Project Title

Predicting Student Dropout Using Classical Machine Learning and TensorFlow Deep Learning

1. Problem Statement

Student dropout is a major challenge in higher education institutions worldwide. Students may discontinue their studies due to academic difficulties, financial pressure, personal issues, or lack of institutional support. High dropout rates negatively impact:

- Institutional funding
- Graduation statistics
- Student success metrics
- Resource planning

To address this problem proactively, institutions need early-warning systems capable of predicting which students are most at risk. Accurate prediction makes timely intervention possible.

This project aims to build a predictive model that classifies students into three outcomes:

- **Dropout**
- **Enrolled (Continuing)**
- **Graduate**

The complexity of the dataset — consisting of 36 socioeconomic, academic, and demographic attributes — makes it an ideal candidate for **advanced AI techniques such as deep learning**.

2. Project Objectives

1. **Develop a data preprocessing pipeline** capable of handling both numerical and categorical variables using ColumnTransformer.

2. **Train and evaluate classical ML models:**
    a. Logistic Regression
    b. Random Forest Classifier
3. **Design, train, and optimize a TensorFlow deep learning model** (Multilayer Perceptron) and compare performance with classical models.
4. **Interpret model outputs** and determine key risk factors associated with student dropout.
5. **Deliver a fully documented end-to-end machine learning solution**, including:
    a. Graphs
    b. Confusion matrices
    c. Training curves
    d. Model comparison
    e. Recommendations for stakeholders

3. Stakeholder & Audience Identification

Primary Stakeholders

- University administrators
- Academic advisors
- Student success and retention departments

Secondary Stakeholders

- Policy makers
- Faculty
- Students and parents

Audience

- Educators seeking data-driven decision tools
- AI/ML practitioners interested in classification models on tabular data
- Researchers studying educational analytics

4. Dataset Description

- **Source:** UCI Machine Learning Repository
- **Instances:** 4,424 students
- **Features:** 36 attributes
- **Target variable:** "Target" with classes:
    - Dropout
    - Enrolled
    - Graduate

Features include:

- Sociodemographic variables
- Academic performance
- Family education level
- Financial status
- Course enrollment metrics

The dataset is **imbalanced**, making weighted metrics and deep learning important.

5. Methodology

5.1 Data Preprocessing

- Feature separation (numeric vs categorical)
- Standardization using StandardScaler
- One-hot encoding via OneHotEncoder
- Complete pipeline with ColumnTransformer

5.2 Classical ML models

1. **Logistic Regression (baseline)**
    a. Softmax output for multi-class classification
    b. Good interpretability
2. **Random Forest**
    a. Handles nonlinear relationships
    b. Provides feature importance

5.3 TensorFlow Deep Learning Model

A fully connected neural network (MLP):

- Input: Preprocessed data (float32)
- Hidden Layer 1: 10 neurons, ReLU
- Hidden Layer 2: 10 neurons, ReLU
- Output Layer: 3 neurons (softmax)
- Loss: Categorical crossentropy
- Optimizer: Adam (LR = 0.001)
- Epochs: 50, batch size: 10
- Validation monitoring

5.4 Evaluation Metrics

- Accuracy
- Precision, Recall, F1-Score
- Confusion matrix
- Training curves (accuracy, loss)

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:
    1. The pipeline from data loading to evaluation can be completed within the course timeline.
    2. The project complexity is appropriate for a capstone-preparation assignment, requiring multi-step preprocessing, model comparison, and analysis.

7. Expected Outcomes

- A robust predictive model enabling early identification of at-risk students
- Insights into key academic and socioeconomic dropout predictors
- A deep learning model outperforming classical models due to nonlinear decision boundaries
- A reusable ML pipeline for future deployment