Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Information retrieval, extraction and integration

Assignment 4: Data Integration and Bias

Authors:

José Antonio Ruiz Heredia

Joseph Tartivel

Álvaro Honrubia Genilloud

Teacher:

Mari Carmen Suárez de Figueroa Baonza

Date:
April 2, 2025

Contents

1 Introduction

This report investigates the feasibility and challenges of combining two rich datasets: the Adult Census Income dataset and the Student Performance Data dataset. The Adult Census Income dataset, offers a comprehensive view of adult demographic and employment characteristics, with a particular focus on income prediction. Moreover, the Student Performance Data dataset, provides a granular perspective on factors influencing student academic achievement, encompassing demographic, social, and academic variables.

The core motivation for merging these datasets is to explore the nuanced relationship between adult socioeconomic factors and student educational outcomes. By linking parental income, education levels, and occupation with student GPA and grade class, we aim to understand the extent to which economic disparities translate into educational inequalities. This analysis has the potential to reveal critical insights into the mechanisms through which socioeconomic factors shape student academic trajectories, informing the development of targeted interventions and policies aimed at promoting educational equity and social mobility.

2 Datasets

2.1 Adult Census Income Dataset

This dataset is primarily designed for binary classification, predicting whether an individual's income exceeds $50,000 per year. The dataset was extracted from the 1994 US Census database.

Detailed Variables:

- **age:** A continuous variable representing the individual's age, potentially revealing age-related income trends.
- **workclass:** A categorical variable denoting employment type (e.g., Private, Self-emp-not-inc), indicating the nature of work.
- **fnlwgt:** A continuous variable, representing the estimated number of individuals in the population that each record represents. It must be considered during analysis that attempts to reflect the population.
- **education:** A categorical variable representing the highest level of education achieved (e.g., Bachelors, HS-grad).
- **education.num:** A continuous variable, numerical encoding of educational attainment, revealing the impact of education on income.
- **marital.status:** A categorical variable representing marital status (e.g., Married-civ-spouse, Never-married).
- **occupation:** A categorical variable representing the individual's occupation (e.g., Tech-support, Craft-repair).
- **relationship:** A categorical variable representing the individual's relationship in their family (e.g., Wife, Husband).
- **race:** A categorical variable representing the individual's race (e.g., White, Black).
- **sex:** A categorical variable representing the individual's gender.
- **capital.gain:** A continuous variable reflecting capital gains.

- **capital.loss:** A continuous variable reflecting capital losses.
- **hours.per.week:** A continuous variable indicating work hours, potentially influencing income levels.
- **native.country:** A categorical variable representing the individual's country of origin.
- **income:** The target variable, a binary classification (¡=50K, ¿50K), indicating income level.

2.2 Student Performance Data Dataset

This dataset aims to analyze the multifaceted factors influencing student academic performance. The dataset contains student information, and information about the students parents, with the goal of being able to predict the students performance.

Detailed Variables:

- **StudentID:** A categorical variable, a unique identifier for each student.
- **Age:** A continuous variable, representing the student's age.
- **Gender:** A categorical variable, representing the student's gender.
- **Ethnicity:** A categorical variable, representing the student's ethnicity.
- **ParentalEducation:** A categorical variable representing the highest level of education attained by the student's parents.
- **StudyTimeWeekly:** A continuous variable, detailing weekly study hours.
- **Absences:** A continuous variable, detailing the number of absences.
- **Tutoring:** A categorical variable, indicating tutoring status.
- **ParentalSupport:** A categorical variable, indicating parental support level.
- **Extracurricular:** A categorical variable, indicating extracurricular participation.
- **Sports:** A categorical variable, indicating sports participation.
- **Music:** A categorical variable, indicating music involvement.
- **Volunteering:** A categorical variable, indicating volunteering activity.
- **GPA:** A continuous variable, representing the student's grade point average.
- **GradeClass:** A categorical variable, representing the student's grade level.

3 Conflicts

The process of merging the Adult Census Income and Student Performance Data datasets, while promising, presents a series of significant challenges. These challenges stem from the inherent differences in the datasets' structure, content, and purpose. The conflicts can be broadly categorized into issues of data heterogeneity, linkage difficulties, temporal discrepancies, and ethical concerns

3.1 Data-level

Variable Name and Definition Inconsistencies:

The "education" column in the Adult Census dataset might categorize educational attainment as "HS-grad," "Bachelors," and "Masters," reflecting formal U.S. educational stages. Conversely, the "ParentalEducation" column in the Student Performance dataset might use terms like "high school," "college," and "graduate degree," which, while semantically similar, lack precise alignment. This discrepancy requires a standardized classification system, involving mapping both sets of categories to a common educational level hierarchy.

Data Type Mismatches:

The "education.num" column in the Adult Census data provides a numerical representation of the "education" column, facilitating quantitative analysis. However, the Student Performance dataset lacks a direct numerical equivalent for "ParentalEducation." Furthermore, binary responses like "Yes/No" are represented as booleans in the Student Performance dataset, but might be strings in the Adult Census dataset, such as "Yes" or "No". Data type mismatches can lead to errors during data integration and analysis. Aligning data types requires transforming variables to a common format, ensuring compatibility for subsequent analytical operations.

Structural Differences:

The Adult Census dataset is structured around individual-level income prediction, with each row representing an adult and their associated demographic and employment data. In contrast, the Student Performance dataset is organized to analyze factors influencing student academic performance, with each row representing a student and their corresponding academic, social, and demographic information. This difference in purpose leads to distinct table structures, making direct merging challenging. The structural differences between datasets would require careful consideration of how to join or combine them. This may involve reshaping the data, creating aggregate features, or using different analytical approaches that can accommodate the varying structures.

Variable Presence/Absence:

The Adult Census dataset includes variables like "workclass" (e.g., Private, Self-emp-not-inc) and "occupation" (e.g., Tech-support, Craft-repair), which are not present in the Student Performance dataset. Conversely, variables like "StudyTimeWeekly" and "Absences" are unique to the student data. The presence of unique variables in each dataset requires decisions on whether to exclude them from the merged dataset, create proxy variables to approximate their information, or use analytical techniques that can handle the varying feature sets.

3.2 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 con-

sideration 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.

4 Distribution Analysis for Bias Evaluation

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

4.1 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.

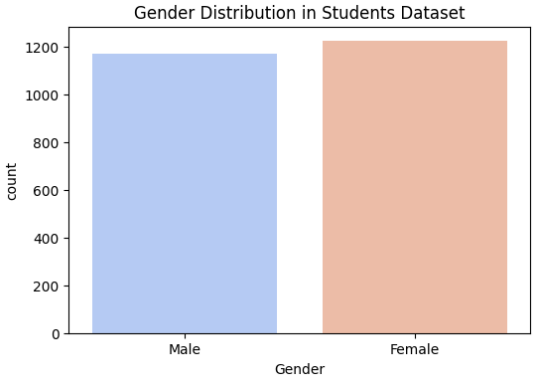

Figure 1: Gender Distribution in Students Dataset

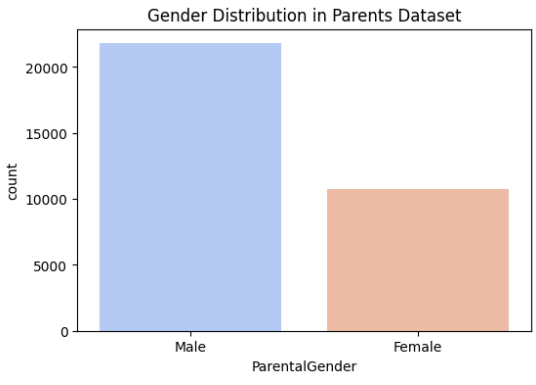

Figure 2: Gender Distribution in Parent Dataset

4.2 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.

4.3 Income Distribution

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

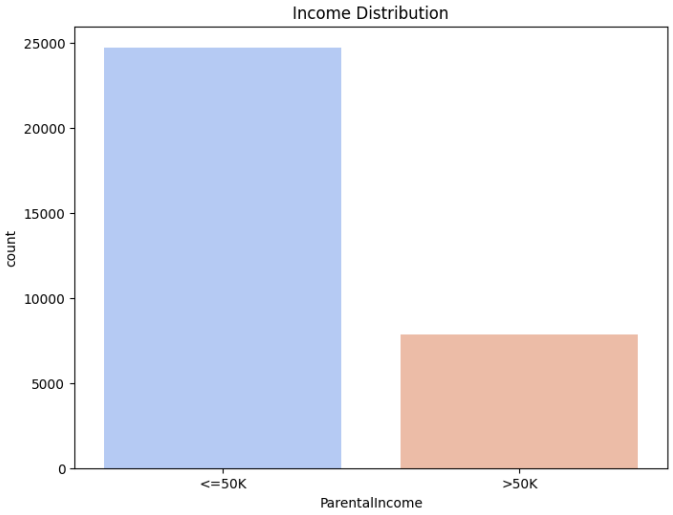

Figure 3: Income Distribution

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 $\geq 50k$ income class. Despite the presence of gender bias, the disparity in gender representation is more pronounced in the higher income class compared to the $\leq 50k$ class.

Additionally, with regard to *Ethnicity*, Figure 5 portraits a significant overrepresentation of *Caucasian* individuals in both income classes, particularly in the $\leq 50k$ 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.

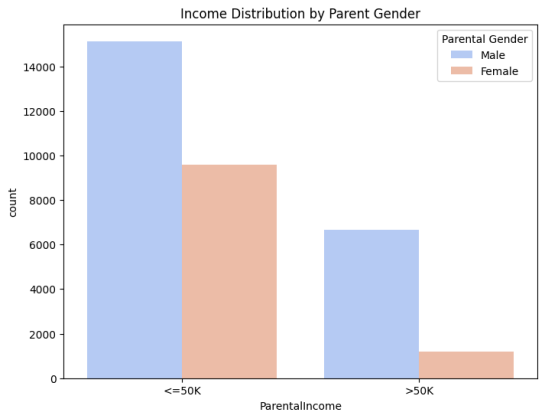

Figure 4: Income Distribution by Parent Gender

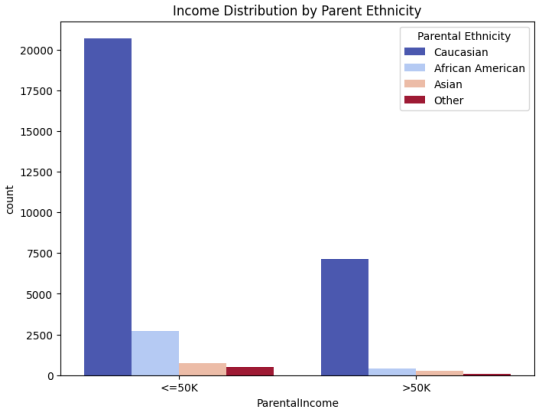

Figure 5: Income Distribution by Parent Ethnicity

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 $\geq 50k$ income class.

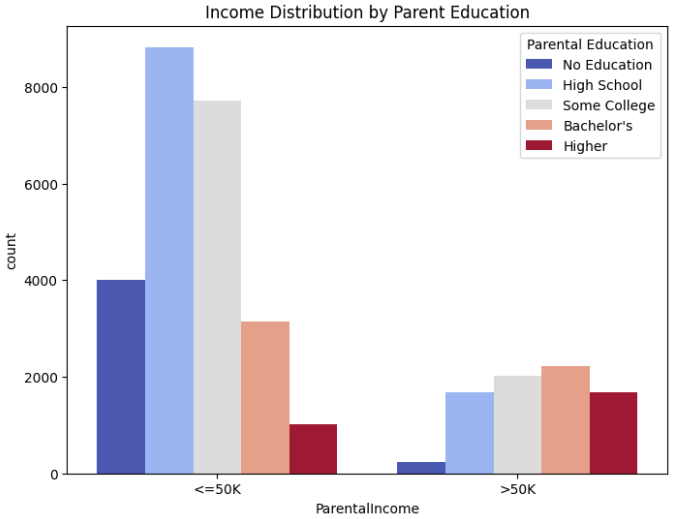

Figure 6: Income Distribution by Parent Education

4.4 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 7.

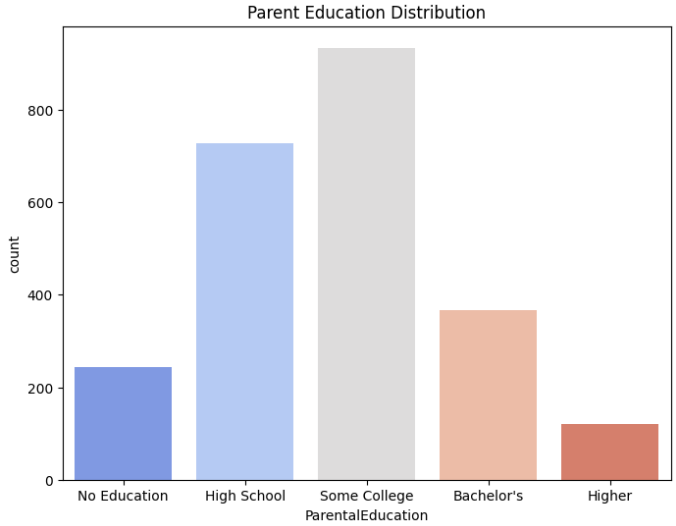

Figure 7: Parent Education Distribution

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.

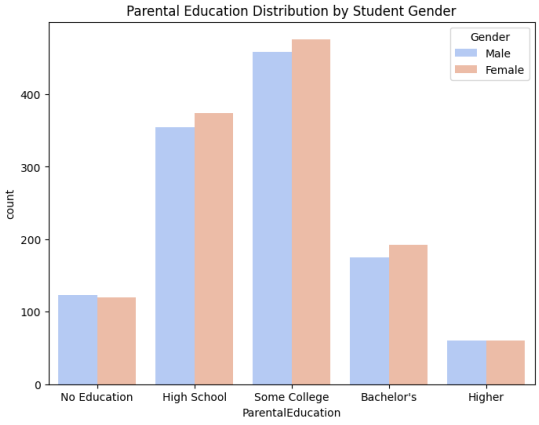

Figure 8: Parent Education Distribution by Gender

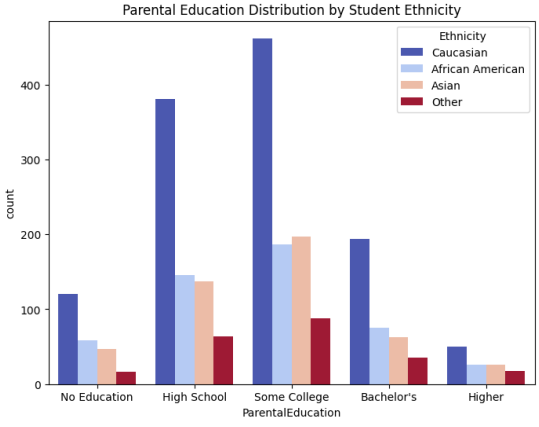

Figure 9: Parent Education Distribution by Ethnicity

5 Bias mitigation & Fairness

To evaluate the fairness of grade predictions in our student dataset, we used the library *IBM AI Fairness 360 (AIF360)* [3] 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.

5.1 Fairness Evaluation by Gender

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 low-performing 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.

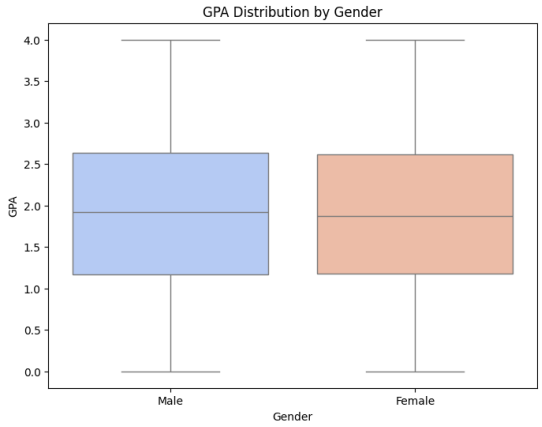

Figure 10: GPA Distribution by Gender

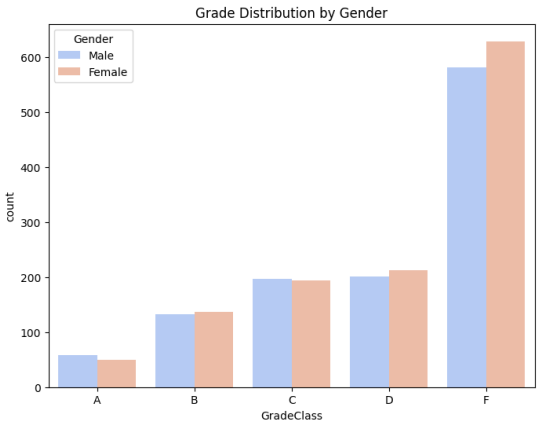

Figure 11: Grade Distribution by Gender

5.2 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.

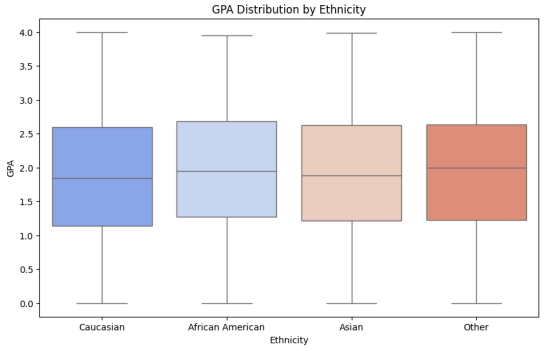

Figure 12: GPA Distribution by Ethnicity

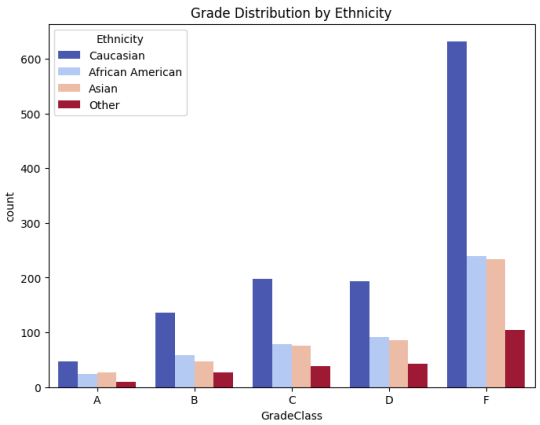

Figure 13: Grade Distribution by Ethnicity

5.3 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.

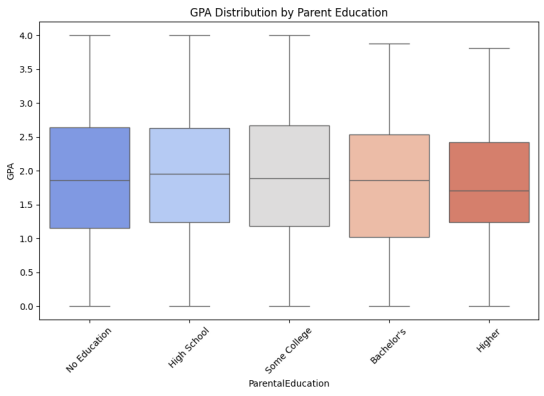

Figure 14: GPA Distribution by Parent Education

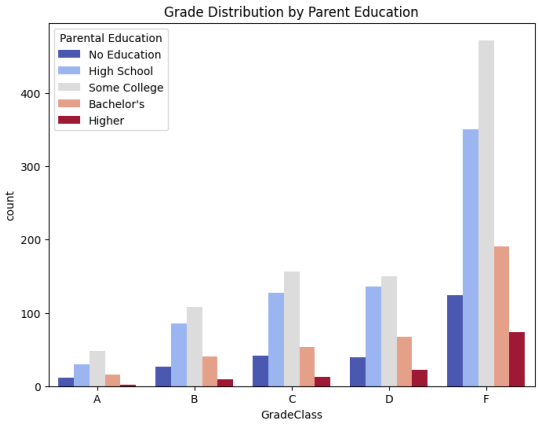

Figure 15: Grade Distribution by Parent Education

6 Conclusion

In this study we evaluated bias and fairness in two datasets by analyzing demographic distributions and applying fairness metrics using *IBM AI Fairness 360*. The results revealed varying levels of bias in gender, ethnicity, and parental education. While gender fairness metrics indicated minimal disparities, ethnicity and parental education showed more significant imbalances, particularly in *Equalized Odds Difference* and *False Negative Rate Difference*.

Visualizing distributions presented challenges, including handling imbalanced data representation and interpreting fairness metrics effectively. The use of *AI Fairness 360* also posed difficulties in selecting appropriate thresholds and understanding metrics. Despite these challenges, the study highlights the importance of fairness to mitigate potential biases and improve equitable outcomes in predictive analytics. Future work could explore bias mitigation strategies such as *reweighting*, *adversarial debiasing*, and *fairness-aware* algorithms to ensure fairer predictions across demographic groups.

References

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

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

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