

Judul : Analisis dan Prediksi Faktor Penentu Prestasi Akademik Siswa Menggunakan Metode Regresi, Klasifikasi, dan Clustering

Anggota Kelompok : 1. Deni Permana (2210512015) (DS-B)
2. Diaz Fahreza Akbar (2210512069) (DS-B)

1. Why? (Business Understanding & Analytic Approach)

Deskripsi Masalah:

Prestasi akademik siswa merupakan salah satu indikator utama keberhasilan pendidikan. Namun, banyak faktor yang memengaruhi hasil belajar, mulai dari aspek akademik, sosial, hingga personal. Dengan meningkatnya tekanan untuk mencapai target prestasi, penting untuk memahami pola dan faktor yang paling signifikan dalam menentukan nilai akhir siswa.

Task yang Diusulkan:

- Regresi: Memodelkan hubungan antara faktor-faktor seperti waktu belajar, dukungan keluarga, dan tingkat kehadiran terhadap nilai akhir siswa (G3).
- Klasifikasi: Mengelompokkan siswa ke dalam kategori prestasi (rendah, sedang, tinggi) berdasarkan nilai akhir.
- Clustering: Menemukan pola tersembunyi dalam data untuk mengidentifikasi kelompok siswa berdasarkan perilaku belajar dan sosial.

2. Data Requirement & Data Collection

Sumber Data :

Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository.
<https://doi.org/10.24432/C5TG7T>.

Jumlah Data:

- student-mat.csv: 395 baris
- student-por.csv: 649 baris

Deskripsi Data

- Atribut Demografis:
 - o school: Sekolah siswa (GP atau MS).
 - o sex: Jenis kelamin siswa (F atau M).
 - o age: Usia siswa (15-22 tahun).
 - o address: Jenis tempat tinggal (U - urban, R - rural).
 - o famsize: Ukuran keluarga (LE3 - ≤3 anggota, GT3 - >3 anggota).
 - o Pstatus: Status hubungan orang tua (T - bersama, A - berpisah).
- Atribut Sosial-Ekonomi:
 - o Medu & Fedu: Tingkat pendidikan ibu dan ayah (0 - tidak ada hingga 4 - pendidikan tinggi).
 - o Mjob & Fjob: Pekerjaan ibu dan ayah (e.g., teacher, health, at_home).
 - o reason: Alasan memilih sekolah (home, reputation, course, atau other).
 - o guardian: Wali siswa (mother, father, atau other).

- Atribut Akademik:
 - studytime: Waktu belajar mingguan (1-4: <2 jam hingga >10 jam).
 - failures: Jumlah kegagalan kelas sebelumnya (0-4).
 - schoolsup & famsup: Dukungan belajar dari sekolah/keluarga (yes atau no).
 - paid: Kelas tambahan berbayar dalam mata pelajaran tersebut (yes atau no).
- Atribut Kehidupan Pribadi & Sosial:
 - freetime, goout: Waktu luang dan frekuensi keluar bersama teman (1-5).
 - Dalc, Walc: Konsumsi alkohol pada hari kerja/akhir pekan (1-5).
 - romantic: Dalam hubungan romantis (yes atau no).
 - health: Status kesehatan saat ini (1 - buruk, 5 - sangat baik).
 - absences: Jumlah absensi siswa (0-93 hari).
- Atribut Nilai Akademik (Target):
 - G1, G2, G3: Nilai periode pertama, kedua, dan akhir (0-20).

Kolom yang Digunakan:

- Atribut Demografis:
 - sex: Jenis kelamin siswa dapat memengaruhi hasil belajar.
 - age: Usia siswa mungkin relevan terhadap kemampuan belajar.
 - address: Jenis tempat tinggal (urban/rural) bisa menunjukkan akses terhadap sumber belajar.
- Atribut Akademik:
 - studytime: Waktu belajar mingguan sebagai indikator komitmen akademik.
 - failures: Jumlah kegagalan sebelumnya untuk melihat risiko akademik.
 - schoolsup & famsup: Dukungan tambahan yang mungkin berkontribusi terhadap prestasi.
- Atribut Sosial:
 - freetime: Waktu luang dapat memengaruhi produktivitas belajar.
 - goout: Frekuensi keluar bersama teman, faktor potensi distraksi.
 - romantic: Kehidupan romantis mungkin memengaruhi konsentrasi belajar.
- Atribut Nilai (Target):
 - G1, G2: Sebagai prediktor untuk G3.
 - G3: Target utama analisis untuk regresi, klasifikasi, dan evaluasi.

Alasan Pemilihan Kolom:

- Regresi: G3 diprediksi menggunakan variabel yang berhubungan langsung seperti G1, G2, studytime, dan faktor sosial.
- Klasifikasi: Menggunakan atribut seperti failures, studytime, dan romantic untuk menentukan kategori prestasi (rendah, sedang, tinggi).
- Clustering: Variabel seperti freetime, goout, dan studytime digunakan untuk membentuk pola kelompok siswa.

3. Modelling & Evaluation

Pemilihan Model:

- Regresi:
 - o Model: Linear Regression atau Random Forest Regression.
 - o Tujuan: Memprediksi nilai akhir siswa (G3) berdasarkan fitur seperti studytime, absences, dan famrel.

- Klasifikasi:
 - o Model: Decision Tree atau Logistic Regression.
 - o Tujuan: Mengelompokkan siswa ke dalam kategori prestasi ("rendah", "sedang", "tinggi").
 - o Transformasi: Kategori ditentukan dengan threshold nilai G3:

Rendah: $G3 < 10$

Sedang: $10 \leq G3 < 15$

Tinggi: $G3 \geq 15$.

- Clustering:
 - o Model: K-Means atau Hierarchical Clustering.
 - o Tujuan: Mengidentifikasi pola tersembunyi dalam kelompok siswa berdasarkan fitur seperti studytime, freetime, dan famrel.

Evaluasi Model:

- Metrik evaluasi Regresi: Mean Absolute Error (MAE), Root Mean Square Error (RMSE).
- Metrik evaluasi Klasifikasi: Akurasi, Precision, Recall.
- Clustering: Evaluasi menggunakan Elbow Method untuk menentukan jumlah cluster optimal, serta Silhouette Score untuk validasi hasil.

4. Analytics Goal

Tujuan:

- Mengidentifikasi faktor utama yang memengaruhi prestasi akademik siswa.
- Memberikan rekomendasi berbasis data untuk membantu sekolah meningkatkan kualitas pembelajaran.
- Memahami pola kelompok siswa untuk merancang intervensi yang sesuai.

Kebermanfaatan:

- Memberikan insight kepada pendidik mengenai faktor risiko dan potensi keberhasilan siswa.
- Membantu perencanaan strategi peningkatan prestasi berdasarkan hasil prediksi dan segmentasi siswa.

