Beyond Performance: Explaining and Ensuring Fairness in Student Academic Performance Prediction with Machine Learning

Abstract:

Use - UCI Student Performance dataset., **AI "brains" (algorithms)**(logical regression, Random Forest, and XGBoost), integrating the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and 5-fold cross-validation for robust model training.

Gives High performance- XGBoost (performance (**accuracy: 0.789**; F1 score: 0.803))

Goal- This study uses XGBoost to predict student grades with 78.9% accuracy while ensuring the AI doesn't discriminate based on gender or family background. By using tools like AIF360 and SMOTE, the researchers proved that machine learning can be both highly effective and socially fair.

Demographics: Gender and what type of school they attend.

SMOTE: Since there were likely fewer "failing" students in the data than "passing" ones, they used a technique called SMOTE to balance the numbers so the AI could learn from both groups equally.

AIF360: This is a specialized toolkit used to measure fairness. They used it to calculate two main things:

Demographic Parity (DP): Does the AI give "passing" scores to different groups at the same rate?

Equalized Odds (EO): Is the AI equally accurate for all groups?

Equalized Odds Difference (EO) measures bias by comparing true and false positive rates across different demographic subgroups.

SHAP explains how each feature (like grades or absences) contributes to a model's prediction in a fair and transparent way.

Adversarial Debiasing: A technique that reduces bias by training the model to make accurate predictions while hiding sensitive attributes like gender.

SMOTE: A method that balances the dataset by creating synthetic samples for underrepresented student outcomes.

Educational Data Mining: The process of analyzing student data to discover patterns that improve learning and performance.

Machine Learning in Education: Using algorithms to predict student outcomes and support early academic intervention.

Model Interpretability: The ability to understand and explain why a model makes a particular prediction.

Introduction:

Many AI models act like "black boxes"—they give a prediction (like "this student might fail") but don't explain **why**, and they can accidentally develop **biases** against students based on their gender or family background.
To solve this, the researchers built a "holistic" framework:

1. **Fairness:** They used **Adversarial Debiasing** and the **AIF360** toolkit to ensure the AI doesn't discriminate based on sensitive factors like parental education or home life.
2. **Clarity:** They used **SHAP** and **LIME** (Explainable AI) to show exactly which factors—like study time or absences—influenced the AI's decision.
3. **Balance:** They used **SMOTE** to balance the data, making sure the AI learns equally well from both passing and failing students.

Educational Data Mining (EDM)

Methodology:

Step 1: Data Preparation

The researchers combined two datasets of Portuguese secondary school students, focusing on grades, absences, and family background. They used a technique called **SMOTE-NC** to create "synthetic" examples of failing students, ensuring the AI had a balanced amount of data to learn from both passing and failing groups.

Step 2: Model Training & Selection

They tested three types of AI "brains": **Logistic Regression**, **Random Forest**, and **XGBoost**. They used 5-fold cross-validation, which means they tested the models on different parts of the data multiple times to make sure the results were consistent and not just lucky guesses.

Step 3: Removing Bias (Fairness)

Using a toolkit called **AIF360**, they checked if the AI was being unfair to students based on gender or parents' education. They applied **Adversarial Debiasing**, which essentially "teaches" the AI to ignore sensitive personal traits while still focusing on academic performance.

Step 4: Explaining the "Why" (XAI)

To make the AI transparent, they used **SHAP** and **LIME**. These tools act like a "translator," showing exactly which factors (like high absences or past failures) led to a specific prediction so teachers can understand the AI's logic.

Step 5: Final Evaluation

The researchers found that **Logistic Regression** was the most effective on this specific data, reaching **96.2% accuracy**. Most importantly, they successfully reduced the bias gap (Demographic Parity) from **0.048 down to 0.021**, making the system much fairer.

Limitations:

Despite the success, the researchers noted a few hurdles:

- **Small Scope:** The data only represents secondary school students in **Portugal**, so the results might not apply to every country.
- **Synthetic Data:** Using **SMOTE** might create "fake" student patterns that don't perfectly match real-life complexities.
- **Timing:** The highest accuracy relied on mid-term grades (G1 and G2); without them, the AI's ability to predict success early in the year is much lower.
-

Future Work

Future research should consider the following:

• Multi-class classification of academic grades (e.g., A–F), enabling finer-grained interventions
and personalized support.

• Longitudinal outcome modeling to track academic progression across years.

• Hybrid ensembles that blend the simplicity of linear models with the robustness of non-linear classifiers.

• External validation across institutions with different demographic and curricular profiles.

Results:

The study compared predictive power against fairness metrics:

- **Top Performance: Logistic Regression** achieved the highest numerical accuracy at **96.2%** with an AUC of 0.987.
- **Key Drivers: G1 and G2 (past grades)**, **absences**, and **past failures** were the strongest predictors of the final outcome.
- **Fairness Gains:** After applying debiasing, the bias gap (Demographic Parity) dropped significantly from **0.048 to 0.021**, with only a tiny 1.8% drop in accuracy.

Conclusion:

The authors concluded that it is possible to build AI systems that are both highly accurate and socially responsible:

- **Transparency:** Simpler models like Logistic Regression are often better for schools because they are easier for educators to trust and interpret.
- **Impact:** These tools can help schools proactively identify at-risk students and distribute resources more equitably.
- **Framework:** The proposed **ETIK-AI guidelines** offer a blueprint for schools to adopt AI ethically.

Part A: Limitations of This Paper

1. Limited Dataset Size and Diversity

- The study uses only **395 students**
- Data comes from **two Portuguese secondary schools**
- Students from other countries, cultures, or education systems are not included

Limitation

The results **cannot be generalized globally**.

Thesis Opportunity

Validate the model on:

- A larger dataset
- A local or multi-country dataset

---

2. Use of G1 and G2 (Data Leakage Risk)

- G1 and G2 are **mid-term exam scores**
- These are not available at the beginning of the academic year

Limitation

The model is **not suitable for true early intervention**.

Thesis Opportunity

Build an **early-warning system** using only:

- Demographic
- Behavioral
- Socioeconomic features

---

3. Reliance on SMOTE (Synthetic Data Risk)

- SMOTE generates **artificial samples**
- Synthetic data may not reflect real student behavior

Limitation

Potential **overfitting** and **artificial patterns**

Thesis Opportunity

Compare:

- SMOTE
- Fair-SMOTE
- GAN-based oversampling
- Cost-sensitive learning

---

4. Limited Fairness Attributes

- Fairness is evaluated mainly on:
    - Gender
    - School type
    - Parental education
    - Family size

Limitation

Psychological, motivational, and mental-health factors are missing.

Thesis Opportunity

Include:

- Engagement
- Attendance patterns
- Learning behavior features

---

5. Limited Fairness Metrics

- Only two fairness metrics are used:
    - Demographic Parity
    - Equalized Odds

Limitation

Other fairness notions are ignored.

Thesis Opportunity

Extend analysis with:

- Individual Fairness
- Counterfactual Fairness
- Predictive Parity

---

6. Single Bias-Mitigation Technique

- Only **adversarial debiasing** (in-processing) is used

Limitation

No comparison with other bias-mitigation strategies.

Thesis Opportunity

Compare:

- Pre-processing (Reweighing, Fair-SMOTE)
- In-processing (Adversarial Debiasing)
- Post-processing (Equalized Odds Post-processing)

---

7. No Real-World or Human Evaluation

- Teachers or students were not involved

- Explainability effectiveness was not tested in practice

Limitation
Practical usefulness is unknown.

Thesis Opportunity
Conduct:

- Human-in-the-loop evaluation
- Teacher-focused explanation analysis

---

8. Binary Classification Only

- Outcome is only **Pass / Fail**

Limitation
Cannot provide fine-grained academic recommendations.

Thesis Opportunity
Use:

- Multi-class grade prediction (A, B, C, D)
- Regression-based performance prediction

Part B: What You Can Do in Your Thesis (Strong Topics)

Thesis Topic 1 (Highly Recommended)

What you do:

- Remove G1 and G2
- Use early features only
- Apply fairness and explainability
- Compare with the original study

Thesis Topic 2

You compare:

- SMOTE
- Fair-SMOTE
- Reweighing
- Adversarial Debiasing

Evaluation:

- Accuracy
- F1 score
- Demographic Parity
- Equalized Odds

Thesis Topic 3 (Ethics-Oriented)

Focus on:

- Gender × Socioeconomic status
- Parental education × Family structure
- Bias amplification analysis

Thesis Topic 4 (Explainable AI)

- SHAP vs LIME
- Case-based explanations
- Trust and interpretability analysis

Thesis Topic 5 (Advanced but Practical)

- Default vs optimized thresholds
- Impact on recall, fairness, and false negatives