
Bridging the STEM Divide: Addressing Educational Disparity in Indiana

By Tianle Liang, Anthony Thieu

AUTHOR BIOGRAPHY

Tianle Liang is a senior at Zionsville Community High School interested in using machine learning to solve global disparities. He's conducted research at the University of Chicago (applied ML) and Stanford University (optimization and energy systems via SNOPT). Through engaging with his communities, he aims to improve access to STEM opportunities and leverage tech to improve lifelong learning.

Anthony Thieu is a senior at Zionsville Community High School passionate about STEM equality and healthcare. He has conducted research at the Indiana University School of Medicine in the field of diabetes, alongside founding clubs dedicated to providing access to STEM opportunities. Through his academic work and community involvement, he is interested in understanding how research and policy can reduce inequities in medicine.

ABSTRACT

This paper proposes and develops a robust, cross-benchmarked machine learning framework for predicting SAT benchmark percentages across unified school districts in Indiana. The study utilizes state-provided datasets from the Indiana Department of Education for 300+ districts and created indices representing early-stage ethnic and gender disparity, average early-stage performance, enrollment, and a variety of funding mechanisms. After dimensionality reduction through principal component analysis (captured 83.35% original variance) and quantile-binning the SAT label, random forest (RF) and XGBoost (XGB) ensemble learning models were trained independently, reaching 0.71 and 0.72 classification accuracy, respectively. A composite model was developed by blending probability distributions from RF and XGB, reaching an optimal accuracy of 0.73. Macro-averaged values for precision, recall, and F1 scores similarly increased to 0.74, 0.69, and 0.71, respectively, enhancing class balance overall. Subsequently, XGBoost gain calculations and Shapley Additive Explanations (SHAP) plots were discussed to examine specific feature relationships with SAT classes. Notable variables including ethnic disparity and enrollment were geospatially graphed to identify key geographical patterns for future mitigation planning. This study concluded by

SPRING 2026

recommending baseline policies encompassing small-district funding prioritization, corporate partnerships, and early-stage intervention.

Keywords: *Machine Learning, Artificial Intelligence, Forecasting, Geospatial Mapping, STEM Education, Disparities, Policy, Indiana*

INTRODUCTION

STEM (Science, Technology, Engineering, and Mathematics) fields fuel innovation, drive economic growth, and create jobs that strengthen Indiana communities. Key industries—like Eli Lilly in biotechnology, Corteva in agricultural technology, and Allison Automotive Transmission in automotive engineering—depend on STEM talent to sustain growth and competitiveness.

As of December 31, 2024, Eli Lilly employs 12,577 individuals in Indianapolis and announced a \$4.5 billion facility in Lebanon, Indiana, projected to open by late 2027 and create 400 full-time STEM jobs, including engineers, scientists, operations personnel, and lab technicians (Eli Lilly and Company, 2024) (Reuters, 2024). Despite such opportunities, a talent gap persists. Purdue University and the US Department of Agriculture estimate 60,000 annual openings in agriculture and related fields, yet only 35,000 graduates meet requirements. This underscores the urgent need to expand STEM education not only in college but also in primary education, to prepare a workforce capable of filling these high-demand roles in biotechnology, agronomy, and data science (Purdue University, 2024).

HISTORICAL CONTEXT

Early Public Education in Indiana

Indiana's 1851 constitution required a "general and uniform system of Common Schools," providing tuition-free education for all children. Funding at the township level, dependent on land values, disadvantaged rural schools: "Because funding for district schools was administered at the township level and was dependent upon lease or sale of school lands, the sums of money available to each district varied greatly. The poorest townships with the greatest need for money generally had the lowest land values and, consequently, the smallest school funds" (U.S. Commission on Civil Rights, 2022). As a result, urban schools thrived with resources and experienced teachers, while rural schools struggled with facilities, instruction, and access to STEM tools. This

SPRING 2026

early funding disparity laid the foundation for persistent educational inequities, impacting not only STEM but all areas of learning and contributing to long-term urban–rural divides.

Racial Segregation and Its Lasting Effects

De jure segregation in Indiana, enforced by laws like the Indiana School Law of 1869, created separate schools for Black children. Judicial rulings, including *Cory v. Carter* (1874), upheld these policies. While Indiana abolished legal segregation in 1949 and *Brown v. Board of Education* (1954) made segregation unconstitutional, de facto segregation persisted through white flight, redlining, and discriminatory school districting.

Although Indiana abolished segregation in 1949 and *Brown v. Board of Education* (1954) outlawed it, de facto segregation persisted through redlining, white flight, and discriminatory districting. In Indianapolis, this left Black families in underfunded urban schools, limiting access to gifted, honors, and AP courses—preparing only about 20% of students for “thinking work.” Black, Latinx, and Native American students, particularly young Black men, were disproportionately placed in alternative or dropout tracks, reflecting entrenched educational inequities that legal mandates alone could not resolve (Indiana State Library, 2023).

Limited access to gifted, honors, and AP courses, along with placement in alternative or dropout tracks, denied Black students early exposure to rigorous STEM curricula. Coupled with historical segregation and inequitable tracking, this creates a self-reinforcing cycle of underrepresentation in both academic achievement and the STEM workforce.

Unequal Funding Based on Property Taxes

Indiana’s heavy reliance on local property taxes creates inequities, as a student’s ZIP code largely determines educational quality. Affluent districts like Carmel Clay Schools benefit from high property values, while districts in East Chicago or Greene County receive significantly less per student—historically over \$2,000 less than wealthier areas (Indiana Department of Education, 2024).

Underfunded schools often lack fully equipped science labs, AP courses, updated textbooks, and modern technology. In STEM education, disparities are particularly stark: students in wealthier districts may have robotics kits, laboratory equipment, and research programs, while peers elsewhere may only experience theoretical instruction with outdated resources. These gaps hinder achievement, restrict career pathways, and perpetuate cycles of poverty and limited social mobility.

National STEM Push and Indiana's Uneven Response

Following *A Nation at Risk* (1983), Indiana sought to strengthen STEM education, but implementation has been uneven. By 2018, only 32% of rural high schools offered AP Physics or AP Computer Science, and in 2017–2018, just 33 schools statewide offered AP Computer Science, with only 920 students taking the exam (Indiana State Board of Education, 2020). In contrast, wealthier suburban districts provide a full range of AP STEM courses and extracurricular programs, highlighting a significant equity gap in access to college-level STEM education.

The disparity in STEM education stems from infrastructure and staffing, as rural and under-funded schools face shortages of qualified teachers due to low salaries and geographic isolation. These challenges deter new teachers and increase turnover. This creates a two-tiered system: students in wealthier districts are prepared for STEM careers, while others lack access, limiting individual prospects, the state's economy, innovation, and social mobility. Without investment in teacher support, labs, and curriculum, thousands of Indiana students risk being left behind in a high-tech future.

The Legacy Today

These disparities have worsened over time, reflecting systemic inequities in Indiana education. In 2023, only 28% of Black students and 33% of low-income students were proficient in math, compared to 58% of white students and 67% of high-income students, highlighting the persistent failure of the system for marginalized communities (Center for Social Determinants of Health, 2023).

STEM disparities in Indiana arise from curriculum gaps, infrastructure, and teacher shortages. Low salaries and geographic isolation drive high turnover in rural schools. This creates a two-tiered system, limiting access for students outside wealthier districts and restricting economic growth, innovation, and social mobility. Addressing these inequities requires early, sustained investment in elementary and middle school STEM education, including funding, qualified teachers, labs, afterschool programs, relevant curricula, mentorship, family engagement, and outreach to help underrepresented students envision and pursue STEM careers.

METHODOLOGY

Besides compiling and examining past literature, this paper aims to develop an effective machine learning framework to evaluate Indiana school districts' STEM quality through various recorded factors. Throughout this section, multiple ensemble learning techniques were optimized and cross-benchmarked to

SPRING 2026

predict math SAT scores from factors like early-stage proficiency percentages, ethnicity and gender disparity, and a variety of funding mechanisms.

Data Description

The data used was primarily compiled from public files provided by the Indiana Department of Education (IDOE), including files detailing per-district SAT benchmark percentages, math and science proficiencies, and 13 funding variables. Information was compiled from over 415 school districts across Indiana in 2022 (most recent complete dataset), although this number was finalized at 317 districts after all data was merged based on intersecting occurrences of districts. For computational simplicity, individual schools within districts were merged into their umbrella districts in the final dataset, with metrics combined based on enrollment sizes. (For example, Zionsville Middle School and Zionsville Community High School were consolidated under Zionsville Community Schools.)

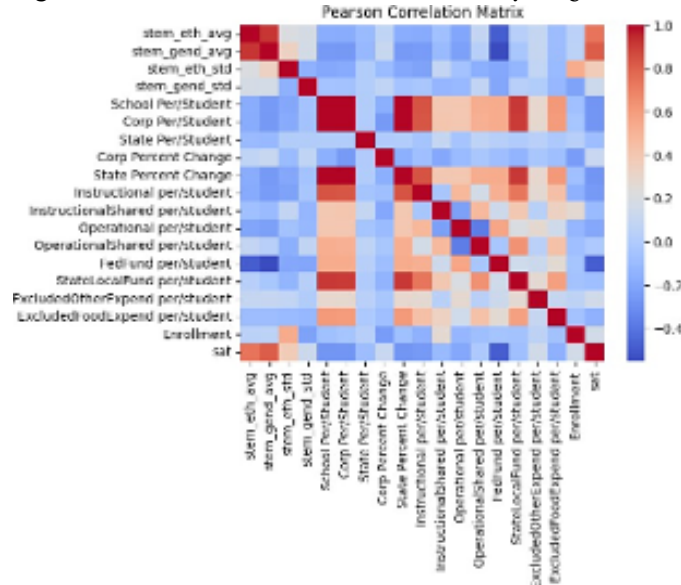
Cleaning and Preprocessing

The first variables processed were STEM disparity variables: specifically, early-stage math and science proficiency percentages by ethnicity and gender, extracted from the IDOE. Proficiency was determined by the standardized ILEARN (Indiana's Learning Evaluation and Assessment Readiness Network) test administered in middle schools. Proficiency percentage averages were calculated as an overall proxy indicator of future college and career readiness. More importantly, to create a disparity index, the standard deviations σ of proficiency

percentages across ethnicity and gender were calculated, where $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$ with n as the number of accounted ethnicities and μ as the mean proficiency percentage. A larger standard deviation would indicate higher early-stage STEM disparity, and vice versa. This index quantifies possible systemic equity concerns, structural bias, and uneven skill development between groups in school, factors that are later driving forces of success in education. The eight starting variables were finally averaged into four overall STEM variables (averaging math and science).

The following Figure 1 displays a Pearson correlation coefficient matrix calculated for all the numerical variables at this stage. Some major collinearity, with correlations of around 0.5-0.9 were observed among funding variables. Many funding variables were also determined to be disorganized and excessive given multiple similar overlapping domains, thus motivating the next approach of principal component analysis (PCA).

Figure 1: Pearson Correlation Matrix of Early-Stage STEM Disparity and School Funding Variables



Note. This figure presents the Pearson correlation coefficients among all numerical variables included after initial cleaning and preprocessing, including early-stage STEM disparity indices (based on ILEARN math and science proficiency standard deviations across ethnicity and gender) and multiple school funding variables. Moderate to high positive correlations ($r = 0.5-0.9$) are observed, particularly among funding-related variables, indicating substantial multicollinearity. These results justify the subsequent application of principal component analysis (PCA) to reduce dimensionality and address redundancy across overlapping funding domains.

Principal Component Analysis of Funding

Principal component analysis (PCA) was utilized subsequently for dimensionality reduction to simplify the 13 funding variables into four broader, more interpretable composite indices. The covariance matrix $S \in \mathbb{R}^{13 \times 13}$ was computed to capture the pairwise linear relationships among the original 13 variables. PCA solves the equation $Sv_i = \lambda_i v_i$ for eigenvectors v (loadings) and eigenvalues λ (explained variance). From the directions of maximum variance, four eigenvectors representing the principal components (PCs) were selected, preserving 83.35% of the original variance. The resultant PCs largely maintained interpretable structure, and Table 1 shows their respective domains.

Table 1: Principal Component Loadings and Explained Variance Ratios for School Funding Variables

PC	Category	Top Variables (Loading Factor)	EVR
----	----------	--------------------------------	-----

SPRING 2026

PC1	Core Funding	State Percent Change (0.41), School Per Student (0.41), Corp Per Student (0.41), StateLocalFund Per Student (0.37)	49.7%
PC2	Operational/Shared Funding	Operational Per Student (0.56), OperationalShared Per Student (0.51), InstructionalShared Per Student (0.41)	17.1%
PC3	Corporate Funding	Corp Percent Change (0.83)	8.83%
PC4	Other Funding	ExcludedOtherExpend Per Student (0.83), FedFund Per Student (0.37), ExcludedFoodExpend Per Student (0.19)	7.75%

Note. This table presents the results of principal component analysis (PCA) conducted on 13 school funding variables to reduce dimensionality and mitigate multicollinearity. Four principal components (PC1–PC4) were retained, collectively explaining 83.35% of the total variance. Each component is labeled according to its dominant funding domain, as interpreted from the highest loading variables. PC1 (Core Funding) captures broad per-student state and local funding measures, while PC2 reflects operational and shared expenditure structures. PC3 is driven primarily by corporate funding percent change, and PC4 captures alternative and federally influenced expenditure categories. The explained variance ratio (EVR) indicates the proportion of total variance accounted for by each component.

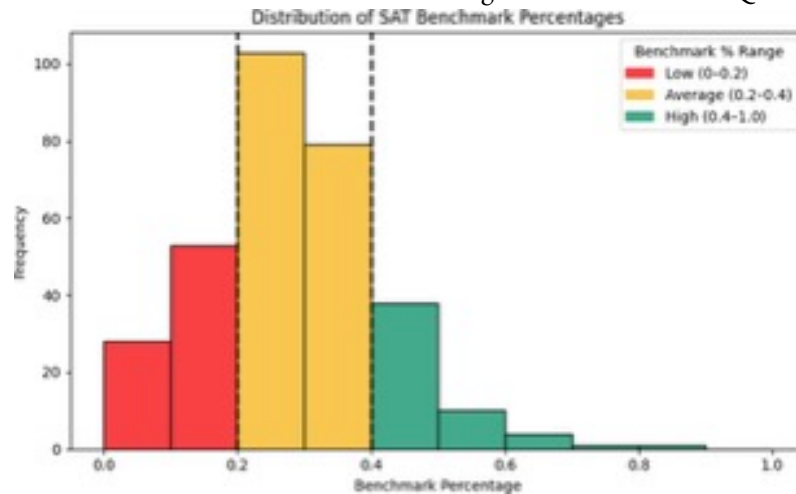
SAT Quantile Binning

The chosen label to be outputted in the model development stage was SAT scores per school, specifically the percentage of students reaching the school SAT benchmark. As a standardized test of college and career readiness, the SAT benchmark percentage was determined to be the optimal indicator of school district resources and academic quality from IDOE’s provided data. However, the heavy concentration of benchmark percentages in the lower ranges creates a risk of predictive bias in ensemble learning structures. Consequently, SAT benchmark percentages were binned into three categories, selecting bins to minimize class imbalance for a relatively normal distribution. The bins represented Low, Average, and High benchmark percentages, corresponding with 0.0–0.2, 0.2–0.4, and 0.4–1.0, respectively (shown in Figure 2).

Overall, a classification-based approach was chosen over regression for multiple reasons. Classification maximizes policy relevance, as policy thresholds often operate in tiers. Additionally, this approach creates technical advantages by preventing overfitting to small percentage noise and severe skewing. In future extensions, an ensemble regression model (e.g., RFR or XGBRegressor) can predict exact SAT benchmark percentages to gauge specific contributions of features. Prioritizing XGBoost’s built-in feature importance over PCA for the original 13 features offers a more granular approach to policy recommendations. As such, future

studies are strongly recommended to extend the ordinal framework of the present methodology.

Figure 2: Distribution of School SAT Benchmark Percentages with Tertile-Based Quantile Binning



Note. This figure illustrates the distribution of school-level SAT benchmark attainment percentages and the corresponding bin thresholds used for model classification. Due to a high concentration of schools in the lower benchmark range, raw percentages were separated into three categories (Low (0.0–0.2), Average (0.2–0.4), and High (0.4–1.0)) to reduce class imbalance and mitigate predictive bias in ensemble learning models. The selected bins approximate a more balanced class distribution while preserving meaningful distinctions in college and career readiness outcomes.

Model Development

Ensemble learning techniques of random forest (RF) and gradient boosting were selected for classification due to their robustness and handling of complex, nonlinear features. RF and gradient boosting differ primarily in the conglomeration of decision trees: random forest averages results across independent, parallel trees, whereas gradient boosting builds trees sequentially and reduces error incrementally with each step. XGBoost (XGB) was chosen as the specific framework of gradient boosting because of effective L1 and L2 regularization for overfitting. Random forest and XGBoost were trained separately on the same variables, as shown in Table 2.

Table 2: Feature Set Used for Random Forest and XGBoost Classification Models

Variable	Description
----------	-------------

SPRING 2026

stem eth avg	Average early-stage STEM proficiency across ethnicities
stem gend avg	Average early-stage STEM proficiency across genders
stem eth std	Standard deviation of early-stage STEM proficiency across ethnicities
stem gend std	Standard deviation of early-stage STEM proficiency across genders
enrollment	Total student enrollment in the district
pc1 core	Core funding component score from PCA
pc2 operational shared	Operational and shared funding component score from PCA
pc3 corporate	Corporate funding component score from PCA
pc4 other	Other funding component score from PCA

Note. This table summarizes the predictor variables used to train both the random forest and XGBoost classification models. The feature set includes early-stage STEM proficiency averages and disparity indices (standard deviations across ethnicity and gender), total district enrollment, and four principal component funding scores derived from PCA. Together, these variables integrate measures of academic readiness, structural disparity, district scale, and consolidated funding domains.

The hyperparameters for random forest and XGBoost were automatically optimized using RandomSearchCV for 25 iterations. The best hyperparameter combination for each model was determined by testing weighted F1 scores. The following Tables 3 and 4 show the selected hyperparameter combinations for random forest and XGBoost, respectively.

Results and Composite Model

Random forest and XGBoost performed moderately well on a 30% validation test set of school districts, reaching 0.71 and 0.72 accuracy respectively. In terms of raw classification accuracy, cross-benchmarking RF and XGB did not reveal substantial differences in performance across the different model structures. However, comparing precision, recall, and overall F1 scores, XGBoost performed with a slight edge in both macro averages and weighted averages (weighted by class instances to mitigate class imbalance). Overall, XGBoost performed with macro-averaged values of 0.72, 0.69, and 0.70 for precision, recall, and F1, respectively, compared to RF's values of 0.72, 0.66, and 0.68. Weighted average values were 0.72 for XGB compared to RF values of 0.71, again indicating better balanced performance across all classes.

To optimize raw accuracy and class balancing across models, a composite model was developed by

blending the probability distribution for predicted labels between RF and XGB.

Optimized Hyperparameter for Random Forest and XGBoost Models

Table 3: RF Hyperparams

Parameter	Value
n estimators	200
min samples split	10
min samples leaf	2
max features	sqrt
max depth	None

Table 4: XGB Hyperparams

Parameter	Value
subsample	1.0
n estimators	200
max depth	5
learning rate	0.01
colsample bytree	0.6

Note. These tables present the hyperparameter configurations selected for the random forest (RF) and XGBoost (XGB) classifiers following 25 iterations of RandomizedSearchCV. Hyperparameters were optimized using weighted F1 score as the selection criterion to account for class imbalance across SAT benchmark categories. The configurations reflect the structural differences between the ensemble methods, including tree depth, sampling strategies, and regularization parameters.

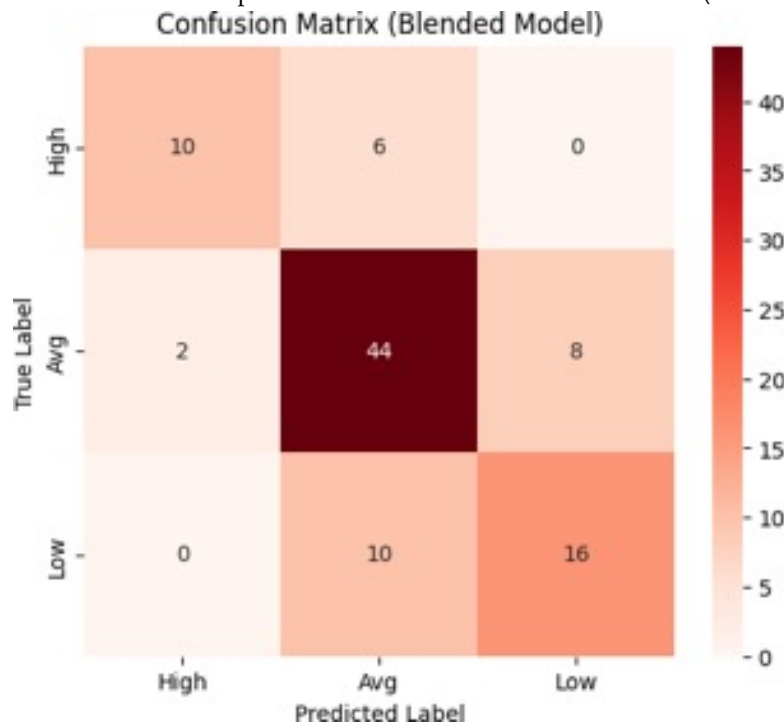
The composite model reached a marginally higher raw accuracy, at 0.73, and similar weighted average values of 0.73. A noticeable increase in macro-averaged values was also observed, with precision, recall, and F1 scores of 0.74, 0.69, and 0.71, respectively. Figure 3 displays the confusion matrix, representing predictions for the 96 unified school districts in the validation test set, for the composite model.

Five-fold cross-validation (approximately 63 districts per fold) was implemented to further confirm the stability of the composite model results. CV yielded a mean accuracy of 0.71 with a standard deviation of 0.02. Macro-averaged F1 scores centered around 0.70 ± 0.03 , while weighted F1 scores at 0.74 ± 0.01 . Weighted metrics were likely more stable than macro-averaged

metrics due to the prominence of the Average label.

In Figure 3, a strong diagonal was observed, corresponding with the satisfactory accuracy of the model. Predictions are especially balanced with classifying Average level school districts, with only a small percentage of misclassification spillover into High and Low categories. The model predicted less confidently the High and Low labels, struggling the most with classifying Low districts. However, spillover was completely near-adjacent toward the Average label, reinforcing strong ordinal awareness overall. Future work may mitigate this spillover by integrating ordinal-aware predictive techniques, including ordinal regression or custom loss penalties.

Figure 3: Confusion Matrix for Composite Ensemble Model on Validation Set (n = 96 Districts)



Note. This figure displays the confusion matrix for the blended composite model combining random forest and XGBoost probability outputs, evaluated on a 30% validation test set of 96 unified school districts. The strong diagonal pattern reflects an overall classification accuracy of 0.73 and balanced macro- and weighted-average performance. Misclassifications occur primarily between adjacent categories (Low - Average and Average - High), indicating ordinal consistency in predictions. The model demonstrates strongest performance for Average-level districts, with comparatively greater difficulty distinguishing Low-performing districts, suggesting opportunities for

future refinement using ordinal-aware modeling techniques.

Feature Importance

To assist in policy recommendations and future mitigation strategies for Indiana legislators, the feature importances of the XGBoost input variables were calculated to determine which areas to prioritize the allocation of funds to. Feature importances were calculated based on XGB's gain formula, as shown in Equation 1.

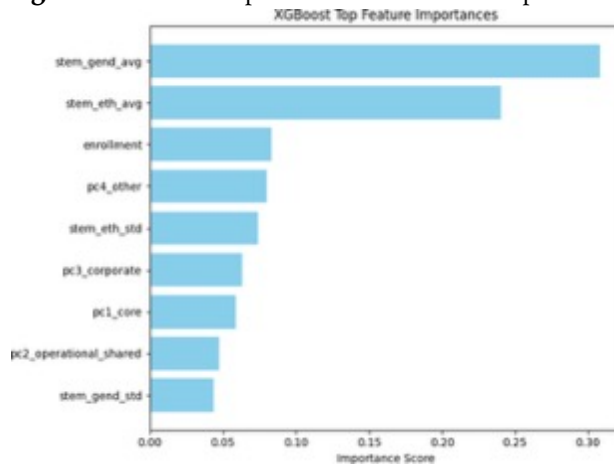
$$I(f) = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right) - \gamma$$

Where:

- I(f) = gain function of feature f
- G, H = sum of gradients and Hessians
- $G_{L,R}$, $H_{L,R}$ = G and H for child nodes
- λ = L2 regularization term
- γ = minimum loss reduction per split

The prominence of proficiency averages across ethnicity and gender was expected, under- lining the importance of early-stage performance as a proxy indicator of college and career readiness. However, disparity indexes ranked relatively lower, although performance disparity across ethnicities was significant at around 0.075. Latent structures formed through the principal components resulting from PCA were observed with considerable effect on predictions. PC4 (other and miscellaneous funding) and PC3 (corporate funding) were significant, thus validating the structural retainment of PCA and suggesting the creation of underlying socioeconomic signals that affect SAT benchmark percentages. Notably, enrollment was ranked third out of nine, at approximately 0.08 importance. The predictive power of enrollment suggests differences in funding allocation to school districts based on size, which may affect SAT scores and overall academic quality.

Figure 4: Feature Importance Scores for Composite Ensemble Model



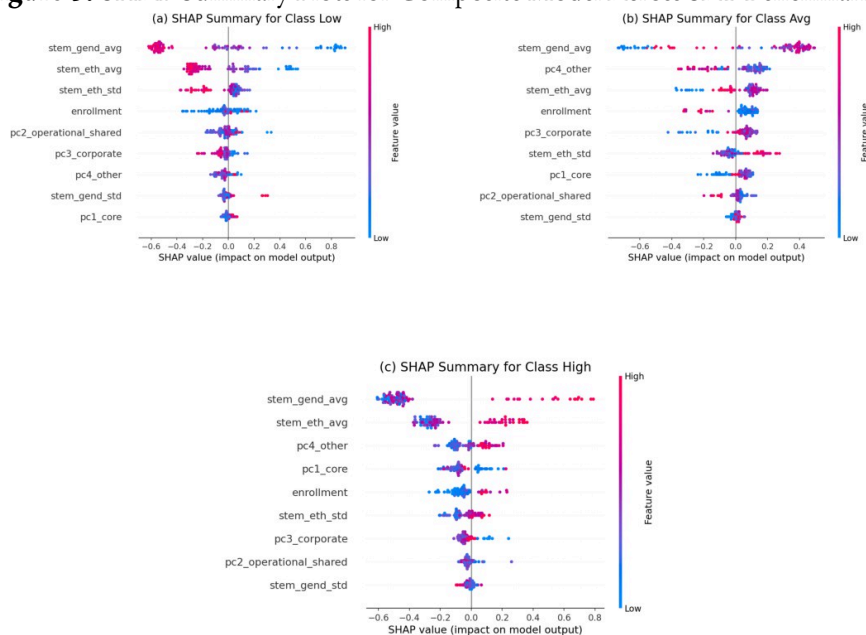
Note. This figure presents the normalized feature importance scores (summing to 1.0) for the nine predictor variables used in the composite model. Early-stage STEM proficiency averages across ethnicity and gender contribute the largest importance values. Funding-related principal components (PC3: Corporate Funding; PC4: Other/Miscellaneous Funding) and district enrollment also demonstrate substantial influence, while disparity indices show comparatively smaller but non-negligible contributions.

To examine the individual relationships between top features in predicted class of SAT benchmark percentages, SHAP (Shapley Additive Explanations) visualizations were created, as shown in Figure 5. The three figures each represent SHAP summary plots for classes Low, Average, and High, respectively. The beeswarm visualizations indicate the concentration of variables, with red points and blue points representing high and low values, respectively. For Low class predictions, the averaged early-stage STEM proficiency across genders is prominent, ranging from -0.6 to 0.9. There is moderate stratification of low and high values: a blue tail was observed around 0.2–0.9, contrasting the red clustered at -0.6 and -0.5. As low blue values of early-stage performance correspond strongly with a Low label, the SHAP summary confirms interpretability. Notably, a red tail of high values in early-stage STEM proficiency disparity across ethnicities from -0.4 to -0.2 raises a slight discrepancy. In the SHAP summary for Average class districts, a similar red tail is observed around 0–0.3, suggesting that the model gravitates toward Average predictions from high ethnic disparity. This tendency is likely due to the prominence of urban school districts, with a wider range of student ethnicities and moderate SAT scores.

Reinforcing model interpretability, the SHAP distributions of early-stage performance indicators (both across ethnicities and genders) remain intuitively consistent in all three classes. For example, a clear red tail of high early-stage performance was observed around 0.2–0.8 in the High class SHAP summary with a discrete,

concentrated of lower or moderate values around -0.5 (for early-stage performance across genders). The SHAP summary for the Average class is noticeably less stratified and correctly corresponds to a moderate mix of early-stage performance with Average SAT scores. Similarly to enrollment's prominence in gain feature importances (Figure 4), enrollment is significant in the SHAPs, suggesting the polarizing effect of enrollment-based school resource allocation on SAT scores. Notably, in the High class SHAP summary, red and blue points are clearly stratified. High values of enrollment are shown to push label predictions toward the High class, aligning with the greater amounts of funding large urban school districts receive from the state.

Figure 5: SHAP Summary Plots for Composite Model Across SAT Benchmark Classes



Note. This figure presents SHAP (Shapley Additive Explanations) summary beeswarm plots for the composite model's three predicted SAT benchmark classes: Low, Average, and High. Each plot displays the distribution of SHAP values for all input features, where red points represent higher feature values and blue points represent lower feature values. The horizontal axis reflects the magnitude and direction of each feature's contribution to class prediction. Early-stage STEM proficiency indicators, district enrollment, and funding-related principal components exhibit clear stratification patterns across classes, reflecting their relative influence in driving model predictions.

The representation of the principal components for funding, especially with other and corporate funding, indicating subtle structural signals that impact SAT performance. Specifically, the stratification of PC3, corporate funding, in the Low class SHAP summary is significant. High funding (red) are concentrated around

-0.25 to 0.0, whereas low funding (blue) extends from 0.0 to 0.2. This confirms a direct relationship between corporate funding amounts and academic quality. Downstream geospatial risk analysis examines corporate funding to determine the impact of geographical patterns in funding.

Geospatial Risk Distribution

A geospatial risk analysis was conducted to examine geographical patterns in significant features and adequately provide structure to future policy decisions. Four variables were mapped by school districts: ethnic disparity in early-stage STEM proficiency and corporate funding due to their predictive contributions (Figure 6), and district enrollment alongside binned SAT classes to visualize general targets of policies (Figure 7). In the ethnic disparity map shown in Figure 6a, key geographic patterns were observed. Disparity risks are not uniformly rural nor urban, but appear in scattered hotspots throughout the state. However, there are concentrations of moderate ethnic disparity in the Southwest corner (encompassing South Bend) and the North-central belt. Districts with high outliers of disparity above 0.35, displayed in dark red tend to gravitate toward central Indiana. Low disparity districts are prominent throughout the state, but especially concentrated in Southern Indiana, as seen in the multiple dark green districts.

Notably, the top districts for corporate funding significantly outperformed the bottom districts, reinforcing the direct relationship between strong corporate funding and performance. However, counterintuitively, the districts with lowest ethnic disparity underperformed compared to the most unequal districts, which actually classified in Average and High SAT bins. This discrepancy can be attributed to the drastic enrollment differences between districts. All three of the most ethnically unequal districts are mega-districts with over 4,000 students, whereas the bottom districts are comparatively tiny, rural schools. In fact, Tippecanoe's enrollment of 13,000+ students is over 60x larger than Medora's enrollment of approximately 200. The large size likely contributes to high disparity, despite moderate SAT scores.

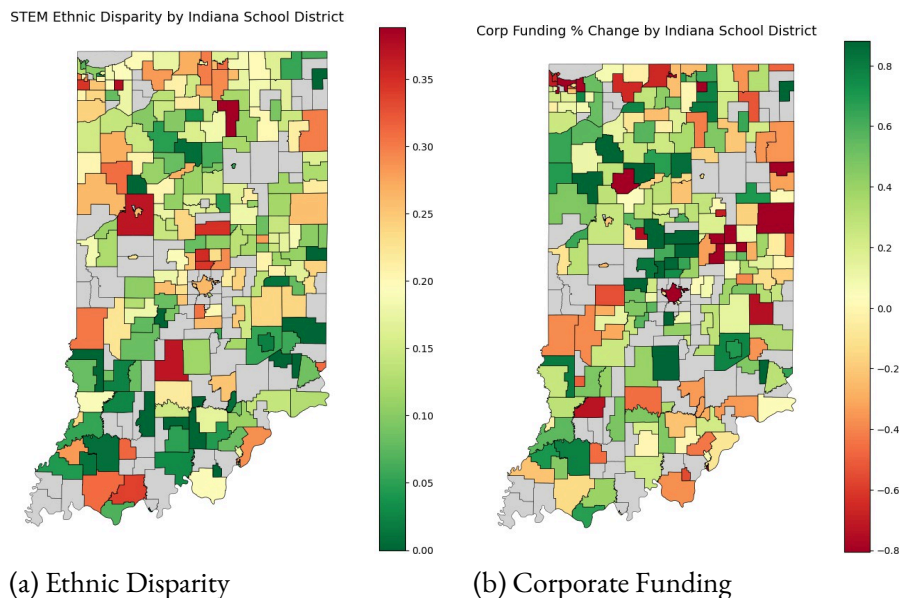
To further examine the effects of enrollment to SAT performance and future policy structuring, enrollment by school district was mapped in Figure 7a. The vast majority of schools have enrollments under 4,000 students, with dark red concentrations likely representing small rural districts. High-enrollment districts were observed scattered throughout the state, especially in Indianapolis (22,000+ enrollment), the Fort Wayne area (28,000+ enrollment), and several other hotspots in both Northern and Southern directions.

Significant patterns were also observed in the geospatial distribution of SAT classes. Most districts corresponded with the Average class, which was prominent throughout the state, and concentrations of high-performing districts were observed near the Indianapolis area. This central-Northeast cluster encompasses well-funded, top-ranked suburban schools encompassing Zionsville and Carmel. High-performance districts

SPRING 2026

were also scattered in Northwest suburbs and Southern peripheries. Low SAT class districts were especially concentrated in Southwestern rural belts, as seen in clusters of red, and the Northwestern industrial districts like Gary. These geographical patterns are realistic, as particularly low-performance districts align with historically underfunded urban or rural areas.

Figure 6: Input Variable Geospatial Distributions.



Note. This figure depicts the annual percentage change of corporate funding is mapped per school district. Key patterns of increasing funding in moderate and dark green are prevalent in Northern rural belts west of Fort Wayne and in central Indiana. Districts with decreasing funding (with -0.4% to -0.8% change annually shown in orange and dark red) were observed around the general area of Fort Wayne, with some districts scattered in the Southwestern direction. Tables 5 and 6 display the top and bottom 3 ranked school districts for the mapped features and corresponding SAT benchmark percentages.

Important observations were noted in Tables 7 and 8, which display the top and bottom 3 ranked districts for enrollment and SAT class. Most noticeably, the bottom three districts for enrollment are exactly the bottom districts for ethnic disparity, affirming previous hypotheses on the correlations of enrollment. The large mega-districts of Hamilton Southeastern and Fort Wayne significantly outperform these underfunded rural areas. The examination of the top and bottom ranked schools for SAT scores aligned with historical trends, with affluent Indianapolis suburbs dominating and the same rural suburbs (Medora, Cannelton, etc.) failing

benchmarks.

Top and Bottom Three Districts Ranked by Ethnic STEM Disparity and Corporate Funding Levels

Table 5: Ethnic Disparity

Top 3		Bottom 3	
Name	SAT	Name	SAT
Tippecanoe	0.42	Medora	0.0
Valparaiso	0.46	Cannelton	0.0
Warsaw	0.36	Eminence	0.21

Table 6: Corporate Funding

Top 3		Bottom 3	
Name	SAT	Name	SAT
Milan	0.25	Indy	0.04
NW Hendricks	0.39	Griffith	0.22
West Central	0.18	Marion	0.10

Note. These tables display the three highest- and lowest-ranked districts based on (Table 5) early-stage STEM proficiency disparity across ethnicities and (Table 6) corporate funding principal component (PC3) scores. Corresponding SAT benchmark percentages are included to illustrate performance differences across extremes of disparity and funding levels. The comparison enables examination of structural relationships between funding, demographic composition, and academic outcomes.

CONCLUSION

This paper reviews Indiana's educational history with a focus on STEM equity in unified districts and applies a machine learning framework for policy analysis. Random forest and XGBoost models were trained, cross-benchmarked, and combined to achieve 0.73 classification accuracy. Key factors, including early-stage ethnic disparities and corporate funding, were identified via XGBoost gain and SHAP plots, then mapped geospatially by district. These findings inform policy recommendations for the IDOE and other stakeholders.

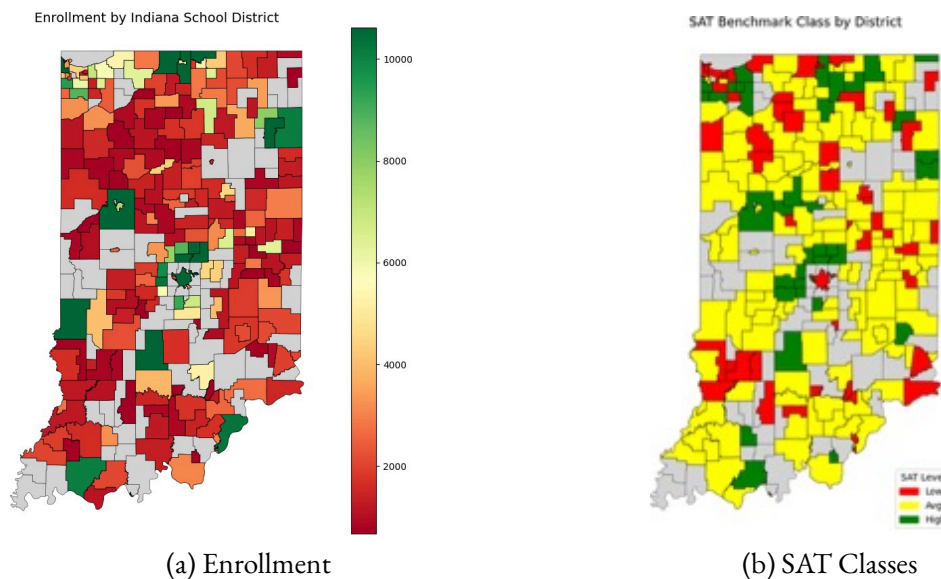
SPRING 2026

Policy Recommendations

Rural and Small District Funding

SHAP analysis and geospatial mapping revealed that district enrollment impacts SAT and STEM performance: small rural districts show low disparity indices but underperform in STEM. Policies should direct additional funds to districts with fewer than 1,000 students. Current IDOE efforts, like the 2024–2025 STEM Integration Grant, provide up to \$25,000 for districts under 1,000 students and \$50,000 for those with 1,000–1,999, covering teacher stipends, professional development, and curriculum integration. Similar support exists through programs like the U.S. Department of Education’s Rural and Low-Income School Program (U.S. Department of Education, 2024).

Figure 7: Geospatial Distribution of District Enrollment and SAT Benchmark Classification Across Indiana



Note. Panel (a) maps total student enrollment by district, with darker shading indicating higher enrollment levels. Panel (b) displays the geographic distribution of predicted SAT benchmark classes (Low, Average, High) from the composite model. The visualization highlights spatial clustering patterns in district size and academic performance, enabling comparison between enrollment scale and SAT classification outcomes.

SPRING 2026

Grant programs could incorporate mobile or remote STEM labs to provide hands-on experiences often missing in underfunded rural districts. A 2019 Chemistry Education Research and Practice study with 50 students found that 33.6% of learning gains were due to lab-based activities. Introducing such labs, even remotely or on a rotating basis, would help promote greater equity between rural and more affluent districts (Royal Society of Chemistry, 2019).

Top and Bottom Three Districts Ranked by Enrollment and SAT Benchmark Performance
Table 7: Enrollment

Top 3		Bottom 3	
Name	SAT	Name	SAT
Hamilton SE	0.60	Medora	0.0
Indy	0.04	Cannelton	0.0
Fort Wayne	0.16	Eminence	0.21

Table 8: SAT Classes

Top 3		Bottom 3	
Name	SAT	Name	SAT
Hamilton SE	0.60	Medora	0.0
Zionsville	0.67	Cannelton	0.0
Carmel	0.70	Gary	0.02

Note. These tables present the three largest and smallest districts by enrollment (Table 7) and the three highest- and lowest-performing districts by SAT benchmark percentage (Table 8). SAT benchmark values are included to facilitate comparison between district size and academic performance, providing additional context for the relationship between enrollment scale and predicted SAT classification.

Corporate and Community Partnerships

To expand hands-on STEM experiences, the IDOE should create state programs that match corporate donations or partnerships for career exposure and resources. Legislators should support existing initiatives and

SPRING 2026

engage new local companies. Programs like DirectEmployers Institute's STEM YES! (6 Indianapolis businesses hosting 12 students each) and RoseSTEM at Rose-Hulman provide underrepresented 9–12th graders in rural and urban Indiana access to STEM professionals (DirectEmployers Institute, 2024). Recently, 65 students from GEO academies in Indianapolis, Gary, Central Indiana, and Terre Haute attended STEM career events at Rose-Hulman. Recognizing top corporate or nonprofit partners can encourage long-term investment and visibility (Rose-Hulman Institute of Technology, 2024).

Early-Stage Intervention

Early-stage performance metrics across ethnicities and genders were strong predictors of SAT outcomes, as shown by SHAP analyses. Interventions are most effective in middle school, aligned with ILEARN testing. A 2022 NSF study at Purdue found that early STEM interventions significantly improved assessment scores, especially for students of color, with knowledge retained for over eight weeks (Wang & Degol, 2017). Implementing such policies allows the IDOE to track and evaluate intervention effectiveness over time.

REFERENCES

Center for Social Determinants of Health. (2023). 2023 State Test Score Results: Indiana (STSR Data Series 2023-02-IN-00) [Accessed January 2026].

DirectEmployers Institute. (2024). Stem yes program. <https://directemployersinstitute.org/our-programs/stem-yes>

Eli Lilly and Company. (2024). Key facts about lilly. <https://www.lilly.com/who-we-are/about-lilly/key-facts>

Indiana Department of Education. (2024). 2024–2025 stem grant guidelines. <https://www.in.gov/doe/files/2024-2025-STEM-Grant-Guidelines.pdf>

Indiana State Board of Education. (2020). 2020 Annual Indiana Advanced Placement Performance Report [Accessed January 2026].

Indiana State Library. (2023). The colored school. <https://www.in.gov/history/state-historical-markers/find-a-marker/the-colored-school/>

Purdue University. (2024). Usda employment opportunities. <https://www.purdue.edu/usda/employment/>

Reuters. (2024). Lilly announces new \$4.5 billion manufacturing site in indiana. [https : / / www.reuters.com/business/healthcare-pharmaceuticals/lilly-announces-new-45-bln-manufacturing-site-indiana-2024-10-02/](https://www.reuters.com/business/healthcare-pharmaceuticals/lilly-announces-new-45-bln-manufacturing-site-indiana-2024-10-02/)

Rose-Hulman Institute of Technology. (2024). Rosestem program preparing historically under- represented students for stem careers. <https://www.rose-hulman.edu/news/2024/Rosestem-program-preparing-historically-underrepresented-students-for-stem-careers.Html>

Royal Society of Chemistry. (2019). Factors influencing student engagement in chemistry edu- cation. <https://pubs.rsc.org/en/content/articlelanding/2019/rp/c9rp00021f>

U.S. Commission on Civil Rights. (2022). Civil rights and school funding: 2022 annual report. <https://www.usccr.gov/files/2022-12/cssbmb-2022-annual-report.pdf>

U.S. Department of Education. (2024). Rural and low-income school program. <https://www.ed.gov/grants-and-programs/formula-grants/rural-insular-areas/rural-and-low-income-school-program>

Wang, M.-T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and math- ematics (stem): Current knowledge, implications for practice, policy, and future directions. *Journal of Teacher Education*, 68 (2), 1–18. <https://onlinelibrary.wiley.com/doi/10.1002/tea.21756>