Análisis exploratorio y selección de atributos relevantes sobre el dataset de las pruebas saber 11 para la ciudad de Cartagena

Exploratory analysis and selection of relevant attributes on the saber 11 test dataset for the city of Cartagena

Análise exploratória e seleção de atributos relevantes no dataset dos exames Saber 11 para a cidade de Cartagena

- Artículo de investigación -

Gabriel Elías Chanchí Golondrino[1]
Universidad de Cartagena

Dayana Alejandra Barrera Buitrago[2]
Nidia Danigza Lugo López[3]
Universidad Nacional Abierta y a Distancia

Recibido: 15.03.2024

Aceptado: 15.05.2024

Resumen

Este trabajo tiene como objetivo el desarrollo de un estudio basado tanto en el análisis exploratorio de datos como en la selección de los mejores atributos que inciden en el rendimiento académico, utilizando el conjunto de datos de las pruebas Saber 11 de 2019 de la ciudad de Cartagena. Para el desarrollo del estudio se hizo uso de una adaptación de la metodología de minería de datos SEMMA, definiendo cuatro fases metodológicas, a saber: F1. Muestreo de los datos; F2. Exploración y modificación de los datos; F3. Aplicación del método de selección de atributos; y F4.

[1] gchanchig@unicartagena.edu.co
https://orcid.org/0000-0002-0257-1988
[2] dayana.barrera@unad.edu.co
https://orcid.org/0000-0001-8867-9705
[3] nidia.lugo@unad.edu.co
https://orcid.org/0000-0002-9096-5767

Análisis de los resultados obtenidos. Como resultados relevantes del estudio, se obtuvo que las áreas con medias más altas fueron lectura crítica y matemáticas. Asimismo, se evidenció que la formación de los padres a nivel posgradual tiene una influencia representativa en el rendimiento del estudiante. Finalmente, se identificó un conjunto de atributos del dataset que inciden en el rendimiento de las cinco áreas de la prueba. Este estudio pretende servir de referencia a nivel investigativo para la caracterización del rendimiento académico en diferentes regiones, con el fin de contribuir al desarrollo de estrategias enfocadas en el fortalecimiento de la calidad.

Palabras clave: análisis estadístico, análisis de correlación, rendimiento escolar, examen, estado, habilidades

Abstract

This work aims to develop a study based on both exploratory data analysis and the selection of the best attributes that affect academic performance, using the data set of the 2019 Saber 11 tests from the city of Cartagena. To develop the study, an adaptation of the SEMMA data mining methodology was used, defining four methodological phases, namely: F1. Data sampling; F2. Exploration and modification of data; F3. Application of attribute selection method; and F4. Analysis of the obtained results. As relevant results of the study, it was obtained that the areas with the highest averages were critical reading and mathematics. Likewise, it was evidenced that parent training at the postgraduate level has a representative influence on student performance. Finally, a set of dataset attributes that affect the performance of the five areas of the test were identified. This study aims to serve as a reference at a research level for the characterization of academic performance in different regions, to contribute to the development of strategies focused on strengthening educational quality.

Keywords: statistical analysis, correlation analysis, school performance, exam, state, skills

Resumo

Este trabalho tem como objetivo desenvolver um estudo baseado tanto na análise exploratória de dados quanto na seleção dos melhores atributos que afetam o desempenho acadêmico, utilizando o conjunto de dados das provas Sabre 11 2019 da cidade de Cartagena. Para desenvolver o estudo foi utilizada uma adaptação da metodologia de mineração de dados SEMMA, definindo quatro fases metodológicas, a saber: F1. Amostragem de dados; F2. Exploração e modificação de dados; F3. Aplicação de método de seleção de atributos; e F4. Análise dos resultados obtidos. Como resultados relevantes do estudo, obteve-se que as áreas com maiores médias foram leitura crítica e matemática. Da mesma forma, evidenciou-se que a formação dos pais em nível de pós-graduação tem influência representativa no desempenho dos alunos. Por fim, foi identificado um conjunto de atributos do conjunto de dados que afetam o desempenho das cinco áreas do teste. Este estudo pretende servir de referência a nível de investigação para a caracterização do desempenho académico em diferentes regiões, de forma a contribuir para o desenvolvimento de estratégias focadas no fortalecimento da qualidade educativa.

Palavras-chave: análise estatística, análise de correlação, desempenho escolar, exame, estado, habilidades

Introduction

According to UNESCO, education, beyond being a fundamental right, entails a transformative process that enables the identification of whether the defined standards align with the demands imposed by society. Consequently, the educational evaluation process becomes pivotal for the progressive enhancement of education (Sanabria James et al., 2020). In this vein, the educational policy advocated by the Ministry of National Education of Colombia aims for the development of academic competencies by educational institutions to be the primary objective in basic, middle, and higher education (Palacios-Mena, 2018), conceiving competencies from the standpoint of a successful life and a developed society. One of the most significant indicators of the quality of education in Colombia is the results attained by students in the Saber tests, which gauge the extent of development of academic competencies across various disciplines (Garizabalo Dávila, 2012). These test outcomes serve as a diagnostic tool facilitating the identification of students' strengths and weaknesses, thereby enabling educational institutions to receive feedback on curricular aspects, fostering strategic decisions aimed at enhancing the effectiveness of the teaching-learning process (Acero et al., 2016; Sanabria James et al., 2020). The ICFES (Colombian Institute for the Evaluation of Education) is the governmental institute tasked with designing and evaluating the Saber tests for the third and fifth grades of primary education, as well as for the ninth and eleventh grades of secondary education and also for the university level, in which case they are referred to as Saber Pro tests (Morales-Piñero et al., 2019; Palacios-Mena & Rodríguez-Márquez, 2019).

The Ministry of National Education (MEN) utilizes the Saber tests across its various levels to monitor and assess the quality of education accessible to students, where the Saber 11 tests are pivotal due to the linkage between secondary and higher education (Iguarán Jiménez et al., 2023; Timarán Pereira et al., 2020). The Saber 11 tests correspond to standardized assessments taken by grade 11 students as a requirement for admission to different undergraduate programs at universities in Colombia, thereby evaluating through closed-ended questions five areas, namely: Critical Reading, Mathematics, Natural Sciences, Social Sciences, and English (Alonso et al., 2012; Ruiz-Escorcia et al., 2017). The competencies assessed by the ICFES through the Saber tests in various areas include interpretation and representation, reasoning and argumentation, formulation and execution, articulation of a text, reflection on content, social thinking, explanation of different phenomena, among others (Díaz Pinzón, 2020).

With the advancement of data science in recent years, educational data mining (EDM) has emerged as an interdisciplinary field tasked with applying computational methods to explore data originating from the educational context. Its purpose is to extract value-added information that benefits decision-making in the educational sector regarding academic performance, the teaching-learning process, or the enhancement of educational quality (Devasia et al., 2016; Nasiri et al., 2012; Pathan et al., 2014; Romero & Ventura, 2010). In this regard, various contributions have been identified in the state of the art concerning the application of EDM in the context of Saber tests. Thus, in (Timarán-Pereira et al., 2019), a classification model based on decision trees is applied to detect factors associated with the academic performance of Colombian students who took the Saber 11 tests in 2015 and 2016. Similarly, in (Timarán Buchely & Timarán Pereira, 2023), a data mining study based on decision trees is conducted to determine patterns associated with academic performance in the generic competencies of the Saber Pro tests among students at the Universidad Javeriana de Cali in the years 2017 and 2018. In (Chanchí-Golondrino et al., 2021), an exploratory and spatial analysis of the data pertaining to the Saber 5 tests of 2016 is conducted to correlate and classify test results with the spatial distribution of the data. In (García-González et al., 2019), a neural network-based model is proposed for predicting performance in the Saber Pro tests at Higher Education Institutions in the city of Barranquilla, based on a dataset containing the results of this test. In (Arboleda-Posada et al., 2022), a data mining study is conducted to identify factors influencing performance levels for military or police training programs in Colombia, utilizing data from the Saber Pro tests between 2012 and 2019. In (Timarán-Pereira et al., 2023), decision tree models are applied to detect patterns influencing academic performance in the mathematics competency of the Saber 5 tests. In (Oviedo Carrascal & Jiménez Giraldo, 2019), a data mining study based on supervised and unsupervised learning is conducted on the results of the Saber Pro tests in the Antioquia department (Colombia) from the year 2016, aiming to determine the economic, social, and demographic factors influencing students' performance. In (Narváez Zúñiga, 2022), various predictive models are tested to determine the most effective one for predicting performance based on different factors from the Saber Pro tests between 2016 and 2019, revealing that the multivariable linear regression model is the most suitable. In (Acevedo et al., 2015), the factors associated with course repetition and graduation delays in Engineering programs at the University of Cartagena are assessed, considering the results obtained in the competencies of the Saber Pro tests, as well as information published in the university's statistical bulletins. In (Gorostiaga & Rojo-Álvarez, 2016), a study based on exploratory analysis and machine learning is conducted to characterize the PISA tests in Spain, revealing that variables such as computer availability and immigration status are key factors in mathematics performance.

Based on the studies, it is important to highlight that various research endeavors have been conducted across different regions, utilizing data mining techniques to identify factors influencing academic performance in standardized tests among primary, secondary, and university-level students. In most of these studies, besides conducting exploratory data analysis, machine learning models are applied, with decision tree models being predominantly used. However, there is a lack of research focused on studying the variables affecting the academic performance of 11th-grade students in the city of Cartagena. Similarly, the explored studies did not reveal a prior investigation into the optimal attributes influencing the prediction of performance across the five areas through heuristic-based selection methods. The objective of studying Cartagena is that, despite being one of the most important and tourist cities in Colombia, it is also one of the major cities with a high index of poverty and inequality (Ayala-García & Meisel-Roca, 2016).

This article proposes both the conduct of exploratory data analysis and the application of a heuristic search method for selecting the best attributes on a representative sample of the dataset from the Saber 11 tests of 2019, corresponding to the city of Cartagena de Indias (13,535 records). The objective of the analysis is to determine possible relationships between the various variables in the academic, social, and demographic dataset with the overall academic performance of students and in the knowledge areas assessed by the test (critical reading, mathematics, natural sciences, social sciences, and English). It is worth mentioning that the dataset used is freely accessible and was obtained from the Colombia open data portal. Similarly, it is important to highlight that both the exploratory data analysis and the process of selecting the best attributes from the dataset were carried out with the support of open-source tools. Thus, for exploratory data analysis, open-source libraries such as pandas, numpy, matplotlib, seaborn, and scikit-learn were used, while for obtaining the best attributes, the BestFirst algorithm implementation provided by the GPL-licensed weka tool was utilized. Hence, the purpose of this study is to identify relevant findings that lead governmental authorities and educational institutions to identify the main factors influencing student performance in the city of Cartagena, aiming to propose strategies that contribute to improving academic performance. Similarly, the identification of relevant attributes is a fundamental contribution to the construction of predictive models associated with the academic performance of students in Cartagena in the Saber 11 tests.

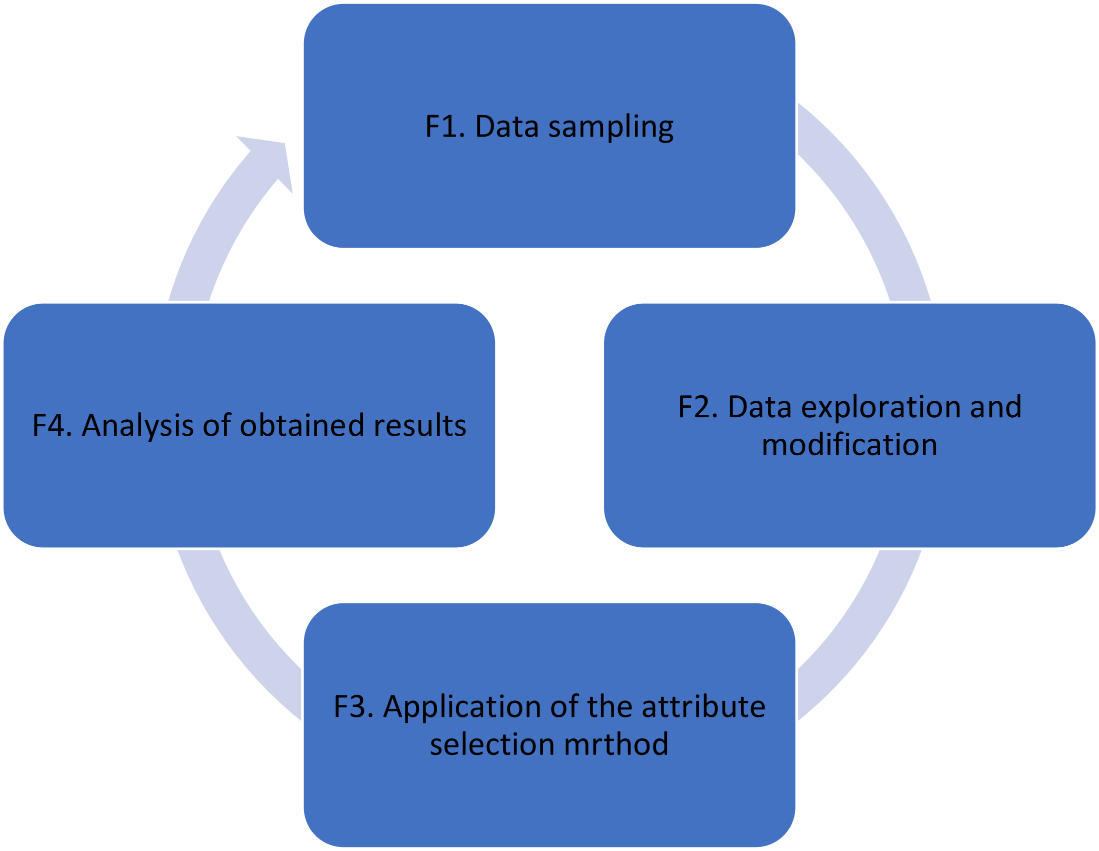
The remainder of the article is organized as follows: firstly, the methodology considered for the development of the present study is presented. Subsequently, the results obtained from the pre-processing of the dataset are described, as well as those related to the application of exploratory data analysis and the selection of relevant dataset attributes. Likewise, this section presents the discussion of the results in relation to some literature works. Finally, in the last section, the conclusions and future work derived from the present research are presented.

Methodology

For the development of the present research, a 4-phase adaptation of the SEMMA methodology (Sample, Explore, Modify, Model, and Assess) (Palacios-Gómez et al., 2016; Tariq et al., 2019) was employed, whereby the following phases were defined: F1. Data sampling, F2. Data exploration and modification, F3. Application of the attribute selection method, and F4. Analysis of obtained results (see Figure 1).

In phase 1 of the methodology, the dataset corresponding to the Saber 11 tests at the national level was obtained from the Colombia open data portal. This dataset initially comprises a total of 546,612 records corresponding to the results of students from different municipalities and departments of Colombia. Through the use of the pandas Python library, the records corresponding to the city of Cartagena were filtered, resulting in a total of 13,535 records. It is worth mentioning that, when filtering for the city of Cartagena, there are fields or columns that must be removed as they are not considered relevant in the analysis.

Figure 1. Methodology considered


F1. Data sampling
F2. Data exploration and modification
F3. Application of the attribute selection mrthod
F4. Analysis of obtained results

Source: Own elaboration

In phase 2 of the methodology, following the filter applied for the city of Cartagena, the first step involved reviewing the 83 columns of the dataset to identify those that would not be considered in the exploratory analysis. Consequently, it was decided to discard 23 variables. These variables were removed as they correspond to identifiers, consecutive values, or columns that all have the same value (country, department, municipality). It is also worth noting that these columns were eliminated utilizing the advantages provided by the pandas library. This step ensured that the dataset was refined and relevant for the subsequent phases of data exploration and attribute selection.

Table 1. Columns of the dataset deleted

Columns deleted
ID, ESTU_NACIONALIDAD, PERIODO, ESTU_CONSECUTIVO, ESTU_ESTUDIANTE, ESTU_PAIS_RESIDE, ESTU_DEPTO_RESIDE, ESTU_COD_RESIDE_DEPTO, ESTU_MCPIO_RESIDE, ESTU_COD_RESIDE_MCPIO, COLE_CODIGO_ICFES, COLE_COD_DANE_ESTABLECIMIENTO, COLE_COD_DANE_SEDE, COLE_COD_MCPIO_UBICACION, COLE_MCPIO_UBICACION, COLE_COD_DEPTO_UBICACION, ESTU_PRIVADO_LIBERTAD, COLE_DEPTO_UBICACION, ESTU_COD_MCPIO_PRESENTACION, ESTU_MCPIO_PRESENTACION, ESTU_DEPTO_PRESENTACION, ESTU_COD_DEPTO_PRESENTACION.

Source: Own elaboration

Once the columns listed in Table 1 were deleted, Table 2 presents the set of 60 columns considered for the development of exploratory data analysis and the application of inference rules. These variables generally include information about the student at the socio-economic, family, educational, and nutritional levels, as well as the performance obtained in the areas of critical reading, mathematics, natural sciences, social sciences, and English.

Table 2. Columns considered in the dataset

Column	Description
ESTU_TIPODOCUMENTO	It corresponds to the type of document of the student (It has been left because in some cases it is an ID card and in others it is an identity card). It is a categorical data.
ESTU_GENERO	It corresponds to the gender of the student (M or F). It is categorical data.
ESTU_FECHANACIMIENTO	It corresponds to the student's date of birth.
ESTU_TIENEETNIA	It indicates whether the student belongs to a particular ethnic group or not. It is categorical data.
FAMI_ESTRATOVIVIENDA, FAMI_PERSONASHOGAR, FAMI_CUARTOSHOGAR	They correspond respectively to the stratum to which the household belongs, the number of people per household, and the number of rooms in the household. They are categorical data.
FAMI_EDUCACIONPADRE, FAMI_EDUCACIONMADRE	They respectively correspond to the educational level of the student's father and mother. They are categorical data.
FAMI_TRABAJOLABORPADRE, FAMI_TRABAJOLABORMADRE	They respectively correspond to the job description of the student's father and mother. They are categorical data.
FAMI_TIENEINTERNET, FAMI_TIENESERVICIOTV	They respectively indicate whether the family has internet or television service. They are categorical data.
FAMI_TIENECOMPUTADOR, FAMI_TIENELAVADORA, FAMI_TIENEHORNOMICROOGAS	They respectively indicate whether the family has a computer or a washing machine or a microwave/gas service. They are categorical

Column	Description
	data.
FAMI_TIENEAUTOMOVIL, FAMI_TIENEMOTOCICLETA	They respectively indicate whether the family has a car or a motorcycle. They are categorical data.
FAMI_TIENECONSOLAVIDEOJUEGOS	It indicates whether the family has a video game console. It is categorical data.
FAMI_NUMLIBROS	It indicates the range of books available to the family. It is categorical data.
FAMI_COMELECHEDERIVADOS	It indicates the range or frequency of milk or dairy consumption within the family. It is categorical data.
FAMI_COMECARNEPESCADOHUEVO	It indicates the range or frequency of meat, fish, or egg consumption within the family. It is categorical data.
FAMI_COMECEREALFRUTOSLEGUMBRE:	It indicates the range or frequency of cereal, fruit, or legume consumption within the family. It is categorical data.
FAMI_SITUACIONECONOMICA	It indicates the level or range of economic status. It is categorical data.
ESTU_DEDICACIONLECTURADIARIA, ESTU_DEDICACIONINTERNET	They indicate the range of time dedicated by the student to daily reading or internet browsing. They are categorical data.
ESTU_HORASSEMANATRABAJA	It corresponds to the range of hours the student dedicates weekly to work. It is categorical data.
ESTU_TIPOREMUNERACION	It corresponds to the type of remuneration received by the student. It is categorical data.
COLE_NOMBRE_ESTABLECIMIENTO	It corresponds to the name of the educational institution to which the student

Column	Description
	belongs. It is categorical data.
COLE_GENERO, COLE_NATURALEZA, COLE_CALENDARIO, COLE_BILINGUE, COLE_CARACTER	They indicate, regarding the student's school, the gender (mixed, male, female), its nature (public or private), if it is bilingual or not, the type of calendar, and the character (academic, technical, technical/academic). They are categorical data.
COLE_NOMBRE_SEDE, COLE_SEDE_PRINCIPAL, COLE_AREA_UBICACION, COLE_JORNADA	They indicate, regarding the student's school, the name of the campus, if the school is a main campus or not, if the school is located in a rural or urban area, and the type of school schedule. They are categorical data.
PUNT_LECTURA_CRITICA, PERCENTIL_LECTURA_CRITICA, DESEMP_LECTURA_CRITICA	They correspond respectively to the score obtained by the student in critical reading, the percentile corresponding to that score, and the level in which the score is classified. The first two attributes are numerical, and the last one is categorical.
PUNT_MATEMATICAS, PERCENTIL_MATEMATICAS, DESEMP_MATEMATICAS	They correspond respectively to the score obtained by the student in mathematics, the percentile corresponding to that score, and the level in which the score is classified. The first two attributes are numerical, and the last one is categorical.
PUNT_C_NATURALES, PERCENTIL_C_NATURALES, PERCENTIL_C_NATURALES	They correspond respectively to the score obtained by the student in natural sciences, the percentile corresponding to that score, and the level in which the score is classified. The first two attributes are numerical, and

Column	Description
	the last one is categorical.
PUNT_SOCIALES_CIUDADANAS, PERCENTIL_SOCIALES_CIUDADANAS, DESEMP_SOCIALES_CIUDADANAS	They correspond respectively to the score obtained by the student in social and citizenship sciences, the percentile corresponding to that score, and the level in which the score is classified. The first two attributes are numerical, and the last one is categorical.
PUNT_INGLES, PERCENTIL_INGLES, DESEMP_INGLES	They correspond respectively to the score obtained by the student in english, the percentile corresponding to that score, and the level in which the score is classified. The first two attributes are numerical, and the last one is categorical.
PUNT_GLOBAL, PERCENTIL_GLOBAL	They correspond to the overall score obtained by the student and the percentile associated with that score. They are numerical data.
ESTU_INSE_INDIVIDUAL	It corresponds to the value of the socioeconomic level indicator of the student. It is numerical data.
ESTU_NSE_INDIVIDUAL, ESTU_NSE_ESTABLECIMIENTO	They indicate the socioeconomic level of the student and the educational institution. They are categorical data.
ESTU_ESTADOINVESTIGACION	It indicates the research status of the student. It is categorical data.
ESTU_GENERACION-E	It indicates the type of benefit that the student receives from the Generación-E program. It is categorical data.

Fuente: Own elaboration

From the columns defined for the dataset and presented in Table 2, the process of cleaning different records in these columns with NA or "-" values was carried out. For categorical attributes with missing data, imputation was performed using the mode, while for numerical attributes with missing data, imputation was carried out using the mean. The aforementioned imputation processes were conducted using the functionalities provided by the panda’s library data frames in Python. With the dataset free of null or missing values, the descriptive and exploratory data analysis continued, which included the following operations: counting categories and determining the mode of discrete variables, obtaining statistical measures (mean, minimum, and maximum values) for numerical variables, analyzing correlations between numerical variables to determine their impact on performance, analyzing scatter plots of numerical variables to determine linear increasing relationships, analyzing the distribution of some numerical variables, generating box and whisker plots for columns associated with performance in the 5 areas, and finally generating violin plots that relate different categorical variables to the student's overall performance in the Saber tests.

In phase 3 of the methodology, following the adjustments and cleaning carried out in phase 2, the BestFirst algorithmic model was applied to determine the best attributes affecting performance in each of the 5 areas of the test and overall performance in the context of Cartagena. This algorithm corresponds to a heuristic search model that explores the space of possible attribute subsets, continuously evaluating the quality of each combination with respect to a predefined evaluation criterion. This algorithm allows the user to specify an evaluator, such as CfsSubsetEval, which determines the relevance and redundancy of the attributes. BestFirst conducts a heuristic-guided search to find subsets that maximize the evaluation criterion. Among its advantages is the ability to handle large attribute spaces and flexibility to adapt to different evaluation criteria. By selecting optimal subsets, BestFirst facilitates improving the performance of machine learning models by reducing dimensionality and highlighting key attributes for the specific task.

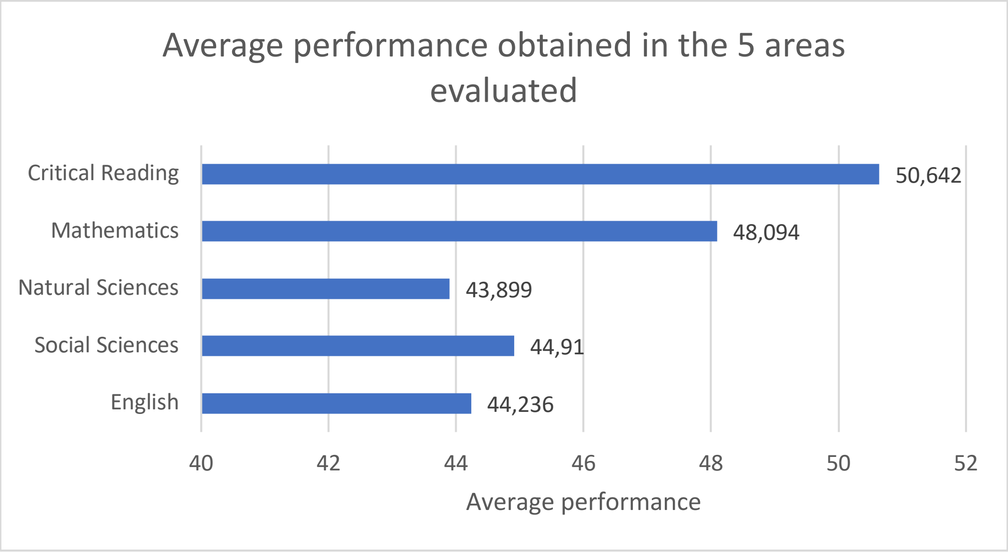
Finally, in phase 4 of the methodology, the attributes determined by the BestFirst algorithm for each of the 5 areas evaluated in the test were collected, along with the analysis of the merit metric of each subset of data obtained. This metric indicates the quality of the subset of attributes selected by the algorithm during the search and refers to the measure of how well a particular set of attributes performs according to the selected evaluation criterion.

Results

In the first instance, at the level of descriptive analysis of the dataset, statistical measures associated with the numerical variables were obtained. A key finding was the average performance achieved globally and in the five disciplinary areas of the test (critical reading, mathematics, natural sciences, social sciences, and English). The average overall score obtained in the city of Cartagena was 236 out of 500 possible points, while the maximum and minimum scores obtained globally were 475 and 195, respectively.

Similarly, the score with the highest mean in the city of Cartagena corresponds to critical reading competence, with a value of 50.642, whereas the score with the lowest mean in Cartagena is in social and civic sciences, with a value of 43.8963. Additionally, the score showing the lowest data dispersion (standard deviation) is in critical reading, with a value of 11.2107. The average results in the five considered areas can be more clearly observed in Figure 2.

Figure 2. Average performance obtained in the 5 areas evaluated



Source: Own elaboration

Similarly, upon analyzing the quantity of maximum values (100 points), it was found that the English area had 28 maximum values, mathematics obtained a total of 16 maximum values, social sciences had 11 maximum values, critical reading had 8 maximum values, while natural sciences obtained 4 maximum values. On the other hand, concerning minimum values (0 points) in the different areas, 9 values were obtained for critical reading, 5 values for the English area, and 1 value for social sciences.

In the same vein, upon conducting quartile analysis, it was found that for critical reading, 75% of students (Q3 quartile) obtained scores equal to or below 59, whereas for social sciences, 75% of students obtained scores equal to or below 53. Furthermore, regarding the count of categorical variables, it is noteworthy in terms of mode that most students who took the test belong to socioeconomic stratum 1, come from families of 3 to 4 members, reside primarily in houses with 2 rooms, and the most common educational level among their parents is high school. Moreover, the fathers mostly work independently, the schools attended by the students are predominantly technical/academic and non-bilingual, many schools are located in urban areas and operate in the morning shift, mostly adhering to schedule A.

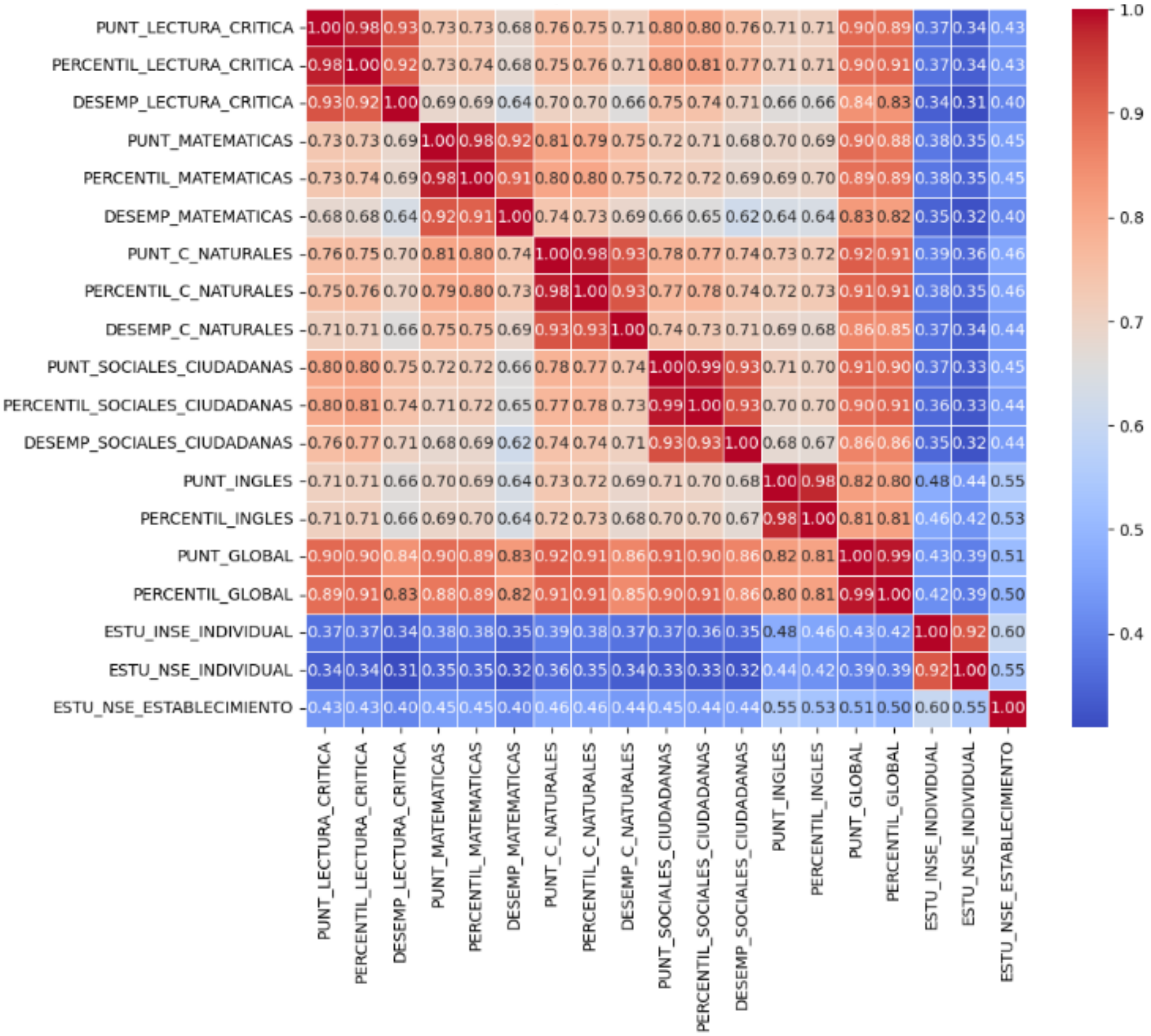
Now, to ascertain a potential correlation among the continuous numerical variables of the dataset and to determine whether two variables exhibit a possible linear relationship that would subsequently allow for the fitting of a regression model, a correlation matrix was obtained among these variables (see Figure 3), which is depicted through a heatmap generated using the seaborn library in Python.

Based on the correlation matrix presented in Figure 3, it becomes pertinent to assess the correlations between performance in the five different areas and the overall score obtained. Likewise, it is important to describe the correlation existing between economic indices and the score of each area of interest.

In accordance with the foregoing, it is observable from Figure 3 that all five disciplinary areas exhibit correlations exceeding 82% in all cases, with social sciences showing the highest correlation with the overall score at 92%, while English displays the lowest correlation at 82%. Additionally, it is noteworthy that naturally, each column of scores for the five areas demonstrates a high correlation (values exceeding 0.9) with their associated attributes (percentile and performance). Likewise, it is important to mention that the highest correlation of critical reading scores is observed with social sciences scores (0.8). Similarly, the highest correlation of mathematics scores is with natural sciences scores (0.81).

On the other hand, the highest correlations of natural sciences scores are with mathematics scores (0.81) and social sciences scores (0.78). Furthermore, the highest correlation of English scores is with natural sciences scores (0.73). Finally, concerning economic variables, it is notable that a high correlation between the overall score and the student's socioeconomic index (INSE) is not observed, with a value of 0.43. Moreover, the highest correlation between the socioeconomic index (INSE) and scores in the different areas is with English (0.48), however, this value is not high. This leads to the conclusion that it is possible to fit linear regression models between some knowledge areas and the overall score, as correlations exceeding 0.8 are evident in these cases.

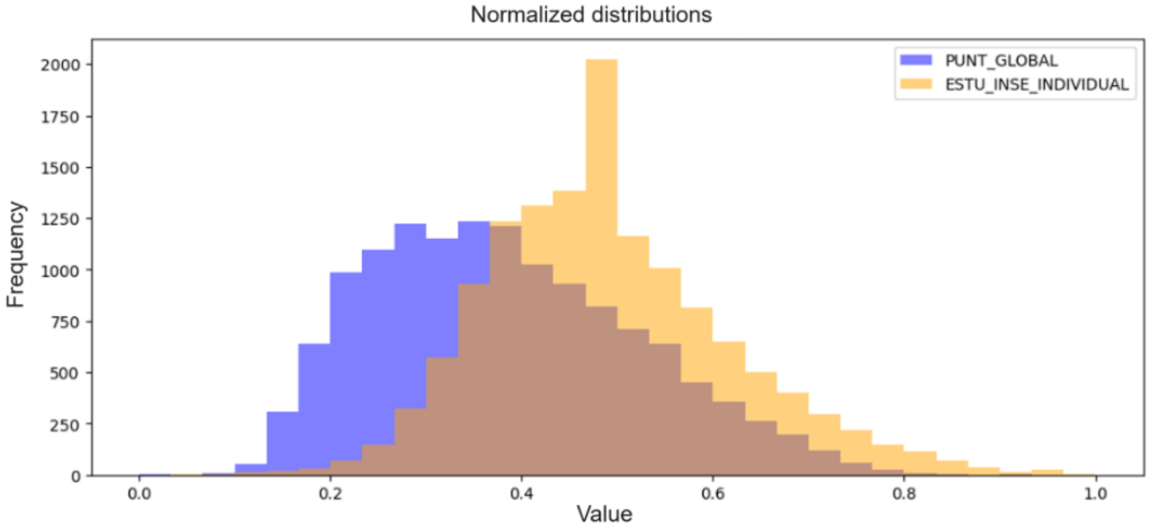
Figure 3. Correlation matrix of the numerical variables of the dataset



Source: Own elaboration

Continuing with the distribution analysis, the variables corresponding to the overall score (PUNT_GLOBAL) and the individual socioeconomic index of the student (ESTU_INSE_INDIVIDUAL) were selected. Consequently, the distribution curves for the normalized values of these variables were obtained, as depicted in Figure 4.

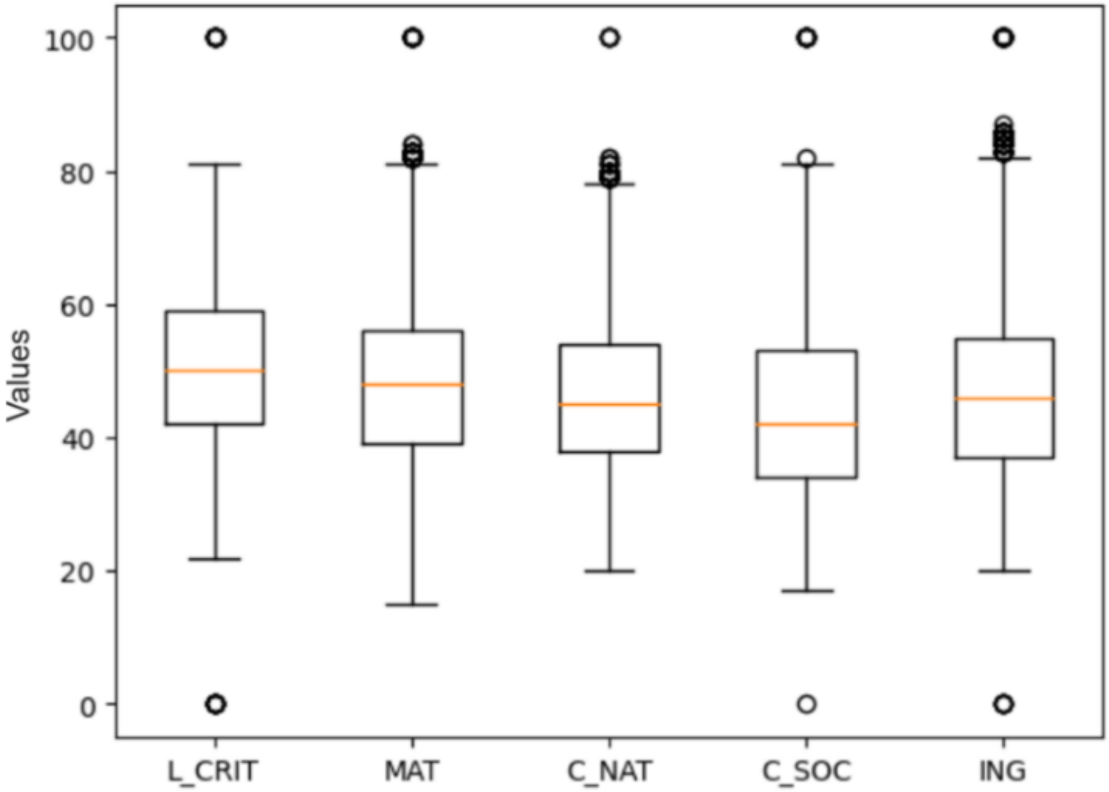
Figure 4. Distribution of the global score and individual socioeconomic index variables



Source: Own elaboration

Through the graph depicted in Figure 4, it is possible to infer that the distribution of the variable PUNT_GLOBAL exhibits a left-skewed pattern in comparison to the variable ESTU_INSE_INDIVIDUAL, which adheres more closely to a normal distribution. This implies that in the case of the variable PUNT_GLOBAL, there is a concentration of higher values towards the left-hand side of the distribution, indicating smaller values than the median. In other words, there is a greater number of overall scores lying below the median of the values. Now, in order to assess the distribution of scores in the five areas (critical reading, mathematics, natural sciences, social sciences, and English) in relation to the median, a box plot was generated, as presented in Figure 5.

Figure 5. Box plot diagram for the 5 evaluated knowledge areas

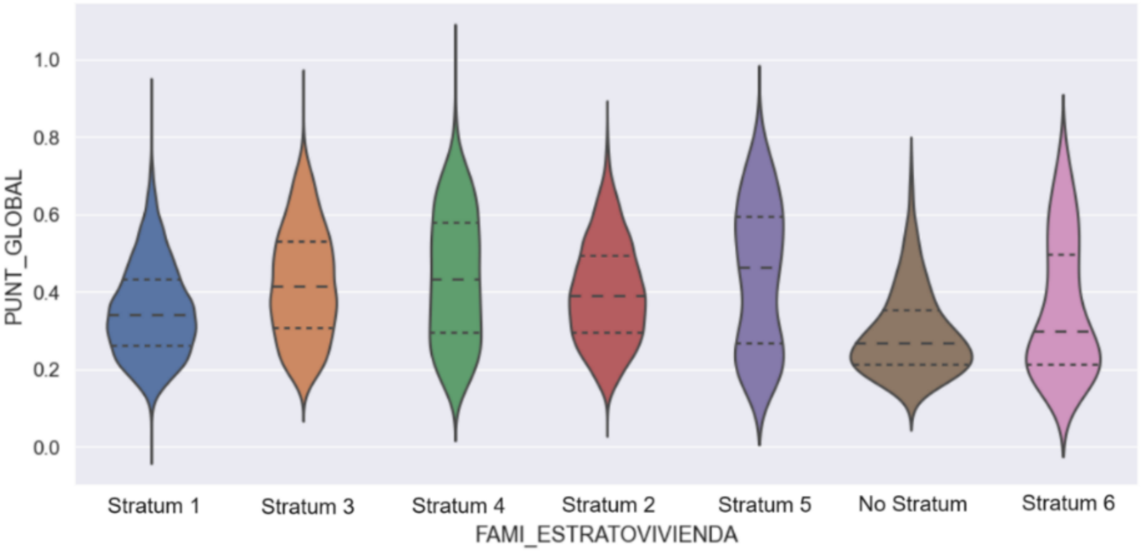


Source: Own elaboration

According to the results obtained from the box plot in Figure 5, the median values of the five areas range between 40 and 60, with the highest median value observed in the critical reading area, while the lowest median value is in the social sciences area. Additionally, the area with the least data dispersion (narrower box) is natural sciences, whereas the area with the most data dispersion (wider box) is social sciences. Concerning data symmetry, the critical reading and mathematics areas demonstrate better symmetry in the data distribution, suggesting a balanced distribution within the interquartile region and an equal amount of data above and below the median for these two areas. In contrast, the natural sciences, social sciences, and English areas exhibit a negative skewness or left tail, indicating that scores in these areas are concentrated below the median. Finally, the areas with a higher number of different outliers are mathematics, natural sciences, and English, corresponding to scores close to or equivalent to the maximum and minimum values.

To identify the relationship between certain socioeconomic categorical variables and overall performance, violin plots were employed. In Figure 6, the violin plot relates the student's socioeconomic stratum (FAMI_ESTRATOVIVIENDA) to the normalized overall score obtained in the Saber tests.

Figure 6. Violin plot for the strata and the normalized overall score

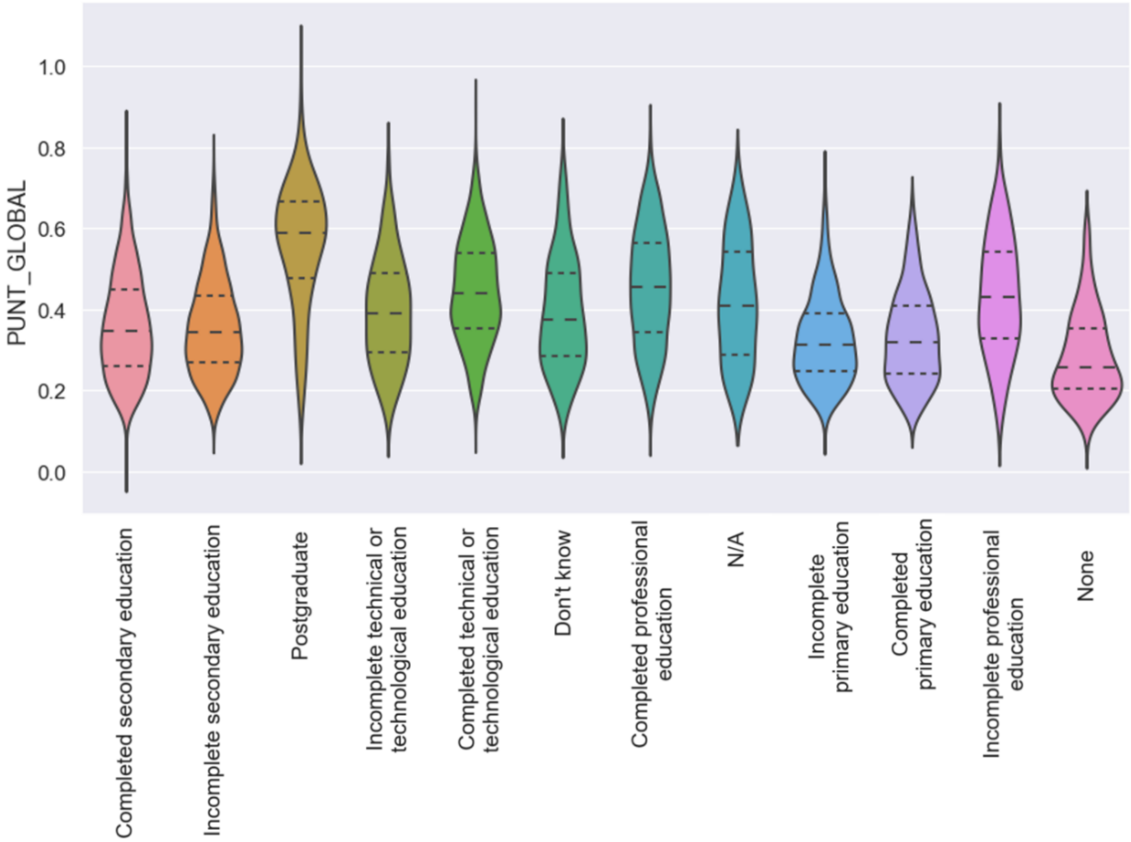


Source: Own elaboration

From Figure 6, it is possible to observe that there is no clear differentiation between socioeconomic stratum and performance in the test, as evidenced by overlaps in the different violins for each area. However, the highest median is observed for stratum 5, while the lowest median is apparent in the categories without stratum and in stratum 6.

On the other hand, Figure 7 depicts the violin plot relating the categorical variable FAMI_EDUCACIONPADRE to the normalized overall score obtained in the test, where FAMI_EDUCACIONPADRE corresponds to the categorized educational level of the student's father. The purpose of this graph is to ascertain whether there is an impact between the various levels of education of the father and the overall score. According to the results in Figure 7, it is evident that students whose fathers have postgraduate education achieved a higher overall score in the test, with a median of 0.6 in the normalized value of the overall score. Similarly, students whose fathers do not have any formal education exhibit the lowest median in the overall score.

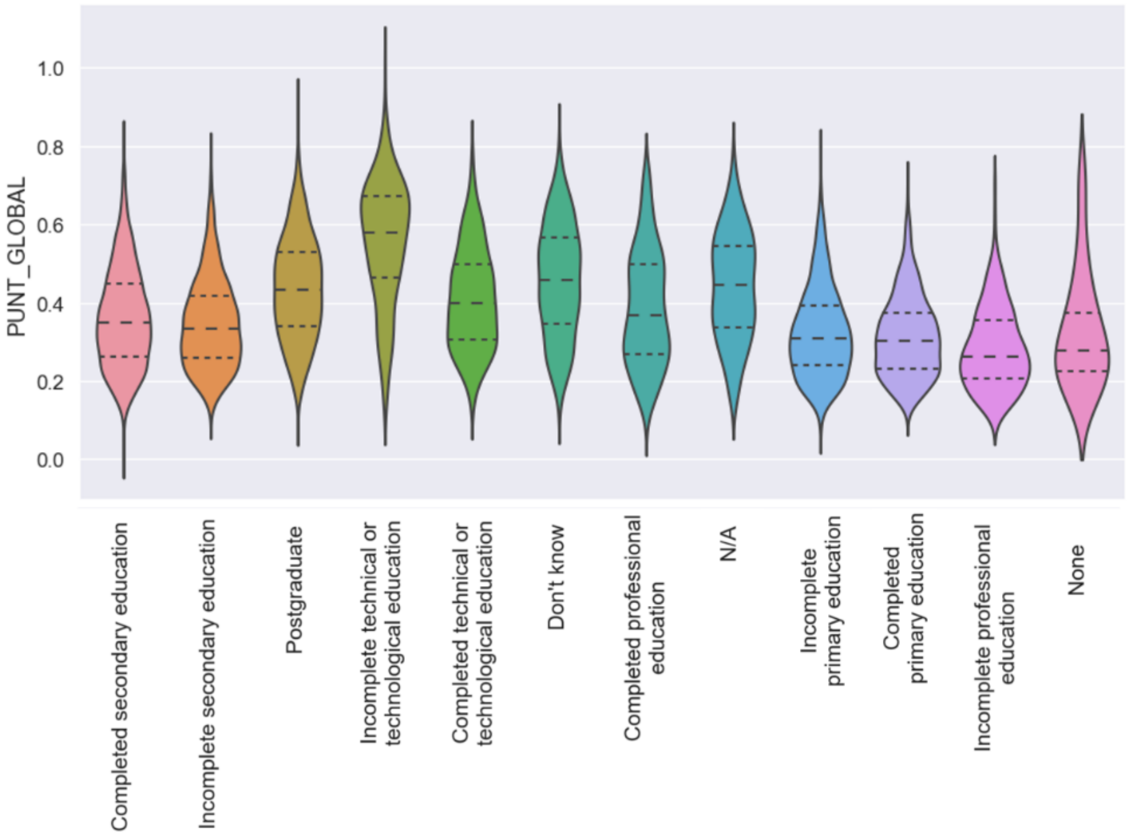
Figure 7. Violin plot for paternal education and normalized overall score



Source: Own elaboration

Similarly, in Figure 8, the violin plot relates the categorical variable FAMI_EDUCACIONMADRE to the normalized overall score obtained in the test, where FAMI_EDUCACIONMADRE represents the categorized educational level of the student's mother. The results in Figure 8 are consistent with those obtained for the father's education level. When the mother has postgraduate education, a higher value is observed in the overall test score, with a median close to 0.6 in the normalized value of the overall score. Conversely, students whose mothers do not have any formal education exhibit the lowest median in the overall score.

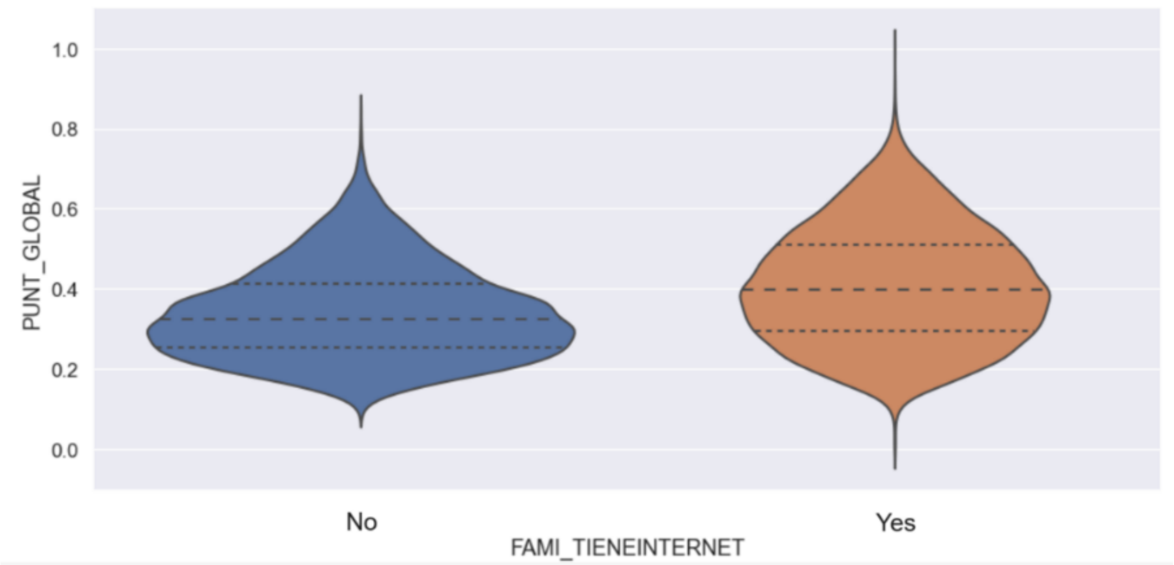
Figure 8. Violin plot for maternal education and normalized overall score



Source: Own elaboration

On another note, Figure 9 presents a violin plot that relates the variable FAMI_TIENEINTERNET to the normalized overall test score. FAMI_TIENEINTERNET corresponds to whether the student's family has internet service.

Figure 9. Violin plot for internet service and normalized overall score

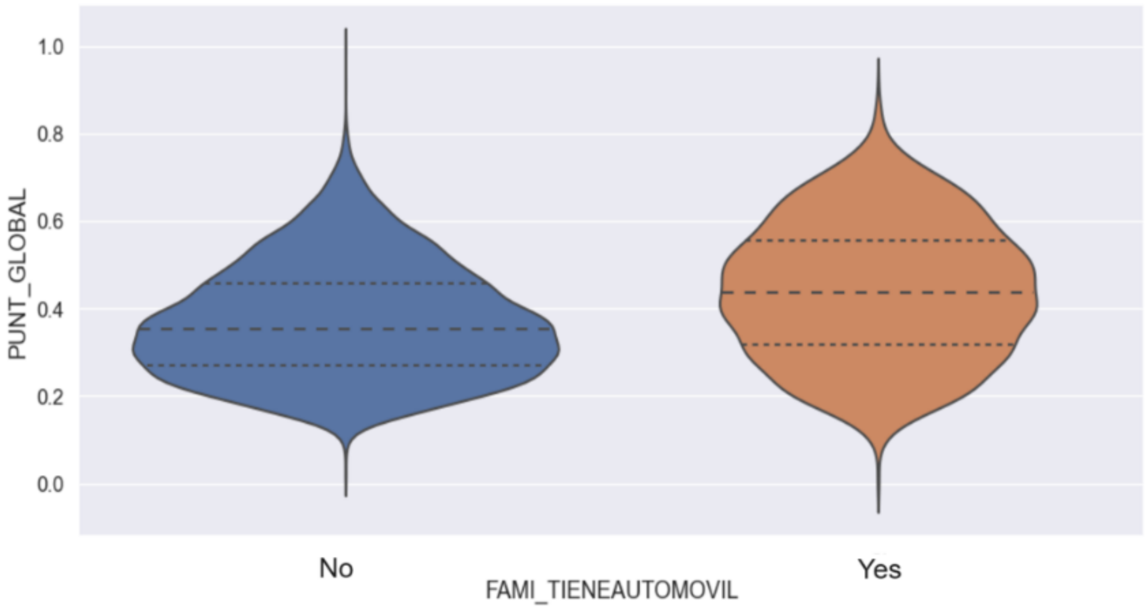


Source: Own elaboration

From Figure 9, it can be observed that although the median (0.4) of students who have internet is slightly higher than the median of students who do not have internet service (0.3), there is significant overlap between the violins, indicating that this attribute is not considered decisive in student performance.

Similarly, in Figure 10, the violin plot relates the variable FAMI_TIENEAUTOMOVIL to the normalized overall test score, where FAMI_TIENEAUTOMOVIL corresponds to whether the student's family has a car or not. From Figure 10, it can be observed that both the median of students whose families own a car and the median of students who do not have one are close to 0.4, with significant overlap between the violins of these categories. Therefore, this attribute is not considered decisive in student performance.

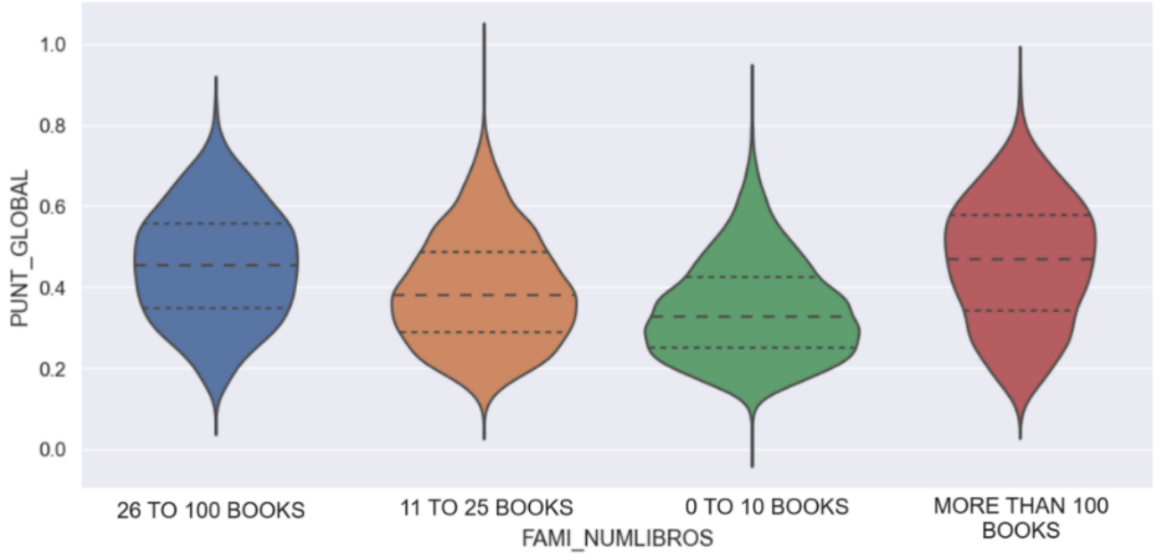
Figure 10. Violin plot for car availability and normalized overall score



Source: Own elaboration

On the other hand, Figure 11 presents a violin plot illustrating the relationship between the variable FAMI_NUMLIBROS and the normalized overall score of the Saber 11 test, where FAMI_NUMLIBROS represents the range of books owned by the student's family. It is possible to observe from Figure 11 that although there is overlap in the violins of the considered categories, families with 26 to 100 books and those with more than 100 books obtained a value higher than 0.4 for the median of the normalized overall score. Conversely, families of students who have between 0 and 10 books exhibit the lowest value in the median of the normalized overall score, indicating that this attribute can be considered a valid indicator of performance in the test.

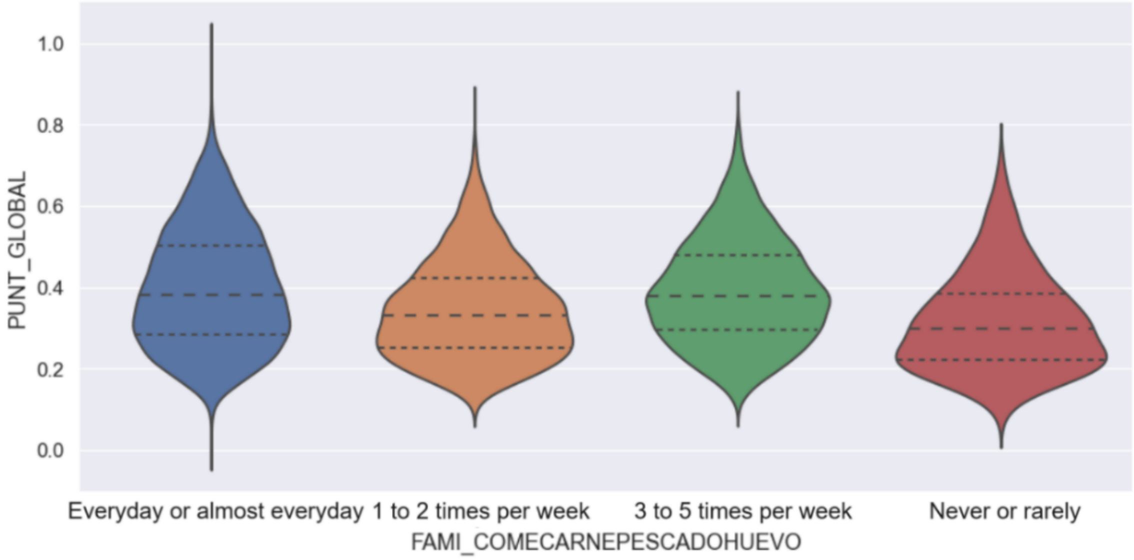
Figure 11. Violin plot for number of books and normalized overall score



Source: Own elaboration

Finally, regarding attributes involving the nutritional aspects of the student, Figure 12 presents a violin plot indicating the relationship between the variable FAMI_COMECARNEPESCADOHUEVO and the normalized overall score of the test, where FAMI_COMECARNEPESCADOHUEVO corresponds to the frequency with which meat, fish, or eggs are consumed in the student's family. According to the results in Figure 12, it is observed that in the case of students' families where meat, fish, or eggs are rarely consumed, the median obtained in the normalized overall score is the lowest, with a value close to 0.3. Similarly, although there is overlap in the other violins of the categories, it is noted that the categories associated with the consumption of meat, fish, or eggs daily or 3 times a week exhibit the highest median in the normalized overall score, with a value of 0.4.

Figure 12. Violin plot for meat/fish/egg consumption frequency and normalized overall score

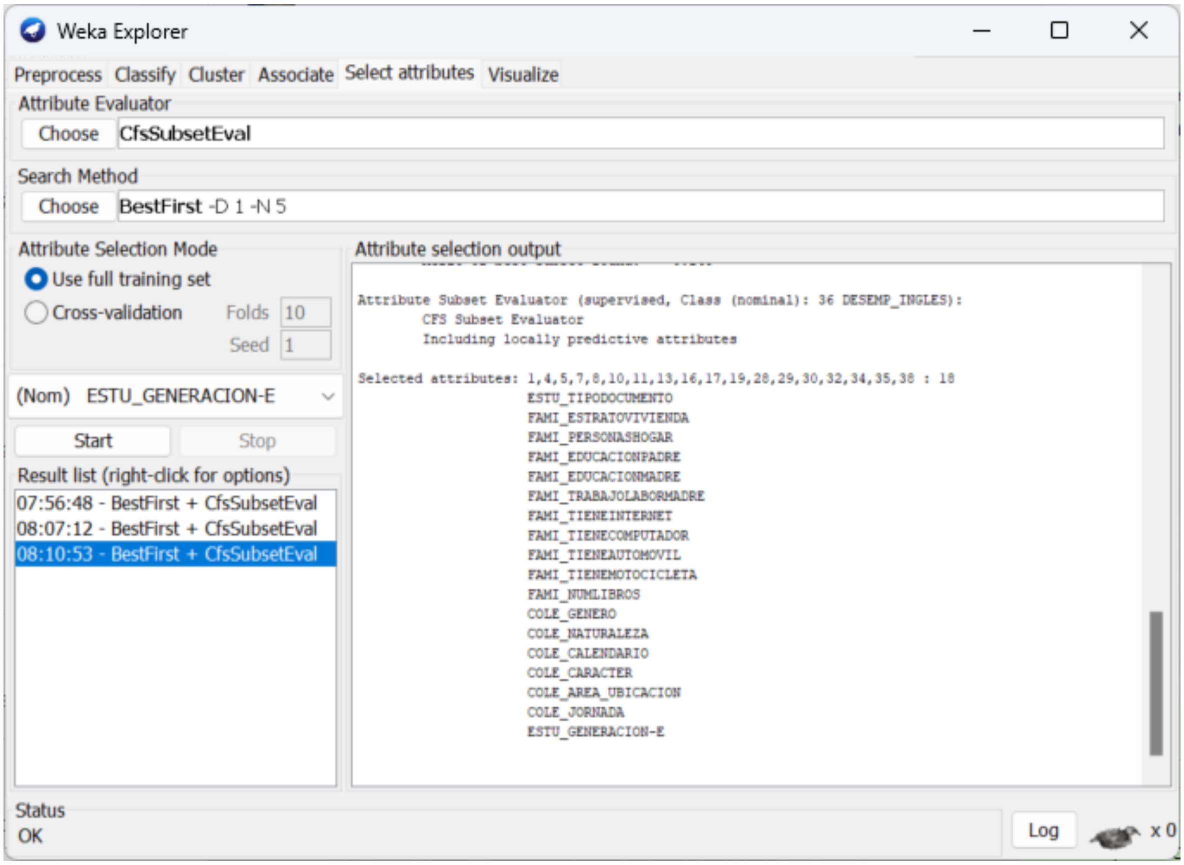


Source: Own elaboration

Next, to determine the best parameters influencing overall performance and performance in the different areas evaluated in the Saber 11 tests, the BestFirst attribute selection algorithm was applied. This algorithm is a heuristic method that explores the space of possible attribute subsets, continually evaluating the quality of each combination with respect to a predefined evaluation criterion. For the implementation of this method, the open-source software tool Weka was utilized.

The dataset was adapted from spreadsheet format to ARFF format. Thus, using Weka, the best attributes influencing performance in the Saber tests at both global and individual area levels were determined (see Figure 13). It is worth mentioning that attributes involving scores and performances in the different areas were discarded for the process so that these attributes can be used in the future as predictor attributes. Both continuous and remaining discrete attributes were considered for the process.

Figure 13. Application of the BestFirst algorithm in determining the optimal parameters



Attribute Evaluator
CfsSubsetEval

Search Method
BestFirst -D 1 -N 5

Attribute Selection Mode
● Use full training set
○ Cross-validation Folds 10 Seed 1

(Nom) ESTU_GENERACION-E

Result list (right-click for options)
07:56:48 - BestFirst + CfsSubsetEval
08:07:12 - BestFirst + CfsSubsetEval
08:10:53 - BestFirst + CfsSubsetEval

Attribute selection output

```
Attribute Subset Evaluator (supervised, Class (nominal): 36 DESEMP_INGLES):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,4,5,7,8,10,11,13,16,17,19,28,29,30,32,34,35,38 : 18
                     ESTU_TIPODOCUMENTO
                     FAMI_ESTRATOVIVIENDA
                     FAMI_PERSONASHOGAR
                     FAMI_EDUCACIONPADRE
                     FAMI_EDUCACIONMADRE
                     FAMI_TRABAJOLABORMADRE
                     FAMI_TIENEINTERNET
                     FAMI_TIENECOMPUTADOR
                     FAMI_TIENEAUTOMOVIL
                     FAMI_TIENEMOTOCICLETA
                     FAMI_NUMLIBROS
                     COLE_GENERO
                     COLE_NATURALEZA
                     COLE_CALENDARIO
                     COLE_CARACTER
                     COLE_AREA_UBICACION
                     COLE_JORNADA
                     ESTU_GENERACION-E
```

Status
OK

Source: Own elaboration

Table 3 presents the results obtained when applying the BestFirst algorithm to overall performance. Prior to this, the PERCENTIL_GLOBAL attribute was categorized into quartiles, creating a new attribute with categories named Q_GLOBAL.

Table 3. Best attributes identified for overall performance

Predictor attribute and metric	Attributes obtained
Q_GLOBAL	ESTU_TIPODOCUMENTO FAMI_EDUCACIONPADRE FAMI_EDUCACIONMADRE FAMI_NUMLIBROS COLE_NATURALEZA

Predictor attribute and metric	Attributes obtained
Merit of the subset obtained: 0.137	COLE_AREA_UBICACION COLE_JORNADA ESTU_INSE_INDIVIDUAL ESTU_NSE_ESTABLECIMIENTO ESTU_GENERACION-E

Source: Own elaboration

The results from Table 3 highlight the relevant attributes to be considered for implementing predictive models associated with the overall score. These attributes predominantly focus on economic indices, school characteristics (such as schedule, nature, location), parental education, the number of books available at home, the student's age, and benefits received through the GENERACION-E program. On the other hand, Table 4 presents the best attributes obtained by the BestFirst algorithm for the five areas evaluated in the Saber tests.

Table 4. Best attributes identified for the 5 evaluated areas

Predictor attribute and metric	Attributes obtained
DESEMP_LECTURA_CRITICA Merit of the subset obtained: 0.115	ESTU_TIPODOCUMENTO FAMI_EDUCACIONMADRE FAMI_NUMLIBROS ESTU_DEDICACIONLECTURADIARIA COLE_NATURALEZA COLE_AREA_UBICACION COLE_JORNADA ESTU_NSE_INDIVIDUAL ESTU_NSE_ESTABLECIMIENTO ESTU_GENERACION-E
	ESTU_TIPODOCUMENTO ESTU_GENERO FAMI_EDUCACIONMADRE

Predictor attribute and metric	Attributes obtained
DESEMP_MATEMÁTICAS Merit of the subset obtained: 0.122	FAMI_NUMLIBROS COLE_NATURALEZA COLE_AREA_UBICACION COLE_JORNADA ESTU_NSE_INDIVIDUAL ESTU_NSE_ESTABLECIMIENTO ESTU_GENERACION-E
DESEMP_CIENCIAS_NATURALES Merit of the subset obtained: 0.135	ESTU_TIPODOCUMENTO ESTU_GENERO FAMI_EDUCACIONMADRE FAMI_NUMLIBROS COLE_NATURALEZA COLE_AREA_UBICACION COLE_JORNADA ESTU_INSE_INDIVIDUAL ESTU_NSE_ESTABLECIMIENTO ESTU_GENERACION-E
DESEMP_SOCIALES_CIUDADANAS Merit of the subset obtained: 0.12	ESTU_TIPODOCUMENTO FAMI_EDUCACIONMADRE FAMI_NUMLIBROS ESTU_DEDICACIONLECTURADIARIA COLE_NATURALEZA COLE_AREA_UBICACION COLE_JORNADA ESTU_INSE_INDIVIDUAL ESTU_NSE_ESTABLECIMIENTO ESTU_GENERACION-E
	ESTU_TIPODOCUMENTO FAMI_ESTRATOVIVIENDA FAMI_EDUCACIONMADRE

Predictor attribute and metric	Attributes obtained
DESEMP_INGLES Merit of the subset obtained: 0.175	FAMI_NUMLIBROS COLE_NATURALEZA COLE_CALENDARIO COLE_JORNADA ESTU_INSE_INDIVIDUAL ESTU_NSE_INDIVIDUAL ESTU_NSE_ESTABLECIMIENTO ESTU_GENERACION-E

Source: Own elaboration

Based on the results obtained in Table 4, it is notable that for the critical reading and social sciences areas in relation to the overall test score, the attribute related to maternal education is more significant than that involving paternal education.

Similarly, in these cases, the attribute referring to the number of hours dedicated to daily reading (ESTU_DEDICACIONLECTURADIARIA) influences performance in these two areas. Regarding the mathematics and natural sciences areas concerning the overall test score, Table 4 allows for the identification that maternal education is more relevant than paternal education for predicting performance in these areas. On the other hand, in the case of the English area, it was determined that concerning the overall score, the attribute related to maternal education is more important than that addressing paternal education. Likewise, in this case, the attributes referring to the school's calendar (COLE_CALENDARIO) and the student's household stratum (FAMI_ESTRATOVIVIENDA) are crucial in predicting performance in these areas. It is worth mentioning that these two attributes (COLE_CALENDARIO and FAMI_ESTRATOVIVIENDA) do not appear as relevant in the other four analyzed areas. Finally, it is important to conclude that the set of attributes presenting a better metric is associated with the English area with a merit metric of 1.75, indicating that this set of attributes can contribute to the implementation of more efficient supervised learning models.

Discussion

In the discussion, it is important to mention that this study constitutes a significant contribution from an educational standpoint for the Caribbean region of Colombia, as there are no evident educational data mining researches focused on determining the socioeconomic or demographic factors affecting the performance of students taking the Saber 11 tests in the city of Cartagena. In this regard, the works proposed by Timarán-Pereira et al. (2019) and Timarán Buchely & Timarán Pereira (2023) focus on analyzing these factors at a national level for the Saber 11 and Saber Pro tests, proposing a predictive model based on decision trees. In contrast to these works, the present proposal utilizes the BestFirst method for selecting the best attributes influencing academic performance in the Saber 11 tests in Cartagena, such that these attributes serve as fundamental inputs for adjusting various supervised learning predictive models that may achieve higher precision in their metrics derived from the confusion matrix.

Conclusions

The exploratory data analysis allowed for significant conclusions regarding the impact on overall test performance. It was determined that parental postgraduate education leads to a marked difference in test results, with a higher median than other categories in the violin plot. Similarly, students who have more than 25 books at home exhibit a higher median in performance compared to other categories. Additionally, a slight difference in the median of the global score in the violin plots was observed among families that regularly consume meat, fish, or eggs.

The study conducted in this article leveraged the significant potential of open-source and free tools, demonstrating their great relevance and effectiveness for use in various application contexts, both in academia and the business environment. The libraries pandas, numpy, matplotlib, seaborn, and scikit-learn proved suitable for conducting studies based on exploratory data analysis.

Similarly, the free tool Weka, through the BestFirst algorithm, was valuable in identifying the best attributes influencing performance in the Saber tests at the global level and in each of the five test areas.

The application of the BestFirst algorithm for determining the relevant attributes impacting performance in each area revealed that, in general terms, maternal education, school characteristics, economic factors, and benefits received through the Generación-E program serve as indicators of performance. In the specific cases of critical reading and social sciences, the parameter associated with hours devoted to daily reading influences performance in these areas. For the English area, two additional attributes related to household stratum and school calendar are significant.

The set of attributes selected for the English area achieved a better merit metric than those obtained in the other four areas, suggesting promising outcomes when implementing supervised learning models to predict performance based on the attributes presented in Table 4. Unlike the other areas, English performance has more categories associated with classifications corresponding to international standards.

As future work stemming from this research, the intention is to evaluate various supervised learning models to determine, either through cross-validation or metrics from the confusion matrix, the best model for predicting performance in the Saber 11 tests. This will be done considering the different attributes selected and presented in Table 4.

Referencias

Acero, W., Sánchez, J. F., Suárez, D., & Téllez, C. (2016). Modelo de recalificación para la prueba Saber 11. *Comunicaciones En Estadística, 9*(1), 43–54.

Acevedo, D., Torres, J. D., & Jiménez, M. J. (2015). Factores asociados a la repetición de cursos y retraso en la graduación en programas de ingeniería de la Universidad de Cartagena, en Colombia. *Formación Universitaria, 8*(2), 35–42. https://doi.org/10.4067/S0718-50062015000200006

Alonso, J. C., Casasbuenas, P., Gallo, B., & Torres, G. (2012). Bilinguismo en Santiago de Cali: Análisis de los resultados de las Pruebas SABER 11 y SABER PRO. Universidad ICESI. https://www.icesi.edu.co/centros-academicos/images/Centros/cienfi/libros/Bilinguismo_en_Santiago_de_Cali.pdf

Arboleda-Posada, G. I., García-Arango, D. A., Vasco-Ospina, A. M., Garizabal, S. R., & Sastoque-Zapata, J. A. (2022). Saber Pro en programas de formación militar y policial: minería de datos para la identificación de factores asociados a los resultados. *Revista Ibérica de Sistemas e Tecnologias de Informação, E19*, 508–516.

Ayala-García, J., & Meisel-Roca, A. (2016). La exclusión en los tiempos del auge: el caso de Cartagena. https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/6947/dtser_246.pdf

Chanchí-Golondrino, G.-E., Ospino-Pinedo, M.-E., & Muñoz-Sanabria, L.-F. (2021). Application of Spatial Data Science on Results of the Saber 5 Test. *Revista Facultad de Ingeniería, 30*(58), e13823. https://doi.org/10.19053/01211129.v30.n58.2021.13823

Devasia, T., Vinushree T. P., & Hegde, V. (2016). Prediction of students performance using educational data mining. *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

Díaz Pinzón, J. E. (2020). Evaluación de la incidencia de un curso preicfes en los resultados de la prueba Saber 11. *Actualidades Pedagógicas, 1*(75), 33–52. https://doi.org/10.19052/ap.vol1.iss75.3

García-González, J. R., Sánchez-Sánchez, P. A., Orozco, M., & Obredor, S. (2019). Extracción de conocimiento para la predicción y análisis de los resultados de la prueba de calidad de la educación superior en Colombia. *Formación*

Universitaria, 12(4), 55–62. https://doi.org/10.4067/S0718-50062019000400055

Garizabalo Dávila, C. M. (2012). Estilos de aprendizaje en estudiantes de enfermería y su relación con el desempeño en las pruebas Saber Pro. *Revista Estilos de Aprendizaje, 9*(9), 1–18. https://redined.educacion.gob.es/xmlui/handle/11162/94536

Gorostiaga, A., & Rojo-Álvarez, J. L. (2016). On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain. *Neurocomputing, 171*, 625–637. https://doi.org/10.1016/j.neucom.2015.07.001

Iguarán Jiménez, A. M., Cabas-Manjarrés, M. F., Paba Barbosa, C., & Diazgranados Rincones, P. (2023). Relación de la prueba Saber 11, examen de admisión, promedio académico, prueba saber pro de estudiantes del programa de psicología de la Universidad del Magdalena. *Revista Digital de Investigación En Docencia Universitaria, 17*(2), e1421. https://doi.org/10.19083/ridu.2023.1421

Morales-Piñero, J. C., Cote-Sánchez, M. C., Molina-Bernal, I. A., & Rodríguez-Jerez, S. A. (2019). Incidencia de las TIC en el mejoramiento de las pruebas saber 11 a partir del modelo TPACK. *Encuentro Internacional de Educación En Ingeniería 2019*. https://acofipapers.org/index.php/eiei/article/view/40/35

[Narváez Zúñiga, A. F. (2022). Modelo estadístico para determinar los factores académicos en los resultados de las pruebas Saber Pro. *Investigación e Innovación En Ingenierías, 11*(1), 3–21. https://doi.org/10.17081/invinno.11.1.6255]

Nasiri, M., Minaei, B., & Vafaei, F. (2012). Predicting GPA and academic dismissal in LMS using educational data mining: A case mining. *6th National and 3rd International Conference of E-Learning and E-Teaching*, 53–58. https://doi.org/10.1109/ICELET.2012.6333365

Oviedo Carrascal, A. I., & Jiménez Giraldo, J. (2019). Minería de datos educativos: análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO. *Revista Politécnica, 15*(29), 128–140.

https://doi.org/10.33571/rpolitec.v15n29a10

Palacios-Gómez, H. J., Pantoja-Hernández, G. A., Navarro-Martínez, A. A., Puetaman, I. M. A., & Toledo Jimenez, R. A. (2016). Comparativa entre CRISP-DM y SEMMA para la limpieza de datos en productos MODIS en un estudio de cambio de cobertura y uso del suelo: Comparative between CRISP-DM and SEMMA for data cleaning of MODIS products in a study of land use and land cover change. *2016 IEEE 11th Colombian Computing Conference (CCC)*, 1–9. https://doi.org/10.1109/ColumbianCC.2016.7750789

Palacios-Mena, N. (2018). El currículo de ciencias sociales y las pruebas Saber 11 en Colombia: consonancias y disonancias. *Voces y Silencios. Revista Latinoamericana de Educación, 9*(2), 80–106. https://doi.org/10.18175/vys9.2.2018.06

Palacios-Mena, N., & Rodríguez-Márquez, M. A. (2019). Los resultados de la prueba Saber 11 de ciencias sociales y las opiniones de los estudiantes: convergencias y divergencias. *Revista Electrónica de Investigación Educativa, 21*, 1–17. https://doi.org/10.24320/redie.2019.21.e28.2116

Pathan, A. A., Hasan, M., Ahmed, M. F., & Farid, D. M. (2014). Educational data mining: A mining model for developing students' programming skills. *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)*, 1–5. https://doi.org/10.1109/SKIMA.2014.7083552

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Ruiz-Escorcia, R. R., Arévalo-Medrano, J. B., & Morillo, G.-P. (2017). Análisis de componentes principales aplicado a la prueba estatal Colombiana Saber 11. *Revista Espacios, 39*(10).

Sanabria James, L. A., Pérez Almagro, M. C., & Riascos Hinestroza, L. E. (2020). Pruebas de evaluación Saber y PISA en la educación obligatoria de

Colombia. *Educatio Siglo XXI, 38*(3 Nov-Feb), 231–254. https://doi.org/10.6018/educatio.452891

Tariq, H. I., Sohail, A., Aslam, U., & Batcha, N. K. (2019). Loan default prediction model using sample, explore, modify, model, and assess (SEMMA). *Journal of Computational and Theoretical Nanoscience, 16*(8), 3489–3503.

Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2019). 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*(2), 363–378. https://doi.org/10.19053/20278306.v9.n2.2019.9184

Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2023). Detección de patrones de desempeño académico en la competencia de matemáticas en las pruebas Saber 5o. *Revista Científica, 47*(2), 127–137. https://doi.org/10.14483/23448350.20908

Timarán Buchely, A., & Timarán Pereira, R. (2023). Minería de datos educativa para descubrir patrones asociados al desempeño académico en competencias genéricas. *Revista Colombiana de Tecnologías de Avanzada (RCTA), 2*(38), 87–95. https://doi.org/10.24054/rcta.v2i38.1282

Timarán Pereira, R., Hidalgo Troya, A., & Caicedo Zambrano, J. (2020). Factores asociados al desempeño académico en lectura crítica en las pruebas Saber 11° con árboles de decisión. *Investigación e Innovación En Ingenierías, 8*(3), 29–37. https://doi.org/10.17081/invinno.8.3.4701