Original Research Article

Evaluating Bias in Machine Learning Predictions of High School Students' Academic Performance in Ontario, Canada

Zihao Ye

White Oaks Secondary School, 1330 Montclair Dr, Oakville, Ontario, L6H 7A4, Canada

ABSTRACT

This study examines whether machine learning models trained on aggregated institutional data exhibit varying prediction accuracy among different school performance groups within Ontario's public education system. Utilizing theoretical frameworks from algorithmic fairness research, the study investigates whether predictive reliability exhibits systematic variation across educational contexts, even in the absence of explicit demographic variables. Using publicly available standardized test data from the Education Quality and Accountability Office, the study created regression models to predict how well schools perform academically based on their achievement levels. Model performance was evaluated using error-based metrics and subgroup error analysis across performance strata. Results indicate that prediction accuracy was not uniform across school groups, with lower-performing schools consistently exhibiting higher prediction error across model configurations. These results show that predictive models created from combined educational data might perform differently depending on the specific context. This underscores the necessity of assessing predictive consistency across institutional contexts when implementing machine learning techniques in education.

Keywords: Academic performance prediction**;** predictive modelling; institutional performance evaluation; algorithmic fairness; educational data mining; regression analysis

INTRODUCTION

The increasing integration of machine learning into education systems has transformed how academic outcomes are analyzed, predicted, and evaluated. Predictive models are now routinely used to estimate student performance and inform institutional policy decisions. Despite their potential for enhanced efficiency and data-driven decision-making, researchers have expressed concerns that these models may not function uniformly across diverse populations or educational settings (1). Unequal prediction accuracy can lead to biased outcomes and exacerbate existing structural inequalities.

These concerns are especially relevant in public education systems that rely on aggregated assessment data for institutional evaluation. In Ontario, Canada, the Education Quality and Accountability Office (EQAO) administers standardized assessments that generate large-scale, publicly available performance indicators used to evaluate school effectiveness. Because these datasets are highly structured and aggregated at the school level, they are increasingly attractive for machine learning applications. However, aggregated data can contain

Corresponding author: Zihao Ye, E-mail: daniel.ye0414@gmail.com.

Accepted April 21 2026
https://doi.org/10.70251/HYJR2348.42426432

underlying variations in performance distributions and school contexts that may affect model reliability (2).

Research on algorithmic fairness has shown that predictive models frequently exhibit unequal performance across groups defined by demographic, socioeconomic, or contextual characteristics. Models trained on structured social data can inherit imbalances present in the training dataset, leading to systematic differences in prediction error (1, 2). This phenomenon has been documented across domains including healthcare, finance, and education. Foundational studies emphasize that overall accuracy alone is insufficient; subgroup performance must also be examined to ensure predictive systems work equitably (1). More recent work reinforces the need for fairness-focused metrics and subgroup analysis to reveal hidden disparities in educational prediction models (2, 3).

While much of the existing literature on fairness in educational prediction has focused on individual demographic variables such as socioeconomic status or gender, many publicly available provincial datasets, including those from Ontario, are aggregated at the institutional rather than the student level. This creates a methodological gap: fairness has been widely studied across demographic groups, but less attention has been paid to whether predictive models show differential accuracy across institutional performance contexts when only aggregated data are available (1, 5). Ontario's EQAO datasets offer an opportunity to address this gap, as school performance is reported in standardized achievement levels. Investigating predictive disparities at the aggregated institutional level can determine whether structural performance differences alone are sufficient to produce unequal predictive reliability.

This study addresses the following research question:

Do machine learning models trained to predict high school academic performance in Ontario demonstrate differential prediction accuracy across school performance groups?

The central hypothesis is that predictive accuracy will not be uniform. Based on established findings in algorithmic fairness and the observed heterogeneity of educational performance distributions, models are expected to exhibit higher prediction error for lower-performing schools. Such a pattern would indicate that predictive models trained on aggregated institutional data may perform less reliably in contexts characterized by greater variability or structural disadvantage.

The primary aims of this research are to develop a baseline regression model for predicting school-level academic achievement using EQAO standardized assessment data, evaluate overall predictive performance with standard error metrics, and, most importantly, conduct subgroup error analysis across school performance strata. By focusing on predictive consistency rather than classification outcomes, the study contributes to the algorithmic fairness literature while respecting the methodological constraints of publicly available aggregated data. In doing so, it seeks to provide evidence-based guidance for the responsible use of predictive analytics in publicly funded education systems.

METHODS AND MATERIALS

Study Design and Data Source

This study investigates whether machine learning models trained on aggregated standardized assessment data exhibit systematic differences in predictive accuracy across school performance contexts in Ontario, Canada. Specifically, the study employed a quantitative, observational research design using publicly available aggregated educational assessment data. The analysis utilized school-level performance data derived from standardized provincial assessments administered by the Education Quality and Accountability Office (EQAO). These datasets report the proportion of students achieving predefined performance levels rather than individual student scores, therefore allowing institutional-level predictive modeling while maintaining student anonymity. The dataset consisted of aggregated achievement-level distributions from 90 Ontario public high schools. Using an 80/20 train-test split, 72 schools are used for training, and 18 for testing. The data were drawn from the 2023 EQAO assessment cycle.

All data used in the study was secondary, de-identified, and publicly accessible. No individual-level information was collected or analyzed; all variables remained in percentage form (0-100). Because the datasets represent aggregated institutional outcomes, the unit of analysis for this study was the school rather than the student.

Variables

The dataset included percentage distributions of students across four standardized achievement levels (Level 1 through Level 4). These levels reflect increasing degrees of academic proficiency according to provincial assessment standards. The data was stored in structured tabular format and processed using

Python-based data analysis tools. Data preprocessing was conducted using pandas (an open-source Python library for data manipulation and analysis). After loading the dataset, structural validation procedures were performed to confirm dataset dimensions, variable types, and completeness of performance-level variables. Only numeric achievement-level percentage variables were used to ensure consistency across observations. To construct a single interpretable outcome variable, a composite measure of academic success was defined. The target variable, termed "High Achievement Rate," was calculated as the sum of the proportion of students achieving level 3 and level 4 performance at the provincial and above provincial standards, respectively. These levels correspond to meeting or exceeding provincial academic standards and therefore provide a meaningful representation of institutional academic success. All variables were retained in percentage form. No normalization or transformation procedures were applied, as the data was already standardized across schools.

Predictive Model Development

A supervised machine learning framework was implemented to model school-level academic performance. The predictive task was formulated as a linear regression problem, where the model estimated the High Achievement Rate for each school based on other achievement-level indicators. The predictor variables consisted of the proportion of students achieving level 1 and level 2 performance, which are both below the provincial standard. These variables were selected to capture the lower-performance distribution within each school and provide explanatory information for predicting higher achievement outcomes.

Model Training and Performance Evaluation

As stated before, the dataset was partitioned into training and testing subsets using an 80/20 split. A fixed random seed was applied to ensure reproducibility of results. Model training was conducted exclusively on the training subset, while evaluation metrics were computed using the held-out testing subset. A linear regression model was selected as the baseline predictive model due to its interpretability and suitability for continuous outcome prediction. Model fitting was performed using standard least-squares optimization. After training, predictions were generated for all observations in the testing dataset. Overall predictive performance was evaluated using mean absolute error (MAE) and root mean squared error (RMSE), which quantify the average magnitude of prediction error.

Subgroup Analysis

To investigate differential predictive performance across educational contexts, subgroup analysis was conducted. Schools in the testing dataset were categorized into three performance strata based on their observed High Achievement Rate. Performance groups were defined using empirical distribution thresholds, with schools partitioned into lower, middle, and upper performance tiers according to quantile cutoffs. The prediction error was then computed separately within each performance group using mean absolute error. This procedure made it possible to directly compare predictive accuracy across institutional contexts. Differences in subgroup error were interpreted as evidence of systematic variation in model performance across school performance environments. All data processing, modeling, and visualization procedures were implemented using Python in a reproducible computational environment. The analysis was conducted at the institutional level to evaluate model behavior rather than individual outcomes. Findings are therefore interpreted as structural patterns in predictive performance rather than characteristics of specific students or demographic groups.

RESULTS

A series of regression models were trained to predict school-level academic performance using aggregated EQAO achievement level distributions. Model performance was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Prediction error was further analyzed across school performance groups categorized as low, medium, and high based on observed achievement rates.

Model Performance Comparison

Three model configurations were evaluated using different feature sets:

Model A used Level 1 and Level 2 achievement percentages; Model B used Level 1, Level 2, and Level 3 percentages; Model C used all four achievement levels.

As shown in Figure 1, Model C demonstrated the lowest prediction error across both evaluation metrics. Model A produced the highest error values, while Model B showed moderate improvement relative to Model A but remained substantially less accurate than Model C.

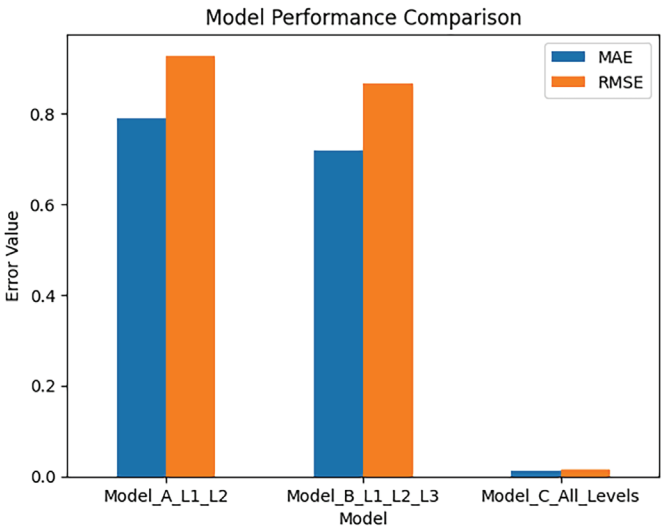

Figure 1. *Overall Predictive Performance (Mean Absolute Error (Mae) and Root Mean Squared Error (Rmse), Expressed in Percentage Points) of Three Linear Regression Models for School-Level High Achievement Rate on the Held-Out Test Set (N = 18 Schools).*

Prediction Error Across School Performance Groups

Prediction error was calculated separately for low, medium, and high-performing school groups for each model configuration. Figure 2 displays the mean absolute error (MAE) for each performance stratum, representing the average absolute difference between the model's predicted and the observed High Achievement Rates.

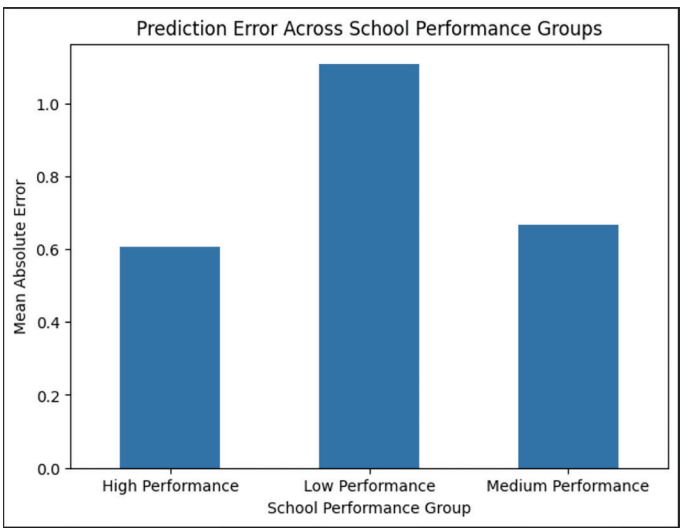

Figure 2. *Mean Absolute Prediction Error (Mae, in Percentage Points) by School Performance Stratum for the Baseline Linear Regression Model (Model A: Levels 1 & 2 Only) on the Held-Out Test Set (N = 18 Schools).*

For Model A, mean absolute error was highest for low-performing schools (1.103), followed by medium-performing schools (0.665), and lowest for high-performing schools (0.605). The SD of the prediction error was also highest in the low-performing group.

For Model B, the same ordering was observed. Mean absolute error was highest for low-performing schools (1.060), followed by medium-performing schools (0.569) and lowest for high-performing schools (0.529).

For Model C, overall prediction error decreased substantially across all groups. Mean absolute error was 0.017 for low-performing schools, 0.011 for medium-performing schools, and 0.009 for high-performing schools. Despite the reduction in overall error magnitude, the relative ordering of prediction error across groups remained consistent. These values, along with the model performance values from Figure 1, are further summarized in Table 1.

Table 1. *Subgroup Mean Absolute Prediction Error (Mae, in Percentage Points) by School Performance Stratum (Low, Medium, High) for Each Model Configuration on the Test Set (N = 18 Schools).*

Model	Performance Group	Mean Error	Std. Error	Sample Size
Model A	Low	1.103	0.639	6
Model A	Medium	0.665	0.502	5
Model A	High	0.605	0.238	7
Model B	Low	1.060	0.642	6
Model B	Medium	0.569	0.447	5
Model B	High	0.529	0.216	7
Model C	Low	0.017	0.010	6
Model C	Medium	0.011	0.008	5
Model C	High	0.009	0.004	7

Hypothesis Outcome

The results indicate that prediction error was not uniform across school performance groups. Across all model configurations, low-performing schools exhibited higher mean prediction error compared to medium and high-performing schools. The observed pattern was consistent across feature configurations and persisted after overall model accuracy improved.

Together, these findings establish that predictive accuracy varied systematically across school performance

groups and was not uniform across institutional contexts. While the results identify consistent patterns in model behavior, they do not by themselves explain why these differences emerged or what they imply for the use of predictive modeling in educational systems. The next part of the discussion explains these findings in terms of differences in academic performance, how the models were set up, and the overall reliability of predictions in combined educational data.

DISCUSSION

The findings of this study indicate that prediction accuracy differed systematically across school performance groups. Across all model configurations, schools categorized within the lower achievement range exhibited higher prediction error than medium- and high-performing schools. This pattern remained consistent even as overall model accuracy improved through expanded feature inclusion. The persistence of this error distribution suggests that predictive performance was not uniformly distributed across institutional contexts.

One explanation for this pattern relates to structural variability in academic performance distributions. Lower-performing schools may exhibit greater heterogeneity in achievement patterns, making them more difficult for models to represent using aggregated indicators (6). When predictive models are trained on structured performance distributions, they tend to approximate dominant patterns in the dataset. As a result, contexts that deviate more strongly from central tendencies may be predicted with lower precision. The error pattern therefore matches the general idea that predictive systems work best for cases that are similar to the average examples in the training data.

These findings align with established principles from algorithmic fairness research, which highlight that predictive systems can perform unevenly for different groups, even if the models aren't specifically trained on demographic information. Previous research has shown that models trained on structured social data can reflect existing imbalances present in that data (1, 7). The present findings extend this insight by showing that differential predictive reliability can emerge at the institutional level using aggregated educational data. This supports the argument that fairness evaluation must consider not only individual-level attributes but also structural performance contexts (8).

The results are also consistent with prior research indicating that conventional performance metrics can obscure important differences in model behavior across subgroups (1, 2, 3). Although overall model accuracy improved substantially when all achievement levels were included as predictors, the relative ordering of prediction error across performance groups remained consistent. This suggests that improvements in aggregate model performance do not necessarily eliminate subgroup disparities. Such a pattern reinforces the methodological position that subgroup evaluation is necessary to fully assess predictive reliability.

At the same time, the magnitude of prediction error decreased considerably under the most comprehensive model configuration. This indicates that model specification plays an important role in determining predictive stability. Incorporating additional information about achievement distributions appears to reduce overall uncertainty in prediction. However, the persistence of relative error differences across school groups indicates that model complexity alone cannot fully resolve structural variability in educational performance (9). These patterns therefore highlight both the explanatory potential of expanded model design and the boundaries of what the available data can capture, and recognizing these boundaries is essential for interpreting the findings appropriately.

Several limitations must be considered when interpreting these findings. First, the study relied on aggregated institutional data rather than individual-level student records. As a result, it was not possible to evaluate demographic sources of predictive disparity directly. Second, the sample size of schools included in subgroup analyses was relatively small (the test set sample size was about 5-7 per subgroup), which may limit the generalizability of the findings. Such limited subgroup sample sizes substantially reduce statistical power, increase sampling variability, and produce less stable and reliable estimates of mean absolute error within each performance stratum. Consequently, while a consistent directional pattern of higher prediction error in lower-performing schools was observed across all model configurations, these subgroup differences should be interpreted with caution. The small sample also limits the generalizability of the findings to the broader population of Ontario secondary schools and raises the possibility that the observed disparities could be influenced by a few influential observations. The analysis relied on cross-sectional aggregated data from a single provincial context and a modest total sample of about 90 schools. Future work should incorporate larger, longitudinal, and individual-level datasets from

multiple jurisdictions. Prediction targets were derived from achievement distribution summaries rather than longitudinal outcome measures, restricting the scope of predictive inference. More importantly, Model C, the one using all achievement levels, produced artificially low error because it incorporated Levels 3 and 4, which are the exact variables used to construct the target High Achievement Rate (Levels 3 + 4). Since percentages sum to 100%, this configuration introduced data leakage, rendering its near-zero error non-generalizable and invalid for true out-of-sample prediction. Models A and B, which use only the lower levels, provide more realistic assessments.

Despite these limitations, the study contributes to ongoing discussions about responsible application of machine learning in education. The results demonstrate that predictive models can exhibit differential reliability across institutional performance contexts even when using standardized and publicly available data. This highlights the importance of evaluating not only whether models predict accurately overall, but also how prediction accuracy is distributed across educational environments.

The findings support the central hypothesis that predictive accuracy would not be uniform across schools. The ongoing trend of higher prediction errors in lower-performing schools indicates that predictive systems using combined data from different institutions may work differently in various educational settings. This reinforces the broader argument that fairness evaluation in educational machine learning must extend beyond demographic attributes to include structural performance differences within education systems (10).

CONCLUSION

This study demonstrates that machine learning models for predicting institutional academic performance in Ontario's public education system must be evaluated not only by overall accuracy but also by how consistently predictions perform across different school performance contexts. The consistent pattern of higher prediction error in lower-performing schools reveals that aggregated EQAO data contain important contextual complexity and heterogeneity that standard linear regression models do not fully capture, even when additional achievement-level predictors are included.

These findings showcase the importance of conducting subgroup-level fairness evaluations whenever predictive analytics are applied to institutional educational data by highlighting systematic differences in predictive reliability across school performance strata. This study provides evidence-based support for more careful and equitable deployment of machine learning in publicly funded education systems. Future research should extend this work by incorporating larger datasets, individual-level student records, and data from multiple provinces or assessment frameworks to better understand how demographic, socioeconomic, and institutional factors interact with model performance.

CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

REFERENCES

1. Kizilcec RF, Lee H. Algorithmic fairness in education. arXiv [preprint]. 2020. Available from: https://arxiv.org/abs/2007.05443
2. Verger M, Lall S, Bouchet F, Luengo V. Is your model "MADD"? A novel metric to evaluate algorithmic fairness for predictive student models. arXiv [preprint]. 2023. Available from: https://arxiv.org/abs/2305.15342
3. Raftopoulos G, Davrazos G, Kotsiantis S. Evaluating fairness strategies in educational data mining: A comparative study of bias mitigation techniques. *Electronics*. 2025; 14 (9): 1856. Available from: https://www.mdpi.com/2079-9292/14/9/1856, https://doi.org/10.3390/electronics14091856
4. Khan S, Mazhar T. Predictive analytics in education-enhancing student achievement through machine learning. *ScienceDirect.* 2024; 15: 3095-3125. Available from: https://www.sciencedirect.com/science/article/pii/S2590291125005522
5. Idowu J. Debiasing Education Algorithms. 2024. Available from: https://link.springer.com/article/10.1007/s40593-023-00389-4, https://doi.org/10.1007/s40593-023-00389-4
6. Bobonis G, Wagner J. The Effects of School Consolidation on Students and Teachers: Evidence from an Underperforming System. University of Toronto. 2022. Available from: https://www.economics.utoronto.ca/gustavo.bobonis/BSW_School_Consolidations_22-12.pdf
7. Chen R, Wang J, Mahmood, F. Algorithm fairness in artificial intelligence for medicine and healthcare. National Library of Medicine. 2023. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10632090/
8. Baker R, Hawn A. Algorithmic Bias in Education. Springer Nature Link. 2021. Available from: https://

link.springer.com/article/10.1007/s40593-021-00285-9, https://doi.org/10.35542/osf.io/pbmvz

9. Sha L, Chen G. Lessons from debiasing data for fair and accurate predictive modeling in education. *ScienceDirect*. 2022. Available from: https://www.sciencedirect.com/science/article, https://doi.org/10.2139/ssrn.4274378
10. Gándara D, Anahideh H, Ison M, & Picchiarini L. Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction. AERA Open. 2024. Available from: https://journals.sagepub.com/doi/10.1177/23328584241, https://doi.org/10.1177/23328584241258741