TUGAS BESAR DATA MINING

PENGEMBANGAN MODEL PREDIKSI KELULUSAN MAHASISWA

Diajukan untuk memenuhi salah satu tugas Mata Kuliah Data Mining yang Dibina oleh: Fauzan Ramadhan S.Kom., M.Kom.

Disusun oleh Kelompok 9:

Difa Ramadhan	220102023
Nabila Tsari Aulia Mahmudah	220102064
Siti Arfi Mutoharoh	220102082

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS MUHAMMADIYAH BANDUNG

2025

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Allah Subhanahu wa Ta'ala atas segala limpahan rahmat dan karunia-Nya, sehingga kami dapat menyelesaikan laporan proyek pembuatan aplikasi prediksi kelulusan mahasiswa ini dengan baik. Laporan ini disusun sebagai bagian dari tugas besar mata kuliah Data Mining.

Laporan ini membahas proses pengembangan aplikasi prediksi kelulusan mahasiswa berbasis data mining, dimulai dari pemahaman bisnis, eksplorasi dan persiapan data, pemodelan, hingga evaluasi hasil serta implementasi dalam bentuk dashboard interaktif menggunakan Streamlit. Diharapkan laporan ini dapat menambah wawasan pembaca dalam menerapkan konsep data mining secara praktis, khususnya dalam membangun solusi berbasis klasifikasi dan visualisasi data.

Pada kesempatan ini, kami mengucapkan terima kasih yang sebesar-besarnya kepada semua pihak yang telah mendukung kelancaran penyusunan laporan ini. Terima kasih khusus kami sampaikan kepada Fauzan Ramadhan S.Kom., M.Kom. selaku dosen pengampu mata kuliah Data Mining atas ilmu, arahan, dan bimbingan yang diberikan. Ucapan teri ma kasih juga kami sampaikan kepada seluruh anggota kelompok serta teman-teman seperjuangan atas semangat, kerja sama, dan dukungan selama proses pengerjaan proyek ini.

Demikian laporan ini kami susun. Kami menyadari bahwa masih terdapat kekurangan baik dalam penulisan maupun isi materi yang disampaikan. Oleh karena itu, kami terbuka terhadap kritik dan saran yang membangun guna perbaikan pada karya kami di masa mendatang.

Bandung, 28 Januari 2025

Kelompok 9

DAFTAR ISI

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pendidikan tinggi merupakan pilar utama dalam pembangunan sumber daya manusia yang berkualitas. Namun, tidak semua mahasiswa mampu menyelesaikan studi mereka tepat waktu. Tingkat kelulusan yang rendah masih menjadi tantangan besar bagi banyak institusi pendidikan tinggi di seluruh dunia (Tinto, 2012). Fenomena ini dapat berdampak negatif tidak hanya bagi mahasiswa secara individu, tetapi juga terhadap reputasi institusi dan efektivitas sistem pendidikan secara keseluruhan.

Dengan berkembangnya teknologi informasi, institusi pendidikan kini memiliki akses terhadap sejumlah besar data akademik mahasiswa, mulai dari nilai ujian, tugas, hingga partisipasi dalam kelas. Data tersebut menyimpan informasi penting yang dapat diolah untuk mengidentifikasi pola-pola yang berkaitan dengan performa akademik dan kelulusan mahasiswa (Siemens & Long, 2011).

Pendekatan Educational Data Mining dan Learning Analytics memungkinkan eksplorasi data pendidikan secara sistematis untuk mendukung pengambilan keputusan akademik yang lebih berbasis data (Baker & Inventado, 2014). Dengan memanfaatkan algoritma machine learning, institusi dapat membangun sistem prediksi yang mampu mengidentifikasi mahasiswa yang berisiko tidak lulus secara lebih awal, serta mengambil langkah intervensi yang lebih tepat (Romero & Ventura, 2020).

Data Science Pipeline menyediakan kerangka kerja yang terstruktur mulai dari pengumpulan, pembersihan, eksplorasi, pemodelan, hingga evaluasi dan visualisasi data (Cao, 2017). Melalui pendekatan ini, data nilai mahasiswa dapat digunakan untuk membangun model klasifikasi yang memprediksi status kelulusan secara akurat, serta segmentasi mahasiswa berdasarkan karakteristik nilai mereka dengan metode unsupervised learning seperti K-Means Clustering.

Dalam proyek ini, digunakan data yang mencakup berbagai fitur akademik seperti nilai ujian tengah semester, nilai ujian akhir, nilai tugas, kuis, proyek, serta partisipasi kelas. Analisis dilakukan menggunakan algoritma Logistic Regression dan Random Forest Classifier untuk klasifikasi kelulusan, serta K-Means

Clustering untuk segmentasi mahasiswa. Penelitian ini diharapkan memberikan wawasan yang komprehensif tentang karakteristik mahasiswa yang lulus dan tidak lulus, serta mendukung pengambilan keputusan akademik berbasis data.

1.2 Rumusan Masalah

Adapula rumusan masalah pada makalah ini adalah sebagai berikut:

1) Apa saja fitur akademik yang paling memengaruhi kelulusan mahasiswa?
2) Bagaimana membangun model klasifikasi yang mampu memprediksi status kelulusan mahasiswa dengan akurasi tinggi?
3) Bagaimana karakteristik klaster mahasiswa berdasarkan nilai-nilai akademik mereka?
4) Bagaimana hubungan antara komponen penilaian seperti nilai tugas, kuis, ujian, dan partisipasi terhadap kelulusan mahasiswa?
5) Bagaimana hasil segmentasi mahasiswa dapat digunakan untuk mendukung intervensi pembelajaran yang lebih tepat sasaran?
6) Bagaimana desain dashboard interaktif yang efektif untuk menampilkan hasil prediksi dan segmentasi kelulusan mahasiswa?
7) Apa metrik evaluasi yang digunakan dan bagaimana performa masing-masing model dalam memprediksi kelulusan?

1.3 Tujuan

Tujuan dari proyek ini adalah:

1) Mengidentifikasi fitur akademik yang memiliki korelasi signifikan terhadap kelulusan mahasiswa.
2) Membangun dan mengevaluasi model klasifikasi menggunakan Logistic Regression dan Random Forest.
3) Mengelompokkan mahasiswa menggunakan K-Means Clustering berdasarkan nilai akademik mereka.
4) Menganalisis kontribusi setiap komponen penilaian terhadap status kelulusan.
5) Menyediakan visualisasi interaktif melalui dashboard berbasis Streamlit untuk mendukung pengambilan keputusan.

6) Menyediakan dasar untuk strategi peningkatan kelulusan berdasarkan hasil analisis prediktif dan segmentasi.
7) Mengevaluasi performa model menggunakan metrik evaluasi klasifikasi yang relevan (akurasi, presisi, recall, F1-score).

1.4 Manfaat Aplikasi

Manfaat yang diharapkan dari aplikasi ini adalah:

1) Memberikan gambaran awal bagi pihak sekolah/universitas dalam mengidentifikasi siswa yang berisiko tidak lulus.
2) Menjadi alat bantu analisis bagi pendidik dalam pengambilan keputusan intervensi akademik.
3) Memberikan siswa umpan balik berdasarkan performa mereka yang bisa menjadi bahan introspeksi.

1.5 Batasan Masalah

Agar proyek ini terfokus dan terukur, maka batasan masalah yang ditetapkan adalah:

1) Dataset yang digunakan adalah Students Performance Dataset dari Kaggle yang berisi data akademik mahasiswa.
2) Analisis hanya mencakup fitur akademik seperti nilai tugas, kuis, proyek, ujian, dan partisipasi; fitur-fitur non-akademik tidak menjadi fokus utama dalam analisis ini.
3) Model klasifikasi dibatasi pada algoritma Logistic Regression dan Random Forest tanpa proses hyperparameter tuning mendalam.
4) Segmentasi menggunakan algoritma K-Means Clustering dengan jumlah cluster ditentukan sebanyak tiga.
5) Status kelulusan dianalisis sebagai variabel biner (lulus/tidak lulus) tanpa mempertimbangkan klasifikasi nilai yang lebih rinci.
6) Dashboard interaktif hanya menampilkan hasil klasifikasi dan segmentasi berdasarkan input nilai, tanpa terhubung dengan sistem informasi akademik eksternal.

7) Tidak dilakukan validasi model pada dataset eksternal sehingga hasil hanya berlaku untuk dataset yang digunakan.
8) Evaluasi model terbatas pada metrik akurasi, presisi, recall, dan F1-score tanpa pendekatan interpretabilitas lanjutan seperti SHAP atau LIME.

BAB II

PEMBAHASAN

2.1 Business Understanding

Kelulusan mahasiswa tepat waktu merupakan indikator penting dari efektivitas penyelenggaraan pendidikan tinggi. Tingkat kelulusan yang tinggi mencerminkan keberhasilan institusi dalam memberikan dukungan akademik, pembelajaran yang efektif, serta pengelolaan mahasiswa secara menyeluruh. Namun pada kenyataannya, tidak semua mahasiswa berhasil menyelesaikan studinya tepat waktu. Faktor-faktor seperti rendahnya nilai ujian, kurangnya partisipasi, dan ketidakhadiran dalam kelas menjadi penyebab umum yang dapat memengaruhi keberhasilan akademik mahasiswa (Tinto, 2012).

Permasalahan utama yang dihadapi oleh institusi pendidikan adalah kesulitan dalam mengidentifikasi mahasiswa yang berisiko tidak lulus sejak dini. Tanpa dukungan sistem yang mampu memprediksi performa mahasiswa secara objektif, intervensi yang diberikan seringkali terlambat atau tidak tepat sasaran. Dalam konteks ini, pemanfaatan data akademik mahasiswa yang telah dikumpulkan secara rutin menjadi sangat penting untuk menghasilkan wawasan yang dapat ditindaklanjuti.

Dengan kemajuan teknologi dalam bidang data science, khususnya machine learning, kini memungkinkan institusi pendidikan untuk menerapkan pendekatan prediktif berbasis data. Model prediktif ini dapat membantu dosen, wali akademik, maupun pengelola institusi untuk:

- Mengidentifikasi mahasiswa dengan risiko tidak lulus.
- Memahami faktor-faktor utama yang memengaruhi kelulusan.
- Merancang strategi pembelajaran atau intervensi yang lebih terarah.
- Proyek ini bertujuan untuk mengembangkan solusi prediktif dan analitis berbasis data mining yang dapat digunakan oleh institusi pendidikan untuk meningkatkan tingkat kelulusan mahasiswa. Fokus utama proyek adalah:

1) Model Prediksi Kelulusan (Classification)

   Menggunakan algoritma Logistic Regression dan Random Forest, model ini dibangun untuk mengklasifikasikan mahasiswa ke dalam dua kategori utama berdasarkan performa akademik mereka: Lulus (Pass) atau Tidak

Lulus (Fail). Model akan dilatih menggunakan data nilai-nilai akademik seperti ujian tengah semester, ujian akhir, tugas, proyek, kuis, dan partisipasi kelas.

2) Segmentasi Mahasiswa (Clustering)

   Untuk mendukung analisis lebih lanjut, dilakukan juga segmentasi mahasiswa menggunakan algoritma K-Means Clustering. Tujuannya adalah untuk mengelompokkan mahasiswa ke dalam beberapa klaster berdasarkan kemiripan pola nilai akademik mereka, yang dapat digunakan untuk memahami tipe-tipe mahasiswa dan memberikan pendekatan pembelajaran yang lebih personal.

3) Implementasi Aplikasi Interaktif

   Untuk memudahkan penggunaan model oleh pihak kampus, hasil analisis dan prediksi dikemas dalam bentuk dashboard interaktif berbasis web menggunakan framework Streamlit. Aplikasi ini memungkinkan pengguna (dosen, admin, atau mahasiswa) untuk:

   - Melihat hasil prediksi kelulusan berdasarkan input nilai individu.
   - Mengakses visualisasi segmentasi mahasiswa berdasarkan klaster.
   - Mengeksplorasi data dan pemodelan secara visual untuk pengambilan keputusan yang lebih cepat dan informatif.

Dengan pendekatan ini, institusi pendidikan diharapkan dapat mengoptimalkan peran data dalam proses pengambilan keputusan akademik, meningkatkan efisiensi intervensi, dan pada akhirnya, meningkatkan tingkat kelulusan mahasiswa secara keseluruhan.

2.2 Model Klasifikasi

Model klasifikasi difokuskan untuk memprediksi status kelulusan berdasarkan total skor akademik yang dihitung dari berbagai komponen penilaian. Target variabel Class dihasilkan berdasarkan kriteria:

- Pass jika total skor ≥ 70
- Fail jika total skor < 70

Fitur utama yang digunakan:

- Projects_Score
- Final_Score
- Midterm_Score
- Assignments_Avg
- Quizzes_Avg
- Participation_Score

Model klasifikasi ini diharapkan mampu:

- Memprediksi kelulusan mahasiswa dengan akurasi tinggi
- Memberikan insight tentang fitur paling berpengaruh
- Digunakan sebagai alat bantu monitoring performa akademik secara kuantitatif

2.2.1 Implementasi Aplikasi Streamlit

Sebagai bagian dari proses deployment, model diintegrasikan ke dalam aplikasi berbasis web menggunakan *Streamlit*. Aplikasi ini terdiri dari 7 halaman utama:

1) Dashboard Overview
2) Business Understanding
3) Data Understanding
4) Exploratory Data Analysis
5) Data Preprocessing
6) Modeling & Evaluation
7) Input Score for Prediction

Aplikasi memungkinkan pengguna memasukkan nilai mahasiswa dan langsung mendapatkan prediksi status kelulusan berdasarkan model yang telah dilatih.

2.2.2 Manfaat untuk Organisasi dan Stakeholder

1) Bagi Institusi Pendidikan:
    - Mempermudah analisis kelulusan mahasiswa secara real-time.
    - Mengurangi tingkat drop-out dengan intervensi dini berbasis data.

2) Bagi Dosen & Wali Akademik:
   - Memberikan informasi berbasis data untuk monitoring kemajuan mahasiswa.
   - Menyusun strategi pembelajaran personalisasi untuk mahasiswa berisiko.

3) Bagi Mahasiswa:
   - Mendapatkan umpan balik terhadap performa akademik.
   - Memahami kelemahan dan kelebihan berdasarkan prediksi model.

2.2.3 Indikator Keberhasilan Model

Keberhasilan proyek ini diukur melalui indikator evaluasi sebagai berikut:

1) Kinerja Model:
   - Akurasi: Persentase prediksi yang tepat dari keseluruhan prediksi.
   - F1-Score: Rata-rata harmonis dari presisi dan recall untuk menangani ketidakseimbangan kelas.
   - Precision & Recall: Presisi mengukur tingkat ketepatan prediksi Pass, sementara recall menilai sejauh mana model mampu mendeteksi semua mahasiswa yang benar-benar Pass.

2) Visualisasi Performa Model:
   - Confusion Matrix: Menunjukkan True/False Positive dan Negative.
   - Barplot Feature Importance: Menampilkan pengaruh relatif setiap fitur terhadap prediksi Class.

3) Kemudahan Penggunaan & Interpretasi:
   - Aplikasi web berbasis *Streamlit* memungkinkan pengguna non-teknis untuk mengakses prediksi dan analisis secara intuitif.
   - Tersedianya visualisasi interaktif memperkuat interpretabilitas hasil model bagi pengambil kebijakan.

2.3 Data Understandings

Tahap *data understanding* adalah proses awal dalam proyek data mining yang bertujuan untuk memahami struktur, isi, kualitas, dan potensi insight dari dataset. Pemahaman ini akan menjadi dasar penting dalam menentukan strategi preprocessing, eksplorasi, dan pemodelan data. Dalam proyek ini, dataset yang digunakan berasal dari sumber nyata dan telah disesuaikan untuk kebutuhan pembelajaran serta eksplorasi analitik prediktif.

2.2.1 Penjelasan Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 23 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Student_ID                 5000 non-null   object
 1   First_Name                 5000 non-null   object
 2   Last_Name                  5000 non-null   object
 3   Email                      5000 non-null   object
 4   Gender                     5000 non-null   object
 5   Age                        5000 non-null   int64
 6   Department                 5000 non-null   object
 7   Attendance (%)             5000 non-null   float64
 8   Midterm_Score              5000 non-null   float64
 9   Final_Score                5000 non-null   float64
 10  Assignments_Avg            5000 non-null   float64
 11  Quizzes_Avg                5000 non-null   float64
 12  Participation_Score        5000 non-null   float64
 13  Projects_Score             5000 non-null   float64
 14  Total_Score                5000 non-null   float64
 15  Grade                      5000 non-null   object
 16  Study_Hours_per_Week       5000 non-null   float64
 17  Extracurricular_Activities 5000 non-null   object
 18  Internet_Access_at_Home    5000 non-null   object
 19  Parent_Education_Level     3975 non-null   object
 20  Family_Income_Level        5000 non-null   object
 21  Stress_Level (1-10)        5000 non-null   int64
 22  Sleep_Hours_per_Night      5000 non-null   float64
dtypes: float64(10), int64(2), object(11)
memory usage: 898.6+ KB
```

Dataset yang digunakan berjudul Students Grading Dataset dari Kaggle (https://www.kaggle.com/datasets/mahmoudelhemaly/students-grading-dataset, ini merupakan data asli sebanyak 5.000 entri yang dikumpulkan dari sebuah penyedia layanan pendidikan swasta. Dataset ini mencerminkan kombinasi antara kinerja akademik dan perilaku mahasiswa, yang sangat relevan dalam membangun model prediksi

kelulusan. Tujuan utama dari dataset ini adalah untuk memberikan representasi realistis tentang faktor-faktor yang memengaruhi keberhasilan akademik. Data mencakup aspek kognitif (nilai ujian), non-kognitif (partisipasi, kegiatan ekstrakurikuler, stres, tidur), dan kondisi sosial (penghasilan keluarga, tingkat pendidikan orang tua, akses internet).

2.2.2 Struktur dan Karakteristik Dataset

Dataset ini terdiri dari 5.000 baris dan 24 kolom fitur, yang dapat diklasifikasikan ke dalam beberapa jenis:

Jenis Data	Contoh Kolom
Identifikasi	Student_ID, First_Name, Last_Name, Email
Demografis	Gender, Age, Department
Akademik	Midterm_Score, Final_Score, Assignments_Avg, Quizzes_Avg, Projects_Score, Participation_Score, Total_Score, Grade
Perilaku & Psikologis	Attendance (%), Study_Hours_per_Week, Sleep_Hours_per_Night, Stress_Level (1-10)
Sosial & Ekonomi	Extracurricular_Activities, Internet_Access_at_Home, Parent_Education_Level, Family_Income_Level

Beberapa kolom memiliki nilai missing/null, seperti:

- Attendance (%)
- Assignments_Avg
- Parent_Education_Level

Selain itu, terdapat bias yang sengaja ditambahkan oleh pemilik dataset, seperti:

- Mahasiswa dengan kehadiran tinggi cenderung memiliki nilai akhir yang sedikit lebih tinggi.
- Distribusi jumlah mahasiswa antar departemen tidak seimbang.

2.2.3 Deskriptif Numerik dan Kategorikal

1) Statistik Numerik:

Statistik Deskriptif Numerik:

	Age	Attendance (%)	Midterm_Score	Final_Score	Assignments_Avg	Quizzes_Avg
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	21.048400	75.356076	70.701924	69.546552	74.956320	74.836214
std	1.989786	14.392716	17.436325	17.108996	14.404287	14.423848
min	18.000000	50.010000	40.000000	40.010000	50.000000	50.000000
25%	19.000000	62.945000	55.707500	54.697500	62.340000	62.357500
50%	21.000000	75.670000	70.860000	69.485000	75.090000	74.905000
75%	23.000000	87.862500	85.760000	83.922500	87.352500	87.292500
max	24.000000	100.000000	99.990000	99.980000	99.990000	99.990000

Participation_Score	Projects_Score	Total_Score	Study_Hours_per_Week	Stress_Level (1-10)	Sleep_Hours_per_Night
5000.000000	5000.00000	5000.000000	5000.000000	5000.000000	5000.000000
49.963720	74.78305	71.652097	17.521140	5.507200	6.514420
28.989785	14.54243	7.230097	7.193035	2.886662	1.446155
0.000000	50.00000	50.602000	5.000000	1.000000	4.000000
25.075000	61.97000	66.533875	11.500000	3.000000	5.300000
49.600000	74.54000	71.696250	17.400000	6.000000	6.500000
75.500000	87.63000	76.711625	23.700000	8.000000	7.800000
100.000000	100.00000	95.091500	30.000000	10.000000	9.000000

2) Statistik Kategorikal:

First_Name:

	Jumlah	count
0	Maria	657
1	Ahmed	651
2	Ali	644
3	Emma	628
4	Sara	612
5	John	608
6	Omar	601
7	Liam	599

Last_Name:

	Jumlah	count
0	Johnson	868
1	Jones	850
2	Davis	829
3	Brown	825
4	Smith	817
5	Williams	811

Gender:

	Jumlah	count
0	Male	2551
1	Female	2449

Department:

	Jumlah	count
0	Engineering	1274
1	Business	1264
2	CS	1239
3	Mathematics	1223

Grade:

	Jumlah	count
0	C	2307
1	D	1760
2	B	638
3	F	279
4	A	16

Extracurricular_Activitie

	Jumlah	count
0	Yes	2512
1	No	2488

Internet_Access_at_Home:

	Jumlah	count
0	Yes	2520
1	No	2480

Parent_Education_Level:

	Jumlah	count
0	Bachelor's	1020
1	PhD	1012
2	Master's	1000
3	High School	943

Family_Income_Level:

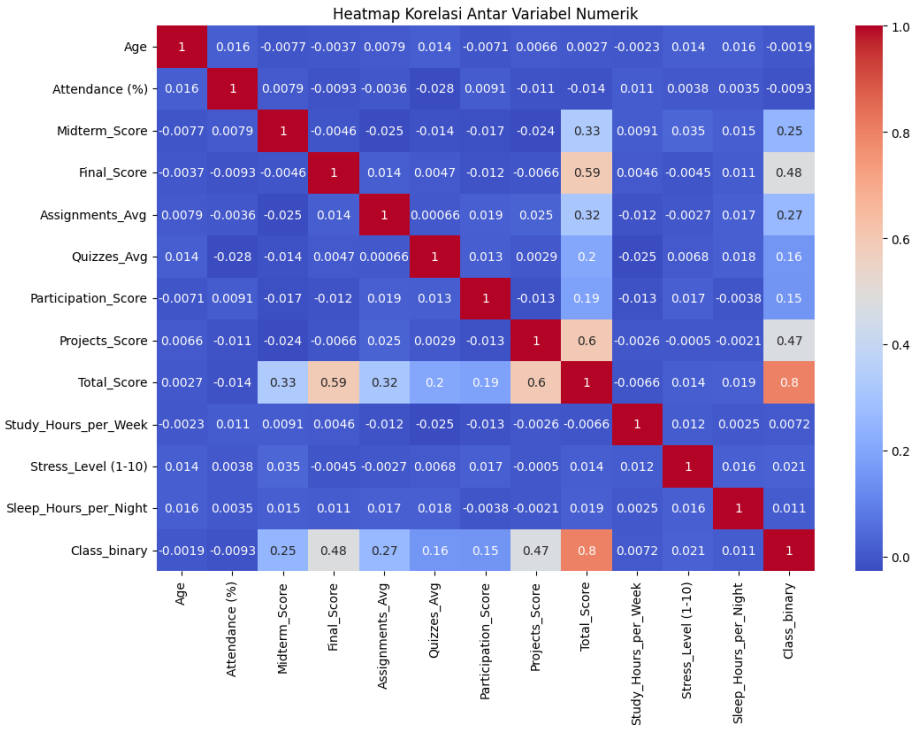
	Jumlah	count
0	Low	1687
1	Medium	1674
2	High	1639

- Gender: Male, Female
- Department: Business, CS, Engineering, Mathematics.
- Extracurricular_Activities: Yes / No.
- Internet_Access_at_Home: Yes / No.
- Parent_Education_Level: None, High School, Bachelor's, Master's, PhD.
- Family_Income_Level: Low, Medium, High.
- Grade: A, B, C, D, F.
- Class *(target)*: Pass / Fail — ditentukan berdasarkan Total_Score ≥ 70.

2.4 Data Exploration

2.3.1 Correlation Heatmap Analysis

Analisis korelasi bertujuan untuk memahami sejauh mana keterkaitan antar fitur numerik dalam dataset. Korelasi dihitung menggunakan metode Pearson yang mengukur hubungan linear antara dua variabel.

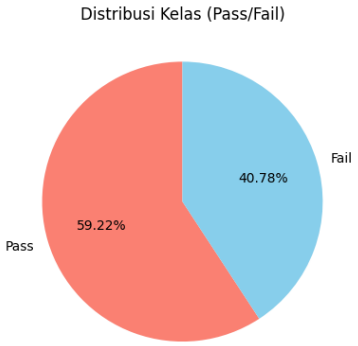


Dari hasil perhitungan korelasi dan visualisasi menggunakan heatmap, ditemukan beberapa hubungan signifikan:

- Fitur dengan korelasi terbesar terhadap kelulusan (Clas_binary) adalah Total_Score.
- Final_Score dan Projects_Score memiliki korelasi sedang
- Assignments_Avg, Midterm_Score, Quizzes_Avg, Participation_Score memiliki korelasi yang lemah
- Attendance (%), Age, Sleep_Hours_per_Night, Stress_Level (1-10), Study_Hours_per_Week memiliki fitur yang sangat lemah.

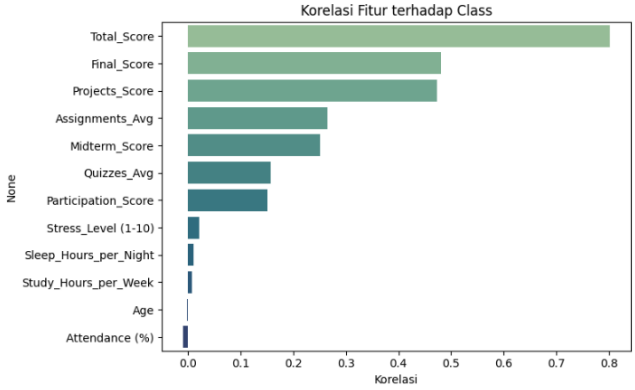
2.3.2 Distribusi Target Class

Distribusi nilai target atau kolom Class menunjukkan proporsi siswa yang lulus (Pass) dan tidak lulus (Fail).



Dari visualisasi diagram pie, diperoleh bahwa mayoritas siswa mendapatkan label "Pass" sebesar 59.22% sementara sisanya "Fail" sebesar 40.78%.

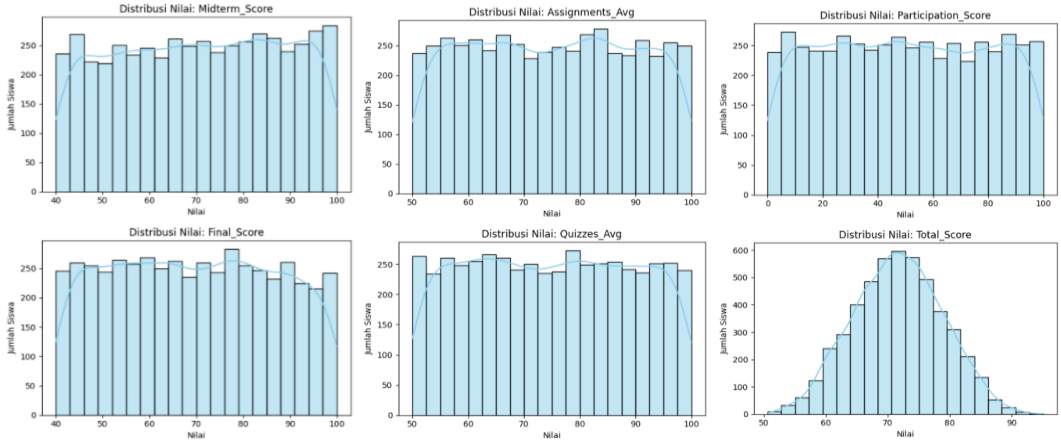
2.3.3 Korelasi Terhadap Class



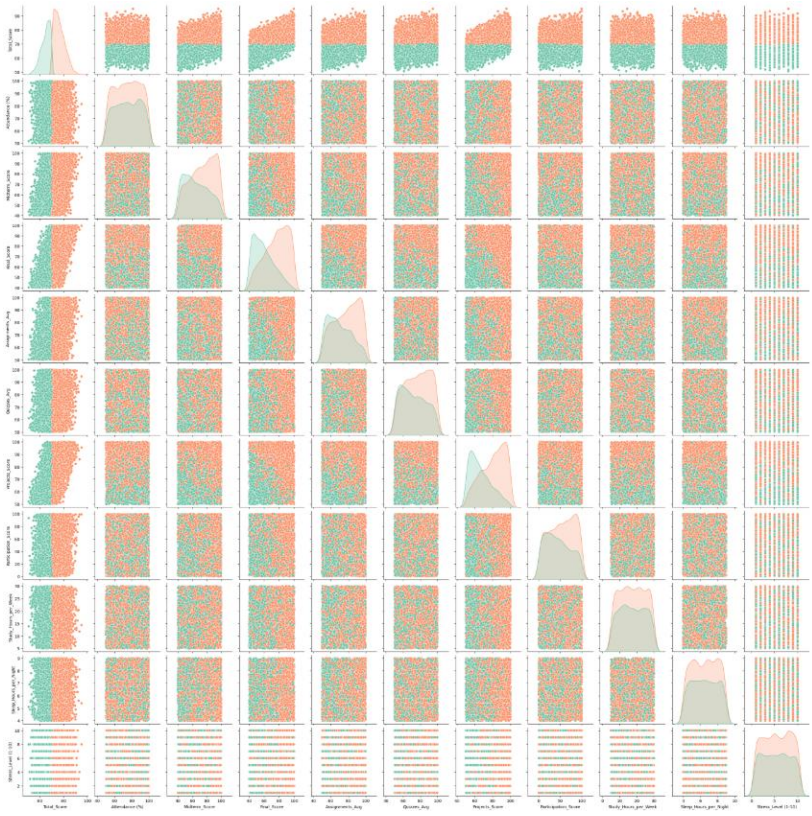
Dapat dilihat Total_score memiliki korelasi paling tinggi, nilai akademik memiliki korelasi yang lumayan besar dibandingkan fitur non-akademik yang sangat rendah dibawah 0.1.

2.3.4 Histogram Distribusi Nilai Fitur

Setiap fitur numerik dalam dataset dianalisis distribusinya menggunakan histogram untuk melihat pola nilai umum.



2.3.5 Pairplot Antar Fitur Penting



Pairplot digunakan untuk memvisualisasikan hubungan antar fitur numerik utama dan membandingkan distribusi kelas Fail (hijau) dan Pass (oranye). Fitur akademik seperti Total_Score, Final_Score, Midterm_Score memiliki pemisahan jelas antar class yang dapat digunakan untuk menentukan class atau kelulusan. Fitur non-akademik seperti Sleep_Hours, Study_Hours, dan Stress_Level tidak menunjukkan pola yang signifikan.

2.5 Data Preprocessing

2.4.1 Penanganan Data Kosong (Missing Values)

```
# Cek Missing Values
print("Missing Values per Kolom:")
print(df.isnull().sum())
```

```
Missing Values per Kolom:
Student_ID                      0
First_Name                      0
Last_Name                       0
Email                           0
Gender                          0
Age                             0
Department                      0
Attendance (%)                  0
Midterm_Score                   0
Final_Score                     0
Assignments_Avg                 0
Quizzes_Avg                     0
Participation_Score             0
Projects_Score                  0
Total_Score                     0
Grade                           0
Study_Hours_per_Week            0
Extracurricular_Activities      0
Internet_Access_at_Home         0
Parent_Education_Level       1025
Family_Income_Level             0
Stress_Level (1-10)             0
Sleep_Hours_per_Night           0
Class                           0
dtype: int64
```

```
[ ] # Parent_Education_Level tidak ak
df = df.dropna()
print(df.isnull().sum())
```

```
Student_ID                      0
First_Name                      0
Last_Name                       0
Email                           0
Gender                          0
Age                             0
Department                      0
Attendance (%)                  0
Midterm_Score                   0
Final_Score                     0
Assignments_Avg                 0
Quizzes_Avg                     0
Participation_Score             0
Projects_Score                  0
Total_Score                     0
Grade                           0
Study_Hours_per_Week            0
Extracurricular_Activities      0
Internet_Access_at_Home         0
Parent_Education_Level          0
Family_Income_Level             0
Stress_Level (1-10)             0
Sleep_Hours_per_Night           0
Class                           0
dtype: int64
```

Pada tahap awal, dilakukan pemeriksaan nilai kosong untuk memastikan bahwa tidak ada data yang hilang pada kolom-kolom penting. Hasilnya menunjukkan bahwa kolom Parent_Education_Level memiliki data kosong dan tidak akan digunakan dalam pemodelan, sehingga dihapus dari dataset df = df.dropna(). Setelah dilakukan penghapusan, dicek ulang untuk memastikan tidak ada nilai kosong pada kolom manapun.

2.4.2 Pemeriksaan dan Penghapusan Duplikat

```
# Cek dan hapus data duplikat
print("\nJumlah baris duplikat:", df.duplicated().sum())
print("\nTidak ada data duplikat")
```

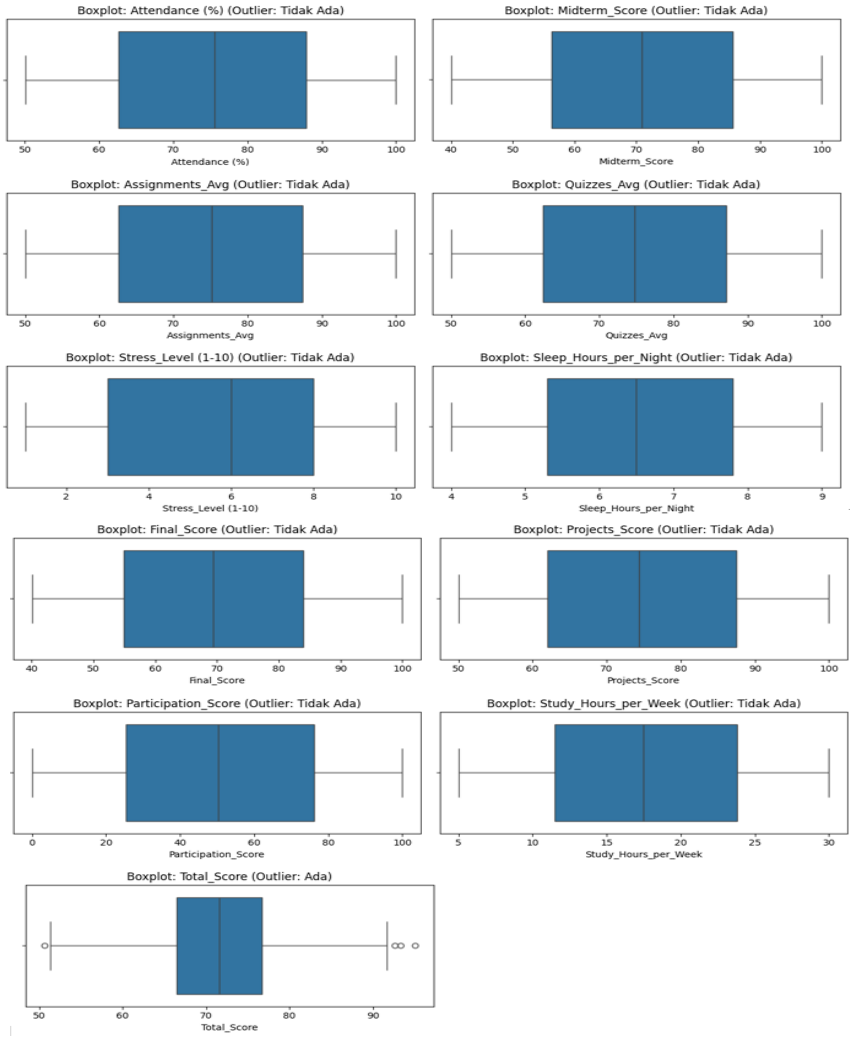
```
Jumlah baris duplikat: 0

Tidak ada data duplikat
```

Pemeriksaan dilakukan untuk memastikan tidak ada data yang tercatat lebih dari satu kali. Hasil pengecekan menunjukkan bahwa tidak terdapat data duplikat.

2.4.3 Deteksi Outlier

Hasil visualisasi outlier menggunakan boxplot menunjukkan bahwa Total_score memiliki autlier sebanyak 4, dan data lainnya tidak ada outlier.



2.4.4 Pemilihan Fitur (Feature Selection)

Kolom Total_Score tidak digunakan karena merupakan gabungan langsung dari fitur-fitur input seperti Midterm, Final, dan Assignments. Jika digunakan dalam pemodelan, hal ini dapat menyebabkan model "curang" dengan belajar dari informasi sudah ada yang memiliki korelasi terlalu kuat, ini disebut data leakage dan Total_score juga dihapus untuk menghindari overfitting. Dan data yang memiliki korelasi terlalu kecil terhadap class (kelulusan) seperti 'Attendance (%)', 'Study_Hours_per_Week', 'Sleep_Hours_per_Night', dan 'Stress_Level (1-10)' juga tidak digunakan untuk menghindari noise dan underfitting. Sehingga fitur-fitur yang dipilih untuk klasifikasi adalah:

- Projects_Score
- Final_Score
- Midterm_Score
- Assignments_Avg
- Quizzes_Avg
- Participation_Score
- Class (target)

2.4.5 Encoding Target

Target Class memiliki dua kategori: Pass dan Fail. Proses encoding dilakukan menggunakan LabelEncoder, di mana:

```
[ ] # Encode target Class
    from sklearn.preprocessing import LabelEncoder
    le = LabelEncoder()
    df['Class'] = le.fit_transform(df['Class'])  # Pass = 1, Fail = 0
    df.head()
```

	Projects_Score	Final_Score	Midterm_Score	Assignments_Avg	Quizzes_Avg	Participation_Score	Class
0	62.84	59.61	40.61	73.69	53.17	73.4	0
1	98.23	74.00	57.27	74.23	98.23	88.0	1
2	91.22	63.85	41.84	85.85	50.00	4.7	0
3	55.48	44.44	45.65	68.10	66.27	4.2	0
4	87.43	61.77	53.13	67.66	83.98	64.3	1

2.4.6 Normalisasi (Scaling)

Fitur input dinormalisasi menggunakan MinMaxScaler agar semua nilai berada pada skala 0–1. Hal ini penting terutama untuk model seperti Logistic Regression dan K-Means.

```
[17] # Pisahkan fitur dan target
     X = df.drop(columns='Class')
     y = df['Class']

     # Reset index agar sinkron saat digabungkan kembali
     X = X.reset_index(drop=True)
     y = y.reset_index(drop=True)

     # Scaling
     scaler = MinMaxScaler()
     X_scaled = scaler.fit_transform(X)

     # Gabungkan kembali
     df_scaled = pd.DataFrame(X_scaled, columns=X.columns)
     df_scaled['Class'] = y

     print("\nSebelum Scaling")
     print(df.head())
     print("\nSetelah Scaling")
     print(df_scaled.head())
```

```
Sebelum Scaling
   Projects_Score  Final_Score  Midterm_Score  Assignments_Avg  Quizzes_Avg  \
0           62.84        59.61          40.61            73.69        53.17
1           98.23        74.00          57.27            74.23        98.23
2           91.22        63.85          41.84            85.85        50.00
3           55.48        44.44          45.65            68.10        66.27
4           87.43        61.77          53.13            67.66        83.98

   Participation_Score  Class
0                 73.4      0
1                 88.0      1
2                  4.7      0
3                  4.2      0
4                 64.3      1

Setelah Scaling
   Projects_Score  Final_Score  Midterm_Score  Assignments_Avg  Quizzes_Avg  \
0          0.2568     0.326830       0.010168         0.473895     0.063413
1          0.9646     0.566783       0.287881         0.484697     0.964793
2          0.8244     0.397532       0.030672         0.717143     0.000000
3          0.1096     0.073870       0.094182         0.362072     0.325465
4          0.7486     0.362848       0.218870         0.353271     0.679736

   Participation_Score  Class
0                0.734      0
1                0.880      1
2                0.047      0
3                0.042      0
4                0.643      1
```

2.4.7 Penyimpanan Data dan Model Scaling

Data yang telah bersih dan terskalakan disimpan dalam file cleaned_data.csv dan scaler disimpan dalam file scaler.pkl untuk digunakan kembali dalam prediksi dan input data baru.

2.6 Modeling dan Evaluation

Bagian ini menjelaskan proses pembangunan model dan evaluasi yang dilakukan untuk memprediksi kelulusan mahasiswa berdasarkan nilai-nilai akademik mereka. Tiga metode digunakan: dua algoritma supervised learning (Logistic Regression dan Random Forest) serta satu metode unsupervised learning (K-Means Clustering). Dataset yang digunakan telah melalui tahap pembersihan dan normalisasi menggunakan MinMaxScaler.

2.5.1 Splitting Data

Sebelum model dilatih, data dibagi menjadi dua bagian yaitu data latih dan data uji. Pembagian ini dilakukan menggunakan fungsi train_test_split dari pustaka sklearn.model_selection dengan proporsi 80% data latih dan 20% data uji. Parameter random_state=42 digunakan untuk memastikan hasil pembagian data yang konsisten.

Stratifikasi digunakan agar distribusi label Class tetap seimbang di antara data latih dan uji. Tidak dilakukan proses balancing menggunakan SMOTE, karena data sudah cukup seimbang sejak awal.

2.5.2 Supervised Learning

2.5.2.1 Logistic Regression

Logistic Regression digunakan sebagai model baseline untuk klasifikasi kelulusan mahasiswa. Model ini dilatih pada data latih, dan hasil prediksinya diuji pada data uji. Nilai maksimum iterasi ditingkatkan ke 1000 untuk memastikan konvergensi.

Hasil evaluasi menunjukkan metrik sebagai berikut:

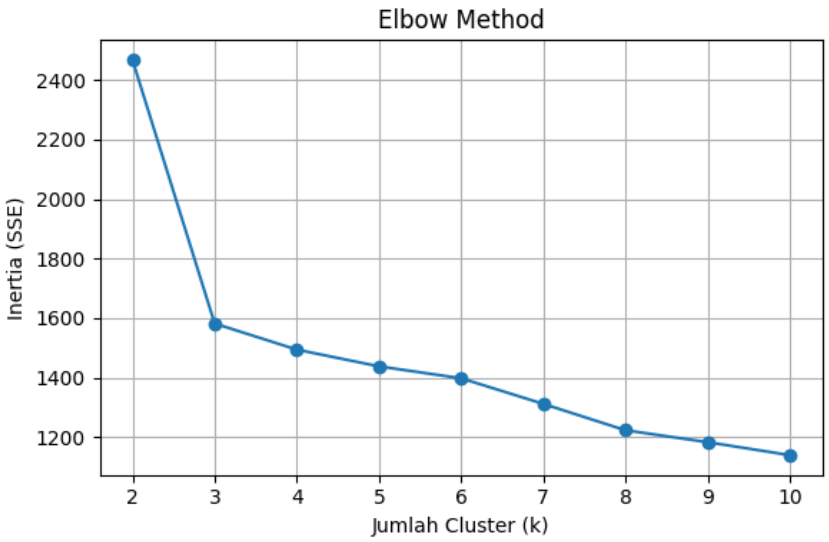
- Akurasi: xx.xx%
- Presisi, Recall, dan F1-Score dihitung dan ditampilkan bersama confusion matrix.
- Visualisasi confusion matrix dibuat untuk menilai prediksi terhadap kelas Pass dan Fail.

2.5.2.2 Random Forest

Random Forest digunakan sebagai model alternatif dengan pendekatan ensemble. Model ini memberikan performa yang lebih baik karena mampu menangani non-linearitas dan mengurangi overfitting melalui agregasi banyak pohon keputusan. Selain evaluasi metrik yang serupa dengan Logistic Regression, visualisasi **feature importance** ditampilkan untuk menunjukkan fitur mana yang paling berpengaruh terhadap prediksi model.

2.5.3 Unsupervised Learning: K-Means Clustering

K-Means digunakan untuk mengeksplorasi pola-pola pengelompokan mahasiswa berdasarkan fitur nilai akademik. Data yang digunakan adalah hasil scaling menggunakan MinMaxScaler, dan tidak termasuk label Class. Jumlah klaster yang dipilih dalam proses K-Means Clustering adalah tiga (k=3). Meskipun nilai Silhouette Score tertinggi terdapat pada k=2 (0.4034), perbedaan skor dengan k=3 (0.3982) relatif kecil. Pemilihan tiga klaster memberikan segmentasi yang lebih informatif, memungkinkan pengelompokan mahasiswa ke dalam kategori berprestasi tinggi, sedang, dan berisiko, sehingga lebih bermanfaat untuk analisis dan pengambilan keputusan akademik dibandingkan hanya dua kelompok yang bersifat biner.



```
Tabel Silhouette Score:
 Jumlah Cluster (k)  Silhouette Score
                  2          0.403448
                  3          0.398215
                  4          0.312919
                  5          0.308234
                  6          0.305059
                  7          0.227215
                  8          0.150830
                  9          0.149659
                 10          0.154128
```

Untuk keperluan visualisasi, dilakukan reduksi dimensi menggunakan PCA menjadi 2 komponen. Hasil cluster divisualisasikan menggunakan scatter plot dengan warna berbeda untuk tiap cluster. Interpretasi cluster berdasarkan rata-rata nilai per fitur:

2.5.4 Penyimpanan Model

Setelah model dilatih dan dievaluasi, model disimpan dalam file .pkl untuk digunakan dalam aplikasi web prediksi menggunakan streamlit .

BAB III

HASIL DAN PEMBAHASAN

3.1 Hasil Evaluasi Model

3.1.1 Hasil Evaluasi Model Klasifikasi

Dalam proyek ini, dua algoritma supervised learning digunakan untuk melakukan klasifikasi kelulusan mahasiswa berdasarkan nilai mereka, yaitu:

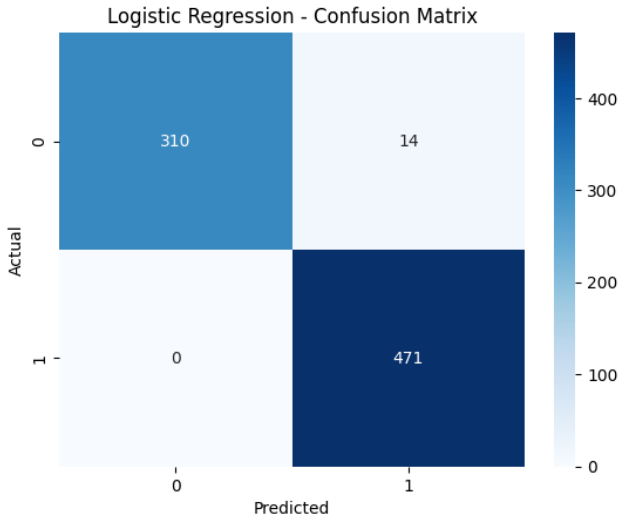
- Logistic Regression
- Random Forest Classifier

Dataset yang digunakan telah melalui proses pembersihan (cleaning), penghapusan outlier, feature selection, dan scaling menggunakan MinMaxScaler. Data kemudian dibagi menggunakan fungsi train_test_split dengan proporsi 80% data latih dan 20% data uji serta stratify=y dan random_state=42 untuk hasil yang konsisten.

1) Evaluasi Logistic Regression

Model Logistic Regression dilatih menggunakan parameter default dan hasil prediksinya diuji pada data uji. Hasil Evaluasi Logistic Regression:

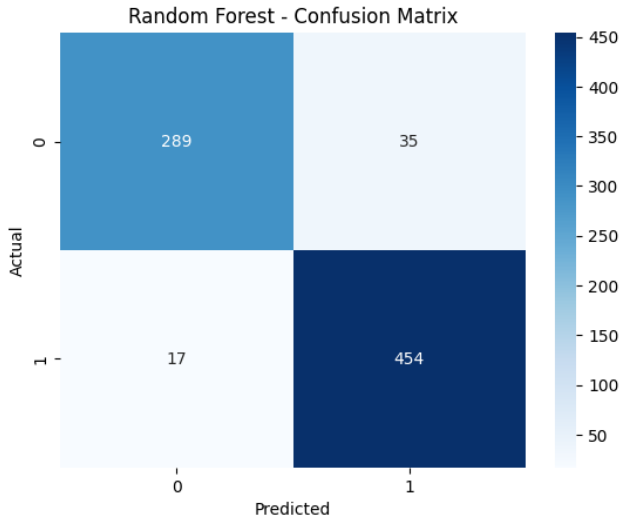
- Akurasi :0.9824
- Presisi :0.9711
- Recall :1.0000
- F1-Score:0.9854



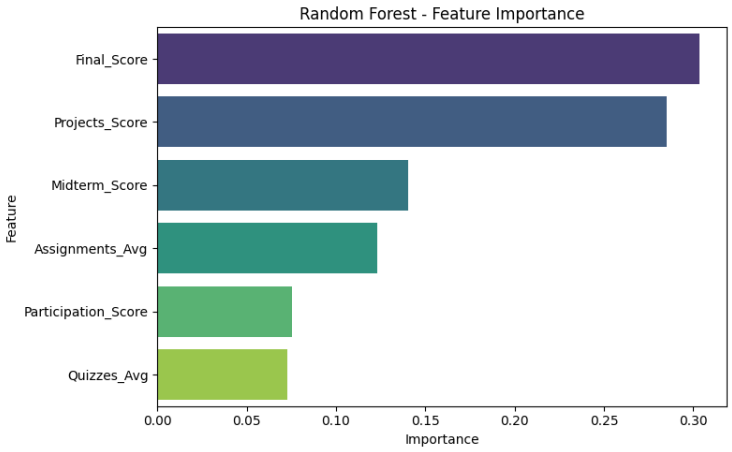
2) Evaluasi Random Forest

Model Random Forest dilatih dengan parameter default dan menunjukkan performa yang lebih baik dibandingkan Logistic Regression. Hasil Evaluasi Random Forest:

- Akurasi : 0.9346
- Presisi : 0.9284
- Recall : 0.9639
- F1-Score : 0.9458



Dan hasil visualisasi feature importance untuk random forest menunjukkan bahwa Final_Score dan Projects_Score memiliki nilai paling besar, Midterm_Score dan Assignment_Score memiliki nilai sedang, dan fitur Participation_Score dan Quizzes_Score memiliki score paling kecil.

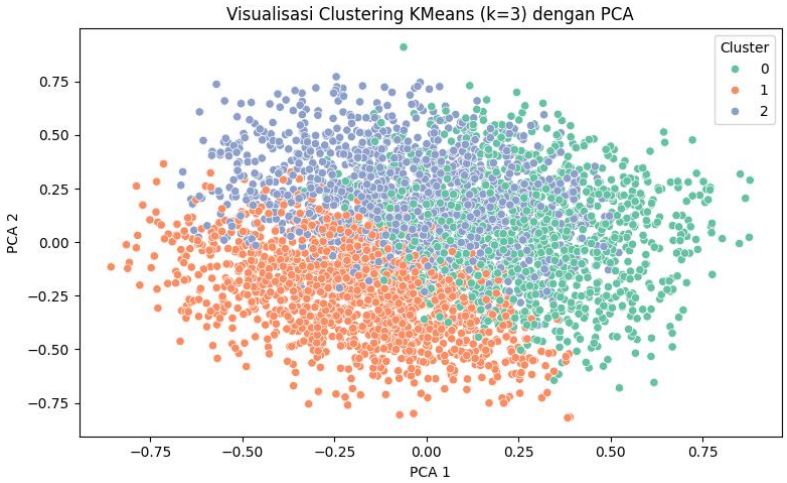


3.1.2 Hasil Evaluasi Model Klustering (K-Means)

Selain klasifikasi, dilakukan juga unsupervised learning menggunakan algoritma K-Means Clustering untuk mengelompokkan mahasiswa berdasarkan pola nilai. K-Means dilakukan dengan 3 cluster berdasarkan eksplorasi awal. Dengan fitur yang Digunakan:

- Projects_Score
- Final_Score
- Midterm_Score
- Assignments_Avg
- Quizzes_Avg
- Participation_Score

Hasil Clustering:




```
Rata-rata fitur per Cluster:
         Projects_Score  Final_Score  Midterm_Score  Assignments_Avg  \
Cluster
0              0.695634     0.403505       0.288860         0.530119
1              0.446870     0.544027       0.616818         0.478204
2              0.354152     0.524321       0.613384         0.494713

         Quizzes_Avg  Participation_Score
Cluster
0           0.575902             0.548159
1           0.460238             0.220267
2           0.455621             0.761530
```

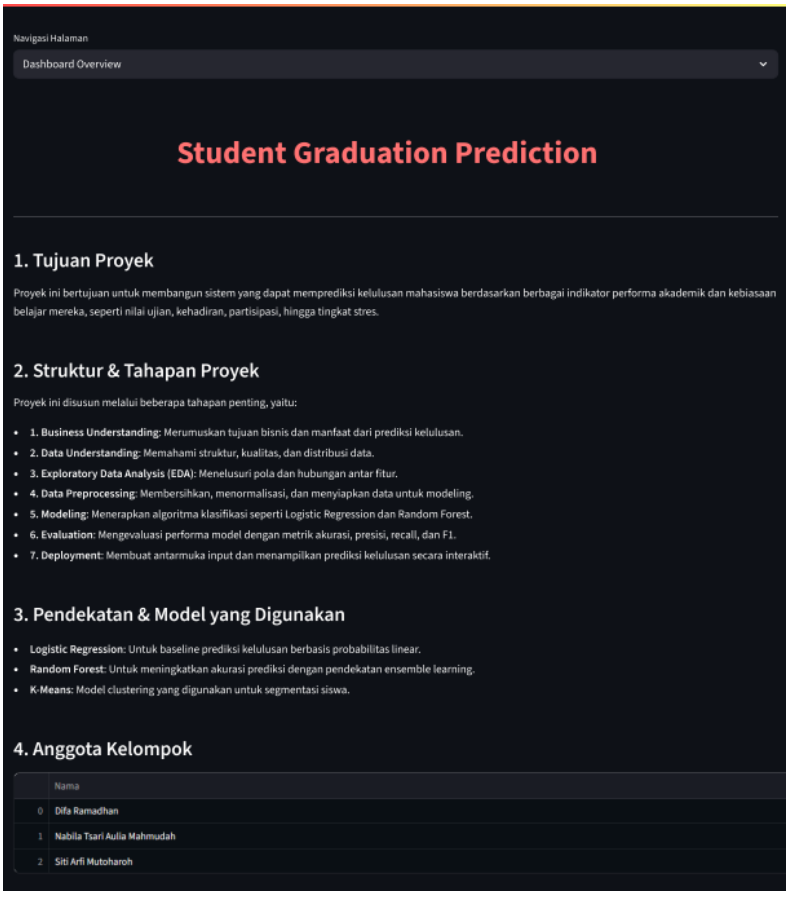
1) Cluster 0:
   - Tinggi di: Projects_Score (0.69), Quizzes_Avg (0.57), Assignments_Avg (0.53)
   - Rendah di: Final_Score (0.40), Midterm_Score (0.28)
   - Interpretasi: Siswa yang rajin dan konsisten dalam tugas dan proyek, tetapi kurang unggul di ujian. Cocok disebut sebagai tipe "pekerja keras" yang mungkin tidak terlalu bagus dalam ujian tulis.
2) Cluster 1:
   - Tinggi di: Midterm_Score (0.61), Final_Score (0.54)
   - Rendah di: Participation_Score (0.22), Projects_Score (0.44)
   - Interpretasi: Siswa dengan nilai ujian tinggi tapi tidak aktif dalam kelas dan tidak terlalu menonjol dalam tugas/proyek. Cocok disebut sebagai tipe independen dan pintar, tapi kurang terlibat secara aktif di kelas.

3) Cluster 2:

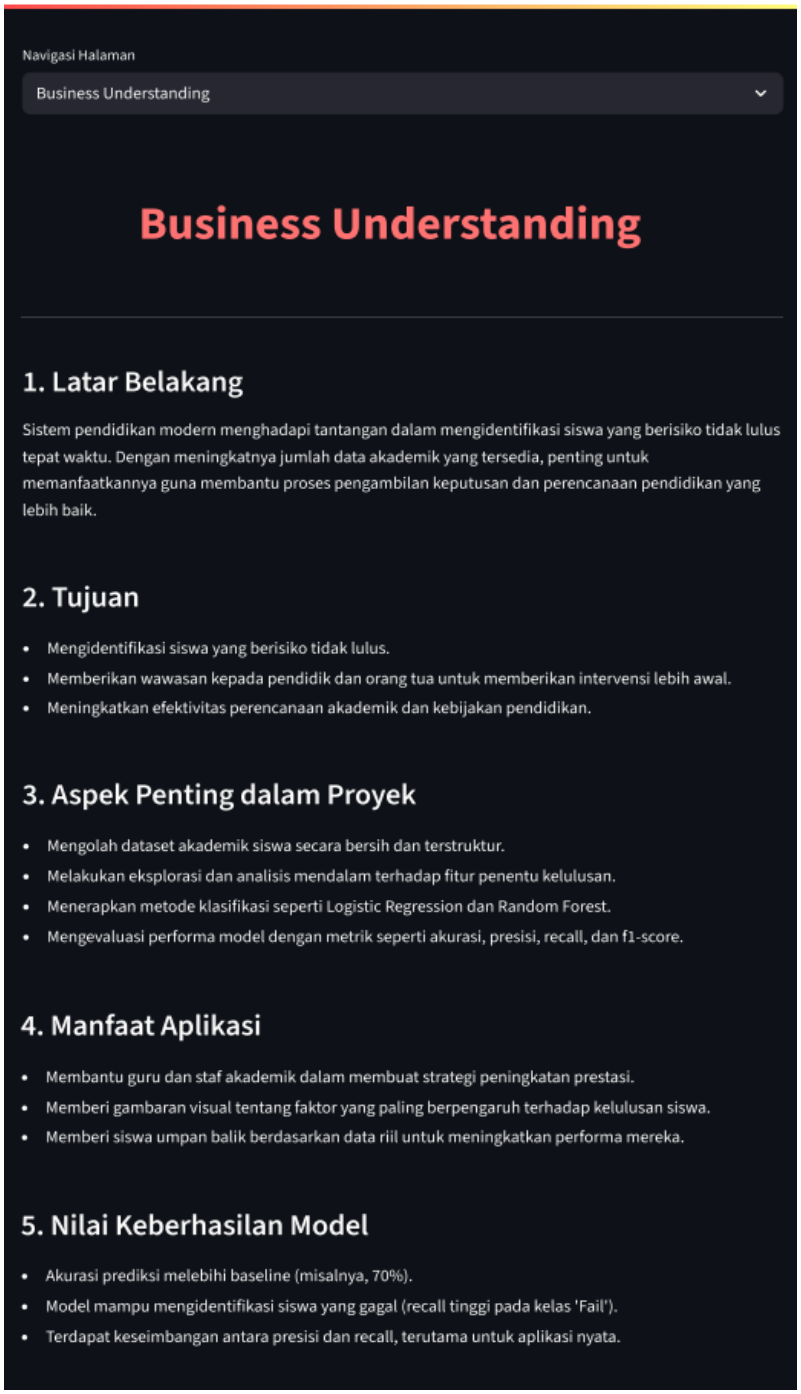
- Tinggi di: Participation_Score (0.76), Midterm_Score (0.61)
- Rendah di: Projects_Score (0.35)
- Interpretasi: Siswa dengan nilai ujian tinggi tapi tidak terlalu menonjol dalam tugas proyek dan memiliki tingkat keaktifan paling tinggi di kelas. Cocok disebut sebagai tipe independen dan pintar.

3.2 Tampilan Dashboard Streamlit

3.2.1 Dashboard (Overview)



3.2.2 Business Understanding



Navigasi Halaman

Data Understanding

Data Understanding

1. Preview Dataset

	Student_ID	First_Name	Last_Name	Email	Gender	Age	Department	Attendance (%)	Midterm_Score	Final_Score	As
0	S1000	Omar	Williams	student0@university.com	Female	22	Mathematics	97.36	40.61	59.61	
1	S1001	Maria	Brown	student1@university.com	Male	18	Business	97.71	57.27	74	
2	S1002	Ahmed	Jones	student2@university.com	Male	24	Engineering	99.52	41.84	63.85	
3	S1003	Omar	Williams	student3@university.com	Female	24	Engineering	90.38	45.65	44.44	
4	S1004	John	Smith	student4@university.com	Female	23	CS	59.41	53.13	61.77	

2. Ukuran Dataset

Jumlah Baris: 5000

Jumlah Kolom: 23

3. Tipe Data Tiap Kolom

	Kolom
0	Student_ID
1	First_Name
2	Last_Name
3	Email
4	Gender
5	Age
6	Department
7	Attendance (%)
8	Midterm_Score
9	Final_Score

4. Missing Values

0	Student_ID
1	First_Name
2	Last_Name
3	Email
4	Gender
5	Age
6	Department
7	Attendance (%)
8	Midterm_Score
9	Final_Score

5. Duplikat

Jumlah Baris Duplikat: 0

6. Statistik Deskriptif Kolom Numerik

	Age	Attendance (%)	Midterm_Score	Final_Score	Assignments_Avg
count	5000	5000	5000	5000	
mean	21.0484	75.3561	70.7019	69.5466	7
std	1.9898	14.3927	17.4363	17.109	1
min	18	50.01	40	40.01	
25%	19	62.945	55.7075	54.6975	
50%	21	75.67	70.86	69.485	
75%	23	87.8625	85.76	83.9225	
max	24	100	99.99	99.98	

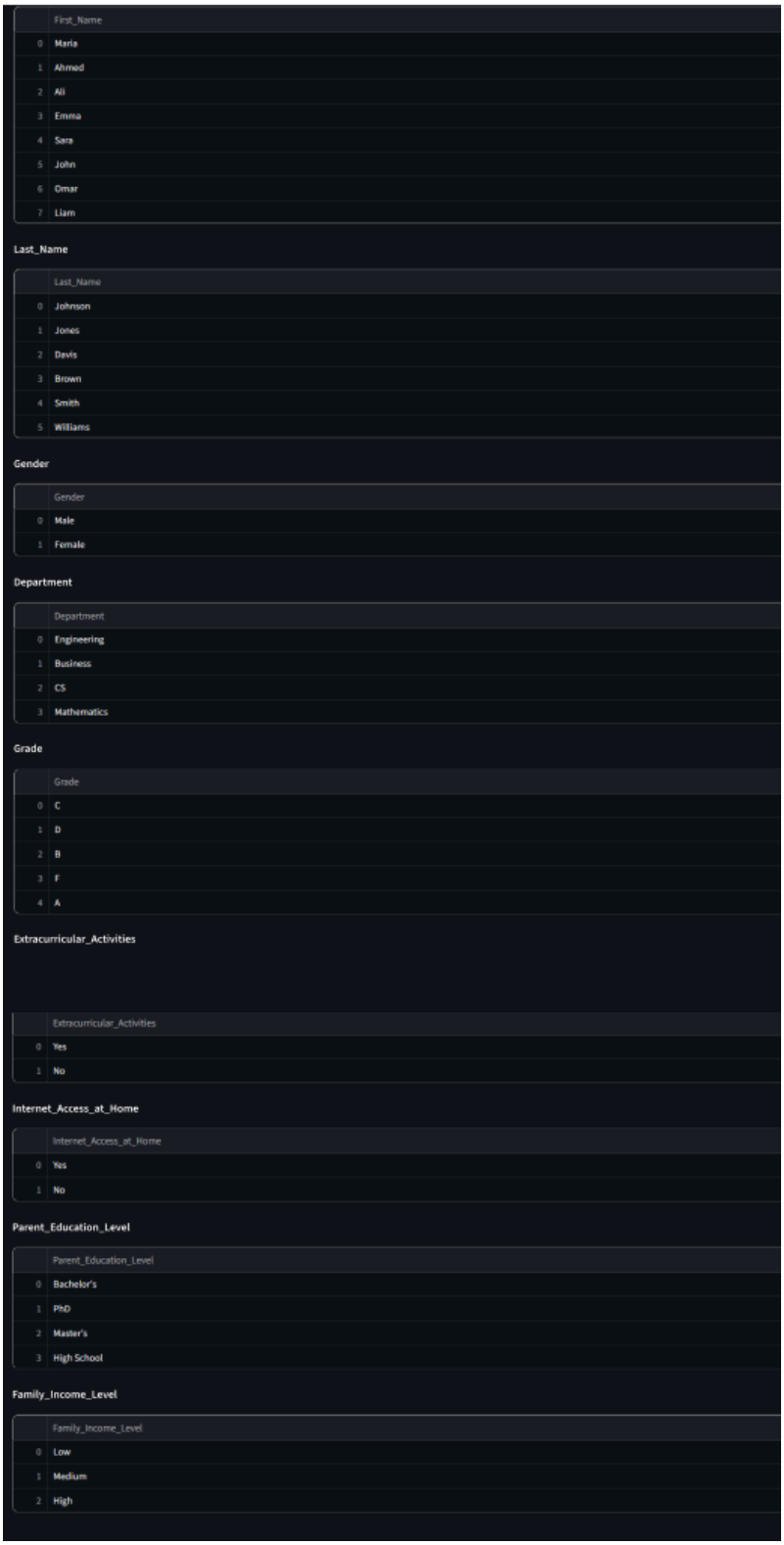
7. Statistik Deskriptif Kolom Kategorikal

	Student_ID	First_Name	Last_Name	Email	Gend
count	5000	5000	5000	5000	5000
unique	5000	8	6	5000	2
top	S5999	Maria	Johnson	student4999@university.com	Male
freq	1	657	868	1	2551

8. Jumlah Kemunculan Unique Data di Kolom Kategorikal

First_Name

3.2.3 Data Understanding



	First_Name
0	Maria
1	Ahmed
2	Ali
3	Emma
4	Sara
5	John
6	Omar
7	Liam

Last_Name

	Last_Name
0	Johnson
1	Jones
2	Davis
3	Brown
4	Smith
5	Williams

Gender

	Gender
0	Male
1	Female

Department

	Department
0	Engineering
1	Business
2	CS
3	Mathematics

Grade

	Grade
0	C
1	D
2	B
3	F
4	A

Extracurricular_Activities

	Extracurricular_Activities
0	Yes
1	No

Internet_Access_at_Home

	Internet_Access_at_Home
0	Yes
1	No

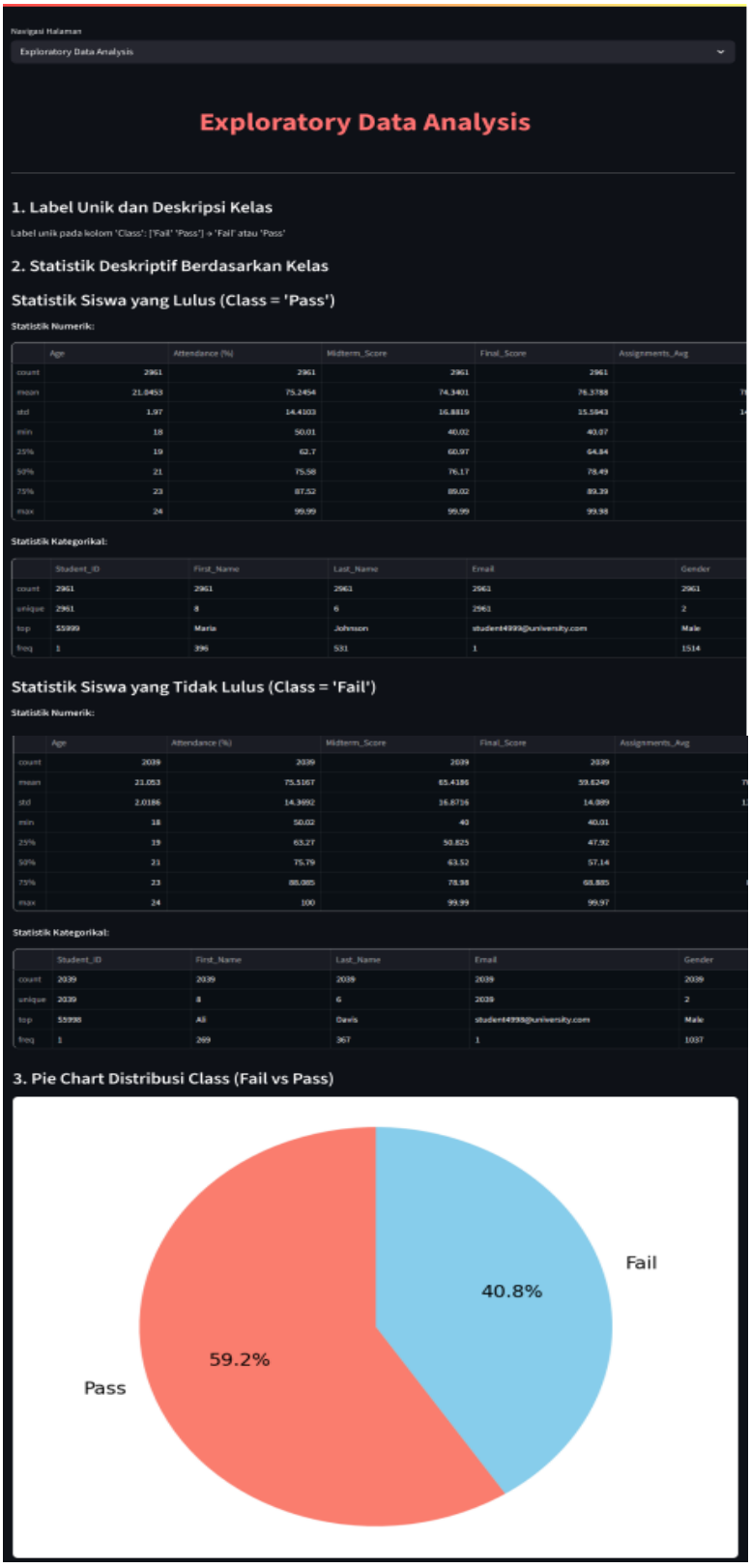
Parent_Education_Level

	Parent_Education_Level
0	Bachelor's
1	PhD
2	Master's
3	High School

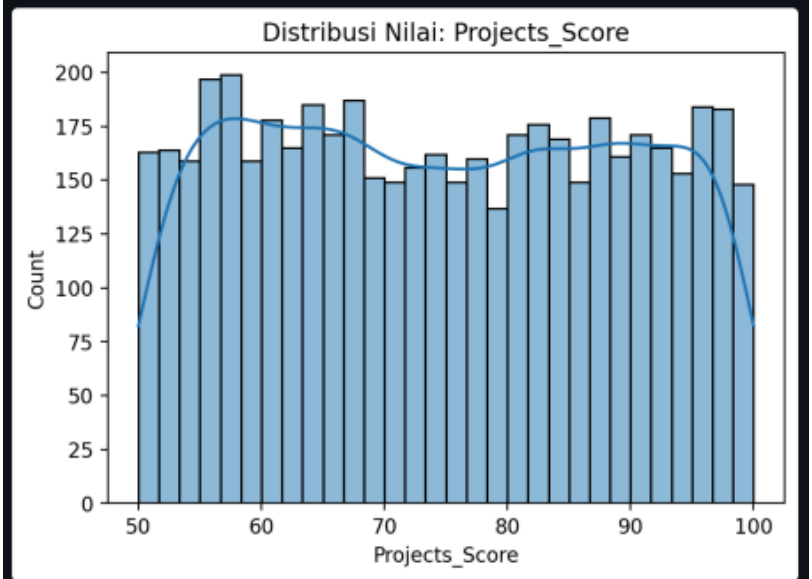
Family_Income_Level

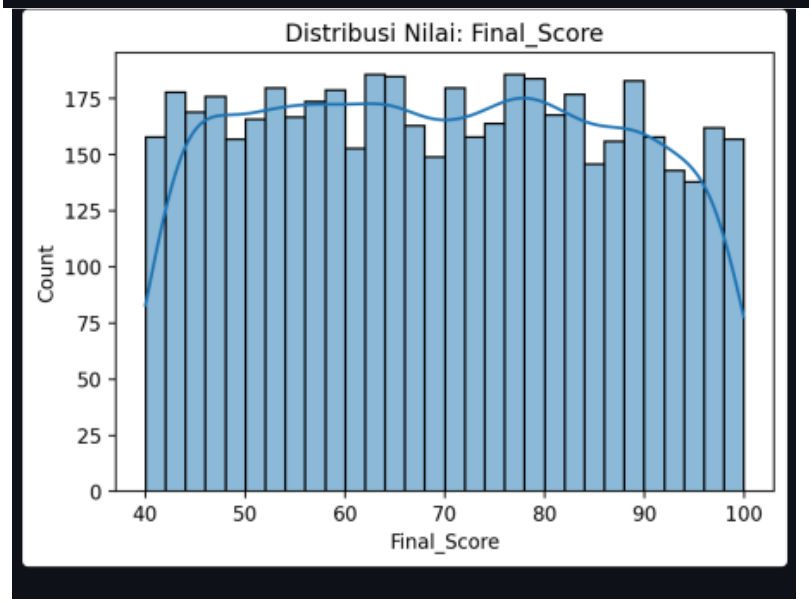
	Family_Income_Level
0	Low
1	Medium
2	High

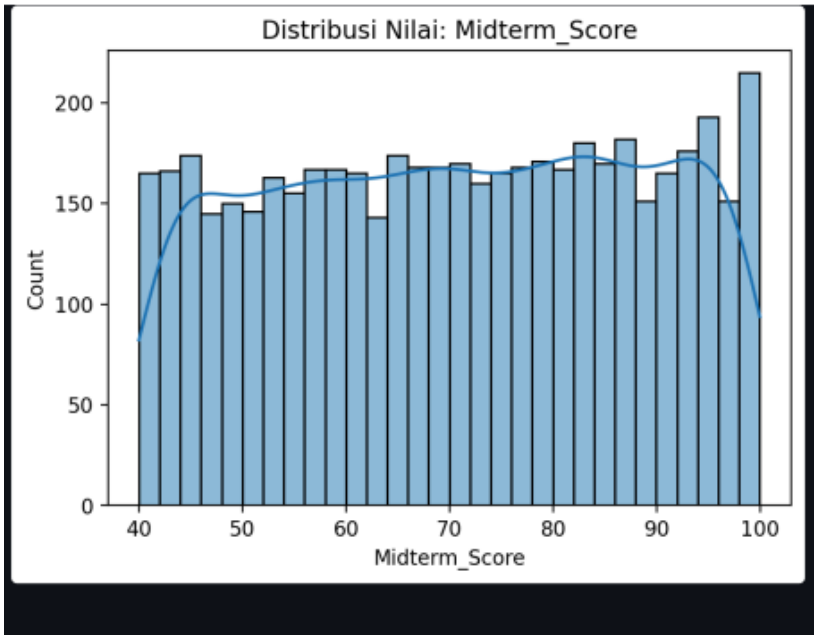
3.2.4 Data Exploration

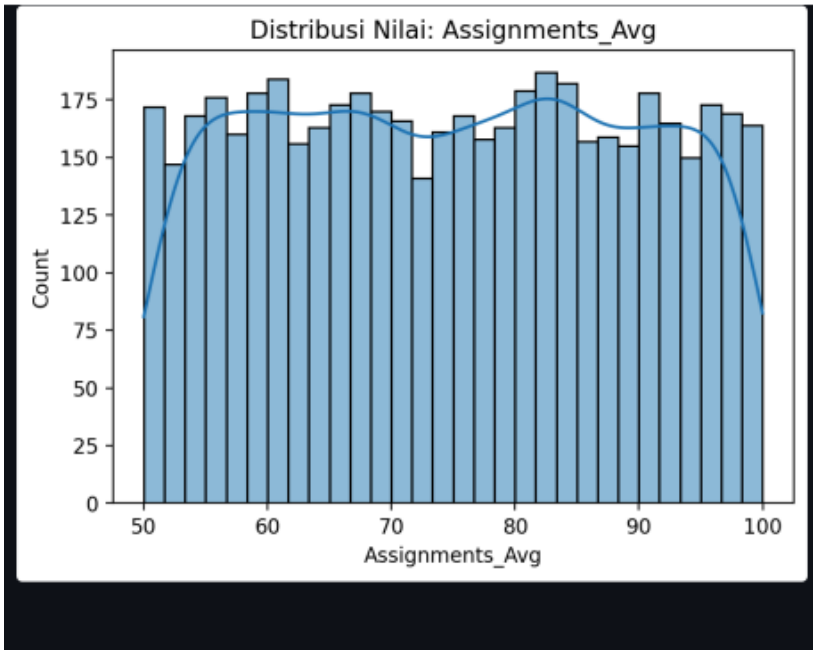


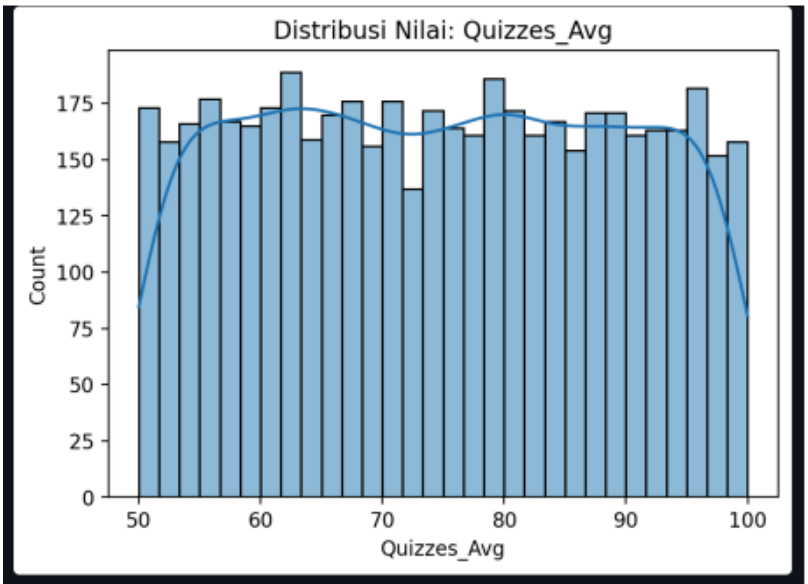
4. Histogram Distribusi Tiap Fitur

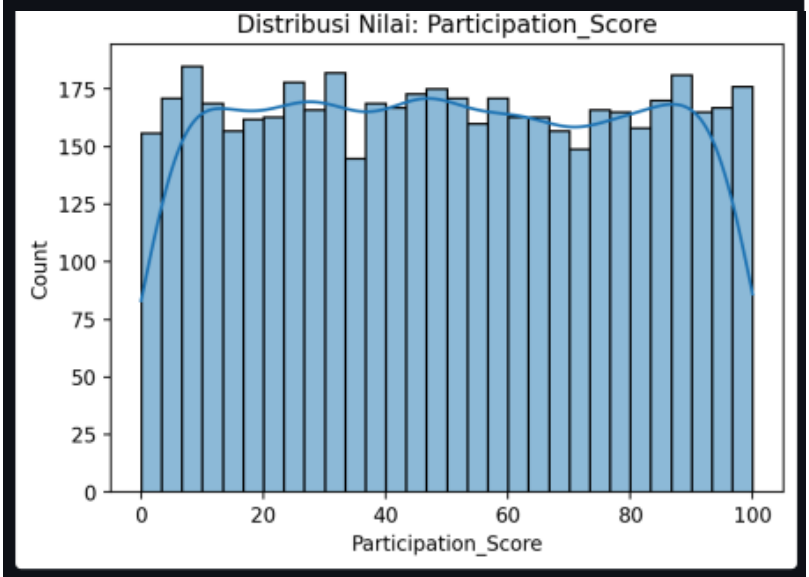





Distribusi Nilai: Midterm_Score
Count
0
50
100
150
200
40
50
60
70
80
90
100
Midterm_Score


Distribusi Nilai: Assignments_Avg
Count
0
25
50
75
100
125
150
175
50
60
70
80
90
100
Assignments_Avg

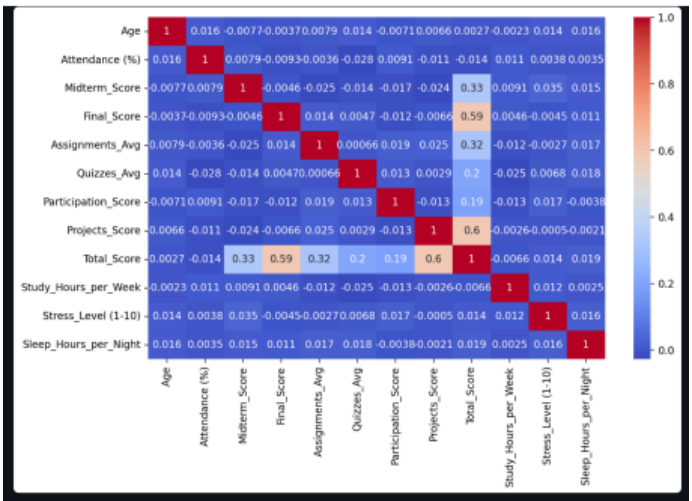




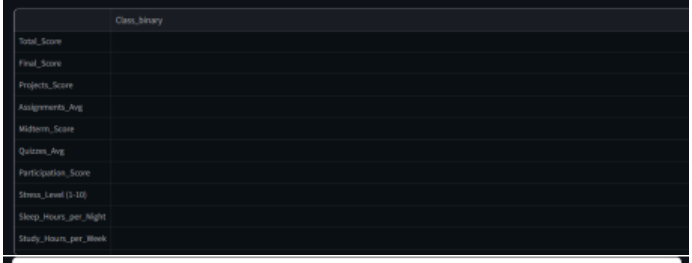
5. Korelasi Antar Fitur Numerik

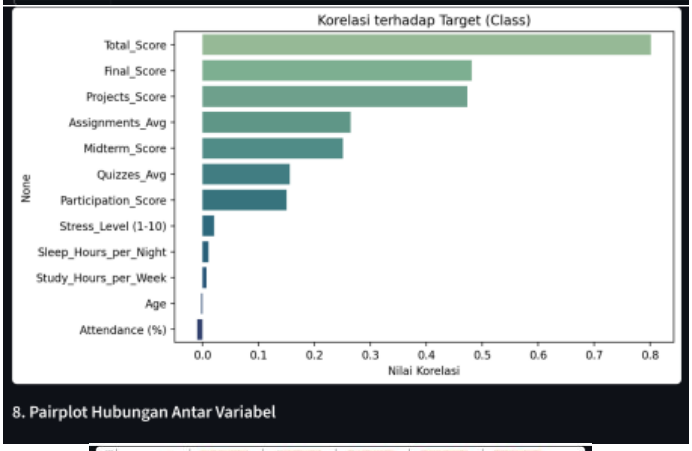
	Age	Attendance (%)	Midterm_Score	Final_Score	Assignments_Avg
Age	1	0.0159	-0.0077	-0.0037	
Attendance (%)	0.0159	1	0.0079	-0.0093	
Midterm_Score	-0.0077	0.0079	1	-0.0046	
Final_Score	-0.0037	-0.0093	-0.0046	1	
Assignments_Avg	0.0079	-0.0036	-0.0252	0.0136	
Quizzes_Avg	0.014	-0.0278	-0.014	0.0047	
Participation_Score	-0.0071	0.0091	-0.0174	-0.0118	
Projects_Score	0.0066	-0.0111	-0.0241	-0.0066	
Total_Score	0.0027	-0.0342	0.3306	0.5888	
Study_Hours_per_Week	-0.0023	0.0112	0.0091	0.0046	

6. Heatmap Korelasi

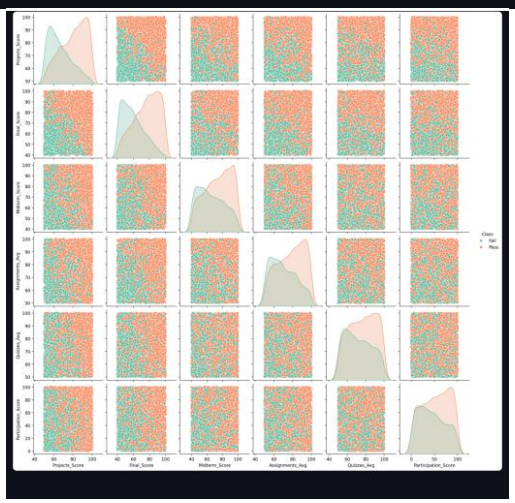


7. Korelasi terhadap Target (Class)

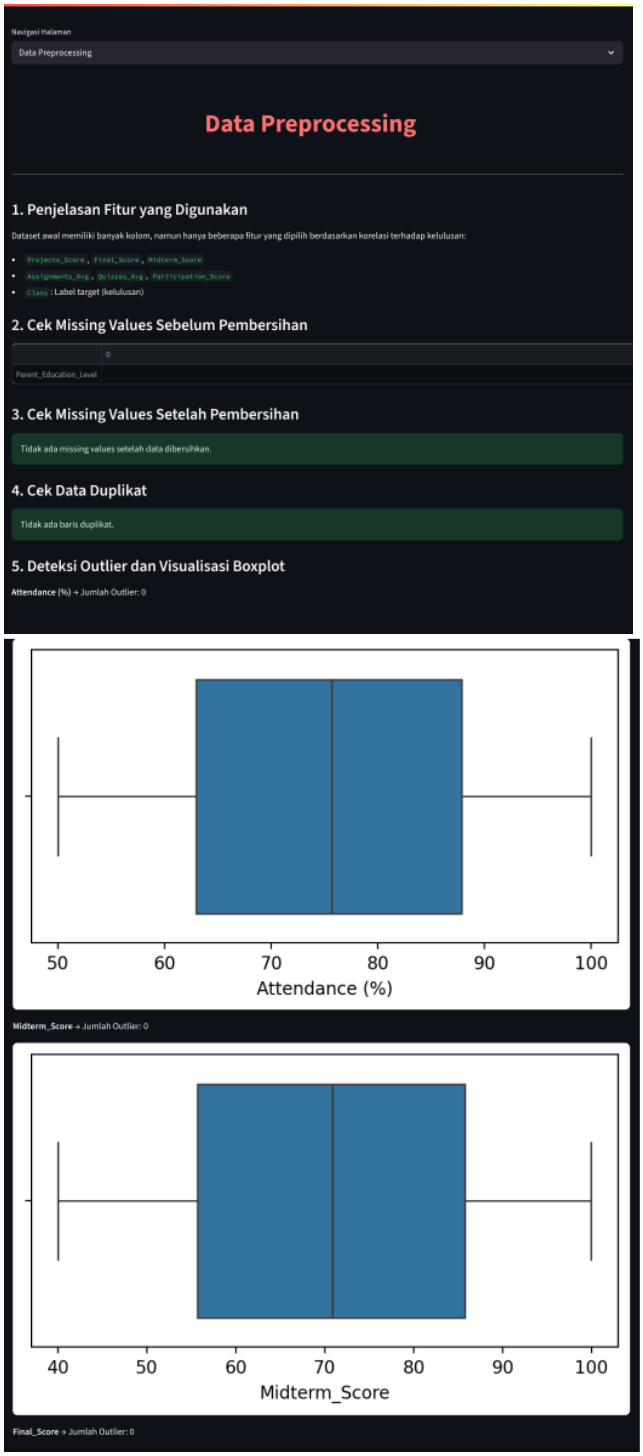


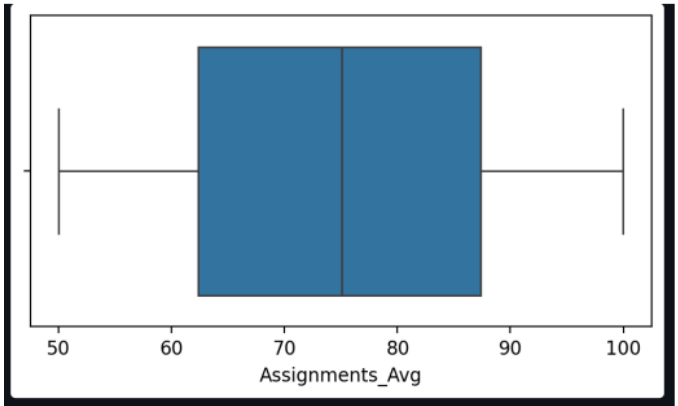


8. Pairplot Hubungan Antar Variabel

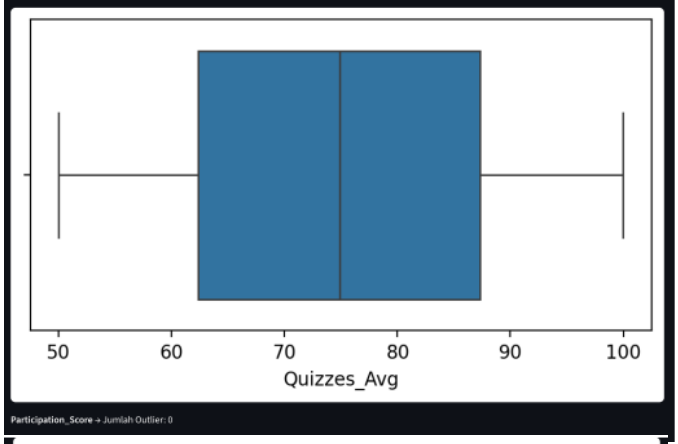


3.2.5 Data Preparation

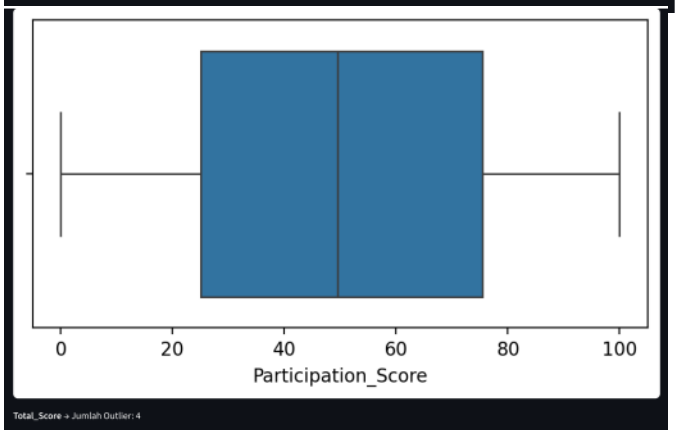




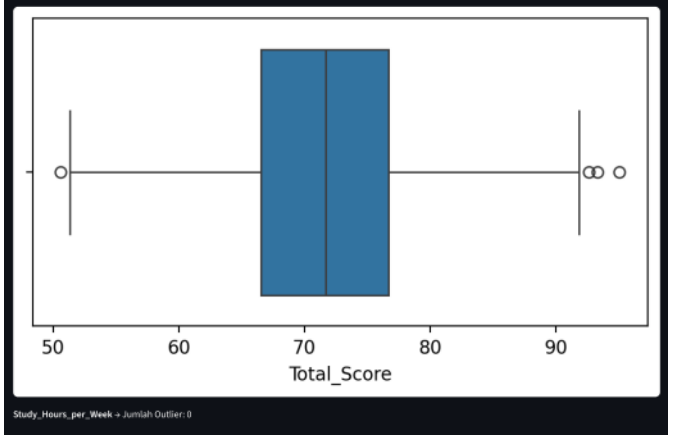
Quizzes_Avg → Jumlah Outlier: 0



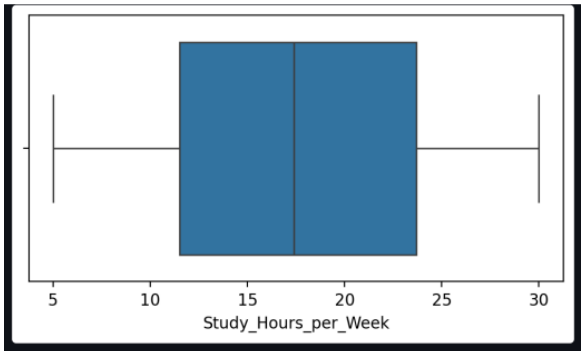
Participation_Score → Jumlah Outlier: 0



Total_Score → Jumlah Outlier: 4

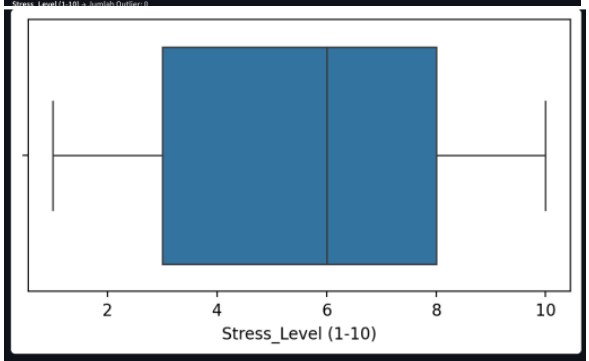


Study_Hours_per_Week → Jumlah Outlier: 0



Sleep_Hours_per_Night → Jumlah Outlier: 0

Stress_Level (1-10) → Jumlah Outlier: 0



6. Encode Target `Class` (Pass/Fail)

Label kelulusan `Class` telah diubah dari teks (`Pass` / `Fail`) menjadi angka (`1` / `0`).

	Class	Jumlah
0	Pass	
1	Fail	

Contoh encode:

	Projects_Score	Final_Score	Midterm_Score	C
0	0.2568	0.3268	0.0102	
1	0.9646	0.5668	0.2879	
2	0.8244	0.3975	0.0307	
3	0.1096	0.0739	0.0942	
4	0.7486	0.3628	0.2189	

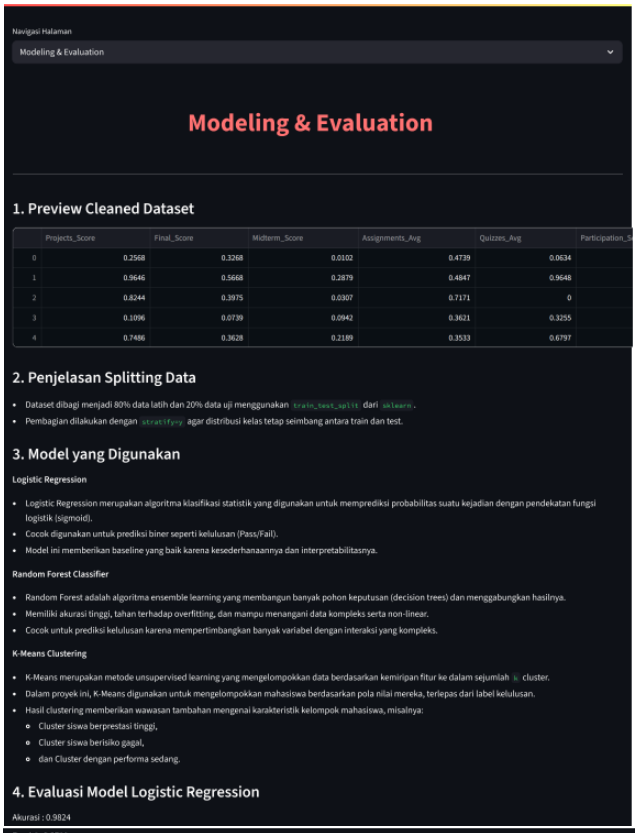
7. Hasil Akhir Setelah Scaling

Dataset hasil preprocessing telah discaling menggunakan MinMaxScaler dan disimpan sebagai `cleaned_data.csv`.

	Projects_Score	Final_Score	Midterm_Score	Assignments_Avg	Quizzes_Avg	Participation_S
0	0.2568	0.3268	0.0102	0.4739	0.0634	
1	0.9646	0.5668	0.2879	0.4847	0.9648	
2	0.8244	0.3975	0.0307	0.7171	0	
3	0.1096	0.0739	0.0942	0.3621	0.3255	
4	0.7486	0.3628	0.2189	0.3533	0.6797	

Data sudah siap digunakan untuk pemodelan.

3.2.6 Modelling & Evaluation

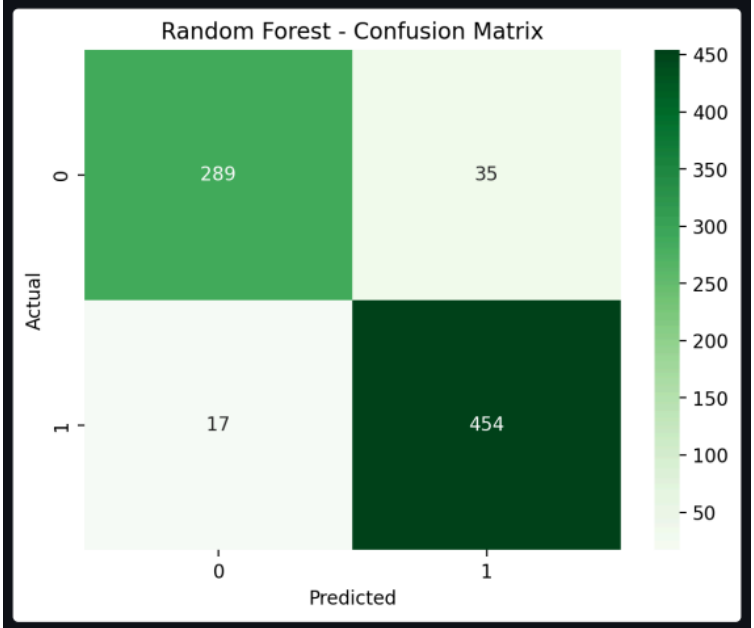


F1-score : 0.9458

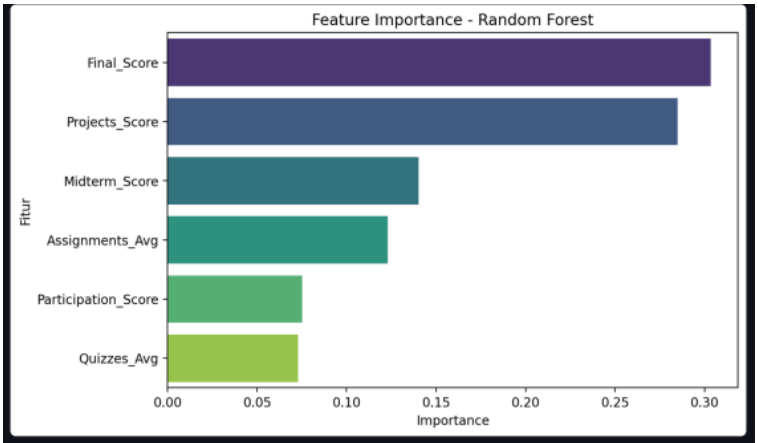
Classification Report:

	precision	recall	f1-score	support
0	0.94	0.89	0.92	324
1	0.93	0.96	0.95	471
accuracy			0.93	795
macro avg	0.94	0.93	0.93	795
weighted avg	0.93	0.93	0.93	795

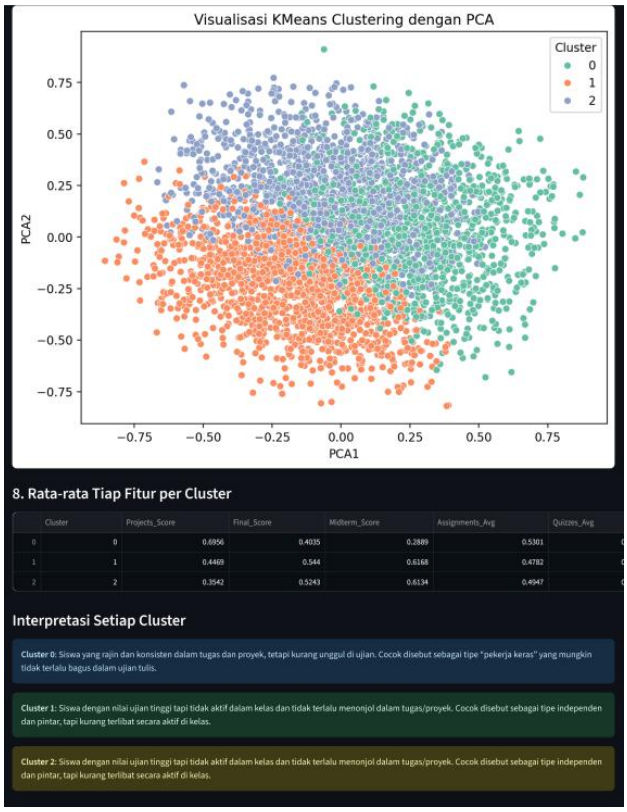
Confusion Matrix:



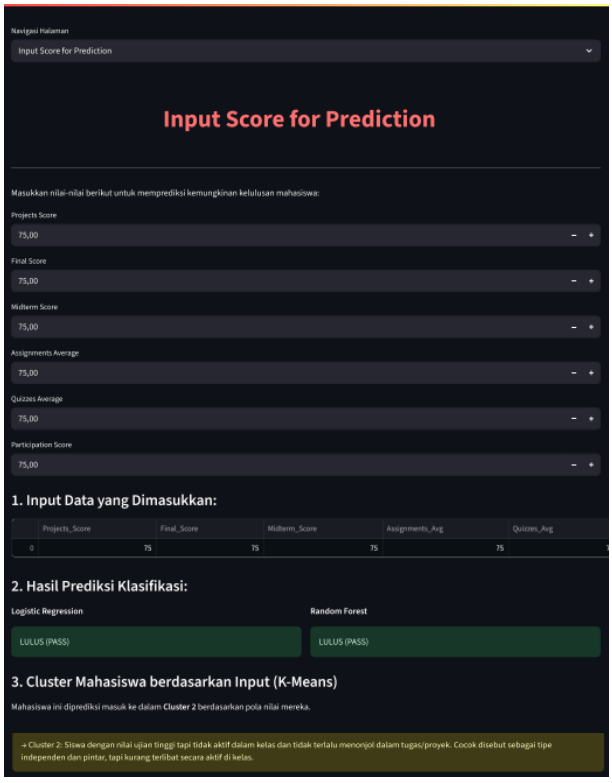
6. Visualisasi Feature Importance (Random Forest)



7. K-Means Clustering (Visualisasi dan Ringkasan)



3.2.7 Input Score for Prediction



BAB IV

PENUTUP

4.1 Kesimpulan

Proyek ini bertujuan untuk memprediksi kelulusan mahasiswa berdasarkan data akademik, perilaku, dan partisipasi mereka dalam proses pembelajaran. Dataset yang digunakan adalah *Students Grading Dataset* dari Kaggle, yang telah melalui tahapan preprocessing seperti pembersihan data, penghapusan fitur yang menyebabkan data leakage, transformasi data, feature selection, dan scaling.

Dalam proses klasifikasi, dua algoritma supervised learning yang digunakan adalah Logistic Regression dan Random Forest Classifier. Berdasarkan evaluasi performa, kedua model menunjukkan hasil yang baik dengan akurasi tinggi, precision, recall, dan F1-score yang seimbang. Model Random Forest secara khusus menunjukkan hasil yang lebih stabil dan performa lebih tinggi dibandingkan Logistic Regression.

Selain klasifikasi, analisis unsupervised learning dilakukan menggunakan K-Means Clustering. Model ini digunakan untuk mengelompokkan mahasiswa ke dalam beberapa segmen berdasarkan kemiripan karakteristik nilai mereka. Visualisasi hasil klaster menggunakan PCA menunjukkan bahwa tiga klaster utama dapat mengelompokkan mahasiswa berdasarkan kecenderungan prestasi, seperti:

1) Cluster 0: Siswa yang unggul dalam tugas proyek, kuis, dan penugasan, tetapi nilai ujian (Final & Midterm) mereka tergolong rendah. Bisa jadi tipe mahasiswa yang aktif dan rajin, namun kurang maksimal dalam ujian tertulis.",
2) Cluster 1: Siswa dengan performa kuat di ujian tertulis (midterm & final), namun tidak terlalu aktif berpartisipasi dan hanya sedang dalam proyek atau kuis. Bisa jadi tipe yang pintar tapi pasif."

3) Cluster 2: Siswa yang sangat aktif berpartisipasi, tapi skor tugas dan proyek cenderung rendah. Mungkin tipe mahasiswa yang aktif di kelas tapi kesulitan saat pengerjaan tugas individu."

Nilai Silhouette Score sebesar 0.6157 menunjukkan kualitas clustering yang cukup baik, dengan pemisahan antar klaster yang memadai. Seluruh hasil analisis dikemas ke dalam sebuah dashboard berbasis web interaktif menggunakan Streamlit, yang memungkinkan pengguna (misalnya, dosen atau staf akademik) untuk:

- Memasukkan nilai mahasiswa dan melihat prediksi kelulusan secara langsung,
- Melihat posisi mahasiswa dalam kelompok klaster berdasarkan pola nilai.

  Proyek ini menunjukkan bahwa pendekatan data mining dan machine learning dapat secara efektif digunakan untuk mendukung proses evaluasi kelulusan mahasiswa secara objektif dan berbasis data.

4.2 Saran

1) Peningkatan Model Klasifikasi:

- Evaluasi Algoritma Tambahan: Meskipun Random Forest telah menunjukkan performa tinggi, eksplorasi algoritma lain seperti Gradient Boosting, XGBoost, atau bahkan pendekatan deep learning dapat diuji untuk mendapatkan hasil yang lebih optimal.

- Feature Engineering Lanjutan: Menambahkan fitur baru yang relevan, seperti pola waktu belajar atau interaksi sosial, bisa meningkatkan kemampuan model dalam mengenali pola yang lebih kompleks.

- Validasi Lintas Dataset: Untuk mengukur generalisasi model, disarankan menggunakan dataset dari institusi pendidikan lain.

2) Pengembangan Implementasi Aplikasi:

- Prediksi Real-Time: Dashboard dapat dikembangkan lebih lanjut untuk menerima input nilai langsung dari sistem akademik (seperti LMS), sehingga prediksi kelulusan dapat dilakukan secara otomatis.
- Integrasi Sistem Akademik: Sistem prediksi dapat diintegrasikan ke dalam portal akademik untuk memberikan rekomendasi dukungan akademik kepada mahasiswa yang berisiko tidak lulus.

3) Evaluasi dan Peningkatan Clustering:

- Uji Metode Clustering Lain: K-Means bekerja baik untuk data terpisah linier. Untuk pola yang lebih kompleks, metode seperti DBSCAN atau Hierarchical Clustering dapat dicoba.
- Penambahan Fitur Non-Nilai: Klasterisasi dapat lebih bermakna jika juga mencakup fitur seperti kehadiran, partisipasi ekstrakurikuler, atau indikator psikologis (motivasi, stres).

4) Penanganan Data:

- Pertimbangan Data Imbang: Meskipun balancing seperti SMOTE tidak digunakan karena data relatif seimbang, untuk kasus nyata yang tidak seimbang, metode seperti class weighting atau ensemble khusus bisa diimplementasikan.
- Data Real Time dan Historis: Menggabungkan data historis dan data terkini dapat memperkuat akurasi model dalam memahami tren mahasiswa dari waktu ke waktu.

DAFTAR PUSAKA

Tinto, V. (2012). Completing college: Rethinking institutional action. University of Chicago Press.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 30-40.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355.

Cao, L. (2017). Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), 1-42.