1. Project Proposal

Title: Early Prediction of Student Dropout in Higher Education Using Machine Learning

1. Problem Statement

Student dropout in higher education is a major concern for universities and policy makers because it leads to wasted educational resources, financial losses, and negative personal and social consequences for students and families. Institutions often identify at-risk students too late, when grades have already dropped significantly and disengagement has escalated.

This project aims to build a machine learning model that predicts whether a newly enrolled student is likely to **drop out, remain enrolled, or graduate**, based on demographic, socioeconomic, and academic factors available at or shortly after enrollment. By providing an early warning signal, the model can support proactive interventions such as tutoring, financial aid counseling, or mentoring programs.

2. Objectives

1. **Predictive Objective**
    a. Develop a multi-class classification model to predict the final status of a student:
        i. **Dropout**
        ii. **Enrolled**
        iii. **Graduate**
2. **Analytical Objective**
    a. Identify which features (e.g., previous qualification, parental education, economic indicators, academic performance in early semesters) are most strongly associated with dropout risk.
3. **Practical Objective**
    a. Produce a prototype that could be integrated into an early-warning system used by universities to prioritize student support.
4. **Evaluation Objective**
    a. Evaluate the model’s performance using metrics such as **accuracy**, **precision**, **recall**, **F1-score (macro and weighted)**, and **confusion matrix**, with special emphasis on correctly detecting the **dropout** class.

3. Stakeholders and Target Audience

- **Primary stakeholders**
    - Higher education institutions (administration, deans, department heads)
    - Student success / retention offices
    - Academic advisors and counsellors
- **Secondary stakeholders**
    - Students and their families
    - Policy makers and educational researchers
- **Target audience for this project**
    - Technical audience: AI/ML practitioners, data scientists in education
    - Non-technical audience: university decision makers who need interpretable insights and actionable outputs.

4. Dataset and Justification

We will use the **"Predict Students' Dropout and Academic Success"** dataset from the **UCI Machine Learning Repository** (ID: 697). archive.ics.uci.edu

Key characteristics:

- **Instances:** 4,424 students
- **Features:** 36 attributes
- **Feature types:** integer, real, categorical
- **Task:** multi-class classification (Dropout, Enrolled, Graduate)
- **Domain:** higher education (multiple undergraduate degrees)
- **Licence:** CC BY 4.0 – suitable for academic projects. archive.ics.uci.edu

Why this dataset meets your course requirements:

- **Complexity**
    - Contains mixed data types (demographic, socioeconomic, academic, course-related, and macroeconomic indicators).
    - Strong **class imbalance**: most students belong to one class, making the prediction of dropout more challenging. archive.ics.uci.edu
    - Allows the use of more advanced techniques such as **class imbalance handling**, **feature importance analysis**, and possibly **ensemble models**.
- **Size**

  - 4,424 rows and 36 features is large enough for training and testing robust models without being too big for typical academic compute resources.
- **Relevance**
  - Directly aligned with the goal: predicting student dropout and success in higher education. archive.ics.uci.edu
- **Public Availability**
  - Publicly accessible via UCI ML Repository with clear usage licence and documentation. archive.ics.uci.edu

5. AI Methodologies

We will treat this as a **multi-class classification** problem.

- **Baseline models**
  - **Logistic Regression (multinomial)**
  - **Decision Tree Classifier**
- **Advanced models**
  - **Random Forest Classifier**
  - **Gradient Boosting model** (e.g., XGBoost or sklearn's GradientBoostingClassifier)
- **Preprocessing & Engineering**
  - Use ucimlrepo to fetch the dataset programmatically. archive.ics.uci.edu
  - One-hot encoding for categorical features, scaling of appropriate numeric features.
  - Train–test split (80/20 as recommended by dataset authors). archive.ics.uci.edu
  - Address class imbalance using class weights or resampling techniques.

6. Feasibility

- **Data access:** Dataset is already hosted and documented on UCI.
- **Tooling:** Implementation will use Python with libraries such as **Pandas, NumPy, scikit-learn, Matplotlib/Seaborn; optionally XGBoost**.
- **Time and complexity:**
  - The pipeline from data loading to evaluation can be completed within the course timeline.

    - The project complexity is appropriate for a capstone-preparation assignment, requiring multi-step preprocessing, model comparison, and analysis.

7. Expected Outcomes & Evaluation Metrics

- A trained, evaluated model that estimates each student's probability of **dropout**, **enrollment**, and **graduation**.
- Clear performance metrics:
    - **Accuracy**
    - **Macro and weighted Precision, Recall, F1-score**
    - **Confusion matrix**
    - Optionally **ROC curves** (one-vs-rest) for each class.
- **Success criteria:**
    - Significantly better predictive performance than a naïve baseline (e.g., always predicting the majority class).
    - Reasonable recall for the **dropout** class, which is the main focus for intervention.