Yıldız Social Science Review, Vol. 10, No. 1, pp. 7–27, 2024

Yıldız Social Science Review
Web site information: https://yssr.yildiz.edu.tr
DOI: 10.51803/yssr.1461030

Original Article / Orijinal Makale

Predicting Student Achievement via Machine Learning: Evidence from Turkish Subset of PISA

Makine Öğrenmesi Yöntemi ile Öğrenci Başarısının Tahmini: PISA Türkiye Örnekleminden Bulgular

Selin ERDOĞAN*, Hüseyin TAŞTAN

Department of Economics, Yildiz Technical University, İstanbul, Türkiye

Yıldız Teknik Üniversitesi, Ekonomi Bölümü, İstanbul, Türkiye

ARTICLE INFO

Article history
Received: 29 March, 2024
Accepted: 8 May, 2024

Keywords:
Economics of education, educational data mining, school effectiveness, student achievement, machine learning

MAKALE BİLGİSİ

Makale Hakkında
Geliş tarihi: 29 Mart 2024
Kabul tarihi: 8 Mayıs 2024

Anahtar kelimeler:
Eğitim ekonomisi, eğitimsel veri madenciliği, PISA, öğrenci başarısı, makine öğrenmesi

ABSTRACT

This study seeks to identify the determinants of academic performance in mathematics, science, and reading among Turkish secondary school students. Using data from the OECD's PISA 2018 survey, which includes several student- and school-level variables as well as test scores, we employed a range of supervised machine learning methods specifically ensemble decision trees to assess their predictive performance. Our results indicate that the boosted regression tree (BRT) method outperforms other methods bagging and random forest regression trees. Notably, the BRT highlights the importance of general secondary education programs over vocational and technical (VAT) education in predicting academic achievement. Moreover, both characteristics specific to student and school environment are demonstrated to be significant predictors of academic performance in all subject areas. These findings contribute to the development of evidence-based educational policies in Turkey.

Cite this article as: Erdoğan, S., & Taştan, H. (2024). Predicting Student Achievement via Machine Learning: Evidence from Turkish Subset of PISA. *Yıldız Social Science Review, 10*(1), 7–27.

ÖZ

Bu çalışma, Türk ortaokul öğrencileri arasındaki matematik, fen ve okuma akademik başarısının belirleyicilerini tespit etmeyi amaçlamaktadır. Bunun için, OECD'nin 2018'de düzenlemiş olduğu PISA çalışmasının öğrenci ve okul anketleriyle birlikte PISA test sonuçları ve gözetimli regresyon tabanlı makine öğrenmesi yöntemleri kullanılarak Türk orta okul öğrencilerinin akademik başarısını en iyi tahmin edebilecek model araştırılmıştır. Sonuçlarımız, yükseltme regresyon ağacı (BRT) yönteminin diğer yöntemler olan torbalama ve rastgele orman regresyon ağaçlarını geride bıraktığını göstermektedir. Yükseltme regresyon ağacı (BRT) yönteminden elde edilen bulgulara göre Türk orta okul öğrencilerinin akademik başarısını tahmin etmede öne çıkan değişkenlerden en önemlisi öğrencinin kayıtlı olduğu okulun program tipidir (Mesleki ve Teknik Orta Öğretim yerine Genel Orta Öğretimdir). Ek olarak, Türk orta okul öğrencilerinin akademik başarısını tahmin etmede hem öğrenci hem de okul düzeyindeki

* **Sorumlu yazar / Corresponding author**
*E-mail address: altayselin.o@gmail.com

değişkenler öne çıkmaktadır. Söz konusu bulgular her ders için geçerlidir. Bu bulgular, Türkiye'de kanıta dayalı eğitim politikalarının geliştirilmesine katkıda bulunmaktadır.

Atıf için yazım şekli: Erdoğan, S., & Taştan, H. (2024). Predicting Student Achievement via Machine Learning: Evidence from Turkish Subset of PISA. *Yıldız Social Science Review, 10*(1), 1–27.

1. INTRODUCTION

This study aims to develop machine learning (ML) models to identify the predictors of academic performance in mathematics, science, and reading among Turkish secondary school students. Using student- and school-level data from the Turkish subset of the OECD's PISA 2018 survey, tree-based machine learning techniques are employed to explore the relationships between academic achievement and various demographic, socioeconomic, and scholastic factors. Unlike previous studies conducted in Turkey, which often used pre-selected models and focused on a particular subject, this study compares different ML models to identify the best predictor of student performance with the highest accuracy. Moreover, this investigation delves deeper into performance predictors across various school types. By utilizing data from a large-scale international assessment and applying supervised ML techniques such as ensemble decision trees, the study aims to construct an accurate model that can predict the academic performance of Turkish students outside of the sample. Not only the best ML method with the highest predictive accuracy for forecasting Turkish students' academic performance but also the most critical predictors of student success within a highly competitive learning system characterized by early tracking of students has been assessed within the context of present study. Additionally, comparison between the most successful high school type in Turkey and the predominant high school types is conducted in terms of the factors influencing students' performance. Furthermore, we assess the robustness of our findings by formulating models for each subject. This research augments the burgeoning field by presenting insights from a developing country, characterized by its youthful demographic and recent educational reforms. Consequently, the findings bear implications not just for similar developing nations but also for developed countries striving to bolster student academic success.

Enhancing the quality of education hinges on the ability to predict student performance accurately. ML techniques applied to educational assessment surveys have recently become widespread in identifying the factors that contribute to students' success. ML approach is particularly suitable to the prediction policy problems where the causal inference is not the primary concern (Kleinberg et al., 2015). Rather than establishing causality between specific variables and student achievement, the focus is on identifying the variables that are most strongly associated with academic performance and using this information to develop accurate predictive models. By applying an ML approach, we can explore complex relationships between a wide range of demographic, socioeconomic, scholastic factors and academic achievement. With this information, educators can make necessary instructional improvements and plan personalized support for students who are likely to perform poorly. By identifying students at risk of poor academic performance beforehand, policymakers and educators can provide better guidance and early interventions, moving away from a one-size-fits-all approach to education. For example, recent studies using ML techniques have highlighted importance of socioeconomic status and self-efficacy in academic achievement (e.g., Chen et al., 2019; Dong & Hu, 2019; Gabriel et al., 2018; Hu et al., 2022; Lee & Lee, 2021; Martinez-Abad et al., 2020; Masci et al., 2018; Puah, 2021; She et al., 2019; Yoo, 2018; among others). Consequently, policymakers can aim to foster the development of self-efficacy in students by channeling more resources to this area, and schools can provide remedial classes to enrich the school environment in terms of arithmetic and literacy. Despite the challenges in mitigating the negative impacts of low socioeconomic status and lack of home endowments, the ultimate goal is to support all students in achieving academic success using available tools.

In Turkish education system, accurately predicting student performance holds particular importance due to its centralized and intensely competitive nature. Students in Turkey are tracked early, undergoing central external exams as young as 13, marking one of the earliest among OECD countries. Moreover, Turkish students are required to take two critical competitive central external exams that determine their high school and university placements. The high school placement is particularly influential since it often predetermines achievement in the consequential university entrance exam. Students are funneled into distinct high school categories based on aptitude and academic history.

Despite these high stakes, Turkish students have lagged behind the OECD average in core subjects, as reflected in the six PISA assessments since 2003. Specifically, the PISA 2018 assessment revealed a stark contrast: merely 3% of Turkish students were top-tier performers in reading, 5% in mathematics, and 2% in science. This pales compared to OECD averages of 9%, 11%, and 7% for the same subjects.

Given Turkey's underperformance relative to OECD average, there's an urgent need to nurture both advanced

and foundational literacy skills in Turkish youth across principal subjects. With a significant portion of Turkey's demographic comprising the youth, forecasting student performance in imminent exams becomes imperative. Appropriate interventions—whether at the individual, school, or societal level—are crucial to ensure every student realizes their potential. Therefore, predictive insights into student academic achievements are indispensable for tailoring the education system to be more sensitive and effective. Engaging in a predictive approach empowers all relevant parties including policymakers, educators, researchers, students and society at large to not only forecast outcomes but also derive novel understandings from the vast expanses of educational data.

The rest of the study is organized as follows: Section 2 reviews the literature, Section 3 introduces the Turkish education system, Section 4 describes and analyses the data, Section 5 outlines the methodology, Section 6 presents results from the ML framework, Section 7 discusses empirical results and policy implications, and Section 8 concludes with a brief summary.

2. LITERATURE REVIEW

Within the field of education research, a considerable amount of work has focused on education production, which typically examines student academic achievement as an outcome influenced by various student background characteristics (e.g., Hanushek, 1979; Hanushek & Kimko, 2000; Lee & Barro, 2001; Şirin, 2005; Woessmann, 2008; among others). Another strand of literature has recently emerged that aims to predict student academic performance and identify the most influential factors. This approach, known as Educational Data Mining (EDM), harnesses data mining, ML, and statistics to derive insights from educational data. The term EDM can be traced back to studies by Romero and Ventura (2007, among others), and Baker and Yacef (2009) who describe it as a research area aimed at solving educational questions by means of analyzing educational datasets through data mining techniques.

Both Educational Production Function (EPF) and EDM are rooted in Walberg's (1981) educational productivity theory, which treats student learning as a production process influenced by various institutional, academic, demographic, and economic factors, akin to the Cobb Douglas economic production function. This theory underscores the importance of input complementarity, rejecting the notion that any single factor independently affects academic performance. The way that these strands differ from each other depends on the purpose of the study. EPF studies aim to establish each variable's significance in student achievement through structural equations, while EDM studies seek to identify and predict relevant variables for achievement. EDM studies contend that due to the complex and nonlinear interactions among numerous factors influencing learning and achievement, ML methods can aid in selecting critical variables without assuming a predefined functional form.

In the realm of EDM literature, investigations into predicting student performance unfold on multiple dimensions, ranging from single country studies that delve deep into the intricacies of local contexts to multi-country analyses that illuminate broader global trends and correlations. The study by Yu et al. (2012) compared science achievement between the USA and Canada using PISA 2006, and emphasizes science enjoyment, use of software for educational purposes, interest in science and home possessions in addition to number of books at home out of school-related variables, student related variables in terms of home resources and attitude towards science, and ICT-familiarity. Kılıç Depren (2018) conducted a similar comparison between Turkey and Singapore in science achievement at PISA 2015 using decision trees and multivariate adaptive regression splines. The most influential factors for Turkish students were environmental optimism and awareness, science learning time, home possessions, epistemological beliefs about science, socioeconomic status, inquiry-based science teaching, learning practices, family wealth, and test anxiety. For Singaporean students, the most pronounced factors were mathematics learning time, teacher-directed science instruction, school disciplinary climate, student self-efficacy, unfair teacher practices, ICT availability at school, environmental optimism and awareness, home possessions, and socioeconomic status. Another cross-country study by Puah (2021) for scientific literacy in 60 countries participating in PISA 2015 identified variables such as students' environmental awareness, technological and educational resources at home, science self-efficacy, epistemological beliefs about science, home possessions, family wealth, ESCS, and total number of science teachers at school as important predictors. Using PISA 2015, Chen et al. (2021) classified top-performing students in science across 58 countries using SVM. They found key factors distinguishing top performers, including feedback on science learning, parental education, teacher-student ratio, teacher's adaptation and instruction, disciplinary climate, early childhood education, and self-confidence in science. Building on this, Hu et al. (2022) used SVM to differentiate high performers from low performers in PISA 2015 science. Besides the factors mentioned earlier, they considered variables like students' educational aspiration, gender, and teacher collaboration, immigration status, gender, home possessions, all affecting students' performance in secondary science education.

Capabilities of ML in this realm extend beyond just science subject. For nine developed countries, including Australia, Canada, France, Germany, Italy, Japan, Spain, the UK, and the US, Masci et al. (2018) provided evidence on the impact of student and school characteristics on mathematics literacy using PISA 2015. Employing BRT, they found that students' self-anxiety toward tests, socioeconomic status and self-motivation were among the most

important variables. Similarly, Kasap et al. (2021) studied reading comprehension of secondary school students in nine countries, including Indonesia, Colombia, Saudi Arabia, Serbia, Hungary, Slovenia, the US, Finland, and Turkey. Home possessions, students' perception of the difficulty level of PISA, parental occupation and education, student's expected occupational status, and enjoyment in reading activities were identified as important predictors of students' reading performance for all countries. It's important to note that Kasap et al.'s (2021) results relied on averaging 10 plausible values in reading, which, according to the PISA Data Analysis Manual, could lead to flawed results compared to using individual values. Lee and Lee (2021) investigated the impact of country and school-level factors on mathematics achievement in 76 countries participating PISA 2018. They used light gradient boosting and found that number of researchers in R&D, proportion of youth not in education, training or employment, student behavior hindering learning, student's mathematics learning time, grade repetition, ESCS, home ICT resources, student's perception of PISA test's difficulty, student's occupational aspiration, self-efficacy and -awareness related to global issues are the most informative in predicting student achievement.

Alongside multi-country studies, there are single-country studies that provide focused insights into specific national contexts. Gorostiaga and Rojo-Alvarez (2016) utilized supervised ML methods to identify successful Spanish students in the top 25th percentile of the PISA 2009 mathematics test. They found that gender, immigration status, parental occupational status, home cultural possessions like books and works of art, ICT availability and usage, reading engagement, regional and communal factors, and out-of-school lessons were associated with successful performance. Gabriel et al. (2018) investigated mathematics competency of Australian students in PISA 2012 using BRT. They found that self-efficacy and socioeconomic status are the key variables. By employing elastic net variable selection model, Yoo (2018) examined mathematical literacy among Korean students in TIMSS 2011, confirming the importance of factors like home possessions, parental support, self-confidence and self-efficacy. She et al. (2019) used the decision tree method to distinguish high- and low-performing Taiwanese students on PISA 2015, highlighting factors such as epistemic beliefs, learning time, and student interest in science. Rebai et al. (2020) analysed Tunisian students' performance using random forest with PISA 2012, identifying characteristics associated with higher performance as school size, competition, class size, parental pressure and the proportion of girls. Dong and Hu (2019) employed SVM to classify readers in Singapore's PISA 2015 test, noting influential factors including socioeconomic status, student learning time, school size, home possessions, and teacher participation. Martinez-Abad et al. (2020) used decision tree analysis to explore effectiveness in Spanish secondary schools across various subject areas in PISA 2015, identifying factors like academic extracurricular activities, ICT use, school lesson hours, and school principal autonomy.

Research on identifying factors that influence student academic performance at large-scale assessments has been conducted in Turkey as well. Kıray et al. (2015) examined variables affecting mathematics and science achievement among Turkish secondary school students using the decision tree with TIMSS 1999, PISA 2003, and PISA 2006. They found that self-concept, interest, motivation, self-efficacy, anxiety, attitude and problem-solving skills are influential. Additionally, under/achievement in one subject impacts the other. In the same vein, Aksu and Güzeller (2016) used PISA 2012 mathematics scores for Turkey and decision trees, highlighting self-efficacy, attitude, and studying discipline as key factors in classifying successful and unsuccessful students. Likewise, Filiz and Öz (2019) analyzed the Turkish subset of TIMSS 2015 science subject using logistic regression and SVM. They identified home educational resources, student self-confidence, computer/tablet ownership, extra science lessons, and educational aspiration as predictors of science achievement. Uğuz et al. (2021) employed Naïve Bayes, K-NN, and Random Forest classification methods with Turkish students' PISA 2018 science scores. They supported previous findings regarding the importance of science learning time, ICT usage at school, and students' perception of ICT proficiency. However, as mentioned previously, averaging plausible values at the student level, as opposed to using individual values, may yield flawed results.

Aligned with previous research on factors affecting student success, the body of work delves into the prediction of student achievement, recognizing the complexity of elements influencing performance in large-scale assessments. Employing supervised ML methods like BRT and elastic net variable selection, as well as classification methods such as logistic regression and SVM, these studies underscore the significance of considering a broad spectrum of factors, including psychological dispositions and demographic variables. This research, adopting a comprehensive approach, investigates the determinants of student performance as students near the end of their compulsory education in Turkey, drawing upon extensive data from individual students gathered through PISA. This research contributes to the knowledge base for policymakers and educators seeking to enhance student academic performance in large-scale assessments, while recognizing the need for further research to explore the influence of these factors in different cultural and socioeconomic contexts.

3. TURKISH EDUCATION SYSTEM

The formal education system in Turkey includes preschool, basic, secondary and higher education, and it is offered by both publicly and privately. Public schools in Turkey provide education free of charge. Before 1997, basic education was five years of primary and three years of

middle school. Since 1980, attending primary school and completing a minimum of five years of education has been mandatory in Turkey. A new law unified primary and secondary schools in 1997 and extended obligatory education from five to eight years. By integrating basic and secondary education in 2012, obligatory education was extended to 12 years by the latest law. The 2012 education system overhaul increased compulsory schooling from 8 to 12 years, with four years for primary, secondary, and high school. Despite government efforts to boost school enrolment, this reform has presented obstacles. The starting age for schooling has decreased, there are more religious schools, the curriculum emphasizes religious material more than secular education, and early school dropouts are higher. These flaws in the Turkish education system have hurt education quality and student achievement. Table 1 presents the number of highschools and students in the education year of 2017-2018 presented by Turkish Ministry of Education (MEB) National Education Statistics. It can be observed from the table that most of the students are enrolled at general secondary education schools yet the proportion of students enrolled at vocational and technical and imam-preacher highschools is substantial.

Figure 1 depicts a snapshot of Turkish students' performance in reading, science and mathematics subjects in PISA 2018 assessment. The relative position of Turkey with respect to other OECD countries in terms of key competencies and knowledge levels of 15 year-old individuals reveals that, students in Turkey scored lower than OECD average in each subject area. Moreover, a lesser percentage of Turkish students demonstrated proficiency at the uppermost tiers (Level 5 or 6) in any subject, while at the same time a reduced proportion of students had a baseline level of proficiency (Level 2 or above) in any subject.

Figure 2 presents the evolution of Turkish students' achievement in reading, mathematics and science subjects at the PISA test over time. When considering results from all years, 2015 is marked as considerably the lowest for all subjects. Besides, although there is an upward trend

Table 1. The Distribution of Highschool Students to School Types in Turkey, 2017-2018 Education Year

	Total Number of Schools	Total Number of Students	Percentage of Students
All Secondary Education Institutions	11783	5.689.427	100
General Secondary Education	5717	3.074.642	54
VAT Secondary Education	4461	1.987.282	35
Imam-Preacher Secondary Education	1605	627.503	11

Note: According to MEB National Education Statistics, general secondary education institutions include Anatolian, Science, Social Science, Sports and Fine Arts highschools, and vocational and technical secondary education institutions include Vocational and Technical, and Multi Programme highschools.

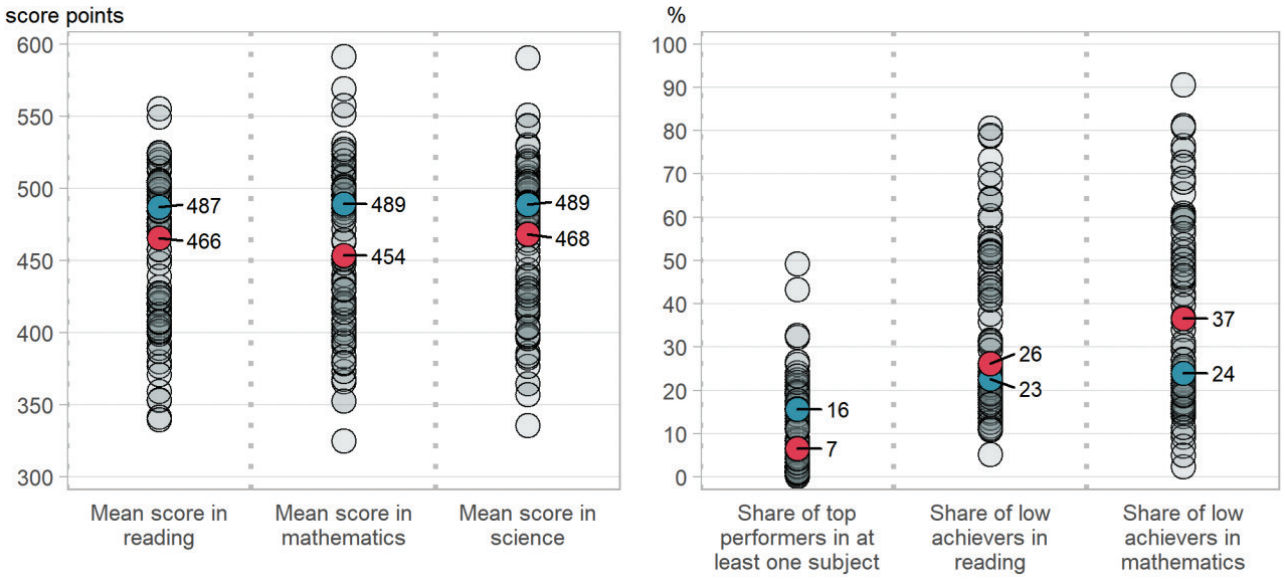

Figure 1. Relative Performance of Turkey with respect to Other OECD Countries. Notes: Red dote represents Turkey. Blue dote represents OECD average. Grey dote represents other OECD member countries. Source: OECD, Turkey Country Note, PISA 2018 Results, Retrieved from: https://www.oecd.org/pisa/publications/PISA2018_CN_TUR.pdf

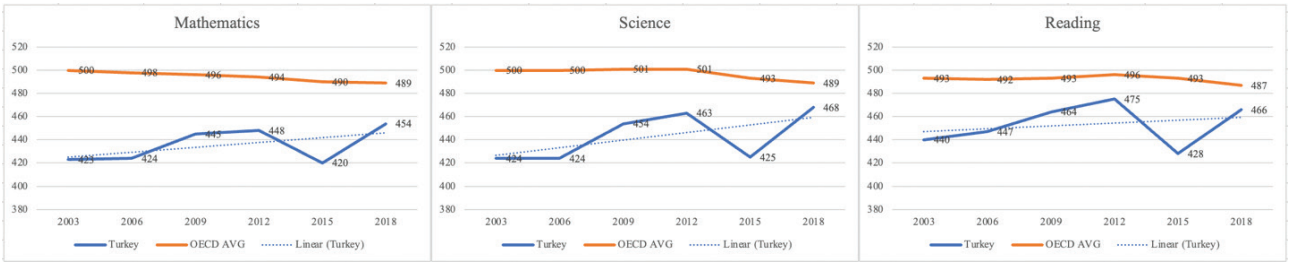

Figure 2. Evolution of Turkish Students' Performance in Reading, Mathematics and Science Subjects Over Time. Notes: Blue line represents Turkey. Orange line represents OECD average. Blue dotted line represents the trend in Turkey's average scores in each subject area.

Table 2. Descriptive Statistics of Outcome Variables in Each Subject from PISA 2018

	PV1MATH	PV1SCIENCE	PV1READING
Minimum	230,881	243,397	241,229
Maximum	778,445	725,275	748,371
Range	547,564	481,878	507,142
Standard Deviation	83,379	78,995	82,672
Mean	470,319	484,097	483,773
Median	466,5055	482,365	482,853

in Turkey's average scores, they have remained below the OECD average.

4. DATA AND DESCRIPTIVE ANALYSIS

The study uses PISA test and surveys conducted by OECD, focusing on the Turkish subset of PISA 2018 with data from 6890 students in 186 schools. The rationale for utilizing the 2018 dataset in this study stems from its coverage of the period preceding the widespread transition to remote learning systems in schools due to the well-known pandemic of 2018. Through this approach, it becomes feasible to sterilize and hence evaluate students' performance antecedent to any potential deleterious impacts attributable to the aforementioned pandemic. The goal is to explore the relationship between student achievement and various educational indicators. Plausible values[1] (PV1MATH, PV1SCIENCE, PV1READ) represent literacy skills in each subject. Input variables include responses from student and school questionnaires, examining 71 measurable student- and school-level factors. After data cleaning, the study analyses 3876 observations from 160 schools. Table A in the Appendix details variables and explanations.

Table 2 presents the descriptive statistics for the first plausible values of each subject from the Turkish subsample of PISA 2018. On average, Turkish students performed better in science and reading compared to mathematics. However, the variability in test scores was highest for mathematics, followed by reading and science. The median was found to be the lowest for mathematics, indicating that half of the students scored below 466.51 in this subject. Overall, these results suggest that Turkish students may need more support and resources to improve their performance in mathematics compared to science and reading.

Figure 3 presents box plots showing the distribution of Turkish students' mathematics test scores in PISA 2018, grouped by the type of school program they are enrolled in. Median values of the first plausible values of mathematics within each highschool type are represented by the horizontal line on the boxes. There is a significant disparity in scores across different high school types. The highest median score, marked at 599.83, is associated with the Science High Schools (Sc), Social Sciences High Schools (SSc) and Anatolian High Schools (An) also have higher average scores than the national average of 470. Since admission to these schools is highly competitive and based on performance on a nation-wide exam, it is not surprising that students from these schools perform well in terms of mathematical literacy. Anatolian Imam and Preacher (An), Anatolian Sports and Fine Arts (SFA), Multi-Programme Anatolian (MP) and Vocational and Technical (VAT), and Lower Secondary Schools (LS) have average scores slightly over 400 points, which is less than the national average.

[1]There is, theoretically, no minimum or maximum score in PISA. Scores are scaled to fit a normal distribution with a mean of 500 points and a standard deviation of 100 points.

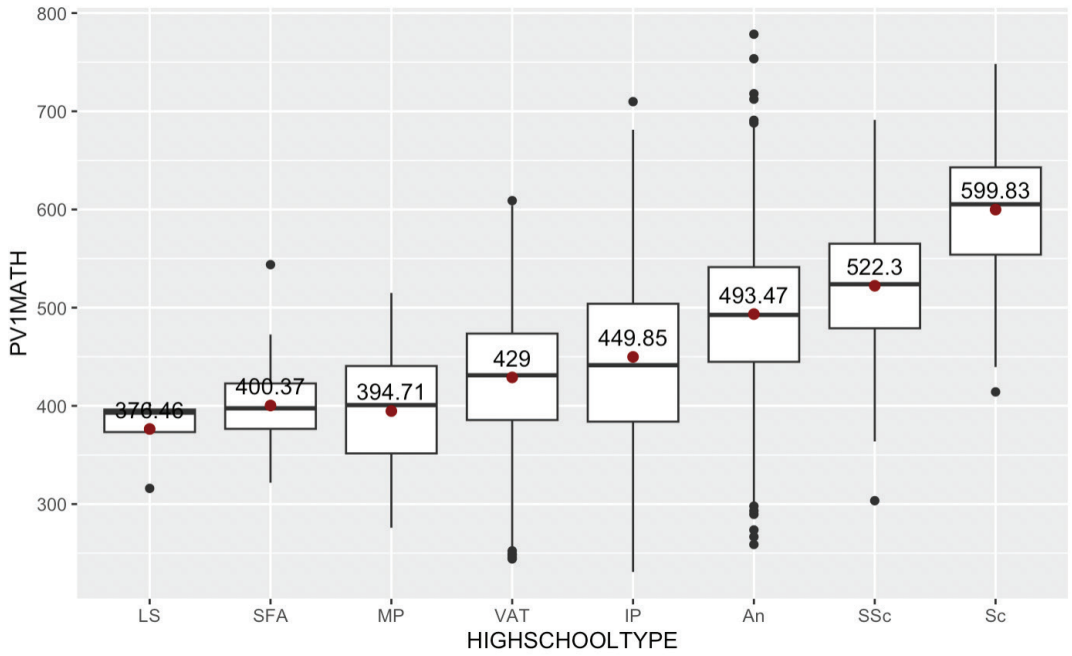

Figure 3. Plausible Values of 2018 PISA Mathematics Subject by School Programme Types Notes: Vertical axis represents mathematics scores. Horizontal axis represents school types; See Table A in Appendix for further information. Red dots and the numbers on the boxes represent average of the first plausible values of mathematics within each highschool type.

A similar pattern in achievement of students by their school programme types can be observed for the science subject. According to Figure 4, students from Science High Schools (Sc), on average, score higher than those from other school programme types, and also score higher than the country average of approximately 485. Social Sciences High Schools (SSc) and Anatolian High Schools (An) follow as the next most successful school types in the PISA 2018 exam, with students from these schools also tending to score above the country average.

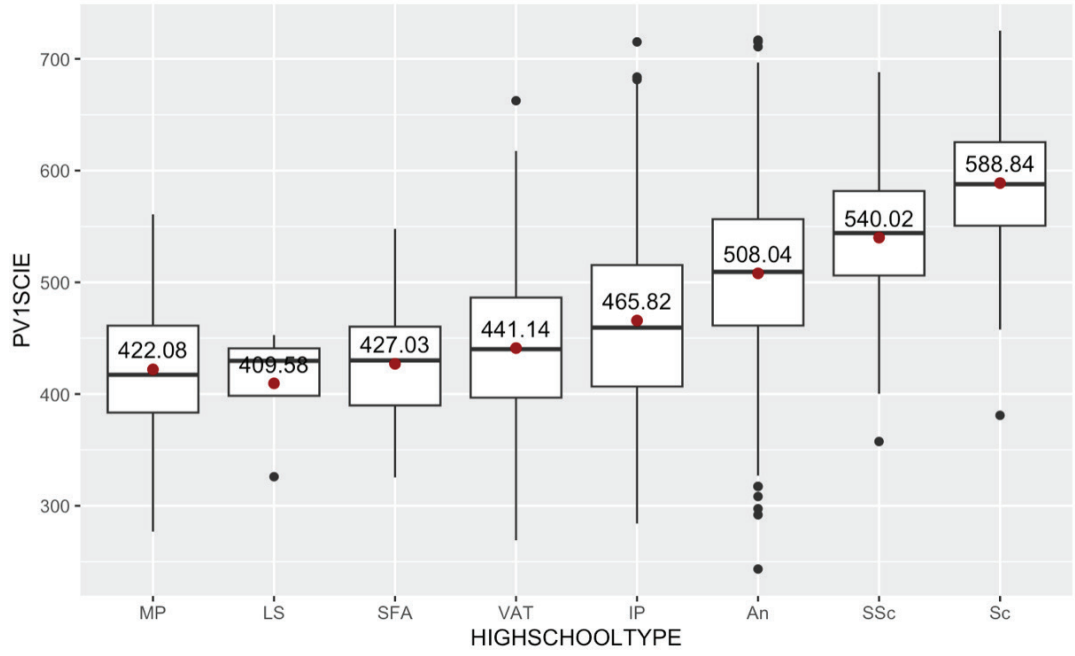

Figure 4. Plausible Values of 2018 PISA Science Subject by School Programme Types Notes: Same notes in Figure 3 applies except that vertical axis represents science scores.

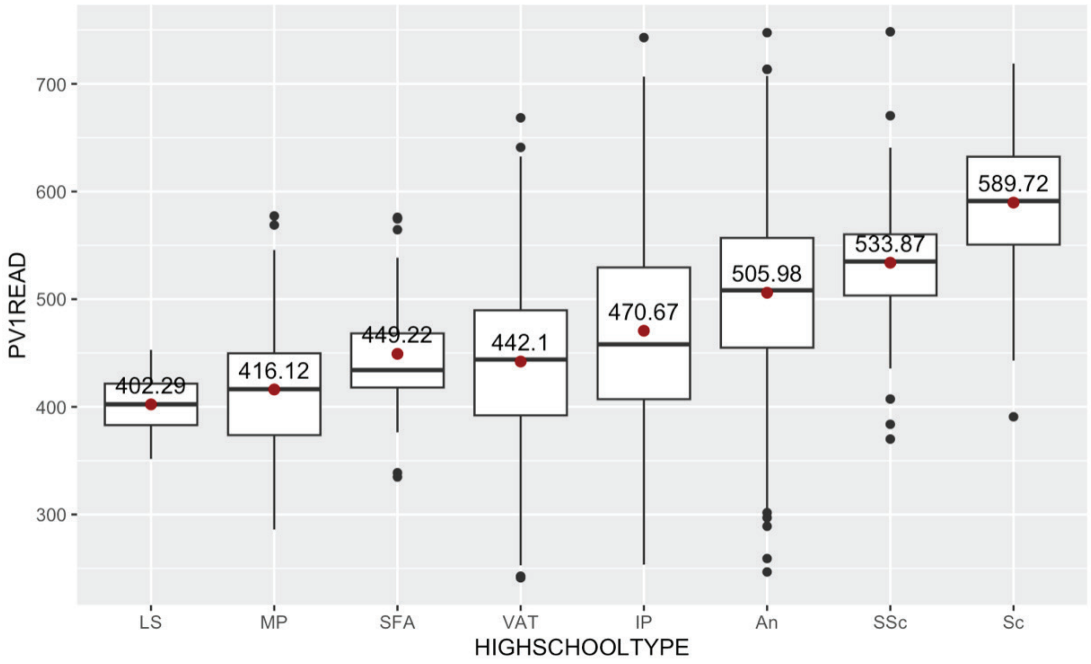

Figure 5. Plausible Values of 2018 PISA Reading Subject by School Programme Types Notes: Same notes in Figure 3 applies except that vertical axis represents reading score as given by reading scores.

The pattern observed in the mathematics and science subject literacy of Turkish students in the PISA 2018 also holds for the reading subject area, as can be seen in Figure 5. Students from Science High Schools (Sc) score, on average, higher than the rest of the school programme types for the reading subject. Additionally, students from Science High Schools (Sc) score higher than the country average, which is approximately 484. Social Sciences High Schools (SSc), Anatolian High Schools (An), and Anatolian Imam and Preacher High Schools (IP) are the next most successful school types in the reading subject of the PISA 2018 exam. Similar to students from Science High Schools (Sc), students enrolled at these schools tend to score higher than the country average.

Table 3 presents summary statistics of the following predictors for the Turkish subsample: economic, social, and cultural status index (ESCS), highest parental occupational status (HISEI), highest parental education in years (PARED), family wealth index (WEALTH), student-teacher ratio (STRATIO), mathematics subject study time in minutes per week (MMINS), science subject study time in minutes per week (SMINS), and reading subject study time in minutes per week (RMINS). On average, parents in the sample had completed 11.07 years of schooling, with a range of 13 years. Turkish students spent more time studying mathematics than science or reading, with a maximum difference of 1040, 1710, and 1600 minutes per week for mathematics, science, and reading, respectively. The standard deviation showed that there was more variability in study time for science subject. The student-teacher ratio in Turkish secondary schools was, on average, 13.73 students per teacher, ranging from 2.34 to 40.76 students per teacher. The median student-teacher ratio was 14.02 students per teacher.

HISEI index is derived from students' responses to questions related to father's and mother's occupations, with higher values indicating higher levels of occupational status (PISA 2018 Technical Report). In the Turkish subsample of PISA 2018, the mean of the HISEI is 39.68, with values

Table 3. Descriptive Statistics of Feature Variables

	ESCS	HISEI	PARED	WEALTH	STRATIO	MMINS	SMINS	RMINS
Minimum	-4,76	11,56	3	-4,94	2,34	40	0	0
Maximum	2,76	88,96	16	3,52	40,76	1080	1710	1600
Range	7,52	77,4	13	8,467	38,42	1040	1710	1600
Standard Deviation	1,18	23,25	4,27	0,91	3,459	81,5	119,6	75,72
Mean	-1,05	39,68	11,07	-1,26	13,73	233,39	208,54	207,51
Median	-1,18	30,34	12	-1,243	14,02	240	240	200

ranging from 11.56 to 88.96. WEALTH is another index derived from students' responses to questions related to home possession variables and home educational resources, including a room of somebody's own, connection to the Internet, possession of works of art, number of books at home, televisions, cars, air conditioner, and affording a holiday. According to Table 3, the average value of WEALTH for Turkish students is -1.26, with a range from -4.94 to 3.52. The index of economic, social and cultural status (ESCS) is calculated by PISA using parents' education, parents' occupation and the index of home possessions which can be considered as proxies of material wealth or cultural capital. ESCS has been built as a weighted average of three standardized components across all countries participating PISA 2018: parental educational attainment, occupational status and possessions at home. The final ESCS variable was transformed, with an average of 0 and standard deviation of 1 across equally weighted OECD countries. Therefore, positive values indicate higher socioeconomic status and negative values indicate lower socioeconomic status with respect to OECD average student's socioeconomic status. Based on Table 3, the average value of ESCS for Turkish students is -1.05, ranging from -4.76 to 2.76.

Table B in Appendix provides information on the distribution of students in the PISA 2018 Turkish subsample based on various characteristics . In Turkey, there are eight different programme types for secondary school students to enroll in. As shown in Table 4, 48% of the sample attend Anatolian High schools, making it the most popular programme type, followed by VAT high schools, and so on. It is noteworthy that this distribution is consistent with the nationwide enrolment figures. Majority of the Turkish sample reside in large cities, followed by cities and towns. Gender distribution is relatively balanced, with no significant differences between male and female students. Most of the schools in the sample are public schools, and Turkish is the most commonly spoken language at home.

5. METHODOLOGY

The purpose of this study is to identify relevant features that predict Turkish secondary school students' success, in order to inform education policies proposed by policymakers. To achieve this, we compared the predictive performances of several supervised ML models, specifically Ensemble Decision Trees. To avoid overfitting, we split the sample of 3876 observations into an 80% training and a 20% test set. For all models, we performed 10-fold cross-validation (CV) on the training set to estimate the model and hyperparameters. The predictive performance of each model was evaluated using the mean squared error on the test set.

Tree-based methods, such as decision trees, are a subset of regression-based ML methods. Classification and regression tree (CART) algorithm is a well-known technique for building decision trees. CART aims to model an outcome variable based on a set of decision rules imposed on predictors (Breiman et al., 1984). CART employs binary recursive partitioning to determine these decision rules by repeatedly splitting the data into successively smaller groups using binary splits based on a single predictor (Prasad et al., 2006). The optimal split for each predictor is the one that produces the smallest residual sum of squares at each partition. In a regression problem, CART predicts the outcome based on the mean of the response values for all observed data that fall in that subgroup.

Ensemble decision trees are built upon decision trees by constructing more than one decision tree. Bagging is a procedure to reduce the variance of a single decision tree. At first, several subsets are created from training sample by choosing randomly with replacement. Each subset is then used to grow their own decision tree. Each of those trees has high variance yet low bias and averaging them reduces the variance:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \qquad (1)$$

where B is the number of subsets chosen randomly with replacement out of training set, $\hat{f}^{b}$ are predictions from those subsets.

Bagging improves prediction accuracy at the expense of interpretability since there is no single tree anymore. In this case overall summary of each variable's importance is obtained by averaging total decrease in RSS due to partitions from a variable over all B trees. A large value indicates that it is an important variable.

Random Forest is an extension of bagging which employs all variables in partitioning the data. Instead of using full set of p variables, random forest chooses a random sample of m predictors out of p as partition candidates. This allows random forest method to overcome the possibility that resulting trees look like each other i.e. they are highly correlated, and hence increased variance. For regression problems, a rule of thumb is to set $\approx p/3$. Consequently, bagging is a special case of random forest such that $m = p$.

BRT method differs from bagging and random forest in that trees are grown sequentially using the information in previously grown trees. More specifically, observations that are previously predicted incorrectly, are chosen more often than correctly predicted observations. Therefore, BRT tries to produce new predictors that are better able to predict observations for which the current ensemble's performance is poor. James et.al (2013) provided a brief description of the process of BRT for supervised regression problems.[2] BRT has two important hyperparameters that need to be

[2]For a detailed procedure for boosted BRT, refer to p.322 of James et al., (2013).

tuned by the researcher. These are learning rate which is the contribution of each tree to the construction of the model and the number of splits (*aka* tree complexity). As learning rate gets smaller many trees are to be built. These two parameters together determine the required number of trees for optimal prediction. The ultimate goal is to find combination of hyperparameters that yields minimum prediction error.

6. RESULTS

6.1. Variable Importance Indicated by Models

Figures 6 and 7 depict top 20 most important variables that come forward in the prediction of student success in mathematics, science and reading subjects of PISA 2018 according to bagging and random forest, respectively. Bagging implies that the most important features that have the highest predictive effect on a Turkish student's mathematics performance in PISA are found as socioeconomic status, number of girls in school, school's capacity using digital devices, number of boys in school, student's science learning time, school size, highschool type, total number of teachers at school, school location, student behavior hindering learning, parental occupation, school's capacity using digital devices, school location, class size at school, proportion of teachers with ISCED level 6, shortage of educational staff, school funding, subject related ICT use during lessons. For the prediction of science performance of a Turkish student; as different from mathematics subject; home possessions, number of available computers per student, extra-curricular creative activities at school, competition from other schools, school programme type and school's policy for grouping students into classes based on ability matter the most. Finally; in order to predict reading performance of a Turkish student, school's policy for grouping students into classes based on ability, use of ICT at home/school, student-teacher ratio, school's capacity using digital devices come forward as different from mathematics and science subjects. To sum, according to bagging, for each subject, a combination of student- and school-level characteristics emerge as major features for prediction.

It is apparent from Figure 7, according to random forest most influential features that have the highest predictive effect on a Turkish student's mathematics performance in PISA are found as socioeconomic status, student's science learning time, school programme type, highschool type, number of girls and boys in school, parental occupation, student behavior hindering learning, usage of ICT at home/school, home possessions, disciplinary climate in lessons, familial wealth, school size, student's occupational aspiration, home cultural possessions. In addition to these, number of available computers per student is found as significant in predicting students' science achievement in Turkey. For predicting reading success of students in Turkey in addition to aforementioned variables, school ICT availability and number of available computers per student are pointed out.

Table 4 summarizes results from 10 fold CV for BRT method on the training set for each subject of PISA. Accordingly, a grid of hyperparameters[3] are constructed

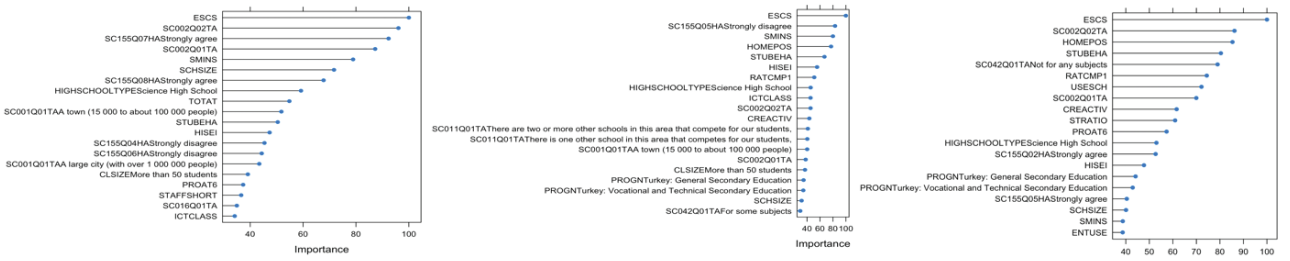

Figure 6. Bagging Variable Importance Plots Note: Mathematics, Science and Reading subjects from left to right.

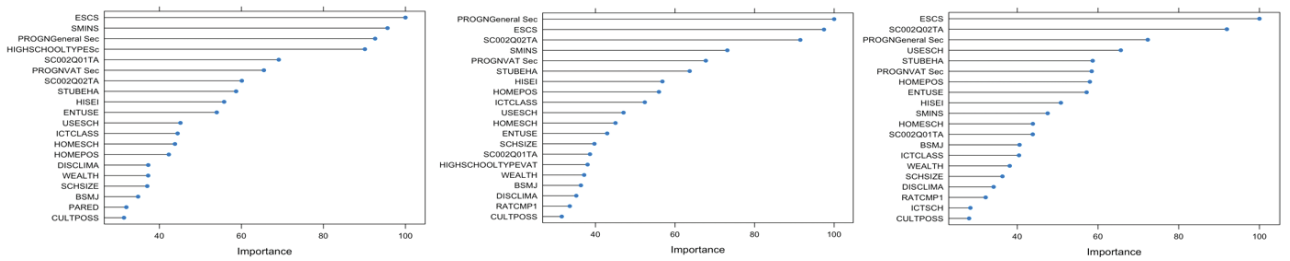

Figure 7. Random Forest Variable Importance Plots Note: Same note in Figure 6 applies.

[3]The set of hyperparameters for learning rate: {0.05, 0.1, 0.2, 0.5} for depth of the tree: {1, 2, 3, 4, 5} for number of trees: {1000, 2000, 5000, 10000}.

Table 4. 10 fold CV Results of BRT

	Mathematics	Science	Reading
Learning Rate	0.1	0.05	0.05
Depth	1	2	2
Number of Trees	2000	1000	1000

and BRT is grown for each combination of hyperparameters and mean squared errors are calculated. The combination of hyperparameters which produced the smallest MSE is chosen.

Figure 8 plots relative influences from top twenty important variables as implied by BRT model which produces the similar results for students' literacy in each subject as in bagging and random forest. Accordingly, highschool type, socioeconomic status, student's science learning time, school programme type, number of boys' enrolment to the school student behavior hindering learning, number of girls in school, student's status of grade repetition in the past, usage of ICT at home/school, percentage of school dropout, parental education, proportion of teachers with ISCED level 6, proportion of teachers with degree ISCED level 5-Master, student's gender, number of students in the school, existence of creative extra-curricular activities, proportion of fully certified teachers, parental occupation status emerge as influential in prediction of mathematical literacy of a Turkish student. For the prediction of scientific literacy of a Turkish student, in addition to previously emphasized characteristics which are mentioned for mathematics except for gender, proportion of certified teachers, girls' enrolment and school size; student's expected occupational status, home possessions and student-teacher ratio matter the most. In order to predict reading literacy of a Turkish pupil; similar to science literacy, student's home possessions and expected occupational status and student-teacher ratio in addition to mathematical literacy characteristics except number of boys in school, school programme type and ICT availability at school come forward.

In order to investigate how values of BRT model inputs affect the model's predictions partial dependence plots, which represents a prediction for a particular value of input variable while averaging out the impact of other variables in the model, are generated. Panels (a), (b) and (c) in Figure 9 show the relationship between top 12 student- and school-level characteristics and predicted student's achievement at PISA mathematics, science and reading subjects, respectively. School programme type appears to be one of the most influential variables on academic achievement. For all subjects, students who are enrolled at high schools of general secondary education tend to be more successful compared to their peers who are enrolled at other category high schools. High school type is another significant variable in the prediction of student success. Accordingly, most successful students are from Science high schools followed by students from Social Science high schools and students from Anatolian high schools. Moreover, this result holds for all subject areas. Student's science learning time emerges as important in predicting student achievement in all subject areas: Turkish secondary school students' performance in each subject is increasing with weekly studying time. Student's socioeconomic status is another variable stressed by the model as important in the prediction of student success in all subjects. Higher socioeconomic status is associated with higher predicted student scores up to one standard deviation ahead the mean. After that, student's scores in mathematics and science -except reading- are predicted to decline. Boys' enrolment to school is associated with higher predicted mathematics scores yet for science and reading scores girls' enrolment to school is much more important. Student behavior hindering learning appears to be another important predictor of student success. Smaller values of the index reflects no/little existence of student behavior hindering learning whereas higher values are associated with frequent occurrence of the event. The school dropout rate seems to be negatively related to student performance prediction for all subjects. Student's grade repetition status appears to be related to prediction of mathematics performance such that student's grade repetition in the past is associated with lower predicted mathematics score. Finally, ICT availability and usage come

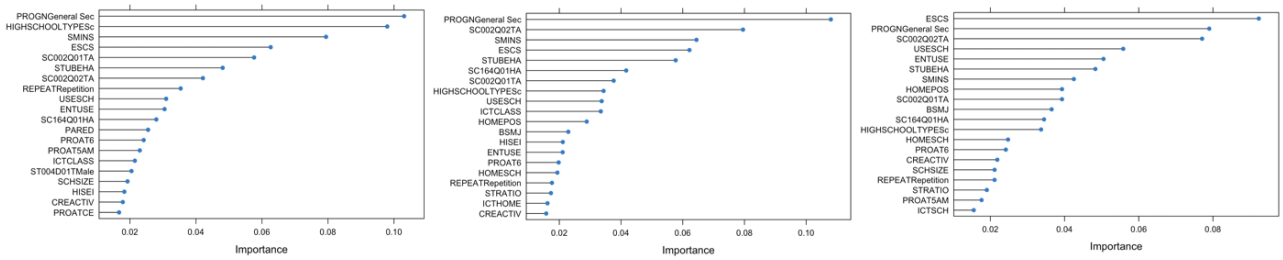

Figure 8. BRT Relative Influence Plots Note: Same note in Figure 6 applies.

forward in prediction of Turkish student scores. ICT-related variables are standardized such that mean score is 0 and standard deviation is 1 (OECD, 2018). Therefore, it can be said that students with negative values are below the average, students with positive values are above the average and students with value of 0 is in the average in relevant category. According to partial dependence plots, the relationship between use of ICT at school and predicted student scores in mathematics and science subjects is U-shaped, whereas it is a negative relationship between the former and reading subject. Students using ICT outside of school for leisure activities above the average are predicted to score higher in mathematics and science yet no such indication can be derived for reading subject. Similarly, students with above the average usage of subject related ICT during lessons are predicted to score higher in science. For reading, students with above the average home endowments are predicted to score higher. For reading and science, BRT estimates a step function for student's expected future occupational status, where higher values are associated with higher occupational status.

6.2. Test Sample Performance Comparison

Table 5 summarizes mean squared errors calculated based on each model in the test sample. It is apparent from the table that, BRT model produces the best predictive performance for Turkish secondary school students' subject literacy in mathematics, science and reading.

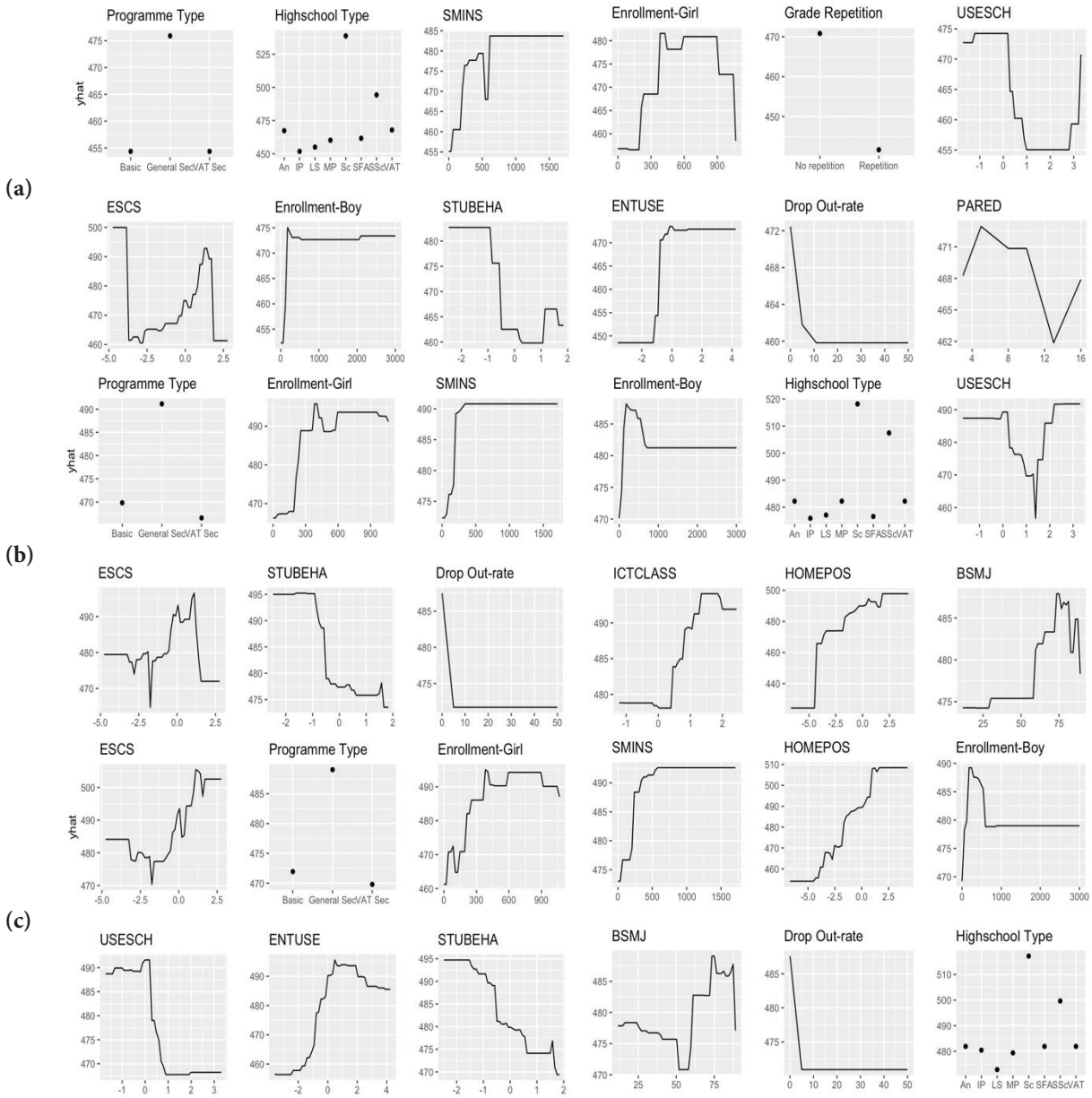

Figure 9. Partial Dependence Plots of Top 12 variables from BRT for Mathematics, Science and Reading Note: yhat stands for prediction of student achievement in PISA 2018. See Table A in Appendix for variable explanations.

Table 5. Prediction Accuracy of Models for Turkish Students' Performance

	Mathematics	Science	Reading
Method	**MSE in Test Sample**		
Bagging	3805.494	3474.429	3731.695
Random Forest	3038.675	2595.649	3040.552
BRT	**2745.782**	**2421.81**	**3033.227**

Note: Model in bold indicates the best predictive performance

6.3. Turkish Students' Academic Achievement by High School Types

Given the insights provided by the BRT model, it becomes evident that for Turkish students high school type stands out as a paramount factor influencing student achievement across all subject areas in the PISA 2018 test. This naturally prompts the question: How do these predictors of student achievement evolve concerning different types of high schools? To explore this phenomenon, we undertake a comparative analysis, using the identical methodology of BRT with a 10-fold CV, focusing on the two most commonly enrolled and the most successful highschools namely VAT, Anatolian and Science.

Table 6 summarizes results from 10 fold CV for BRT method on the training set for each subject and for VAT, Anatolian and Science highschools. Similar to the previous execution, a grid of hyperparameters[4] are constructed and BRT is grown for each combination of hyperparameters and the combination of hyperparameters which produced the smallest MSE is chosen.

Figure 10 plots relative influences from top twenty important variables for the prediction of mathematics achievement of students from VAT, Anatolian and Science highschools as implied by BRT. In order to predict the mathematics achievement of Turkish students who are enrolled at VAT highschools, shortage of educational staff , student-teacher ratio, ICT availability and usage at home/school, school's capacity using digital devices, home possessions, number of boys' in the school, student's science learning time, grade repetition in the past and gender, student behavior hindering learning, class size, proportion of teachers with degree ISCED level 5-Master, parental occupation, percentage of teaching staff attended a professional development programme and proportion of teachers with degree ISCED level 5A-Bachelor emerge as influential.

Similar to VAT highschool students, the predictors of Anatolian highschool students' mathematics performance are mostly related to number of boys' enrolment to the school, student behavior hindering learning, to number of girls' enrolment to the school, student's science learning time, total number of teachers at school, student's gender, percentage of total funding of school, socioeconomic status, percentage of teaching staff attended a professional development programme, ICT usage at home/school, parental occupation and education, shortage of educational staff, proportion of teachers with degree ISCED level 5-Master, student's reading learning time, proportion of certified teachers. For the two most commonly enrolled highschool

Table 6. 10 fold CV Results for BRT Method by Highschool Type

	Mathematics	Science	Reading
VAT Highschool			
Learning Rate	0.05	0.05	0.05
Depth	1	1	1
Number of Trees	1000	1000	1000
Anatolian Highschool			
Learning Rate	0.05	0.05	0.05
Depth	1	1	4
Number of Trees	2000	1000	1000
Science Highschool			
Learning Rate	0.1	0.1	0.05
Depth	4	5	1
Number of Trees	1000	1000	1000

[4]See Footnote 2.

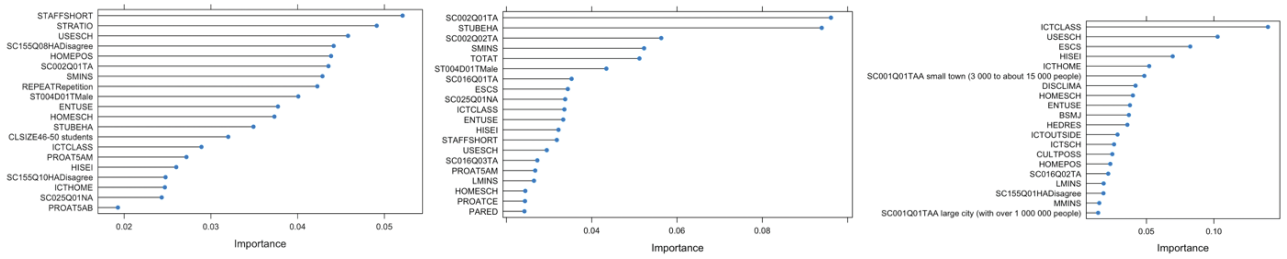

Figure 10. BRT Model Relative Influence Plots for Mathematics Subject by Highschool Type Note: VAT, Anatolian, and Science High schools from left to right.

types, twelve out of twenty most important characteristics are school related. In fact they are in essence indicators of quality of education in those schools such that quality of teachers (i.e. their education, their status of attendance to professional development programs etc.), school's facility to enable better education such that class size, ICT availability and usage etc.

For the Science highschool students' mathematics performance it is observed that eleven out of twenty most important characteristics are student based. These include socioeconomic status, parental education, ICT availability and usage at home, student's expected occupational status, home endowments as educational and cultural resources, student's reading and mathematics learning time. In terms of school-based characteristics ICT availability/usage and disciplinary climate, location of school and percentage of total funding of school appear as important in predicting science highschool students' mathematics performance.

Figure 11 replicates the procedure as in Figure 8 for science subject. In order to predict the science achievement of Turkish students who are enrolled at VAT highschools, percentage of student drop out from school, school's capacity using digital devices, ICT availability and usage at home/school, shortage of educational staff, student's science learning time, student-teacher ratio, parental education, student's occupational aspiration, home educational resources, student's grade repetition in the past, school size, number of boys in the school, socioeconomic status, proportion of teachers with degree ISCED level 5-Master and total number of teachers at school appear as important.

Like VAT highschool students, the predictors of Anatolian highschool students' science performance mostly include aforementioned characteristics as well as student behavior hindering learning, parental occupational status, number of available computers per student, proportion of certified teachers, home possessions and number of books at home. The prediction of science highschool students' science performance involves similar characteristics as their peers in other types of highschools and student's mathematics learning time, student's familial wealth and grade level and home cultural possessions.

Figure 12 replicates the same procedure as in Figures 8 and 9 for reading subject. In order to predict reading achievement of Turkish students who are enrolled at VAT highschools, availability and usage of ICT at home/school, number of boys' in the school, home possessions, school size, school's capacity using digital devices, shortage of educational staff, the percentage of student dropout from school, student-teacher ratio, student's occupational aspiration and science learning time, parental occupation, and percentage of total funding of school come forward.

Like VAT highschool students, the predictors of Anatolian highschool students' reading performance mostly include aforementioned characteristics together with home educational resources, teacher behavior hindering learning, familial wealth, number of books at home, disciplinary climate in lessons, socioeconomic status, ICT usage and percentage of total funding of school. The prediction of science highschool students' reading performance involves similar characteristics as their peers in other types of highschools and percentage of teaching staff attended professional development and percentage of total funding of school.

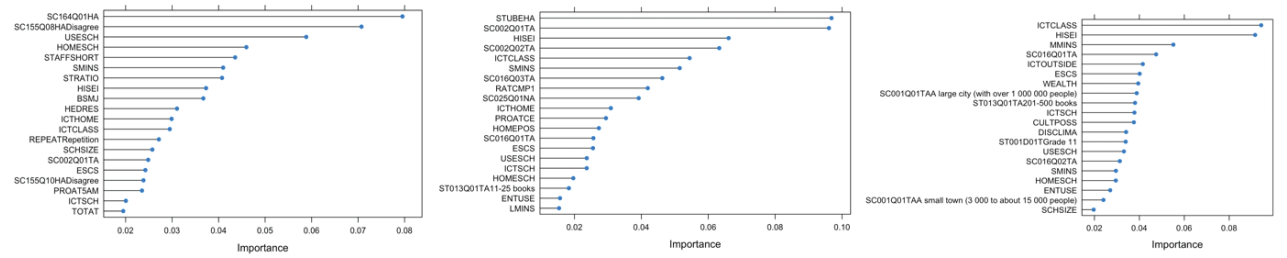

Figure 11. BRT Model Relative Influence Plots for Science Subject by Highschool Type Note: Same note in Figure 10 applies.

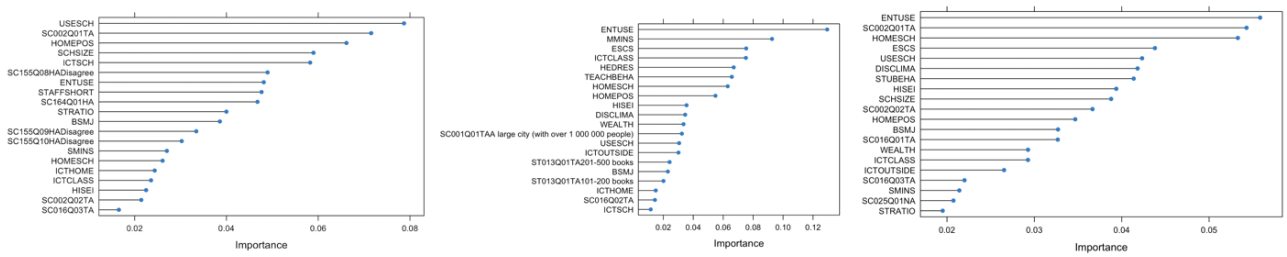

Figure 12. BRT Model Relative Influence Plots for Reading Subject by Highschool Type Note: Same note in Figure 10 applies.

7. RESULTS AND DISCUSSION

Our findings resonate with a growing body of literature on Educational Data Mining, especially in the context of a developing country like Turkey with a young demographic and recent educational reforms. Notably, the significance of a student's socioeconomic status, as reflected by the ESCS index, in predicting success echoes results from various international student assessment tests (e.g., Gabriel et.al., 2018; Masci 2018; Kılıç Depren, 2018; Puah 2021;). This emphasis on factors like socioeconomic status, home possessions, and the influence of ICT availability both in and outside school settings has been reiterated across multiple studies (e.g., Yu et.al., 2012; Gorostiaga & Rojo-Alvarez, 2016; Kılıç Depren, 2018; Yoo, 2018; Dong & Hu, 2019; Filiz & Öz, 2019; Hu et.al., 2022). Moreover, student's weekly learning time appears to be influential in other studies as well (She et.al., 2019; Lee & Lee, 2021).

While our research aligns with past findings, it offers unique insights into the academic landscape of Turkish secondary schools. We comprehensively examined the performance predictors across three subject area literacies. Our data reveals that students from high schools of general secondary education, particularly Science high schools, consistently outperform their peers. This consistency is expected given these schools' selective admission criteria based on nationwide exams.

Furthermore, our study is pioneering in disentangling how performance predictors vary across different high school types. We delved into the two most common high school types in Turkey: VAT and Anatolian, as well as Science high schools, using the BRT methodology. Across all these school types, influential factors like student learning time per week and parental occupational status emerged as prominent. More academically successful schools like Anatolian and Science high schools also underscored the importance of a student's socioeconomic situation. Parents with stable occupations and income often prioritize their children's education, investing in educational and cultural resources to further their academic development. This underscores the potential for policy interventions targeting parental awareness and engagement. In VAT high schools, enhancing school quality can further boost student success. Our findings suggest that by addressing challenges like educational staff shortage, student-teacher ratio imbalances, and optimizing ICT usage, significant improvements can be achieved. The role of funding in Anatolian and Science high schools further underscores the importance of resource allocation in shaping academic outcomes.

8. CONCLUSION

Using ML methods, this study examined the impact of student, family, and school level attributes on the mathematics, science, and reading literacy of Turkish secondary school students on a well-known international large-scale assessment test PISA (2018). The focus of this work is threefold: the first objective is to determine the best ML method which has the highest predictive accuracy in the context. The second objective is to provide evidence for the most influential variables that help predicting student success within a learning system which is highly competitive and early tracking of students is present. Finally this research aims to compare the most successful high school type in Turkey with the mostly enrolled high school types in terms of the factors that affect students' performance.

Results suggest that, for each subject area in PISA 2018, Boosted Regression Tree is chosen to be the method with the best predictive accuracy of Turkish secondary schoolers' academic performance. In relation to the second objective of the study, it is concluded that the type of high school, socioeconomic status, students' weekly learning time, school program type, number of boys in the school, student behavior hindering learning, student's grade repetition in the past, use of ICT at school in general, percentage of recent school dropout, students' use of ICT for leisure, parental education and occupational status, students' gender, school size, teachers' educational degree, subject-related ICT usage in class during lessons, existence of creative extracurricular activities, home possessions, students' expected occupational status, and student-teacher ratio are among the most important predictors of secondary schoolers' academic performance in Turkey, with school program type being the foremost.

Regarding the final objective of the study, the findings suggest that while school quality indicators are crucial for student success in Vocational and Technical and Anatolian high schools, individual student characteristics

take precedence in Science high schools. This nuance underscores the importance of tailored educational interventions instead of blanket policies. Moreover, empirical evidence in this study provides educational policy makers with indications related to which of the student and school level characteristics work to improve education performance. In doing so, Turkish young can be better targeted with educational opportunities based on their endowments at personal/ familial and school level. For all subjects, student's enrolment to schools of general secondary education instead of VAT education is by far the most important in higher predicted scores of students. This result is also evident in higher predicted scores of students from Science, Social Science and Anatolian High schools. Therefore, there is still room for improvement in the nature of educational activities implemented in vocational and technical education schools.

Although this study demonstrated the best ML method to predict Turkish secondary school students' academic achievement at a well-known international student assessment test and the most successful predictors of the outcome, it has certain limitations in terms of generalizability of the results to other exams, countries and years. Besides, it is essential to recognize that these findings do not establish causal relationships between academic achievement of Turkish secondary schoolers and examined factors rather they present an outline of individual, family and school based attributes in which the former is common. Future studies would benefit from an in-depth examination of student and school-related factors using administrative data, shedding light on the nuances within the layered Turkish education system, further advancing the insights and policy recommendations drawn from this study.

Contribution Rate of Authors: The authors contributed to the study equally.

Conflict of interest disclosure: The authors report there are no competing interests to declare.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement: The data that support the findings of this study are openly available in PISA 2018 Database at https://www.oecd.org/pisa/data/2018database/

REFERENCES

Aksu, G., & Güzeller, C. O. (2016). Classification of PISA 2012 mathematical literacy scores using decision-tree method: Turkey sampling. *TED Eğitim ve Bilim 41*(185). [CrossRef]

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Breiman, L. (2017). *Classification and regression trees*. Routledge. [CrossRef]

Chen, J., Zhang, Y., Wei, Y., & Hu, J. (2021). Discrimination of the contextual features of top performers in scientific literacy using a machine learning approach. *Research in Science Education*, *51*(1), 129–158. [CrossRef]

Dong, X., & Hu, J. (2019). An exploration of impact factors influencing students' reading literacy in Singapore with machine learning approaches. *International Journal of English Linguistics*, *9*(5), 52–65. [CrossRef]

Filiz, E., & Öz, E. (2019). Finding the Best Algorithms and Effective Factors in Classification of Turkish Science Student Success. *Journal of Baltic Science Education*, 18(2), 239–253. [CrossRef]

Gabriel, F., Signolet, J., &Westwell, M. (2018). A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. *International Journal of Research & Method in Education*, 41(3), 306–327. [CrossRef]

Gorostiaga, A., & Rojo-Álvarez, J. L. (2016). On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain. *Neurocomputing*, *171*, 625–637. [CrossRef]

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, 351–388. [CrossRef]

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American Economic Review*, *90*(5), 1184–1208. [CrossRef]

Hu, J., Peng, Y., & Ma, H. (2022). Examining the contextual factors of science effectiveness: a machine learning-based approach. *School Effectiveness and School Improvement*, *33*(1), 21–50. [CrossRef]

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. [CrossRef]

Kasap, Y., Doğan, N., & Koçak, C. (2021). PISA 2018'de Okuduğunu anlama başarısını yordayan değişkenlerin veri madenciliği ile belirlenmesi. *Manisa Celal Bayar Üniversitesi Sosyal Bilimler Dergisi*, *19*(4), 241–258. [CrossRef]

Kılıç Depren, S. (2018). Prediction of students' science achievement: an application of multivariate adaptive regression splines and regression trees. *Journal of Baltic Science Education*, *17*(5), 887–903. [CrossRef]

Kıray, S. A., Gök, B., &Bozkır, A. S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *Journal of Education in Science Environment and Health*, *1*(1), 28–48. [CrossRef]

Kleinberg, J., Ludwig, J., Mullainathan, J., and Obermeyer, Z. (2015). "Prediction Policy Problems", *American Economic Review, Papers and Proceedings*, *105*(5), 491–495. [CrossRef]

Lee, J. W., & Barro, R. J. (2001). Schooling quality in a cross–section of countries. *Economica*, *68*(272), 465–488. [CrossRef]

Lee, H., & Lee, J. W. (2021). *Why East Asian students perform better in mathematics than their peers: An investigation using a machine learning approach.* CAMA Working Paper No. 66/2021. [CrossRef]

Martínez-Abad, F., Gamazo, A., & Rodríguez-Conde, M. J. (2020). Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment. *Studies in Educational Evaluation, 66,* Article 100875. [CrossRef]

Masci, C., Johnes, G., &Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, *269*(3), 1072–1085. [CrossRef]

MEB (2019). *PISA 2018 ulusalönraporu*. Ankara: http://pisa.meb.gov.tr/eski%20dosyalar/wpcontent/uploads/2020/01/PISA_2018_Turkiye_On_Raporu.pdf

OECD. (2009). PISA Data Analysis Manual. https://www.oecd-ilibrary.org/docserver/9789264056275-en.pdf?expires=1680205505&id=id&accname=guest&checksum=-11DAE831D022F23D8FF8E094F9E7AB8C

OECD (2019), PISA 2018, https://www.oecd.org/pisa/data/2018database/ accessed on 25 October 2021.

OECD. (2019). *PISA 2018 Technical Report.* https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2019). Turkey - Country Note - PISA 2018 Results. https://www.oecd.org/pisa/publications/PISA2018_CN_TUR.pdf

Prasad, A. M., Iverson, L. R., &Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, *9*, 181–199. [CrossRef]

Puah, S. (2021). Predicting Students' Academic Performance: A Comparison between Traditional MLR and Machine Learning Methods with PISA 2015. Preprint. doi: 10.31234/osf.io/2yshm [CrossRef]

Rebai, S., Yahia, F. B., &Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Economic Planning Sciences*, *70*, Article 100724. [CrossRef]

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. [CrossRef]

She, H. C., Lin, H. S., & Huang, L. Y. (2019). Reflections on and implications of the Programme for International Student Assessment 2015 (PISA 2015) performance of students in Taiwan: The role of epistemic beliefs about science in scientific literacy. *Journal of Research in Science Teaching*, *56*(10), 1309–1340. [CrossRef]

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453. [CrossRef]

Uğuz, E., Şahin, S., & Yılmaz, R. (2021). PİSA 2018 fen bilimleri puanlarının değerlendirilmesinde eğitsel veri madenciliğinin kullanımı. *Bilgi ve İletişim Teknolojileri Dergisi*, *3*(2), 212–227. [CrossRef]

Walberg, H. J. (1981). A psychological theory of educational productivity. In F. H. Farley & N. Gordon (Eds.), Psychology and education (pp. 81–110). Berkeley, CA: McCutchan.

Woessmann, L. 2008. "How equal are educational opportunities? Family background and student achievement in Europe and the United States." *ZeitschriftfürBetriebswirtschaft, 78*(1), 45–70.

Yoo, J. E. (2018). TIMSS 2011 student and teacher predictors for mathematics achievement explored and identified via elastic net. *Frontiers in Psychology*, *9*, Article 317. [CrossRef]

Yu, C. H., Kaprolet, C., Jannasch-Pennell, A., & DiGangi, S. (2012). A data mining approach to comparing American and Canadian grade 10 students' PISA science test performance. *Journal of Data Science, 10*(24), 441–464. [CrossRef]

Appendix

Table A. Variable list considered in the study

Variable Code	Variable Description	Variable Type
Student Level		
ST001D01T	Student International Grade	
ST004D01T	Student Gender	Nominal
ST013Q01TA	How many books are there in your home?	Ordinal
PROGN	Unique national study programme code	
HISEI	Index highest parental occupational status	Numerical, Index
LANGN	Language at home	3 digit code
PARED	Index highest parental education in years of schooling	Numerical (Min:3 Max:18)
IMMIG	Index Immigration status	Ordinal, 3 point Likert Scale: Native, Second Generation, First Generation
REPEAT	Grade Repetition	Nominal: Yes, No
BSMJ	Student's expected occupational status	Numerical, Index
MMINS	Mathematics learning time (minutes per week)	Numerical
SMINS	Science learning time (minutes per week)	Numerical
LMINS	Reading learning time (minutes per week)	Numerical
ESCS	Index of economic, social and cultural status	Numerical, Index
ICTHOME	ICT available at home	Numerical, Index
ICTSCH	ICT available at school	Numerical, Index
HOMEPOS	Home Possessions	Numerical, Index
CULTPOSS	Cultural Possessions	Numerical, Index
HEDRES	Home Educational Resources	Numerical, Index
WEALTH	Family Wealth	Numerical, Index
DISCLIMA	Disciplinary Climate in lessons	Numerical, Index
HOMESCH	Use of ICT outside of school (for school work activities)	Numerical, Index
ENTUSE	ICT use outside of school (leisure)	Numerical, Index
USESCH	Use of ICT at school in general	Numerical, Index
ICTCLASS	Subject-related ICT use during lessons	Numerical, Index
ICTOUTSIDE	Subject-related ICT use outside of lessons	Numerical, Index
School Level		
SC001Q01TA	Which of the following definitions best describes the community in which your school is located?	Ordinal
PRIVATESCH	Private or Public	Nominal
SC016Q01TA	Percentage of total funding from: Government	Percentage
SC016Q02TA	Percentage of total funding from: Student fees or school charges paid by parents	Percentage
SC016Q03TA	Percentage of total funding for school year from: Benefactors, donations etc.	Percentage
SC016Q04TA	Percentage of total funding for school year from: Other	Percentage

Table A. Variable list considered in the study (*continued*)

Variable Code	Variable Description	Variable Type
SC155Q01HA	School's capacity using digital devices: The number of digital devices connected to the Internet is sufficient	Ordinal, 4 point Likert Scale: Strongly Agree, Agree, Disagree, Strongly Disagree
SC155Q02HA	School's capacity using digital devices: The school's Internet bandwidth or speed is sufficient	
SC155Q03HA	School's capacity using digital devices: The number of digital devices for instruction is sufficient	
SC155Q04HA	School's capacity using digital devices: Digital devices are sufficiently powerful in terms of computing capacity	
SC155Q05HA	School's capacity using digital devices: The availability of adequate software is sufficient	
SC155Q06HA	School's capacity using digital devices: Teachers have the skills to integrate digital devices in instruction	
SC155Q07HA	School's capacity using digital devices: Teachers have sufficient time to prepare lessons integrating digital devices	
SC155Q08HA	School's capacity using digital devices: Effective professional resources for teachers to learn how to use digital device	
SC155Q09HA	School's capacity using digital devices: An effective online learning support platform is available	
SC155Q10HA	School's capacity using digital devices: Teachers are provided with incentives to integrate digital devices	
SC155Q11HA	School's capacity using digital devices: The school has sufficient qualified technical assistant staff	
SC011Q01TA	Which of the following statements best describes the schooling available to students in your location?	Ordinal, 3 point Likert Scale: More than one, One, None
SC012Q01TA	Student admission to school: Student's record of academic performance (including placement tests)	
SC042Q01TA	School's policy for 15yearolds: Students are grouped by ability into different classes.	
SC042Q02TA	School's policy for 15yearolds: Students are grouped by ability within their classes.	
STUBEHA	Student behavior hindering learning	Numerical, Index
TEACHBEHA	Teacher behavior hindering learning	Numerical, Index
EDUSHORT	Shortage of educational material	Numerical, Index
STAFFSHORT	Shortage of educational staff	Numerical, Index
CREACTIV	Creative extra-curricular activities	Numerical, Index
PROATCE	Index proportion of all teachers fully certified	Numerical, Index
PROAT6	Index proportion of all teachers ISCED LEVEL 6	Numerical, Index
PROAT5AB	Index proportion of all teachers ISCED LEVEL 5A Bachelor	Numerical, Index
PROAT5AM	Index proportion of all teachers ISCED LEVEL 5A Master	Numerical, Index
SCHSIZE	School Size (Sum of boys and girls)	Numerical
SC002Q01TA	What was the total school enrolment? Number of boys	Numerical
SC002Q02TA	What was the total school enrolment? Number of girls	Numerical
TOTAT	Total number of all teachers at school	Numerical

Table A. Variable list considered in the study (*continued*)

Variable Code	Variable Description	Variable Type
SC025Q01NA	During the last three months, what percentage of teaching staff attended a programme of professional development?	Percentage
CLSIZE	Class size	Ordinal
SC164Q01HA	In the last full academic year, what proportion of students in final grade left school without a certificate?	Percentage
SC152Q01HA	Does your school offer additional test language lessons during the usual school hours?	Nominal: Yes, No
SC052Q01NA	For 15year old students, school provides study help: Room(s) where the students can do their homework	Nominal: Yes, No
SC052Q02NA	For 15year old students, school provides study help: Staff help with homework	Nominal: Yes, No
SC052Q03HA	For 15year old students, school provides study help: Peer-to-peer tutoring	Nominal: Yes, No
SCHLTYPE	School Ownership: Public, Private, Private- Government dependent	Nominal
STRATIO	Student-Teacher Ratio	Percentage
RATCMP1	Number of available computers per student at modal grade	Numerical
HIGHSCHOOLTYPE	Type of Highschool: Anatolian (An), Imam-Preacher (IP), Vocational-Technical (VAT), Science(Sc), Social Science (SSc), Sports-Fine Arts (SFA), Multi-Programme (MP), Lower Secondary (LS)	Categorical

Table B. Total Number and Percentage of Students in PISA 2018 Turkish Subsample based on Different Characteristics

	Total Number	**Percentage**
Highschool Type	**3876**	**100**
Secondary School	4	0.1
Anatolian Preacher	464	12
Multi Programme Anatolian	126	3.3
Vocational and Technical	1145	29.5
Anatolian Fine Arts	21	0.5
Social Sciences	101	2.6
Science	162	4.2
Anatolian	1853	47.8
Location of Residency	**3876**	**100**
A Village	73	1.8
A Small Town	101	2.6
A Town	956	24.7
A City	1143	29.5
A Large City	1603	41.4
Gender	**3876**	**100**
Male	1868	48.2
Female	2008	51.8
School Type	**3876**	**100**
Public School	3487	90
Private Independent School	374	9.5
Private Government Dependent	15	0.5
Language at Home	**3876**	**100**
Turkish	3672	95
Another Language	204	5
Enrolments by Grade	**3876**	**100**
Grade 8	4	0.1032
Grade 9	620	15.9
Grade 10	3132	80.80
Grade 11	118	3.04
Grade 12	2	0.0516
Grade Repetition Status	**3876**	**100**
Repeated a grade	207	5
Did not repeat a grade	3669	95