Constante-Amores, A., Arroyo Resino, D., Navarro Asencio, E., and Castro Morera, M., (2026). Determinants of Spanish students' low performance in reading comprehension in PISA: a machine learning approach. *Revista de Investigación Educativa, 44.* https://doi.org/ https://doi.org/ 10.6018/rie.650791

Determinants of Spanish Students' Low Performance in Reading Comprehension in PISA: A Machine Learning Approach

Determinantes del bajo rendimiento del alumnado español en comprensión lectora en PISA: un enfoque de *machine learning*

Alexander Constante-Amores[1] *, Delia Arroyo Resino**, Enrique Navarro Asencio y María Castro Morena**

* Department of Education. Camilo José Cela University (Spain)
** Department of Research and Psychology in Education. Complutense University of Madrid (Spain)

Abstract

Low performance in reading comprehension is one of the major challenges in the Spanish educational system. Therefore, the objective of this study is to investigate the contextual determinants associated with this low reading performance. The sample consists of a total of 35,943 Spanish students and 1,089 educational centers that participated in the 2018 PISA assessment. The criterion variable is reading competence, which has been dichotomized (0 = medium and high performance, 1 = low performance). A total of 721 predictors were selected as independent variables. For data analysis, Random Forest machine learning algorithm was applied, and a multilevel binary logistic regression was conducted. The 30 most important variables related to students and school center explain 46% and 24% of the criterion variable, respectively. The final model (comprising both predictors) explains 47%. Among the main conclusions, the significance of educational process variables and non-cognitive and meta-cognitive constructs in low reading performance stands out. Therefore, the importance of addressing this educational phenomenon from a perspective less linked to socio-economic determinants and more focused on pedagogical aspects is emphasized.

Keywords: PISA, machine learning, low performance, reading competence

[1] **Correspondence:** Delia Arroyo Resino (garroy01@ucm.es). Department of Research and Psychology in Education, Faculty of Education, Complutense University of Madrid.

Resumen

El bajo rendimiento en comprensión lectora es uno de los grandes problemas del sistema educativo español. Debido a esto, el objetivo de dicho trabajo es estudiar los determinantes de contexto que se asocian a este bajo desempeño lector. La muestra se encuentra conformada por un total de 35943 estudiantes españoles/as y 1089 centros educativos que participaron en PISA 2018. La variable criterio es la competencia lectora, la cual se ha dicotomizado (0 = medio y alto rendimiento y 1 = bajo rendimiento). Como variables independientes se seleccionaron un total de 721 predictores. Para el análisis de los datos se aplicó el algoritmo de machine learning Random Forest y se realizó una regresión logística binaria multinivel. Las 30 variables más importantes relacionadas con los y las estudiantes y centro escolar explican el 46% y 24% de la variable criterio, respectivamente. El modelo final (formado por ambos predictores) explican un 47%. Entre las principales conclusiones se destaca la relevancia que tienen las variables de proceso educativo y los constructos no cognitivos y meta-cognitivos en el bajo rendimiento lector. Por lo tanto, se subraya la importancia de trabajar este fenómeno educativo desde una perspectiva menos vinculada a determinantes socioeconómicos y más orientada hacia aspectos pedagógicos.

Palabras clave: PISA; *machine learning*; bajo rendimiento; competencia lectora.

Introduction

One of the major problems of the Spanish education system is the low level of student achievement in compulsory education (Antelm et al., 2018; Garrido et al., 2020). The State System of Education Indicators argues that this leads to early school leaving (Ministry of Education and Vocational Training, 2020a). It also leads to demotivation, low academic self-concept (Fernández-Lasarte et al., 2019), low chances of entering the labour market (Marchesi, 2003), higher risk of social exclusion (European Commission, 2016) and lower economic and cultural growth of the country (Astakhova et al., 2016).

Reading literacy is one of the skills with the highest rate of underperforming students. In PISA 2018, where it was the main skill tested, 23% of Spanish students showed a low level, with an average student score of 477, significantly lower than the OECD average (487) and the European Union overall (489) (Ministry of Education and Vocational Training, 2020b).

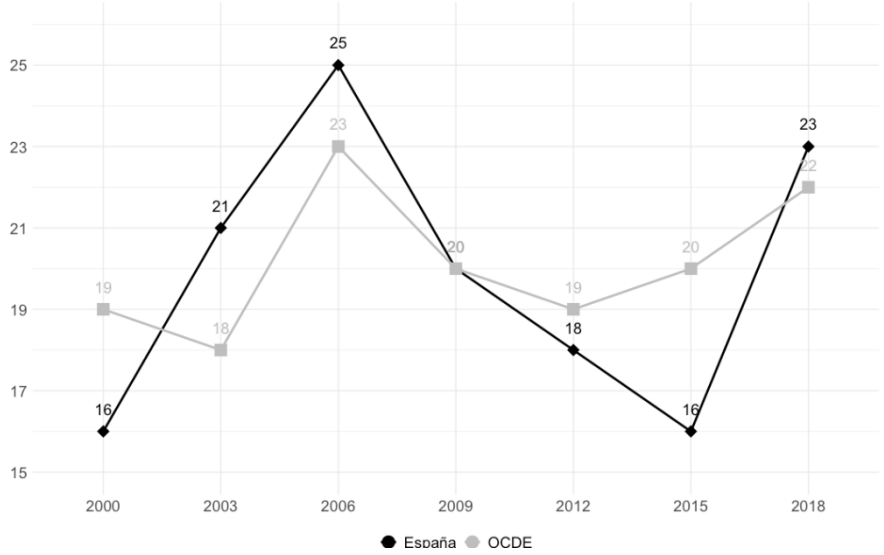
Reading literacy in PISA 2018 is defined as "understanding, using, evaluating, reflecting on and engaging with texts to achieve one's own goals, develop knowledge and participate in society" (OECD, 2018, p. 9). It is assessed on a scale with mean 500 and standard deviation 100, distributed over six levels (1 to 6). Level 1, considered low performers, is subdivided into level 1a students who identify relevant information in simple texts; level 1b students who understand short texts and locate fragments; and level 1c students who master basic vocabulary but do not understand complex relationships or meanings. From level 2 onwards (considered the minimum for adequate performance and successful social and economic inclusion), it is required to interpret relationships in the

text, while higher levels involve understanding complex texts (level 3), long texts (level 4), distinguishing content and purpose (level 5) and analysing abstract texts (level 6) (OECD, 2016, 2019) (OECD, 2016, 2019).

In Spain, as shown in Figure 1, the percentage of students with low performance (level 1) has risen from 16% in 2015 to 23% in 2018, which is similar to the average for OECD countries (22%) (Ministry of Education and Vocational Training, 2020b). This trend with respect to the percentage of students in the low-performing levels is mitigated in the years 2009, 2012 and 2015, but again appears to be increasing.

Figure 1

Evolution of the percentage of students with low performance in reading literacy in Spain and the OECD



Note: Prepared by the authors based on Ministry of Education and Vocational Training (2020b).

Low performance in reading literacy has a number of negative consequences. At the academic level, as it is closely linked to performance in other subjects (mathematics, science and language), it can have a negative impact on these subjects (Viramontes et al., 2019), as there may be a positive and significant relationship between them (García et al., 2018). It also hinders the possibility of obtaining meaningful learning (Molina Ibarra, 2020). At a cognitive and personal level, a lack of reading comprehension is linked to a certain inability to think critically. Reading comprehension allows for analysing, inferring and evaluating information, so its absence limits the development of these essential skills (Huamán Herreros, 2024). All of this has repercussions on self-concept and reduces interest in learning (Martín-Ruiz and González-Valenzuela, 2022). At the occupational and social level, poor reading comprehension affects the ability to interpret documents,

follow instructions and communicate effectively, impacting on employability and the exercise of citizenship (Molina Ibarra, 2020), since, as García et al. (2018) point out, the ability to learn autonomously throughout life is based on adequate reading comprehension.

Therefore, early identification of factors that contribute to low reading achievement is crucial, taking into account its multidimensional nature (Choi and Calero, 2013; Hu et al., 2022).

Following PISA 2018, such factors can be grouped around three areas: student background, educational processes, and non-cognitive and meta-cognitive constructs (Figure 2).

Figure 2

Modular structure of context variables in PISA 2018

<table>
<tr><th></th><th colspan="2">Constructos de antecedentes de los estudiantes</th><th colspan="3">Constructos asociados a la escuela</th><th>Constructos no cognitivos y metacognitivos</th></tr>
<tr><td rowspan="4">Competencia lectora</td><td colspan="2" rowspan="4"></td><td colspan="3">ENSEÑANZAS Y APRENDIZAJE</td><td rowspan="4">4. Medidas relacionadas con la lectura: actitudes, motivación y estrategias</td></tr>
<tr><td>5. Experiencia científica fuera de la escuela</td><td>1. Cualificación y conocimiento profesional del profesor</td><td>2. Práctica de enseñanza de las ciencias</td><td>11. Tiempo de aprendizaje y currículo</td></tr>
<tr><td colspan="3">POLÍTICAS ESCOLARES</td></tr>
<tr><td colspan="3">3. Entorno de aprendizaje escolar de la lectura</td></tr>
<tr><td rowspan="4">Categorías generales</td><td rowspan="4">6. Estatus socioeconómico del estudiante y la familia

7. Migración y cultura</td><td rowspan="4">8. Recorrido educativo en la primera infancia</td><td>13. Implicación de los padres</td><td>12. Clima escolar: relaciones interpersonales, confianza y expectativas</td><td>14. Contextosy recursos escolares</td><td rowspan="4">9. Variables disposicionales y enfocadas en la escuela

10. Disposiciones para la competencia global</td></tr>
<tr><td colspan="3">GOBERNANZA</td></tr>
<tr><td colspan="2">16. Evaluación del estudiante, evaluación institutional y responsabilidad</td><td>15. Asignación, selección y elección</td></tr>
<tr><td colspan="3"></td></tr>
</table>

Note: OECD (2019, p. 220).

Students' backgrounds stand out in the literature as the most important variables in explaining low performance in reading literacy (Calero et al., 2010; Garrido et al., 2020) , especially students' socio-economic background (Choi and Calero, 2013). Research by Franco et al. (2016) indicates that low socioeconomic status is linked to low reading comprehension scores. The work of Gil-Flores and García-Flores (2017) shows the importance of student and school socio-economic status on PISA performance. In addition, the OECD (2015) highlights that schools with a high percentage of students from disadvantaged socio-economic backgrounds are more likely to have low scores in reading literacy.

Other variables related to this area are grade repetition, gender and reading habits. In relation to the former, research by Calero and Choi (2013) and Garrido et al. (2020) shows that students who have repeated a grade are more likely to have a low performance in reading literacy . In relation to the latter, it is boys who have a higher low performance (Cordero et al., 2011; Instituto Nacional de Estadística, 2018; Ostria et al., 2014). With regard to reading habits, work by Akande and Oyedapo (2018) indicates that students with lower performance in reading comprehension spend less time reading. In this sense, when there are no consolidated reading habits, we find low levels of reading performance (Jiménez-Pérez et al., 2020).

In relation to the area of *teaching and learning processes,* a set of variables that have a direct impact on reading comprehension stand out, such as the quality of the school environment (González and Jackson, 2014), including family involvement (Murillo and Hernández-Castilla, 2020), teacher training (Cutimbo, 2008; Vélez et al., 1994), available resources (Beltrán et al., 2011; Rouse and Krueger, 2004; Albelda, 2019) and the institutional climate (Albelda, 2019), among others. Improving these processes is key to transforming academic performance and moving towards a more inclusive and effective school.

Finally, among the *non-cognitive and metacognitive* constructs, students' educational aspirations stand out. Garrido et al. (2020) show that students who expect to complete university studies are more likely to achieve high performance in reading literacy . Furthermore, students' attitudes and motivation are also relevant, and the OECD (2016) shows that students with negative academic attitudes are more likely to perform poorly. Research by Franco et al. (2016) shows that motivation together with attitude towards reading and the use of free time are fundamental in the development of reading literacy. Also, Guerra-García et al. (2021) indicate that demotivation is one of the main causes of low performance when interpreting a written text.

Therefore, performance depends not only on academic skills, but also on personal and emotional expectations, which requires an education that fosters commitment, self-esteem and a sense of purpose in students.

Analyses of these factors have traditionally been approached using multilevel models (Calero et al., 2010; Choi and Calero, 2013; Garrido, 2020), valued for their balance between parsimony and fit to the data (Pardo and Ruiz, 2013). However, the selection of predictors in these models is often based on theoretical criteria, which may imply a limited choice of variables and the exclusion of other potentially relevant variables due to lack of empirical support. In this context, data mining and machine learning approaches gain relevance, allowing the identification, among large volumes of information, of the variables with the greatest predictive capacity (Constante-Amores et al., 2024). This methodological integration helps to reduce specification errors and avoids excluding significant variables (Gaviria and Castro, 2005).

Following the above, the general objective of this research is to analyse, in the Spanish context, the determinants (student background, educational processes and non-cognitive

and meta-cognitive constructs) associated with low performance in reading comprehension in the framework of the PISA 2018 assessment. This general objective is specified in the following specific objectives:

- To identify the student and school variables that have the greatest influence on low performance in reading comprehension.
- To determine the variables that most influence the probability of a student's low reading achievement, taking into account the hierarchical structure of the data (level 1: student and level 2: school).

Method

Design

A secondary analysis of the PISA 2108 assessment was carried out for the study of factors influencing low reading literacy performance, as it was the core subject of that edition. The research methodology is characterised by its quantitative nature, being a non-experimental, cross-sectional and predictive design.

Participants

The sample consisted of 35943 Spanish students and 1089 schools. The mean age of the subjects was 15 years (SD = 0.28), of which 49.95% were female and 50.04% were male.

Variables

The response variable is reading comprehension in PISA 2018, which has been dichotomised (0 = medium and high performance and 1 = low performance). A student is considered low achiever when he/she does not reach level 2 in reading comprehension.

As predictor variables, the questionnaires administered to students (student, use of information and communication technologies, student well-being, educational pathway) and to school principals (school questionnaire) were used. The variables from the financial literacy questionnaire were not included, as they focus on knowledge that is not related to reading literacy. Of the 739 initial predictor variables, 18 were eliminated as they exceeded 70 % of missing values. Finally, 558 student variables and 163 school variables were used. Moreover, the predictors were kept at their original level, without aggregation or disaggregation between levels.

Procedure and data analysis

In order to answer the first research objective, aimed at identifying the student and school variables that have a greater association with low performance in reading comprehension, Breiman's (2001) *Random Forest* classification algorithm was used. First, the data were preprocessed following Arroyo et al. (2024a). The student and school

database were randomly divided into two sets: a training set (70%), used to fit the model, and a validation set (30%), used to evaluate its performance. The most important *Random Forest* hyperparameters (number of trees, number of variables in each tree and maximum tree depth) were then jointly optimised in the training sample by cross-validation of 10 folds. The hierarchical structure of the data was not considered in the implementation of the algorithm, as the multilevel version of Random Forest for classification is still under development. Therefore, schools with less than 20 students were not excluded, which is a fundamental criterion when building multilevel models (Gaviria and Castro, 2005). Nor were weights applied according to sample weight, as this is an exploratory analysis.

When implementing a machine learning algorithm, there is no unified criterion for handling Plausible Values (PV). For example, the study by Hu et al. (2022) used the first PV. Arroyo et al. (2024a) and Arroyo et al. (2024b) opted to select the PV they considered most accurate (lowest prediction error). The present research followed the methodology adopted by the latter authors, but in this research, being a categorical variable, the value that correctly classified a higher percentage of cases in both categories (0 = not low performance and 1 = low performance) for both the student and centre databases was used. Table 1 presents the most accurate VPs together with their respective hyperparameters, jointly optimised on the training sample. An excellent performance of the models is observed in the validation sample.

Table 1

Random Forest hyperparameters and accuracy (level 1: student and level 2: centre)

	Hyperparameters (training sample)			***Accuracy* (validation sample)**
	Number of variables	Number of trees	Maximum depth	
VP 6 (level 1: student)	5	1000	20	86%
VP 5 (level 2: centre)	45	1000	10	81%

Once the most accurate VPs were identified, the most important variables were selected. One of the limitations of Random Forest and *machine learning* algorithms in general is that there is no specific cut-off point when selecting the most relevant predictors (Raschka and Mirjalili, 2019). Therefore, following Gorostiaga and Rojo-Alvarez (2016), three sets of variables (15, 20 and 30) were evaluated to determine the optimal number of predictors. The analysis showed that the set of 15 variables, selected using the Gini index, provided the highest accuracy in both the student and school databases.

To address the second objective, which was to identify the variables that affect the probability of a student's low performance, weights were applied according to sample weights, as recommended by the OECD to avoid biased estimates (OECD, 2025) using the multilevel logistic regression technique, taking into account the hierarchical structure of the data (level 1: student and level 2 = school), as the random variance between schools in the null model was statistically significant and the Intraclass Correlation Coefficient was above 10% (Lee, 2000). Prior to modelling, variables were correlated to eliminate those with a correlation higher than 0.80 (Kassambara, 2018). In this sense, the predictors, *number of girls* and *number of boys* in the school had a high relationship, so we proceeded to eliminate the latter as it had a lower influence on our variable of interest. The variable *What do you think you will be doing in 5 years' time?* was also dropped, as it provided the same information as Do you expect to *complete university studies?* and Do you expect to *complete a baccalaureate or intermediate degree?* Then, to ensure a proper multilevel analysis, schools with less than 20 students were eliminated (Gaviria and Castro, 2005), which meant the exclusion of 1532 subjects from 133 schools. The final sample consisted of 34411 students from 976 schools. The analysis was conducted following the methodological recommendations of Laukaityte and Wiberg (2017) for the use of multilevel models using the 10 VPs.

A total of four models were estimated. Model 0 is the null model (no predictors). Model 1 includes only the student variables and model 2 only the school predictors. In this way, the contribution of each set of variables can be observed separately. Finally, model 3 includes all variables. It should be noted that the order in which the independent variables were introduced into the predictive model was that obtained with the Random Forest algorithm. The interpretation of the results was based on the *odds ratio*, while the explanatory power of the model was assessed using the marginal and conditional R^2. The former reflects the variance explained by the fixed effects and the latter that of the full model, including random effects, which allows assessing both the contribution of the predictors and the hierarchical structure (Nakagawa and Schielzeth, 2013).

Finally, the AIC, BIC and Deviance indices were used to check the fit of the models, also, in order to compare nested models, the significance of this reduction statistic was calculated and the percentage of reduced variance (R^2) was estimated (Cameron and Windmejier, 1997). All analyses were performed with R software, version 4.3.0.

Results

The results are presented below according to the objectives.

Objective 1: Identification of the predictors that most influence low performance.

Table 2 shows the most important student variables. As can be seen, the predictor with the greatest influence on low performance in reading comprehension, , is grade repetition, followed by educational aspirations.

Table 2

Relationship of variables most associated with low performance in Reading Comprehension (student level)

	Independent variable	Gini index
1	Having repeated a grade in primary school.	497.431
2	Student expectations: do you expect to complete university studies?	353.285
3	Having repeated a grade in secondary school.	288.373
4	Expectations of the student: do you expect to complete baccalaureate or intermediate degree?	244.538
5	What do you think you will be doing in 5 years?	225.329
6	Do you currently attend supplementary lessons in the subject of Language?	221.352
7	I carefully check whether the most important facts are represented when I summarise.	214.545
8	In the PISA test, how do you feel about the reading tasks: I found many texts difficult.	204.985
9	I find the Internet a great resource for information that interests me (news, sports, dictionary, etc.).	182.776
10	How many books do you have at home?	181.399
11	In the PISA test, how do you feel about the reading questions: there were many words I could not understand.	175.301
12	How informed are you about climate change and global warming?	166.685
13	I verify the sender's email address.	163.458
14	I try to copy as many sentences as possible accurately when summarising.	147.937
15	How long, on average, do your lessons last?	145,35

In relation to the school variables (see Table 3), it can be seen that the predictors with the greatest influence on low performance are those linked to parental involvement and teacher training.

Table 3

Relationship of variables most associated with the risk of low performance in reading literacy (school level)

	Independent variables	Gini index
1	Proportion of parents: volunteers in physical or extracurricular activities.	409.943
2	Teachers with a full-time Master's degree or equivalent.	369.328

3	Part-time teaching staff with a Master's degree or equivalent.	297.286
4	Full-time teachers with a bachelor's degree or equivalent.	287.791
5	Teachers with a bachelor's degree or equivalent on a part-time basis.	230.467
6	Full-time certified teachers	228.013
7	Percentage of students from socio-economically disadvantaged homes	228.013
8	School financing for the school year: student fees or school charges paid by parents	198.487
9	Percentage of students whose language of origin is different from Spanish.	197.764
10	Percentage of students with special educational needs	197.535
11	Part-time certified teachers	191.345
12	In the last academic year, what proportion of students left school without the Compulsory Secondary Education qualification?	143.522
13	Total number of students	132.686
14	Number of girls	117.493
15	Number of boys	106.28

Objective II: Probability of low performance

Once the variables with the greatest impact on low performance had been identified, the multilevel binary logistic regression models were run, but first the independent variables of a categorical nature were dichotomised, as recommended by Pardo and Ruiz (2013) (see Table 4).

Table 4

Categorical independent variables in the multilevel logistic regression model

Label	Independent variable	Recoded values
Repetition in primary school	Having repeated a grade in primary school.	0 = No 1 = Yes, once or more
Secondary repetition	Having repeated a grade in secondary school.	
University studies	Student expectations: Do you expect to complete university studies?	0 = No 1 = Yes
Baccalaureate or intermediate level	Student's expectations: do you expect to complete high school?	
Attendance at supplementary lessons	Do you currently attend supplementary lessons in the subject of Language?	

I check summary	I carefully check whether the most important facts are represented when I summarise.	0 = Not very useful, hardly useful, not useful at all and not useful at all 1 = Useful, quite useful and very useful
Copy summary	I try to copy as many sentences as possible accurately when summarising.	
Reading tasks	In the PISA test, how do you feel about the reading tasks: I found many texts difficult.	0 = Strongly Disagree or Disagree 1= Strongly agree or agree
Internet	I find the internet a great resource for information that interests me (news, sports, dictionary, etc.).	
Reading	In the PISA test, how do you feel about the questions on reading: there were a lot of words I couldn't understand.	
Books at home	How many books do you have at home?	0 = 25 books or less 1 = 26 books or more
I check mail	I check the sender's email address.	0 = Not very appropriate, hardly appropriate and not appropriate at all 1 = Appropriate, fairly appropriate and very appropriate
Science	How well informed are you about climate change and global warming?	0 = I have never heard of it or I have heard of it, but, I can't explain what it is really about. 1 = I know how to explain it or I am familiar with it.

Table 5 shows the results of the four estimated models. In model 1 all variables are statistically significant and explain 43% of the variance of the response variable. The predictors with the largest effect sizes are *grade repetition in primary education, grade repetition in secondary education*, *educational expectations* and *attendance at supplementary language lessons.* In model 2 the school variables explain 14% of the variance, the independent variable with the largest effect is the *percentage of students from socio-economically disadvantaged households* . In model 3, where all variables are included, it is observed that all student level predictors remain significant, however, the only one that is significant at level 2 is the *percentage of students from socio-economically disadvantaged households*. The variables with the largest effect are *grade repetition in primary education and*

grade repetition in secondary education. Specifically, students who have repeated at least one grade are 216% and 113% more likely to belong to the low reading achievement group in primary and secondary education, respectively. The percentage of variance explained by introducing school variables increases by 1%.

Finally, as far as the fit of the models is concerned, it is observed that the model with the lowest AIC and BIC is the model composed of all the predictors (model 3), where a significant reduction of variance equivalent to aR^2 of 28% is achieved with respect to model 1. The reduction of variance of model 1 with respect to model 0 (27%) and of model 2 with respect to the null (1%) was also significant.

Table 5

Probability (odds ratio) of low reading comprehension achievement

	Model 0 (null)	**Model 1 (students)**	**Model 2 (school) (school)**	**Model 3 (all) (all)**
Primary repetition		3.133 ***		3.158 ***
University studies		0.560 ***		0.559 ***
Secondary school repetition		2.132 ***		2.134 ***
Baccalaureate or intermediate level		0.609 ***		0.607 ***
Attendance at supplementary classes		2.086 ***		2.084 ***
Verified summary		0.502 ***		0.503 ***
Reading tasks		1.808 ***		1.809 ***
Internet		0.573 ***		0.575 ***
Books at home		0.659 ***		0.660 ***
Reading		1.666 ***		1.667 ***
Climate change and global warming		0.718 ***		0.714 ***
I check mail		0.584 ***		0.584 ***
Copy summary		1.738 ***		1.739 ***
Time		1.006 ***		1.006 ***
Parental involvement in school activities			1.001	1.000
Full-time teachers (Master)			0.999	1.000
Part-time teaching staff (Master's)			1.010	1.004
Full-time teaching staff (Bachelor)			1.000	1.001
Part-time teachers (Bachelor's degree)			0.973	0.967

Full-time certified teachers			1.001	1.000
Percentage of students from socio-economically disadvantaged households			1.001***	1.003***
School funding			0.998***	1.000
Percentage of students whose language of origin is other than Spanish.			0.992	0.967
Students with special needs			1.029	1.000
Part-time certified teachers			1.011	0.997
Without secondary education degree			1.001	1.004
Total number of students			1.000	0.999
Number of girls			0.996**	0.992
Intercept	0.223***	0.744*	0.225***	0.149***
Variance	0.9718	0.715	0.970	0.709
Marginal R2		35%	1%	38%
Conditional R2	23%	46%	24%	47%
Adjustment				
AIC	383286.414	281850.241	383025,585	281450,708
BIC	383303.301	281985.379	383160,723	281704,092
Log likelihood	191641.1245	- 140909.121	-191496,792	- 140695,354
Deviation	383282.125	281818.214	382994.457	275718.114
Pr(>Chisq)		0.000	0.000	0.000

Discussion

Due to the impact that low reading achievement can have on students' academic, social and working lives, the *first objective* of this study was to identify the student and school variables that are most closely related to low performance in reading comprehension.

In relation to the student, it was observed that the variables with the greatest weight were: *having repeated a grade in primary school* (student background) and the *student's expectations with respect to completing university studies* (non-cognitive measures). Regarding the first variable, research by Choi and Calero (2013) and Garrido et al. (2020) indicates that students who have repeated a grade are more likely to have a low performance in reading literacy, and it is also closely related to early school leaving and school failure (Crespo, 2018). Regarding the second variable, it seems that students who have high and positive expectations have a high level of achievement (Gonzalo et al., 2022), specifically Garrido et al. (2020) highlight that students who expect to complete university studies are more likely to have a high performance in reading literacy.

Therefore, this reality requires a critical look at the education system, where in order to prevent grade repetition, individualised reinforcement plans can be established based on the results obtained in periodic diagnostic evaluations, support teachers, inclusive methodologies (cooperative learning, differentiated teaching , etc.) and, of course, the participation of families so that there is adequate school-family communication. This participation is also fundamental to cultivate high academic expectations in students, encouraging a positive academic self-concept from the school, generating a motivating school climate, where students are helped to establish clear and accessible goals.

Regarding the school variables, the predictors that have the greatest influence are, first, *the proportion of parents: volunteers in physical or extracurricular activities, in* this sense, González and Jackson (2014) find that the participation of parents in the school has a positive and significant impact on reading, likewise Murillo and Hernández-Castilla (2020) show that the participation of parents in extracurricular activities has a positive impact on reading. The second variable that has a greater impact is the variable referring to *teachers with a master's degree or full-time equivalent* (teaching-learning process), in this sense, several studies conclude that there is a positive and significant correlation between the level of teacher training and the academic performance of students (Cutimbo, 2008; Vélez et al., 1994).

Given these results, it seems that family participation not only strengthens the school-home bond, but also translates into significant improvements in reading comprehension. Therefore, it is important to invite families to collaborate in different projects, offering flexible schedules and even facilitating virtual participation. Similarly, it is also important to provide and facilitate quality training for teachers, where they are encouraged to keep updating their skills, as this has an impact on the academic performance of their students. These findings underline the need for policies that promote family involvement and continuous teacher training as pillars for improving educational outcomes.

Regarding this first objective, of the 30 most important student and school-related variables, 10 are student background variables, 12 are teaching-learning process variables and 8 are non-cognitive measures. These results differ from a large body of literature that advocates the importance of student background predictors in reading comprehension performance, as opposed to other areas (Barrera et al., 2019; Choi and Calero, 2013; Franco et al., 2016; Guio and Choi, 2014). Thus, this study provides empirical evidence that questions the traditional view of the literature, since the results show that, although the variables linked to antecedents are relevant, those related to the teaching-learning process and non-cognitive measures have an even greater weight. Therefore, it would be advisable to consider school and motivational factors, which are easily modifiable, in educational policies to improve reading comprehension.

As for the *second objective*, based on determining the variables that most influence the probability that a student has a low performance in reading comprehension, taking into account the hierarchical structure of the data, we found that in model 1 (student variables) all were statistically significant. We found that in model 1 (student variables) all the

variables were statistically significant. In relation to the coefficients , those with the greatest effect are those linked to the student's background, specifically *repeating a year*. Thus, students who repeat a year in Primary Education are 216% more likely to have a low performance in reading comprehension compared to those who have never repeated a year, and this probability is reduced in the following educational stage to 113%. Along these lines, Asensio et al. (2018), using data from PISA 2015, find that repetition is indeed associated with low achievement, and Hattie (2009), as a synthesis of 800 meta-analyses, concluded that repetition is one of the most important variables in achievement levels, with an effect of 16%. Another of the variables with the greatest impact, in this case in the area of teaching and learning processes, refers to *attendance in supplementary language classes*; those students who attend have a 108% lower PISA score than those who do not attend. This may be due to the fact that the students who attend these classes are those who need reinforcement or support in this subject, in this sense, when there are no consolidated reading habits, we find low levels of performance in this competence (Jiménez-Pérez et al., 2020). With respect to model 2, referring to the school variables, the variable with the greatest significant weight belongs to the area of student background and refers to the *percentage of students from socioeconomically disadvantaged homes and without a secondary education qualification.* In this sense, for each additional student in the school *from socio-economically disadvantaged households,* the probability of underachievement increases by about 1%. This is in line with Barrera et al. (2019) who show that students from low socio-economic backgrounds achieve lower cognitive performance in tasks such as reading comprehension. Likewise, Franco et al. (2016) show that low socioeconomic status is associated with low scores in reading comprehension. Finally, in model 3, where both student and school variables are introduced, it is observed that while all the student variables are significant, at the school level only the *percentage of students from socioeconomically disadvantaged households* was significant, thus, the percentage of variance explained by the student variables is higher than that of the school (46% and 24%, respectively).

The results confirm that student variables, especially grade repetition, are the most important predictors of low performance in reading comprehension, which is evidence of the persistence of an educational model that fails to compensate for initial inequalities. However, school factors, such as the concentration of socio-economically disadvantaged students, are also found to influence performance, although to a lesser extent. This finding raises the need to rethink repetition policies, replacing them with preventive and early support strategies, and to strengthen equity between schools through additional resources and compensatory programmes. Only a comprehensive intervention that combines actions on the student and the school context will make it possible to reduce the gaps in reading comprehension.

Thus, following the classification made by PISA 2018, it seems that it is the variables related to the student's background that have the greatest impact on the response variable, especially grade repetition and academic expectations, both variables inversely related

since, as Arroyo et al. (2019) indicate, it is observed that the probability of repeating a grade is lower when students have higher educational aspirations. Based on the above, it would be appropriate for schools to work on improving student self-concept, since this has an impact on students' academic expectations and therefore on their performance in different subjects, as shown by Carrillo-López et al. (2022). It would also be interesting to develop activities that reinforce students' motivation towards reading in order to improve reading comprehension performance and, thus, mitigate the effect that the lack of mastery of this skill may have on other subjects and on social skills.

In addition to the educational contributions, this research is also characterised by its methodological contribution, as it analyses the influence of 721 independent variables on low reading performance in PISA using the Random Forest machine learning algorithm, whereas traditionally, this phenomenon has been studied using multilevel logistic regression techniques (Calero et al., 2010; Choi and Calero, 2013; Garrido, 2020). In this sense, this work may lay the groundwork for future secondary PISA studies using this methodological approach characterised by its precision.

In relation to the limitations of the study, it should be noted that, although this study does not make international comparisons and the purpose of the first analysis was purely exploratory, one methodological limitation is not having applied sampling weights when identifying which student and school variables are associated with low performance in reading literacy, which could lead to some bias in the estimates. Also, because of its cross-sectional nature, PISA does not allow causal relationships to be established, limiting inference about the factors associated with performance. In order to establish conclusions of this type, it would be necessary to apply structural equation modelling to test hypotheses about the direct and indirect effects between variables. On the other hand, the cross-sectional design of this study prevents us from assessing the evolution of results over time, which is a methodological limitation of PISA studies.

As a prospective research project, it would be interesting to determine the predictors associated with low reading literacy performance internationally in order to see whether in other European countries the process variables and cognitive and metacognitive measures are as determinant as in this research.

References

Arroyo Resino, D.; Constante Amores, I. A., y Asensio Muñoz, I. (2019). La repetición de curso a debate: un estudio empírico a partir de PISA 2015. *Educación XX1*, *22*(2), 69-92. https://doi.org/10.5944/educxx1.22479

Akande, S. O., y Oyedapo, R. O. (2018). Developing the Reading Habits of Secondary School Students in Nigeria: The Way Forward. *International Journal of Library Science*, *7*(1), 15-20. Developing the Reading Habits of Secondary School Students in Nigeria: The Way Forward

Albelda Esteban, B. (2019). Contribución de las bibliotecas escolares a la adquisición de competencias en comprensión lectora en educación primaria en España: una aproximación a partir de los datos del estudio PIRLS 2016. *Revista de educación*, *384*, 1-9. https://doi.org/10.4438/1988-592X-RE-2019-384-408

Antelm Lanzat, A. M., Gil López, A. J., Cacheiro González, M. L., y Pérez Navío, E. (2018). Causas del fracaso escolar: Un análisis desde la perspectiva del profesorado y del alumnado. *Enseñanza & Teaching: Revista Interuniversitaria de Didáctica*, *36*(1), 129-149. https://doi.org/10.14201/et2018361129149

Arroyo Resino, D., Constante-Amores, A., Castro, M., y Navarro, E. (2024b) School effectiveness and high reading achievement of Spanish students in PISA 2018: a machine learning approach. *Educación XX1*, *27*(2), 223-251. https://doi.org/10.5944/educxx1.38634

Arroyo Resino, D., Constante-Amores, A., Gil-Madrona, P., y Carrillo-López, P. J. (2024a). Student well-being and mathematical literacy performance in PISA 2018: a machine-learning approach. *Educational Psychology*, *44*(3), 340–357. https://doi.org/10.1080/01443410.2024.2359104

Asensio Muñoz, I., Carpintero Molina, E., Expósito Casas, E., y López Martín, E. (2018). ¿Cuánto oro hay entre la arena? Minería de datos con los resultados de España en PISA 2015. *Revista Española de Pedagogía (REP)*, 76(270), 225-245. https://doi.org/10.22550/REP76-2-2018-02

Astakhova, K. V., Korobeev, A. I., Prokhorova, V. V., Kolupaev, A. A., Vorotnoy, M. V., y Kucheryavaya, E. R. (2016). The Role of Education in Economic and Social Development of the Country. International *Review of Management and Marketing*, *6*(1S), 53-58. https://econjournals.com/index.php/irmm/article/view/1865

Barrera, J. E. C., Polanco, J. G., y Acosta, J. D. (2019). Comprensión lectora de estudiantes universitarios. Factores asociados y mecanismos de acción. *Revista Venezolana de Gerencia*, *24*(87), 874-889. https://www.redalyc.org/articulo.oa?id=29060499015

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Beltrán, A. C., Seinfeld, J. C., Narro Llacza, O., y Lisboa Vásquez, C. (2011). *Hacia una educación de calidad: La importancia de los recursos pedagógicos en el rendimiento escolar*. Universidad del Pacífico, CIES.

Calero, J., Choi, A., y Waisgrais, S. (2010). Determinantes del riesgo de fracaso escolar en España: una aproximación a través de un análisis logístico multinivel aplicado a PISA-2006. *Revista de Educación*, *numero extraordinario*, 225-256. https://www.educacionfpydeportes.gob.es/revista-de-educacion/numeros-revista-educacion/numeros-anteriores/2010/re2010/re2010-09.html

Choi de Mendizábal, Á., y Calero Martínez, J. (2013). Determinantes del riesgo de fracaso escolar en España en PISA-2009 y propuestas de reforma. *Revista de Educación*, 362, 562-593. https://doi.org/10.4438/1988-592X-RE-2013-362-242

Constante-Amores, A., Arroyo-Resino, D., Sánchez-Munilla, M., y Asensio-Muñoz, I. (2024). Contribution of machine learning to the analysis of grade repetition in Spain: A study based on PISA data [Contribución del machine learning al análisis de la repetición escolar en España: un estudio con datos PISA]. *Revista Española de Pedagogía*, *82* (289), 539-562. https://doi.org/10.22550/2174-0909.4014

Cameron, A. C., y Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, *77*(2), 329-342. https://doi.org/10.1016/S0304-4076(96)01818-0

Carrillo-López, P. J., Constante-Amores, A., Arroyo-Resino, D., y Sánchez-Munilla, M. (2022). Self-concept and academic achievement in primary school: A predictive study. *International Journal of Education in Mathematics, Science, and Technology (IJEMST)*, *10*(4), 1057-1073. https://doi.org/10.46328/ijemst.2303

Comisión Europea. (2016). *Abandono Escolar. Fichas temáticas del semestre europeo*. Consejería de Educación. european-semester_thematic-factsheet_early-school-leavers_es.docx

Crespo Cebada, E. (2018). *Fracaso escolar español: Una comparativa internacional con datos de PISA*. Universidad de Extremadura, Servicio de Publicaciones. http://hdl.handle.net/10662/20543

Cordero, J. M., Crespo, E., y Pedraja, F. (2011). Rendimiento educativo y determinantes según PISA: Una revisión de la literatura en España. *Revista de Educación*, 362, 361-388. http://dx.doi.org/10.4438/1988-592X-RE-2011-362-161

Cutimbo Estrada, P. M. (2008). *Influencia del nivel de capacitación docente en el rendimiento académico de los estudiantes del Instituto Superior Pedagógico Público de Puno: caso de la Especialidad de Educación Primaria* IX Semestre-2008 [Trabajo de fin de Grado, Universidad Nacional Mayor de San Marcos]. https://hdl.handle.net/20.500.12672/2395

Fernández-Lasarte, O., Goñi, E., Camino, I., y Zubeldia, M. (2019). Ajuste escolar y autoconcepto académico en la Educación Secundaria. *Revista de Investigación Educativa*, *37*(1), 163-179. http://dx.doi.org/10.6018/rie.37.1.308651

Franco Montenegro, M. P., Cárdenas Rodríguez, R., y Santrich Sánchez, E. R. (2016). Factores asociados a la comprensión lectora en estudiantes de noveno grado de Barranquilla. *Psicogente*, *19*(36), 296-310. https://doi.org/10.17081/psico.19.36.1299

García, M. A.; Arévalo D., M. A., y Hernández, C. A. (2018). La comprensión lectora y el rendimiento escolar. *Cuadernos de Lingüística Hispánica*, (32), 155-174. https://doi.org/10.19053/0121053X.n32.2018.8126

Garrido Yserte, R., Gallo-Rivera, M. T., y Martínez-Gautier, D. (2020). ¿Cuáles son y cómo operan los determinantes del fracaso escolar? Replanteando las políticas públicas para el caso de España y sus regiones. *Revista Internacional de Ciencias del Estado y de Gobierno*, *1*(4), 509-540.

Gaviria Soto, J. L., y Castro Morera, M. (2005). *Modelos jerárquicos lineales*. La Muralla

Gil-Flores, J., y García-Gómez, S. (2017). *Importancia de la actuación docente frente a la política educativa regional en la explicación del rendimiento en PISA*. Revista de Educación, *378*, 187-210. http://dx.doi.org/10.4438/1988-592x-RE-2017-378-361

González, R. L., y Jackon, C. L. (2014). Engaging with parents: The relationship between school engagement efforts, social class, and learning. *School Effectiveness and School Improvement*, *24*(3), 225–316. http://dx.doi.org/10.1080/09243453.2012.680893

Gorostiaga, A., y 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. https://doi.org/10.1016/j.neucom.2015.07.001

Guerra-García, J., Saldívar-Llanos, A., y Sandria-López, S. (2021). Evaluación de comprensión lectora, uso de estrategias y su relación con variables académicas y sociodemográficas en estudiantes universitarios. *Revista Innova Educación*, *3*(2), 360-373. https://doi.org/10.35622/j.rie.2021.02.005

Guio, J. M., y Choi, A. (2014). Evolución del riesgo de fracaso escolar en España durante la década del 2000: Análisis de los resultados de PISA con un modelo logístico de dos niveles. *Estudios Sobre Educación*, (26), 33-63. https://doi.org/10.15581/004.26.33-62

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* Taylor & Francis. https://doi.org/10.4324/9780203887332

Hu, J., Peng, Y., y Ma, H. (2022). Examining the contextual factors of science effectiveness: a machine learning-based approach. *School Effectiveness and School Improvement*, *33*, 21-50. https://doi.org/10.1080/09243453.2021.1929346

Huamán Herreros, J. A. (2024). *Comprensión lectora y su relación con el rendimiento académico en estudiantes de secundaria*. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo* (RIDE), *15*(29), e743. https://doi.org/10.23913/ride.v15i29.2109

Instituto Nacional de Estadística (2018). *Panorama educativo de México. Indicadores del Sistema Educativo Nacional 2017. Educación básica y media superior*. Instituto nacional para la evaluación de la educación. Pnorama_educativo.pdf

Jiménez-Pérez, E. D., Martínez León, N., y Cuadros Muñoz, R. (2020). La influencia materna en la inteligencia emocional y la competencia lectora de sus hijos. *Revista de Estudios Sobre Lectura*, *19*(1), 80-89. https://doi.org/10.18239/ocnos_2020.19.1.2187

Kassambara, A. (2018). *Machine learning essentials: Practical guide in R*. STHDA.

Laukaityte, I., y Wiberg, M. (2017) Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics - Theory and Methods*, *46*(22), 11341-11357. https://doi.org/10.1080/03610926.2016.1267764

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist*, *35*(2), 125-141. https://doi.org/10.1207/S15326985EP3502_6

Marchesi, A. (2003). *El fracaso escolar en España*. Fundación Alternativas-Documento de Trabajo 11/2003. https://fundacionalternativas.org/wp-content/uploads/2022/07/xmlimport-ZPRlx1.pdf

Martín-Ruiz, I., y González-Valenzuela, M. J. (2022). *Autoconcepto académico, motivación y rendimiento escolar en estudiantes de Educación Secundaria Obligatoria*. *Anales de Psicología*, *38*(2), 251–259. https://dx.doi.org/10.6018/analesps.419111

Ministerio de Educación y Formación Profesional. (2020a). *El sistema estatal de indicadores de la Educación*. Instituto Nacional de Evaluación Educativa. https://www.libreria.educacion.gob.es/libro/sistema-estatal-de-indicadores-de-la-educacion-2020_180617/

Ministerio de Educación y Formación Profesional. (2020b). *PISA 2018. Resultados de lectura en España.* Instituto Nacional de Evaluación Educativa. https://www.libreria.educacion.gob.es/libro/pisa-2018-resultados-de-lectura-en-espana_181801/

Molina Ibarra, C. A. (2020). Comprensión lectora y rendimiento escolar. *Revista Boletín Redipe*, *9*(1), 121-131. https://doi.org/10.36260/rbr.v9i1.900

Murillo, F. J., y Hernández-Castilla, R. (2020). ¿La implicación de las familias influye en el rendimiento? Un estudio en educación primaria en América Latina. *Revista de Psicodidáctica*, *25*(1), 13-22. https://doi.org/10.1016/j.psicod.2019.10.002

Nakagawa, S., y Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2015). *Education Policy Outlook 2015: Making Reforms Happen.* OECD Publishing. http://dx.doi.org/10.1787/9789264225442-en

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2016). *PISA. Low-Performing Students. Why they fall behind and how to help them succeed*. OECD Publishing. https://doi.org/10.1787/9789264250246-en

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2018). *Marco teórico de lectura PISA 2018*. Instituto Nacional de Evaluación Educativa, Ministerio de Educación, Cultura y Deporte, España

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

Organización para la Cooperación y el Desarrollo Económicos (OCDE). (2025). *PISA 2022 Technical Report*. OECD Publishing. https://doi.org/10.1787/01820d6d-en

Ostria Baltazar, C. E., López Padilla M. G., y Valenzuela González, J.R. (2014). Caracterización de las competencias transversales. La competencia lectora: identificación del nivel de logro. *Investigaciones Sobre Lectura*, *2*, 32-43. http://hdl.handle.net/10481/33864

Pardo, M., y Ruiz, M. (2013). *Análisis de Datos en Ciencias Sociales y de la Salud III*. Síntesis.

Raschka, S., y Mirjalili, V. (2019). *Python Machine Learning: aprendizaje automático y aprendizaje profundo con Python. Scikit-Learn y TensorFlow.* Marcombo.

Rouse, C. E., y Krueger, A. B. (2004). Putting computerized instruction to the test: A randomized evaluation of a "scientifically based" reading program. *Economics of Education Review*, *23*(4), 323-338. https://doi.org/10.1016/j.econedurev.2003.10.005

Vélez, E., Schiefelbein, E., y Valenzuela, J. (1994). Factores que afectan el rendimiento académico en la educación primaria: Revisión de la literatura de América Latina y el Caribe. *Revista latinoamericana de Innovaciones Educativas*, *17*, 1-16. https://hdl.handle.net/20.500.12799/4317

Viramontes, E., Amparán, A., y Núñez, L. D. (2019). Comprensión lectora y el rendimiento académico en Educación Primaria. *Investigaciones Sobre Lectura*, *12*, 65-82. http://hdl.handle.net/10481/60379

Date received: 20 February, 2025
Date reviewed: 10 March, 2025
Date accepted: 20 February, 2026