Iowa Initiative for Artificial Intelligence
Final Report

Project title:	Predictive Insights into U.S. Students' Mathematics and Science Performance on PISA Using Ensemble Tree-Based Machine Learning Models
Principal Investigator:	Hyesun You (PI, Teaching and Learning Department)
Prepared by (IIAI):	Zhi Chen
Other investigators:	
Date:	1/6/2026
Were specific aims fulfilled:	Y
Readiness for extramural proposal?	Y
If yes … Planned submission date	Drs. You and Chen submitted a Spencer Small Research Grant proposal on December 8, 2025, focused on AI-generated feedback for individualized science learning.
Funding agency	Spencer small grant (https://www.spencer.org/grant_types/small-research-grant)
Grant mechanism	The Spencer Small Research Grants program supports education research projects. The program provides funding of up to $50,000 per project. Proposals are evaluated based on the significance of the research problem, the strength of the conceptual framework, methodological rigor, and the project's potential contribution to the field of education research.
If no … Why not? What went wrong?	

Research report:

Specific Aim 1. To identify the most influential predictor variables (X) affecting U.S. students' mathematics and science performance on the PISA assessments from 2000 to 2022 by using ensemble tree-based machine learning models, including Random Forest, XGBoost, and LightGBM.

Specific Aim 2. To compare the accuracy and efficiency of machine learning (ML) methods with traditional statistical approaches for predictor selection in large-scale assessment data.

Specific Aim 3. To provide a detailed analysis of the relationships between predictor variables and student outcomes using descriptive statistics and clustering methods.

Introduction:
PISA (Programme for International Student Assessment) evaluates 15-year-olds' students' proficiency in reading, math and science across dozens of countries every three years (Organization for Economic Co-operation and Development, n.d.). This large-scale data informs policy makers and highlights the important factors affecting learning. The 2006 and 2015 years focused on school- and student-level predictors of science achievement, including the index of economic, social and cultural status (escs), representing "students' access to family resources (Avvisati, 2020)". To examine the complex relationships among these predictors, this study used

machine learning (ML) models to address the below questions:
1. What are the key factors that predict students' science achievement in PISA assessments?
2. How do these predictors differ between low- and high-performing students?
3. How do the relationships between the top 3 predictors and science achievement differ between 2006 and 2015?

By bridging machine learning models with traditional statistical methods, this study provides information about the key predictors of science achievement in addition to their change across two cycles.

Theoretical Framework:

This study is supported by the opportunity-propensity (O-P) framework (Byrnes & Miller, 2007) in which antecedent, opportunity, and propensity factors are categorized as predictors of student academic achievement. Background characteristics such as ESCS, which is one of the important variables predicting science achievement, are represented by antecedent factors (Desoete & Baten, 2022). Opportunity factors are the ones related to the school-level variables such as school size that provides some exposure to the resources (Ceulemans et al., 2017). Lastly, propensity factors refer to internal conditions of students such as self-efficacy (Hallman, 2014). The O-P framework is helpful when analyzing large scale assessments since it can deal with the complexity of different factors both on individual and contextual levels (Byrnes & Miller, 2007; Lewis & Farkas, 2017). Many PISA studies using traditional and ML methods have identified socioeconomic status (SES), self-efficacy, and motivation as key predictors of science performance (Avvisati, 2020; Hallman, 2014). XGBoost has proven to be effective while identifying relevant predictors across contexts (Acisli-Celik & Yesilkanat, 2023), while a study from Hong Kong highlighted the effect of SES on student performance (Ho, 2010). Integrating ML with statistical methods, as in You et al. (2025) shows that when these approaches are used together, they can uncover the contributions of each predictor for student achievement.

Method:

In this study, PISA 2006 and 2015 science datasets from the United States, which include both student- and school-level variables were used. Students' science achievement was measured by five plausible values (PV1SCIE to PV5SCIE) in each year, which provides multiple imputations of students' true ability.

In both years, only variables common to the PISA 2006 and 2015 were used. If a variable was missing in either year, it was excluded. The data were also filtered by the "USA" country code to include only U.S. students. Predictors included factors such as parental education and school size, while plausible values were excluded from the predictor set to avoid feature leakage. Figure 1 shows that missing rates were higher than 10% for "bsmj", "schsize" and "stratio" for school-level, whereas it was lower for student-level variables. Based on these, continuous variables were replaced by mean values to preserve sample size.

Moreover, correlation matrices show consistent patterns across years in Figure 2. Schoollevel variables had weaker correlations. Only PV1SCIE was used to represent science achievement, given the high correlation among all PVs

Three ML models were applied (Random Forest, XGBoost, and Support Vector Regression) to predict science performance using Python version 3.11. Students were split into low- and high-performing groups using the median of PV1SCIE. Data were split into training (80%) and testing (20%) for each year. To evaluate model performance, RMSE (Root Mean Square Error) was used on test data and SHAP summary plots were used to interpret the effects of each variable on our prediction. Lastly, regression models predicting science achievement were run separately for each of the five plausible values and relevant effect sizes were reported.

Results:

RQ1: What are the key factors that predict students' science achievement in PISA assessments?

The performance of ML models (Random Forest, XGBoost and Support Vector Regressor) was explored to see their predictive accuracy for science achievement scores for both 2006 and 2015 cycles of PISA data. Science achievement was measured using five plausible values (PV1SCIE to PV5SCIE). In addition, SHAP (SHapley Additive exPlanations) plots were used to show the summary of feature importance plots. Before the model training, all plausible values were removed from the feature set to prevent feature leakage. The models were also evaluated separately for low- and high-performing groups, using the median split of "PV1SCIE" each year.Table 1 presents average science scores (means) and RMSE values separately with groups (low vs high) and year (2006 and 2015) for each plausible value in addition to the comparison of each model. In terms of model performance, "RF" is seen to have the lowest RMSE values across most of the PVs for both groups, specifically for low-performing students. In contrast, "XGB" and "SVR" results showed comparable RMSE values across high-performing students.

Table 2 provides the average RMSE values for five different plausible values (PV1SCIE to PV5SCIE) for each ML model and student group. XGBoost model produced the lowest average RMSE values in the overall datasets for each year. However, the Random Forest model had slightly better performance for low-performing students. In addition, mean scores were higher in 2015 data with lower RMSE values for all models.

Since XGBoost model showed the best predictive performance across years and groups, SHAP plots for this model were selected to illustrate feature contribution. Table 3 highlights the top three important features of the SHAP analysis for XGBoost model. This table presents different trends between low- and high-performing students, showing the changes in the prediction of science performance on PISA in time.

Figure 3 SHAP summary plots show the comparison of overall feature importance for XGBoost model in each year. In both 2006 and 2015, "escs" (economic, social and cultural status) are highlighted as the top predictor overall. However, there is a shift in the distribution of other features. For instance, "envaware" (environmental awareness) is seen influential each year, but "stratio" (student-teacher ratio) is a more effective feature in 2015.

RQ2. How do these predictors differ between low- and high-performing students?
Figure 4 shows the comparison of feature importance for low vs high performing students. This figure complements the findings on Table 3 and reveals changes in learning environments across performance levels.

RQ3. How do the relationships between the top 3 predictors and science achievement differ between 2006 and 2015?
Figure 5 shows the actual vs predicted values in 2006 and 2015 for XGBoost model. The red dashed lines represent the linear regression fits between the actual and predicted scores while blue dots represent the residuals. The R2 value for 2006-year highlights that 48% of the variance in science achievement scores can be explained by the model, while it is around 40% in 2015-year. The overall graphs suggest that the model can produce relatively accurate predictions for the overall data though the grey points show some under- or over-estimation issues.

Figure 6 illustrates the effect of the number of estimators on RMSE for RF and XGB models in each year. XGB is seen to outperform RF when the number of trees increased specifically around 100 estimators, which was selected for each model prediction. The SVR was not included here since it does not rely on estimators.

In addition, Figure 7 shows the coefficients of the top 3 predictors of science achievement using SHAP each year with confidence intervals and the relevant coefficients for each variable. ESCS was seen as an important variable for each year while science motivation (scieff in 2006 and joyscie in 2015) stayed as a strong positive predictor in each cycle. Environmental awareness (envaware) in 2006 had a positive relationship whereas environmental optimism (envopt) in 2015 had a significant negative relationship. The adjusted R-squared values across the models were consistent, approximately 0.30 and 0.21 for 2006 and 2015 respectively. This indicates these predictors can explain about 21.6 % (escs, envaware and scieff in 2006) and 30 % (escs, envopt, and joyscie in 2015) of the variance in students' science achievement. The model equations below represent the relationship between science achievement and the top three SHAP features. and represents the predicted science score in 2006 and 2015 respectively.

Discussion:
In this study, ML methods were combined with traditional statistical applications to understand key variables affecting science achievement in PISA. The importance of students'

socioeconomic status (ESCS) in predicting science achievement highlights the effect of educational inequality and the need for policies that can address these disparities. Differences in predictors across student groups (low vs high) and years suggest a shift in the factors influencing science achievement. These findings show that model choice can be effective while evaluating learning patterns, specifically for low-performing students. The use of SHAP values was beneficial to uncover the contribution of predictors which might be missed by traditional methods, especially when the variables have weak correlations but meaningful effects.

Figure 1. Missing Data Summary Graph

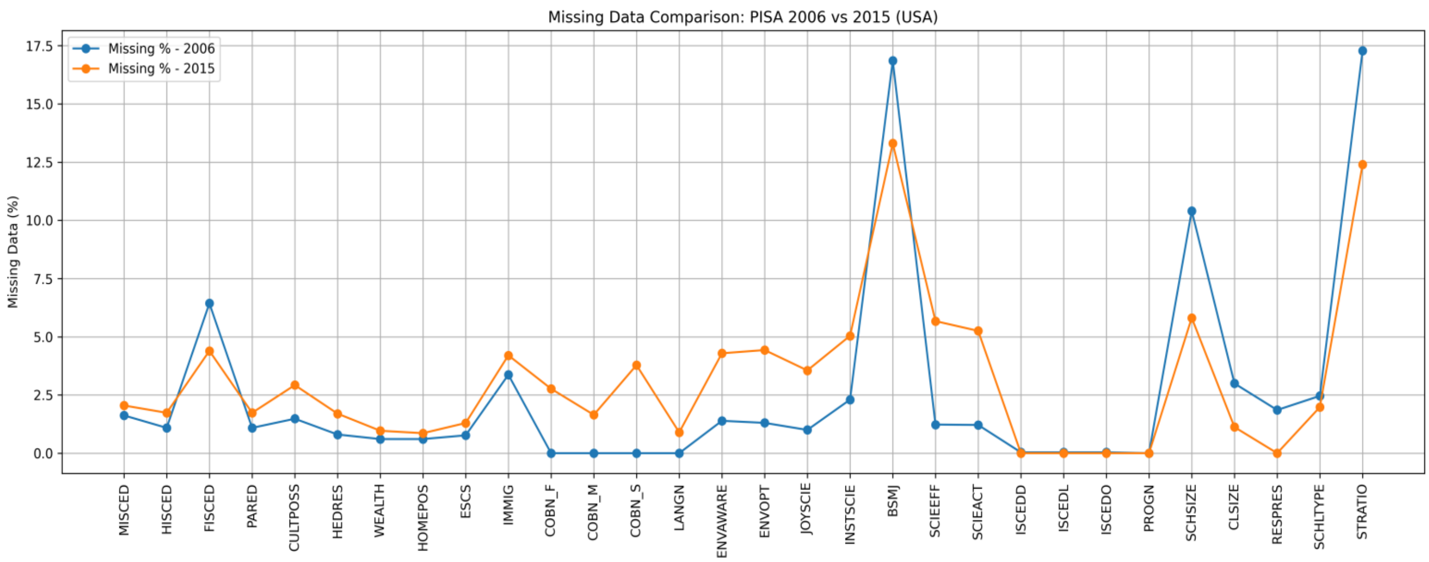


Figure 2. Correlation matrices for 2006 and 2015

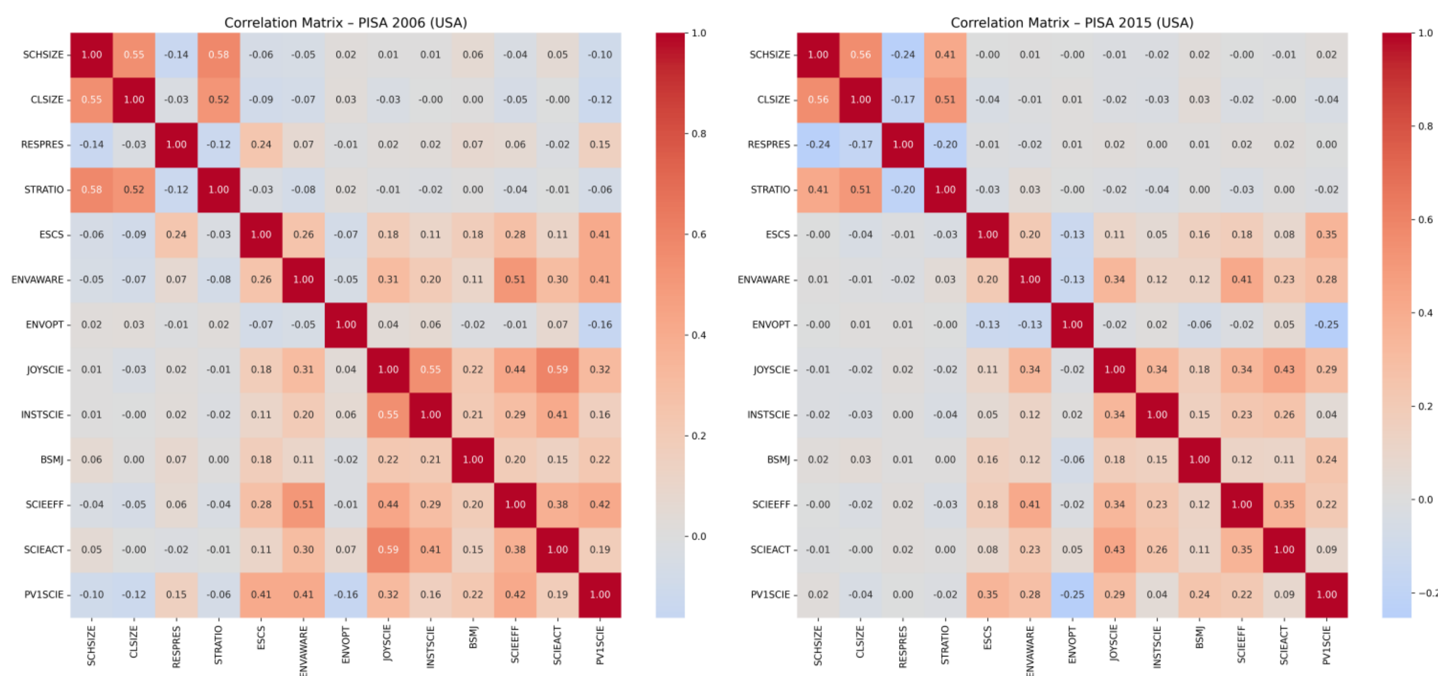


Table 1. Average Scores of Each Plausible (PVs) values

Year	PV	Group	Mean Score	RMSE (RF)	RMSE (XGB)	RMSE (SVR)
2006	PV1SCIE	Low	406.92	51.30	59.93	60.22
2006		High	569.69	53.43	59.99	57.92
2006	PV2SCIE	Low	406.98	49.68	58.03	58.95
2006		High	569.85	55.42	62.46	62.66
2006	PV3SCIE	Low	406.73	55.18	55.30	54.92
2006		High	569.80	53.96	57.15	56.52
2006	PV4SCIE	Low	407.03	56.29	65.82	66.69
2006		High	570.21	54.06	65.18	64.80
2006	PV5SCIE	Low	401.55	54.65	57.10	56.95
2006		High	574.29	56.06	60.65	60.87
2015	PV1SCIE	Low	416.29	51.69	60.08	59.53
2015		High	575.45	51.04	59.06	57.32
2015	PV2SCIE	Low	420.23	53.00	55.08	53.20
2015		High	570.16	52.83	56.03	55.47
2015	PV3SCIE	Low	420.88	52.46	60.22	59.42
2015		High	572.60	54.19	60.25	59.39
2015	PV4SCIE	Low	420.54	50.57	56.19	55.38
2015		High	572.01	54.09	63.22	62.51
2015	PV5SCIE	Low	421.31	52.86	56.72	55.92
2015		High	570.49	54.27	58.37	56.79

Note: Number of students in the low-performing group is 2930 while it is 2681 for the high-performing group in 2006, whereas their number is 2856 for both high-performing and low-performing groups in 2015.
(RF= Random Forest, XGB= XGBoost SVR= Support Vector Regression)

Table 2. Average Summary Table for all plausible values (PV1SCIE to PV5SCIE).

Year	*Group*	*Avg. Mean*	*Avg. RMSE (RF)*	*Avg. RMSE (XGB)*	*Avg. RMSE (SVR)*
2006	Low	405.44	53.82	59.66	59.54
	High	570.37	54.58	61.49	60.57
Overall	-	484.12	78.40	**76.45**	86.47
2015	Low	419.85	52.12	57.26	56.69
	High	572.54	53.29	59.39	58.30
Overall	-	496.20	74.63	**72.67**	83.38

Table 3. Showing Top 3 SHAP Features per year and low- vs high-performing group for XGBoost

Year	*Group*	*Top 3 SHAP Features*
2006	Low	School size (schsize), home educational resources (hedres), highest level of education attained by father (hisced)
	High	Student-teacher ratio (stratio), school size (schsize), economic, social and cultural status (escs)
Overall	-	Economic, social and cultural status (escs), environmental awareness (envaware), science self-efficacy (scieff)
2015	Low	School size (schsize), student-teacher ratio (stratio), economic, social and cultural status (escs).
	High	Science activities (sciact), student-teacher ratio (stratio), wealth index (wealth)
Overall	-	Economic, social and cultural status (escs), enjoyment of science (joyscie), optimism about the environment (envopt)

Figure 3. SHAP Summary Plots for 2006 and 2015

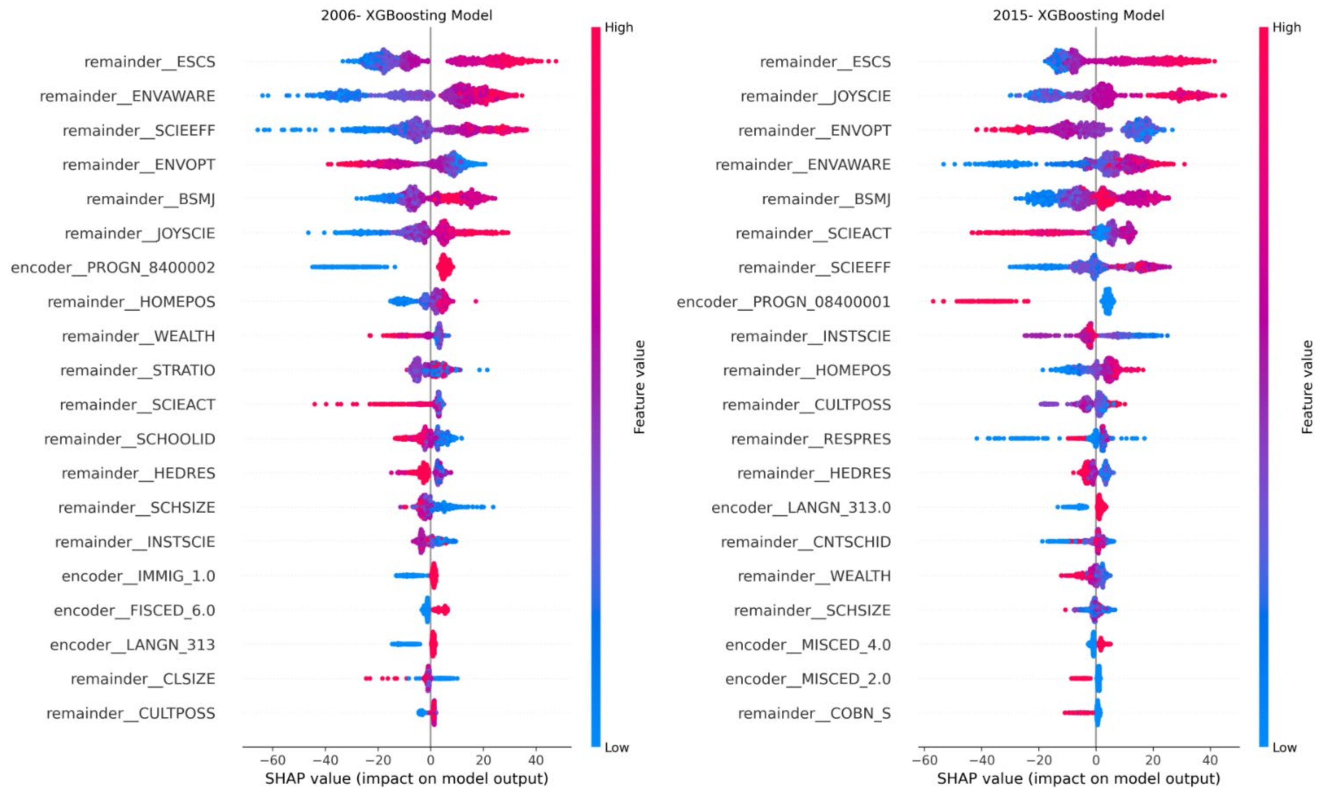


Figure 4. SHAP Bar Plots by Group and Year

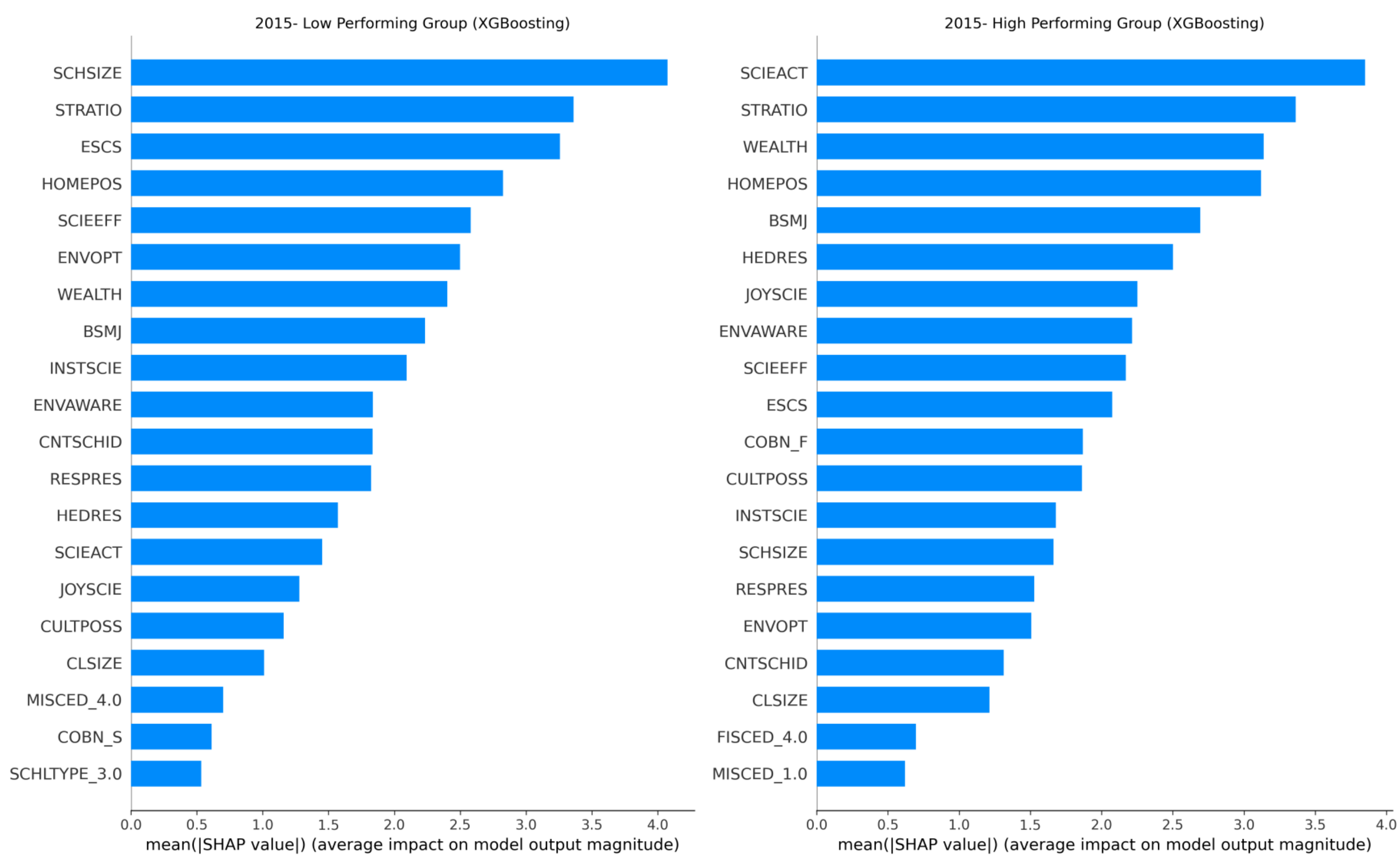


Figure 5. Regression Plots for XGBoost Model in 2006 and 2015

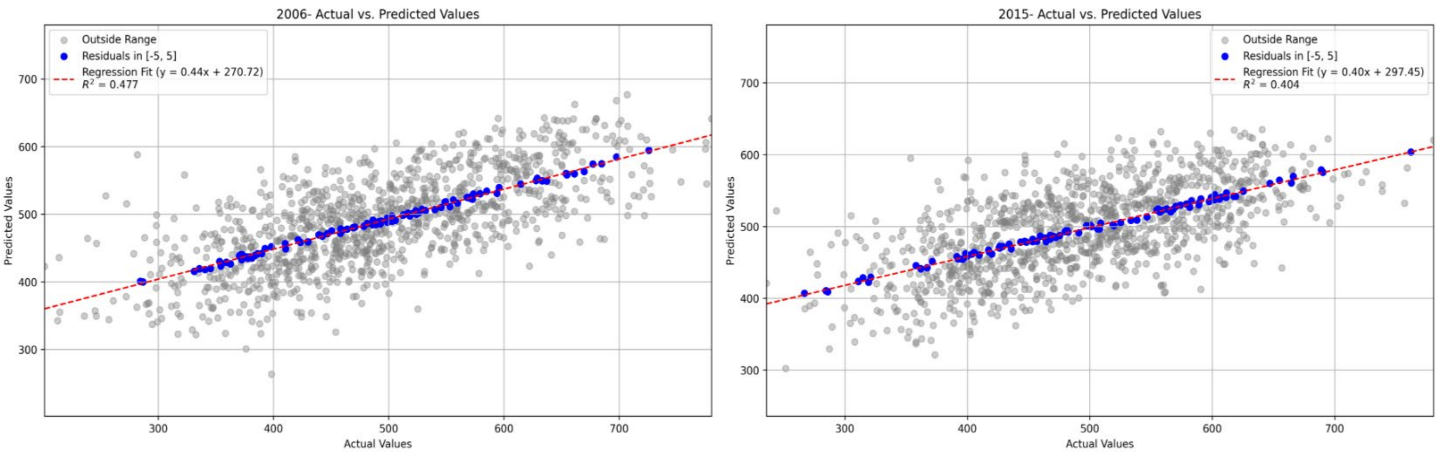


Figure 6. Parameter Tuning for 2006 and 2015

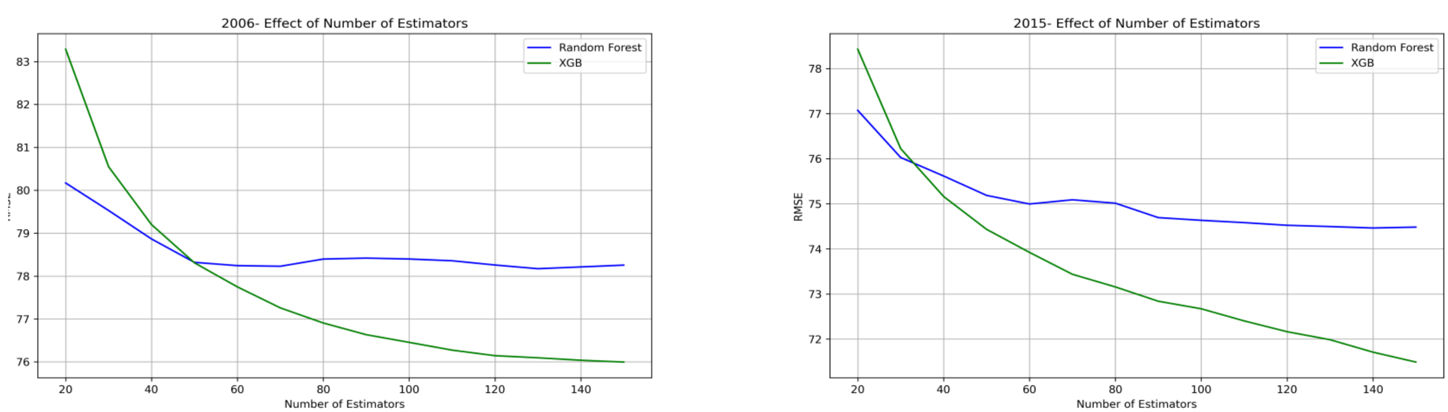
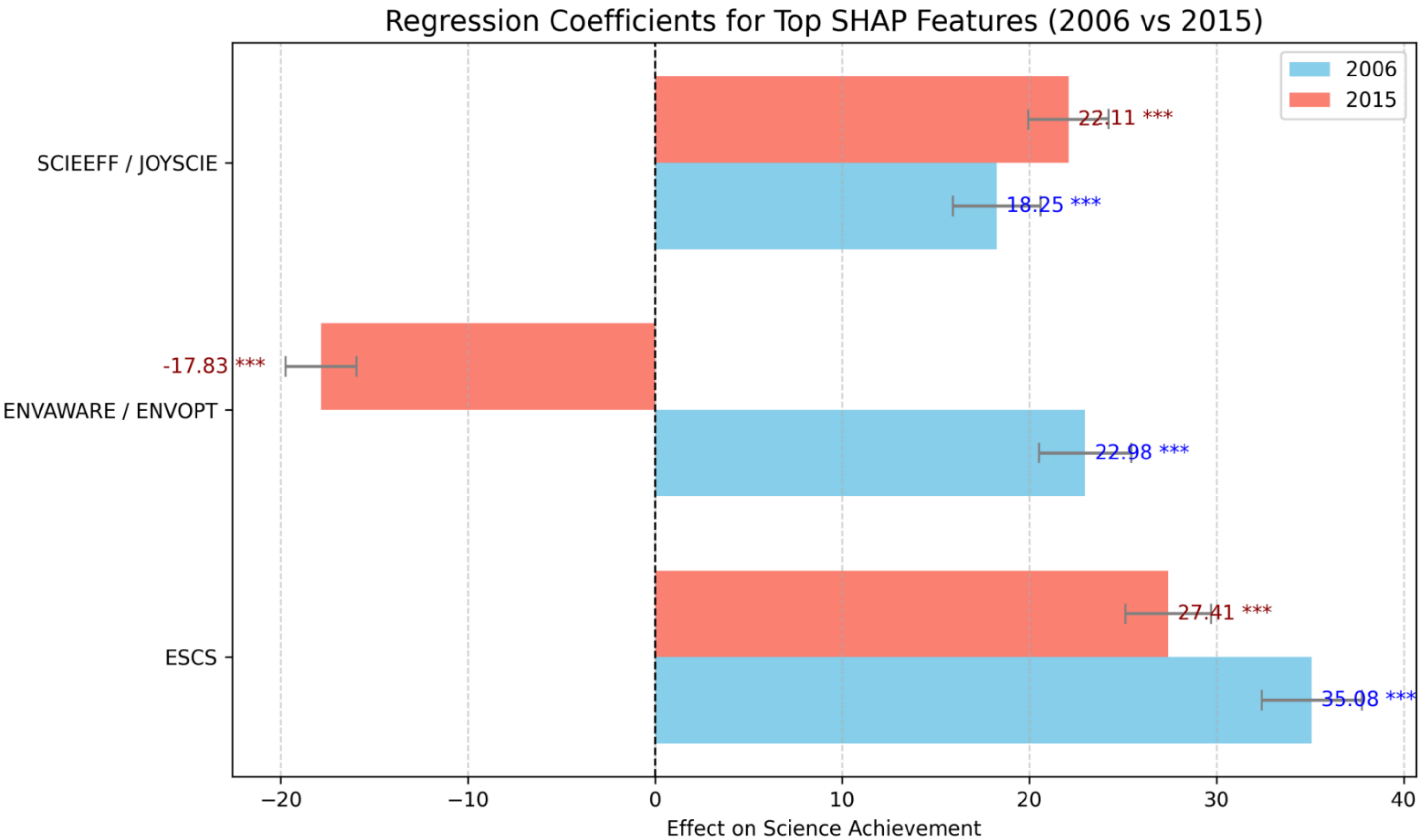


Figure 7. Regression Analysis for Top 3 SHAP Features in 2006 and 2015 (Note: *** shows the p values < 0)



References

Acıslı-Celik, S., & Yesilkanat, C. M. (2023). Predicting science achievement scores with machine learning algorithms: A case study of OECD PISA 2015–2018 data. Neural Computing and Applications, 35(28), 21201–21228.

Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. Large-Scale Assessments in Education, 8(1), 8.

Byrnes, J. P., & Miller, D. C. (2007). The relative importance of predictors of math and science achievement: An opportunity–propensity analysis. Contemporary Educational Psychology, 32(4), 599–629.

Ceulemans, A., Baten, E., Loeys, T., Hoppenbrouwers, K., Titeca, D., Rousseau, S., & Desoete, A. (2017). The relative importance of parental numerical opportunities, prerequisite knowledge, and parent involvement as predictors for early math achievement in young children. Interdisciplinary Education and Psychology, 1(1).

Desoete, A., & Baten, E. (2022). Math learning in Grade 4 and 5: What can we learn from the Opportunity–Propensity Model? International Electronic Journal of Elementary Education, 14(3), 213–225.

Hallman, L. M. (2014). Do inquiry-based science classroom experiences promote science achievement for public high school students? An application of the Opportunity–Propensity framework (Doctoral dissertation). Saint Joseph's University.

Ho, E. S. C. (2010). Family influences on science learning among Hong Kong adolescents: What we learned from PISA. International Journal of Science and Mathematics Education, 8(3), 409–428.

Lewis, R. W., & Farkas, G. (2017). Using an opportunity–propensity framework to estimate individual-, classroom-, and school-level predictors of middle school science achievement. Contemporary Educational Psychology, 51, 185–197.

Organisation for Economic Co-operation and Development. (n.d.). PISA (Programme for International Student Assessment). Retrieved from https://www.oecd.org/en/about/programmes/pisa.html

You, H., Hong, M., Zhu, L., & Zhenhan, F. (2025). Machine learning approaches for predicting U.S. students' scientific literacy: An analysis of key factors across performance levels and socioeconomic statuses. International Journal of Science and Mathematics Education, 1–29.

Ideas/aims for future extramural project:
Traditional machine learning methods are limited in their ability to effectively handle categorical data commonly found in educational contexts. As artificial intelligence tools—particularly large language models (LLMs)—rapidly enter higher education classrooms, educators face urgent questions regarding the quality, trustworthiness, and pedagogical value of AI-generated content. Our future study will address these concerns by examining how carefully engineered AI systems, fine-tuned on authentic student work and corresponding feedback, can support students' conceptual understanding, provide actionable guidance, and promote more equitable learning opportunities.

Publications resulting from project: Ms. Metesoglu and Dr. You submitted a paper titled Predicting Science Achievement: A Machine Learning Approach to PISA Data Analysis Across 2006 and 2015. This study leverages advanced machine learning techniques to analyze large-scale international assessment data from the Programme for International Student Assessment (PISA), focusing on trends and predictors of students' science achievement across two assessment cycles. By comparing data from 2006 and 2015, the project examines how key student, school, and contextual variables contribute to science performance over time, offering both methodological and substantive insights into educational achievement.

The conference proposal was submitted to the American Educational Research Association (AERA) Annual Meeting, one of the most prestigious and internationally recognized conferences in the field of education. AERA serves as a leading venue for disseminating cutting-edge educational research and for engaging with scholars from around the world.