**Laporan Pertemuan 4 – Machine Learning**

***Analisis Data Kelulusan Mahasiswa***

**Nama:** Galih Naufal Faturrohman  
**NIM:** 231011402731  
**Program Studi:** Informatika – Universitas Pamulang

**1. Pendahuluan**

Analisis data merupakan tahap awal yang penting dalam proses *Machine Learning*.  
Pada tugas pertemuan ke-4 ini, dilakukan eksplorasi terhadap dataset kelulusan mahasiswa untuk memahami struktur data, mendeteksi anomali, serta menyiapkan data agar siap digunakan pada proses pemodelan.  
Analisis difokuskan pada hubungan antara IPK, jumlah absensi, dan waktu belajar terhadap status kelulusan mahasiswa.

**2. Tujuan**

Tujuan dari analisis ini adalah:

1 Melakukan eksplorasi awal dan pembersihan data kelulusan mahasiswa.

2 Mengidentifikasi nilai kosong, duplikasi, dan anomali data.

3 Melakukan visualisasi menggunakan *boxplot*, *histogram*, *scatterplot*, dan *heatmap*.

4 Menambahkan fitur turunan baru untuk meningkatkan kualitas dataset.

5 Membagi dataset menjadi data latih, validasi, dan uji untuk tahap pemodelan berikutnya.

**3. Metode Analisis**

Analisis dilakukan menggunakan bahasa pemrograman **Python** dengan pustaka *pandas*, *seaborn*, dan *scikit-learn*.  
Tahapan utama meliputi:

**• Membaca dataset** kelulusan\_mahasiswa.csv dan menampilkan informasi struktur data.

**• Membersihkan data** dari nilai kosong dan duplikasi.

**• Visualisasi** dilakukan untuk memahami distribusi IPK dan hubungan antar variabel.

**• Feature Engineering:** menambahkan fitur baru

◦ Rasio\_Absensi = Jumlah\_Absensi / 14

◦ IPK\_x\_Study = IPK \* Waktu\_Belajar\_Jam

**• Membagi data** menjadi tiga bagian:

◦ Training (70%)

◦ Validation (15%)

◦ Testing (15%)

**4. Hasil dan Pembahasan**

Hasil eksplorasi menunjukkan bahwa dataset tidak memiliki nilai kosong maupun data duplikat.  
Visualisasi dengan *boxplot* dan *histogram* memperlihatkan distribusi IPK yang relatif normal dengan sedikit outlier.  
Scatterplot menunjukkan adanya korelasi positif antara IPK dan waktu belajar, di mana mahasiswa dengan waktu belajar lebih tinggi cenderung memiliki IPK lebih baik.  
Selain itu, *heatmap* korelasi menunjukkan bahwa IPK dan jumlah absensi memiliki hubungan yang kuat dengan variabel target **Lulus**.

Fitur baru seperti Rasio\_Absensi dan IPK\_x\_Study memberikan wawasan tambahan terhadap karakteristik mahasiswa dan diharapkan dapat meningkatkan performa model klasifikasi pada tahap selanjutnya.

**5. Kesimpulan**

1 Dataset kelulusan mahasiswa telah dibersihkan dan siap digunakan untuk pemodelan.

2 Analisis visual membantu memahami hubungan antara IPK, absensi, dan waktu belajar terhadap kelulusan.

3 Penambahan fitur baru memperkaya representasi data yang akan digunakan pada model prediktif.

4 Dataset telah dibagi menjadi tiga bagian (train, validation, test) untuk menjamin proses pelatihan dan evaluasi model berjalan optimal.

Hasil analisis ini menjadi fondasi penting untuk proses pemodelan *Machine Learning* di pertemuan berikutnya, menggunakan algoritma seperti Logistic Regression dan Random Forest.