Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/jptcp.v30i6.2472

PREDICTIVE MODEL OF SCORES IN STANDARDIZED HIGH SCHOOL TESTS FOR BOGOTÁ, COLOMBIA.

Oscar Jardey Suarez[1]*, Edier Hernán Bustos Velazco[2], Jaime Duván Reyes Roncancio[3]

[1]*Professor Universidad de Nariño, Email: ojsuarez@udenar.edu.co,
https://orcid.org/0000-0001-8780-595X
[2]Professor Universidad Distrital Francisco José de Caldas, Email:
ehbustosv@udistrital.edu.co,https://orcid.org/0000-0003-0072-8598
[3]Professor Universidad Distrital Francisco José de Caldas, Email: jdreyesr@udistrital.edu.co,
https://orcid.org/0000-0002-9229-1196

***Corresponding Author: -** Oscar Jardey Suarez
*Professor Universidad de Nariño, Email: ojsuarez@udenar.edu.co, https://orcid.org/0000-0001-8780-595X

Abstract

Scores on standardized tests are part of the measure of educational quality. The objective is to construct a predictive model, applying linear regression and multilevel linear regression, for the scores of the Saber 11 tests, using statistically significant factors collected from surveys administered to test takers between 2019 and 2022 in Bogotá, Colombia. The methodological approach is quantitative. The information organization procedure follows the Standard Process for Data Mining across Industries. The modeling utilized the open database of the Colombian Institute for the Promotion of Higher Education, consisting of 129,087 records (over 81% of the total). Women account for 53% of the records in the study. The Inter Class Correlation in the multilevel linear regression model among the 20 localities is 1.54. The Mean Absolute Percentage Error in the linear regression model is 11.79% for the entire dataset and 11.81% when using 70% of the data for training. The statistically significant factors include gender, socio-economic status, resources for studying at home, nutrition, student employment, and the institution. In conclusion, the possession and access to technological resources, hardware, and software, as well as the urban location of the institution, had a positive impact on scores during the COVID-19 pandemic, providing empirical evidence of a wider educational gap in populations with limited technological access or residing in rural areas.

Bogotá is the capital city of Colombia, politically organized into 20 localities, with a population exceeding 10 million people and a population density of around 5,000 individuals per square kilometer. The locality of Sumapaz has the largest rural area with the lowest population density.

Standardized tests are part of some educational systems, and their results contribute to assessing the quality of education and its continuity within the education system. In Colombia, the name of these tests are ¨Saber tests¨, while in Chile, they are known as Pruebas de Acceso a la Educación Superior (PAES), and in Ecuador, they are call to as SER (Sistema de Evaluación y Rendición de la Educación). Brazil conducts similar tests through its state educational evaluation system, among others. At the international level, standardized tests compare educational systems, such as the Programme for International Student Assessment (PISA) administered to countries belonging to the Organisation for Economic Co-operation and Development (OECD).

Theoretical Foundations

Various studies indicate that gender is one of the variables that influence scores in standardized tests, with males often obtaining higher scores (Beckham et al., 2023; Ekubo & Esiefarienrhe, 2022; Maisarah-Samsudin et al., 2021; Posada & Mendoza, 2014; Qazdar et al., 2019; Ramírez & Teichler, 2014). The economic situation of students and their families is a significant factor affecting test results in numerous studies (Atlam et al., 2022; Contreras Bravo et al., 2022; Galster et al., 2016; Ibourk & Amaghouss, 2016; Martínez-Mateus, 2015; Masci et al., 2018; Murillo & Carrillo, 2021; Orjuela, 2014; Posada & Mendoza, 2014; Qiu & Wu, 2019; Romero, 2009). Other factors are associated with the family environment (Galster et al., 2016; Lisboa- Bartholo & Da-Costa, 2016; Orjuela, 2014; Salal & Abdullaev, 2020), access to information and communication technologies (Atlam et al., 2022; Kumari et al., 2018; Posada & Mendoza, 2014; Wandera et al., 2019), study habits (Beckham et al., 2023; Giannakas et al., 2021), institutional characteristics (Baashar et al., 2022; Contreras Bravo et al., 2022; Masci et al., 2018; Qiu & Wu, 2019; Rebai et al., 2019; Romero, 2009), and student motivation (Atlam et al., 2022; Masci et al., 2018; Rebai et al., 2019; Shah et al., 2019).

Investigating the factors that predict scores in the Saber 11 standardized tests can contribute to understanding the student population in the city of Bogotá. This knowledge has the potential to inform policy decisions and the design of plans and programs aimed at improving test results. It is important to note that these tests are requirements for accessing higher education institutions, whether public or private, making them crucial for accessing the higher education system.

This research aims to answer the following question: "What statistically significant factors contribute to a predictive model of Saber 11 test scores for students in Bogotá between 2019 and 2022?"

Academic achievement is often related to students' efforts or scores in various tests or evaluations within the educational setting. However, understanding academic performance or learning achievement involves studying human beings' different dimensions or factors when interacting with specific areas or the education system as a whole. That is a complex and evolving concept (Navarro, 2003; Parra-Castillo et al., 2021), which even includes aspects such as nutrition (Gimeno-Tena & Esteve-Clavero, 2021; Santos- Holguín & Barros Rivera, 2022).

Methodology

The research approach is quantitative, employing statistical modeling techniques such as linear regression and multilevel linear regression, which are suitable for the study (Murillo-Torrecilla, 2008).

The methodological process includes the following stages: understanding the object of study, comprehending the data, data preparation, data modeling, model evaluation, and conclusion (Chapman et al., 2000).

Understanding the object of study involves establishing the research objective, consolidating previous research studies, and selecting the appropriate data model or models along with relevant tools.

In the data comprehension stage, we obtain data from the *Instituto Colombiano para la Evaluación de la Educación* (ICFES) and its data dictionary. It is considering its structure in the review process.

During data preparation, we selected data related to Bogotá from 2019, 2020, 2021, and 2022. Also, we did Queries to ensure data consistency and address missing data by making appropriate adjustments or predicting and eliminating missing data.

For data modeling, the Multilevel Linear Regression Model (MRLM) and the Linear Regression Model (LRM) have been selected, with parameters corresponding to those in the ICFES database data dictionary. Then we run the model using 70% of randomly selected data on a computer with an Intel Core i5 processor, 16 gigabytes of RAM, a solid-state drive, and an integrated graphics card.

Regarding model evaluation and conclusion, we evaluated the obtained model using the remaining 30% of the data. Subsequently, we tested it using 70% of the data, and its predictive capability was verified.

Data Dictionary

According to the public information provided by ICFES, the database and data dictionary of the Saber 11 tests are used, valid for 2019 to 2021, as shown in Table 1.

Table 1. Data Dictionary of Saber 11 Standardized Tests.

Variable	Description	Response Options
ESTU_GENERO	Gender	F - Female, M - Male
FAMI_ESTRATOVIVIENDA	Refers to the socioeconomic stratum of the place of residence according to the electricity bill.	Stratum 1, Stratum 2, Stratum 3, Stratum 4, Stratum 5, Stratum 6, No Stratum.
FAMI_PERSONASHOGAR	Inquires about the number of people, including the examinee, living in the household.	1 to 4, 5 to 8, 9 or more.
FAMI_CUARTOSHOGAR	Inquires about the number of rooms where people sleep in the examinee's household.	1 to 3, 4 to 5, 6 or more.
FAMI_EDUCACIONPADRE	Father's highest level of education.	(1) Complete Primary Education, (2) Incomplete Primary Education, (3) Complete Secondary Education (High School), (4) Incomplete Secondary Education (High School), (5) Complete Higher Education, (6) Incomplete Higher Education, (7) Complete Technical or Technological Education, (8) Incomplete Technical or Technological Education, (9) Postgraduate, (10) None.
FAMI_EDUCACIONMADRE	Mother's highest level of education.	(1) Complete Primary Education, (2) Incomplete Primary Education, (3) Complete Secondary Education (High School), (4) Incomplete Secondary Education (High School), (5) Complete Higher

Variable	Description	Response Options
		Education, (6) Incomplete Higher Education, (7) Complete Technical or Technological Education, (8) Incomplete Technical or Technological Education, (9) Postgraduate, (10) None.
FAMI_TRABAJOLABORPADRE	Refers to the father's primary occupation during the last year.	(1) Small business owner (has few or no employees, e.g., shop, stationery store, etc.), (2) Works as a professional (e.g., doctor, lawyer, engineer), (3) Machine operator or driver (taxi driver, chauffeur), (4) Large business owner, holds a managerial or executive position, (5) Retired, (6) Homemaker, unemployed, or studying, (7) Salesperson or works in customer service, (8) Self-employed (e.g., plumber, electrician), (9) Farmer, fisherman, or laborer, (10) Has an administrative support job (e.g., secretary or assistant), (11) Works in cleaning, maintenance, security or construction.
FAMI_TRABAJOLABORMADRE	Refers to the mother's primary occupation during the last year.	(1) Small business owner (has few or no employees, e.g., shop, stationery store, etc.), (2) Works as a professional (e.g., doctor, lawyer, engineer), (3) Machine operator or driver (taxi driver, chauffeur), (4) Large business owner, holds a managerial or executive position, (5) Retired, (6) Homemaker, unemployed, or studying, (7) Salesperson or works in customer service, (8) Self-employed (e.g., plumber, electrician), (9) Farmer, fisherman, or laborer, (10) Has an administrative support job (e.g., secretary or assistant), (11) Works in cleaning, maintenance, security or construction.

Variable	**Description**	**Response Options**
FAMI_TIENEINTERNET	Inquires if the household has internet service.	Yes, No.
FAMI_TIENECOMPUTADOR	Inquires if the household has a computer.	Yes, No.
FAMI_NUMLIBROS	Establishes the number of physical or electronic books present in the place, excluding publications such as newspapers, magazines, phone directories, or books from the educational institution.	0 to 10 BOOKS, 11 to 25 BOOKS, 26 to 100 BOOKS, MORE THAN 100 BOOKS.
FAMI_COMELECHEDERIVADOS	Asks how many times milk or dairy products (cheese, yogurt, etc.) are consumed in the household.	1 or 2 times a week, 3 to 5 times a week, Rarely or never eat that, Almost every day.
FAMI_COMECARNEPESCADOHUEVO	Asks how many times proteins (meat [chicken, turkey, beef, lamb, pork, rabbit, etc.], fish, or eggs) are consumed in the household.	1 or 2 times a week, 3 to 5 times a week, Rarely or never eat that, Almost every day.
FAMI_COMECEREALFRUTOSLEGUMBRE	Asks how many times cereals (oats, granola), nuts (almonds, peanuts), or legumes (beans, chickpeas, lentils) are consumed in the household.	1 or 2 times a week, 3 to 5 times a week, Rarely or never eat that, Almost every day.
FAMI_SITUACIONECONOMICA	Inquires about the economic situation of the household in the immediately preceding year.	Same, Better, Worse.
ESTU_DEDICACIONLECTURADIARIA	Asks how much time is dedicated to reading for entertainment per day.	Do not read for entertainment, 30 minutes or less, Between 30 and 60 minutes, Between 1 and 2 hours, More than 2 hours.
ESTU_DEDICACIONINTERNET	Asks how much time is spent on the Internet for non-academic activities per day.	Do not browse the Internet, 30 minutes or less, Between 30 and 60 minutes, Between 1 and 3 hours, More than 3 hours.
ESTU_HORASSEMANATRABAJA	Asks how many hours the examinee worked during the	0, Less than 10 hours, Between 11 and 20 hours,

Variable	Description	Response Options
	week before the interview.	Between 21 and 30 hours, More than 30 hours.
VARIABLES RELATED TO THE EDUCATIONAL INSTITUTION		
COLE_CALENDARIO	Refers to the academic calendar of the educational institution to which the examinee belongs.	A, B, OTHER.
COLE_AREA_UBICACION	Refers to the geographical location of the educational institution's headquarters.	RURAL, URBAN.
COLE_JORNADA	Specifies the educational schedule in which the examinee is studying at the educational institution.	SINGLE, MORNING, NIGHT, SATURDAY, AFTERNOON

Source: Open database of ICFES (https://www.icfes.gov.co/publicación-de-datos-abiertos1).

Data

The open database of ICFES for the years 2019 (23,441), 2020 (19,846), 2021 (44,060), and 2022 (42,460) consists of 160,148 records. After cleaning and organizing the database for the study, a total of 129,087 records remained, representing over 81% of the total. 53% of the records correspond to females, and 47% to males.

Information Processing - Linear Regression Modeling (Non-significant Multilevel)

To organize the information, we used Microsoft Access® and Excel® software and R® for running the modeling. The libraries used for the MRLM and MRL algorithms were lme4 and stats, with the R-4.3.0 version for Windows®.

We calculated the Intraclass Correlation (ICC) to identify the agreement between the level variable, which in this case is the georeferencing of students based on their locality (Donner & Koval, 1980). The Mean Absolute Percentage Error (MAPE) is used to assess the predictive model of Saber 11 test scores (De-Myttenaere et al., 2016). After running the MRLM model and based on the ICC, we evaluated the need to run the MRL model.

Results and Discussion

After processing the information with the MRLM, we found that the score variation between localities is 17.13. The Interclass Correlation (ICC), which measures the agreement due to localities, is 1.54. Therefore, the score differences do not correspond to the locality of the test takers. Consequently, we run the linear regression model, which identifies 21 factors.

Linear regression models were run separately for 2019, 2020, 2021, and 2022. We created a combined database for four years and constructed a global linear regression model. For the linear regression model, the MAPE is 11.79% for the training dataset (70%) and 11.81% for the remaining 30% test dataset.

The variables are described below, grouped according to their affinity.

Gender

In this category, the scores are statistically significant and higher for males than females in each of the yearly models: 2019 (+9.26), 2020 (+8.55), 2021 (+8.61), 2022 (+9.41), and the global model (+9). These results confirm previous studies conducted at different educational levels (Beckham et al., 2023; Ekubo & Esiefarienrhe, 2022; Maisarah-Samsudin et al., 2021; Posada & Mendoza, 2014; Qazdar et al., 2019), including higher education, where a positive Pearson correlation has been identified (Ramírez & Teichler, 2014).

Socio-economic Status

Socio-economic stratification is a classification system for residential properties that determines the allocation of public services and differential billing based on strata. This classification allows for the allocation of subsidies and contributions according to the economic capacity of each stratum (Alcaldía de Bogotá, 2021). In this way, those with higher economic resources pay more for public services, which helps lower-strata households cover their bills (Departamento Nacional de Planeación, 2014). The main objective of socio-economic stratification is to ensure equitable access to essential public services such as water, electricity, and natural gas in all residential areas. Establishing different rates based on socio-economic strata aims to reduce inequalities and support low-income households facing difficulties in covering the costs of public services (Congreso de Colombia, 1994). This stratification approach is legally grounded in Colombia and supported by regulations such as Law 142 of 1994, which establishes the regulatory framework for domiciliary public services, and National Decree 298 of 2014, which regulates socio-economic stratification in the country (Congreso de Colombia, 1994; Departamento Nacional de Planeación, 2014).

In summary, socio-economic stratification in Colombia is a system used to differentially allocate and charge for domiciliary public services to ensure equitable access to these services and support low-income households. This classification system allows those with higher economic capacity to pay more for services, enabling lower strata to cover their bills.

In the social and economic aspects of the test-taker population, we grouped some variables present in the data dictionary: socio-economic stratum, number of people in the household, education level of both father and mother, and employment status.

Stratum 1 references all variable values in different models (Table 2). Stratum 5 shows the highest score gain in the global model (+13.38), followed by strata 6 (+11.61) and 4 (+10.77). In the 2019 model, the score of stratum 2 is not significant.

Table 2. Estimated Behavior of the Housing Stratum Variable.

Variable	**2019**	**2020**	**2021**	**2022**	**Global**
FAMI_ESTRATOVIVIENDAEstrato 2	1,55*	2,70	2,96	1,67*	2,33
FAMI_ESTRATOVIVIENDAEstrato 3	4,50	4,11	4,62	2,30	3,75
FAMI_ESTRATOVIVIENDAEstrato 4	10,77	11,72	13,08	8,31	10,92
FAMI_ESTRATOVIVIENDAEstrato 5	14,13	15,44	14,84	10,39	13,38
FAMI_ESTRATOVIVIENDAEstrato 6	11,61	10,11	12,96	4,61	9,53
FAMI_ESTRATOVIVIENDASin Estrato	-6,01*	-9,99	-6,22	-15,41	-10,01

These values are not significant in the corresponding linear regression model.
Source: Own elaboration with R.

In the global model, the number of people in the household and the number of rooms where people sleep in the household, which are statistically significant, negatively affect the test-takers' scores when there are more than five people in the household or when there are more than four rooms in the household (Table 3).

Table 3. Number of people in the household and rooms where people sleep in the test-takers household.

Variable	2019	2020	2021	2022	Global
FAMI_PERSONASHOGAR5 a 8	-2,12	-1,52	-0,68*	-0,73*	-1,07
FAMI_PERSONASHOGAR9 o mas	-3,02*	-6,08	0,13*	-2,92*	-2,59
FAMI_CUARTOSHOGAR4 a 5	-4,77	-5,13	-7,52	-6,55	-6,41
FAMI_CUARTOSHOGAR6 o mas	-8,17	-6,83	-11,97	-15,33	-11,28

These values are not significant in the corresponding linear regression model.
Source: Own elaboration with R.

Regarding the education level of the father (Table 4), the highest scores in the different yearly models and the global model were for those test-takers whose fathers have a Postgraduate degree, with the highest increase observed in 2019, prior to the two years of confinement in Colombia due to the COVID-19 pandemic. We found the same pattern when the father works as an administrative or professional assistant. In 2019 and 2020, the fact that the father is self-employed in occupations such as plumber or electrician is not significant, while in 2021, it has a negative impact (-3.24), ultimately resulting in a negative effect in the global model (-1.90).

Table 4. Values of the Father's Education variable.

Variable	2019	2020	2021	2022	Global
FAMI_EDUCACIONPADRE2	-1,31*	0,26*	-1,65*	-0,05*	-0,74*
FAMI_EDUCACIONPADRE3	0,77*	2,52*	1,71*	1,68*	1,73
FAMI_EDUCACIONPADRE4	1,60*	0,65*	0,43*	-0,04*	0,45*
FAMI_EDUCACIONPADRE5	12,85	10,62	10,96	11,66	11,48
FAMI_EDUCACIONPADRE6	12,17	10,51	13,50	13,07	12,75
FAMI_EDUCACIONPADRE7	8,25	9,92	9,14	10,87	9,75
FAMI_EDUCACIONPADRE8	5,50	6,03	2,93*	5,87	4,87
FAMI_EDUCACIONPADRE9	21,79	18,52	22,51	21,31	21,42
FAMI_EDUCACIONPADRE10	-9,62	-2,24*	-7,42	2,36*	-4,07

These values are not significant in the corresponding linear regression model.
Source: Own elaboration with R.

Table 5 shows in the global model that the mother's education, particularly postgraduate education, contributes the most to the prediction of test results, followed by complete professional education, incomplete professional education, and complete technical or technological education. The contribution to the scores from the father and the mother is similar in the global model, with a combined contribution exceeding 57 points on the test. In 2020, the model indicated a slightly lower contribution of around 55 points. In the global model, parents' contribution due to their postgraduate education exceeds 44 points.

Table 5. Values of the Mother's Education variable.

Variable	2019	2020	2021	2022	Global
FAMI_EDUCACIONMADRE2	-1,09*	-0,39*	0,76*	-1,69*	-0,60*
FAMI_EDUCACIONMADRE3	3,86	3,86	3,20	3,01	3,38
FAMI_EDUCACIONMADRE4	1,73*	0,90*	0,30*	-1,16*	0,15*
FAMI_EDUCACIONMADRE5	16,17	12,90	14,65	12,88	14,07
FAMI_EDUCACIONMADRE6	14,27	11,63	12,75	11,47	12,39
FAMI_EDUCACIONMADRE7	12,76	10,01	11,00	10,18	10,85
FAMI_EDUCACIONMADRE8	6,83	6,60	6,05	6,37	6,44
FAMI_EDUCACIONMADRE9	24,93	21,74	22,93	23,81	23,37
FAMI_EDUCACIONMADRE10	-1,85*	-7,67	-4,36*	-9,11	-5,95

These values are not significant in the corresponding linear regression model.
Source: Own elaboration with R.

Previous studies indicate that parental education has a direct impact or association (Ramírez & Teichler, 2014) with students' academic performance in exams (Masci et al., 2018; Posada & Mendoza, 2014), perhaps due to their ways of interacting with their children or the pressure they exert on them (Rebai et al., 2019). This study indicates that when both parents have completed their professional training, the examinees' scores concerning educational parent's level are statistically significant.

Table 6. Values of the Father's Occupation variable.

Variable	2019	2020	2021	2022	Global
FAMI_TRABAJOLABORPADRE2	-2,50*	0,88*	-3,65	-2,17	-2,18
FAMI_TRABAJOLABORPADRE3	-2,38*	-1,91*	-3,68	-4,51	-3,42
FAMI_TRABAJOLABORPADRE4	-6,50	-1,79*	-6,40	-6,56	-5,68
FAMI_TRABAJOLABORPADRE5	-3,87*	-1,73*	-1,36*	-3,21	-2,38
FAMI_TRABAJOLABORPADRE6	-3,65*	0,12*	-3,71	-3,02	-2,77
FAMI_TRABAJOLABORPADRE7	-5,02	-2,91	-7,24	-6,85	-6,03
FAMI_TRABAJOLABORPADRE8	-1,70*	0,34*	-3,24	-1,86	-1,90
FAMI_TRABAJOLABORPADRE9	-5,25	-2,17*	-6,20	-6,82	-5,47
FAMI_TRABAJOLABORPADRE10	-7,26	-6,07	-9,51	-9,03	-8,43
FAMI_TRABAJOLABORPADRE11	-4,59	-3,55	-4,80	-4,63	-4,41

These values are not significant in the corresponding multilevel linear regression model.
Source: Own elaboration with R.

The fact that the father works harms the test taker's scores (Table 6). The activities that have the most negative impact in the global model are when the father works as an administrative assistant (-8.43), followed by when the father works as a salesperson or in customer service (-6.03), or when the father is a business owner, holds a managerial or executive position (-5.68). In the MRL model for 2019 and 2020, most of the father's occupations are not statistically significant, which coincides with the lockdown period due to the COVID-19 pandemic.

Table 7. Values of the Mother's Occupation variable by year.

Variable	2019	2020	2021	2022	Global
FAMI_TRABAJOLABORMADRE2	-5,43	1,36*	-2,93	-2,47	-2,37
FAMI_TRABAJOLABORMADRE3	-3,06*	3,73*	-2,08*	-2,51*	-1,32*
FAMI_TRABAJOLABORMADRE4	-4,57	-6,21	-4,32	-6,63	-5,29
FAMI_TRABAJOLABORMADRE5	-5,18*	4,53*	-6,04	-4,64*	-3,40
FAMI_TRABAJOLABORMADRE6	-1,63*	3,08	1,03*	0,24*	0,80*
FAMI_TRABAJOLABORMADRE7	-3,81	-2,42*	-2,22	-3,15	-2,68
FAMI_TRABAJOLABORMADRE8	1,28*	0,39*	3,43	1,31*	2,03
FAMI_TRABAJOLABORMADRE9	0,57*	-0,11*	-4,94*	-2,93*	-2,44*
FAMI_TRABAJOLABORMADRE10	-6,65	-1,76*	-4,03	-4,79	-4,30
FAMI_TRABAJOLABORMADRE11	-5,72	-4,92	-6,17	-6,05	-5,79

These values are not significant in the corresponding multilevel linear regression model.
Source: Own elaboration with R

On the other hand, unlike the father's occupations, the mother's occupations (Table 7) negatively impact the test-takers scores in the global model. The occupations that have the most negative impact are cleaning, maintenance, security, and construction (-5.79), followed by being a business owner or holding a managerial or executive position (-5.29). Notice that when the mother is self-employed, it positively contributes to the test-takers score (+2.03).

Table 8. Values of the Socio-economic Status variable, with the reference value being equal.

Variable	2019	2020	2021	2022	Global
FAMI_SITUACIONECONOMICAMejor	-7,64	-4,24	-6,26	-7,55	-6,79
FAMI_SITUACIONECONOMICAPeor	9,10	7,11	6,73	8,26	7,48

Source: Own elaboration with R

Compared to the previous year, the perception of the test-takers household economic situation (Table 8) appears to affect the scores in the different models when there is variation. When the test-taker perceives an improvement in the economic situation, their score decreases, while if the situation worsens, the score increases.

With the above, the fact that parents have certain occupations, such as the father being a business owner or holding a managerial position, owning a small business, operating machinery or driving a vehicle, or working as an assistant, among others, and the mother being a business owner, holding a managerial position, operating machinery or driving a vehicle, working in cleaning, maintenance, security, or construction, or being retired, negatively affects the scores obtained by students in the global model, except for the mother's self-employment. The parents' occupations impact a better economic situation for the test-takers household, negatively contributing to the prediction of the Saber 11 test results. The results coincide partially with the findings of Posada Ramos and Mendoza Martínez (Posada & Mendoza, 2014), who had warned in their study that parental occupation reduces test takers' academic achievement.

The family environment becomes a predictive variable for the test-takers scores in the Saber 11 tests, which aligns with other studies (Ekubo & Esiefarienrhe, 2022; Lisboa- Bartholo & Da-Costa, 2016; Orjuela, 2014; Posada & Mendoza, 2014; Ramírez & Teichler, 2014; Salal & Abdullaev, 2020; Wandera et al., 2019), where the presence of parents and the family environment contribute to the academic performance or achievement of students.

Resources for studying at home.

In the global model that predicts the scores in the Saber 11 tests, variables related to the resources available for studying at home, such as the internet, computer, and the presence of physical or electronic books, positively contribute to the scores.

Having internet access ("FAMI_TIENEINTERNETSi") at home for 2019, 2020, and 2022 was not statistically significant, which coincides with some years of greater confinement due to the COVID-19 pandemic. However, for the year 2021, having internet access at home contributes to the scores (+3.81), and in the global model, also there is a positive contribution to the prediction (+1.84)

When the test-taker has a computer at home "FAMI_TIENECOMPUTADORSi," it significantly contributes to the student's scores in the models for the different years: 2019 (+2.41), 2020 (+3.72), 2021 (+7.92), 2022 (+5.98), as well as the global model (+5.78).

The presence and possibility of interacting with books at home (Table 9) is a factor that positively contributes to the test-takers scores in the Saber 11 tests. There is a positive and incremental contribution in the predictive models from 2019 to 2022 and the global model as more books are available for interaction.

Table 9. Values of the Book variable, with a reference value of 0 to 10 books, 30 minutes or less for recreational reading, and regarding internet browsing dedication, the reference is no browsing or browsing less than 30 minutes.

Variable	**2019**	**2020**	**2021**	**2022**	**Global**
FAMI_NUMLIBROS11 A 25 LIBROS	5,94	6,57	3,93	4,31	4,70
FAMI_NUMLIBROS26 A 100 LIBROS	13,23	13,62	11,62	11,22	11,89
FAMI_NUMLIBROSMAS DE 100 LIBROS	15,51	17,95	13,30	13,36	14,36
ESTU_DEDICACIONLECTURADIARIAEntre 1 y 2 horas	7,09	8,77	9,59	7,32	8,40
ESTU_DEDICACIONLECTURADIARIAEntre 30 y 60 minutos	7,00	6,16	7,68	6,66	7,04
ESTU_DEDICACIONLECTURADIARIAMas de 2 horas	13,58	10,15	13,04	11,06	12,09
ESTU_DEDICACIONLECTURADIARIANo leo por entretenimiento	-4,63	-5,20	-4,93	-4,94	-5,01
ESTU_DEDICACIONINTERNETEntre 1 y 3 horas	9,79	10,74	11,19	11,75	11,16
ESTU_DEDICACIONINTERNETEntre 30 y 60 minutos	4,89	6,17	5,09	4,74	5,14
ESTU_DEDICACIONINTERNETMas de 3 horas	7,25	9,06	8,28	9,78	8,77

Source. Self-made with R.

Study habits contribute to academic achievement measured in the Saber 11 tests, which is consistent with other studies that have identified the association between study habits and students' academic performance (Beckham et al., 2023; Ekubo & Esiefarienrhe, 2022; Giannakas et al., 2021).

Moderate Internet browsing in activities other than academic ones contributes to students' scores with formative elements. This empirical result provides a basis for further research, as some studies have delved into the potential risk of excessive internet use and addiction among young people (Peris et al., 2018; Suárez et al., 2022).

Nutrition and Scores

Some previous studies have investigated the relationship between academic performance and nutrition. The global model, as well as the models for the years 2019, 2020, 2021, and 2022, indicate that not consuming or rarely consuming proteins derived from fish, eggs, or meat (like chicken, turkey, beef, lamb, pork, rabbit) does not have statistically significant significance. On the other hand, excluding "milk and its derivatives" and "cereals, fruits, or legumes" from the diet appears to have a statistically significant impact on the predictive scores of the Saber 11 tests, according to the 2021 models and the global model.

Table 10. Values of the variables FAMI_COMELECHEDERIVADOS (consumes milk and its derivatives), FAMI_COMECARNEPESCADOHUEVO (consumes meat, fish, and eggs), and FAMI_COMECEREALFRUTOSLEGUMBRE (consumes cereals, fruits, or legumes).

Variable	**2019**	**2020**	**2021**	**2022**	**Global**
FAMI_COMELECHEDERIVADOS3 - 5 times per week	6,79	5,49	5,86	4,75	5,48
FAMI_COMELECHEDERIVADOSNunca – or Rarely or seldom eat that	5,16	-3,89	-5,62	-3,77	-3,39
FAMI_COMELECHEDERIVADOSTodos or Almost every day	12,13	8,95	8,62	7,30	8,65
FAMI_COMECARNEPESCADOHUEVO3 -5 times per week	2,61	3,53	2,69	0,70*	2,11
FAMI_COMECARNEPESCADOHUEVONunca - Rarely or seldom eat that	-3,19*	-0,20*	0,74*	-1,47*	-0,77*
FAMI_COMECARNEPESCADOHUEVOTodos or Almost every day	5,94	3,74	4,72	2,80	4,04
FAMI_COMECEREALFRUTOSLEGUMBRE3 - 5 times per week	3,50	3,83	3,19	4,28	3,74
FAMI_COMECEREALFRUTOSLEGUMBRENunca or Rarely or seldom eat that	-0,89*	-1,26*	-4,94	-5,62	-3,95
FAMI_COMECEREALFRUTOSLEGUMBRETodos or Almost every day	0,06*	1,50*	0,49*	1,99	1,12

These values are not significant in the corresponding linear regression model.
Source: Self-made with R.

The influence of dietary habits and the possibility of having a diet rich in dairy products, proteins, fruits, and vegetables impacts academic achievement. This finding is consistent with various studies that highlight how a healthy diet can contribute to better brain function at different educational levels (Gimeno-Tena & Esteve-Clavero, 2021; Parra-Castillo et al., 2021; Santos-Holguín & Barros-Rivera, 2022; Taras, 2005; Woodhouse & Lamport, 2012), across different subjects, such as mathematics and language (Mora et al., 2019).

Further research is needed to understand the relationship between nutrition and academic achievement in different geographical areas. The new research will help inform social policies and actions that promote equitable education, closely aligned with the United Nations Sustainable Development Goals of "Zero Hunger" and "Quality Education."

Student Employment

Students are sometimes required to work due to their families' economic situation. The predictive linear models of learning achievement for 2019, 2020, 2021, and 2022 and the global model indicate that the time spent working is a statistically significant variable with a negative effect (Table 10). This finding holds even during the period of the COVID-19 pandemic.

Table 11. Values of the variable ESTU_HORASSEMANATRABAJA (student's weekly working hours), with the reference value being "does not work."

Variable	2019	2020	2021	2022	Global
ESTU_HORASSEMANATRABAJA	-13,31	-12,57	-13,28	-12,54	-12,89
ESTU_HORASSEMANATRABAJA Between 11 and 20 hours.	-10,50	-9,49	-8,80	-12,82	-10,44
ESTU_HORASSEMANATRABAJAMore tan 30 hours	-6,98	-11,53	-11,76	-11,28	-10,75
ESTU_HORASSEMANATRABAJALess tan 20 hours	-10,66	-8,41	-12,21	-12,01	-11,36

Source: Self-made with R.

The current regulations in Colombia govern the work of minors. According to Pedraza-Avella & Ribero-Medina (2006), work among adolescent youth hurts their schooling, increasing grade repetition rates and negatively affecting their health. From a social and political perspective, the implications of child and youth labor on educational quality should be considered, with an understanding that child and youth labor, as defined by the International Labour Organization, refers to activities that do not compromise their health, personal development, or educational attainment.

The Institution

Regarding the characteristics of the educational offerings provided by the institution and its location, the B-type schedule obtains higher scores than the A-type schedule in the predictive linear regression models for scores in 2019 (+26.31), 2020 (+18.84), 2021 (+23.98), 2022 (+18.06), and the overall model (+21.77). Institutions classified as schedules other than A or B only show a statistically significant difference during 2019 (-11.09) and 2022 (-7.22) models. The COVID-19 pandemic does not show variations in scores across different models, except for those institutions that do not fall under the A or B schedule.

Urban institutions may contribute higher scores than rural institutions. This situation is evident in the linear model for the years 2019 (+9.99), 2020 (+16.55), 2021 (+13.79), 2022 (+8.79), and the overall model (+11.53). The years with the most negative impact on scores for rural areas coincide with the COVID-19 pandemic. It is worth noting that during confinement, due to the pandemic, interactions (student-student, teacher-teacher, student-teacher) were limited for those in rural areas due to deficiencies in the means utilized.

The Colombian education system, addressing coverage issues since the 1960s, has established three educational sessions: morning (30 Hours per Week [HPW]), afternoon (30 HPW), and evening (17.5 HPW). Additionally, as mentioned above, a Saturday session is available for a specific population whose daily schedule does not allow session attendance. Since 2017, Colombian regulations have included the concept of a unified session across different educational levels (preschool [25 HPW], primary [30 HPW], secondary, and high school [35 HPW]), allowing students to spend more time in the educational institution to engage in various learning activities (Bocanegra-Acosta & Huertas-Bustos, 2018).

According to the predictive linear regression model of the Saber 11 tests, students in the afternoon session do not show statistically significant differences compared to students in the morning session (Table 12). However, students in the evening or Saturday sessions are negatively affected. Students in public sessions positively impact scores, with a +15.31 effect in the overall model.

Table 12. Values of the variable COLE_JORNADA (school session), with the reference being COMPLETE or MORNING.

Variable	2019	2020	2021	2022	Global
COLE_JORNADANOCHE	-20,62	-18,03	-15,22	-15,73	-16,92
COLE_JORNADASABATINA	-16,35	-20,35	-16,91	-16,09	-16,90
COLE_JORNADATARDE	1,19*	-0,83*	-1,81	1,02*	-0,22*
COLE_JORNADAUNICA	15,91	11,75	14,95	16,95	15,31

Source: Self-made with R.

The above findings indicate that the time students dedicate to their education during the school session impacts the results of the Saber 11 tests. Based on the findings presented here, the educational policy of transitioning to a unified session anticipates improved scores in standardized tests.

Conclusions

The current study provides empirical evidence that the COVID-19 pandemic has hurt rural communities with limited access to information and communication technologies. These findings suggest that the educational gap has widened among disadvantaged population groups.

Students from socio-economic strata 1 and 2, whose parents have jobs with specific schedules, exhibit lower results in the Saber 11 tests than those whose parents are also employed but belong to higher socio-economic strata. This situation is worst when students from these lower socio-economic strata must work to support their families. Therefore, students from lower socio-economic strata, whose parents have jobs with specific schedules and who also need to work to support their families, face multiple challenges that have long-term implications for their academic and socio-economic development.

Firstly, the combination of work and academic responsibilities can negatively affect the academic performance of these students. Balancing work and studies may result in less time and energy dedicated to learning, as can be reviewed in lower scores in the Saber 11 tests. In this sense, they can limit their opportunities for higher education and secure better-paying jobs.

Additionally, the need to work to support families from lower socio-economic strata can lead to reduced availability of economic resources to invest in the education of these students. The situation results in a lack of access to educational materials, extracurricular programs, and tutoring, negatively impacting their academic development and ability to compete equally with students from higher socio-economic strata.

These difficulties can perpetuate socio-economic inequality over time. Suppose students from lower socio-economic strata fail to achieve strong academic outcomes and access better educational and employment opportunities. In that case, they will likely remain trapped in poverty and socio-economic disadvantage, with limited prospects for upward social mobility.

Finally, the long-term implications of low scores in the Saber 11 tests for the lower socio-economic strata population are worst when these students also have to work to support their families. It is crucial to address these structural inequalities and provide additional support to these students in terms of educational resources and policies that promote more favorable working conditions to break the cycle of socio-economic disadvantage and open development opportunities for this population.

References

1. Atlam, E. S., Ewis, A., El-Raouf, M. M. A., Ghoneim, O., & Gad, I. (2022). A new approach to identifying the psychological impact of COVID-19 on university student’s academic performance. Alexandria Engineering Journal, 61(7), 5223–5233. https://doi.org/10.1016/j.aej.2021.10.046

2. Alcaldía de Bogotá. (2021). Estratificación socioeconómica. Recuperado de https://www.bogota.gov.co/sisjur/normas/Norma1.jsp?i=94978
3. Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., Alsariera, Y. A., Ali, A. Q., Hashim, W., & Tiong, S. K. (2022). Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs). Applied Sciences, 12(3). https://doi.org/10.3390/app12031289
4. Beckham, N. R., Akeh, L. J., Mitaart, G. N. P., & Moniaga, J. V. (2023). Determining factors that affect student performance using various machine learning methods. Procedia Computer Science, 216, 597–603. https://doi.org/10.1016/j.procs.2022.12.174
5. Bocanegra-Acosta, H., & Huertas-Bustos, A. P. (2018). La política de jornada única escolar: los referentes y la experiencia de una Institución Educativa Distrital. Revista Republicana, 25, 199–240. https://doi.org/https://doi.org/10.21017/rev.repub.2018.v25.a56
6. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.O step-by-step data mining guide. NCR System Engineering Copenhagen, DaimlerChrysler AG, SPSS and Verzekeringen en Bank Groep B.V. http://www.crisp-dm.org/CRISPWP-0800.pdf
7. Congreso de Colombia. (1994). Ley 142 de 1994. Por la cual se establece el régimen de los servicios públicos domiciliarios y se dictan otras disposiciones. Diario Oficial No. 41.257, de 11 de julio de 1994
8. Contreras Bravo, L. E., Nieves-Pimiento, N., & Gonzalez-Guerrero, K. (2022). Prediction of University-Level Academic Performance through Machine Learning Mechanisms and Supervised Methods. Ingeniería, 28(1). https://doi.org/10.14483/23448393.19514
9. De-Myttenaere, A., Golden, B., Le-Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. Neurocomputing, 192, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114
10. Departamento Nacional de Planeación. (2014). Decreto Nacional 298 de 2014. Por el cual se reglamenta la estratificación socioeconómica. Diario Oficial No. 49.172, de 23 de diciembre de 2014.
11. Donner, A., & Koval, J. J. (1980). The Estimation of Intraclass Correlation in the Analysis of Family Data. Biometrics, 36(1), 19. https://doi.org/10.2307/2530491
12. Ekubo, E. A., & Esiefarienrhe, B. M. (2022). Using machine learning to predict low academic performance at a Nigerian university. The African Journal of Information and Communication (AJIC), 30. https://doi.org/10.23962/ajic.i30.14839
13. Galster, G., Santiago, A., Stack, L., & Cutsinger, J. (2016). Neighborhood effects on secondary school performance of Latino and African American youth: Evidence from a natural experiment in Denver. Journal of Urban Economics, pp. 93, 30–48. https://doi.org/10.1016/j.jue.2016.02.004
14. Giannakas, F., Troussas, C., Voyiatzis, I., & Sgouropoulou, C. (2021). A deep learning classification framework for early prediction of team-based academic performance. Applied Soft Computing, 106. https://doi.org/10.1016/j.asoc.2021.107355
15. Gimeno-Tena, A., & Esteve-Clavero, A. (2021). Relación entre los hábitos saludables y el rendimiento académico en los estudiantes de la Universitat Jaume I. Nutricion Clinica y Dietética Hospitalaria, 41(2), 99–106. https://doi.org/https://doi.org/10.12873/412gimeno
16. Ibourk, A., & Amaghouss, J. (2016). Convergence éducative et déterminants socioéconomiques: Analyse spatiale sur des données marocaines. Mondes en Developpement, 176(4), 93–116. https://doi.org/10.3917/med.176.0093
17. Kumari, P., Jain, P., & Pamula, R. (2018). An Efficient use of Ensemble Methods to Predict Students Academic Performance. 4th Int’l Conf. on Recent Advances in Information Technology. https://doi.org/10.1109/RAIT.2018.8389056
18. Lisboa- Bartholo, T., & Da-Costa, M. (2016). Evidence of a school composition effect in Rio de Janeiro public schools. Ensaio, 24(92), 498–521. https://doi.org/10.1590/S0104-40362016000300001
19. Maisarah-Samsudin, N., Milleana-Shaharudin, S., Filza-Sulaiman, N., Mohd-Fuad, M., Fareezuan-Zulfikri, M., & Hila-Zainuddin, N. (2021). Modeling student’s academic performance

during Covid-19 based on classification in support vector machine. Turkish Journal of Computer and Mathematics Education, 12(5), 1798–1804. https://doi.org/10.17762/turcomat.v12i5.2190

20. Martínez-Mateus, W. (2015). Análisis de distribución geográfica y espacial de los resultadosde las Pruebas Saber 11 del Instituto Colombiano para el Fomento de la Educación Superior -ICFES-. 2005-2012. Colombia. Cuadernos Latinoamericanos de Administración, 11(21), 39–50. https://doi.org/10.18270/cuaderlam.v11i21.1618
21. Masci, C., Johnes, G., & Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. European Journal of Operational Research, 269(3), 1072–1085. https://doi.org/10.1016/j.ejor.2018.02.031
22. Mora, J. I., Mosqueira, C. M. H., & Ventura-Vall-Llovera, C. (2019). Hábitos alimentarios y rendimiento académico en escolares adolescentes de Chile. Revista Española de Nutricion Humana y Dietética, 23(4), 292–301. https://doi.org/https://doi.org/10.14306/renhyd.23.4.804
23. Murillo-Torrecilla, J. (2008). Los modelos multinivel como herramienta para la investigación educativa. Magis Revista Internacional de Investigación en Educación, 1(1), 45–62. https://www.redalyc.org/pdf/2810/281021687004.pdf
24. Murillo, J., & Carrillo, S. (2021). Incidencia de la Segregación Escolar por Nivel Socioeconómico en el Rendimiento Académico. Un Estudio desde Perú. Archivos analíticos de políticas educativas, 29(49), 3–11. https://doi.org/10.14507/epaa.29.5129
25. Navarro, R. E. (2003). EL rendimiento académico: concepto, investigación y desarrollo. REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 1(2), 1–15. https://doi.org/2152
26. Orjuela, J. (2014). Análisis del Desempeño Estudiantil en las Pruebas de Estado para Educación Media en Colombia mediante Modelos Jerárquicos Lineales. Ingeniería, 18(2). https://doi.org/10.14483/udistrital.jour.reving.2013.2.a04
27. Parra-Castillo, A., Morales-Canedo, L. M., & Medina-Valencia, M. M. (2021). Relación entre los hábitos alimentarios y el rendimiento académico en estudiantes de universidades públicas y privadas de la localidad de Chapinero, Bogotá. Perspectivas en Nutrición Humana, 23(2), 183–195. https://doi.org/https://doi.org/10.17533/udea.penh.v23n2a05
28. Pedraza-Avella, A. C., & Ribero-Medina, R. (2006). El trabajo infantil y juvenil en Colombia y algunas de sus consecuencias claves. Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 4(1), 7. https://dialnet.unirioja.es/servlet/articulo?codigo=4657561&info=resumen&idioma=SPA
29. Peris, M., Maganto, C., & Garaigordobil, M. (2018). Escala de riesgo de adicción-adolescente a las redes sociales e internet: fiabilidad y validez (ERA-RSI). Revista de psicología Clínica con Niños y Adolescentes, 5(2), 30–36. https://doi.org/10.21134/rpcna.2018.05.2.4
30. Posada, J., & Mendoza, F. (2014). Determinantes del logro académico de los estudiantes de grado 11 en el periodo 2008 – 2010 . Una perspectiva de género y región. Universidad del Valle, 1–48.
31. Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. Education and Information Technologies, 24(6), 3577–3589. https://doi.org/10.1007/s10639-019-09946-8
32. Qiu, X., & Wu, S. sheng. (2019). Contextual variables of student math proficiency and their geographic variations in Missouri. Applied Geography, 109, 102040. https://doi.org/10.1016/j.apgeog.2019.102040
33. Ramírez, C. E., & Teichler, T. U. (2014). 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. 26.
34. Rebai, S., Ben Yahia, F., & Essid, H. (2019). A graphically based machine learning approach to predict secondary schools performance in Tunisia. Socio-Economic Planning Sciences, 70(August 2018), 100724. https://doi.org/10.1016/j.seps.2019.06.009

35. Romero, C. (2009). Eficacia aprendizaje y de instituciones saber 2009 -1 Análisis de la eficacia de aprendizaje y eficacia de las instituciones educativas mediante el uso de los datos de la Prueba Censal SAB. SaberInvestigar.
36. Salal, Y., & Abdullaev, S. (2020). Deep learning based Ensemble Approach to Predict Student Academic Performance: Case Study. En 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). https://doi.org/10.1109/ICISS49785.2020.9316044
37. Santos-Holguín, S. A., & Barros-Rivera, S. E. (2022). Influencia del Estado Nutricional en el Rendimiento Académico en una institución educativa. Revista De investigación en salud Vive, 5(13), 154–169. https://doi.org/https://doi.org/10.33996/revistavive.v5i13.138
38. Santos Holguín, S. A., & Barros Rivera, S. E. (2022). Influencia del Estado Nutricional en el Rendimiento Académico en una institución educativa. Revista De investigación en salud Vive, 5(13), 154–169. https://doi.org/https://doi.org/10.33996/revistavive.v5i13.138
39. Shah, M., Kaistha, M., & Gupta, Y. (2019). Student Performance Assessment and Prediction System using Machine Learning. 4th International Conference on Information Systems and Computer Networks, ISCON 2019, 386–390. https://doi.org/10.1109/ISCON47742.2019.9036250
40. Suárez, O., Urbina-Cárdenas, J., & Suárez-Riveros, D. (2022). Factores de Riesgo en Jóvenes Escolarizados Asociados al Uso de las Redes Sociales y la Internet. revista Perspectivas, 7(1), 87–97. https://revistas.ufps.edu.co/index.php/perspectivas/article/view/3392
41. Taras, H. (2005). Sleep and student performance at school. Journal of School Health, 75(6), 199.
42. Wandera, H., Marivate, V., & Sengeh, M. (2019). Predicting national school performance for policy making in South Africa. 6th International Conference on Soft Computing and Machine Intelligence, ISCMI 2019, 23–28. https://doi.org/10.1109/ISCMI47871.2019.9004323
43. Woodhouse, A., & Lamport, M. (2012). The Relationship of Food and Academic Performance: A Preliminary Examination of the Factors of Nutritional Neuroscience, Malnutrition, and Diet Adequacy. Christian Perspectives in Education, 5(1), 1.