# LAPORAN PERTEMUAN 4 DATA PREPARATION

1. Pendahuluan  
Data preparation adalah tahapan penting dalam proses machine learning yang bertujuan untuk memastikan bahwa data yang digunakan dalam proses pelatihan model memiliki kualitas yang baik. Tahapan ini mencakup proses pembersihan data, analisis kualitas data, hingga transformasi data menjadi bentuk yang lebih siap untuk diolah. Data yang tidak diolah dengan baik dapat menyebabkan model menjadi bias, tidak akurat, atau bahkan gagal menjalankan fungsi prediksi secara optimal.

2. Sumber Data  
Pada pertemuan ini, data yang digunakan merupakan dataset terkait kelulusan mahasiswa berdasarkan beberapa parameter yang diambil dari perkuliahan sebelumnya. Dataset ini disimpan dalam bentuk file CSV dan berisi kolom seperti IPK, jumlah absensi, waktu belajar, serta status kelulusan mahasiswa. Data ini akan diolah kembali untuk menemukan pola yang dapat membantu proses klasifikasi pada pertemuan berikutnya.

3. Pemeriksaan Struktur Dataset  
Sebelum dilakukan pembersihan, struktur dataset diperiksa menggunakan fungsi info() dan head() untuk memastikan bahwa kolom-kolom terbaca dengan baik oleh program. Dari hasil pemeriksaan awal, dapat diketahui tipe data masing-masing atribut, jumlah data setiap kolom, serta identifikasi awal terhadap kemungkinan nilai kosong.

4. Pengecekan Missing Value  
Missing value merupakan masalah yang sering dijumpai pada dataset. Pada dataset ini diperiksa menggunakan fungsi isnull().sum() untuk mengetahui jumlah nilai yang hilang pada setiap kolom. Berdasarkan hasil pemeriksaan, tidak ditemukan missing value, sehingga data dapat langsung diproses tanpa perlu dilakukan imputasi nilai.

5. Duplikasi Data  
Duplikasi dapat memengaruhi hasil pelatihan model karena nilai yang sama dapat menambah bias terhadap pola tertentu. Pemeriksaan duplikasi dilakukan dan apabila ditemukan data yang identik, sistem akan menghapus entri tersebut. Pada dataset ini tidak ditemukan duplikasi, sehingga tidak memerlukan penghapusan data.

6. Pengecekan Outlier  
Outlier merupakan nilai ekstrem yang berpotensi memengaruhi distribusi data. Pada dataset ini dilakukan visualisasi menggunakan boxplot terhadap nilai IPK. Visualisasi ini membantu untuk mengidentifikasi apakah terdapat mahasiswa dengan nilai IPK yang jauh melenceng dari rentang rata-rata. Dari hasil boxplot, tidak ditemukan outlier signifikan, sehingga tidak diperlukan proses penyesuaian atau penghapusan data.

7. Exploratory Data Analysis (EDA)  
EDA dilakukan untuk mendapatkan gambaran umum mengenai distribusi data. Histogram digunakan untuk melihat distribusi IPK secara keseluruhan, sedangkan scatterplot membantu menunjukkan hubungan antara IPK dan waktu belajar terhadap status kelulusan mahasiswa. Selain itu, heatmap korelasi digunakan untuk mengidentifikasi hubungan antarvariabel, yang nantinya dapat membantu dalam pembentukan model machine learning yang lebih baik.

8. Feature Engineering  
Untuk meningkatkan kualitas dataset, dilakukan proses feature engineering berupa penambahan atribut baru yaitu Rasio Absensi dan IPK\_x\_Study. Rasio Absensi dihitung berdasarkan jumlah kehadiran terhadap total pertemuan, sedangkan IPK\_x\_Study merupakan hasil perkalian antara IPK dan waktu belajar per minggu. Penambahan fitur ini diharapkan dapat memberikan sinyal tambahan mengenai pola kelulusan mahasiswa.

9. Penyimpanan Dataset Baru  
Setelah dilakukan proses pembersihan dan penambahan fitur, dataset disimpan kembali ke dalam file baru dengan nama processed\_kelulusan.csv. Dataset ini nantinya akan digunakan pada pertemuan-pertemuan berikutnya sebagai dataset yang telah siap untuk diproses lebih lanjut menggunakan teknik machine learning.

10. Kesimpulan  
Melalui tahapan data preparation ini, data yang awalnya berbentuk mentah kini telah diproses dan siap untuk digunakan dalam pengembangan model prediksi. Tidak ditemukan missing value maupun duplikasi, namun proses feature engineering dilakukan untuk memperkaya fitur. Dalam proses ini juga dilakukan analisis distribusi data dan hubungan antarvariabel untuk meningkatkan pemahaman terhadap dataset. Tahapan ini merupakan fondasi penting sebelum memasuki proses pemodelan yang akan dilakukan pada pertemuan berikutnya.