2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC

Host Sponsor UNC Charlotte

Paper ID #51717

An Explainable Machine Learning Framework for K–12 Artificial Intelligence Education

Sizhu Qu, Northeastern University

Sizhu Qu is a computer science graduate student at Northeastern University and holds a master's degree in education from Johns Hopkins University. Her interdisciplinary research explores explainable AI and data-driven educational equity, with a focus on the responsible integration of AI in K-12 education.

Prof. Huihui H Wang, Northeastern University

Dr. Huihui Wang is a Full Teaching Professor and Director of Computing Programs at Northeastern University in Arlington, VA. She was a Program Director at the NSF since August 2021 until June 2024. She was a tenured Associate Professor and Director of Cybersecurity Programs at St. Bonaventure University and the Founding Department Chair of Engineering at Jacksonville University, FL.

An Explainable Machine Learning Framework for K–12 Artificial Intelligence Education

1. Introduction

Significant disparities exist across U.S. K-12 school districts in providing equitable computer science and artificial intelligence (AI) education. Differences in district size, student demographics, and organizational structures impact students' learning opportunities and long-term educational outcomes [1], [2]. Understanding the relationship between these structural factors and educational equity is crucial for educators and policymakers committed to responsibly promoting AI education [5] and expanding equitable access to computer science learning opportunities [31], [32]. While machine learning has been used to analyze educational disparities on a large scale, much of this research relies on "black box" models that lack in-depth explanations of the underlying mechanisms contributing to inequality [6], [7] and may unintentionally amplify inequities when decisions are not transparent [4]. Transparency and interpretability are just as important as predictive accuracy for educational decision-making [6], [15].

This study proposes an interpretable machine learning framework to examine structural and demographic patterns in K-12 school districts. Because direct indicators of K–12 AI/CS course access are not consistently available in the CCD, this study treats graduation rate as a system-level outcome proxy that reflects district capacity and opportunity structures (e.g., institutional scale and demographic composition) relevant to equitable AI/CS education. Using nationally representative district-level data, this framework combines ensemble modeling with local explanation techniques to reveal how structural characteristics influence educational outcomes across districts of different sizes (small, medium, and large) [17], [24]. The goal of this research is not merely prediction, but to support educational decision-making by making structural inequalities visible and interpretable [6], [27].

2. Related works

Machine learning and artificial intelligence have been widely used in education to predict student performance, analyze learning behaviors, and inform resource allocation. However, concerns remain that opaque "black box" models may reinforce existing educational inequalities. Prior research shows that structural factors, including funding gaps and uneven resource distribution, strongly shape students' educational opportunities and outcomes [1], [2]. Large-scale datasets such as EdNet [3] support detailed behavioral analysis, but their educational value depends on whether the analytical models are interpretable and contextually grounded.

At the same time, computer science education research increasingly emphasizes equity as a central concern. Equity-oriented pedagogical studies highlight how learning environments and instructional practices can either support or marginalize learners [30]. Policy and regional analyses further show that expanding access to computer science education depends on institutional capacity, resource allocation, and local implementation conditions, as illustrated by large-scale urban initiatives such as those in New

York City [31]. Recent work frames “computer science education for all” as a systemic goal that requires addressing disparities in access and participation across K–12 systems [32].

Interpretability has therefore become a key issue in educational AI. Holstein and Doroudi [4] argue that whether AI systems reduce or exacerbate inequality largely depends on the transparency of their decision-making processes, a principle also emphasized in national AI education policies [5]. Methodologically, prior work proposes interpretable AI frameworks tailored to educational contexts [6], while noting persistent conceptual and technical challenges [7].

Empirical studies suggest that interpretable models can reveal important factors influencing student outcomes [8], but also involve trade-offs between interpretability and predictive accuracy [9], [10]. Research has explored interpretable tools for teacher feedback [11], examined the stability of explanation methods across student groups [12], and shown that interpretable feedback can improve trust and usability among educators and learners [13], [14]. At the same time, critical reviews caution that explanation techniques such as LIME or SHAP may produce unstable or misleading results without sufficient contextual grounding [15], [16].

Overall, prior research establishes interpretability as both a technical requirement and a pedagogical necessity in educational AI, while equity-focused studies underscore the importance of structural and systemic conditions. However, most existing work focuses on individual learners or specific platforms. There has been limited exploration of how interpretable machine learning can be applied at the school district level to examine structural inequalities across K–12 education systems. This study addresses this gap by proposing an interpretable framework to analyze the relationship between district-level structural characteristics and educational equity.

3. Approach and Uniqueness

This study presents a clear machine learning framework for examining educational equity in K-12 school districts. The framework examines how structural and demographic factors affect student learning outcomes within each school district. It uses a random forest model to capture nonlinear relationships in large-scale administrative data. It also applies the Local Interpretable Model-agnostic Explanations (LIME) method to give case-level explanations that go beyond the overall model trends.

The analysis is based on school district-level data from the National Center for Education Statistics (NCES) Common Core Data (CCD). It constructs feature variables representing enrollment, demographic composition, and institutional size. Unlike previous equity studies that focused solely on prediction, this framework prioritizes interpretability and policy relevance, aiming to reveal structural inequalities across school districts of different sizes.

4. Data Source and Selection

This study utilizes the Common Core of Data (CCD) [17] released by the National Center for Education Statistics (NCES). This is a nationally representative administrative dataset covering all public K-12 school districts in the United States, providing standardized information on district structure, enrollment, demographics, and graduation outcomes. It is suitable for equity-oriented and policy-relevant analyses. We focused on the Local Education Agency (LEA) census data for the 2018-2019 school year to avoid structural disruptions caused by the COVID-19 pandemic and to ensure temporal consistency among variables.

The reason we ultimately chose CCD over other large educational datasets initially considered (such as EdNet [3]) is that it reflects the US public education system and includes macro-level structural and outcome variables. EdNet, for example, while providing detailed micro-level learning behavior data from private tutoring platforms, lacks district-level institutional characteristics and policy-relevant outcomes, limiting its applicability in equity analyses of K-12 education structures.

The three components of the CCD: directory, membership, and graduation datasets, were cleaned and integrated into a single unified district-level dataset. Records unrelated to standard public-school districts (e.g., state agencies or correctional institutions) were removed. Key identifiers were standardized, numerical variables were converted to a consistent format, and unreasonable values (e.g., graduation rates outside a reasonable range) were excluded. Missing numerical values were imputed using the median to maintain distribution stability.

Five variables were ultimately retained for analysis: total student enrollment, number of operating schools, grade span, percentage of female students, and graduation rate. The graduation rate was treated as the outcome variable, while the other four variables served as structural and demographic predictors. Based on enrollment size and prior literature on the effects of school size [20-23], we categorized school districts into three groups: small districts (fewer than 1,500 students), medium districts (1,500-10,000 students), and large districts (more than 10,000 students). These thresholds reflect the actual enrollment distribution and existing research findings that the impact of school size on outcomes is context-dependent and non-linear. The final dataset comprises 12,673 valid school districts nationwide, encompassing significant variations in district size, institutional structure, and student demographics. This preparatory work supports subsequent modeling and interpretability analyses by ensuring that structural differences across regions are consistently represented and suitable for examining equity-related patterns.

5. Modeling and Analysis

This study employs the Random Forest (RF) regression model as the primary analytical method. The RF model, proposed by Breiman [24], is suitable for structured tabular data, capable of capturing non-linear relationships, and robust against overfitting. We chose to investigate the relationship between structural and demographic factors and K-12 school district graduation rates. Since graduation rate is a continuous variable, we modeled this task as a regression problem, consistent with previous educational economics research that uses regression-based methods to analyze school district-level performance [23], [25].

Compared to more complex models such as gradient boosting or deep neural networks, the RF model strikes a good balance between predictive performance and interpretability. Global feature importance scores help us examine how factors such as school district enrollment size, gender composition, and grade span influence graduation rates, thus supporting policy-relevant interpretations.

To complement the global explanations, this study introduces Local Interpretable Model-agnostic Explanations (LIME) [27], [28] to provide case-level explanations, highlighting how specific features influence predictions for individual school districts, thereby enhancing transparency at the local level. For benchmarking purposes, we implemented a K-Nearest Neighbors (KNN) regression model [26]. The KNN model serves as a distance-based baseline model for comparing predictive performance and illustrating the limitations of simpler models in capturing non-linear structural relationships.

All analyses were conducted in a Jupyter Notebook environment using Python and the Scikit-Learn library. The workflow includes data preprocessing, model training and testing, performance evaluation (R^2, MSE, MAE), and model interpretation using global feature importance and LIME-based local explanations to enhance transparency and allow for future extension to other models or datasets.

6. Results and Contributions

6.1. Model Performance Overview

This analysis aims to determine whether easily observable structural and demographic characteristics at the school district level can provide useful information about student graduation outcomes. We used four school district-level indicators: total enrollment, the percentage of female students, the number of schools, and the grade span. We wanted to see how much variation in graduation rates these indicators can explain without using detailed socioeconomic or instructional data.

The Random Forest (RF) model yielded an R^2 value of 0.34, indicating that a considerable portion of graduation outcomes is related to school district-level structural conditions. While this predictive accuracy is not high, the result is still noteworthy considering the limited range of variables and the complexity of educational achievement. To better understand this result, we also evaluated a K-Nearest Neighbors (KNN) regression model. Its performance ($R^2 = 0.16$) suggests that simple similarity-based methods cannot capture the complex nonlinear relationships found in school district educational systems. This comparison highlights the need for modeling approaches that can account for differences across school districts.

Importantly, the goal of this analysis was not to maximize predictive accuracy, but rather to assess whether structural indicators can provide interpretable insights into educational inequality. The results show that, even without factors such as family income, school funding, or teacher traits, the structures of school districts can exhibit clear patterns in graduation outcomes. This finding encourages more research into how the size of institutions and their demographic backgrounds affect educational fairness.

6.2. Performance Across District Size Categories

To see whether the link between school district structure and student achievement varies with system size, we divided school districts into small, medium, and large groups based on total enrollment. We trained random forest models for each category to understand how structural features work in different educational settings.

Significant differences appeared between school districts of various sizes. In small school districts with fewer than about 1,500 students, the relationship between structural indicators and graduation rates was fairly consistent, with an $R^2 = 0.34$, similar to the overall results. In contrast, the model's performance in medium-sized school districts was much weaker, showing an $R^2 = -0.02$. This indicates more internal diversity that the basic structural variables do not fully explain. Large school districts showed the least stability, with an $R^2 = -0.14$. Negative R^2 indicates that, within the medium and large groups, the model performs worse than a within-group mean baseline, which can occur when outcome variance is limited and unobserved contextual factors dominate. This reflects the complexity of large educational systems. In these systems, many factors beyond the school district's structure affect students' academic results.

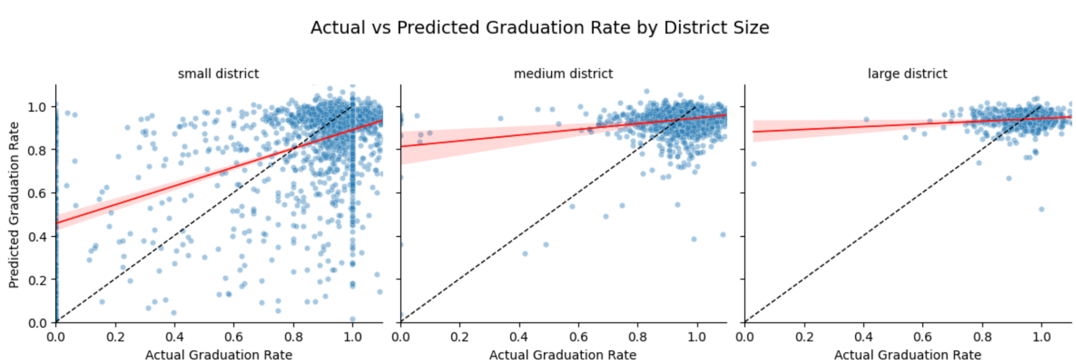
These results indicate that school district size moderates the relationship between structural characteristics and educational outcomes. In smaller, more structured systems, structural indicators provide more information. However, their explanatory power drops as institutional complexity increases. This pattern suggests that the difficulties large school districts face in achieving educational equity may stem from contextual and socioeconomic factors that basic administrative data do not capture.

Table 1. Random Forest predictive performance by district size.

District Size	n_train	n_test	R^2	MSE	MAE
Small	3972	1972	0.3383	0.1715	0.0706
Medium	3236	1622	-0.0198	0.0757	0.0152
Large	1282	589	-0.1392	0.0703	0.0101

Small districts show the highest model stability. In medium and large districts, the model shows lower R^2 values (often negative), suggesting that unobserved contextual factors dominate within-group variation even when absolute errors (MSE/MAE) appear small.

Figure 1. Actual vs. Predicted Graduation Rates Across School District Size Groups.



The faceted visualization in Table 1 summarizes the predictive performance for different school district sizes, and Figure 1 illustrates the increasing dispersion between predicted and actual graduation rates as school district size increases. These results highlight how system size impacts the interpretability and educational significance of structural indicators.

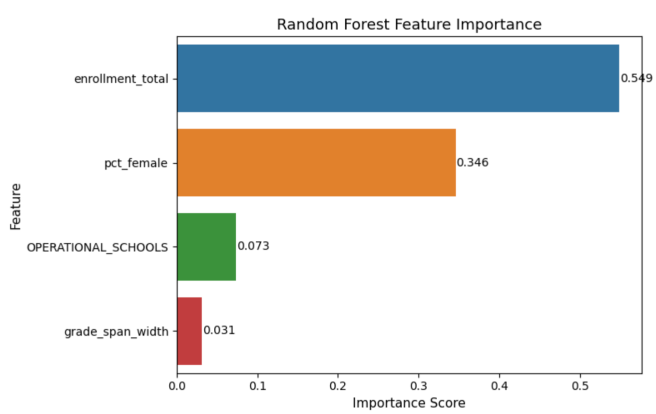
6.3. Feature Importance: Global Structural Patterns

The random forest model showed a clear global pattern in how structural and demographic factors relate to graduation outcomes. Figure 2 indicates that, of the four predictor variables, total enrollment and the proportion of female students accounted for almost 90% of the feature importance. This suggests that larger characteristics have a greater impact than internal structural factors, such as grade span or the number of schools.

Enrollment size, which indicates the scale of an institution, had the biggest impact. This finding aligns with earlier research showing that very small and very large school districts face organizational challenges that affect student support and academic continuity. The significance of gender composition suggests that, even at the school district level, demographic structure matters. This may reflect differences in student engagement, involvement, or access to academic paths.

In contrast, the impact of organizational structure variables was minimal. Large-scale educational inequality is more closely linked to the population served by the school district and the district's size than to decisions about its administration. From an educational policy perspective, these results suggest that equity-oriented interventions need to prioritize demographic and scale-related factors rather than structural reorganization.

Figure 2. Feature importance scores of structural and demographic variables.



The feature importance distribution indicates that size-related indicators and demographic indicators (especially total enrollment and gender composition) contribute far more to model performance than organizational structure variables such as the number of schools or grade span. This result is consistent with previous findings that institutional size and demographic composition are better predictors of educational outcomes than administrative configuration.

6.4. Performance Comparison of Random Forest and K-Nearest Neighbors Models

To better understand the random forest model's prediction results, we compared its performance with that of a K-nearest neighbors (KNN) regression model using the same standardized features. The random forest model had a predictive performance of $R^2 = 0.34$. This was significantly higher than the

KNN model's $R^2 = 0.16$. This finding suggests that graduation outcomes rely on non-linear interactions among features at the school district level.

This performance gap became more noticeable as the school district size grew. The KNN model worked fairly well in small districts. However, its accuracy dropped significantly in medium- and large-sized districts, which had greater internal differences. These findings show that distance-based models cannot fully understand the complexities of large educational systems.

In conclusion, this comparison shows that non-linear ensemble models are more effective for studying structural inequalities at the school district level. It also stresses the need to combine these models with interpretability methods, such as LIME. This combination ensures that better modeling capabilities do not compromise educational transparency and policy relevance.

6.5. Local Interpretability Based on LIME

Global feature importance analysis shows general structural trends; however, it does not clarify how these factors work in individual school districts. To overcome this limitation, we used the Local Interpretable Model-agnostic Explanations (LIME) method on the trained random forest model. We then examined representative school districts in each size category.

Figure 3 (a). Small district – prediction driven mainly by low enrollment

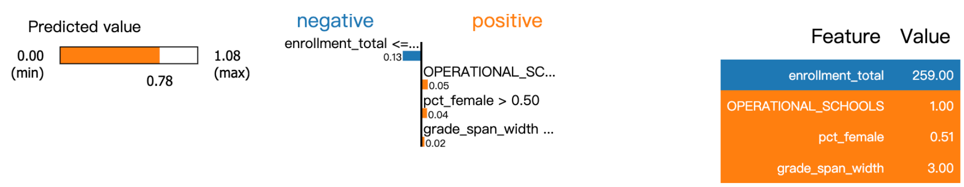


Figure 3 (b). Medium district – mixed and partially conflicting structural contributions

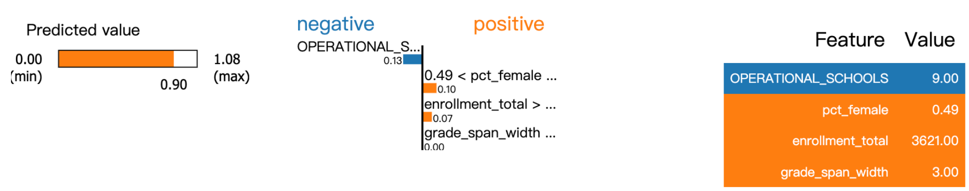
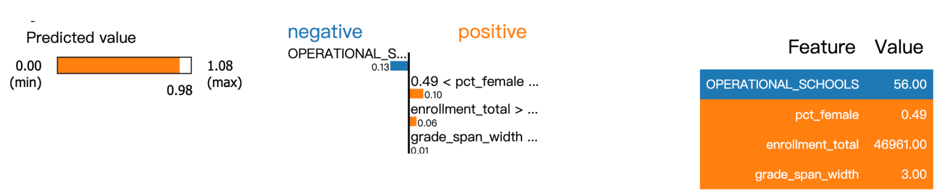


Figure 3 (c). Large district – unstable feature effects for enrollment, number of schools, and gender composition



Figures 3(a)–(c) show three typical cases. In small school districts (3a), the LIME analysis indicates that predictions are primarily influenced by low enrollment rates and a higher proportion of female students, reflecting relatively stable and consistent structural conditions. In medium-sized school districts (3b), LIME reveals the complexity of feature contributions, sometimes even showing contradictory effects. The same structural factors can have either supportive or detrimental effects on graduation outcomes, depending on the specific context. In large school districts (3c), feature contributions are unstable and partially cancel each other out, suggesting that fundamental structural indicators fail to capture the complexity of large educational systems.

This instability echoes previous findings that structural indicators alone are insufficient to describe complex urban areas. In summary, these local interpretations show that the effect of structural characteristics depends heavily on context and is influenced by regional factors. This supports using LIME as an interpretable layer over a nonlinear integrated model in education analysis focused on equity.

These explanations show that structural factors affect education differently across school districts of different sizes and organizational structures. By highlighting these context-dependent patterns, LIME makes it easier to understand non-linear models and helps with more detailed and fair educational analysis. LIME empowers educators and policymakers to investigate how structural conditions influence educational outcomes in specific contexts.

7. Summary

This study shows that interpretable machine learning can reveal meaningful patterns in educational equity using only school district-level structural and demographic data. Across analyses, graduation outcomes were more closely linked to district size and student composition, especially total enrollment and gender composition, than to administrative structure variables like grade span or number of schools.

The results indicate that structural indicators are most informative in smaller school districts, where institutional conditions are relatively stable. In contrast, predictive performance declines in medium and large districts, often yielding low or negative R^2 values, suggesting that within-group variation is dominated by unobserved contextual factors not captured in national administrative data.

Methodologically, this work contributes an interpretable, policy-relevant framework that combines random forest models with global feature importance and LIME-based local explanations. Rather than offering causal claims, the framework supports transparency-driven analysis by helping researchers and policymakers identify structural patterns, surface equity-related questions, and motivate more context-sensitive investigations of K–12 AI and computer science education.

8. Future Work

This study is an initial look at how interpretable machine learning methods can help us understand structural inequalities in K-12 artificial intelligence and computing education. Future research will dig deeper into context and relevance for education.

First, the analysis will shift from aggregated national data to targeted, representative school district case studies. We will choose several school districts using solid criteria, such as size (small, medium, large), geographic location, and variations in their computing and artificial intelligence education policies. This approach will help us investigate how structural patterns relate to local governance, institutional capacity, and policy environments. Second, future research will integrate policy and contextual analysis with structural data. We will look at school district and state policies related to computer science and artificial intelligence education. This includes curriculum rules, graduation requirements, and technology initiatives. We aim to explain the similarities or differences between what we see in practice and the goals of these policies. Third, we will expand this framework by adding several years of CCD data. This will help analyze the stability and changes of structural patterns over time, thus exploring how structural conditions and policy interventions jointly shape educational outcomes.

In summary, these extensions seek to change the current framework from a nationwide structural scan to a context-sensitive, policy-relevant analytical tool. This tool will support educational research and decision-making regarding fair access to artificial intelligence and computing education.