Assignment: Understanding the AI Development Workflow

Course: AI for Software Engineering
Duration: 7 days
Total Points: 100

Group 131 members Feb2025-Cohort

◆ 1. Problem Definition (6 pts)

Problem Statement: *"Predicting Student Dropout Risk in Online Learning Platforms."*

Objectives:

1. Identify students at risk of dropping out early in the course.
2. Support targeted interventions to improve retention.
3. Help instructors personalize learning experiences.

Stakeholders:

- Learners
- Course Administrators

KPI:

- *Intervention Success Rate:* Percentage of at-risk students who complete the course after receiving targeted support.

◆ 2. Data Collection & Preprocessing (8 pts)

Data Sources:

- Learning Management System (LMS) logs (clickstream data, login frequency, quiz scores)
- Student demographics and prior academic history

Potential Bias:

- Students in regions with limited internet access may be inaccurately flagged as disengaged due to connectivity issues.

Preprocessing Steps:

1. *Imputation* of missing quiz or login data.
2. *Normalization* of engagement metrics (e.g., rescale time spent to a 0–1 range).
3. *One-hot encoding* of categorical variables such as course language or learning preferences.

◆ 3. Model Development (8 pts)

Model Choice:

- *Random Forest Classifier* — Ideal for tabular data, resistant to overfitting, and provides feature importance for explainability.

Data Splitting:

- 70% training, 15% validation, 15% testing
- Stratified sampling to ensure proportional dropout rates in each set.

Hyperparameters to Tune:

1. *n_estimators* (number of trees) — affects performance and robustness.
2. *max_depth* — controls model complexity and risk of overfitting.

◆ 4. Evaluation & Deployment (8 pts)

Evaluation Metrics:

- *Precision:* Minimizes false positives—important when interventions cost time/resources.
- *Recall:* Captures as many true dropouts as possible—crucial for early warnings.

Concept Drift:

- It's the shift in data patterns over time (e.g., new learning platform updates may affect engagement).
- Mitigation: Periodically retrain the model and monitor drift using statistical tests like Kolmogorov–Smirnov.

Technical Deployment Challenge:

- *Scalability:* The model must handle thousands of users in real time—requires cloud deployment and load balancing.