www.psye.com

© Psychology, Society, & Education, 2019. Vol. 11(3), pp. 299-311
ISSN 2171-2085 (print) / ISSN 1989-709X (online)
Doi 10.25115/psye.v10i1.2056

Predicting Academic Performance Based on Students' Family Environment: Evidence for Colombia Using Classification Trees

Juan David GARCÍA-GONZÁLEZ[1] y Anastasija SKRITA[2]

[1]Facultad de Ciencias Económicas y Empresariales, Universidad de Almería (España)
[2]Department of Arts, Latvian Academy of Culture (Latvia)

(Received on October 22, 2018; Accepted on Decembre 25, 2018)

ABSTRACT: Family environment, parental economic and social conditions influence students' academic performance. However, in Colombia, research and methods that have been used to study these variables are rather limited. This research predicts the academic performance of students who took the 2016 State exam for access to higher education (*Saber 11*) based on observations and characteristics of the students' families. The data comes from the database of the Colombian Institute Educational Evaluation (*ICFES*), and classification trees are generated in order to predict academic results. The results show that the family variables that best predict academic results are, respectively, the educational level of the mother, the socioeconomic stratum of the household, the number of books in the home, the educational level of the father and the presence of a computer in the household.

Keywords: *academic outcome; human capital; machine learning; Saber 11.*

*Prediciendo el desempeño académico según el entorno familiar de los estudiantes:
evidencia para Colombia usando árboles de clasificación*

Resumen: El entorno familiar, las condiciones económicas y sociales propias de las familias influyen en el desempeño académico de los estudiantes y, por ende, en los resultados de las pruebas académicas. No obstante, en Colombia es limitada la investigación y los métodos que se han usado en el estudio de estas variables. Esta investigación predice el desempeño académico de los estudiantes que presentaron el examen de Estado de 2016 para acceder a la educación superior (Saber 11) a partir de las observaciones y características familiares propias de los estudiantes. Los datos provienen de la base de datos del Instituto Colombiano para la evaluación de la educación (ICFES) y se realizan árboles de clasificación para predecir los resultados académicos. Los resultados muestran que las variables familiares que mejor predicen los resultados académicos son, en su orden: el nivel educativo de la madre, el estrato socioeconómico de la vivienda, el número de libros, el nivel educativo del padre y el poseer computador en la vivienda.

Palabras clave: *rendimiento académico, capital humano; machine learning; Saber 11.*

Correspondencia: Juan David García González. Facultad de Ciencias Económicas y Empresariales, Universidad de Almería (España), email: judgarciago@unal.edu.co

Introduction

In Colombia, research on family environment and academic results of students has not been sufficiently addressed. As at the international level, there is a tendency to use descriptive statistics, case studies, factor analysis, principal component analysis and regressions, while little use has been made of prediction techniques based on decision trees.

On the other hand, ambitious educational objectives were stipulated within the framework of the National Development Plan 2014-2018 under Law 1753 of 2015; Colombia is expected to be the most educated country in Latin America in 2025 (Ministerio de Educación Nacional, 2015). However, there is still a long way to go before reaching those objectives; especially in terms of equal opportunities and academic results, there are large differences in the average municipal percentiles of the State examination between the central region and other regions of the country.

This article tries to determine, by means of classification trees, what family characteristics best predict the academic performance of students in the *Saber 11* exams are. In addition, it is intended to characterize the family environment according to the results of state exams.

This topic is appealing since extensive debates about academic gaps in Colombia and the characteristics and causes of students' academic performance are being generated from academics. In addition, the subject is important because it can help to not only characterize and hierarchize the variables of the family environment of students with “good and bad" academic results, but it can also be an input for further research, predict academic results, facilitate public policy decisions, and even impute missing data. Furthermore, the use of methods based on decision trees is considered innovative in the educational sector, making it possible to present results that are easy to interpret, robust against atypical data, and that allow non-linear relationships (Assis & Almeida, 2017).

Family Environment and Academic Performance

Until the mid-twentieth century, academic performance was explained by individual characteristics of a student; however, school and family are crucial environments in education (Coleman, 1966). At the international level, PISA tests and reports, new analysis techniques and greater availability of statistics have multiplied the research that tries to compare, explain, and predict the academic performance of students based on their family factors.

Beneyto (2015) and Ruiz de Miguel (2001) divide family factors into two groups: dynamic family characteristics, associated with the family-student relationship, and structural family characteristics that include income level, family structure, cultural resources, and the formation of the parents.

The dynamic family characteristics suggest that income is not the only family factor that influences the academic performance. Some articles determined that the type of parent-child relationship, focus, control, follow-up and high parental expectations for child’s academic achievement affect the academic results (Beneyto, 2015; Gil, 2009; Hernando, Oliva &

Pertegal, 2012; Robledo & García, 2012; Ruiz de Miguel, 2001; Shumow, Vandell & Kang, 1996; Steinberg, Lamborn, Dornbusch & Darling, 1992).

In relation to the structural family characteristics, several studies conclude that family income and its variations throughout the school stage strongly influence the parent-child relationships and student academic performance. Generally, the best results come from students with high-income families, while dropout rates and low scores are associated with students from low-income families (Arranz, et al., 2010; Cavanagh & Fomby, 2012; Duncan & Brooks-Gunn, 1997; Gil, 2011; Rojas, Alemany & Ortiz, 2011; Shumow et al., 1996).

In addition, the availability of minimum resources in the family environment accompanied by the level of preparedness and psychological health of the parents have a considerable impact on the student's learning experiences at home as well as on the quality of parent-child teamwork, which all together are reflected in student academic performance (De Jorge Moreno, 2016, Duncan & Brooks-Gunn, 1997; Gil, 2011; Mensah & Kiernan, 2010; Rojas et al., 2011; Shumow et al., 1996; Steinberg et al., 1992; Yeung, Linver & Brooks-Gunn, 2002). Within these minimum resources, it is worth highlighting books and access to communication and information tools; there is sufficient evidence to claim that access and use of these tools facilitates learning and improves academic results (De Ferranti & Perry, 2003; Hall & Maffioli, 2008).

Other structural family variables such as race, number of siblings or parents' marital status are important to take into consideration when observing variations in academic performance (Rojas et al., 2011; Shumow et al., 1996).

In the Colombian case, the inherent characteristics of the student as well as their socio-economic characteristics and family context are directly related to academic achievement (García, Espinosa, Jiménez & Parra, 2013; Ramírez & Teichler, 2011; Sarmiento & Silva, 2013). For example, the educational level of parents seems to affect significantly student academic performance (Gaviria & Barrietos, 2001).

To sum up, in Colombia there is a lack of research that relates family environment and academic performance. Furthermore, on a national scale, the tree-based prediction techniques have never been applied to determine the importance and the level of prediction of family environment variables respect to student academic performance.

Therefore, this research aims to contribute to the discussion in two ways: on the one hand, it intends to determine the variables of family environment that can predict academic performance of students using classification trees and, on the other hand, it aims to characterize the family environment according to performance of students in *Saber 11* State exams 2016.

Saber 11 State Exam

The Colombian Institute for Educational Evaluation (*ICFES*) is an entity specialized in offering educational evaluation services, including the *Saber 11* state exam. The *Saber 11* State exam is an official test that all senior high school students must take in order to both graduate from high school and to get access higher education.

The exam evaluates five fields of knowledge: 1. Math, 2. Critical reading, 3. Social and civic competences 4. Natural science, and 5. English. The results show both absolute and relative scores in each of the five fields of knowledge and, from these five scores, the total score that represents the student's final qualification is derived.

Data

The data comes from the *ICFES* databases. Specifically, the data of *Saber 11* is taken with the cross section of 2016. Table 1 shows the selected variables that are associated with academic performance and family environment. The last ones were chosen and classified as proposed by (Beneyto, 2015; Ruiz de Miguel, 2001).

Table 1. *Variables of academic performance and family environment*

VARIABLES ASSOCIATED TO ACADEMIC PERFORMANCE	
1. Global percentile	
2. Percentile in Critical Reading	
3. Percentile in Math	
4. Percentile in Natural Science	
5. Percentile in Social and Civic Competences	
6. Percentile in English	
VARIABLES ASSOCIATED TO FAMILY ENVIRONMENT	
Socioeconomic dimension	**Parents' education dimension**
1. Residence area (rural or urban)	1. Father's level of education
2. Number of rooms	2. Mother's level of education
3. Stratum of the household	3. Mother's occupation
4. Monthly income	4. Father's occupation
5. Predominant material of floors	
6. Goods (landline phone, washing machine, microwave, oven, and car)	
Cultural resources dimension	**Family structure dimension**
1. Quantity of books	1. Number of siblings
2. Goods (DVD, internet, television services, and computer)	2. Number of people residing in the household

Source: own elaboration based on (Beneyto, 2015; Ruiz de Miguel, 2001).

With the aim of generating a "clean" database, the cross sections of *Saber 11* of 2016 I and 2016 II are added. Subsequently, the database is filtered to only display the last-year "student" individuals (eleventh grade). Finally, individuals who do not have complete data and who are difficult to impute are removed since they generally belong to indigenous or Afro-descendant minorities. In total, a matrix of 582,282 individuals per 30 variables (6 dependent and 24 independent) is the main input of the investigation, Table 1.

Method

Classification trees are used in order to predict academic results based on the student's family characteristics. Classification trees are made for the six variables associated with academic performance, table 1.

According to Reche (2017), a classification tree T represents a recursive partition of the sample space X based on a set of prototypes S. Methods based on classification trees allow stratifying or segmenting the predictor space into a number of simple regions in order to make predictions based on observations and characteristics of individuals or groups (Breiman, Friedman, Stone & Olshen, 1984; James, Witten, Hastie & Tibshirani, 2015).

The classification trees are similar to the regression trees, except the fact that the former are used to predict qualitative data while the latter predict quantitative values; both of them are hierarchical classification methods. Classification trees predict based on the class or value that repeats the most in the prediction space.

Given a node t, the partition that increases the homogeneity of the resulting groups will be chosen. In order to determine the size of the classification tree and the best division of the predictor space, impurity measures are used as a criterion. These measures try to estimate the accuracy of prediction in the resulting groups (Breiman et al., 1984; James et al., 2015; Reche, 2017).

Let p(j|t) be the probability that a case of node t is of class j, where:

$$p(j|t) = \frac{N_j(t)}{N(t)}$$

Some of the impurity measures are: GINI Index: Measures the "diversity of classes in a node".

$$i(t) = 1 - \sum_{j=1}^{J} p(j|t)^2$$

Entropy measure: Measures the "disorder" within a node.

$$i(t) = -\sum_{j=1}^{J} p(j|t) log p(j|t)$$

The advantages of using classification trees are diverse. It is worth pointing out that it is a non-parametric technique that does not imply any model. It can handle non-linear relationships; it cleans variables by itself; it allows working with missing data; it is robust against anomalous data, it works simultaneously with all kinds of variables, it is easy to interpret and allows hierarchizing and identifying the most important variables of the model (Reche, 2017).

On the other hand, it requires a large number of data, which is not a problem in case of this research, but it must be acknowledged that it is difficult to choose the optimal tree and, unlike traditional regression methods, it generates a rule-based model, not a global function of the variables (Reche, 2017).

Classification trees do not represent causal relationships and can repeatedly be a simplification of the true relationship between dependent and independent variables, thus generating low accuracy levels of prediction.

The free R software is used to analyze the data, thanks to the fact that the programming, lecture notes and texts published by experts in the subject facilitate the work and interpretation (James et al., 2015; 2009; Reche, 2017).

Results

Descriptive analysis of predictor variables

The classification trees determine that the family variables that best predict academic results are the educational level of the mother, the socioeconomic stratum of the household, the number of books, the educational level of the father, and the presence of a computer. Table 2 shows the relative frequency of each of the five significant variables.

Table 2. *Relative frequency of the significant variables*

VARIABLE 1 y 2	**edu.dad**	**edu.mom**
Description	**Educational level of the father**	**Educational level of the mother**
1. None	5,08%	2,29%
2. Incomplete primary	17,96%	14,99%
3. Complete primary	15,63%	15,21%
4. Incomplete secondary	13,42%	15,28%
5. Complete secondary	24,63%	28,35%
6. Incomplete Technical or Technological	1,59%	2,07%
7. Complete Technical or Technological	5,77%	7,89%
8. Incomplete professional	1,19%	1,42%
9. Complete professional	8,30%	8,96%
10. Postgraduate	2,11%	2,08%
11. Doesn't know	4,32%	1,46%
VARIABLE 3	**soc.eco**	
Description	**Socioeconomic stratum of the household**	
1. Low-low	44,40%	
2. Low	33,80%	
3. Medium-low	16,10%	
4. Medium	3,60%	
5. Medium-high	1,40%	
6. High	0,80%	
VARIABLE 4	**Books**	
Description	**Number of books in the home**	
0 to 10	46,32%	
11 to 25	29,02%	
26 to 100	18,78%	
More than 100	5,87%	
VARIABLE 5	**Pc**	
Description	**Computer in the household**	
Yes	58,91%	
No	44,40%	

Source: own elaboration based on ICFES database.

In general, the women have higher educational levels than men do, while the two groups coincide in that the secondary and primary educational levels are the most common. Around

45% of the students who took the *Saber 11* exam in 2016 are from stratum 1, the lowest socioeconomic level.

Finally, the average number of books in the home is low, while around 40% of Colombian students still do not have access to computer equipment.

Classification trees

The main result of the classification trees is that of the 24 initial variables associated with the family environment, five are the most relevant when it comes to predicting the academic performance of students and characterizing their family environment. These family variables are, respectively, the educational level of the mother, the socioeconomic stratum of the home, the number of books in the household, the educational level of the father and the presence of a computer in the household.

The variables demonstrate the expected effect: higher educational levels of the parents, better socioeconomic stratum, greater number of books and owning a computer in the household have a positive impact on student academic performance, table 3.

Table 3. *Significant family environment variables and their impact on academic performance*

VARIABLE	PREDICTS SCORE (TREES)	IMPACT	EXPLANATION
Education of the mother (edu.mom)	Global, English, Social and civic competences, Natural science, Math, and Critical reading	Positive	Higher educational level of the mother predicts better academic results
Socioeconomic stratum of housing (soc.eco)	Global, English, Social and civic competences, Natural science, Math, and Critical reading	Positive	Higher socioeconomic stratum predicts better academic results
Number of books in the home (books)	Global, English, Social and civic competences, Natural science, Math, and Critical reading	Positive	Greater number of books in the household predicts better academic results
Education of the father (edu.dad)	Global, Natural science, and Critical reading	Positive	Higher educational level of the father predicts better academic results
Personal computer in the household (pc)	Math	Positive	Presence of a computer in the household predicts better academic results

When it comes to predicting academic results, the most important variable is the educational level of the mother, this variable is the root node for all the dependent variables. In fact, this variable reappears again in lower nodes. Thus, the educational level of the mother is the variable of family environment that best divides the predictor space. Higher educational levels of the mother are associated with a better academic performance at a global level and within each of the five knowledge areas, it evidently happens for reasons of causality rather than chance. The trees predict that higher educational levels of the mother (technical, technological, university or postgraduate) are associated with high academic results of their children, while the mode is belonging to the highest decile of academic performance (between 90% and 100%).

In addition, although the educational level of the father is not as important as the educational level of the mother, it is true that higher educational levels of the father are related to better academic results. It is reflected in the fact that this variable appears in the trees that predict the global score, natural science and critical reading.

These findings are consistent with the evidence that the educational level of parents has a considerable influence on student academic performance (Beneyto, 2015; Gaviria & Barrietos, 2001). However, for the Colombian case, it is clear that the mother's educational level has a greater importance than the father's educational level in order to determine the academic performance of a student.

The second most important variable is the stratum or socioeconomic level of the household. It is located on the left side of the classification trees complementing the profile of low educational levels of the mother. It is noteworthy that the socioeconomic level always separates the students of the stratum 1 from the rest of the students and, on all occasions, the model predicts that these students of the lowest socioeconomic level (stratum 1) present low academic results, while the mode is belonging to the lowest decile of academic performance (between 0% and 10%). Clearly, poor academic performance is associated with low-income, low-stratum families in a situation of poverty and with limited access to goods and services.

The third most important variable is the number of books in the student's household. In all cases, the model predicts that having less than 25 or 10 books will lead to low academic results, while the mode is belonging to the lowest decile of academic performance (between 0% and 10%).

In addition, it is worth noting that owning computer equipment is one of the variables that predicts academic results in the area of mathematics. The profile comprised of the low educational level of the mother and not having a computer in the household exhibits low academic results, while the mode is belonging to the lowest decile of academic performance (between 0% and 10%).

The results are consistent with the research demonstrating that access and use of cultural resources (books or communication and information tools) facilitates learning and improves academic results (De Ferranti & Perry, 2003; Duncan & Brooks-Gunn, 1997). In Colombia, there are still great difficulties to access these goods, which are necessary and complement the learning of the school environment.

The classification trees presented below predict the deciles of the following variables: global score, score in natural science and score in math. As for the three remaining score trees in English, critical reading and social and civic competences, it has been decided to not include, given that they have great similarities with the classification tree that predicts the deciles of the global score.

To sum up, the top three most important family environment variables when predicting academic performance in Colombia are the educational level of the mother, the socioeconomic stratum of the housing and the number of books in the household.

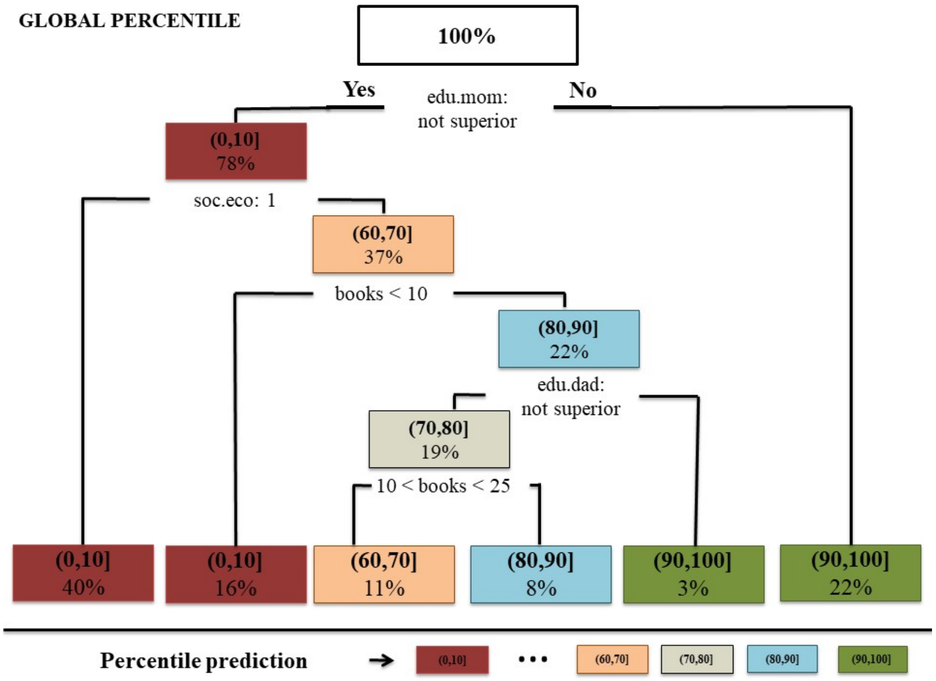

Figure 1. Classification tree for the deciles of the global score

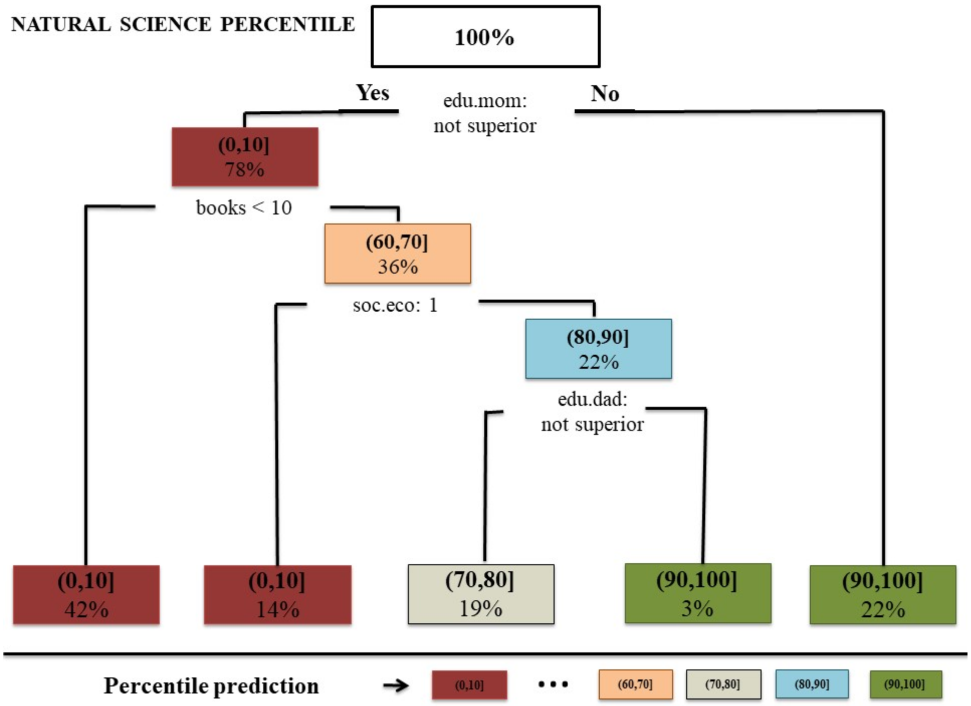

Figure 2. Classification tree for the deciles of the score in Natural Science.

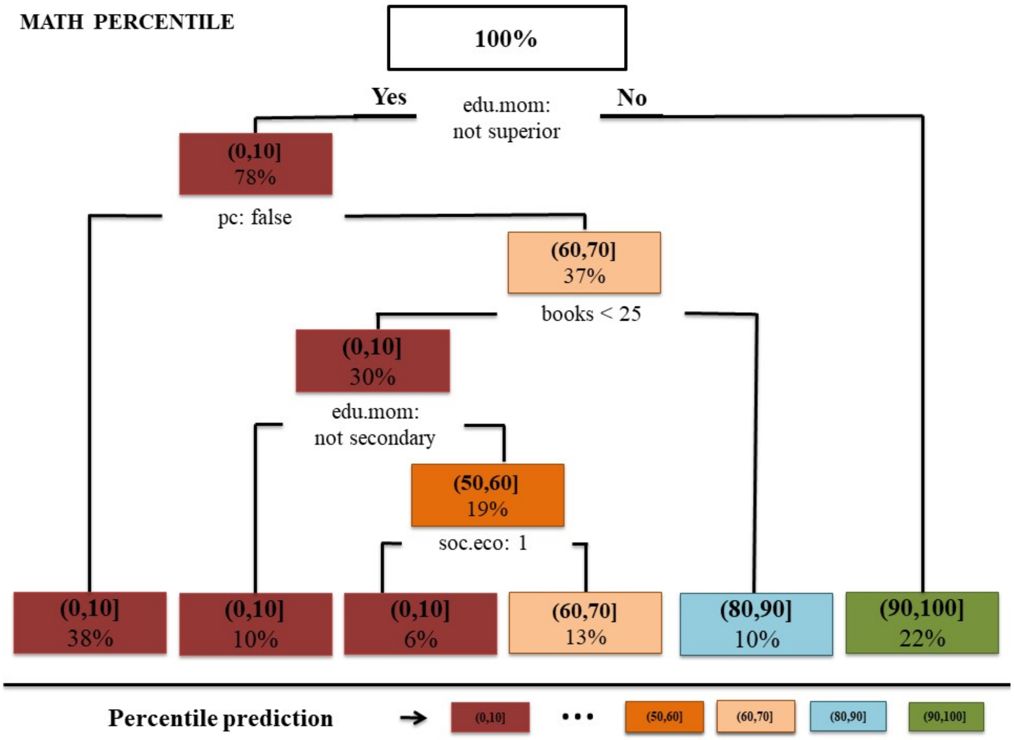

Figure 3. Classification tree for the deciles of the score in mathematics.

Conclusion

Family environment strongly influences academic performance of students. However, in Colombia, the relationship between family environment and academic results has not been sufficiently addressed, limiting public action, decision-making at both macro and micro levels, and leaving many questions about the causes of academic performance in Colombia.

This research contributes in two ways. On a theoretical level, it is innovative since it makes use of classification trees. On a practical level, the results facilitate decision-making, characterize the family environment based on academic performance, predict academic results and even impute missing data.

Classification trees turned out to be an useful theoretical tool to segment and hierarchize the variables since they accept non-linear relationships, operate with all kinds of variables and are easy to interpret, among other advantages (Assis & Almeida, 2017).

Apart from that, it is a valuable contribution to the academic debate since this research finds that the education of the mother, the socioeconomic stratum and the number of books in the home determine and predict largely the global academic results as well as the results within each area of knowledge. The education of the father and the access to computer are important factors according to the field of knowledge.

On a practical level, the results draw attention to the long road Colombia has to travel before reaching the goal of having a quality and egalitarian educational system. The differences between center-periphery persist and the exclusion of the population of low socio-economic stratum in terms of access to cultural and technological goods and to opportunities for quality study and social mobility is still notorious.

It is imperative to join both private and public efforts in order to provide educational opportunities to students belonging to the lower strata access to cultural, technological and quality academic training. The difficulties related to the amount and quality of time that parents dedicate to their children for academic activities should be minimized with the help of programs that provide personal support, as well as by generating mechanisms and spaces to increase the social capital of students in vulnerable situations.

Disclosure statement: We thank to Fernando Reche, Jaime de Pablo Valenciano and José Santiago Gómez for their valuable contribution to this work.

References

Arranz, E., Oliva, A., De Miguel, M., Olabarrieta, F., & Richards, M. (2010). Quality of family context and cognitive development: Across sectional and longitudinal study. *Journal of Family Studies, 16 (2),* 130-142.

Assis, C.M., & Almeida, L.S. (2017). *Advocating the Broad Use of the Decision Tree Method in Education. Practical Assessment, Research & Evaluaiton, 22*, 10.

Beneyto, S. (2015). *Entorno familiar y rendimiento escolar*. Logroño: Universidad de la Rioja

Breiman, L., Friedman, J., Olshen, R.A., & Stone, C.J. (1984). Classification and decision trees. *Wadsworth, Belmont, 378.*

Cavanagh, S. E., & Fomby, P. (2012). Family instability, school context, and the academic careers of adolescents. *Sociology of Education, 85(1),* 81-97.

Coleman, J. (1966). *Equality of educational opportunity Study*. Washington DC: US Department of Health, and Welfare, Office of Education.

De Ferranti, D.M., & Perry, G. E. (2003). *Closing the Gap in Education and Technology.* Washington DC: World Bank Publications.

De Jorge Moreno, J. (2016). Factores explicativos del rendimiento escolar en Latinoamérica con datos PISA 2009. *Revista de Métodos Cuantitativos Para La Economía Y La Empresa, (22)*, 216–229.

Duncan, G.J., & Brooks-Gunn, J. (1997). *The Effects of Poverty on Children. The Future of Children, 7*(2), 55–71. doi: https://doi.org/10.2307/1602387

García, M., Espinosa, J. R., Jiménez, F., & Parra, J. D. (2013). *Separados y desiguales. Educación y clases sociales en Colombia*. Bogotá DC: Colección de Justicia, Ediciones Antropos.

Gaviria, A., & Barrietos, J.H. (2001). *Determinantes de la calidad de la educación en Colombia.* Bogotá DC: fedesarrollo.

Gil, J. (2009). Family habits and attitudes towards reading and students' basic competences. *Revista de Educación, (350)*, 301–322.

Gil, J. (2011). Estatus socioeconómico de las familias y resultados educativos logrados por el alumnado. *Cultura y Educación, 23 (1)*, 141-154. doi: 10.1174%2F113564011794728597.

Hall, B., & Maffioli, A. (2008). Evaluating the impact of technology development funds in emerging economies: evidence from Latin America. *The European Journal of Development Research, 20(2),* 172–198.

Hernando, Á., Oliva, A., & Pertegal, M.-Á. (2012). Variables familiares y rendimiento académico en la adolescencia. *Estudios de Psicología, 33(1),* 51–65. doi: https://doi.org/10.1174/021093912799803791

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning with applications in R*. Performance Evaluation. New York: springer.

Mensah, F. K., & Kiernan, K. E. (2010). Gender differences in educational attainment: influences of the family environment. *British Educational Research Journal, 36(2),* 239–260. doi: https://doi.org/10.1080/01411920902802198

Ministerio de Educación Nacional. (2015). *Colombia, la mejor educada en el 2025*. Bogotá DC: MinEducación.

Ramírez, C., & Teichler, U. (2011). *Factores socioeconómicos y educativos asociados con el desempeño académico según nivel de formación y género de los estudiantes que presentaron la prueba SABER PRO 2009*. Bogotá: ICFES.

Reche, F. (2017). *Árboles de clasificación*. Almería: Universidad de Almería.

Robledo, P. & García, J. N. (2012). Implicación parental en la educación del alumnado de diferentes edades y sexos. *International Journal of Development and Educational Psychology (INFAD) Revista de Psicología, 1,* (2) 371-380.

Rojas, G., Alemany, I., & Ortiz, M. (2011). Influencia de los factores familiares en el abandono escolar temprano. Estudio de un contexto multicultural. *Electronic Journal of Research in Educational Psychology, 9*(3), 1377–1402.

Ruiz de Miguel, C. (2001). Factores familiares vinculados al bajo rendimiento. *Revista Complutense de Educación, 12*(1), 81–113.

Shumow, L., Vandell, D.L., & Kang, K. (1996). School choice, family characteristics, and home-school relations: Contributors to school achievement? *Journal of Educational Psychology, 88*(3), 451–460. doi: https://doi.org/10.1037/0022-0663.88.3.451

Sarmiento, J., & Silva, A. (2013). *Desigualdad de oportunidades en el logro educativo en Colombia: evolución del desempeño en las pruebas SABER 11 y SABER PRO*.

Steinberg, L., Lamborn, S.D., Dornbusch, S.M., & Darling, N. (1992). Impact of Parenting Practices on Adolescent Achievement: Authoritative Parenting, School Involvement, and Encouragement to Succeed. *Child Development, 63*(5), 1266–1281. doi: https://doi.org/10.1111/j.1467-8624.1992.tb01694.x

Yeung, W.J., Linver, M.R., & Brooks-Gunn, J. (2002). How Money Matters for Young Children's Development: Parental Investment and Family Processes. *Child Developement, 73*(6), 1861–1879. doi: https://doi.org/10.2307/3696422