlu ditangani agar tidak mengganggu analisis. Metode penanganannya bisa berupa menghapus baris/data yang mengandung nilai kosong, mengisi nilai kosong dengan rata-rata, median, atau modus (untuk data numerik/kategorik), dan menggunakan teknik imputasi lanjutan seperti prediksi nilai berdasarkan fitur lain. Berikut rumus untuk handling missing value:



Keterangan:

* = nilai rata-rata dari fitur tersebut
* *n* = jumlah data yang tidak kosong (bukan missing)
* *x* = nilai ke-*i* dari fitur yang tersedia (tidak missing)

3) Encoding

Data kategorikal (seperti jenis kelamin, kategori produk, atau label teks) diubah menjadi bentuk numerik agar bisa diproses oleh algoritma machine learning. Contoh label encoding adalah mengubah kategori menjadi angka (misal: "Merah" = 0, "Biru" = 1). Contoh one-hot encoding adalah dengan membuat kolom baru untuk setiap kategori dengan nilai 0 atau 1. Berikut rumusnya:

Jika  adalah nilai kategori, maka:

Encode 

4) Scaling Data

Proses menyeragamkan skala data numerik agar tidak ada fitur yang mendominasi akibat perbedaan rentang nilai. Contoh metodenya adalah

1. Standardization, yaitu dengan menyesuaikan data sehingga rata-rata = 0 dan deviasi standar = 1. Rumus:



Keterangan:

* z = nilai setelah distandarkan (z-score)
* x = nilai asli
* = rata-rata (mean) dari semua nilai di fitur tersebut
* = standar deviasi dari fitur tersebut

1. Normalization, yaitu dengan mengubah rentang nilai menjadi 1 hingga -1. Rumus:



Keterangan:

* = nilai yang sudah dinormalisasi (berada di antara 0 dan 1)
* = nilai asli
* min = nilai minimum dari fitur tersebut
* max = nilai maximum dari fitur tersebut

d. Pemilihan Nilai K Optimum

Nilai K terbaik ditentukan menggunakan metode *Elbow Method* dan *Silhouette Score* untuk mengetahui jumlah cluster yang optimal.

e. Cluster (*K-Means*)

Model *clustering K-Means* digunakan untuk mengelompokkan mahasiswa berdasarkan faktor-faktor akademik dan sosial.

f. Model Evaluasi

Evaluasi model dilakukan dengan mengukur skor *silhouette* dan visualisasi hasil klaster menggunakan PCA (*Principal Component Analysis*).

g. Visualisasi dan Analisis

Pada tahap ini, hasil kelompok divisualisasikan dan dianalisis untuk membuat kesimpulan tentang hasil kelompok masing-masing[5].

2.2 *K Means Clustering*

*K-Means* adalah salah satu teknik clustering pada Data Mining proses pemodelan tanpa supervisi dan metode pengelompokkan data secara partisi. Data yang dikelompokan metode K-Means menjadi beberapa kelompok dan setiap kelompok memiliki karakteristik yang mirip atau sama dengan lainnya tetapi dengan kelompok lainnya memiliki karakteristik berbeda. Tujuannya untuk meminimalisasi perbedaan setiap data di dalam satu cluster serta memaksimalkan perbedaan dengan cluster yang lain[6].

2.3 *Elbow Method*

Metode Elbow adalah metode yang umumnya dipergunakan untuk penentuan jumlah cluster yang akan dibuat pada proses clustering. Metode elbow dipakai untuk memaksimalkan atau menentukan jumlah cluster yang optimal untuk proses clustering yang akan dilakukan. Cara kerja metode elbow adalah dengan menghitung kohesi dan pemisah klister[7]. Rumus dari elbow method adalah:

WCSS 

Keterangan:

* Untuk setiap cluster j
* Ambil semua titik xi, dalam cluster itu
* Hitung kuadrat jaraknya ke centroid j
* Jumlahkan semua jaraknya

## 2. 4 Silhouette Score

Metode Silhouette menggunakan koefisien siluet yang menggabungkan pemisahan dan kohesi [8]. Semakin besar koefisien Silhouette, semakin baik klaster tersebut. Rumus *silhouette score:*

## 

Keterangan:

* = rata-rata jarak titik i ke semua titik di cluster-nya sendiri
* = rata rata jarak i ke titik-titik cluster terdekat lainnya
* = skor silhouette untuk titik i
* hasil akhir adalah rata-rata semua 

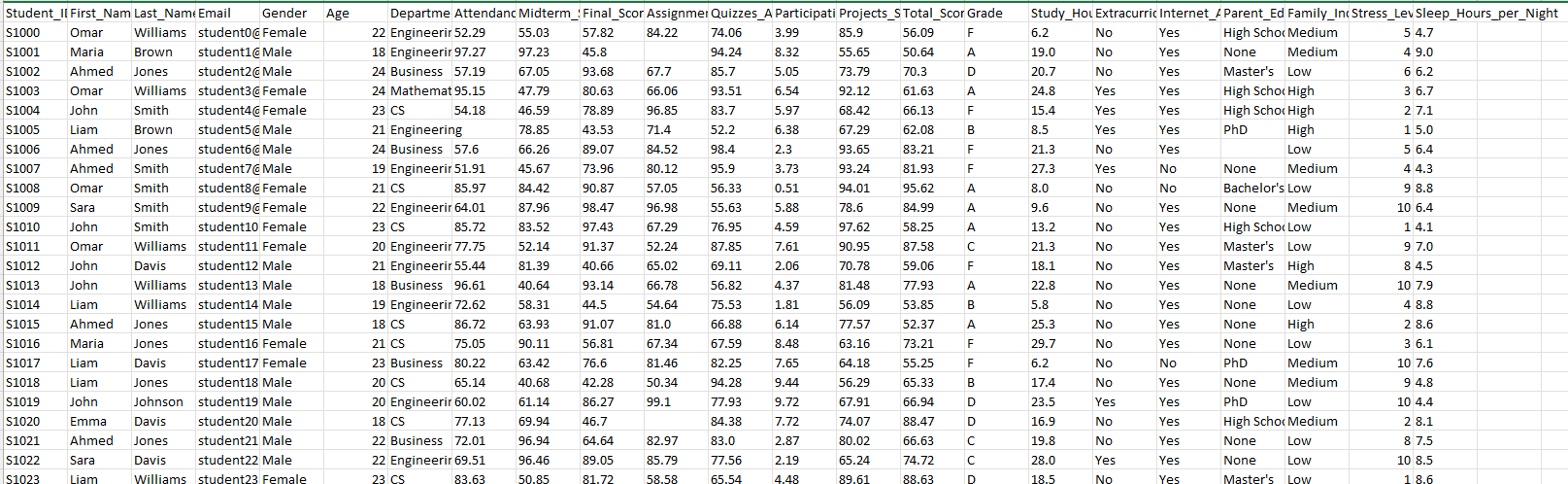
3. HASIL DAN PEMBAHASAN

Implementasi metode K-Means dalam segmentasi mahasiswa berdasarkan performa akademik dan karakter pribadi dilakukan mengikuti tahapan-tahapan yang telah digambarkan dalam diagram alur sebelumnya. Proses dimulai dari pemilihan data, eksplorasi awal, pra-pemprosesan data, pemilihan jumlah cluster yang optimal, hingga visualisasi hasil clustering.

Seluruh proses ini diimplementasikan menggunakan platform Google Collaboratory dan bahasa pemrograman Python. Beberapa library yang digunakan dalam proses analisis dan implementasi model antara lain Pandas, Numpy, Seaborn, Matplotlib, dan Scikit-Learn.

A. Pengumpulan Dataset

Kumpulan data ini adalah data nyata dari 5000 catatan yang dikumpulkan dari penyedia pembelajaran swasta.



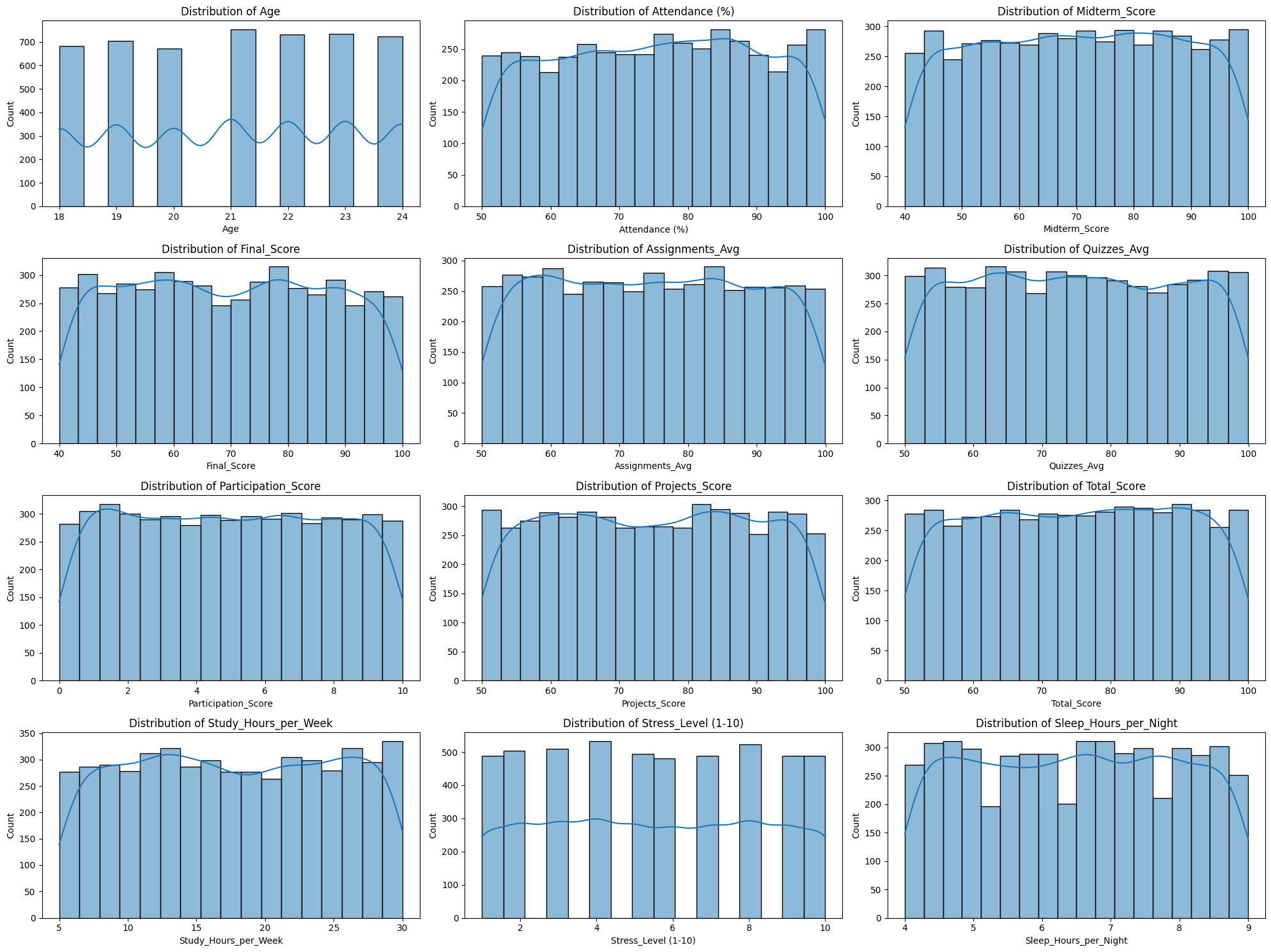
Gambar 3 Dataset Mahasiswa

B. *Exploratory Data Analysis* (EDA)

Berdasarkan dataset pada Gambar 2, dilakukan *Exploratory Data Analysis* (EDA) untuk memahami struktur data secara keseluruhan.

1. *Univariate Analysis* untuk Setiap Kolom Numerik

Adapun hasil visualisasi *Univariate Analysis* untuk setiap kolom numerik dalam dataset tersebut adalah:



Gambar 4 Hasil Visualisasi Distribusi Data Numerik

Gambar ini menampilkan sebaran (distribusi) berbagai variabel yang terkait dengan profil akademik dan kebiasaan siswa. Analisis distribusi ini membantu memahami karakteristik sampel secara menyeluruh, mulai dari aspek demografis, pencapaian akademik, hingga faktor penunjang belajar seperti manajemen waktu dan tingkat stres.

1. Distribusi Usia (Age)

Grafik usia memperlihatkan komposisi siswa berdasarkan kelompok umur. Jika distribusi condong ke usia muda (misalnya 18-22 tahun), dapat disimpulkan bahwa sebagian besar responden adalah mahasiswa tingkat awal. Sebaliknya, jika sebarannya merata, populasi mungkin terdiri dari mahasiswa dengan variasi tingkat semester atau latar belakang berbeda.

1. Kemampuan Finansial (Affordance)

Variabel ini ditampilkan dalam persentase, mengindikasikan proporsi siswa dengan tingkat kemampuan ekonomi tertentu. Misalnya, jika 70% termasuk dalam kategori "mampu", hal ini dapat mempengaruhi akses mereka terhadap sumber belajar tambahan atau fasilitas pendukung akademik.

1. Distribusi Nilai Akademik

Nilai Ujian (Midterm\_Score dan Final\_Score). Grafik ini menunjukkan sebaran pencapaian siswa. Distribusi normal (berbentuk lonceng) mengindikasikan variasi alami, sementara condong ke kiri/kanan menandakan dominasi nilai rendah/tinggi.

Rata-rata Tugas (Assignments\_Avg) dan Kuis (Quizzes\_Avg). Jika nilai tugas lebih tinggi daripada kuis, siswa mungkin lebih konsisten dalam pekerjaan panjang dibanding ujian spontan.

Total Nilai (Total\_Score). Merefleksikan performa keseluruhan. Pola yang seragam dengan sedikit outlier menandakan konsistensi, sedangkan variasi besar bisa disebabkan oleh perbedaan motivasi atau metode pembelajaran.

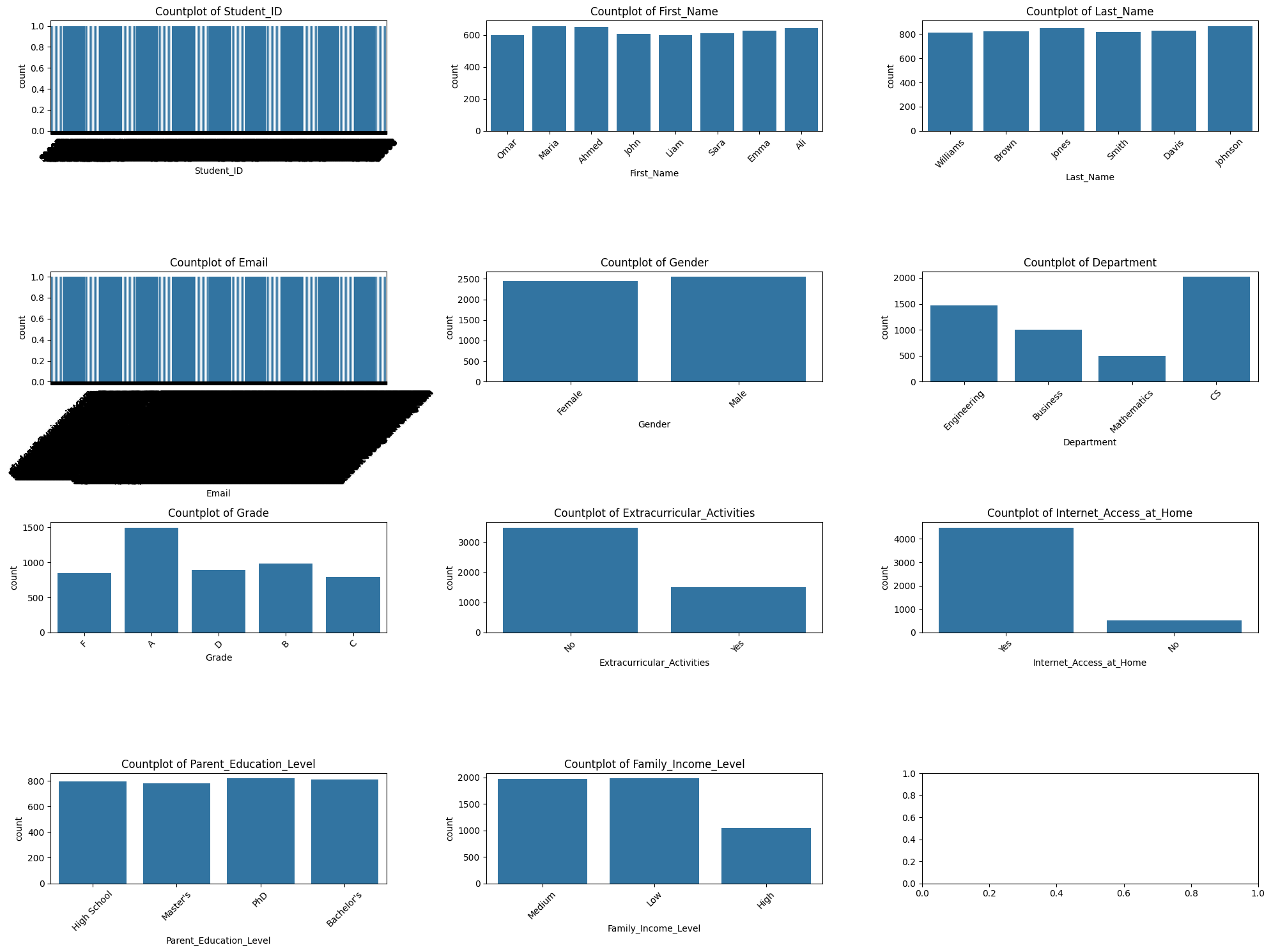
1. Kebiasaan Belajar dan Kesehatan Mental

Jam Belajar per Minggu (Study\_Hours\_per\_Week). Distribusi jam belajar mengungkap kedisiplinan siswa. Misalnya, puncak pada 10-15 jam/week mengindikasikan kebiasaan belajar yang moderat.

Tingkat Stres (Stress\_Level 1-10). Skor tinggi (7-10) pada sebagian besar responden perlu menjadi perhatian, karena stres berlebihan dapat menurunkan produktivitas.

Jam Tidur per Malam (Sleep\_Hours\_per\_Night). Kurang tidur (misalnya <6 jam) yang dominan dapat dikaitkan dengan beban akademik atau manajemen waktu yang buruk.

1. *Univariate Analysis* untuk Setiap Kolom Kategorikal



Gambar 5 Hasil Visualisasi Distribusi Data Kategorikal

Countplot tersebut merepresentasikan distribusi frekuensi dari berbagai atribut dalam dataset mahasiswa. Pada atribut Student\_ID dan Email, setiap batang menunjukkan jumlah yang konsisten, yang menandakan bahwa data bersifat unik dan tidak ada duplikasi. Ini penting untuk menjamin integritas identitas tiap mahasiswa.

Untuk First\_Name dan Last\_Name, terlihat bahwa nama-nama seperti "Omar", "Nina", "Ali", serta "Johnson", "Smith", dan "Davis" muncul dengan frekuensi yang tinggi. Ini menunjukkan adanya pola umum dalam penamaan, meskipun tetap ada keragaman.

Distribusi Gender cukup seimbang, dengan jumlah mahasiswa perempuan sedikit lebih banyak dibandingkan laki-laki. Komposisi ini mencerminkan keterwakilan gender yang relatif adil dalam data. Sedangkan pada atribut Department, mayoritas mahasiswa berasal dari jurusan Teknik (Engineering) dan Ilmu Komputer (CS), disusul oleh Matematika dan Bisnis. Hal ini mengindikasikan minat besar terhadap bidang teknologi dan sains.

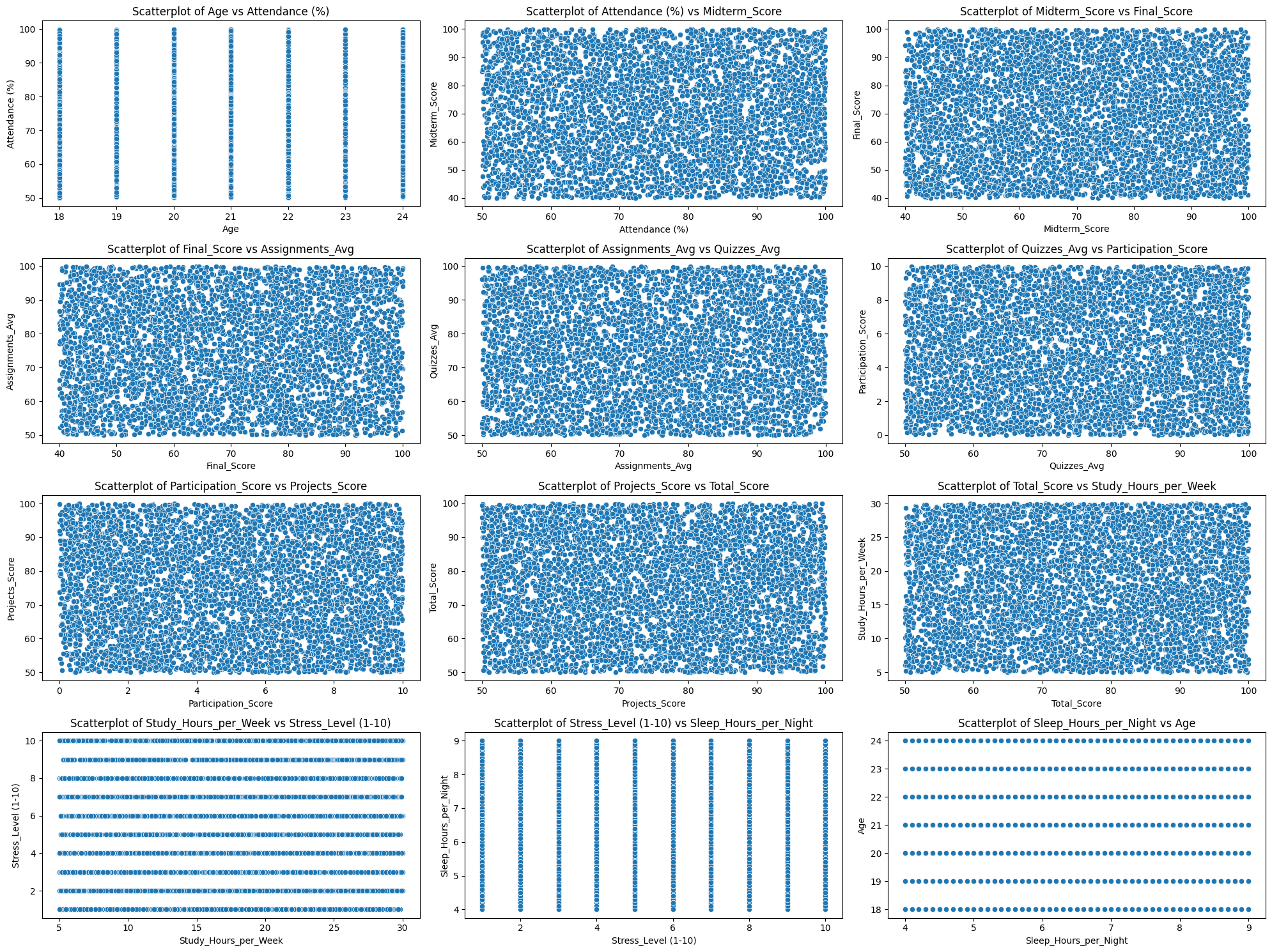
Pada kategori Grade, nilai A menjadi yang paling sering muncul, menunjukkan bahwa sebagian besar mahasiswa memiliki pencapaian akademik yang baik. Nilai F yang relatif rendah menunjukkan bahwa hanya sedikit mahasiswa yang mengalami kegagalan akademik.

Untuk Extracurricular\_Activities, sebagian besar mahasiswa tidak mengikuti kegiatan ekstrakurikuler. Ini menjadi informasi penting dalam melihat bagaimana keterlibatan non-akademik berperan dalam perkembangan mahasiswa. Sementara itu, sebagian besar mahasiswa memiliki *Internet\_Access\_at\_Home,* yang menjadi faktor pendukung penting dalam proses pembelajaran daring atau pencarian informasi tambahan.

Atribut *Parent\_Education\_Level* memperlihatkan distribusi yang merata antara berbagai jenjang pendidikan orang tua, dari SMA hingga PhD. Hal ini menunjukkan bahwa mahasiswa datang dari latar belakang pendidikan keluarga yang beragam. Sementara itu, *Family\_Income\_Level* memperlihatkan bahwa sebagian besar mahasiswa berasal dari keluarga dengan penghasilan menengah dan rendah, dengan jumlah yang lebih kecil dari kalangan berpenghasilan tinggi.

Satu subplot tidak menampilkan grafik apapun, kemungkinan karena data kosong atau kesalahan teknis saat visualisasi.

d) *Bivariate Analysis*



Gambar 6 Hasil *Bivariate Analysis*

Scatterplot tersebut menampilkan hubungan antar variabel kuantitatif yang berkaitan dengan perilaku belajar, performa akademik, dan kondisi psikologis mahasiswa. Pola-pola titik pada grafik memperlihatkan seberapa kuat hubungan antar dua variabel dalam setiap pasangan.

Scatterplot antara Age vs Attendance (%) dan Stress Level vs Sleep Hours per Night menunjukkan pola vertikal yang mencolok, yang menandakan bahwa variabel-variabel tersebut berupa data diskrit atau dikategorikan (misalnya umur hanya dari 18–24, stress level hanya 1–10, dan jam tidur dalam bilangan bulat). Tidak ada tren korelasi yang jelas, karena data tersebar merata di setiap kategori umur atau tingkat stres.

Relasi antara Attendance (%) dengan Midterm Score, serta Midterm Score dengan Final Score menunjukkan persebaran yang cukup menyebar namun mengindikasikan adanya korelasi positif lemah hingga sedang. Artinya, mahasiswa dengan tingkat kehadiran dan nilai ujian tengah semester yang tinggi cenderung memiliki nilai akhir yang lebih baik.

Hubungan antara Final Score dengan Assignments\_Avg, serta antara Assignments\_Avg dengan Quizzes\_Avg, memperlihatkan pola distribusi yang juga cukup tersebar. Meskipun tidak sepenuhnya linier, ada indikasi bahwa rata-rata nilai tugas dan kuis berkaitan dengan nilai akhir, meskipun tidak dalam bentuk hubungan yang kuat.

Scatterplot antara Quizzes\_Avg dengan Participation\_Score, dan Participation\_Score dengan Projects\_Score menunjukkan hubungan yang lemah. Partisipasi tampaknya tidak terlalu berpengaruh secara langsung terhadap skor proyek atau kuis.

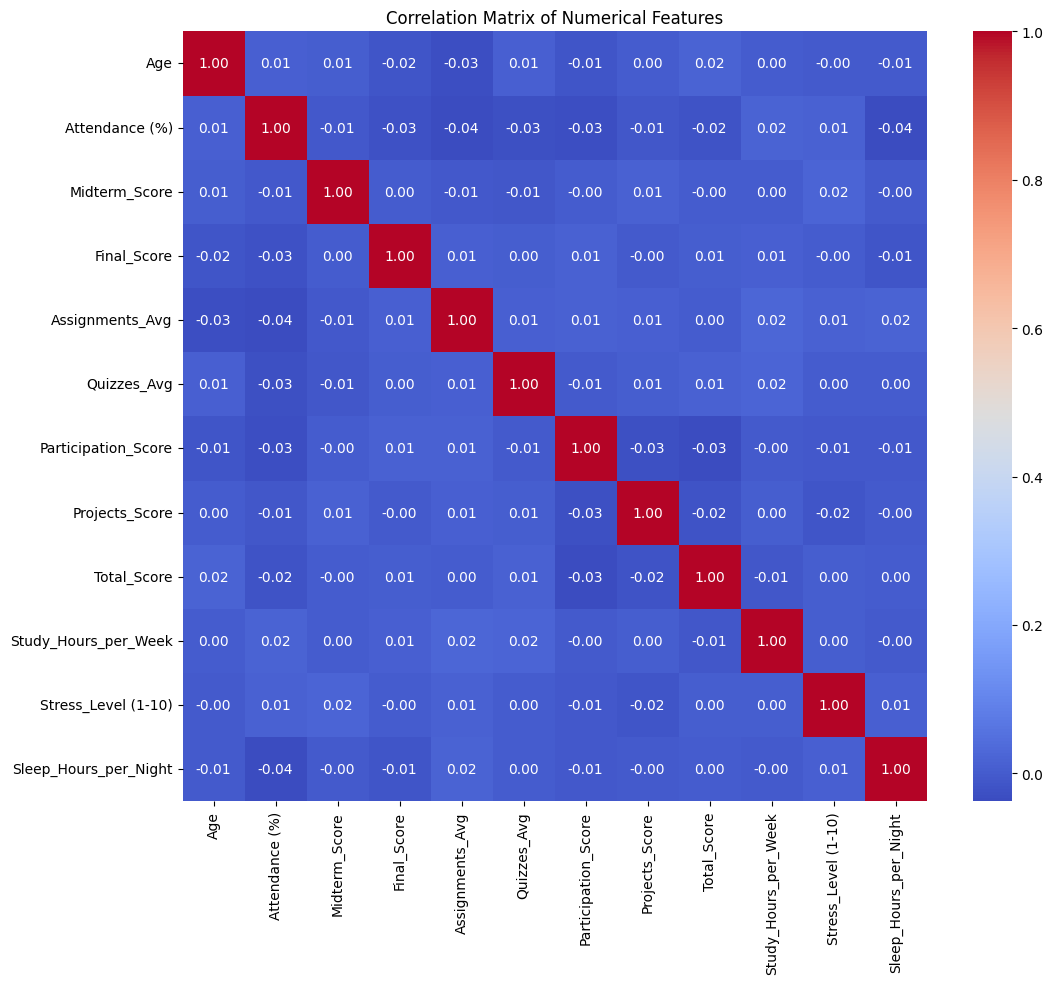
Sementara itu, grafik Projects\_Score vs Total\_Score menunjukkan korelasi yang relatif lebih jelas. Skor proyek tampaknya merupakan komponen signifikan dalam penentuan skor total mahasiswa, terlihat dari kecenderungan naiknya nilai total seiring peningkatan nilai proyek.

Hubungan antara Total\_Score dengan Study\_Hours\_per\_Week menunjukkan persebaran yang luas, menandakan bahwa durasi belajar per minggu tidak selalu sejalan dengan nilai total. Hal ini dapat disebabkan oleh faktor lain seperti efisiensi belajar atau kualitas waktu belajar yang tidak dapat ditangkap oleh jumlah jam semata.

Pada grafik Study\_Hours\_per\_Week dengan Stress Level, terdapat penyebaran horizontal, menunjukkan bahwa mahasiswa pada berbagai tingkat stres menghabiskan waktu belajar yang beragam. Begitu juga dengan grafik Stress Level vs Sleep Hours per Night, tidak menunjukkan hubungan negatif yang konsisten antara stres dan tidur, meskipun secara teori, semakin tinggi stres seharusnya mengurangi waktu tidur.

Terakhir, scatterplot antara Sleep\_Hours\_per\_Night dan Age tidak menunjukkan pola khusus, yang berarti waktu tidur tidak secara langsung dipengaruhi oleh usia dalam rentang umur mahasiswa yang relatif sempit ini.

e) *Multivariate Analysis*



Gambar 7 Matriks Korelasi Variabel Numerik

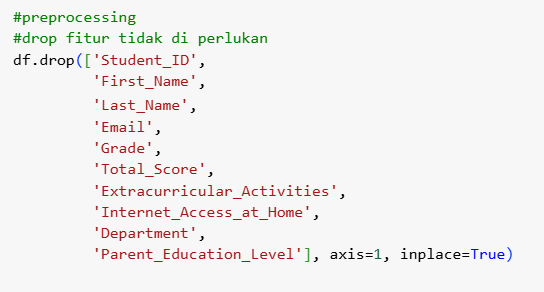
Matriks korelasi di atas menganalisis hubungan linear antara dua belas variabel numerik yang terkait dengan kinerja akademik dan kebiasaan siswa. Nilai korelasi yang ditampilkan berkisar antara -0,04 hingga 0,02, menunjukkan bahwa hampir semua variabel memiliki hubungan yang sangat lemah atau tidak berkorelasi secara signifikan. Sebagai contoh, korelasi antara Usia (Age) dan Kehadiran (Attendance) hanya sebesar 0,01, yang mengindikasikan bahwa usia tidak memiliki pengaruh linear terhadap tingkat kehadiran siswa. Demikian pula, Nilai Ujian Tengah Semester (Midterm\_Score) dan Nilai Ujian Akhir (Final\_Score) memiliki korelasi 0,00, menegaskan bahwa performa di kedua ujian tersebut tidak saling terkait secara linear dalam dataset ini. Beberapa korelasi negatif lemah juga terlihat, seperti antara Partisipasi Kelas (Participation\_Score) dan Nilai Proyek (Projects\_Score) (-0,03) atau antara Kehadiran (Attendance) dan Jam Tidur per Malam (Sleep\_Hours\_per\_Night) (-0,04). Namun, nilai-nilai ini terlalu kecil untuk dianggap bermakna secara statistik atau praktis.

Total Skor (Total\_Score) sebagai variabel kunci juga tidak menunjukkan korelasi yang signifikan dengan variabel lain, seperti Rata-Rata Tugas (Assignments\_Avg) (0,00) atau Jam Belajar per Minggu (Study\_Hours\_per\_Week) (-0,01). Hal ini mengisyaratkan bahwa pencapaian akademik siswa dalam dataset ini mungkin merupakan hasil akumulasi dari berbagai faktor yang seimbang, atau ada faktor eksternal (seperti motivasi intrinsik, kualitas pengajaran, atau lingkungan belajar) yang tidak tercakup dalam analisis. Rendahnya korelasi antar-variabel juga dapat disebabkan oleh ukuran sampel yang terbatas, variasi data yang kecil, atau metode pengukuran yang kurang presisi (misalnya, data tingkat stres yang dilaporkan sendiri). Selain itu, analisis ini hanya mengukur hubungan linear, sehingga pola non-linear (seperti hubungan berbentuk U antara stres dan kinerja) mungkin tidak terdeteksi. Untuk penelitian lanjutan, disarankan untuk menggabungkan teknik analisis multivariat, memperluas cakupan variabel (misalnya, latar belakang sosio ekonomi atau gaya belajar), atau menggunakan metode non-parametrik untuk mengeksplorasi hubungan yang lebih kompleks. Temuan ini menegaskan pentingnya pendekatan holistik dan multidisiplin dalam memahami dinamika kinerja akademik siswa.

C. *Pra-Processing*

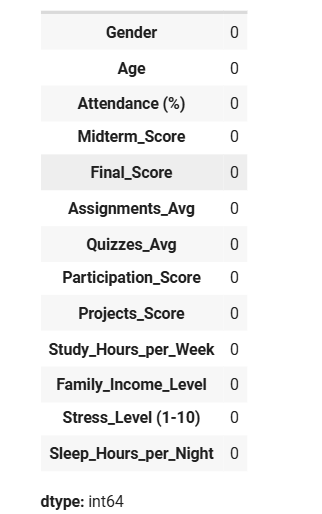
Pada tahap *pra-processing* dilakukan dengan menghapus baris yang mengandung nilai kosong (*missing values*) guna menjaga kualitas data. Selanjutnya, dipilih hanya variabel-variabel yang bertipe numerik, mengingat analisis korelasi hanya dapat diterapkan pada data jenis ini. Tahapan ini dilakukan untuk memastikan bahwa data yang digunakan bersih, relevan, dan siap untuk dianalisis secara statistik.

1. Menghapus Fitur yang Tidak Diperlukan

Gambar 8 Kode Menghapus Fitur yang Tidak Diperlukan

Pada tahap pra-pemrosesan, beberapa kolom dalam data dihapus karena dianggap tidak memiliki pengaruh langsung terhadap tujuan analisis. Kolom-kolom seperti identitas siswa, informasi kontak, nilai total, serta data tambahan seperti aktivitas ekstrakurikuler, akses internet di rumah, dan pendidikan orang tua tidak digunakan dalam analisis. Penghapusan ini dilakukan untuk menyederhanakan data dan memfokuskan analisis hanya pada informasi yang benar-benar dibutuhkan.

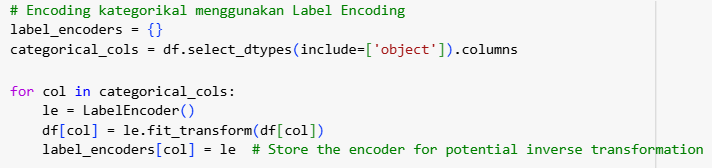
1. *Handling Missing Values*



Gambar 9 Hasil dari Menangani Nilai yang Kosong

Setelah data disederhanakan, langkah selanjutnya adalah menangani nilai yang kosong . Untuk kolom bertipe kategorik, nilai yang kosong diisi dengan nilai yang paling sering muncul (modus). Sementara itu, untuk kolom numerik, nilai kosong diisi dengan rata-rata dari kolom tersebut. Langkah ini dilakukan agar data menjadi lengkap dan siap untuk dianalisis tanpa mengurangi jumlah data yang tersedia.

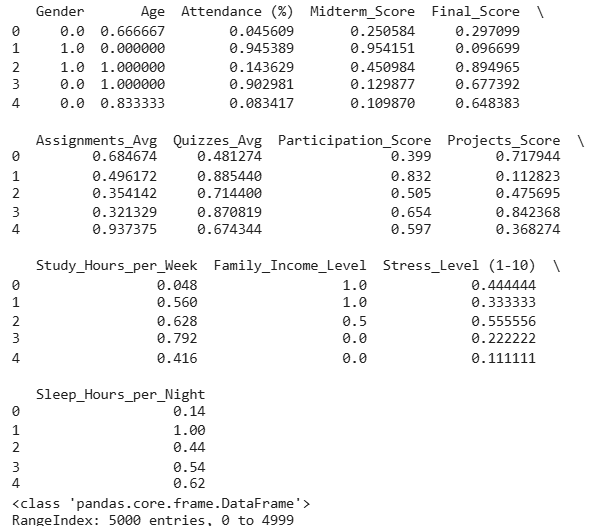
1. Encoding Variabel Kategorikal



Gambar 10 Hasil *Encoding* Variabel

Sebelum proses clustering dilakukan, variabel kategorikal pada dataset dataset diubah menjadi format numerik menggunakan metode Lebel Encoding. Setiap kolom bertipe kategorik dikodekan secara otomatis menggunakan *LabelEncoder* dari *scikit-learn*. Encoder disimpan untuk keperluan transformasi balik jika diperlukan. Proses ini memastikan data siap digunakan dalam algoritma pembelajaran mesin yang hanya dapat memproses data numerik.

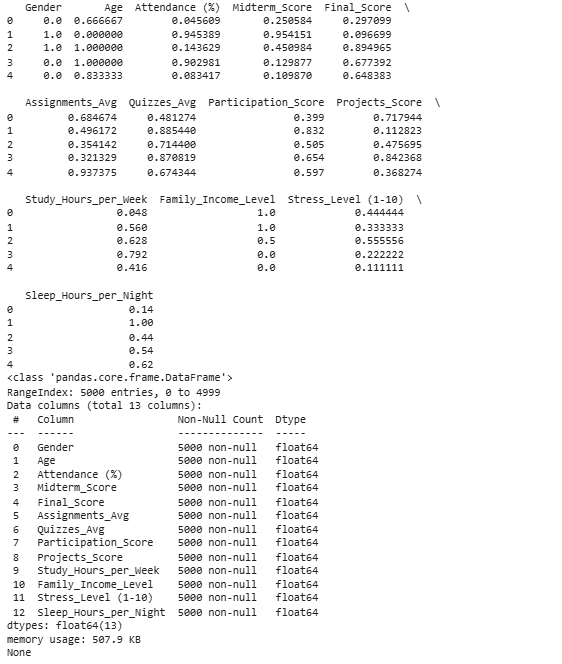
1. *Scaling* Data



Gambar 11 Hasil *Scaling* Data

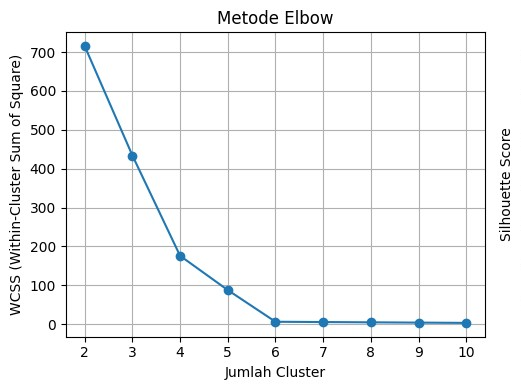
Setelah data lengkap, dilakukan proses penskalaan (*scaling*) pada kolom-kolom numerik menggunakan MinMaxScaler. Teknik ini digunakan untuk mengubah nilai-nilai numerik ke dalam rentang 0 hingga 1, sehingga semua variabel memiliki skala yang sama. Hal ini penting agar tidak ada satu variabel pun yang mendominasi analisis hanya karena perbedaan skala.

Berikut tampilan data frame yang telah diproses.



Gambar 11 Tampilan Data Frame yang Telah Diproses

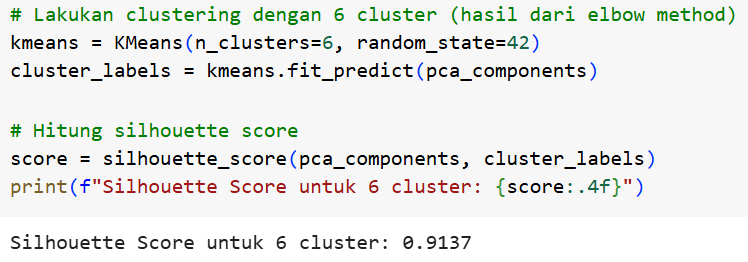
D. Pemilihan Nilai K Optimum



Gambar 12 Metode Elbow

Metode Elbow menggunakan nilai Within-Cluster Sum of Square (WCSS) yang mengukur total jarak kuadrat antara titik data dalam klaster dengan pusat cluster (centroid) untuk mengevaluasi kepadatan klaster. Semakin rendah nilai WCSS, semakin rapat pengelompokan data. Pada gambar, nilai WCSS turun dari 700 (pada 2 klaster) menjadi 0 (pada 10 klaster), menunjukkan bahwa penambahan jumlah klaster mengurangi dispersi data. Namun, tujuan utama metode ini adalah menemukan "siku" (elbow point), yaitu titik di mana penurunan WCSS mulai melandai secara signifikan. Misalnya, jika penurunan WCSS tajam hingga klaster ke- 6 dan kemudian melambat, klaster ke-6 dianggap optimal karena penambahan klaster setelahnya tidak memberikan peningkatan kepadatan yang berarti. Metode ini membantu menghindari overfitting dengan memilih jumlah cluster yang seimbang antara efisiensi dan akurasi.

E. Cluster (K-Means)



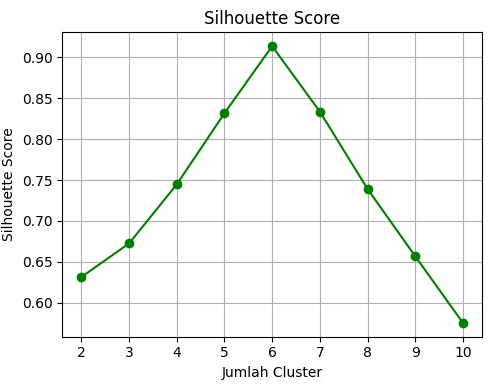
Gambar 13 Cluster (K-Means)

Untuk melakukan segmentasi data, digunakan algoritma K-Means dengan jumlah cluster sebanyak 6, yang diperoleh berdasarkan analisis metode *Elbow*. Proses clustering dilakukan pada data hasil reduksi dimensi menggunakan PCA (*Principal Component Analysis*), yang bertujuan untuk menyederhanakan kompleksitas data dan mempercepat proses komputasi.

Model K-Means berhasil membagi data ke dalam 6 kelompok atau cluster. Untuk mengevaluasi kualitas dari hasil clustering, digunakan nilai *Silhouette Score*. Nilai ini menunjukkan seberapa baik suatu data cocok dengan klasternya sendiri dibandingkan dengan klaster lainnya. Nilai Silhouette Score berada dalam rentang -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan hasil clustering yang lebih baik.

Berdasarkan hasil perhitungan, diperoleh nilai Silhouette Score sebesar 0.9137, yang menunjukkan bahwa pembagian cluster yang dilakukan memiliki kualitas yang sangat baik. Nilai ini mendekati angka 1, sehingga dapat disimpulkan bahwa pemisahan antar cluster cukup jelas dan kohesi dalam masing-masing cluster juga kuat.

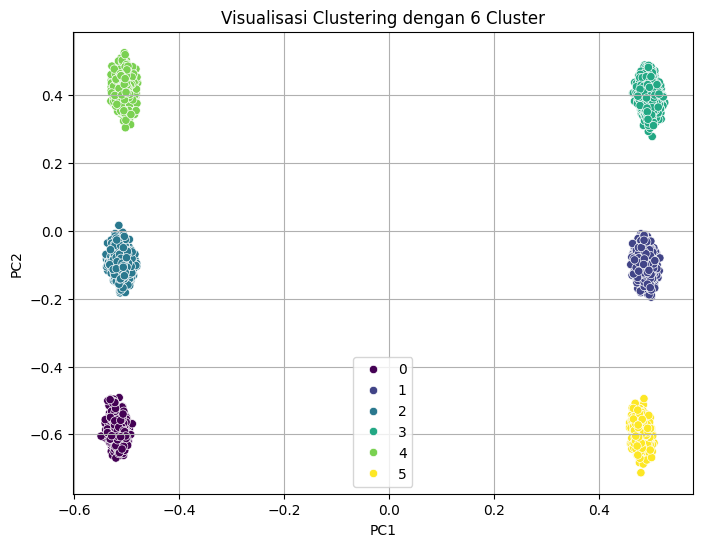
E. Model Evaluasi



Gambar 14 Silhouette Score

Silhouette Score mengukur seberapa baik suatu objek cocok dengan klaster sendiri (cohesion) dibandingkan klaster terdekat lainnya (separation). Skor berkisar antara -1 hingga 1, di mana nilai tinggi (mendekati 1) menunjukkan bahwa objek sangat sesuai dengan klaster sendiri dan terpisah jelas dari klaster lain. Pada gambar, skor tertinggi adalah 0.90 (pada 2 klaster), kemudian menurun seiring penambahan jumlah klaster hingga 0.60 (pada 10 klaster). Dalam praktik, jumlah klaster dengan skor tertinggi (misalnya, 2 atau 3 klaster) sering dianggap optimal. Namun, perlu diingat bahwa skor tinggi pada jumlah klaster kecil mungkin mencerminkan pengelompokan yang terlalu umum, sehingga perlu dikombinasikan dengan metode lain seperti Elbow untuk validasi.

F. Visualisasi dan Analisis



Gambar 15 Visualisasi Clustering

dengan 6 cluster

Visualisasi ini menampilkan hasil pengelompokan mahasiswa ke dalam enam klaster yang dibentuk menggunakan algoritma clustering, dengan bantuan teknik reduksi dimensi bernama Principal Component Analysis (PCA). Setiap titik dalam grafik mewakili satu mahasiswa, dan warna yang berbeda menunjukkan klaster atau kelompok yang berbeda. Secara visual, terlihat bahwa keenam klaster terpisah dengan cukup jelas — tiga klaster muncul di sisi kiri grafik, sementara tiga lainnya berada di sisi kanan.

Pola penyebaran ini menunjukkan bahwa proses clustering berhasil membedakan mahasiswa berdasarkan karakteristik tertentu. Sumbu horizontal (PC1) kemungkinan besar merepresentasikan dimensi utama yang membedakan mahasiswa, seperti antara faktor akademik dan non-akademik, sementara sumbu vertikal (PC2) menangkap variasi tambahan antar sub kelompok.

* Klaster 0 (ungu tua, kiri bawah):

Ini adalah kelompok *rentan secara akademik dan sosial*. Mahasiswa di sini biasanya memiliki nilai yang rendah, kehadiran yang kurang, dan mengalami tantangan sosial seperti minimnya dukungan keluarga atau tingkat stres yang tinggi.

* Klaster 1 (biru muda, kiri tengah)

Ini adalah *mahasiswa yang ideal dan seimbang*. Mereka mempunyai performa akademik yang baik, stres yang terkendali, dan kehidupan yang cukup seimbang antara tugas kuliah dan aktivitas sosial.

* Klaster 2 (hijau terang, kiri atas)

Kelompok ini bisa disebut *pejuang akademik bertekanan tinggi*. Mereka sangat termotivasi dalam hal belajar, namun dibarengi dengan stres yang tinggi. Semangat mereka besar, tapi rentan kelelahan.

* Klaster 3 (hijau kebiruan, kanan tengah)

Ini adalah *mahasiswa stabil dan rata-rata*. Mereka tidak terlalu menonjol dalam satu aspek, tapi juga tidak berada di bawah. Bisa dibilang, kelompok yang konsisten dan cukup aman.

* Klaster 4 (hijau kekuningan, kanan atas)

Klaster ini berisi *mahasiswa berprestasi yang mendapat dukungan sosial kuat*. Mereka unggul secara akademik dan juga punya lingkungan yang mendukung, baik dari keluarga maupun pertemanan. Tingkat stres mereka cenderung rendah meskipun performanya tinggi.

* Klaster 5 (kuning cerah, kanan bawah) Kelompok ini punya *tekanan tinggi tapi juga potensi besar*. Mereka memiliki kemampuan yang menonjol, namun berada dalam tekanan yang cukup berat. Jika diberikan dukungan yang tepat, mereka bisa berkembang luar biasa.

Hasil clustering secara detail ditunjukkan pada gambar



Gambar 16 Sampel Hasil Clustering

4. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan, segmentasi mahasiswa menggunakan algoritma K-Means berhasil mengelompokkan mahasiswa ke dalam enam klaster berdasarkan kombinasi faktor akademik dan sosial. Hasil pengelompokan ini menunjukkan bahwa terdapat perbedaan karakteristik yang jelas antar kelompok mahasiswa, mulai dari kelompok yang berprestasi dan memiliki dukungan sosial yang kuat, hingga kelompok yang rentan secara akademik maupun sosial. Kualitas hasil clustering tergolong sangat baik, ditunjukkan dengan nilai Silhouette Score sebesar 0.9137 yang mendekati 1, menandakan pemisahan antar klaster yang optimal dan kohesi yang kuat dalam tiap klaster.

Penelitian ini memiliki beberapa kelebihan, antara lain penggunaan tahapan analisis yang komprehensif mulai dari EDA, penentuan nilai K optimum, hingga visualisasi hasil menggunakan PCA. Selain itu, data diproses secara sistematis melalui tahapan pra-pemrosesan yang baik sehingga meningkatkan kualitas analisis. Namun, penelitian ini juga memiliki beberapa kekurangan, seperti penggunaan data sekunder yang mungkin kurang relevan dengan konteks lokal, serta variabel yang masih terbatas dan belum mencakup aspek psikologis dan sosial yang lebih dalam. Korelasi antar variabel yang lemah juga menjadi salah satu keterbatasan dalam memberikan interpretasi yang lebih dalam. Untuk pengembangan selanjutnya, penelitian ini dapat ditingkatkan dengan menggunakan data primer dari mahasiswa di institusi lokal, memperluas cakupan variabel yang digunakan, serta membandingkan algoritma K-Means dengan metode clustering lainnya untuk memperoleh hasil yang lebih akurat dan aplikatif dalam menyusun kebijakan pendidikan yang inklusif dan berbasis data.

5. SARAN

Untuk penelitian selanjutnya, disarankan agar menggunakan data primer yang diperoleh langsung dari institusi pendidikan lokal agar hasil segmentasi lebih relevan dan kontekstual dengan kondisi nyata mahasiswa. Selain itu, perlu menambahkan variabel-variabel lain yang lebih mendalam, seperti aspek psikologis (motivasi belajar, kecemasan, dukungan emosional) dan lingkungan sosial (hubungan dengan teman sebaya, kondisi tempat tinggal) agar segmentasi lebih holistik.

Penelitian selanjutnya juga dapat mengadopsi algoritma clustering alternatif seperti DBSCAN atau Hierarchical Clustering untuk dibandingkan dengan hasil K-Means, sehingga dapat diketahui metode mana yang memberikan segmentasi paling optimal. Di samping itu, penggunaan metode evaluasi tambahan seperti Davies-Bouldin Index dapat memperkaya penilaian terhadap kualitas hasil clustering. Terakhir, penelitian lanjutan dapat mengembangkan sistem rekomendasi berbasis hasil clustering untuk mendukung pengambilan keputusan dalam merancang intervensi atau program pembinaan mahasiswa secara lebih personal dan tepat sasaran.

# UCAPAN TERIMA KASIH

DAFTAR PUSTAKA

[1] J. S. Manajemen and S. Rabiah, “MANAJEMEN PENDIDIKAN TINGGI DALAM MENINGKATKAN MUTU PENDIDIKAN Management of Higher Education in Improving the Quality of Education,” *J. Sinar Manaj.*, vol. 6, no. 1, pp. 58–67, 2019, [Online]. Available: http://jurnal.unismuhpalu.ac.id/index.php/JSM

[2] Y. C. Jimmy, “Perancangan Model Predıksı Performa Akademık Mahasıswa Menggunakan Algorıtma K - Means Clusterıng ( Studı Kasus : Unıversıtas Xyz ),” vol. 1, no. 1, pp. 643–649, 2021.

[3] Y. Suhanda, I. Kurniati, and S. Norma, “Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik,” *J. Teknol. Inform. dan Komput.*, vol. 6, no. 2, pp. 12–20, 2020, doi: 10.37012/jtik.v6i2.299.

[4] M. Radhi, A. Amalia, D. R. H. Sitompul, S. H. Sinurat, and E. Indra, “Analisis Big Data Dengan Metode Exploratory Data Analysis (Eda) Dan Metode Visualisasi Menggunakan Jupyter Notebook,” *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 4, no. 2, pp. 23–27, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2475.

[5] R. Ishak and A. Bengnga, “Clustering Prestasi Akademik Lulusan Menggunakan Metode K-Means,” *Jambura J. Electr. Electron. Eng.*, vol. 6, no. 1, pp. 76–81, 2024, doi: 10.37905/jjeee.v6i1.23967.

[6] P. Sari, B. Pramono, and L. ode H. S. Sagala, “Improve K-Means Terhadap Status Nilai Gizi Pada Balita,” *semanTIK*, vol. 3, no. 1, pp. 143–148, 2017, doi: 10.1063/1.2957900.

[7] M. Guntara and N. Lutfi, “Optimasi Cacah Klaster pada Klasterisasi dengan Algoritma KMeans Menggunakan Silhouette Coeficient dan Elbow Method,” *JuTI “Jurnal Teknol. Informasi,”* vol. 2, no. 1, p. 43, 2023, doi: 10.26798/juti.v2i1.944.

[8] P. Vania and B. Nurina Sari, “Perbandingan Metode Elbow dan Silhouette untuk Penentuan Jumlah Klaster yang Optimal pada Clustering Produksi Padi menggunakan Algoritma K-Means,” *J. Ilm. Wahana Pendidik.*, vol. 9, no. 21, pp. 547–558, 2023, [Online]. Available: https://doi.org/10.5281/zenodo.10081332