

Tugas UTS  
Machine Learning

## 1. Deskripsi Dataset

Dataset yang digunakan adalah dataset variabel target kategorikal '**Student Performance**' yang merupakan hasil pengamatan dari 2 SMA di Portugis. Dataset ini digunakan untuk menganalisis faktor-faktor yang mempengaruhi prestasi belajar siswa pada 2 mata pelajaran: matematika dan bahasa portugis. Berikut adalah penjelasan dari salah satu contoh siswa dari dataset.

Kolom	Nilai	Arti / Penjelasan
school	GP	Siswa berasal dari sekolah Gabriel Pereira.
sex	F	Jenis kelamin siswa adalah perempuan.
age	18	Usia siswa adalah 18 tahun.
address	U	Siswa tinggal di daerah perkotaan (Urban).
famsize	GT3	Ukuran keluarga lebih dari 3 anggota.
Pstatus	A	Orang tua siswa tidak tinggal bersama (Apart).
Medu	4	Pendidikan ibu mencapai universitas.
Fedu	4	Pendidikan ayah juga mencapai universitas.
Mjob	at_home	Pekerjaan ibu adalah ibu rumah tangga.
Fjob	teacher	Pekerjaan ayah adalah guru.
reason	course	Alasan memilih sekolah adalah karena kualitas pelajaran (course).
guardian	mother	Wali utama siswa adalah ibu.
traveltime	2	Waktu tempuh ke sekolah sekitar 15–30 menit.
studytime	2	Waktu belajar di luar kelas sekitar 2–5 jam per minggu.
failures	0	Siswa tidak pernah gagal dalam kelas sebelumnya.
schoolsup	yes	Siswa mendapat dukungan tambahan dari sekolah (misal: bimbingan belajar).
famsup	no	Siswa tidak mendapat dukungan tambahan dari keluarga.
paid	no	Siswa tidak mengikuti kursus berbayar tambahan.
activities	no	Siswa tidak aktif dalam kegiatan ekstrakurikuler.
nursery	yes	Siswa pernah bersekolah di taman kanak-kanak.
higher	yes	Siswa berencana melanjutkan pendidikan ke tingkat yang lebih tinggi.
internet	no	Siswa tidak memiliki akses internet di rumah.
romantic	no	Siswa tidak memiliki hubungan romantis.
famrel	4	Hubungan keluarga dinilai baik (skala 1–5).
freetime	3	Waktu luang setelah sekolah berada di tingkat sedang.
goout	4	Frekuensi keluar bersama teman cukup sering.
Dalc	1	Konsumsi alkohol di hari kerja sangat rendah.
Walc	1	Konsumsi alkohol di akhir pekan juga sangat rendah.
health	3	Kondisi kesehatan cukup baik.

absences	4	Siswa jarang absen (4 kali).
G1, G2, G3	11, 11, 11	Nilai ujian pada setiap tahun pembelajaran adalah 11.

## 2. Model yang Digunakan

Model yang digunakan untuk membuat model klasifikasi untuk memprediksi kelas dari dataset Student Performance adalah **Decision Tree** dan **Support Vector Machine (SVM)**.

Alasannya adalah:

- Decision Tree mampu menangani data campuran (numerik dan kategorikal), mudah diinterpretasikan, serta dapat menunjukkan faktor paling berpengaruh terhadap performa siswa (seperti jam belajar atau dukungan keluarga).
- SVM efektif untuk mencari batas pemisah terbaik antar kelas dengan akurasi tinggi, terutama ketika data memiliki dimensi tinggi atau hubungan antar fitur yang kompleks.

## 3. Hasil Evaluasi dan Pembahasan

Dataset displit menggunakan random state dan stratify menjadi 80% data untuk training dan 20% untuk testing. Hasilnya adalah:

Decision Tree:

- Model mampu mengenali kelas Pass dengan baik, tetapi performanya sangat lemah untuk kelas Fail (precision & recall rendah).
- Dari confusion matrix, terlihat banyak kasus Fail diprediksi sebagai Pass (17 dari 20).
- AUC hanya 0.606, menandakan kemampuan pemisahan antar kelas masih rendah.
- Decision Tree masih overfitting ke kelas mayoritas (Pass) dan kurang mampu mendeteksi siswa yang gagal.

SVM:

- SVM memiliki akurasi dan F1-score lebih tinggi dibanding Decision Tree.
- Kemampuan mengenali kelas Pass meningkat signifikan (Recall = 0.95).
- Walau performa untuk kelas Fail tetap rendah, SVM lebih seimbang dan konsisten.
- AUC meningkat menjadi 0.671, menunjukkan pemisahan antar kelas yang lebih baik.
- SVM memberikan kinerja klasifikasi yang lebih baik secara keseluruhan, terutama dalam membedakan siswa yang lulus, meskipun masih kesulitan mengenali siswa yang gagal.

Kesimpulannya, SVM outperform Decision Tree dalam hal akurasi, recall, dan kemampuan generalisasi. Namun, Decision Tree tetap berguna untuk interpretasi faktor-faktor penting yang memengaruhi performa siswa. Kombinasi keduanya bisa digunakan: Decision Tree untuk insight, SVM untuk prediksi terbaik.

