# Human AI Interaction CS698Y
# Assignment 2

Tsewang Namgail (241110093)
Sevak Shekokar (241110065)

September 2025

## Report on Fairness Evaluation and Mitigation on "Predict Students' Dropout and Academic Success" dataset

### 1. Introduction:

In education, predicting student dropout early can transform lives by enabling timely support. However, without care, predictive models can reinforce existing gender disparities, unintentionally holding back those who need help most. This project is dedicated to building a dropout prediction model that is not only **accurate** but also **fair**—striving for equal opportunity across genders using the UCI "Predict Students' Dropout and Academic Success" dataset.

### 2. Dataset Overview

Before modeling, we took a close look at the data to understand the story it tells.

- The dataset includes **4424 student records** with 37 variables each.
- Outcomes are categorized as **Graduate**, **Dropout**, or **Enrolled**. We excluded the "Enrolled" group because their final outcomes remain unclear, and including them could confuse the model.
- Gender is coded as 0 for Male and 1 for Female.

### 3. Initial Bias Assessment

The data revealed striking disparities that mirror real-world inequities.

- Females faced a graduation rate of just **44%**, while males graduated at nearly **70%**.
- Dropout rates told a similar tale: **56% for females**, almost twice the **30% for males**.
- Outcome labels align with performance: for both genders, graduates score approximately **5**–**7** points higher than dropouts (Female **+5.33** & **+6.59**; Male **+5.08** & **+6.60**), males slightly outperform within the same outcome (approximately **0.75**–**0.83**), and dropouts' grades fall from $1^{st} \rightarrow 2^{nd}$ semester (**-1.27**$_F$ / **-1.44**$_M$) while graduates remain stable (**-0.01**$_F$ / **+0.07**$_M$).
- Scholarship status further differentiated success, with non-scholarship females graduating at only **38%**.

This stark picture sets the stage: any predictive model ignoring these gaps risks perpetuating unfairness.

## 4. Model Building & Bias Quantification

The baseline logistic regression model, trained with all features including gender, performed well overall—but uncovered deep biases:

- Accuracy reached **75.3%**, with strong recall.
- Yet fairness metrics revealed imbalance: Statistical Parity Difference (SPD) of **-0.375** and Equal Opportunity Difference (EOD) of **-0.265** showed females were under-predicted and under-identified as graduates.
- Males had approximately **78%** accuracy with **94%** recall; females trailed at **70%** accuracy and **68%** recall.

This baseline was an important reality check, showing that accuracy alone isn't enough.

## 5. Corrective Measures to Mitigate Gender Bias

We explored three key interventions to restore fairness:

In selecting bias mitigation strategies, we aimed to balance maintaining predictive accuracy and improving fairness for gender groups. Feature elimination (dropping gender) removes explicit gender information, testing gender-blind decision-making effects. However, proxies in data could still encode bias, necessitating reweighting using the Kamiran–Calders A×Y method, which corrects underlying sample imbalances to equalize model focus across groups. Finally, calibration (isotonic regression) adjusts predicted probabilities post-modeling to reduce disparities in score distributions and enhance prediction reliability. Together, these approaches address fairness broadly across different modeling stages.

### 5.1 Feature Elimination (Drop Gender)

Removing gender from the model features gave it a "blind" perspective:

- Accuracy dipped slightly to **74.1%**.
- Female recall improved to **82.4%**.
- Bias metrics improved notably (SPD **-0.169**, EOD **-0.091**).

A modest trade-off, but an important step toward fairness.

### 5.2 Reweighting (Kamiran–Calders A×Y)

Adjusting sample weights balanced the dataset's underlying inequalities:

- Accuracy slightly rebounded to **74.5%**.
- Female recall surged to **89.7%**.
- Bias disparities nearly vanished (SPD **-0.113**, EOD **-0.017**).

This demonstrated the power of thoughtful data balancing.

### 5.3 Calibration (Isotonic Regression)

Calibrating model predictions refined fairness post-modeling:

- Accuracy climbed back to **75.2%**.
- Bias reduced moderately (SPD **-0.193**, EOD **-0.106**).
- Female recall remained better than baseline but lower than reweighting.

Calibration was the finishing touch improving prediction reliability.

## 6. Before/After Comparison and Per-Group Analysis

Mitigation efforts restored hope for female students:

- Female recall nearly doubled from **68%** to **90%** after reweighting.
- Male students' performance remained stable, preserving high accuracy and recall.
- Improved female precision and F1 scores closed historical gaps.

## 7. Clarifying Key Fairness Metrics

- **Statistical Parity Difference (SPD):** Measures how often each gender receives positive predictions. Values closer to zero reflect balanced treatment.
- **Equal Opportunity Difference (EOD):** Measures parity in true positive rates (recall) across genders. Zero means equal capability to identify successes.

"We report SPD and EOD using the sign Female  Male; negative values indicate worse outcomes for female students."

## 8. Summary Table of Results

| Model | Accuracy | SPD | EOD | Female Recall | Male Recall |
|---|---|---|---|---|---|
| Baseline (with Gender) | 0.753 | -0.375 | -0.265 | ∼0.68 | ∼0.94 |
| Drop Gender | 0.741 | -0.169 | -0.091 | ∼0.82 | ∼0.92 |
| Reweighted (A×Y) | 0.745 | -0.113 | -0.017 | ∼0.90 | ∼0.90 |
| Calibrated (No Gender) | 0.752 | -0.193 | -0.106 | ∼0.79 | ∼0.87 |

Table 1: Key metrics for baseline and mitigated models.

## 9. Conclusion

Through detailed examination and targeted bias mitigation, this work charts a path to dropout prediction that balances predictive accuracy with fairness. By elevating female student success predictions while maintaining overall model strength, the project supports more equitable educational interventions.

## Code Availability

The full code and resources for this project are available on GitHub at:
`https://github.com/TSEN4M/HAI-CS698Y-Assignment-2`.

## Contributors

- **Tsewang Namgail (241110093)**
  - Data exploration and bias evaluation
  - Implementation of mitigation B (Reweighting) and mitigation C (Calibration)
  - Evaluation and results analysis
  - README and code documentation
- **Sevak Shekokar (241110065)**

- Data exploration and bias evaluation
- Implementation of baseline model and mitigation A (Drop Gender)
- Evaluation and results analysis
- Dockerfile and code documentation

Contributions are equally ditributed across the major components of the project, collaborating on data analysis, modeling, fairness mitigations, evaluation, and documentation.