Ujian Akhir Semester - Data Science

Nama : Romy Ramandika

NIM : 230401010260

Kelas : IF 405

Prodi : PJJ Informatika

Dosen : Alun Sujjada, S.Kom, M.T

1. EDA (Exploratory Data Analysis)

Dataset yang digunakan adalah **Student Performance Data Set** dari UCI, yang terdiri dari dua file: `student-mat.csv` dan `student-por.csv`. Dataset ini berisi informasi siswa seperti data demografis, pendidikan orang tua, kebiasaan belajar, dan nilai akhir.

Fungsi dan Contoh Penjelasan Kolom:

Nama Kolom	Penjelasan
`school`	Nama sekolah (GP atau MS)
`sex`	Jenis kelamin siswa (F = female, M = male)
`age`	Umur siswa
`address`	Alamat tinggal (U = urban, R = rural)
`studytime`	Waktu belajar per minggu (1 = <2 jam, 4 = >10 jam)
`failures`	Jumlah kegagalan sebelumnya
`absences`	Jumlah ketidakhadiran
`G1`, `G2`, `G3`	Nilai ujian tahap 1, 2, dan akhir

Temuan Umum dari EDA:

- Tidak ada missing value.
- Korelasi kuat antara `G1`, `G2`, dan `G3` (nilai bertahap).
- Variabel seperti `studytime` dan `failures` cukup berpengaruh terhadap `G3`.

---

2. Regresi Linear (2 variabel bebas)

Tujuan:

Memprediksi nilai akhir (`G3`) berdasarkan 2 variabel bebas.

Variabel yang dipilih:

- `studytime` (jumlah waktu belajar per minggu)
- `failures` (jumlah kegagalan sebelumnya)

Hasil Regresi:

- Hubungan **positif** antara `studytime` dan `G3`, artinya semakin banyak waktu belajar, semakin tinggi nilai akhir.
- Hubungan **negatif** antara `failures` dan `G3`, artinya semakin sering gagal sebelumnya, semakin rendah nilai akhir.

Visualisasi scatter plot dan garis regresi menunjukkan pola ini dengan jelas.

---

3. Clustering Segmentasi (absensi / waktu belajar)

Tujuan:

Mengelompokkan siswa berdasarkan pola belajar atau ketidakhadiran.

Variabel:

- `absences` (jumlah absen)
- `studytime` (jumlah waktu belajar)

Metode:

Menggunakan **KMeans Clustering** dengan 3 klaster.

Hasil:

- **Cluster 0**: Siswa rajin (absensi rendah, studytime tinggi)
- **Cluster 1**: Siswa tidak aktif (absensi tinggi, studytime rendah)
- **Cluster 2**: Siswa sedang (keseimbangan antara keduanya)

Visualisasi scatter plot menunjukkan pengelompokan yang jelas antar siswa.

---

4. Klasifikasi (3 variabel bebas)

Tujuan:

Memprediksi apakah siswa akan **lulus (G3 ≥ 10)** atau **tidak lulus (G3 < 10)**.

Variabel yang digunakan:

- `studytime`
- `failures`
- `G1` (nilai ujian tahap awal)

Metode:

Menggunakan **Random Forest Classifier**.

Hasil:

- Akurasi tinggi (>80%)
- Precision dan recall baik, terutama untuk siswa yang lulus
- `G1` adalah fitur paling berpengaruh