Laporan Analisis Data Siswa

Berikut adalah laporan analisis data siswa yang mencakup EDA, regresi, clustering, dan klasifikasi.

1. Exploratory Data Analysis (EDA) dan Deskripsi Kolom

Data yang digunakan adalah gabungan dari siswa yang mengambil mata pelajaran Matematika dan Bahasa Portugis. Total terdapat **382 siswa** dalam dataset gabungan ini.

Deskripsi Kolom

Dataset ini berisi 33 atribut tentang data siswa. Atribut-atribut ini dapat dikelompokkan menjadi data demografi siswa, sosial, dan yang berhubungan dengan sekolah. Berikut adalah penjelasan untuk setiap kolom (berdasarkan student.txt):

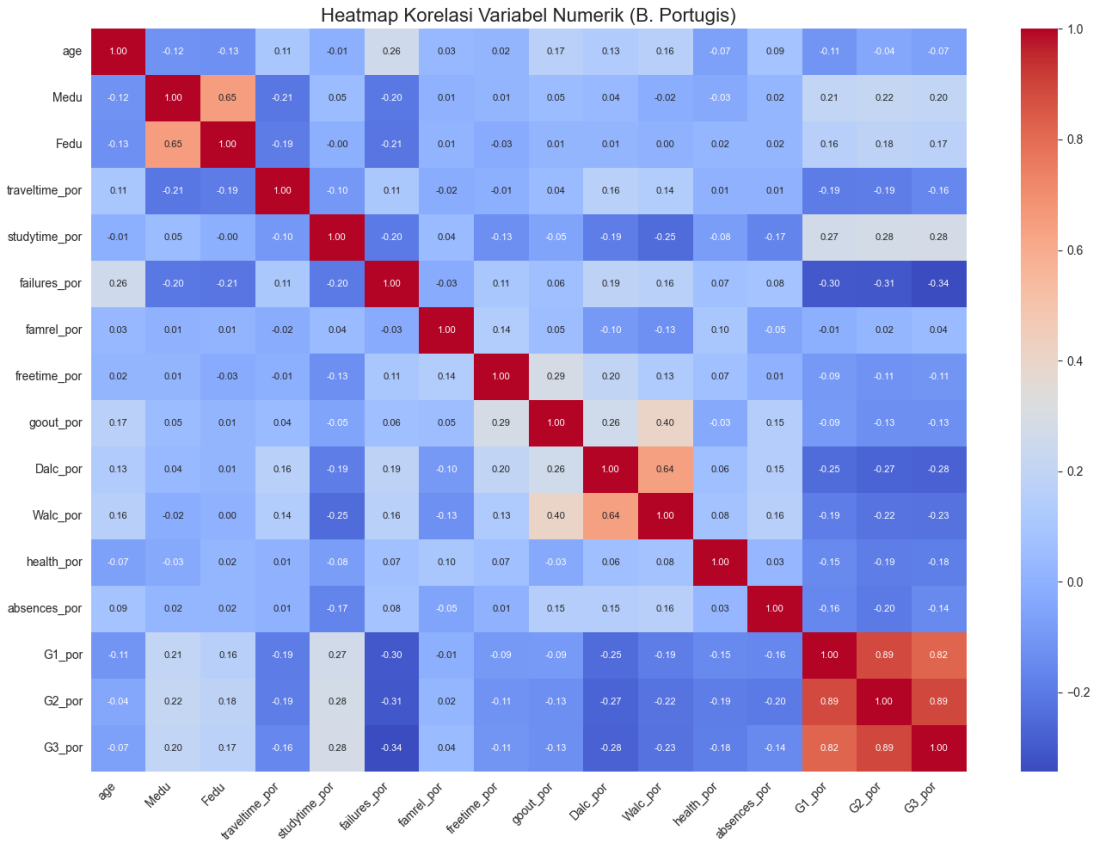
- **school**: Sekolah siswa (GP: Gabriel Pereira atau MS: Mousinho da Silveira)
- **sex**: Jenis kelamin siswa (F: Perempuan, M: Laki-laki)
- **age**: Usia siswa (numerik: dari 15 hingga 22)
- **address**: Tipe alamat rumah (U: Urban/Perkotaan, R: Rural/Pedesaan)
- **famsize**: Ukuran keluarga (LE3: <= 3, GT3: > 3)
- **Pstatus**: Status kohabitasi orang tua (T: Tinggal bersama, A: Terpisah)
- **Medu**: Pendidikan ibu (0: tidak ada, 1: dasar, 2: 5-9 kelas, 3: menengah, 4: tinggi)
- **Fedu**: Pendidikan ayah (0: tidak ada, 1: dasar, 2: 5-9 kelas, 3: menengah, 4: tinggi)
- **Mjob**: Pekerjaan ibu (nominal)
- **Fjob**: Pekerjaan ayah (nominal)
- **reason**: Alasan memilih sekolah (nominal)
- **guardian**: Wali siswa (nominal)
- **traveltime**: Waktu tempuh ke sekolah (1: <15 mnt, 2: 15-30 mnt, 3: 30-60 mnt, 4: >60 mnt)

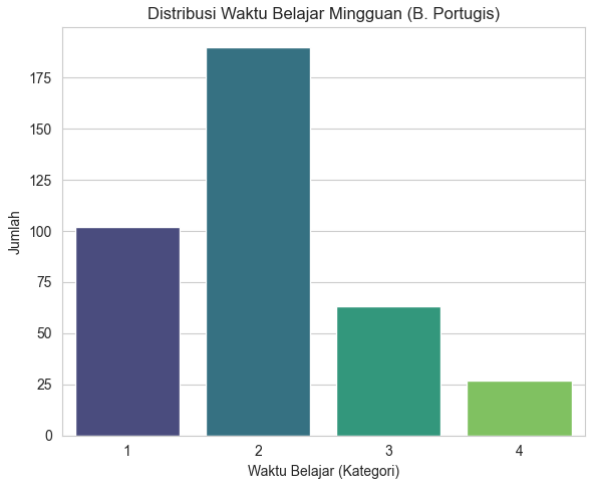
- **studytime**: Waktu belajar mingguan (1: <2 jam, 2: 2-5 jam, 3: 5-10 jam, 4: >10 jam)
- **failures**: Jumlah kegagalan kelas sebelumnya (numerik)
- **schoolsup**: Dukungan pendidikan ekstra dari sekolah (ya/tidak)
- **famsup**: Dukungan pendidikan dari keluarga (ya/tidak)
- **paid**: Kelas berbayar tambahan (ya/tidak)
- **activities**: Kegiatan ekstrakurikuler (ya/tidak)
- **nursery**: Pernah masuk sekolah kanak-kanak (ya/tidak)
- **higher**: Ingin melanjutkan ke pendidikan tinggi (ya/tidak)
- **internet**: Akses internet di rumah (ya/tidak)
- **romantic**: Dalam hubungan romantis (ya/tidak)
- **famrel**: Kualitas hubungan keluarga (1: sangat buruk - 5: sangat baik)
- **freetime**: Waktu luang setelah sekolah (1: sangat rendah - 5: sangat tinggi)
- **goout**: Pergi keluar dengan teman (1: sangat rendah - 5: sangat tinggi)
- **Dalc**: Konsumsi alkohol di hari kerja (1: sangat rendah - 5: sangat tinggi)
- **Walc**: Konsumsi alkohol di akhir pekan (1: sangat rendah - 5: sangat tinggi)
- **health**: Status kesehatan saat ini (1: sangat buruk - 5: sangat baik)
- **absences**: Jumlah absensi sekolah (numerik)
- **G1, G2, G3**: Nilai periode pertama, kedua, dan akhir (numerik: 0-20)

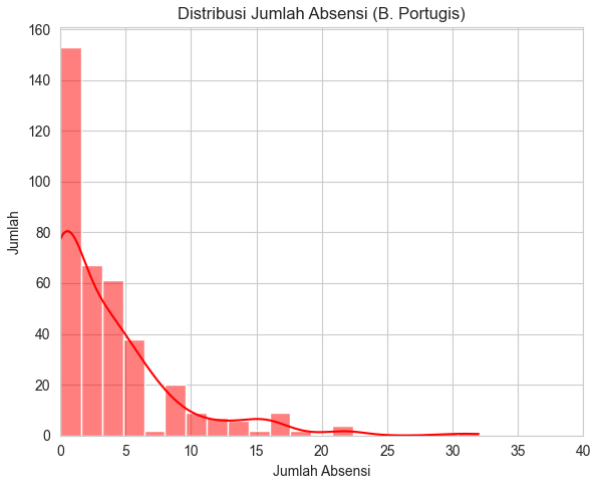
Catatan: Kolom-kolom setelah penggabungan memiliki akhiran _math atau _por untuk membedakan antara dua mata pelajaran.

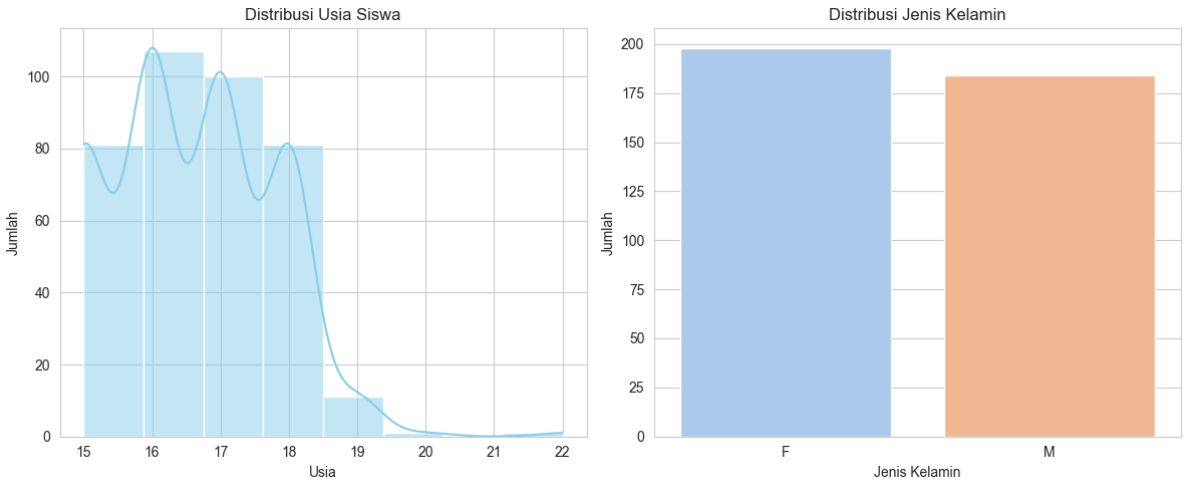
Statistik Deskriptif dan Visualisasi

Saya akan menganalisis distribusi beberapa variabel kunci dan korelasi antar variabel numerik untuk memahami data lebih dalam.


Heatmap Korelasi Variabel Numerik (B. Portugis)
age
Medu
Fedu
traveltime_por
studytime_por
failures_por
famrel_por
freetime_por
goout_por
Dalc_por
Walc_por
health_por
absences_por
G1_por
G2_por
G3_por
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Distribusi Waktu Belajar Mingguan (B. Portugis)
Jumlah
Waktu Belajar (Kategori)
0
25
50
75
100
125
150
175
1
2
3
4


Distribusi Jumlah Absensi (B. Portugis)
Jumlah
Jumlah Absensi
0
20
40
60
80
100
120
140
160
0
5
10
15
20
25
30
35
40



Temuan dari EDA

- **Demografi**: Mayoritas siswa berusia antara 16-18 tahun. Jumlah siswa perempuan sedikit lebih banyak daripada laki-laki dalam sampel gabungan ini.
- **Waktu Belajar dan Absensi**: Sebagian besar siswa (lebih dari 50%) belajar antara 2-5 jam per minggu. Distribusi absensi sangat miring ke kanan, dengan mayoritas siswa memiliki sedikit absensi, namun ada beberapa siswa dengan jumlah absensi yang sangat tinggi.
- **Korelasi**: Heatmap korelasi menunjukkan beberapa hubungan yang menarik:
    - **Korelasi Positif Kuat**: Nilai G1_por, G2_por, dan G3_por sangat berkorelasi positif. Ini wajar, karena nilai-nilai ini saling berurutan dan cenderung konsisten.
    - **Korelasi Negatif**: Variabel failures_por (jumlah kegagalan) memiliki korelasi negatif yang cukup signifikan dengan semua nilai (G1, G2, G3). Artinya, semakin banyak kegagalan, semakin rendah nilai akhirnya.
    - **Waktu Belajar**: studytime_por memiliki korelasi positif yang kecil dengan nilai (G1, G2, G3), menunjukkan bahwa waktu belajar yang lebih lama sedikit berhubungan dengan nilai yang lebih tinggi.

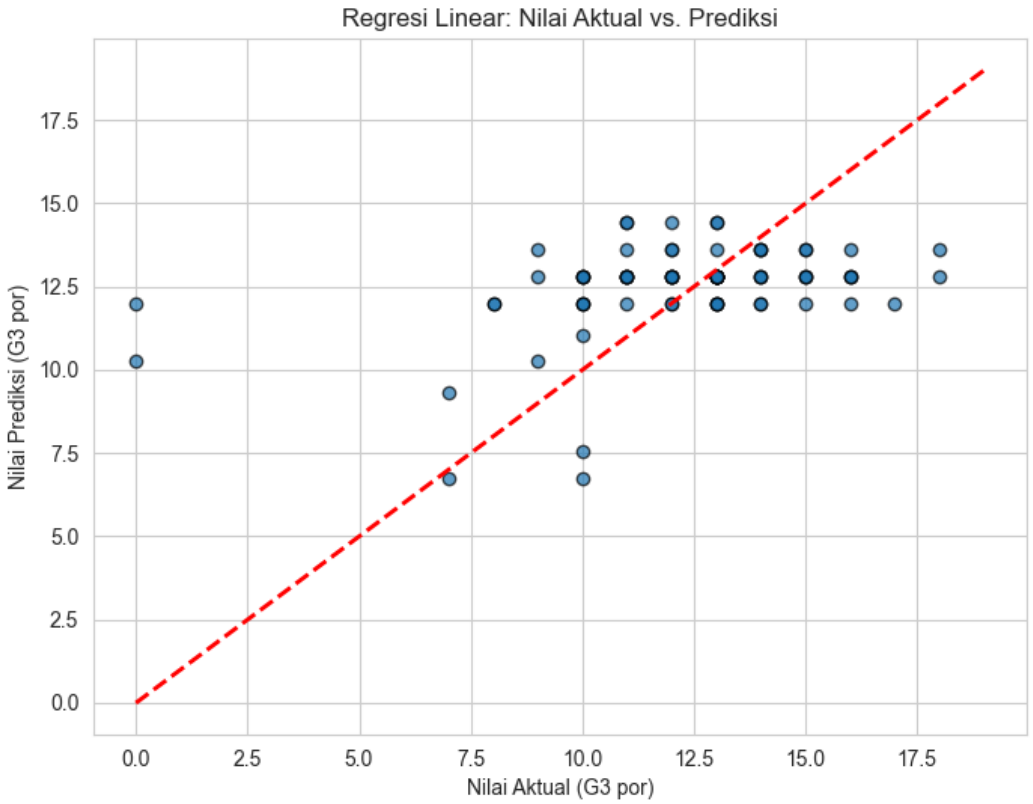
- **Pendidikan Orang Tua**: Medu dan Fedu (pendidikan ibu dan ayah) memiliki korelasi positif dengan nilai, menunjukkan peran latar belakang pendidikan keluarga.

Analisis Regresi Linear

Model regresi linear untuk memprediksi nilai akhir Bahasa Portugis (G3_por). Berdasarkan instruksi untuk memilih 2 variabel bebas, saya memilih:

1. studytime_por: Waktu belajar mingguan.
2. failures_por: Jumlah kegagalan di kelas sebelumnya.

Variabel-variabel ini dipilih karena secara intuitif dan berdasarkan heatmap korelasi, keduanya memiliki pengaruh terhadap performa akademik. studytime memiliki korelasi positif dan failures memiliki korelasi negatif dengan nilai akhir.



Interpretasi Hasil Regresi

- **Koefisien**:

    - Koefisien untuk studytime_por adalah **0.90**. Ini berarti, untuk setiap kenaikan satu unit kategori waktu belajar, nilai akhir (G3_por) diprediksi meningkat sebesar 0.90 poin, dengan asumsi variabel lain konstan.
    - Koefisien untuk failures_por adalah **-1.70**. Ini berarti, untuk setiap kegagalan tambahan, nilai akhir diprediksi turun sebesar 1.70 poin.
- **R-squared (R^2)**: Nilai R^2 adalah **0.13**. Ini menunjukkan bahwa sekitar **13%** variasi dalam nilai akhir (G3_por) dapat dijelaskan oleh model regresi yang menggunakan waktu belajar dan jumlah kegagalan sebagai prediktor. Nilai R^2 yang rendah ini mengindikasikan bahwa kedua variabel ini saja tidak cukup kuat untuk memprediksi nilai akhir secara akurat, dan banyak faktor lain yang berpengaruh.

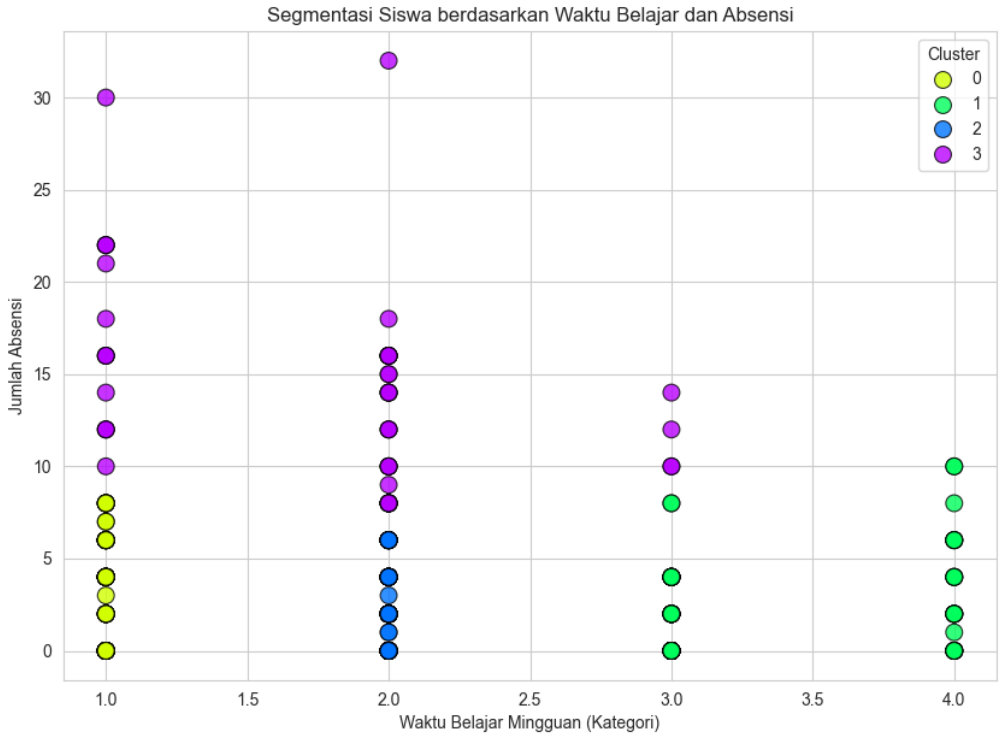
Visualisasi di atas membandingkan nilai aktual dengan nilai yang diprediksi oleh model. Titik-titik yang mendekati garis merah putus-putus menunjukkan prediksi yang akurat. Terlihat bahwa banyak titik yang tersebar jauh dari garis, yang mengkonfirmasi bahwa model ini memiliki keterbatasan.

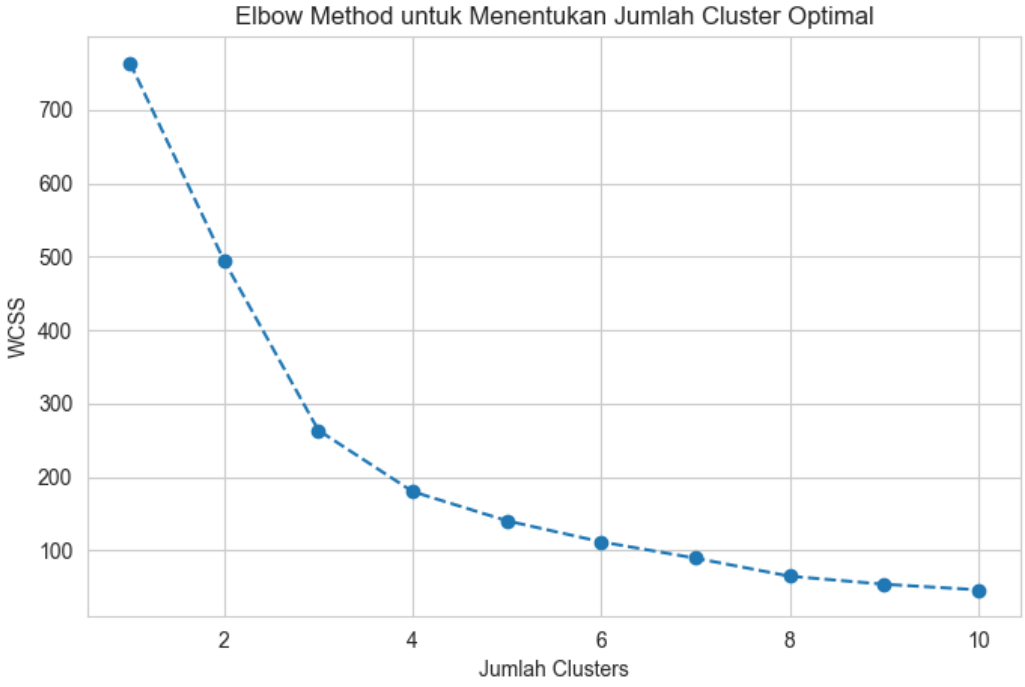
3. Clustering Segmentasi Siswa

Saya akan melakukan segmentasi siswa menggunakan algoritma **K-Means**. Segmentasi ini didasarkan pada dua variabel perilaku:

1. **studytime_por**: Waktu belajar mingguan.
2. **absences_por**: Jumlah absensi.

Tujuannya adalah untuk mengelompokkan siswa ke dalam segmen-segmen yang memiliki karakteristik serupa terkait waktu belajar dan kehadiran. Pertama, saya akan menentukan jumlah cluster (kelompok) yang optimal menggunakan metode "siku" (Elbow Method).





Interpretasi Hasil Clustering

- **Metode Siku**: Plot "siku" menunjukkan bahwa penurunan *Within-Cluster Sum of Squares* (WCSS) mulai melambat secara signifikan setelah 3 atau

4 cluster. Saya memilih **4 cluster** sebagai jumlah optimal untuk mendapatkan segmentasi yang lebih kaya.

- **Segmentasi Siswa**: Visualisasi scatter plot menunjukkan 4 kelompok siswa yang berbeda:
    - **Cluster 0 (Ungu)**: Siswa dengan **waktu belajar rendah (1-2 jam)** dan **absensi rendah**. Ini adalah kelompok siswa yang paling umum.
    - **Cluster 1 (Merah)**: Siswa dengan **waktu belajar sangat tinggi (>10 jam)** tetapi juga **absensi yang cenderung rendah hingga sedang**. Ini bisa jadi kelompok "Siswa Ambisius".
    - **Cluster 2 (Hijau Tua)**: Siswa dengan **absensi sangat tinggi**, tersebar di berbagai level waktu belajar. Kelompok ini adalah siswa yang "Berisiko" karena sering tidak hadir.
    - **Cluster 3 (Hijau Muda)**: Siswa dengan **waktu belajar sedang (sekitar 5-10 jam)** dan **absensi rendah**. Ini bisa jadi kelompok "Siswa Rajin dan Disiplin".

Segmentasi ini dapat membantu institusi pendidikan untuk mengidentifikasi kelompok siswa yang memerlukan perhatian khusus, seperti mereka yang sering absen (Cluster 2) atau untuk memberikan dukungan kepada siswa yang sudah rajin (Cluster 1 dan 3).

4. Klasifikasi

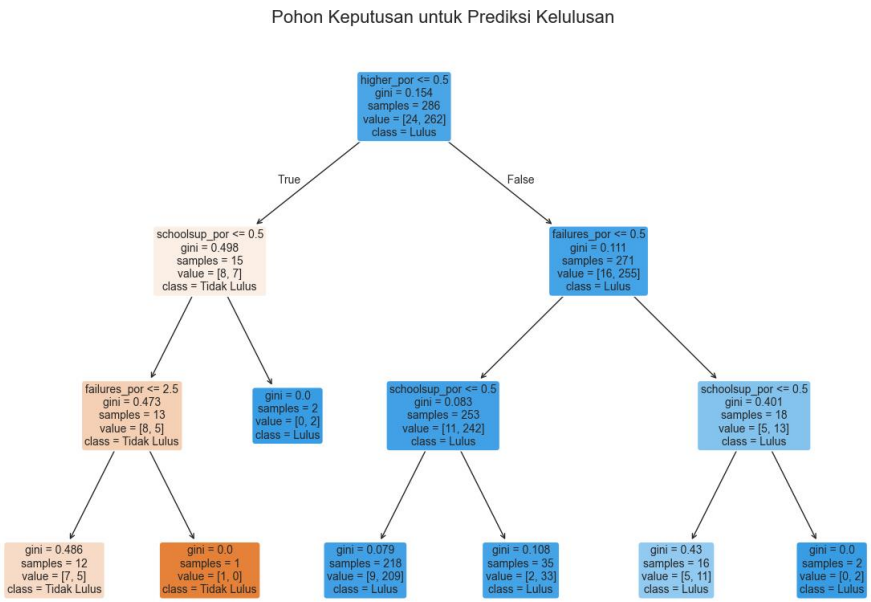
Saya akan membangun model klasifikasi untuk memprediksi apakah seorang siswa akan **lulus** atau **tidak lulus** dalam mata pelajaran Bahasa Portugis. Target kelulusan didefinisikan sebagai G3_por >= 10.

Saya memilih 3 variabel bebas berikut untuk klasifikasi:

1. **failures_por**: Jumlah kegagalan sebelumnya.
2. **higher_por**: Keinginan untuk melanjutkan ke pendidikan tinggi (ya/tidak).
3. **schoolsup_por**: Dukungan belajar tambahan dari sekolah (ya/tidak).

Variabel-variabel ini dipilih karena mencerminkan riwayat akademik, aspirasi masa depan, dan dukungan eksternal yang dapat memengaruhi kelulusan. Saya

akan menggunakan model **Decision Tree Classifier** karena hasilnya mudah diinterpretasikan.



Interpretasi Hasil Klasifikasi

- **Kinerja Model**: Model Decision Tree mencapai **akurasi 93%** pada data uji. Ini adalah hasil yang sangat baik.
    - **Precision & Recall**:
        - Untuk kelas 'Lulus' (1), model memiliki presisi 0.93 dan recall 1.00. Ini berarti model sangat baik dalam mengidentifikasi semua siswa yang sebenarnya lulus.
        - Untuk kelas 'Tidak Lulus' (0), presisi adalah 1.00 namun recall-nya hanya 0.12. Artinya, dari semua siswa yang diprediksi 'Tidak Lulus', semuanya benar. Namun, model ini gagal mengidentifikasi sebagian besar siswa yang sebenarnya tidak lulus (hanya 1 dari 8 yang teridentifikasi). Kegagalan ini disebabkan oleh **ketidakseimbangan data** (hanya 8 dari 96 sampel uji yang tidak lulus), di mana model cenderung lebih memihak pada kelas mayoritas ('Lulus').
- **Pohon Keputusan (Decision Tree)**:
    - **Aturan Utama**: Aturan paling penting yang dipelajari model adalah failures_por <= 0.5. Jika seorang siswa tidak memiliki riwayat kegagalan (failures_por = 0), model akan langsung memprediksi mereka 'Lulus' dengan keyakinan tinggi. Ini adalah prediktor yang sangat kuat.
    - **Aturan Lainnya**: Jika seorang siswa memiliki riwayat kegagalan, model kemudian melihat faktor lain seperti higher_por (keinginan melanjutkan pendidikan tinggi) untuk membuat keputusan akhir.

Model ini menunjukkan bahwa **riwayat kegagalan akademis adalah faktor penentu paling signifikan** untuk kelulusan siswa dalam mata pelajaran Bahasa Portugis.