Activity 1 – Dataset Design & Student Profiling

Objective

The objective of Activity 1 is to design a realistic student mentoring dataset that captures academic performance, wellness, productivity, and career readiness dimensions. The dataset is intended to support AI/ML-based analysis for a Dedicated Mentoring System.

Dataset Overview

The dataset (students.csv) is a structured, tabular dataset where each row represents an individual student and each column represents a specific attribute related to the student's academic and behavioral profile. Synthetic data has been generated to ensure privacy while maintaining realism.

Minimum number of records: 50 students (implemented with a larger sample for robustness).

Data Dictionary

Column Name	Description	Scale / Range
student_id	Unique student identifier	Categorical (e.g., S001)
age	Age of the student	Numeric (18–25)
program	Degree or academic stream	Categorical (B.Tech, MBA, etc.)
semester	Current semester of study	Numeric (1–8)
gpa	Academic performance indicator	0–10
attendance	Attendance percentage	0–100
assignments_completion	Percentage of assignments completed	0–100
stress_level	Self-reported stress level	1–10
sleep_hours	Average sleep hours per night	0–10
mental_wellbeing	Mental wellbeing score	1–10
productivity_score	Time management and productivity level	1–10
distractions	Level of distractions faced	1–10
career_clarity	Clarity of career goals	1–10
skill_readiness	Job skill preparedness	1–10
engagement_score	Platform or academic engagement	0–100

Synthetic Data Generation Approach

Synthetic student data was generated using controlled randomization within realistic ranges for each feature. Logical correlations were implicitly maintained, such as:

- Higher stress levels often co-occurring with lower sleep hours
- Higher engagement sometimes compensating for lower GPA
- Strong academic scores not always aligning with career clarity

This approach ensures that the dataset closely resembles real student behavior while remaining privacy-safe.

Identified Student Behavior Patterns

The dataset intentionally reflects common real-world student patterns, including:

1. **High Stress + Low Productivity**
   Students with high stress levels and high distraction scores tend to show reduced productivity scores.
2. **Low GPA + High Engagement**
   Some students exhibit strong engagement despite lower academic performance, indicating motivation but possible learning gaps.
3. **Strong Academics + Unclear Career Goals**
   Students with high GPA and attendance may still show low career clarity, highlighting the need for career-focused mentoring.

These patterns validate the dataset's suitability for mentoring intelligence and intervention analysis.

Conclusion

The designed dataset successfully captures the multi-dimensional nature of student performance and behavior. It provides a strong foundation for downstream rule-based scoring, machine learning analysis, and mentor recommendation systems within the HEPro AI+ mentoring framework.