Project Report: Making Fair Predictive Models - Mitigating Bias

Team Details:

1. Mohd Nasar Siddiqui 220661 – contribution 50%, designing model architecture, evaluating fairness
2. Kartik 220503 – contribution 50%, estimating bias and feature engineering

Github repo- https://github.com/mohd-nasar/CS698-EthicsOFAI

1. Introduction

This project focused on building fair and accurate predictive models for student outcomes in education. The goal is to predict whether a student will **Dropout**, remain **Enrolled**, or **Graduate** using the **UCI "Predict Students' Dropout and Academic Success" dataset**. A key challenge in building such models is the potential for inheriting or amplifying real-world biases from the data, which can lead to unfair outcomes for different demographic groups. This project addresses these concerns by not only aiming for high accuracy but also by actively identifying, evaluating, and mitigating bias.

1.1 Objectives

- **Bias Identification**: Detect potential sources of unfairness in the dataset.
- **Model Development**: Train classification models, such as XGBoost and Random Forest, to predict student outcomes.
- **Fairness Evaluation**: Assess model performance across demographic groups using specific fairness metrics.
- **Bias Mitigation**: Apply strategies like feature elimination, reweighting, and calibration to reduce bias.
- **Ethical Reflection**: Analyze the trade-offs between a model's accuracy and its fairness.

The ultimate aim is to create a model that is both **accurate** and **equitable** to support fair decision-making in an educational context.

---

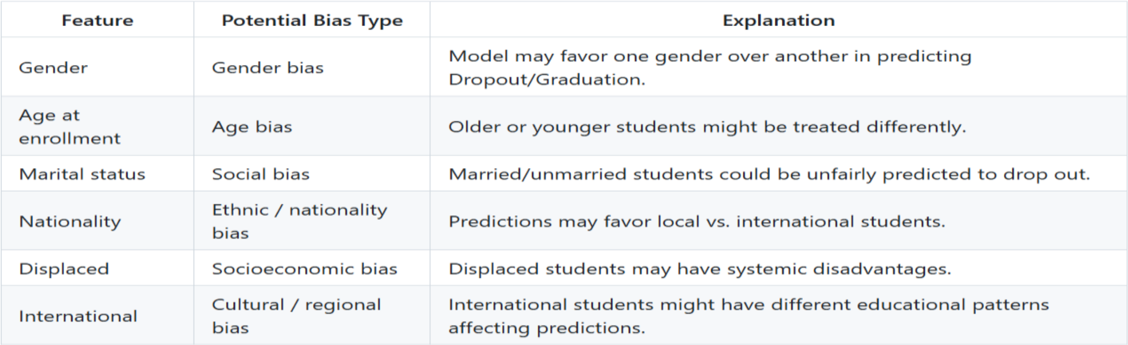
2. Dataset and Bias Analysis

The project uses the UCI "Predict Students' Dropout and Academic Success" dataset, which contains 37 attributes. These attributes can be categorized into four main groups: Demographic, Socioeconomic, Academic, and Macroeconomic.

2.1 Potential Bias in Dataset Features

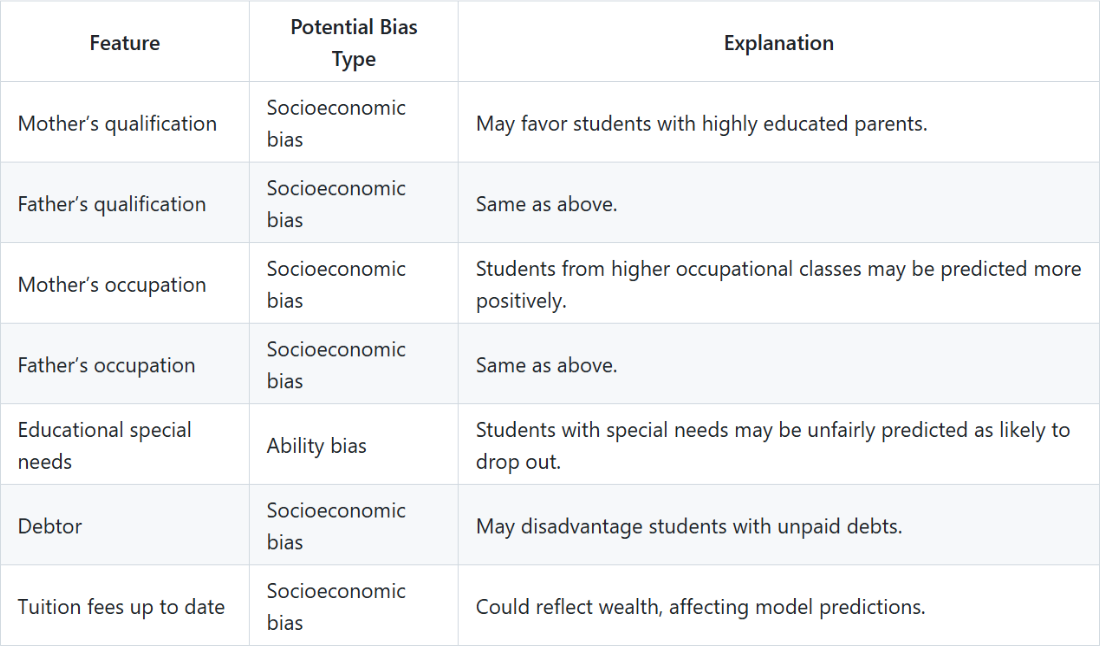
Several features within the dataset are identified as potential sources of bias, as they relate to sensitive or protected attributes.

- **Demographic Features**



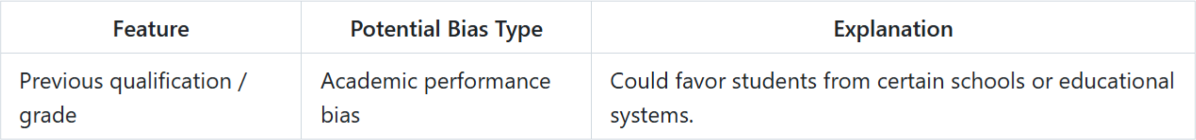
Feature	Potential Bias Type	Explanation
Gender	Gender bias	Model may favor one gender over another in predicting Dropout/Graduation.
Age at enrollment	Age bias	Older or younger students might be treated differently.
Marital status	Social bias	Married/unmarried students could be unfairly predicted to drop out.
Nationality	Ethnic / nationality bias	Predictions may favor local vs. international students.
Displaced	Socioeconomic bias	Displaced students may have systemic disadvantages.
International	Cultural / regional bias	International students might have different educational patterns affecting predictions.

Socioeconomic Features:



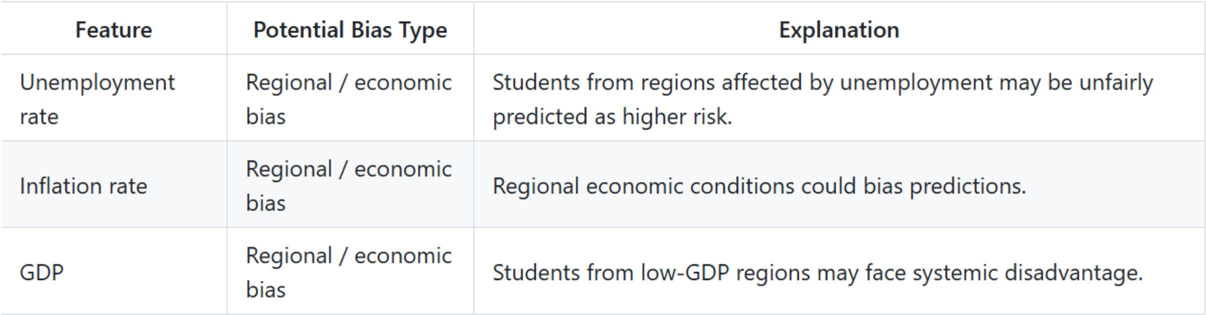
Feature	Potential Bias Type	Explanation
Mother's qualification	Socioeconomic bias	May favor students with highly educated parents.
Father's qualification	Socioeconomic bias	Same as above.
Mother's occupation	Socioeconomic bias	Students from higher occupational classes may be predicted more positively.
Father's occupation	Socioeconomic bias	Same as above.
Educational special needs	Ability bias	Students with special needs may be unfairly predicted as likely to drop out.
Debtor	Socioeconomic bias	May disadvantage students with unpaid debts.
Tuition fees up to date	Socioeconomic bias	Could reflect wealth, affecting model predictions.

- **Academic Features:**



Feature	Potential Bias Type	Explanation
Previous qualification / grade	Academic performance bias	Could favor students from certain schools or educational systems.

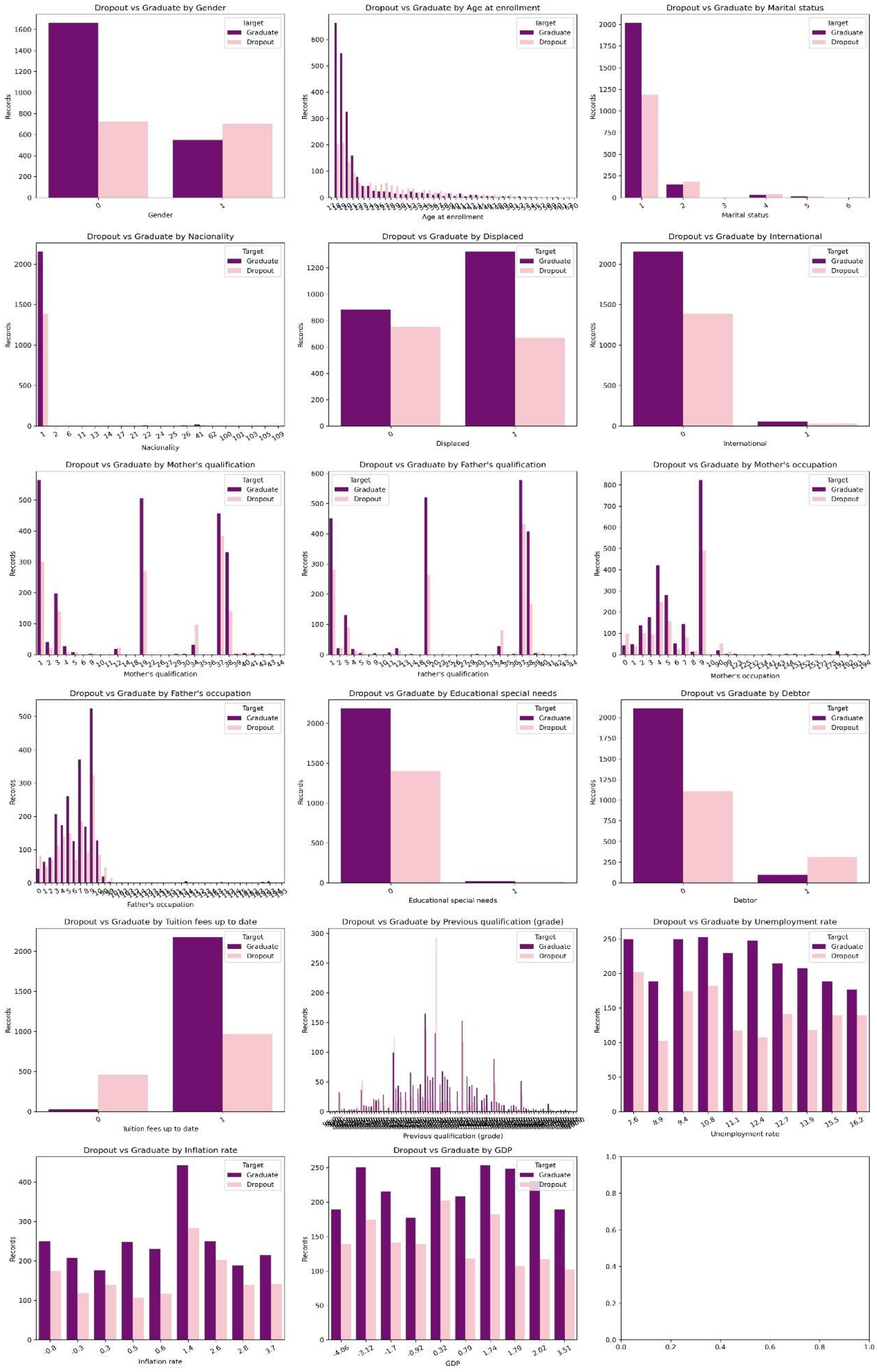
- **Macroeconomic Features:**



Feature	Potential Bias Type	Explanation
Unemployment rate	Regional / economic bias	Students from regions affected by unemployment may be unfairly predicted as higher risk.
Inflation rate	Regional / economic bias	Regional economic conditions could bias predictions.
GDP	Regional / economic bias	Students from low-GDP regions may face systemic disadvantage.

2.2 Bias Analysis Graphs

Here is the distribution of **Dropout vs Graduate** across bias-sensitive features:



Insights

- **Gender** The data is biased to Femaale Gender for more Dropouts than Graduated, whcih ML model can consider Gender as important for feature for predicting dropout or biased to dropout.
- **Age** More Graduates to Droput ratio for lower age as age increases Droputs exceeds Graduates
- **Martial Status** Clearly uneven Graduates Droput proportion, biased for Dropout for marriage persons
- **Nacionality, International, Educational Special Nees** have very few data points for minority classes thereby supressing it and not giving good prediction for them.
- **Tution Fee, Debitor** The data is biased towards one class of people for Droput creating imbalance so the model will takes these features for predicting Dropouts creating biases.
- **GDP, Inflation, Unemployment Rates** I think its Generational Global Universal Problem hence don't have any effect whatsoever neither in making predictions nor biases.
- **Previous Qualification** More Dropouts for low values

3. Defining and Measuring Fairness

3.1 Definition of Fairness

In predictive modeling, fairness means that a model's decisions should not systematically discriminate against individuals or groups based on protected attributes like gender, age, or socioeconomic status. A fair model ensures that individuals with a similar likelihood of a target outcome are treated similarly, regardless of their group membership. This goes beyond simple accuracy by ensuring that the model's errors and benefits are not disproportionately distributed.

3.2 Fairness Metrics

To ensure the model is fair and equitable, the following metrics will be used for evaluation:

1. **Statistical Parity**: Measures whether the positive outcome rate (predicting academic success) is similar across different groups. A **Statistical Parity Ratio** close to **1** indicates fairness. A ratio below **0.8** is a sign of concern.

$$\text{Ratio} = P(Y^\wedge = 1 \mid A = \text{privileged}) / P(Y^\wedge = 1 \mid A = \text{unprivileged})$$

2. **Equal Opportunity**: Assesses if the model is equally effective at identifying successful students across all groups by focusing on the **True Positive Rate (Recall)**.

$$P(Y^\wedge = 1 \mid Y = 1, A = a1) = P(Y^\wedge = 1 \mid Y = 1, A = a2)$$

3. **Equalized Odds**: A more comprehensive metric that ensures the model is fair in identifying both positive cases (True Positive Rate) and negative cases (False Positive Rate).

$$P(\hat{Y} = 1 \mid Y = 1, A = a1) = P(\hat{Y} = 1 \mid Y = 1, A = a2)$$

$$P(\hat{Y} = 1 \mid Y = 0, A = a1) = P(\hat{Y} = 1 \mid Y = 0, A = a2)$$

4. Modeling and Evaluation

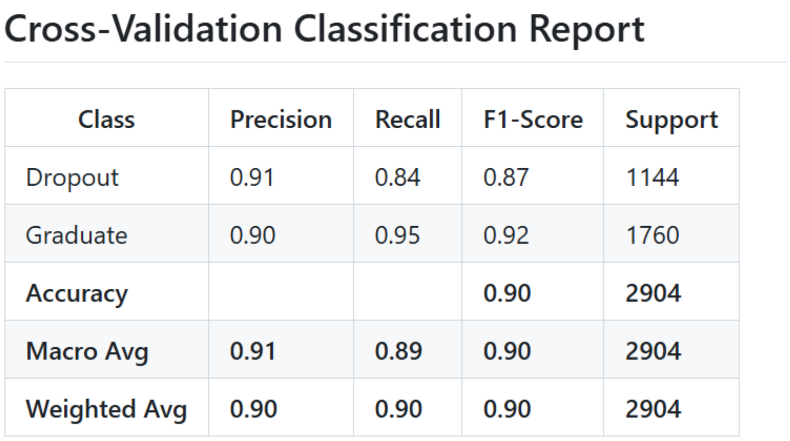
4.1 Modeling Architecture

The project's modeling process involves several key steps:

- **Data Preparation**: Load the dataset, filter for Dropout and Graduate outcomes, and encode the target labels as binary (0 = Dropout, 1 = Graduate).
- **Feature Engineering & Scaling**: Standardize features using StandardScaler to ensure they contribute equally to the model.
- **Model Training**:
  - A **XGBoost Classifier** is trained using **5-fold K-Fold Stratified Cross-Validation** to maintain balanced class distributions.
  - During each fold, both performance and fairness metrics are computed.
- **Fairness Assessment**: The Fairlearn's MetricFrame and demographic_parity_ratio are used to measure fairness, with a Demographic Parity Ratio (DPR) below **0.8** indicating a possible disparate impact.
- **Final Evaluation**: The model is retrained on the entire training set and evaluated on a held-out test set to report overall accuracy, F1 score, and fairness metrics.

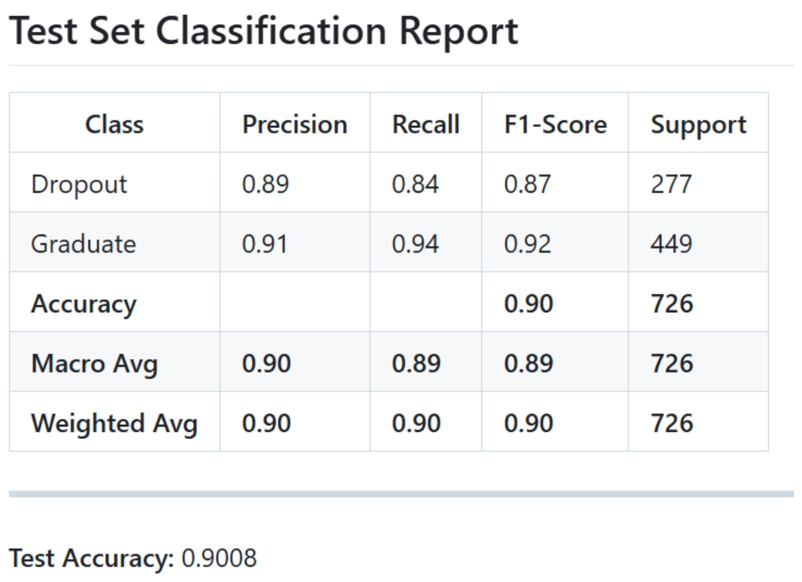
5. Results

5.1 Modeling on Biased Dataset



Cross-Validation Classification Report

Class	Precision	Recall	F1-Score	Support
Dropout	0.91	0.84	0.87	1144
Graduate	0.90	0.95	0.92	1760
Accuracy			**0.90**	**2904**
Macro Avg	**0.91**	**0.89**	**0.90**	**2904**
Weighted Avg	**0.90**	**0.90**	**0.90**	**2904**

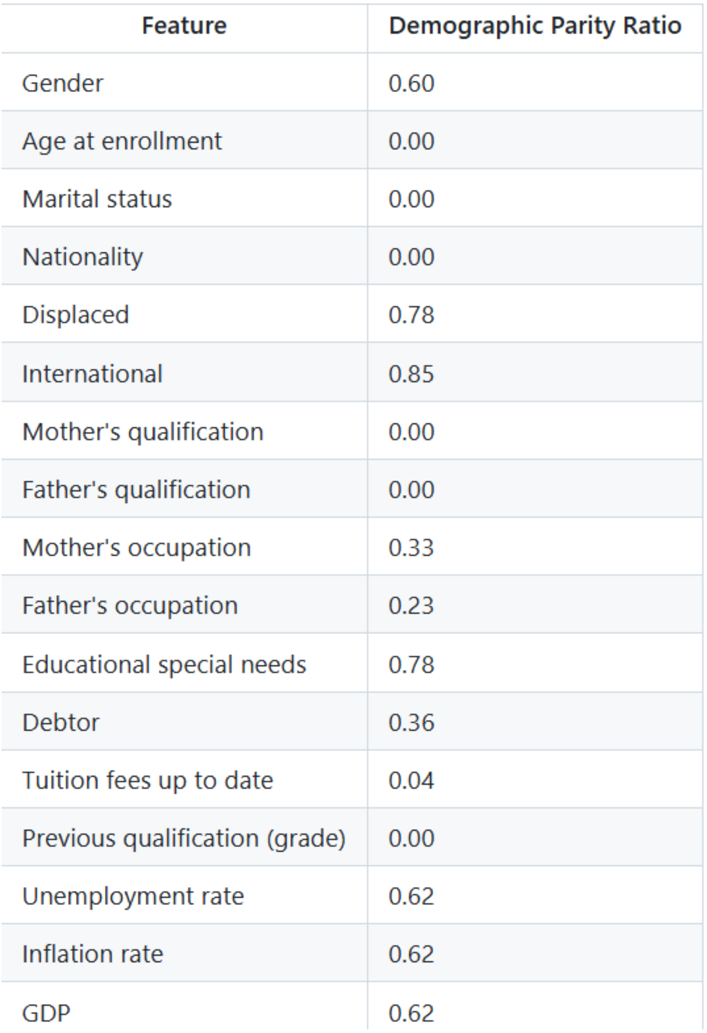


Test Set Classification Report

Class	Precision	Recall	F1-Score	Support
Dropout	0.89	0.84	0.87	277
Graduate	0.91	0.94	0.92	449
Accuracy			**0.90**	**726**
Macro Avg	**0.90**	**0.89**	**0.89**	**726**
Weighted Avg	**0.90**	**0.90**	**0.90**	**726**

Test Accuracy: 0.9008

Fairness Across Protected Groups (Demographic Parity Ratio) The Demographic Parity Ratio (DPR) for many features was significantly below the acceptable threshold of 0.8, indicating major fairness concerns:



Feature	Demographic Parity Ratio
Gender	0.60
Age at enrollment	0.00
Marital status	0.00
Nationality	0.00
Displaced	0.78
International	0.85
Mother's qualification	0.00
Father's qualification	0.00
Mother's occupation	0.33
Father's occupation	0.23
Educational special needs	0.78
Debtor	0.36
Tuition fees up to date	0.04
Previous qualification (grade)	0.00
Unemployment rate	0.62
Inflation rate	0.62
GDP	0.62

5.2 Bias Mitigation

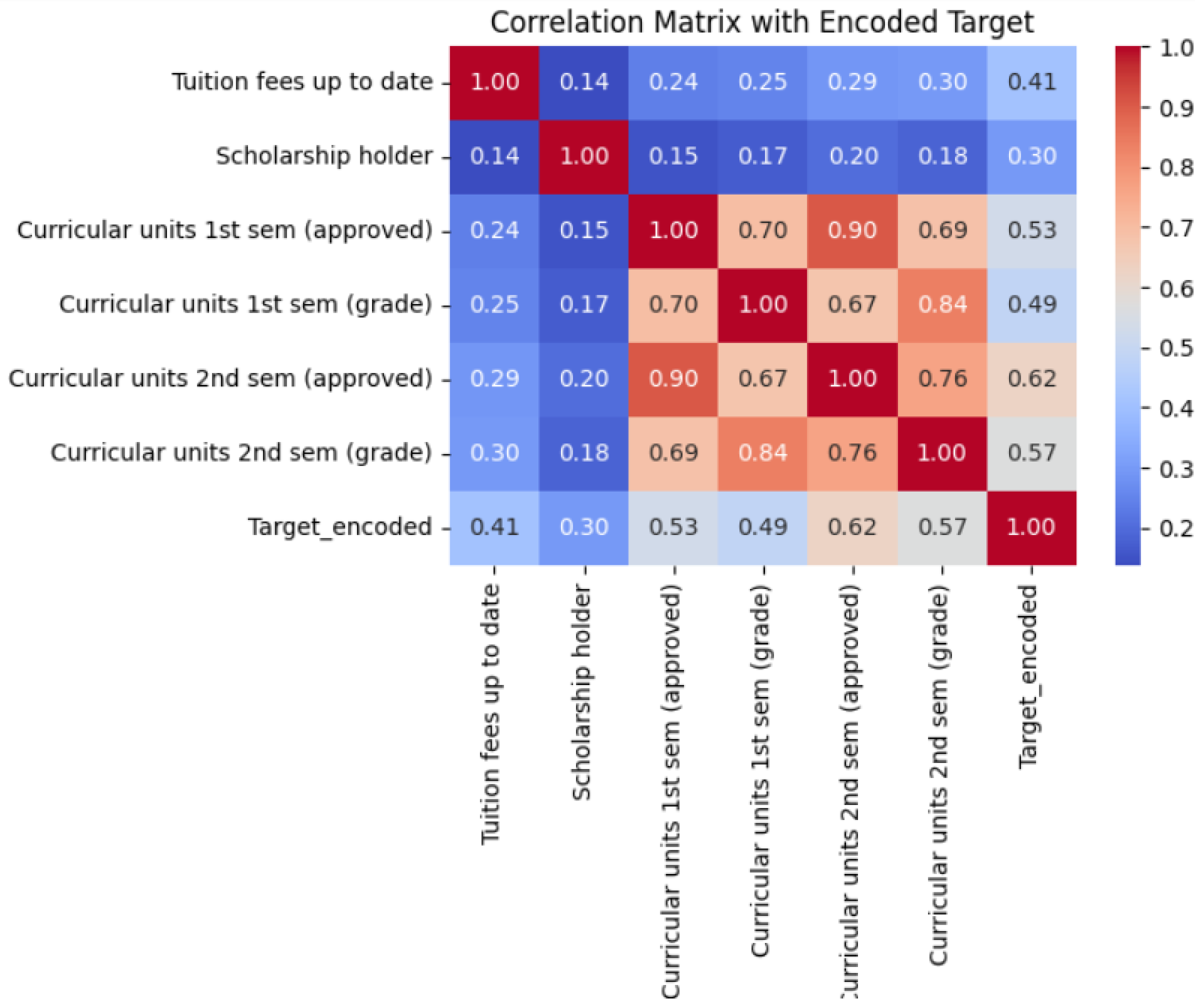
Bias mitigation was attempted by dropping features that were identified as directly or indirectly introducing bias, along with reweighting minority classes. The results of the model after these mitigation steps are as follows:

Overall Metrics (After Mitigation)

- **Accuracy**: 0.8636
- **Precision**: 0.86
- **Recall**: 0.85
- **F1 Score**: 0.8624

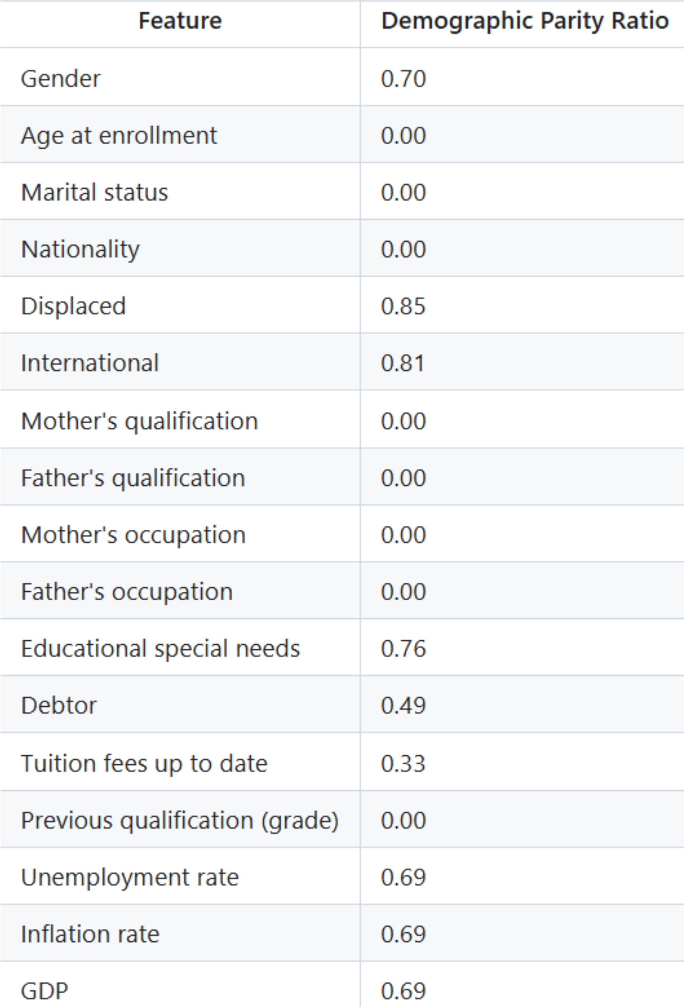
The overall accuracy and F1 score decreased from around **0.90** to **0.86**.

Correlation Matrix:



	Tuition fees up to date	Scholarship holder	Curricular units 1st sem (approved)	Curricular units 1st sem (grade)	urricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Target_encoded
Tuition fees up to date	1.00	0.14	0.24	0.25	0.29	0.30	0.41
Scholarship holder	0.14	1.00	0.15	0.17	0.20	0.18	0.30
Curricular units 1st sem (approved)	0.24	0.15	1.00	0.70	0.90	0.69	0.53
Curricular units 1st sem (grade)	0.25	0.17	0.70	1.00	0.67	0.84	0.49
Curricular units 2nd sem (approved)	0.29	0.20	0.90	0.67	1.00	0.76	0.62
Curricular units 2nd sem (grade)	0.30	0.18	0.69	0.84	0.76	1.00	0.57
Target_encoded	0.41	0.30	0.53	0.49	0.62	0.57	1.00

Statistical Parity by Feature (After Mitigation) While the model's overall performance decreased, some fairness metrics improved:



Feature	Demographic Parity Ratio
Gender	0.70
Age at enrollment	0.00
Marital status	0.00
Nationality	0.00
Displaced	0.85
International	0.81
Mother's qualification	0.00
Father's qualification	0.00
Mother's occupation	0.00
Father's occupation	0.00
Educational special needs	0.76
Debtor	0.49
Tuition fees up to date	0.33
Previous qualification (grade)	0.00
Unemployment rate	0.69
Inflation rate	0.69
GDP	0.69

For features like Gender, Displaced, International, Debtor, Tuition fees up to date, and the macroeconomic features, the Demographic Parity Ratio improved, indicating a reduction in bias. The results highlight a clear **trade-off between model accuracy and fairness**.

---

6. Conclusion and Future Work

This project successfully demonstrated that predictive models can inherit and amplify biases present in real-world data, leading to unfair outcomes. The initial XGBoost model, while highly accurate, showed significant bias across various demographic and socioeconomic features. By applying bias mitigation strategies, the project was able to improve the model's fairness, although this came at the cost of a decrease in overall accuracy.