Comunicaciones en Estadística
Diciembre 2025, Vol. 18, No. 2, pp. 33–42

Data Science in Secondary Education: Exploring Correlations and Predicting Saber 11 Test Results from the Formative Process

Ciencia de Datos en Educación Media: Explorando Correlaciones y Prediciendo Resultados Saber 11 a Partir del Proceso Formativo

Erick Antonio Quintero Chitiva[a]
eqquintero@unbosque.edu.co

Danny Samuel Martinez Lobo[b]
dsmartinez@unbosque.edu.co

Resumen

Esta investigación explora la relación entre las calificaciones obtenidas por una Institución Educativa Distrital (IED) en diferentes asignaturas y los resultados alcanzados en las áreas evaluadas en las pruebas Saber 11. Se realizó un Análisis de Correlaciones Canónicas (ACC) tomando como variables explicativas las calificaciones obtenidas por los estudiantes en las asignaturas que cursaron en una IED logrando explicar un 25 % de la varianza de los resultados de las cinco áreas que evalúa el Instituto Colombiano para la Evaluación de la Educación (ICFES).

La interpretación de los resultados del ACC reveló que en la IED las asignaturas de Lengua Castellana, Idioma Extranjero y Convivencia tienen una influencia significativa en los resultados de las áreas evaluadas con las pruebas Saber 11, lo que destaca la importancia de las habilidades comunicativas. Adicionalmente, la identificación de relaciones débiles entre los resultados de las pruebas y asignaturas específicas como física, química, matemáticas, entre otras; que desde la práctica pedagógica se esperaría que tuvieran altas magnitudes directas de correlación con el grupo de resultados en las áreas evaluadas por el ICFES, evidenciando una dinámica de multicausalidad entre las asignaturas.

Finalmente, para la predicción del puntaje total en la prueba Saber 11, se construyó un Modelo Lineal Generalizado (MLG) que indicó que las calificaciones de las asignaturas Química, Lengua Castellana y Tecnología e Informática tienen un mayor impacto en los resultados obtenidos por los estudiantes de la IED en dichas pruebas.

***Palabras clave*:** Análisis de Correspondencias Canónicas, Modelos Lineales Generalizados, Instituciones Educativas Distritales, Pruebas Saber 11..

Abstract

This research examines how student grades in various subjects at a District Educational Institution (DEI) relate to their performance on the Saber 11 tests. Using Canonical Correlation Analysis (CCA), the study found that grades in DEI subjects explained 25 % of the variance in results across five key areas evaluated by the Colombian Institute for the Evaluation of Education (ICFES).

The CCA results revealed that Spanish Language, Foreign Language, and coexistence existence subjects strongly influence Saber 11 test outcomes, emphasizing the value of communicative skills. In contrast, the study found only weak correlations between test performance and subjects such as physics, chemistry, and mathematics. This unexpected result suggests a complex, multi-faceted relationship between subject mastery and standardized test scores.

[a] Universidad El Bosque
[b] Universidad El Bosque

Additionally, the research used a Generalized Linear Model (GLM) to predict overall Saber 11 scores. The analysis showed that student performance in Chemistry, Spanish Language, and Technology and Informatics had the greatest impact on total test scores at the DEI.

***Keywords*:** Canonical Correlation Analysis, Generalized Linear Models, Secondary Education, Educational Assessment, Saber 11..

1. Introduction

The District Development Plan 2020-2024 aims to close the educational quality gap between public and private schools through curricular and pedagogical reform (SMECE, 2023). This gap is primarily measured by student performance on the Saber 11 tests, administered semiannually by ICFES. To achieve more equitable outcomes, the Directorate of Education Evaluation, with the Multidimensional System for the Evaluation of Educational Quality (SMECE), has prioritized the strategic use of educational data centralizing, organizing, and analyzing information from educational institutions (SMECE, 2023).

Saber 11 is the key assessment tool for evaluating the quality of Colombia's educational system. It is also a requirement for completing secondary school and accessing higher education. Extensive research has used Saber 11 test results, highlighting that the most influential factors are socio-demographic: parents' education and income, student gender, school type (public or private), cultural participation, and technology access (Chica G. et al., 2010; Montes et al., 2014; Alvarez O., 2015; Timarán P. et al., 2019; Caucali M., 2020; Rodríguez R. et al., 2021; Otálora S. and Torres L., 2022; Sáenz C. and Toro V., 2023).

A growing body of research uses data science techniques, including machine learning models like XGBoost, logistic model trees, and K-Nearest Neighbors, to predict academic performance at various educational levels, both nationally and regionally (Vargas C. and Ardila, 2024; Acosta S. et al., 2021; Aguilera P. et al., 2021; Contreras et al., 2020). These studies also rely heavily on socio-demographic data.

Notably, there is limited research using data from within educational institutions—such as student grades and formative assessments to analyze or predict Saber 11 results at the secondary level. Socio-demographic factors are often beyond the control of schools, while internal data on academic performance can offer actionable insights for improving curriculum and teaching practices. Martínez L. (2013), for example, explored the link between the coursework of math majors and their Saber Pro test results.

Given the centrality of socio-demographics in most analyses and the potential for data-driven decision making within schools, this study proposes applying statistical methods to variables directly related to the school environment—specifically, students' grades and their Saber 11 outcomes. The study seeks to answer: To what extent do grades in various subjects for eleventh-grade students at a District Educational Institution relate to their Saber 11 test results? And, can these metrics be combined in a predictive model of overall test scores?

2. Theoretical Framework

Canonical Correlation Analysis (CCA) and the Generalized Linear Model (GLM) with an identity link function from the Gaussian family are robust multivariate analysis tools. These methods help uncover relationships between variables and model complex phenomena across a range of distributional assumptions.

2.1. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) examines the relationship between two sets of variables: predictors (independent variables, X) and responses (dependent variables, Y). Unlike conducting multiple separate regressions, CCA identifies pairs of linear combinations—called canonical variates—that maximize the correlation between the two groups.

Consider two sets of variables: $X_1, X_2, ..., X_p$ and $Y_1, Y_2, ..., Y_q$. CCA seeks $r = min(p, q)$ pairs of canonical variates, each pair capturing the strongest correlation between the sets while remaining independent from previous pairs. These linear combinations provide a clear, comprehensive view of how the groups of variables interact.

$$\begin{aligned}
U_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
V_1 &= b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q \\
U_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
V_2 &= b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q \\
&\vdots
\end{aligned}$$

Each combination is carefully chosen to maximize correlations:

- U_1 and V_1 are paired for the strongest possible correlation.
- U_2 and V_2 have the next highest correlation, ensuring they remain independent from U_1 and V_1.
- This process continues with each subsequent pair, always selecting the next best correlation while maintaining independence from previous pairs.

This approach ensures that each set delivers the highest possible value without overlap. Canonical Correlation Analysis (CCA) simplifies relationships between two variable sets by creating pairs of canonical variates, each forming an independent equation. The first pair typically holds the strongest correlation, making it most crucial for interpretation.

Key points for interpreting CCA results include:

- Magnitude of Canonical Relationship R^2: Indicates the shared variance between canonical variates in each equation.
- Redundancy Index: Summarizes how well one variable set explains the other.
- Canonical Loadings: Show the direct correlation between each original variable and its corresponding canonical variate.
- Canonical Cross-Loadings: Reflect the correlation of an original variable from one set and the canonical variate of the opposite set.
- For samples of 100–150 observations, loadings of 0.40 or greater are considered significant.

2.2. Generalized Linear Models

Generalized Linear Models (GLMs), introduced by Nelder and Wedderburn in 1972, extend classical linear models to work with a broader range of dependent variable distributions. As described by Dunn and Smyth (2018), GLMs have two main components

- **Random component**: Selects the probability distribution (from the exponential family) that best matches the response variable Y.
- **Systematic component**: Uses a link function, such as $g(\mu) = \eta$ allowing a linear combination of predictors $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$ to best explain the expected value of Y.

For the specific context discussed:

- The response variable Y is normally distributed.
- The identity link function is used: $\mu = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$

Assumptions for linear models include:

- **Normality**: Model residuals should follow a normal distribution.
- **Homoscedasticity**: Residual variance should be constant, with no discernible trend.
- **Independence of residuals**: Residuals should have minimal correlation
- **Linearity**: Predictors and outcome variables should maintain a linear relationship.

2.3. Methodology

This study uses a quantitative, correlational approach to examine the relationship between students' final school grades at a DEI and their results on the ICFES Saber 11 test. We applied both Canonical Correlation Analysis (CCA) and a Generalized Linear Model (GLM) for our analysis.

Our sample included 121 students graduating from a DEI in 2024. For each student, we collected final grades in all core curriculum subjects and scores from each area of the Saber 11 exam.

To assess how academic grades relate to Saber 11 outcomes, we conducted a CCA using R. The process included:

- Installing and running the 'yacca' R package, standardizing variables for comparability.
- Selecting the most representative canonical equation, based on R^2.
- Analyzing and interpreting canonical loadings and cross-loadings.

We then used the Python 'pycaret' library to test various machine learning models for predicting Saber 11 results, including Random Forest Regressor, Extra Trees Regressor, Bayesian Ridge, and linear regression. The GLM was selected for its low root mean square error (RMSE) and straightforward interpretation. The model was constructed in R, following these steps:

- Randomly splitting the 121 observations into training (80 %) and validation (20 %) groups using the R sample command.
- Building the GLM with all independent variables and the identity link function.
- Optimizing the model by iteratively selecting and removing variables, guided by the Akaike Information Criterion (AIC).
- Verifying model assumptions.
- Interpreting the final model.

3. Results

Applying the CCA technique identified five pairs of canonical variates. The table below summarizes their key characteristics.

Tabla 1: Significance test for canonical equations and canonical R-Squared

Canonical Equation	R^2	Chisq	df	Pr(>X)
CV 1	0,497	139,344	65	0,000
CV 2	0,222	63,503	48	0,066
CV 3	0,130	35,827	33	0,337
CV 4	0,111	20,471	20	0,429
CV 5	0,066	75,319	9	0,582

Table 1 highlights that the first pair of canonical variates in the analysis have an R^2 value of 0.497. This strong relationship, reflected by a canonical correlation of 0.71, stands out among the canonical equations, as subsequent pairs show much lower R^2 values. Therefore, the primary insights from the CCA—particularly regarding canonical loadings and cross-loadings are drawn from this first equation. The strength and significance of these findings are further confirmed by robust statistical tests, including Wilk's Lambda, Pillai's Trace, Lawley-Hotelling Trace, and Roy's Maximum Root.

3.1. Canonical Loadings

Table 2 displays the canonical loadings for the first canonical equation in the CCA.

Tabla 2: Canonical Loadings

Subjects	
Physics	-0,167
Chemistry	-0,221
Mathematics	-0,332
Advantage Mathematics	-0,340
Spanish Language	-0,593
Foreign Language: English	-0,707
Advanced English	-0,521
Social History	-0,288
Axiological-Humanistic	-0,274
Arts Education	-0,133
Physical Education, Recreation and Sports	-0,098
Technology and Informatics	0,040
Coexistence	-0,434
Areas evaluated in Saber 11	
Spanish Language Score	-0,519
Mathematics Score	-0,548
Social studies & Citizenship Score	-0,730
Natural Science Score	-0,687
English Score	-0,978

Table 2, highlights that Spanish Language, Foreign Language: English, Advanced English, and Coexistence have the strongest influence on the canonical variate related to DEI taught subjects. Additionally, all areas of the Saber 11 tests significantly impact the corresponding canonical variate, with English and Social Studies and Citizenship showing the most notable effects.

These results show a clear connection between academic performance in communicative skills and achievement on the Saber 11 exam, particularly in English and Social Studies and Citizenship Competencies. This analysis underscores the importance of strong language and social studies skills for success in these standardized assessments.

3.2. Canonical Cross-Loadings

Table 3 presents the canonical cross-loadings for canonical equation 1, as determined by the CCA.

Tabla 3: Cross-Loadings

Institution Subjects	**Canonical Loadings**
Physics	-0.118
Chemistry	-0.156
Mathematics	-0.234
Advanced Mathematics	-0.239
Spanish Language	-0.418
Foreign Language: English	-0.498
Advanced English	-0.367
Social History	-0.203
Axiological-Humanistic	-0.193
Arts Education	-0.094
Physical Education, Recreation and Sports	-0.069
Technology and Informatics	0.028
Convivencia	-0.306

According to the table 3, Spanish Language and Foreign Language: English demonstrate the strongest influence on the first canonical variate among the Saber 11 test areas, based on their loading magnitude.

Furthermore, the redundancy index indicates that the independent variables explain approximately 25 % of the variability in the Saber 11 test results. This means that about a quarter of the test performance can be attributed to the grades in subjects taught by the DEI.

3.3. Generalized Linear Model

The constructed Generalized Linear Model (GLM) produced a root mean square error (RMSE) of 37.512 and an Akaike Information Criterion (AIC) score of 973.12. Initial results indicated that subjects such as physics, mathematics, advanced mathematics, advanced English, axiological-humanistic, arts education, physical education, recreation and sports, and coexistence were not statistically significant predictors.

After optimizing the GLM, the AIC score improved to 962.08. The most statistically significant variables were chemistry, Spanish language, foreign language (English), social history, arts education, and technology and informatics (see Table 4).

Tabla 4: Results of the Generalized Linear Model

Term	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
Intercept	113.947	41.275	2.761	0.00699**
Chemistry	15.204	7.256	2.095	0.03896*
Spanish Language	49.177	15.894	3.094	0.00263**
Foreign Language: English	19.677	9.335	2.108	0.03783*
Social History	21.903	11.803	1.856	0.06675$^{\cdot}$
Arts Education	-17.343	8.612	-2.014	0.04701*
Technology and Informatics	-41.070	9.173	-4.477	2.21e-05***

Signif. codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$
Dispersion parameter for gaussian family: 1086.234
Null deviance: 169046 on 96 degrees of freedom
Residual deviance: 97761 on 90 degrees of freedom
AIC: 962.08

The optimized model used these significant subjects as explanatory variables. For example, increasing the valuation of chemistry by one unit, while holding other variables constant, is associated with an approximate increase of 15 points on the Saber 11 test.

Raising the Spanish Language grade by one unit is linked to an increase of about 49 points on the Saber 11 test, assuming all other subjects remain unchanged. A one-unit improvement in English boosts the Saber 11 score by approximately 20 points under the same conditions.

Similarly, increasing the Social History grade by one unit results in a gain of around 22 points on the Saber 11 test. However, a one-unit rise in Arts Education leads to a decrease of about 17 points, and a similar increase in Technology and Informatics reduces the score by approximately 41 points, provided other grades stay constant.

The model's relative error is just 8 %, offering strong confidence in its predictions for educational decision-making. Additional statistical analyses—including the Durbin-Watson and Breusch-Pagan tests and the calculation of variance inflation factors—confirm there is no autocorrelation, homoscedasticity concerns, or multicollinearity among predictors. Tests of normality further indicate that the residuals follow a normal distribution.

4. Discussion

The study reveals a clear and statistically significant link between students' school grades and their performance on the Saber 11 tests. Our Generalized Linear Model (GLM) analysis highlights that subjects focused on developing communicative skills are strong predictors of overall test scores.

4.1. Insights from the Canonical Correlation Analysis (CCA)

The canonical correlation of 0.71 demonstrates substantial overlap between academic performance in DEI subjects and Saber 11 test results. This confirms that the curricular design in the DEI is intentionally aligned with the skills measured by the Saber 11. Notably, courses such as English, Social Studies and Citizenship, Spanish Language, and Advanced English show high canonical loadings, indicating their significant influence on global test performance. These findings underscore the impact of communicative skills on academic success, as reflected in both internal DEI evaluations and external assessments.

4.2. Insights from the Generalized Linear Model (GLM)

The GLM provides a closer look at how individual DEI subjects influence Saber 11 outcomes. Spanish Language, Foreign Language: English, and Chemistry all show positive and statistically significant effects on overall test scores. This reinforces the CCA results, highlighting the importance of strong communication skills in academic achievement.

Interestingly, Arts Education and Technology and Informatics display negative, significant coefficients. This unexpected result may be because the skills taught in these subjects are not directly measured by the Saber 11 tests. As a result, the way these subjects are evaluated internally may not align with the assessment criteria used by ICFES.

When using the Generalized Linear Model (GLM), it's important to note that the analysis reveals predictive associations—not direct cause-and-effect relationships. Negative coefficients for Arts Education and Technology & Informatics do not mean these subjects harm Saber 11 test results. In fact, research by Caucali M. (2020) and Timarán P. et al. (2019) demonstrates the positive impact of arts and technology on student performance.

Surprisingly, Mathematics and Physics did not show a significant influence in any of the methods applied in this study. This outcome may reflect differences in teaching styles or assessment methods. In reviewing the data, attention shifted to the strong and positive correlations between Spanish Language, English, and the exact sciences—Mathematics, Physics, and Advanced Mathematics.

Table 5 highlights these robust relationships, suggesting that subjects in the sciences support performance in language areas. This points to multicausal interactions where disciplines complement each other; for example, studying Mathematics and Physics can strengthen reasoning, analysis, and the interpretation of information, which are critical skills in language studies. These findings align with Junca Rodríguez (2019), who noted the impact of mathematics on language through the development of semiotic representation skills.

Tabla 5: Correlations between independent variables

Subjects	**Spanish Language**	**Foreign Language: English**
Physics	0,601	0,593
Mathematics	0,625	0,623
Advanced Mathematics	0,513	0,469

5. Contributions of the research

This study offers a fresh perspective on analyzing performance in the Saber 11 tests by focusing solely on internal variables produced by educational institutions. Unlike most existing research, which primarily examines socio-demographic factors outside the direct influence of DEIs, our approach highlights variables that schools can actively address.

The research demonstrates how Data Science techniques can prompt meaningful reflection within EIs, encouraging them to consider key questions such as:

- What defines the teaching practices of educators whose students excel in the Saber 11 areas?
- How can these effective practices be applied to other subjects?

Based on predictive modeling adjusted for academic performance at the institution, several targeted pedagogical strategies can be implemented to improve Saber 11 test outcomes.

For the DEI, where the study was conducted, both Canonical Correlation Analysis (CCA) and Generalized Linear Models (GLM) identified Spanish Language and Foreign Language: English as the subjects most strongly associated with high Saber 11 results. These areas offer valuable insights. By sharing effective teaching methods from these subjects, other departments can adapt and integrate similar strategies, potentially elevating overall test performance.

The models developed also enable early identification of students who are excelling or those at risk. This allows educators to tailor support and interventions, ensuring students are positioned for success.

6. Conclusions

- There is a clear and statistically significant link between subjects taught at the DEI and performance in Saber 11 test areas.
- Communicative skills play a pivotal role in test outcomes. Spanish Language and English, in particular, are strong predictors of student success.
- While Mathematics and Physics were not direct predictors, they show strong positive correlations with language subjects. This suggests that analytical skills developed in the exact sciences may indirectly strengthen communication abilities, benefiting overall test performance.

- Analyzing internal data empowers institutions to make informed pedagogical and curricular decisions. At DEI, the models support early identification of students in need and enable the implementation of personalized strategies.

- These findings provide DEIs with actionable statistical tools. They support curriculum alignment with Saber 11 requirements and promote continuous improvement using institution-specific data.

These results provide DEIs with statistical tools that are directly applicable to their pedagogical reality, facilitating the alignment of their curricula with the demands of the Saber 11 tests and supporting the re-evaluation of their educational and evaluative practices; all this using their own data as a resource for continuous improvement in educational quality.

6.1. Limitations of the research

The grading data from the DEI may differ in accuracy and interpretation compared to other institutions. As a result, these findings are currently applicable to the DEI. Replication in other settings is needed to confirm the broader applicability of the models and conclusions.

Recibido: Agosto de 2025
Aceptado: Diciembre de 2025

Referencias

J. Acosta S., D. J. Lancheros C., S. F. Umaña I., and J. R. Coronado H. Predictive models assessment based on crisp-dm methodology for students performance in Colombia - Saber 11 test. In *Procedia Computer Science*, volume 198, pages 512–517. Elsevier B.V., 2021.

M. A. Aguilera P., D. E. Martínez C., and O. J. Salcedo P. Forecasting model with machine learning in higher education ICFES exams. *Article in International Journal of Electrical and Computer Engineering*, 11:5402–5410, 2021.

A. Alvarez O. Análisis de la incidencia del contexto escolar en la prueba Saber 11. 2015.

J. D. Caucali M. Impacto positivo de la participación cultural en los resultados académicos de las pruebas Saber 11 en Bogotá en los años 2015 y 2017. Master's thesis, Universidad de los Andes, Colombia, 2020.

S. M. Chica G., D. M. Galvis G., and A. Ramirez H. Determinantes del rendimiento académico en Colombia. Pruebas ICFES - Saber 11°, 2009. *Revista Universidad EAFIT*, 46:48–72, 2010.

L. E. Contreras, H. Fuentes, and J. I. Rodríguez. Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación universitaria*, 13:233–246, 2020.

L. G. Díaz M. and M. A. Morales R. *Análisis Estadístico de Datos Multivariados*, volume 1. Universidad Nacional de Colombia, Bogotá, 2012.

J. E. Díaz-Pinzón. Analysis of regression and correlation in the Saber 11 test in the period 2011 to 2017. *IJERI: International Journal of Educational Research and Innovation*, (13):96–110, 2020.

P. K. Dunn and G. K. Smyth. *Generalized Linear Models With Examples in R.* Springer Texts in Statistics, New-York, 2018.

G. A. Junca Rodríguez. Desempeño académico en las pruebas Saber 11. *Panorama Económico*, 27:8–38, 2019.

D. S. Martínez L. Análisis de la relación entre las pruebas Saber Pro y los cursos realizados por estudiantes de la licenciatura en matemáticas utilizando correlación canónica. 2013.

I. Montes, J. D. Garcés, and A. J. Jaramillo. Academic achievement: which role plays the institutional factors? *Cetro de Investigaciones Económicas y Financieras*, 2014.

M. J. Otálora S. and D. F. Torres L. Mapa de brecha de evidencia de los factores asociados al aprendizaje sobre el desempeño en la prueba Saber 11. 2022.

M. T. Rincón Cabrera. Formación en competencias comunicativas en educación media y su incidencia en la educación superior. 2014.

D. D. Rodríguez R., R. E. Ordoñez O., and M. E. Hidalgo V. Academic performance determinants of high school students in the department of Nariño, Colombia. *Lecturas de Economia*, pages 87–126, 2021.

D. P. Sáenz C. and S. Toro V. Acciones de mejora para la formación en educación media según el análisis de resultados en Saber 11. *Revista de Investigación y Pedagogía Praxis Saber*, 14, 2023.

SMECE. Informe de resultados de evaluación 2023. 2023.

R. Timarán P., J. CaicedoZ., and A. Hidalgo T. Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. *Revista de Investigación, Desarrollo e Innovación*, 9:363–378, 2019.

V. Vargas C. and L. F. Ardila. Predicción del desempeño en las pruebas Saber 11 utilizando variables del contexto socio-económico de los aplicantes mediante un análisis estadístico con técnicas de machine learning policy brief. 2024.