mahendra kumar chandra

Final_research_report.docx

SRM University AP Amravati

Document Details

Submission ID

trn:oid:::8044:106157507

Submission Date

Jul 28, 2025, 12:33 PM GMT+5:30

Download Date

Jul 28, 2025, 12:34 PM GMT+5:30

File Name

Final_research_report.docx

File Size

5.0 MB

25 Pages

4,001 Words

26,769 Characters

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 10 words)

Exclusions

- 3 Excluded Matches

Match Groups

14 Not Cited or Quoted 6%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

1%	Internet sources
0%	Publications
5%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

14 Not Cited or Quoted 6%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

1%	Internet sources
0%	Publications
5%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	**University of East London on 2025-05-09**	**<1%**
2	Submitted works	**New York Institute of Technology on 2024-12-18**	**<1%**
3	Submitted works	**Johns Hopkins Unversity on 2024-06-25**	**<1%**
4	Submitted works	**University of Central Florida on 2023-12-04**	**<1%**
5	Submitted works	**Indian institute of Management, Udaipur on 2024-10-01**	**<1%**
6	Submitted works	**University of Westminster on 2024-07-28**	**<1%**
7	Internet	**www.deus.ai**	**<1%**
8	Submitted works	**Alliance University on 2025-03-24**	**<1%**
9	Submitted works	**York St John University on 2024-12-04**	**<1%**
10	Submitted works	**University of Hull on 2024-01-10**	**<1%**

11	Submitted works	
University of Salford on 2023-12-08		<1%
12	Internet	
ijsrem.com		<1%
13	Submitted works	
Kaplan College on 2025-02-02		<1%
14	Internet	
link.springer.com		<1%

A RESEARCH PROJECT REPORT

ON

BIAS & FAIRNESS IN EDUCATIONAL AI

Prepared in the partial fulfillment of the

Summer Internship Course

Submitted by

Ugrapalli Mukesh - AP23110010175

Bhanu Kiran Annavarapu - AP23110010256

Mahendra Kumar Chandra - AP23110010328

Bikki Bindu Venkata Priya - AP23110010365

Syed Mohammad Sameer - AP23110010829

Under the Guidance of

Mr. Ayush Bijoura

Assistant Professor, Department of CSE, SRM University–AP

Neerukonda , Mangalagiri, Guntur

Andhra Pradesh – 522 240

[July, 2025]

Table of Contents

1.Abstract

This study investigates the feasibility and challenges of combining two datasets: the Adult Census Income dataset and the Student Performance dataset to examine bias and fairness in educational AI systems. The Adult Census Income dataset provides comprehensive demographic and employment characteristics focused on income prediction, while the Student Performance dataset offers insights into factors influencing academic achievement through demographic, social, and academic variables. The research methodology involves merging these datasets to explore the relationship between adult socioeconomic factors and student educational outcomes. By analyzing the connection between parental income, education levels, occupation and student GPA and grade classifications, this investigation aims to understand how economic disparities translate into educational inequalities. The study employs machine learning algorithms including logistic regression and decision trees, evaluated through fairness metrics such as demographic parity and equal opportunity. Results reveal the mechanisms through which socioeconomic factors shape student academic trajectories. The findings provide critical insights for developing targeted interventions and policies aimed at promoting educational equity and social mobility in AI-driven educational systems.

2.Introduction

This report investigates the feasibility and challenges of combining two rich datasets: the Adult Census Income dataset and the Student Performance dataset to examine bias and fairness in educational AI systems. The integration of these datasets represents a crucial step toward understanding how algorithmic bias manifests in educational contexts and impacts student outcomes.The Adult Census Income dataset offers a comprehensive view of adult demographic and employment characteristics, with particular focus on income prediction across diverse population segments. This dataset provides essential context for understanding the socioeconomic landscape that influences educational opportunities and outcomes. Moreover, the Student Performance dataset provides a granular perspective on factors influencing student academic achievement, encompassing demographic, social, and academic variables that directly impact learning outcomes.The core motivation for merging these datasets is to explore the nuanced relationship between adult socioeconomic factors and student educational outcomes within the framework of algorithmic fairness. By linking parental income, education levels, and occupation with student GPA and grade classifications, this investigation aims to understand the extent to which economic disparities translate into educational inequalities through AI-mediated decision-making processes.

This analysis has significant potential to reveal critical insights into the mechanisms through which socioeconomic factors shape student academic trajectories. The research addresses fundamental questions about how bias propagates through educational AI systems and how these systems may inadvertently perpetuate existing inequalities. Through systematic evaluation of fairness metrics and bias detection methodologies, this study contributes to the development of more equitable AI systems in educational contexts.The findings from this investigation will inform the development of targeted interventions and policies aimed at promoting educational equity and social mobility, ensuring that AI technologies serve to reduce rather than amplify existing educational disparities.

3.Literature Review

The intersection of bias, fairness, and artificial intelligence has emerged as a critical area of research, particularly within the context of educational systems where algorithmic decisions can profoundly impact student opportunities and outcomes. This literature review synthesizes existing research on fairness-aware machine learning methodologies, with particular emphasis on their application to educational datasets and the challenges of generalizing fairness frameworks across different domains.

Bias in Training Data and AI Systems

The foundation of algorithmic bias lies predominantly in the training data utilized for machine learning model development. Research has consistently demonstrated that biased datasets serve as the primary source of discriminatory outcomes in AI systems. Training datasets often contain systematic biases that reflect historical inequalities and societal prejudices, creating what researchers term the "bias in and bias out" phenomenon. These biases manifest across multiple dimensions, including gender, race, and socio-economic status, and are particularly problematic when datasets inadequately represent certain demographic groups.

Educational contexts present unique challenges for bias detection and mitigation

Studies examining the UCI Adult Census Income dataset have revealed pronounced gender-based disparities in income predictions, with models exhibiting systematic wage prediction differences when gender attributes are modified. Specifically, empirical evidence demonstrates substantial gender-based wage prediction disparities of $128.60 when switching from male to female attributes, with Kullback-Leibler divergence scores exceeding 0.13 in tree-based models. Similarly, educational AI systems have been shown to produce false negative rates of 19% for Black students and 21% for Latinx students, incorrectly predicting academic failure for significant portions of these populations.

Fairness-Aware Algorithms and Methodologies

The development of fairness-aware machine learning algorithms has emerged as a primary response to bias concerns. These methodologies are typically categorized into three approaches: pre-processing, in-processing, and post-processing techniques. Pre-processing methods focus on data modification to create more balanced datasets, while in-processing approaches incorporate fairness constraints directly into the learning algorithm, and post-processing techniques adjust model outputs to achieve fairness objectives.Recent research has demonstrated that fairness-aware preprocessing techniques, such as the DB-VEA system developed by MIT researchers, can automatically reduce bias through data re-sampling approaches. Similarly, adversarial debiasing methods and counterfactual fairness approaches have shown promise in educational contexts, with studies exploring the causality of sensitive attributes in educational data.

The evaluation of fairness-aware algorithms relies on multiple metrics including demographic parity, equalized odds, and disparate impact measures. Research has shown that different fairness definitions can be contradictory, making it difficult to achieve universal fairness across all demographic groups simultaneously. This challenge is particularly pronounced in educational settings where multiple sensitive attributes may interact in complex ways.

Educational Datasets and Bias Manifestation

Educational datasets present unique challenges for fairness assessment due to their multifaceted nature and the complex interplay of socioeconomic, demographic, and academic factors. The Student Performance dataset and similar educational resources contain variables encompassing school type, family background, gender, and academic outcomes that can serve as sources of bias. Research examining educational AI systems has revealed systematic disparities based on school type and family background, with institutional bias being as prominent as demographic bias. Studies utilizing educational datasets have demonstrated that machine learning models can exhibit lower performance for students from underrepresented demographic groups, particularly in predictive models for academic success and at-risk student identification.

The implementation of fairness-aware techniques in educational contexts has shown mixed results. While counterfactual approaches have demonstrated effectiveness in reducing gender bias on datasets like the UCI Adult dataset, the application to educational datasets requires careful consideration of domain-specific factors and the causality of educational outcomes.

Challenges in Generalizing Fairness Frameworks

A critical gap identified in the literature concerns the difficulty of generalizing fairness frameworks across different domains and contexts. Research has highlighted that fairness metrics and bias mitigation strategies developed for one domain may not transfer effectively to other contexts, particularly when moving from demographic prediction tasks to educational assessment scenarios.The challenge of domain generalization in fairness-aware machine learning has gained attention recently, with researchers developing theoretical frameworks for transferring both accuracy and fairness across domains. Studies have shown that traditional domain generalization methods focus primarily on accuracy transfer while neglecting fairness considerations, leading to potential discrimination when models are deployed in new environments.

Recent work has proposed novel frameworks such as DCFDG (Disentanglement for Counterfactual Fairness-aware Domain Generalization) that attempt to address fairness preservation across changing domains. However, these approaches remain largely theoretical and require extensive empirical validation in educational contexts.

Research Gaps and Future Directions

Despite significant progress in fairness-aware machine learning, several critical research gaps remain. First, there is limited research on fairness in many educational domains, with studies concentrated primarily in general academic performance prediction rather than specialized educational applications. Second, most fairness research focuses on individual sensitive attributes rather than the complex interactions between multiple demographic and socioeconomic factors that characterize real educational environments.The literature reveals a significant disconnect between technical fairness solutions and practical educational applications. While numerous fairness metrics and algorithms have been developed, their integration into real educational systems remains limited, partly due to the complexity of balancing fairness with accuracy in high-stakes educational decisions.

Furthermore, the field lacks comprehensive frameworks for evaluating fairness across different educational contexts and populations. The dominance of group fairness approaches centered on model performance equality limits the consideration of individual fairness and causal relationships in educational outcomes.The synthesis of existing research indicates that while substantial progress has been made in developing fairness-aware machine learning techniques, significant challenges remain in their application to educational contexts and the generalization of fairness frameworks across domains. The complex interplay between socioeconomic factors and educational outcomes, combined with the high-stakes nature of educational decisions, necessitates continued research focused on domain-specific fairness considerations and the development of robust evaluation frameworks for educational AI systems.

4.Datasets Overview

As part of my research on bias and fairness in educational AI systems, I conducted an extensive investigation into publicly available datasets suitable for analyzing socioeconomic and academic factors that impact fairness in machine learning models. Through this research process, I identified and selected two comprehensive datasets: the Adult Census Income dataset and the Student Performance dataset. These datasets were chosen for their relevance, data richness, and capacity to shed light on demographic and academic influences within AI systems applied to education and socioeconomic mobility.

14

4.1 Adult Census Income Dataset

The Adult Census Income dataset, which we identified during my research, is a benchmark dataset widely used for binary income classification tasks. Extracted from the 1994 US Census database, it helps predict whether an individual's annual income exceeds $50,000. The dataset provides a combination of continuous and categorical variables representing a broad spectrum of demographic and socioeconomic characteristics.

11

Key Variables:

4

- age: Continuous; age of the individual.
- workclass: Categorical; type of employment (e.g., Private, Self-emp-notinc).
- fnlwgt: Continuous; final weight estimating the number of people represented by each record for population-level analysis.
- education: Categorical; highest educational attainment (e.g., Bachelors, HS-grad).
- education.num: Continuous; numeric encoding of education level.
- marital.status: Categorical; marital status (e.g., Married-civ-spouse, Never-married).
- occupation: Categorical; occupation type (e.g., Tech-support, Craft-repair).
- relationship: Categorical; familial relationship (e.g., Wife, Husband).

1

- race: Categorical; racial identity of the participant.
- sex: Categorical; gender of the individual.
- capital.gain: Continuous; capital gains from investments.
- capital.loss: Continuous; capital losses from investments.
- hours.per.week: Continuous; weekly work hours.
- native.country: Categorical; country of origin.
- income: Target variable; binary classification indicating income as either ≤$50K or >$50K.

4.2 Student Performance Dataset

Through my research, we also discovered the Student Performance dataset, designed to examine the multifaceted factors influencing academic achievement. This dataset contains detailed demographic, social, and academic information, allowing for the prediction and analysis of student performance outcomes.

Key Variables:

- StudentID: Categorical; unique identifier for each student.
- Age: Continuous; age of the student.
- Gender: Categorical; gender of the student.
- Ethnicity: Categorical; student's self-identified ethnicity.
- ParentalEducation: Categorical; highest education level attained by a parent.
- StudyTimeWeekly: Continuous; number of hours spent studying per week.
- Absences: Continuous; count of school absences.
- Tutoring: Categorical; indicates whether the student receives tutoring.
- ParentalSupport: Categorical; level of parental support reported.
- Extracurricular: Categorical; involvement in extracurricular activities.
- Sports: Categorical; sports participation status.

- Music: Categorical; participation in music activities.
- Volunteering: Categorical; involvement in volunteering.
- GPA: Continuous; student's grade point average.
- GradeClass: Categorical; student's grade level.

By systematically identifying and selecting these datasets, my research establishes a strong foundation for analyzing how socioeconomic and educational factors contribute to bias and fairness challenges in AI-driven educational systems. The insights gained from these rich datasets are vital for conducting a rigorous and comprehensive fairness assessment in my analytical study.

5. Methodology

5.1 Data Preprocessing

The research followed a systematic approach to ensure accuracy in data integration and fairness analysis. Initial exploration revealed that the Adult Census Income and Student Performance datasets contained missing values in both categorical and continuous variables. Categorical gaps, such as those appearing as "?" or "NaN", were addressed using mode imputation techniques to ensure consistent representation and minimal information loss . For continuous features—including age and GPA—mean or median imputation was applied depending on the skewness, thereby preserving the original data distributions.

In preparation for modeling, categorical variables with low cardinality (e.g., gender, race) were one-hot encoded to avoid introducing ordinal relationships. Data inconsistencies, such as variations in string formatting and capitalization, were standardized. High-cardinality variables, like occupation, were target encoded with regularization to reduce overfitting and preserve equitable treatment across categories. Continuous features were standardized using z-score normalization to mitigate scaling issues and improve model robustness.

5.2 Dataset Integration and Analysis

Merging the two datasets presented significant challenges due to structural differences. Variables like "education.num" in the Census dataset and "ParentalEducation" in the student dataset varied in type and granularity. Additionally, the datasets differed in terms of their scope—individual-level versus institutional or classroom-level records.

Where direct integration was impractical, a comparative analysis strategy was adopted. Common features (e.g., education level, gender, ethnicity) were harmonized wherever possible. In the absence of direct record linkage, indirect statistical alignment and categorical mapping were used to facilitate meaningful comparisons across datasets. Each dataset was then individually examined for fairness using sensitive variables, enabling unbiased and domain-relevant interpretation.

5.3 Model Development and Fairness Evaluation

Supervised machine learning models—primarily logistic regression—were trained separately on each dataset to predict key outcomes such as income level and student grade classification. The choice of interpretable models was intentional to allow transparency and accountability in bias audits.

9 The fairness of each model was evaluated using both traditional performance metrics (accuracy, precision, recall) and fairness-specific indicators:

- **Demographic Parity Difference** – Evaluates differences in positive outcome rates between sensitive groups.
- **Disparate Impact Ratio** – Compares favorable prediction ratios across groups.
- **Equalized Odds Difference** – Captures differences in true positive and false positive rates among demographic segments.

The IBM AI Fairness 360 toolkit was employed to calculate these metrics and generate visual insights. Fairness evaluations were visualized through group-wise performance plots and disparity metrics, enabling the identification of underlying biases and informing future mitigation strategies.

6.Schema Level

Data Heterogeneity: The Adult Census dataset represent missing values as”?”, while the Student Performance dataset uses” NaN”. Additionally, categorical variables could exhibit inconsistencies in capitalization or spelling (e.g.,” white” vs.” White”). This requires handling missing values, standardizing categorical variable encodings, and ensuring data consistency across both datasets.

Variable Discrepancy: The” fnlwgt” variable in the Adult Census dataset represents population weights, indicating the number of individuals each row represents in the population. This contrasts with the individual student records in the Student Performance dataset. The” fnlwgt” variable makes it difficult to link individual records directly, as it introduces a population-level weighting factor that is not applicable to individual student data. This requires careful consideration of how to incorporate population-level information without distorting individual-level relationships.

Income to Performance Correlation: Income in the Adult Census dataset is a continuous numerical value, while student GPA is also continuous, and GradeClass is categorical. There is no direct, obvious way to correlate these values, and any attempt to do so requires careful consideration of what metrics to use. This is a data level conflict, as it is a variable that is very hard to match between the two datasets. This will require careful thought about how to treat categorical versus numerical data.

Data Linkage Challenges: The absence of direct identifiers requires reliance on indirect linkage methods, such as geographic matching or statistical matching, which introduce potential inaccuracies in data-level relationships. Temporal Discrepancies: The datasets may represent different time periods, complicating the establishment of causal relationships at the data level.

Temporal Discrepancies: The datasets may represent different time periods, complicating the establishment of causal relationships at the data level.

7. Result and Analysis

Distribution Analysis for Bias Evaluation

For a first approach to analyze potential bias in these two datasets, we computed different plots to analyze distribution of individuals grouped by categories such as Gender, Ethnicity, Income or Education.

Gender Distribution:

The gender distribution analyzed in both datasets reveals differences that may lead to gender bias. Firstly, in the Students Dataset, the gender distribution in Figure 1 is shown with practically no imbalance between males and females, leading to no bias in this aspect. However, a different pattern is observed in the Parents Dataset, where gender representation is mainly represented by the Male class as shown in Figure 2. This gender imbalance can potentially impact the accuracy of predictive models, as models trained on gender-imbalanced data might favor one gender over the other.

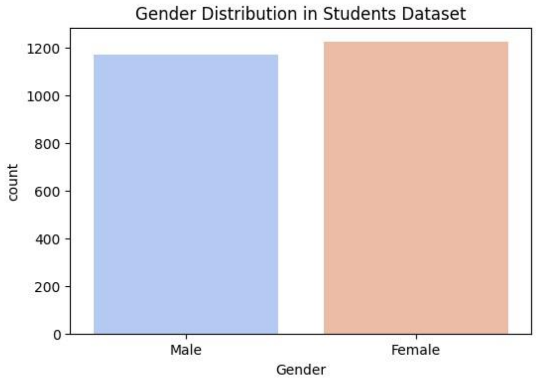


Fig1. Gender Distribution in Students Dataset

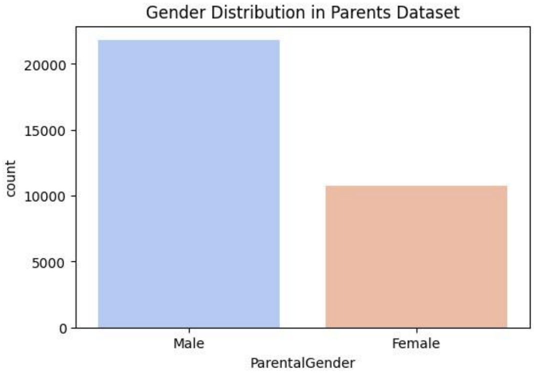


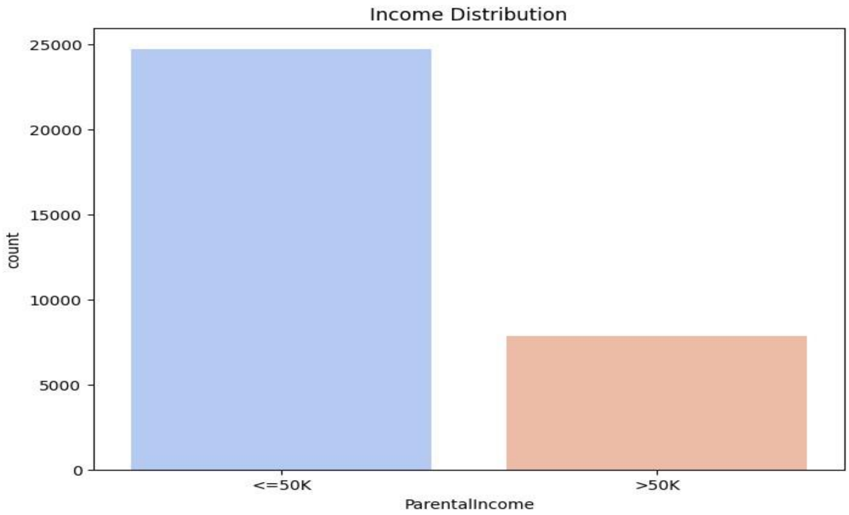
Fig2. Gender Distribution in Parents Dataset

Ethnicity Distribution

The ethnicity distribution analysis in the *Students Dataset* shows a clear demographic breakdown across multiple ethnicities such as Caucasian, African American, Asian, and Other. However, as presented in Figure **??**, the dataset still show over representation of the *Caucasian* class. For the case of the *Parents Dataset*, Figure **??** shows a hicher bias related with the over representation of the *Caucasian* class. If any particular ethnicity is significantly over represented with respect to the others, the model may struggle to generalize fairly across all groups, potentially introducing ethnic bias into predictions or decisions based on ethnicity.

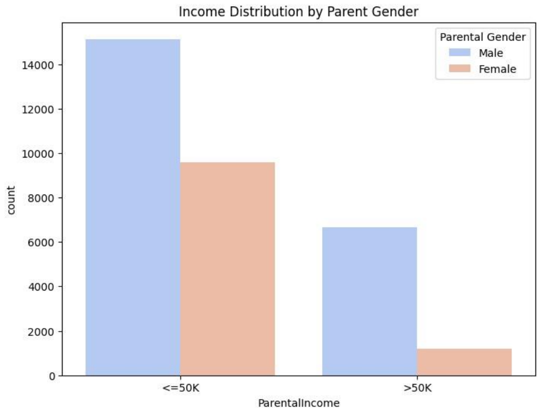
Income Distribution

The *Income Distribution* analysis of parents in the *Parents Dataset* reveals a predominance of the ≤*50k* class as shown in Figure 3.



The breakdown of parental income across various categories could indicate that certain income groups are overrepresented or underrepresented, especially when segmented by parental gender or ethnicity. In the case of *Gender*, Figure 4 reveals a greater underrepresentation in the ≥*50k* income class. Despite the presence of gender bias, the disparity in gender representation is more pronounced in the higher income class compared to the ≤*50k* class.

Additionally, with regard to *Ethnicity*, Figure 5 portraits a significant overrepresentation of *Caucasian* individuals in both income classes, particularly in the ≤*50k* class. However, there appears to be a slight underrepresentation of *African American* individuals in the higher income class, as their presence is comparable to that of *Asian* or *Other* ethnicities even though there are more *African American* individuals in the dataset than the other two. This suggests a potential bias towards lower income levels for *African American* individuals, despite their higher overall representation.



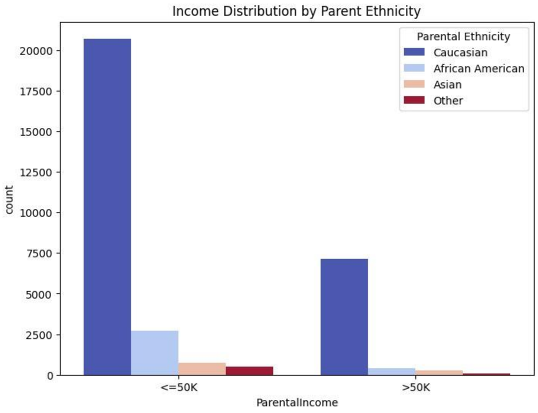


Fig4. Income Ditsrubution by Parent Gender

Fig5. Income Distribution by Parent ethnicity

Furthermore, in the case of *Education*, Figure 6 clearly illustrates that individuals with lower or no education are predominantly found in the lower income class, while a significant proportion of individuals with a Bachelor's degree or higher are classified in the *≥50k* income class.

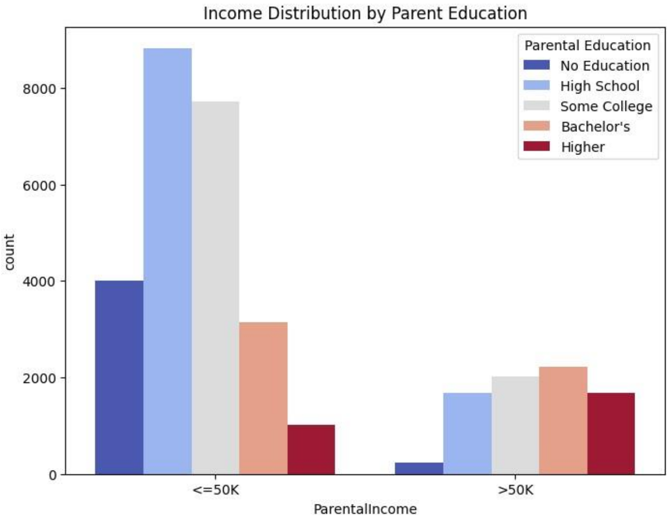


Fig7. Income Ditsrubution by Parent Education

Parent Education Distribution

The *Parent Education Distribution* analysis of parents in the *Students Dataset* reveals a predominance of *High School* and *Some College* classes as shown in Figure 8.

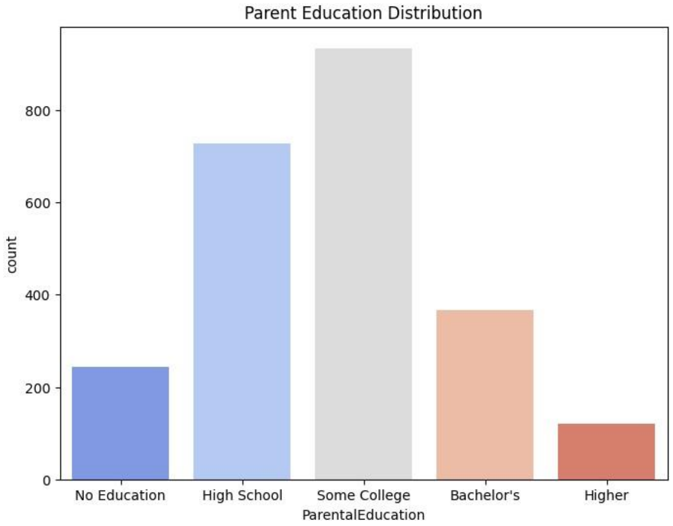
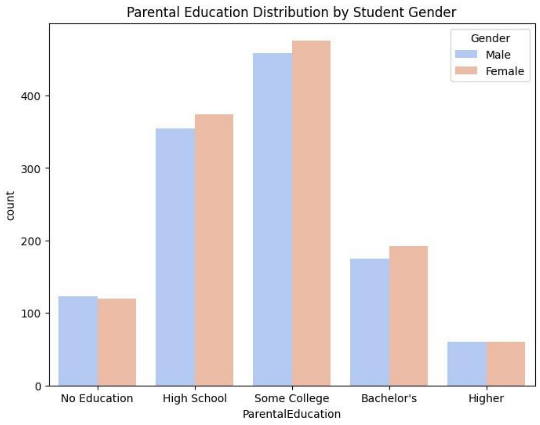


Fig8. Parent Education System

The *Parent Education Distribution* analysis by *Student Gender* and *Student Ethnicity* shows the following trends. In Figure 8, the distribution is fairly balanced between male and female students, though a slight tendency towards higher parental education levels in *Female* students can be observed.

Figure 9 reveals more significant disparities, with *African American* students being underrepresented in higher parental education categories, especially *Higher* education which refers to *Master* and *Doctorate*. In contrast, *Asian* students tend to have parents with *Some College* education level. Aditionally, for the case of *Caucasia* students, their parent education is mainly classified in *High School* and *Some College*. These ethnic differences could introduce bias in the model, potentially disadvantaging students from ethnic groups with lower parental education levels.



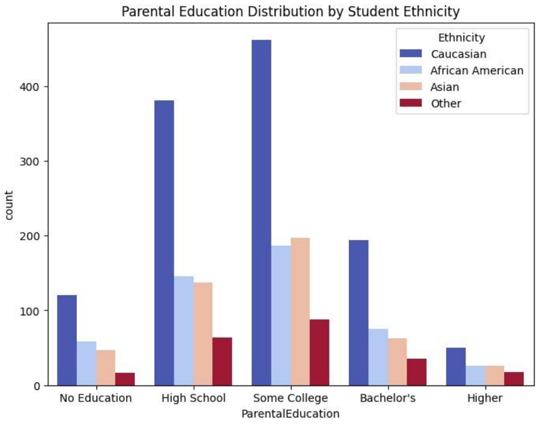


Fig9. Parental Education Distribution by Student Gender

Fig10. Parental Education Distribution by Student Ethnicity

8. Bias Mitigation & Fairness

12

To evaluate the fairness of grade predictions in our student dataset, we used the library *IBM AI Fairness 360 (AIF360)* in a *Python* implementation. Our analysis focused on fairness metrics, including *demographic parity*, *equalized odds*, and *disparate impact*. To develop this assessment, we employed a *Logistic Regression* model to predict *Grades* and compared the predictions against the actual scores in the dataset.

Fairness Evaluation by Gender

10

To assess gender fairness, we analyzed multiple fairness metrics, including *Demographic Parity Difference*, *Disparate Impact Ratio*, *Equalized Odds Difference*, and *False Negative Rate Difference*.

Our results indicate a *Demographic Parity Difference* of **-0.0102**, suggesting only a minimal imbalance in the likelihood of receiving high grades between genders. Additionally, the *Disparate Impact Ratio* of **0.9373** is close to the ideal value of 1.0, meaning that both genders have nearly equal probabilities of obtaining high grades.

Furthermore, the *Equalized Odds Difference* of **0.0099** indicates that the model maintains relatively consistent performance across genders. The *False Negative Rate Difference* of **-0.0009** is nearly zero, suggesting that the likelihood of misclassifying a high-performing student as lowperforming is not significantly different between genders.

Despite these favorable results, Figures 10 and 11 offer additional insights into the distribution of *GPA* and *Grade* across genders.

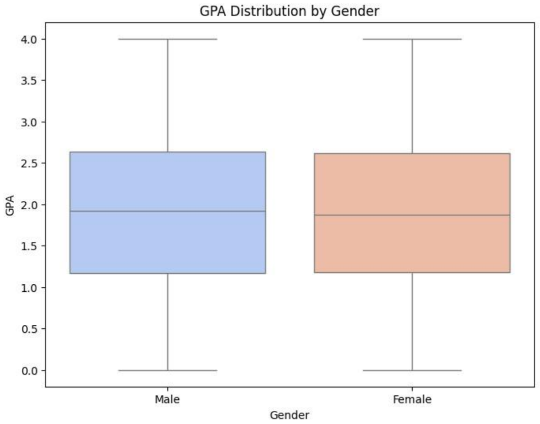


Fig11. GPA Distribution by Gender

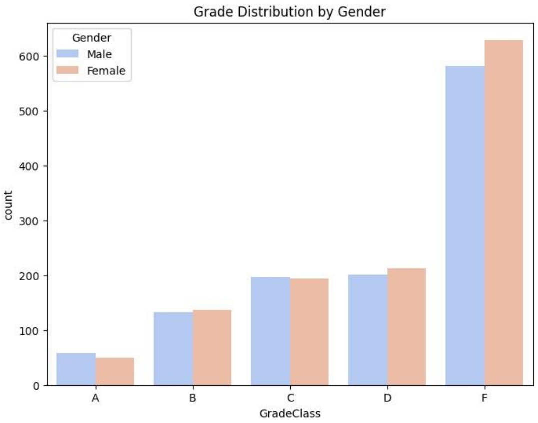


Fig12. Grade Distribution by Gender

Fairness Evaluation by Ethnicity

To evaluate fairness across ethnic groups, we also analyzed the fairness metrics explained previously. Our results show a *Demographic Parity Difference* of **0.0105**, indicating a slight imbalance in the likelihood of receiving high grades among different ethnicities. However, the *Disparate Impact Ratio* of **1.0696** is close to the ideal value of 1.0, suggesting that no single ethnic group is disproportionately advantaged or disadvantaged in high-grade predictions.

Despite this, the *Equalized Odds Difference* of **0.0838** and the *False Negative Rate Difference* of **0.0838** highlight a more significant disparity in model performance across ethnic groups. A higher False Negative Rate for some ethnicities suggests that certain groups are more likely to be misclassified as low-performing, which may indicate underlying bias in the model's decision-making process.

To further analysis, Figures 12 and 13 provide visual insights into the distribution of *GPA* and *Grade* across ethnic groups.

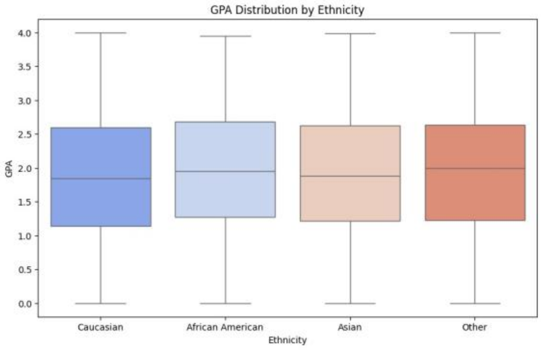


Fig13. GPA Distribution by Ethnicity

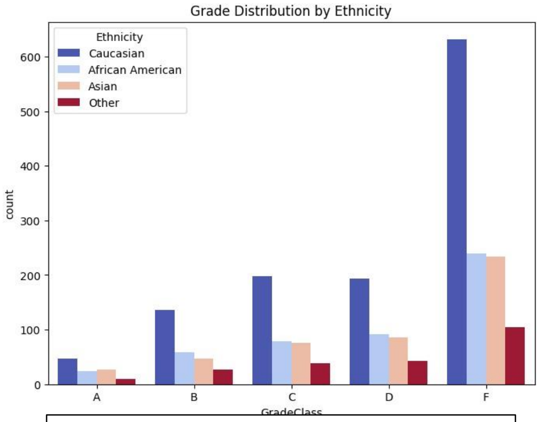


Fig14. Grade Distribution by Ethnicity

Fairness Evaluation by Parent Education

Similar fairness metrics were applied to evaluate fairness by *Parent Education*. Our results show a *Demographic Parity Difference* of **0.0609**, indicating a higher imbalance in fairness among classes than the case of *Gender* or *Ethnicicty*. The *Disparate Impact Ratio* of **1.6646** suggests a significant disparity, meaning that students from higher parental education backgrounds are slightly favored in high-grade predictions compared to those with lower parental education.

Furthermore, the *Equalized Odds Difference* of **0.0588** and the *False Negative Rate Difference* of **-0.0588** indicate that students from lower parental education backgrounds may face a higher rate of misclassification into lower grades. This suggests that parental education may introduce systematic bias.

In addition to the metric, Figures 14 and 15 provide visual insights into the distribution of *GPA* and *Grade* across different parental education levels.

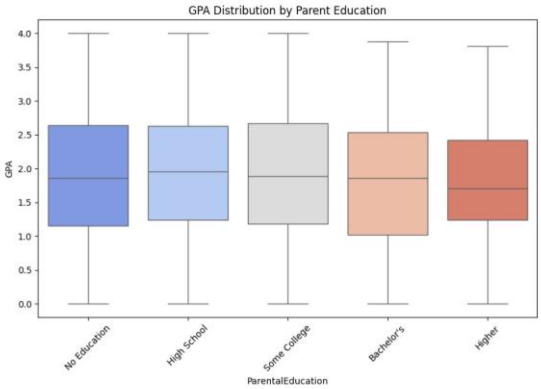


Fig14. GPA Distribution by Parent Education

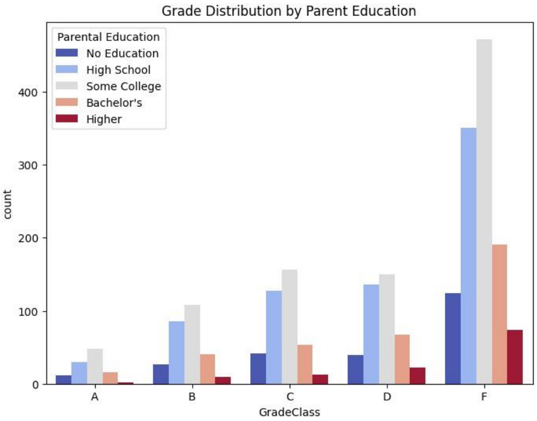


Fig15. Grade Distribution by Parent Education

9.Conclusion

This research aimed to uncover and understand the presence of bias in educational prediction systems using the Adult Census Income and Student Performance datasets. Through comprehensive analysis and the application of fairness metrics via the IBM AI Fairness 360 toolkit, the study identified varying degrees of unfairness across different demographic attributes, with the most notable disparities observed in ethnicity and parental education levels. In contrast, gender-related bias appeared to be relatively minimal.

While the use of fairness evaluation tools added structure and credibility to the process, challenges arose in managing class imbalance and in interpreting complex fairness metrics like Equalized Odds and False Negative Rate Difference. Additionally, the selection of appropriate thresholds often required iterative tuning and careful judgment.

Despite these complexities, the findings reaffirm the critical need for fairness assessments in machine learning workflows—especially in domains like education, where biased decisions can have long-term consequences. This project emphasizes that accuracy alone is not sufficient; equitable outcomes must be a key performance goal in AI systems.

Looking ahead, future work can explore integrating advanced bias mitigation strategies such as reweighting techniques, adversarial training, or embedding fairness-aware algorithms during model development. Such approaches can contribute to building AI systems that not only perform well but also promote inclusivity and trust across diverse learner populations.

10.References

[1] Moreno-Monroy, A. I. et al. *Handling Missing Data in Education Research*, JMIR Med Inform,2021.

Pedregosa, F. et al. *Scikit-learn: Preprocessing Module*, Journal of Machine Learning Research,2011.

KeyLabs. *Merging Multiple Labeled Datasets: Ensuring Consistency and Quality*, Technical Report, 2023.

Datasets:

Rabie ElKharoua, “Students Performance Dataset“ https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset/data

UCI Machine Learning, “Adult Census Income Dataset“ https://www.kaggle.com/datasets/uciml/ adult-census-income/data

IBM Research, “AI Fairness 360 Toolkit“ https://ai-fairness-360.org