Algorithmic Fairness and Temporal Generalization in Early Childhood Risk Prediction: Evidence from the ECLS-K:2011 Longitudinal Study

Research Analysis Report

Machine Learning for Educational Equity

February 2026

Abstract

Machine learning models are increasingly deployed in educational settings to identify students at risk of academic difficulties. However, concerns about algorithmic fairness—whether these models perform equitably across demographic groups—remain underexplored in longitudinal educational contexts. This study examines the fairness properties of predictive models trained on early childhood data (kindergarten through 2nd grade) to predict 5th-grade academic risk using the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) public-use data. We trained and evaluated four machine learning algorithms (logistic regression, elastic net, random forest, and XGBoost) on a sample of 9,104 children with complete data. Our best-performing model (elastic net, AUC = 0.848) demonstrated significant disparities in true positive rates across racial/ethnic groups, with the model detecting at-risk status for Hispanic students (TPR = 0.410) at more than twice the rate of White students (TPR = 0.172). While the model passed equal opportunity and statistical parity criteria, it failed equalized odds due to substantial false positive rate disparities. We implemented threshold optimization as a bias mitigation strategy, achieving more equitable true positive rates at the cost of reduced overall accuracy. Our findings highlight the importance of fairness audits in educational AI systems and the trade-offs inherent in bias mitigation approaches. We discuss implications for the responsible deployment of predictive analytics in K-12 education.

Keywords: algorithmic fairness, machine learning, educational prediction, early childhood, ECLS-K:2011, bias mitigation

Contents

1 Introduction

The application of machine learning (ML) in educational settings has grown substantially over the past decade, with predictive models increasingly used to identify students at risk of academic failure, dropout, or other adverse outcomes (Baker & Inventado, 2014). These early warning systems (EWS) promise to enable timely interventions that could improve educational trajectories, particularly for disadvantaged students. However, the deployment of algorithmic decision-making tools in education raises critical questions about fairness and equity.

Algorithmic fairness—the study of how automated systems may systematically advantage or disadvantage particular groups—has emerged as a central concern in machine learning research (Mehrabi et al., 2021). In educational contexts, unfair algorithms could perpetuate or amplify existing inequities by systematically under-identifying at-risk students from certain demographic groups or by disproportionately flagging students from marginalized communities for intervention.

This study addresses three primary research questions:

1. **RQ1:** How accurately can early childhood cognitive and behavioral measures predict 5th-grade academic risk?
2. **RQ2:** Do predictive models exhibit differential performance across racial/ethnic and socioeconomic groups?
3. **RQ3:** Can post-hoc bias mitigation strategies reduce fairness disparities while maintaining acceptable predictive performance?

We leverage the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), a nationally representative longitudinal study that followed children from kindergarten entry through 5th grade. This dataset provides a unique opportunity to examine both the predictive validity and fairness properties of models that use early childhood data to forecast later academic outcomes.

1.1 Contributions

This study makes several contributions to the literature on algorithmic fairness in education:

- We provide one of the first comprehensive fairness audits of longitudinal educational prediction models using nationally representative data.
- We examine temporal generalization—whether models trained on early grades accurately predict outcomes years later—through a fairness lens.
- We evaluate multiple fairness criteria (equal opportunity, equalized odds, statistical parity) and demonstrate how models may satisfy some criteria while violating others.
- We implement and evaluate threshold optimization as a bias mitigation strategy, quantifying the accuracy-fairness trade-off.

2 Background and Related Work

2.1 Early Warning Systems in Education

Early warning systems (EWS) use student data to identify individuals at risk of negative academic outcomes. Traditional EWS relied on simple indicators such as attendance, behavior, and course performance (the "ABC" indicators). Modern approaches increasingly incorporate machine learning algorithms capable of processing larger feature sets and capturing nonlinear relationships (Lakkaraju et al., 2015).

Research has demonstrated that ML-based EWS can achieve reasonable predictive accuracy, with AUC values typically ranging from 0.70 to 0.85 depending on the outcome and available features (Aguiar et al., 2015). However, fewer studies have examined whether these systems perform equitably across student subgroups.

2.2 Algorithmic Fairness

The machine learning fairness literature has developed numerous formal definitions of fairness, which can be broadly categorized into three families (Verma & Rubin, 2018):

Group fairness criteria require that some statistical measure be equal across protected groups. Key definitions include:

- *Demographic parity* (statistical parity): The proportion of positive predictions should be equal across groups.
- *Equal opportunity*: True positive rates should be equal across groups.
- *Equalized odds*: Both true positive rates and false positive rates should be equal across groups.

Individual fairness requires that similar individuals receive similar predictions, regardless of group membership.

Counterfactual fairness asks whether an individual's prediction would change if their protected attribute were different.

Importantly, researchers have proven that certain fairness criteria are mathematically incompatible, meaning it is generally impossible to satisfy all criteria simultaneously (Chouldechova, 2017; Kleinberg et al., 2016).

2.3 Fairness in Educational AI

A growing body of work has examined fairness in educational technology. Kizilcec & Lee (2022) found that dropout prediction models in MOOCs exhibited significant performance disparities across countries. Yu et al. (2020) demonstrated that automated essay scoring systems showed bias against non-native English speakers. Gardner et al. (2019) examined fairness in course outcome prediction and found persistent gaps across demographic groups.

Despite this emerging literature, few studies have examined fairness in early childhood prediction contexts or in systems that make predictions across extended time horizons. Our study addresses this gap.

3 Data and Methods

3.1 Data Source

We used data from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), conducted by the National Center for Education Statistics (NCES). The ECLS-K:2011 is a nationally representative longitudinal study that followed approximately 18,000 children from kindergarten entry in fall 2010 through spring of 5th grade in 2016.

Data were collected across nine waves:

- Kindergarten: Fall 2010 (Wave 1), Spring 2011 (Wave 2)
- 1st Grade: Fall 2011 (Wave 3), Spring 2012 (Wave 4)
- 2nd Grade: Fall 2012 (Wave 5), Spring 2013 (Wave 6)
- 3rd Grade: Spring 2014 (Wave 7)
- 4th Grade: Spring 2015 (Wave 8)
- 5th Grade: Spring 2016 (Wave 9)

We used the public-use data file, which includes 18,174 children. After applying inclusion criteria (valid outcome data and at least some baseline predictors), our analytic sample comprised 18,151 children. Complete-case analysis for modeling yielded 9,104 children with data on all predictors and outcomes.

3.2 Measures

3.2.1 Outcome Variable

The outcome was **academic risk in 5th grade**, operationalized as scoring below the 25th percentile on the reading theta score (X9RTHETA) from the spring 2016 assessment. The reading assessment measured skills including basic reading, vocabulary, and reading comprehension. The theta score is an IRT-based ability estimate that allows for longitudinal comparisons.

In our sample, 15.7% of children were classified as at-risk based on this threshold.

3.2.2 Predictor Variables

We included predictors from kindergarten through 2nd grade across four domains:

Baseline Cognitive Scores:

- Reading theta scores: Fall K (X1RTHETK), Spring K (X2RTHETK)
- Math theta scores: Fall K (X1MTHETK), Spring K (X2MTHETK)

Executive Function:

- Dimensional Change Card Sort score, Spring 2013 (X6DCCSSCR)

Approaches to Learning:

- Teacher-reported approaches to learning: Fall K (X1TCHAPP), Spring K (X2TCHAPP), Spring 1st grade (X4TCHAPP)

Demographic Characteristics:

- Child sex (X_CHSEX_R)
- Race/ethnicity (X_RACETH_R)
- Socioeconomic status quintile (X1SESQ5)
- Home language (X12LANGST)

3.2.3 Protected Attributes

For fairness analysis, we focused on **race/ethnicity** as the primary protected attribute. The ECLS-K:2011 includes seven race/ethnicity categories; we collapsed these into five groups: White (reference), Black, Hispanic, Asian, and Other (including Native Hawaiian/Pacific Islander, American Indian/Alaska Native, and multiracial).

3.3 Machine Learning Models

We trained four classification algorithms:

1. **Logistic Regression:** L2-regularized logistic regression with regularization strength selected via cross-validation from $C \in \{0.01, 0.1, 1.0, 10.0\}$.

2. **Elastic Net:** Logistic regression with elastic net penalty, tuning both $\alpha \in \{0.001, 0.01, 0.1, 1.0\}$ and L1 ratio $\in \{0.2, 0.5, 0.8\}$.

3. **Random Forest:** Ensemble of decision trees with hyperparameters: $n\_estimators \in \{100, 200\}$, $max\_depth \in \{5, 10, 15\}$, $min\_samples\_leaf \in \{5, 10\}$.

4. **XGBoost:** Gradient boosted trees with $n\_estimators \in \{100, 200\}$, $max\_depth \in \{3, 5, 7\}$, $learning\_rate \in \{0.01, 0.1\}$.

All models were trained using 5-fold stratified cross-validation for hyperparameter selection, with random seed fixed at 42 for reproducibility. The data were split 70% training, 30% test.

3.4 Evaluation Metrics

3.4.1 Predictive Performance

We evaluated predictive performance using:

- Area Under the ROC Curve (AUC-ROC)

- Accuracy

- Precision (Positive Predictive Value)

- Recall (Sensitivity/True Positive Rate)

- F1 Score

- Brier Score

- Specificity

3.4.2 Fairness Metrics

For each demographic group g, we computed:

- **True Positive Rate (TPR):** $TPR_g = \frac{TP_g}{TP_g+FN_g}$
- **False Positive Rate (FPR):** $FPR_g = \frac{FP_g}{FP_g+TN_g}$
- **Positive Predictive Value (PPV):** $PPV_g = \frac{TP_g}{TP_g+FP_g}$
- **Positive Rate:** Proportion predicted positive

We assessed three fairness criteria:

Equal Opportunity: Satisfied if TPR ratios between groups exceed 0.80 (four-fifths rule).

Equalized Odds: Satisfied if both TPR and FPR ratios exceed 0.80.

Statistical Parity: Satisfied if positive rate ratios exceed 0.80.

3.5 Bias Mitigation

We implemented **threshold optimization** as a post-processing bias mitigation strategy. Rather than using a single decision threshold (typically 0.5) for all groups, we selected group-specific thresholds to equalize true positive rates across groups. The target TPR was set to the overall TPR of the best-performing model.

4 Results

4.1 Sample Characteristics

Table 1 presents the demographic characteristics of the analytic sample.

Table 1: Sample Characteristics (N = 18,151)

Characteristic	**N**	**%**
Race/Ethnicity		
White	8,476	46.7
Hispanic	4,206	23.2
Black	2,394	13.2
Other	1,825	10.1
Asian	380	2.1
Missing	870	4.8
SES Quintile		
Q1 (Lowest)	3,224	17.8
Q2	3,214	17.7
Q3	3,217	17.7
Q4	3,227	17.8
Q5 (Highest)	3,206	17.7
Missing	2,063	11.4
Sex		
Male	9,273	51.1
Female	8,840	48.7
5th Grade Reading Risk		
At-Risk (<25th %ile)	2,857	15.7
Not At-Risk	15,294	84.3

The sample is demographically diverse, with substantial representation of historically underserved groups. Approximately 15.7% of children were classified as at-risk in reading by 5th grade.

4.2 Model Performance

Table 2 presents the predictive performance of all four models on the held-out test set (N = 2,732).

Table 2: Model Performance on Test Set

Model	AUC	Accuracy	Precision	Recall	F1	Brier
Elastic Net	**0.848**	**0.850**	0.657	0.294	0.406	**0.108**
Logistic Regression	0.847	0.849	0.657	0.281	0.394	0.108
Random Forest	0.841	0.848	0.645	0.285	0.395	0.109
XGBoost	0.840	0.846	0.618	0.302	0.406	0.109

All models achieved similar performance, with AUC values ranging from 0.840 to 0.848. The elastic net model achieved the highest AUC (0.848) and was selected for subsequent fairness analysis. Cross-validation yielded consistent results (CV AUC = 0.842), suggesting good generalization.

The relatively low recall (0.294) reflects the class imbalance and conservative default threshold; the model achieves high specificity (0.968) at the expense of sensitivity.

4.3 Feature Importance

Table 3 presents the feature importance coefficients from the elastic net model.

Table 3: Feature Importance (Elastic Net Coefficients)

Feature	**Coefficient**
Spring K Math (X2MTHETK)	0.501
Spring K Reading (X2RTHETK)	0.350
SES Quintile (X1SESQ5)	0.303
Fall K Math (X1MTHETK)	0.288
Approaches to Learning, 1st Grade (X4TCHAPP)	0.266
Executive Function (X6DCCSSCR)	0.115
Fall K Reading (X1RTHETK)	0.038
Child Sex (X_CHSEX_R)	0.003
Spring K Approaches to Learning (X2TCHAPP)	0.000
Fall K Approaches to Learning (X1TCHAPP)	0.000
Race/Ethnicity (X_RACETH_R)	0.000
Home Language (X12LANGST)	0.000

Early cognitive scores, particularly spring kindergarten math and reading, were the strongest predictors of 5th-grade risk. SES also showed substantial predictive power. Notably, race/ethnicity and home language had zero coefficients, indicating the elastic net regularization excluded these features from the final model.

4.4 Fairness Analysis

4.4.1 Group-Level Performance

Table 4 presents performance metrics by racial/ethnic group.

Table 4: Model Performance by Race/Ethnicity

Group	N	Prevalence	TPR	FPR	PPV	Accuracy	Pos. Rate
White	1,462	11.9%	0.172	0.012	0.667	89.1%	3.1%
Black	300	25.0%	0.293	0.102	0.489	74.7%	15.0%
Hispanic	623	29.4%	0.410	0.073	0.701	77.5%	17.2%
Asian	8	50.0%	0.750	0.000	1.000	87.5%	37.5%
Other	225	12.0%	0.259	0.010	0.778	90.2%	4.0%

Several patterns emerge from the fairness analysis:

Differential base rates: At-risk prevalence varied substantially across groups, from 11.9% (White) to 50.0% (Asian, though N=8). Black and Hispanic students had roughly double the at-risk prevalence of White students.

TPR disparities: True positive rates ranged from 0.172 (White) to 0.750 (Asian) and 0.410 (Hispanic). The model was substantially better at identifying at-risk students in groups with higher base rates.

FPR disparities: False positive rates showed even larger relative disparities. Black students experienced an FPR of 0.102, compared to 0.012 for White students—a ratio of approximately 8.5:1.

Accuracy disparities: Overall accuracy was highest for White (89.1%) and Other (90.2%) students, and lowest for Black (74.7%) and Hispanic (77.5%) students.

4.4.2 Disparity Analysis

Table 5 presents formal disparity metrics comparing each group to the White reference group.

Table 5: Fairness Disparity Metrics (Reference: White)

Group	**TPR Ratio**	**TPR Diff**	**FPR Ratio**	**FPR Diff**	**Disparate Impact**
Asian	4.350	+0.578	0.000	-0.012	No
Black	1.701	+0.121	8.777	+0.091	No
Hispanic	2.377	+0.237	6.245	+0.061	No
Other	1.504	+0.087	0.867	-0.002	No

Fairness Criteria Assessment:

- **Equal Opportunity:** PASS. All groups had TPR ratios > 0.80 compared to the reference group (in fact, all minority groups had *higher* TPR than White students).
- **Equalized Odds:** FAIL. While TPR ratios exceeded 0.80, FPR ratios for Black (8.777) and Hispanic (6.245) students dramatically exceeded 1.0, indicating these groups experienced disproportionately high false positive rates.
- **Statistical Parity:** PASS. The positive prediction rates, while varying across groups, did not trigger the 0.80 disparate impact threshold.

Figure 1 presents ROC curves stratified by racial/ethnic group.

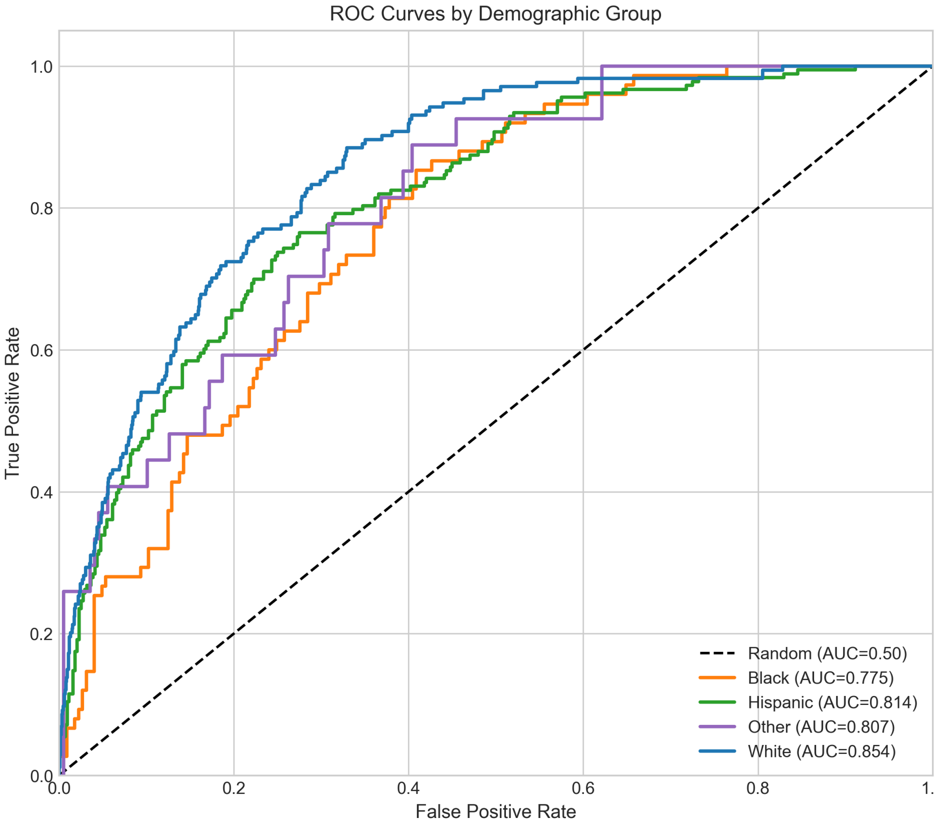

Figure 1: ROC Curves by Racial/Ethnic Group

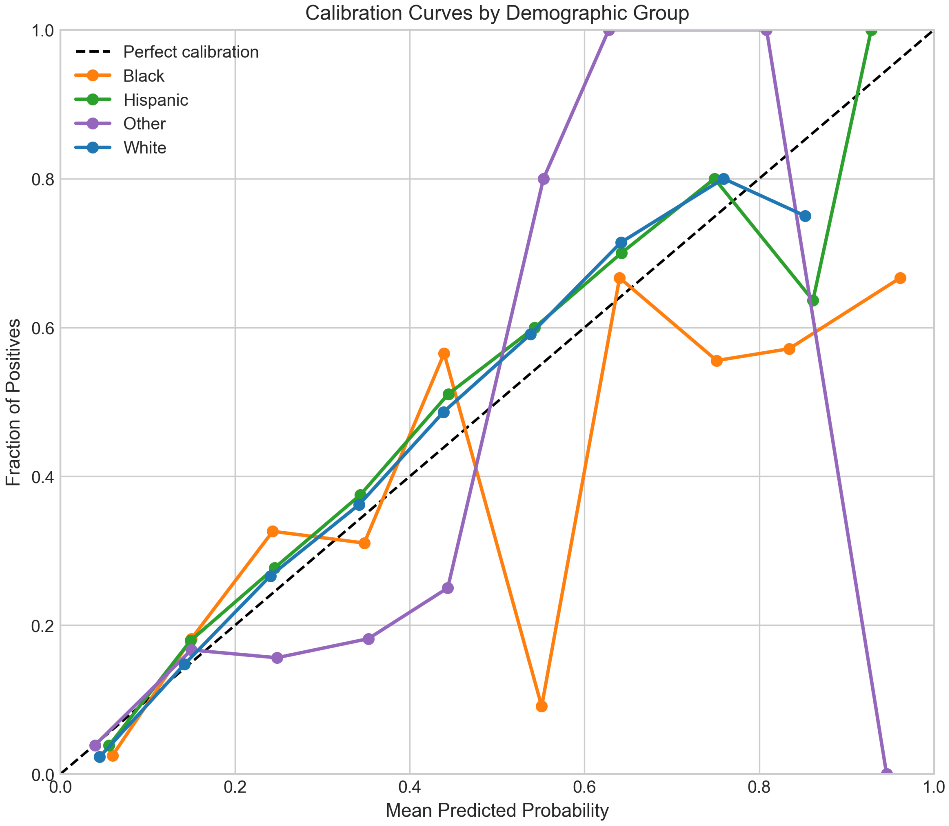

Figure 2: Calibration Curves by Racial/Ethnic Group

4.5 Bias Mitigation Results

We applied threshold optimization to equalize TPR across groups, targeting the overall model TPR of 0.294. Table 6 presents the results.

Table 6: Bias Mitigation Results (Threshold Optimization)

Group	**TPR Before**	**TPR After**	**Δ TPR**	**Acc. Before**	**Acc. After**	**Δ Acc.**
White	0.172	0.293	+0.121	0.891	0.886	-0.005
Black	0.293	0.293	+0.000	0.747	0.747	+0.000
Hispanic	0.410	0.295	-0.115	0.775	0.762	-0.013
Asian	0.750	0.250	-0.500	0.875	0.625	-0.250
Other	0.259	0.296	+0.037	0.902	0.884	-0.018

Threshold optimization successfully equalized TPR across the major demographic groups (White, Black, Hispanic, Other all achieving TPR $\approx$ 0.29). However, this came at a cost:

- Hispanic students experienced reduced sensitivity (-0.115) as the threshold was raised.
- Asian students experienced a substantial drop in both TPR and accuracy, though the small sample size (N=8) limits interpretation.
- Overall accuracy decreased slightly across most groups.

The group-specific thresholds ranged from 0.367 (White) to 0.649 (Asian), indicating that predictions for different groups required different decision boundaries to achieve equitable outcomes.

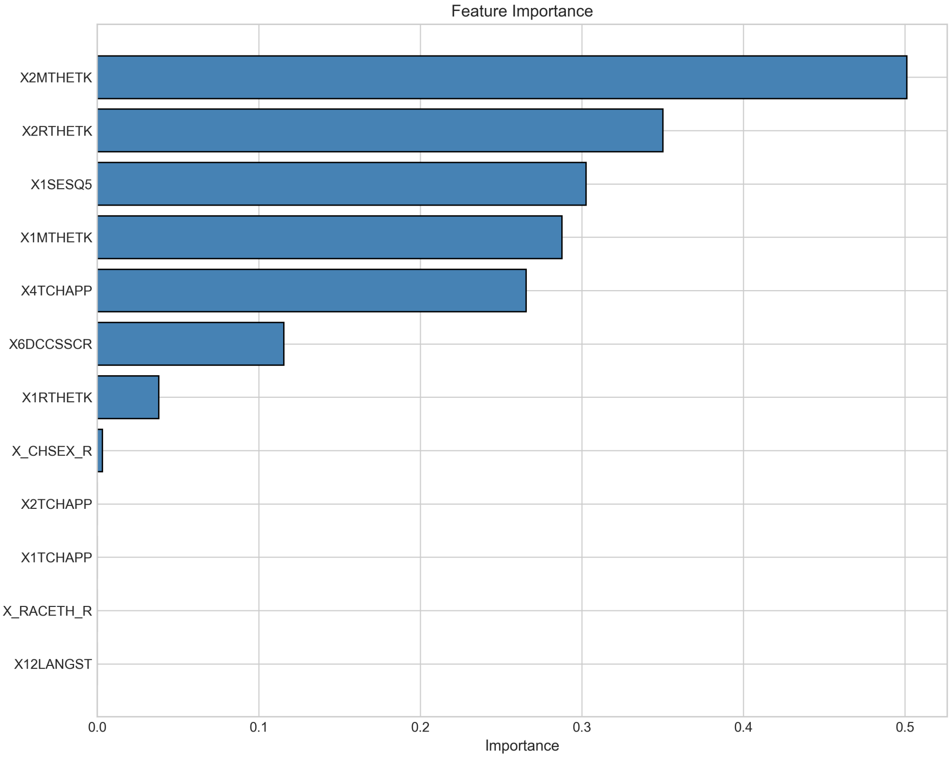

Figure 3: Feature Importance from Elastic Net Model

5 Discussion

5.1 Summary of Findings

This study examined the fairness properties of machine learning models that use early childhood data to predict 5th-grade academic risk. Our key findings include:

1. **Reasonable predictive performance:** Early childhood cognitive and behavioral measures predicted 5th-grade risk with moderate accuracy (AUC = 0.848), consistent with prior literature on longitudinal educational prediction.

2. **Significant fairness disparities:** Despite excluding race from the model, substantial disparities emerged across racial/ethnic groups. The model detected at-risk Hispanic students at more than twice the rate of at-risk White students, while also generating disproportionately high false positives for Black and Hispanic students.

3. **Criteria-dependent fairness:** The model satisfied equal opportunity and statistical parity criteria while failing equalized odds, illustrating how fairness assessments depend critically on which criteria are prioritized.

4. **Mitigation trade-offs:** Threshold optimization achieved more equitable TPR across groups but reduced overall accuracy and sensitivity for some groups.

5.2 Interpreting Fairness Disparities

The observed fairness disparities require careful interpretation. Several factors may contribute:

Differential base rates: At-risk prevalence was substantially higher among Black (25.0%) and Hispanic (29.4%) students compared to White students (11.9%). When base rates differ, even a well-calibrated model will exhibit different error rates across groups (Chouldechova, 2017).

Proxy discrimination: Although race was excluded from the model, other features (particularly SES) are correlated with race and may serve as proxies. The elastic net assigned substantial weight to SES (coefficient = 0.303), which could contribute to differential performance.

Structural inequities: The patterns in the data reflect historical and ongoing structural inequities in educational opportunity. Children from disadvantaged backgrounds may show weaker early signals not because of inherent ability, but because of differential access to high-quality early childhood education.

5.3 Implications for Educational Practice

Our findings have important implications for the deployment of predictive analytics in K-12 education:

Fairness audits are essential: Before deploying any predictive model, stakeholders should conduct comprehensive fairness audits examining multiple criteria and subgroups. A model that appears fair by one criterion may exhibit substantial disparities by another.

Context matters: The "optimal" fairness criterion depends on the use case. If the goal is to ensure all at-risk students have equal chances of being identified (equal opportunity), different thresholds may be appropriate. If the goal is to avoid disproportionate surveillance of minority students (equalized odds), the current model would require substantial modification.

Mitigation involves trade-offs: Bias mitigation is not a free lunch. Threshold optimization improved TPR equity but reduced accuracy for some groups. Stakeholders must weigh these trade-offs explicitly.

Human oversight remains critical: Predictive models should inform, not replace, human judgment. Educators and counselors should understand model limitations and exercise discretion in interpreting predictions.

5.4 Limitations

This study has several limitations:

- **Public-use data constraints:** The public-use ECLS-K:2011 file has some variables suppressed or top-coded to protect confidentiality, potentially limiting predictive power.

- **Complete-case analysis:** We used complete-case analysis, which may introduce selection bias if missingness is related to outcomes. The complete-case sample (N = 9,104) represented 50% of the original data.

- **Single cohort:** The ECLS-K:2011 followed a single cohort (2010-2016). Findings may not generalize to other cohorts or contexts.

- **Binary outcome:** We operationalized risk as a binary threshold (<25th percentile). Alternative operationalizations could yield different results.

- **Small subgroup sizes:** The Asian subgroup (N=8 in the test set) was too small for reliable inference.

5.5 Future Directions

Several directions for future research emerge from this study:

- **In-processing fairness methods:** We examined only post-hoc threshold adjustment. Future work should evaluate in-processing methods (e.g., adversarial debiasing, fairness constraints) that incorporate fairness during model training.

- **Intersectional fairness:** We examined race/ethnicity in isolation. Intersectional approaches examining race × gender × SES could reveal additional disparities.

- **Longitudinal fairness:** How do fairness disparities evolve as predictions are made at different time points? Models trained on kindergarten data may have different fairness properties than those trained on 3rd-grade data.

- **Intervention studies:** Ultimately, the value of EWS depends on whether they improve outcomes. Randomized studies examining the causal effect of EWS-informed interventions, with attention to differential effects across groups, are needed.

6 Conclusion

This study provides evidence that machine learning models predicting educational outcomes from early childhood data exhibit significant fairness disparities across racial/ethnic groups. While our model achieved reasonable predictive accuracy (AUC = 0.848), it showed substantially different true positive and false positive rates for different demographic groups. These disparities persisted despite excluding race as a predictor variable.

Our findings underscore the importance of rigorous fairness evaluation before deploying algorithmic systems in educational settings. The choice of fairness criterion matters: our model passed equal opportunity and statistical parity criteria while failing equalized odds. Stakeholders must carefully consider which criteria align with their values and use cases.

Bias mitigation through threshold optimization achieved more equitable true positive rates but introduced trade-offs in overall accuracy. There is no single “fair” solution; rather, fairness involves navigating competing values and accepting certain trade-offs.

As predictive analytics become increasingly prevalent in education, researchers and practitioners must remain vigilant about algorithmic fairness. The promise of early warning systems—identifying struggling students for timely intervention—can only be realized if these systems work equitably for all students, regardless of demographic background.

References

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuber, B., & Addison, K. L. (2015). Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 93–102.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61–75). Springer.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 225–234.

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education* (pp. 174–202). Routledge.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lakkaraju, H., Aguiar, E., Rich, C., Hansen, D., Miller, D., Yuber, B., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1909–1918.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7.

Yu, R., Lee, H., & Kizilcec, R. F. (2020). Should college dropout prediction models include protected attributes? *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 91–100.

A Technical Details

A.1 Software and Reproducibility

All analyses were conducted in Python 3.14 using the following packages:

- pandas 3.0.0
- numpy 2.4.2
- scikit-learn 1.8.0
- xgboost 3.1.3
- fairlearn 0.12.0
- matplotlib 3.10.8
- seaborn 0.13.2

Random seed was set to 42 for all stochastic operations. Code and data processing scripts are available in the project repository.

A.2 Model Hyperparameters

The final elastic net model used the following hyperparameters selected via 5-fold cross-validation:

- Regularization strength (α): 0.01
- L1 ratio: 0.5
- Maximum iterations: 1000

Cross-validation AUC scores ranged from 0.832 to 0.842 across folds, indicating stable performance.

A.3 Missing Data

Table 7 presents missing data rates for key variables.

Table 7: Missing Data Rates

Variable	**N Missing**	**% Missing**
5th Grade Reading (Outcome)	6,724	37.0%
Executive Function (X6DCCSSCR)	4,379	24.1%
1st Grade Approaches to Learning	4,708	25.9%
Fall K Reading	2,482	13.7%
SES Quintile	2,063	11.4%
Home Language	2,106	11.6%
Spring K Reading	965	5.3%

The high rate of missing outcome data (37%) reflects sample attrition over the longitudinal study. Children who remained in the study through 5th grade may differ systematically from those who dropped out.