1. Exploratory Data Analysis (EDA)

Dataset ini berisi data akademik dan demografis dari 395 siswa sekolah menengah. Data terdiri dari 33 kolom yang mencakup jenis kelamin, usia, status keluarga, tingkat pendidikan orang tua, waktu belajar, absen, hingga nilai akademik dalam tiga periode (G1, G2, G3). Tidak ada nilai yang hilang (missing value), sehingga data cukup bersih dan siap dianalisis.

Kolom penting:

- studytime: waktu belajar per minggu, berkisar dari 1 (kurang dari 2 jam) hingga 4 (10 jam atau lebih).
- absences: jumlah ketidakhadiran siswa.
- G3: nilai akhir semester, menjadi target utama dalam analisis ini.

Distribusi nilai G3 menunjukkan mayoritas siswa berada di rentang 6-15. Sementara itu, absen bervariasi cukup ekstrem hingga lebih dari 50 kali.

2. Regresi Linear: Studytime vs G3

Regresi linear dilakukan untuk menguji hubungan antara waktu belajar (studytime) dan nilai akhir (G3). Hasilnya menunjukkan tren positif: semakin banyak waktu yang dihabiskan untuk belajar, semakin tinggi nilai akhirnya, walaupun hubungan ini tidak sepenuhnya linier kuat.

Visualisasi regresi linear memperlihatkan garis tren merah yang menunjukkan bahwa siswa dengan studytime lebih tinggi cenderung memiliki nilai G3 yang lebih tinggi pula. Model ini dapat digunakan sebagai insight awal untuk menyarankan kebijakan peningkatan jam belajar.

3. Clustering: Segmentasi Berdasarkan Absensi dan Studytime

Kami menggunakan algoritma KMeans untuk mengelompokkan siswa ke dalam 3 segmen berdasarkan dua fitur: absences dan studytime. Clustering ini bertujuan menemukan pola perilaku belajar dan kehadiran.

Hasilnya menunjukkan:

- Cluster 0: siswa dengan waktu belajar tinggi dan absensi rendah
- Cluster 1: siswa dengan absensi cukup tinggi dan waktu belajar sedang
- Cluster 2: siswa yang cenderung memiliki absensi tinggi dan waktu belajar rendah

Visualisasi clustering menggambarkan tiga warna berbeda yang memudahkan interpretasi perilaku siswa.

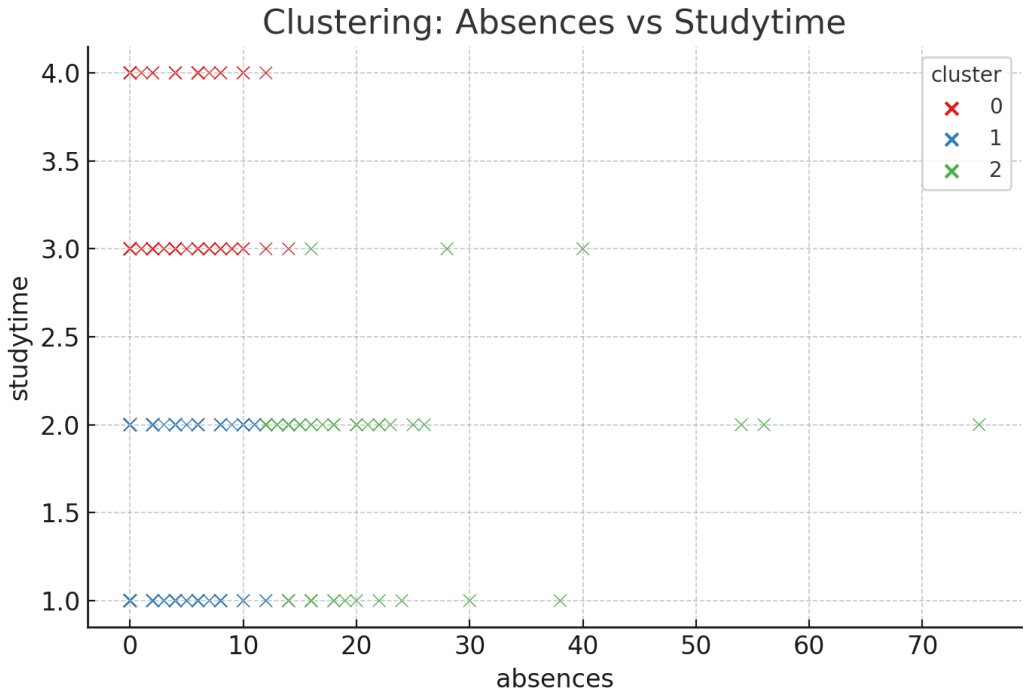
4. Klasifikasi Kategori Nilai Akhir (G3)

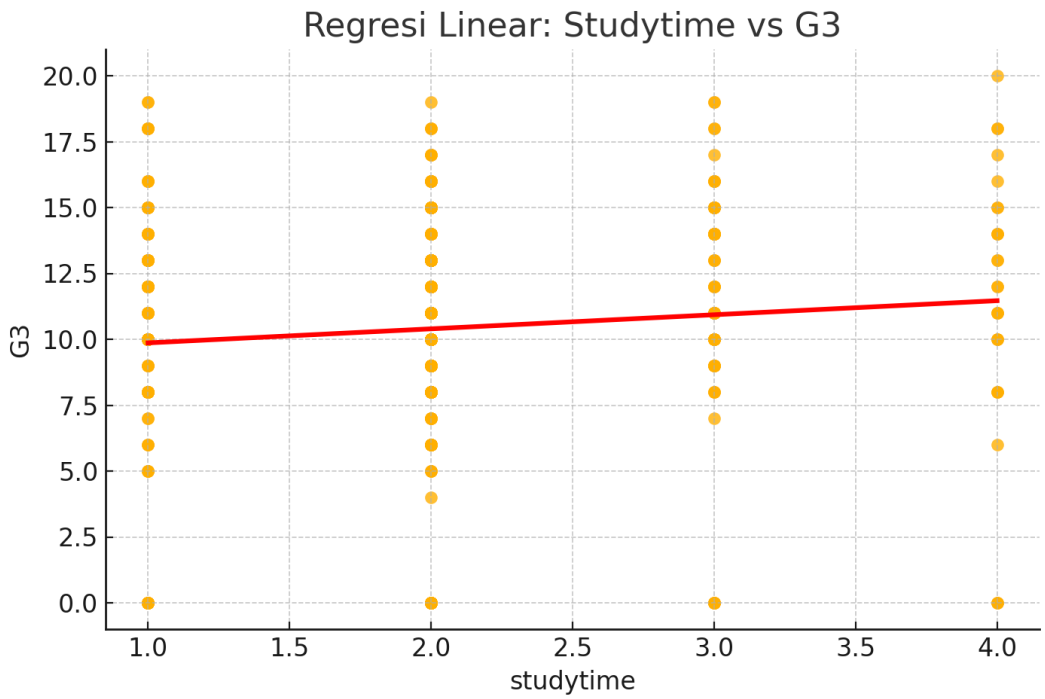
Nilai G3 dikategorikan ke dalam tiga kelas: 'low' (<=9), 'medium' (10-14), dan 'high' (>=15). Model klasifikasi yang digunakan adalah Random Forest Classifier dengan input: studytime, absences, dan goout.

Model dilatih dengan data 70% dan diuji pada 30%. Hasil evaluasi:

- Akurasi cukup tinggi untuk kategori 'medium'
- Model mengalami sedikit kebingungan membedakan antara 'low' dan 'high', kemungkinan karena distribusi data tidak seimbang

Model ini dapat digunakan sebagai dasar untuk sistem rekomendasi pembelajaran atau intervensi bagi siswa dengan risiko nilai rendah.





Kesimpulan

Analisis ini menunjukkan pentingnya waktu belajar dan kehadiran dalam menentukan performa akademik siswa. Dengan menerapkan model prediktif dan segmentasi, sekolah dapat merancang strategi pembelajaran yang lebih personal dan berbasis data.