Random Forest Analysis of Factors Predicting Science Achievement Groups: Focusing on Science Activities and Learning in School

Jeehye Hong | ORCID: 0000-0003-4450-8116  
Department of Chemistry Education, Seoul National University,  
Seoul, 08826, Republic of Korea  
ilovejh12@snu.ac.kr

Hyunjung Kim | ORCID: 0000-0001-6223-9816  
Department of Chemistry Education, Kongju National University,  
Chungnam, 32588, Republic of Korea  
chem95@kongju.ac.kr

Hun-Gi Hong | ORCID: 0000-0001-5578-3565  
Corresponding author  
Department of Chemistry Education, Seoul National University,  
Seoul, 08826, Republic of Korea  
hghong@snu.ac.kr

Received: 26 September 2022 | Revised: 15 November 2022 |  
Accepted: 17 November 2022

Abstract

This study explored science-related variables that have an impact on the prediction of science achievement groups by applying the educational data mining (EDM) method of the random forest analysis to extract factors associated with students categorized in three different achievement groups (high, moderate, and low) in the Korean data from the 2015 Programme for International Student Assessment (PISA). The 57 variables of science activities and learning in school collected from PISA questionnaires for students and parents were analyzed. Variables related to students’ past science activities, science teaching and learning methods, and environmental awareness were found to played important roles in predicting science achievement. When checking partial
dependence plots for major variables, science activities and instructional strategies had a high probability of changing the prediction of an achievement group. This study focused on science-related contextual variables that can be improved through government policies and science teachers’ efforts in the classroom.

Keywords

science achievement – science activities – learning science in school – random forest – PISA 2015

1 Introduction

In 2016, the World Economic Forum presented scientific literacy as one of the basic literacies in the era of the Fourth Industrial Revolution. Science is a core subject for strengthening national competitiveness, and science literacy and competencies are being emphasized for students who will live in a rapidly changing future society. Moreover, science literacy is required not only of experts but of all citizens. Accordingly, countries around the world are making various policy efforts to foster creative and integrative thinking people with science literacy. In Korea, the proportion of students that belong to the top rank has been low in the Programme for International Student Assessment (PISA) science domain, and the number of students going on into science and engineering fields has been decreasing (Korea Institute for Curriculum and Evaluation [KICE], 2016). Improvements are needed to foster future scientific and technological talents who will lead the national industrial development in line with the rapidly changing society with advanced science and technology.

In addition, the proportion of scoring students below the science literacy baseline has more than doubled since the previous PISA assessment cycle, a level which has been maintained since the PISA 2015 assessments (Kim, 2021). If adequate educational support is not provided to students, learning deficits may accumulate over time. This may lead to students experiencing greater difficulties in fostering the skills necessary for living in the society of the future. The field of education in Korea, which has focused on the raising achievement among average students, is currently trying to increase the low percentage of students with high proficiency in science and is trying to decrease the percentage of students scoring below the baseline in science. To address these concerns, the Korean government has been preparing educational policy support.
(Ministry of Science and ICT – Korea, 2021) to identify problems and help foster future leaders with science literacy. This study seeks to provide information that can help policy-makers prepare an effective support plan by analyzing factors affecting the science achievement group prediction of Korean students.

Therefore, in this study, PISA data were analyzed using the educational data mining (EDM) method of the random forest to extract factors associated with students categorized as being in high and low science achievement groups. In particular, by comparing and analyzing the high and low science achievement groups with the moderate achievement group, we tried to identify different factors according to the characteristics of each science achievement group. This study focused on science-related contextual variables that can be improved through government policies and teachers’ efforts. In addition, item-level responses were used as predictor variables to provide a specific rationale for effective learning support. This study aimed to identify:

1. What factors affect the predicted probability of a high-achievement group of Korean students,
2. What factors affect the predicted probability of a low-achievement group of Korean students, and
3. The common and differentiating factors that affect the predicted probability of the high- and low-achievement groups of Korean students.

2 Theoretical Background

2.1 PISA Educational Contextual Variables

PISA assesses the extent to which 15-year-old students, near the end of their compulsory education, have acquired scientific literacy and competencies that are essential for participation in modern societies. Moreover, the conceptual model of educational contextual variables in PISA is structured to understand the educational context of participating countries and its relationship with students’ achievement. A survey is conducted concurrently with students, parents, teachers, and so on (OECD, 2017). The questionnaire comprises general and major domain-related questions about student background, teaching and learning, and non-cognitive achievement constructs. In PISA 2015, the most recent cycle in which science was the major domain, a questionnaire was conducted on general contextual and science-related contextual variables. PISA published the results on its website, and researchers have analyzed these results in various ways to draw meaningful implications.
2.1.1 Science Achievement

The OECD calculated the amount of explanation for science achievement in the PISA contextual variable in the results report to examine the factors that affect students’ science achievement (OECD, 2017). Individual researchers created a model based on their theoretical background and verified it with PISA data to explain and interpret the relationship between the variables and scientific achievement (Forbes et al., 2020; Oliver et al., 2019). In addition, based on the recent development of EDM, studies have been done that aimed to derive variables affecting scientific achievement by applying it to PISA data with many variables and complex interactions between variables (Hong et al., 2022; Hu et al., 2021; Hu & Hu, 2021). These previous studies have shown that students’ scientific psychological variables, environmental awareness, and the PISA index of economic, social, and cultural status (ESCS) influence scientific achievement.

After classifying groups according to students’ science achievement level or ESCS, factors that could belong to a higher achievement level groups were also explored (OECD, 2011). Since the factors affecting a student’s proficiency in reading, math, and science differed according to their achievement level, it was suggested that there is a need for differentiated support based on achievement level (Song et al., 2015). In another study, Ku and Koo (2018) analyzed the effects of educational contextual variables that affect the science achievement of Korean students according to achievement level group. They found that the effects of the environment and strategy for teaching and learning on students’ affective and cognitive characteristics differed according to achievement levels. These results suggested the need to develop customized teaching strategies according to students’ scientific achievement levels. However, there is a limit to examining the effects of specific teaching and learning strategies through research because standardized indices calculated from the results of only three to nine individual questions were used.

2.1.2 Student Participation in Science Activities

In addition to achievement, PISA also investigates how often students engage in science-related activities. Researchers found that students with frequent past science activities (from around the age of 10) showed high intrinsic motivation for learning science and for high achievement in science. The results of a previous study (Tang & Zhang, 2020) suggested that frequency of participation in science activities in the past positively affected current participation in science activities and attitudes toward science. Students who participated more
often in science-related activities had better attitudes and higher achievement than their counterparts who participated less (Lau & Ho, 2022). The overall frequency of Korean students' science activities was relatively lower than in other OECD countries, and within this group, there was a large difference in frequency of participation in science activities between students categorized as high- and low-achieving in science (KICE, 2016).

### 2.1.3 Science Teaching and Learning Methods
How science is taught in schools has been considered a major variable in student achievement; PISA classified four types of science teaching methods: inquiry-based science instruction, teacher-directed science instruction, perceived feedback from science teachers, and adaptive instruction in science lessons (OECD, 2016). Since the effect of teaching method in science class should be considered comprehensively with the classroom environment, issues such as, teacher's support of student choice in class and the disciplinary climate in class, were also considered science-related contextual variables. By analyzing publicly available data from countries where science was a major domain, researchers have compared PISA 2006 and PISA 2015 data from among top-ranking countries to identify trends and to determine the state of science classes in each country (Kim & Koo, 2019; Lau & Lam, 2017; OECD, 2016). Several studies have applied statistical techniques (e.g., regression analysis and latent profile analysis) to explain and interpret the relationship between science learning and achievement (Forbes et al., 2020; Oliver et al., 2019). Since many variables could not be entered at the same time using traditional statistical techniques, standardization indices (Bae & Sohn, 2018; Kim & Sohn, 2018) or selected questionnaire items according to a research hypothesis (Lim et al., 2018; Park et al., 2018) have been used for these studies. However, the standard indices provided by PISA do not contain specifics. Also, studies using some individual questionnaire items have a limitation in that they cannot fully reflect the complex situations in different educational contexts.

### 2.1.4 Environmental Awareness
Environmental awareness variables have not received much attention in science education. However, recently, as global environmental issues (e.g., climate change mitigation) have become critical, the need to re-establish scientific literacy related to the environment has been discussed. In addition, Education for Sustainable Development 2030 was approved recently by the United Nations General Assembly to establish sustainable development goals (SDGs), encourage global participation, and to reflect SDGs in education standards and learning objectives (UNESCO, 2020). The environmental awareness
variable has shown a high correlation between science achievement and enjoyment of science (List et al., 2020; Oliver & Adkins, 2020). This was derived as an important variable predicting the science achievement of Korean students (Hong et al., 2022).

2.2 Random Forest
Since data mining and machine learning techniques have emerged, many researchers have introduced various methods using these techniques to do analyses in the educational field (Peña-Ayala, 2014; Kamath, 2016). In conventional statistical analysis, we follow the steps of building a new hypothesis based on knowledge from the theoretical background and verifying the hypothesis with the data analysis. In data mining, however, we can extract information from a massive dataset. Random forest is one of the popular machine learning techniques that is known to be fast, flexible, and robust (Lantz, 2019). Random forest is widely used for classification and regression analysis. It was recognized as an acceptable alternative to the traditional regression-based model by showing high accuracy and sensitivity even without explicitly modeling the hierarchical nature of the educational data (Yi & Na, 2020).

Random forest predicts the outcome by ensembling multiple decision trees, thus minimizing the error of the decision tree and providing high prediction accuracy. The first step in the random forest technique is to generate multiple sample datasets to build multiple decision trees. To make a sample dataset, random forests use bootstrapping techniques, a random restoration sampling method, to select random samples. Then, random predictors when splitting nodes are used to maximize randomness and create multiple decision trees (Breiman, 2001). Finally, the results are derived from an ensemble of results obtained from multiple decision trees, each embedded in a bootstrap sample (bootstrap aggregation, also known as “bagging”). This approach provides stable and accurate results even with many predictors. An analysis is also possible even if a specific predictor has a large influence (Gareth et al., 2013). For these reasons, random forest has been actively used in various fields.

Researchers in education have also used random forests to analyze PISA data with many variables and complex interactions between variables (Bulut & Yavuz, 2019). They have applied PISA 2015 data to three different data mining techniques. Among these data mining techniques, random forest has demonstrated superior performance, although the actual dataset contains both numerical and categorical predictors. Son et al. (2020) explored the major factors according to reading literacy level group using random forest analysis with PISA 2018 Korean dataset. When there is imbalanced data, the results tend to be biased toward a group with a large number of cases, so a method of
correcting this by adjusting the sampling rate was used. Based on the importance index, major variables affecting the predicted probability of the low-achievement groups were identified. This also revealed new variables that had never been presented in previous studies. In addition, they suggested the necessity of providing educational support to improve the reading literacy of students by visualizing a partial dependence plot (PDP) of the predicted probability of the low-achievement groups according to individual variables.

PISA data analysis using random forest has been conducted in various studies. The random forest technique can provide multiple implications for interpreting students’ achievements by deriving new predictive variables related to students’ achievement and considering multiple variables simultaneously.

3 Methods

3.1 Data Set Used for Study
PISA selected participants based on age rather than grade level, with 15-year-old students being assessed. Therefore, in Korea 523 middle school students and 5,058 high school students participated in PISA 2015 (KICE, 2016). In this study, only high school students were selected as the data set considering the gap between the curriculum and system for middle school and high school in Korea. Among the selected students, 424 students with systematic missing values in science-related questionnaire items were excluded from the analysis. Therefore, only 4,634 students were analyzed.

3.2 Variables
3.2.1 Result Variable
Students’ achievement groups were categorized by the PISA 2015 science cognitive domain assessment scores. PISA assesses students’ ability to provide explanations and evaluate and design scientific inquiries and interpret data using content, procedures, and epistemological scientific knowledge in various situations. The results of the PISA science cognitive domain are calculated as a score through scaling by the item response theory. PISA scores were designed to have an average score of 500 points and a standard deviation of 100 across OECD countries. In PISA 2015, students were divided based on their science scores into proficiency levels of 1–6. Science proficiency was divided into three groups according to the PISA achievement level classification criteria: Proficiency Levels 5 and 6 were high; 2, 3, and 4 were moderate; and 1 was low (OECD, 2016). High-achievement group students in science are those who can
apply scientific knowledge in a variety of complex life situations in some of the high cognitive demands. On the other hand, low-achievement group students in science are those who use a little scientific knowledge in a few familiar life situations that require a low level of cognitive demand (OECD, 2017).

The achievement level group ratios of the selected 4,634 students were 10.8% (500 students), 78% (3,616 students), and 11.2% (518 students) in the high, moderate, and low groups, respectively.

3.2.2 Predictor Variable
In this study, we focused on the nine science-related contextual variables that could improve through government policies and teachers’ efforts. The 57 items, not standardized indices, included in the PISA 2015 student and parent questionnaires were used as predictor variables (students’ past science activities, current science activities, environmental awareness, inquiry-based science teaching and learning practices, teacher-directed science instruction, perceived feedback, the adaption of instruction, teacher support in a science class of a student’s choice, and disciplinary climate in science classes; see Table A1 in the Appendix). Each question was answered based on a 4-point Likert scale. Some questions were reverse coded to indicate the direct relationship between score and frequency of activities and classes. The data used in the analysis showed an effective response rate of ≥ 97%.

3.3 Data Analysis
This study focused on the factors that affect the science achievement level of high school students (mainly 15-year-old students) in Korea using the PISA 2015 dataset. Since we intended to identify different factors according to achievement level, we divided the original data into high-, moderate-, and low-achievement groups. The target group was set to high- and low-achievement groups and the control group was set to the moderate-achievement group. Generally, datasets based on achievement tend to be imbalanced in terms of the number of cases in the target group and the control group. The dataset used in the current study was an imbalanced dataset that had a big difference between the target groups (500 students in the high-achievement group and 518 students in the low-achievement group) and the control group (3,616 students in the moderate-achievement group). To overcome this problem, we adapted the method of random forest with hybrid sampling and a hyper-ensemble approach (ensemble of ensembles). The schematic workflow of the random forest with a hyper-ensemble method is presented in Figure 1.
3.3.1 Generating Multiple Balanced Datasets by Hybrid Sampling
In this study, we used a hybrid random sampling technique for balancing the dataset using the ROSE package (Lunardon et al., 2014), which under-samples the majority cases and over-samples the minority cases. The moderate-achievement group also includes PISA proficiency Levels 2, 3, and 4, so it has various ranges in relation to individual students. To avoid intentional selection of the random forest models, we generated 100 pseudo-random numbers for seed numbers that are applied to build these balanced datasets. One hundred balanced datasets were generated for the random forest training.

3.3.2 Random Forest Training
The random forest algorithm was performed with the R software using the random forest package, and absent data were calculated using the na.roughfix functions (Breiman, 2001). Each balanced dataset underwent random forest
training by building 50,000 decision trees and ensembling the results. We used 70% of the cases from each balanced dataset with bagging to train the random forest algorithm. Meanwhile, the remaining cases were used as test datasets to evaluate the random forest models. Seven (the square root of the number of total variables; \( \sqrt{57} \approx 7 \)) randomly selected variables were used to split each node from the bootstrapped datasets to generate a decision tree in a random forest training (Breiman, 2001). Accuracy, sensitivity, specificity, and area under the curve (AUC) are commonly used parameters in evaluating the random forest model, and it has been shown that closer to 1 means the model is capable of better prediction (Lantz, 2019).

We also used mean decrease accuracy (MDA) to evaluate the importance of each variable in its impact on predicting science achievement level. MDA presents how much the accuracy of the random forest model decreases when each variable is excluded in building the decision tree. The larger the MDA value, the more important the variable (Breiman, 2001). We also calculated a PDP in the important variables using the edarf package (Jones & Linde, 2016) to visualize the relationship between the student's response for each variable and the change in the predicted probability of each group, plotted on the x-axis and y-axis, respectively. During Step 2, 100 random forest models were constructed from 100 balanced data sets of the high versus the moderate groups, and 100 model evaluation values, MDA values of important variables, and PDPs were derived.

3.3.3 Hyper-Ensemble Combination of the Results of Random Forests
The random forest results from 100 balanced datasets were hyper-ensembled. The final important factors were determined by the average value of the MDA derived from each balanced dataset; 15 items, about 30% of the total predictor variables, were presented as the results. In addition, we overlapped the PDPs for the important factors obtained from 100 random forests. Both the high-versus moderate-achievement groups and the low-versus moderate-achievement groups share the same three steps of the process mentioned above.

4 Results
The ranges and average values of parameters for evaluation results from random forest models are summarized in Table 1. Although there was a difference for each balanced dataset, the accuracy, sensitivity, specificity, and AUC of group classification were close to 1, so they can be considered well-trained
random forest models. These various ranges of results also confirm that hyper-ensemble random forest models show good accuracy without overfitting problems.

4.1 Important Variables in Predicting the High-Achievement Group

Drawing from the results of the hyper-ensemble random forest, the top 15 predictor variables according to the mean of the MDA from the high-achievement group in Korea, are presented as a box plot (Figure 2).

The maximum, upper quartile, median, lower quartile, and minimum from the box plot show various ranges from each variable. Each random forest model from a balanced dataset presents a different MDA and different ranks in variables. The derived PDPs of important variables are shown in Figure 3.

The 15 rankings with the highest average importance index included past and current science activities, teacher directed, adaptive instruction, inquiry based, and environmental awareness. Among the predictor variables related to past science activities, read books on scientific discoveries; watched, read, or listened to science fiction; and attended a science club, which were sixth, eighth, and 11th in importance, respectively, were derived as the major predictors. Although the predictions of the achievement group were not clear for hardly ever and sometimes, students who performed regularly and very often were more likely to be predicted into the high group.

In the science activities, four activities were derived as major predictors of the high group. Students with a high frequency of science activities were more likely to be predicted to be in the high group. For borrowing or buying books on science topics, attending a science club, and reading scientific articles, which are fourth, fifth, and seventh in importance, respectively, students who regularly or very often did these things had a high probability of being
predicted in the high group. However, the predictions of achievement groups corresponding to hardly ever and sometimes was unclear. Although watching TV programs about science was less important than the other predictors (12th in importance), we found that students who sometimes, regularly, and very often did so were predicted into the high group.

In the case of a teacher demonstrating ideas, students who recognized the high frequency of occurrence in science class were likely predicted into the high group (10th in importance). In the adaptive instruction item, providing individual help for a student who had difficulties with understanding, 100 PDPS were mixed; thus, achievement group prediction was unclear (14th in importance).

Among the predictors related to inquiry-based science instruction, being allowed to do their own experiments and being asked to do an investigation to test an idea, which were ninth and 15th in importance, respectively, were derived as major predictors. There was a difference in the probability of group prediction according to the variable level; however, the students who answered that they were rarely allowed to design their own experiments and investigate in science classes showed a higher probability of predicted into the high group.

In the environmental awareness category, four topics were derived as important variables in group prediction, including, use of genetically modified organisms, the increase of greenhouse gases in the atmosphere, nuclear waste,
Figure 3  Partial dependence plots of the top 15 predictors for the high-achievement group
and forest clearing, which were first, second, third, and 13th in importance, respectively. A student who answered that they knew each topic and could explain a topic was more likely to be predicted into the high group.

4.2 Important Variables in Predicting the Low-Achievement Group

The box plot of the top 15 variables in the average importance index derived as a result of predicting the low-achievement group is shown in Figure 4.

The 15 rankings with the highest average importance index included past science activities, teacher directed, perceived feedback, adaptive instruction, inquiry based, and environmental awareness. Students’ perceptions of science teaching and learning played an important role in predicting; however, variables related to teacher support and disciplinary climate were not included. In addition, variables of students’ past science activities were included, but not current science activities. The PDPS of important variables derived in this study are summarized in Figure 5.

Among the predictor variables related to past science activities, the lower the frequency of reading books on scientific discoveries, the higher the probability of being predicted into the low-achievement group (seventh in importance). In the case of watched, read, or listened to science fiction, students who had almost never done it showed a slightly higher probability of being predicted into the low-achievement group; however, because the 100 PDPS were

![Figure 4](https://example.com/figure4.png)

**Figure 4** Top 15 predictors box plot of low-achievement group.
Figure 5
Partial dependence plots of the top 15 predictors for the low-achievement group.
mixed, it was a challenge to find a significant difference in the probability (15th in importance).

Students who answered that many or almost all lessons had a whole-class discussion with the teacher were also more likely to be predicted into the low-achievement group (13th in importance). In perceived feedback, students who perceived that they received feedback on their strengths or which areas they could improve, were more likely to be predicted into the low-achievement group (fourth and 11th in importance). In the adaptive instruction variable category, students who reported that the teacher almost never adapted the lesson to their classes’ needs and knowledge were more likely to be predicted into the low-achievement group (10th in importance).

In the predictor variables related to inquiry-based instruction, students who felt that the teacher almost never explained how school science ideas can be applied to different phenomena or who never made clear the relevance of broad science concepts to students’ lives were more likely to be predicted into the low-achievement group (second and fifth in importance). Interestingly, students who reported doing practical experiments and who were allowed to design their own experiments in many and almost every lesson were also more likely to be predicted into the low-achievement group (eighth and ninth in importance).

In the prediction of the low-achievement group, the predictor variable in the environmental awareness category was an important variable, with importance indices of first, third, sixth, 12th, and 14th. Based on the PDPs, students who answered that they had never heard of it or had heard it but could not explain the exact meaning were more likely to be predicted into the low-achievement group.

4.3 Common and Differentiating Variables in Predicting Achievement Group

Table 2 shows the predictors for the high- and low-achievement groups derived as a result of this study according to science-related contextual variable. The ranks of importance derived from the random forest results of the high- and low-achievement groups are displayed together.

Among the variables related to a child's past science activities, read books on scientific discoveries and watched, read, or listened to science fiction were derived as common factors. Students who did more of these activities were more likely to be predicted into higher achievement level groups. In addition, the importance ranks and PDPs were more prominent in the read books on
scientific discoveries than in the watched, read, or listened to science fiction. In addition, current science activities were derived as factors only in the high-achievement group.

Among the teaching and learning strategies and environmental awareness items included in the PISA questionnaire, predictor variables related to teacher support and disciplinary climate were not included in both groups. Among the predictor variables related to inquiry-based science instruction, allowing students to do their own experiments was derived as a common factor in both groups. It was found that the students who answered that the frequency of being allowed their own experiments was low were more likely to be

<table>
<thead>
<tr>
<th>Contextual variables</th>
<th>Question</th>
<th>High group</th>
<th>Low group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child's past science activities</td>
<td>Read books</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Watched or read</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Science club</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Current science activities</td>
<td>Borrow/buy books</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Science club</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Read articles</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Watch TV</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Teacher directed</td>
<td>Demonstrates an idea</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Whole-class discussion</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Perceived feedback</td>
<td>Strengths</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improve my performance</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Adaptive instruction</td>
<td>Individual help</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Needs/knowledge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inquiry based</td>
<td>Designing experiments</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Investigation to test</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explanation of applying</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explain how concepts relate to life</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Practical experiments</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Environmental awareness</td>
<td>Increase of greenhouse gases</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Use of GMO</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Clearing forests</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Nuclear waste</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Air pollution</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Water shortage</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>
predicted into the higher achievement level groups. Other than this, different predictors were derived according to science achievement levels. Perceived feedback-related items included only low-achievement group predictors. The teacher explains scientific ideas variable, which is in the teacher-directed instruction category, was not derived as an important variable for predicting the low-achievement group. Meanwhile, the applications of scientific concepts variable was derived as an important variable.

In the environmental awareness category, the increase of greenhouse gases in the atmosphere, the use of genetically modified organisms, and forest clearing were derived as common factors. A student who answered that they knew each topic and could explain a topic was more likely to be predicted into the higher-achievement group. Although there are slight differences between the topics, the majority of the topics showed a similar pattern. Moreover, for students who answered that they had never heard of a topic or had heard of it but could not explain the exact meaning, the range of change in the prediction probability was small. However, the probability of being predicted into the moderate-achievement group increased when students answered that they knew and could explain the general issues.

5 Conclusions and Discussion

The current study derived variables that predict high- and low-achievement groups in Korea using the EDM method of the random forest to extract factors from PISA data to associated with students categorized as being in high and low science achievement groups. The random forest revealed the specific predictors by inputting detailed questionnaire-level variables. This is an advantage that could not be confirmed in previous research methods, such as structural equations and multi-layered analysis, that have been used to explain and interpret the relationships between science-related contextual variables and achievement. Students’ past science activities (read books on scientific discoveries and watched, read, or listened to science fiction), science learning methods (which included designing experiments), and environmental awareness were all derived as variables that predict both high- and low-achievement groups.

5.1 Student Participation in Science Activities

The results of this study show that past science activities were derived as predictors for both groups, but that participation in current science activities were derived only from the high-achievement group. This is consistent
with previous studies (Tang & Zhang, 2020), which have shown that science activities during childhood positively affected current scientific activities and achievement. Thus, to improve the science achievement of students in Korea, it is vital to check the past science activities of students. As the current science curriculum in Korea starts from the third grade of elementary school, it is likely that the type and quality of science activities experienced before the age of 10, will tend to vary depending on the home background of the student. Based on the importance of past scientific activities derived from the current results, parents should be encouraged to support children to continually develop a variety of science activities from an early age. Active and systematic support from families, schools, and local communities is required. These include the continuous operation of various experiential programs in informal science learning settings, such as local science and natural history museums, that can promote and facilitate out-of-school science activities.

In this study, reading books on scientific discoveries around the age of 10 was an important predictor in both analyses. We observed through the PDPs that students who read more about scientific discoveries around age 10 showed higher scientific achievement. These results are consistent with previous studies (Lee & Chung, 2019), which have presented the positive effect of reading activities on reading, math, and science achievement in elementary school. In Korea, the importance of reading has been emphasized in all subjects. Moreover, a necessity to provide a science reading program for students with low-achievement in science has been suggested (Kim, 2021). In addition, Chung and Park (2022) applied random forest analysis to PISA data and empirically showed an effect of reading in science where reading strategy and enjoyment of reading were identified as major indicators of science achievement. Therefore, systematic support is needed to increase the opportunity to access many science books that fit a student’s proficiency level and interests from an early age.

5.2 Science Teaching and Learning Methods
In science teaching and learning methods, different predictors were derived from the high-and low-achievement groups. This is consistent with the results of previous studies (Ku & Koo, 2018), and from the results of this study, which were analyzed by entering item-level variables rather than standardized indices, and were able to confirm the specific strategies as important. In Korea, students below the science proficiency baseline showed difficulties in applying scientific knowledge to real life; thus, feedback on how to apply acquired scientific knowledge to a new context is required. Our study also found that students predicted to be in low-achievement groups also reported that they
engaged in more whole-class discussion, were supported to do practical experiences, and received routine feedback from their teachers. However, these same students were less likely to report that their teacher explained connections between school science and the real world phenomena. It has been reported that inquiry-based activities may lead to misconceptions if activities do not lead to the understanding of scientific concepts (Lau & Lam, 2017) and that some students have difficulty applying knowledge to real life, even when they have been taught a similar level of knowledge in the school curriculum (Kang & Cogan, 2020). These findings suggest it is necessary that teachers be better supported to organize the connection between experimental activities and scientific concepts and principles more systematically to better enable the application of laboratory experiences to the science learning of students in the low-achievement group.

However, this may be challenging, as studies have previously shown that science teachers in Korea have low confidence for designing hands-on activities, facilitating discussion among students, and assigning differentiated instructions for high- and low-achievement students (Kim, 2022). Therefore, it is necessary to prepare a plan that will improve the professionalism of science teachers to be able to effectively implement these strategies in their classrooms. At the same time, there is a need to develop and disseminate content related materials to support various experiments or investigations to be conducted by students in various achievement levels. This may allow teachers to more effectively implement hands-on science activities that are more appropriate for all learners their science classes.

Concerning perceived feedback, predictors were included only in the low-achievement group. Students who responded that they received feedback in some, many, and almost every lesson were more likely predicted into the low-achievement group. This may be explained by the tendency of teachers to provide more feedback to struggling students (Beaman et al., 2006); however, it is necessary to determine whether feedback to the other student groups is insufficient. While the 2015 revised curriculum in Korea emphasized process-oriented evaluation, studies have revealed that science teachers face problems implementing these evaluations as are they already burdened by many administrative tasks and a general lack of time (Kim et al., 2019). Since the purposes of feedback is to confirm the achievement of instructional goals and facilitate student learning, feedback should be received equally by all students in school science classes, regardless of achievement level. Therefore, it is necessary to build an efficient system that enables science teachers to provide appropriate feedback according to student characteristics which will allow them to provide specific and diverse feedback for students at each achievement level.
5.3 Environmental Awareness

Environmental awareness was derived as an important variable to predict high- and low-science proficiency in Korea. Around the world, research on efforts to reflect the SDGs in the curriculum as detailed by the Education 2030 guidelines (UNESCO, 2030) has become more important. In Korea, we recommend that the curriculum be structured so that the knowledge learned in a subject should be reflected in values and attitudes and should also help students participate in the society in which they will live in the future. Environment-related content can be used as teaching and learning materials related to students’ lives by reflecting the knowledge learned in science subjects in their values and attitudes. Therefore, it is necessary to provide Korean students with educational experiences that enable them to participate in environment-related social issues based on an accurate understanding of scientific terms, concepts, and principles. Moreover, active cooperation with schools, families and local communities is required so that students can grow into active participants when faced with environmental problems in their daily life. In this way, we may find that Korean students who attain a high level of environmental awareness will lead to improved scientific achievement.

6 Limitations and Implications

In this study it is important to note that due to the nature of the PISA questionnaire data, we were unable to conclude a causal relationship based on the interpretation of the student responses. Nevertheless, we could confirm the science teaching methods that caused relative difficulties in students below the baseline. So while there are some limitations for the study, the findings also provide rich opportunities for exploring factors predicting science achievement among different groups of students.

Using an ensemble machine learning model called random forest, we analyzed a large and diverse data set related to student science achievement. While random forest analysis is a relatively new method being applied in educational research, random forest analysis has a lot of potential as it allows many variables at the item level to be input, meaning more specific results can be derived. In addition, by applying the hyper-ensemble approach, more reliable and stable conclusions can be drawn, even from highly imbalanced student achievement group data, such as that presented in the PISA data sets.

Using this method, we were able to derive results that can be helpful for policy makers as they consider methods for further developing effective learning support according to the achievement levels of all students. Specifically, as it was determined that as there was an increased incidence of participation in
science activities among students in the high-achievement group, it could be argued that there is a need for increased provision of educational opportunities for more students through educational policy and school initiatives. This could prove especially beneficial if the learning opportunities also considered environmental issues and were designed with an understanding of the benefits and limitations of different teaching and learning strategies for different groups of science learners.

Abbreviations

AUC Area Under the ROC Curve
EDM Educational Data Mining
ESCS Economic, Social, and Cultural Status
MDA Mean Decrease Accuracy
PDP Partial Dependence Plot
PISA Programme for International Student Assessment
SDGS Sustainable Development Goals

Ethical Considerations

The data reported in this study do not require human subjects' approval.

About the Authors

Jeehye Hong is a doctoral student in the Department of Chemistry Education at Seoul National University in Seoul, Republic of Korea. Her research focuses on chemistry education, science teacher expertise, and data analysis with the application of machine learning.

Hyunjung Kim is an associate professor in the Department of Chemistry Education at Kongju National University. Before joining Kongju National University, she was a researcher at the Korea Institute for Curriculum and Evaluation (KICE). She is interested in science achievement, science teacher expertise, assessment methods, and international comparative research.

Hun-Gi Hong is a professor in the Department of Chemistry Education at Seoul National University in Seoul, Republic of Korea. He received his bachelor’s degree in chemistry education, and a master’s degree in chemistry, both
from Seoul National University. He received his doctoral degree in analytical chemistry from the University of Texas at Austin in the United States. After graduation, he completed a postdoctoral fellowship at the Beckman Institute of Science and Technology at the University of Illinois at Urbana-Champaign. His research interests cover various fields of analytical electrochemistry, chemistry education, science gifted education, and educational technology based on artificial intelligence, augmented reality, and virtual reality.

References


Appendix

### Table A1 Composition of predictor variables

<table>
<thead>
<tr>
<th>Contextual variables</th>
<th>Question (4-point Likert scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Child’s past science activities</strong></td>
<td>Watched TV programs about science</td>
</tr>
<tr>
<td></td>
<td>Read books on scientific discoveries</td>
</tr>
<tr>
<td></td>
<td>Watched, read, or listened to science fiction</td>
</tr>
<tr>
<td></td>
<td>Visited web sites about science topics</td>
</tr>
<tr>
<td></td>
<td>Attended a science club</td>
</tr>
<tr>
<td></td>
<td>Construction play, e.g., Lego bricks</td>
</tr>
<tr>
<td></td>
<td>Took apart technical devices</td>
</tr>
<tr>
<td></td>
<td>Fixed broken objects or items, e.g., broken electronic toys</td>
</tr>
<tr>
<td></td>
<td>Experimented with a science kit, electronics kit, or chemistry set, used a microscope or telescope</td>
</tr>
<tr>
<td></td>
<td>Played computer games with a science content</td>
</tr>
<tr>
<td><strong>Current science activities</strong></td>
<td>Watches TV programs about broad science</td>
</tr>
<tr>
<td></td>
<td>Borrows or buys books on broad science topics</td>
</tr>
<tr>
<td></td>
<td>Visits web sites about broad science topics</td>
</tr>
<tr>
<td></td>
<td>Reads broad science magazines or science articles in newspapers</td>
</tr>
<tr>
<td></td>
<td>Attends a science club</td>
</tr>
<tr>
<td></td>
<td>Simulates natural phenomena in computer programs/virtual labs</td>
</tr>
<tr>
<td></td>
<td>Simulates technical processes in computer programs/virtual labs</td>
</tr>
<tr>
<td></td>
<td>Visits web sites of ecology organizations</td>
</tr>
<tr>
<td></td>
<td>Follows news of science, environmental, or ecology organization via blogs and microblogging</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Contextual variables</th>
<th>Question (4-point Likert scale)</th>
</tr>
</thead>
</table>
| Teacher-directed science instruction | The teacher explains scientific ideas.  
A whole-class discussion takes place with the teacher.  
The teacher discusses our questions.  
The teacher demonstrates an idea. |
| Perceived feedback from science teachers | The teacher tells me how I am performing in this course.  
The teacher gives me feedback on my strengths in this school science subject.  
The teacher tells me in which areas I can still improve.  
The teacher tells me how I can improve my performance.  
The teacher advises me on how to reach my learning goals. |
| Adaptive instruction in science lessons | The teacher adapts the lesson to my class's needs and knowledge.  
The teacher provides individual help when a student has difficulties understanding a topic or task.  
The teacher changes the structure of the lesson on a topic that most students find difficult to understand. |
| Inquiry-based science instruction | Students are given opportunities to explain their ideas.  
Students spend time in the laboratory doing practical experiments.  
Students are required to argue about science questions.  
Students are asked to draw conclusions from an experiment they have conducted.  
The teacher explains how a school science idea can be applied to a number of different phenomena.  
Students are allowed to design their own experiments.  
There is a class debate about investigations.  
The teacher clearly explains the relevance of broad science concepts to our lives.  
Students are asked to do an investigation to test ideas. |
| Teacher support in a science class of the student's choice | The teacher shows an interest in every student's learning.  
The teacher gives extra help when students need it.  
The teacher helps students with their learning.  
The teacher continues teaching until the students understand.  
The teacher gives students an opportunity to express opinions. |
### Table A1  Composition of predictor variables (cont.)

<table>
<thead>
<tr>
<th>Contextual variables</th>
<th>Question (4-point Likert scale)</th>
</tr>
</thead>
</table>
| Disciplinary climate in science classes | Students don’t listen to what the teacher says.  
There is noise and disorder.  
The teacher has to wait a long time for students to quiet down.  
Students cannot work well.  
Students don’t start working for a long time after the lesson begins. |
| Environmental awareness | The increase of greenhouse gases in the atmosphere  
The use of genetically modified organisms (GMO)  
Nuclear waste  
The consequences of clearing forests for other land use  
Air pollution  
Extinction of plants and animals  
Water shortage |