

# **Research Project Proposal**

Name: Goitom Abirha

## **Fair and Explainable AI: Bias Detection and Transparency in Student Performance Prediction**

### **1. Introduction and Background**

Artificial intelligence is increasingly used to predict student outcomes and inform educational policy. However, these models often reflect historical inequities embedded in training data, resulting in biased predictions that disproportionately disadvantage certain demographic groups. This project investigates how fairness-aware machine-learning and explainable-AI techniques can ensure transparency and equity in academic-performance prediction. The research aligns with Responsible AI and social-impact frameworks that emphasize accountability, interpretability, and ethical deployment.

Ensuring fairness in educational AI systems can support equitable access to learning resources and unbiased student assessment.

Research Questions:

1. How can algorithmic bias in student performance prediction be identified and quantified?
2. What fairness metrics best capture disparities across demographic attributes?
3. How can explainable AI methods improve the interpretability and trustworthiness of these models?

### **2. Objectives**

1. Detect and measure bias in predictive models across sensitive features such as gender, parental education, and socioeconomic status.
2. Evaluate fairness trade-offs using quantitative metrics (Demographic Parity, Equal Opportunity, Disparate Impact).
3. Apply explainable-AI techniques (SHAP, LIME) to interpret model outputs.
4. Develop a reproducible, fair, and transparent AI pipeline suitable for educational decision support.

### **3. Literature and Theoretical Foundation**

Fairness in ML: Barocas et al. (2019) discuss algorithmic fairness principles such as equality of opportunity and demographic parity. Mehrabi et al. (2021) provide a comprehensive review of fairness challenges in AI systems.

Educational Prediction Models: Cortez and Silva (2008) used data mining to predict secondary school student performance, demonstrating the potential of ML for education but not addressing fairness.

Explainable AI (XAI): Lundberg & Lee (2017) introduced SHAP for model interpretability, and Ribeiro et al. (2016) proposed LIME to explain individual predictions. These frameworks inspire this research's integration of fairness and transparency.

## 4. Methodology

### 4.1 Data Source

Portuguese Student Performance dataset (UCI Repository), containing demographic, behavioral, and academic variables (G1–G3).

### 4.2 Exploratory Data Analysis (EDA)

- Descriptive statistics, correlation and association tests (Pearson, Cramer's V).
- Visual analytics: histograms, boxplots, heatmaps, pairplots.
- Bias inspection: group comparisons, chi-square, ANOVA.
- Outlier detection: IQR, Z-score, Isolation Forest.
- Fairness-oriented EDA: subgroup performance gaps and disparity visualizations.

### 4.3 Data Preparation

- Missing-value imputation: mean/median/mode, KNN/IterativeImputer.
- Encoding: One-Hot, Ordinal, Target Encoding.
- Scaling: StandardScaler, RobustScaler, QuantileTransformer.
- Balancing: SMOTE, ADASYN, or class weights.
- Feature engineering: interaction terms, binning, log transforms.

### 4.4 Modeling

- Baseline: Logistic Regression, Random Forest, SVM.
- Fairness-aware models: Reweighting, Prejudice Remover, or Adversarial Debiasing (IBM AIF360).

### 4.5 Evaluation

- Performance metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC.
- Fairness metrics: Demographic Parity Difference, Equal Opportunity Difference, Disparate Impact Ratio.
- Explainability: SHAP summary plots, LIME local explanations, PDP and ICE plots.
- Visualization Dashboard: Streamlit or Flask interface for stakeholders.

All code will be implemented in Python and shared via GitHub for transparency and reproducibility.

## 5. Expected Results

- Quantified bias across sensitive features.
- Fairness-audited predictive model balancing accuracy and equity.
- XAI visualizations improving interpretability and accountability.
- Guidelines for deploying transparent AI in education.

## 6. Project Timeline

Week	Activities
Week 1	Literature review and dataset familiarization
Week 2	Data cleaning, EDA, bias inspection
Week 3	Feature engineering and preprocessing
Week 4	Model development and baseline evaluation
Week 5	Fairness auditing and mitigation
Week 6	Explainability integration (SHAP, LIME)
Week 7	Dashboard, report, and presentation

## 7. Tools and Software

Python 3.12, pandas, numpy, scikit-learn, matplotlib, seaborn, AIF360, SHAP, LIME, imblearn, Streamlit/Flask, GitHub.

## 8. Significance and Contribution

The study contributes to the Responsible-AI discourse by demonstrating a methodology for fair, explainable, and ethically aligned prediction systems in education. It bridges data-science rigor with social-equity impact—offering both technical and policy-relevant insights. The project's open-source nature enables educators and policymakers to adopt transparent AI tools.

## 9. Key References

- Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning.
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? KDD Conference.
- Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. EWSN Conference.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys.