Identification of factors that affect the academic performance of high school students in Peru through a machine learning algorithm

Lady Denisse Infante Acosta, Ing.[1], Jonatán Edward Rojas Polo, Mg.[1]
[1]Pontificia Universidad Católica del Perú, Perú, lady.infantea@pucp.edu.pe, jrojasp@pucp.pe

***Abstract*– The Peruvian Ministry of Education annually conducts the Student Census Evaluation (ECE, for its acronym in Spanish) to evaluate the level of learning achievement in the subjects of mathematics, reading and science and technology, both in public and private schools. The results are classified as Before beginning, Beginning, In process or Satisfactory. According to the results of the ECE 2019, it is observed that the academic performance achieved in the area of mathematics presents the highest percentage of students at the Satisfactory level (17.7%); however, in turn, said field of study is also the one that groups the highest percentage of students at the Before beginning level (33.0%).**

***Considering the aforementioned, this research aims to identify those variables that affect the learning achievements in mathematics of high school students. Thus, for the proposed analysis, a classification model was built for each of the mentioned levels, through an ensemble machine learning algorithm that uses the gradient boosting method. As a result of the modeling, the importance of the variables analyzed was obtained, which finally identified those that have greater relevance in the prediction of the classification of each level of learning achievement.**

***Keywords*-- Machine learning, classification model, student academic performance, educational system.**

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2021.1.1.68
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

Identification of factors that affect the academic performance of high school students in Peru through a machine learning algorithm

Lady Denisse Infante Acosta, Ing.[1], Jonatán Edward Rojas Polo, Mg.[1]
[1]Pontificia Universidad Católica del Perú, Perú, lady.infantea@pucp.edu.pe, jrojasp@pucp.pe

***Abstract*– The Peruvian Ministry of Education annually conducts the Student Census Evaluation (ECE, for its acronym in Spanish) to evaluate the level of learning achievement in the subjects of mathematics, reading and science and technology, both in public and private schools. The results are classified as Before beginning, Beginning, In process or Satisfactory. According to the results of the ECE 2019, it is observed that the academic performance achieved in the area of mathematics presents the highest percentage of students at the Satisfactory level (17.7%); however, in turn, said field of study is also the one that groups the highest percentage of students at the Before beginning level (33.0%).**

Considering the aforementioned, this research aims to identify those variables that affect the learning achievements in mathematics of high school students. Thus, for the proposed analysis, a classification model was built for each of the mentioned levels, through an ensemble machine learning algorithm that uses the gradient boosting method. As a result of the modeling, the importance of the variables analyzed was obtained, which finally identified those that have greater relevance in the prediction of the classification of each level of learning achievement.

***Keywords*-- Machine learning, classification model, student academic performance, educational system.**

I. Introduction

Education plays a fundamental role in the development of a country, as it is one of the most important instruments for reducing poverty and guaranteeing equal opportunities. However, the effort a country may make to provide children and adolescents with access to educational centers does not guarantee that the expected level of learning will be achieved [1]. As a result, countries have established different educational quality assessment systems aimed at groups of students at certain grades of schooling and focused on subjects such as reading comprehension, mathematics and science [2].

In Peru, the Ministry of Education, in collaboration with the Office of Quality Measurement of Learning (UMC, for its acronym in Spanish), measures the performance and learning achieved by elementary and high school students through census or sample evaluations. In the case of the Student Census Evaluation (ECE, for its acronym in Spanish), this is focused on obtaining information from all educational institutions and students from fourth grade of elementary school of Regular Basic Education (EBR, for its acronym in Spanish), fourth grade of Intercultural Bilingual Education (EIB, for its acronym in Spanish) and second grade of high school of EBR [3].

The assessment focused on second grade high school students was included as of 2015 and is aimed at measuring performance in the competencies of reading, mathematics and science and technology. The results obtained are classified according to one of the following four levels: Before beginning, Beginning, In process and Satisfactory. According to the last assessment ECE 2019, it is observed that the percentage of students who are part of the group qualified as Satisfactory is higher in mathematics with respect to the other two assessment areas; however, it is also in this competency where there is also the highest percentage of students whose results are qualified as Before beginning: 17.7% Satisfactory and 33.0% Before beginning in mathematics, 14.5% Satisfactory and 17.7% Before beginning in reading, and 9.7% Satisfactory and 10.1% Before beginning in science and technology [4].

Based on the preceding information, there is a clear indication that performance in mathematics is one of the topics where there are the most contrasting results. For that reason, the present research seeks to identify and analyze the factors that affect the academic performance in mathematics of high school students attending schools in the Lima Metropolitan Area. For the analysis in mention, basic student information such as gender and socioeconomic status will be considered, given their disposition through the ECE 2019; nevertheless, the focus will be mainly directed to the characteristics of the educational centers, such as location, management, percentage of permanence and progress of students, and the teaching environment. From this, a machine learning algorithm will be employed to analyze the ranking based on the four levels of learning achievement and identify those factors that have higher importance for each of these levels.

II. State of the art

Education

Academic performance and improvements in education are important research topics, since, by analyzing and identifying those factors that have a positive or negative impact, schools can take measures to improve the level of learning and the State can also generate public policies to reduce existing inequalities, as is clearly observed with the gap between schools in urban versus rural areas, private versus public schools, the distinction between different socioeconomic groups, or access to schools with adequate infrastructure and human capital [5].

There are different factors related to academic performance that can be grouped as factors associated with the student (gender, age, mother language, weight, height, and so forth), factors associated with family and home (socioeconomic status, access to basic services at home, educational level of parents, and so forth) and factors associated with the educational institution (title and years of experience of the teacher, multi-teaching institutions, internet access, condition of classrooms, and so forth). It is important to identify the factors that have the greatest effect on school performance, as this allows the definition of efficient public policies. Hence, if socioeconomic status is a relevant factor, the objective should focus on improving the conditions in which students live or improving the unsatisfied needs of households; on the contrary, in the case of factors associated with schools, the focus should be on identifying the attributes that contribute to students in one school having better academic performance than another school with students of similar characteristics [6].

As part of the public policies applied by the Peruvian State to the education sector between 2000 and 2015, strategies for the promotion of learning, public policies for teacher development and decentralized educational management have been established. These strategies have the common objective of improving the quality of education and reducing educational gaps, mainly the aspiration is to provide quality and equitable education; however, despite focusing efforts on this goal, the gaps are very marked and unfavorable in rural areas and among the extremely poor population. For this reason, part of the analysis to establish public policies is aimed at monitoring academic performance through national and international evaluations that finally allow the identification of those population groups where more resources and strategic programs are needed; in addition, with periodic evaluations, the results make it possible to measure the impact of the educational policies implemented in previous years [7].

Analytics

The application of machine learning techniques in the understanding of academic performance has provided a variety of effective tools that can help in the improvement of education in general, given its use in analytics from different approaches: descriptive, diagnostic, predictive and prescriptive. The main algorithms employed are the following: linear regression, logistic regression, k-nearest neighbor, decision tree (regressor and classification), random forest (regressor and classification), support vector machine (regressor and classification) and ensemble methods.

Classification algorithms allow predicting a discrete or categorical variable from the modeling of independent variables. One of the most commonly used machine learning algorithms to predict the class of an object are decision trees, which in essence are multilevel structures based on a hierarchy learned through model training and whose arrangement of variables from top to bottom is based on the best possible division of the available variables. This basic model can be scaled to a Random Forest model, which is an ensemble method consisting of a set of decision trees, each of which is trained using a subset of the total variables analyzed. This criterion allows having a wide choice of variants for the construction of each decision tree [8].

Additionally, in the search for improving the performance of the classification model, methods based on Boosting algorithms are used, which result from the combination of simple classifiers sequentially and take as a reference the accuracy obtained from the previous model to generate a more robust model, since special attention is paid to those observations that are still being misclassified, resulting in a final model with better predictive power and greater stability in the results [9]. A widely used optimization algorithm is the Gradient Descent, which is executed iteratively to find the optimal values of the parameters of each simple model in such a way that the cost function of the machine learning algorithm is minimized; in other words, it seeks to minimize the differences between the result obtained through the model and the real result [10].

III. Definition of Variables for the Analysis

In order to evaluate those variables that play an important role in the academic performance, it is necessary to include those related to the characteristics of each of the students surveyed, the attributes of the educational centers, the teaching environment, the percentage of permanence and progress of the students, the particularities of the geographic areas where the institutions are located, among others. In order to obtain the aforementioned data, three sources have been used: the Student Census Evaluation (ECE), the Educational Census and the Information System Support Management of the Educational Institution (SIAGIE, for its acronym in Spanish).

The information from the ECE 2019, obtained through the UMC, includes variables of census students, such as gender and socioeconomic index, which is calculated based on five indicators: years of schooling of parents, house building materials (walls, floors and roofs), access to basic services (electricity, water and sewage), possession of assets and other services at home (telephony, internet, etc.) [11]. Regarding the educational center, the census presents variables such as the area (urban or rural), the district, the management (state or non-state) and the Local Educational Management Unit (UGEL) to which it belongs.

Fig. 1 shows part of the diffusion campaign for the ECE 2019 in the different schools of the country. It indicates the scheduled dates and the subjects to be evaluated according to each grade of schooling.

Fig. 1 Dates of the ECE 2019
Source: UGEL (2019)

In addition, a variable has been included referring to the district area within the Lima Metropolitan Area to which each district belongs: Lima Norte (Ancón, Carabayllo, Comas, Independencia, Los Olivos, Puente Piedra, San Martín de Porres, Santa Rosa), Lima Este (Ate, Chaclacayo, Cieneguilla, El Agustino, Lurigancho, San Juan de Lurigancho, Santa Anita), Lima Sur (Chorrillos, Lurín, Pachacamac, Pucusana, Punta Hermosa, Punta Negra, San Bartolo, San Juan de Miraflores, Santa María del Mar, Villa el Salvador, Villa María del Triunfo), Lima Centro (Breña, La Victoria, Lima, Rímac, San Luis) and Lima Moderna (Barranco, Jesús María, La Molina, Lince, Magdalena del Mar, Miraflores, Pueblo Libre, San Borja, San Isidro, San Miguel, Santiago de Surco, Surquillo).

As for the Educational Census, it collects detailed information from educational institutions across the country on aspects such as enrolled students, levels of backwardness, promotion, repetition and dropout, number of teaching and administrative staff, and educational infrastructure [12]. From the information of the 2019 Educational Census, obtained through the Education Statistics Unit (ESCALE, for its acronym in Spanish), variables have been extracted per district referring to the permanence and progress of students in the second grade of high school (percentage of students passed, failed, repeated, withdrawn and with school backwardness), to the teaching environment in high school (certified teachers, ratio of students per teacher, ratio of students per computer, average class size, percentage of schools with internet access, percentage of educational institutions with EBR, with at least one student with Special Educational Needs (SEN) receiving the Service of Support and Advice of the Special Educational Needs (SAANEE, for its acronym in Spanish)).

Finally, with regard to SIAGIE information, obtained through the ESCALE, variables have been extracted per district referring to high school education in terms of interannual and permanent dropout rate, percentage of interannual transfer of educational service and interannual transfer from private to public, percentage of students who have connectivity at home, percentage of students who study in the same province where they were born and percentage of students whose mothers have completed higher university education.

Table I presents the dictionary of variables previously described. It also includes those that have been affected by transformations, as is the case of the categorical variables that have been worked as dummy variables (represented in values of 0 or 1, to indicate the absence or presence of a qualitative attribute, respectively).

IV. Application of the Classification Model

In order to identify the most important factors that affect the academic performance of high school students in the area of mathematics, the analysis is based on a classification model. For this purpose, the level of achievement obtained (Before beginning, Beginning, In process or Satisfactory) is considered as the target variable and, given the intention of knowing those variables that are determinant for each level, a model is generated for each of these.

As a first stage of the modeling, variables are constructed from the results of the ECE 2019 for the second grade of high school and these are filtered so that only the observations corresponding to the Lima Metropolitan Area are available. The variables obtained through the Educational Census and the SIAGIE are included in this base, and additionally, the variables referring to the UGEL code, the district code and the district area are added. Subsequently, the attributes considered as potential explanatory variables are selected and the results obtained for the mathematics assessment for each of the four levels of learning achievement are defined as the target variable. As a final part of this first stage, missing values are eliminated and the categorical variables of management, area, sex, district area and level of achievement are converted into dummy variables.

As a second stage of the modeling, a dataset is generated for each of the four levels of achievement and the following steps are applied for each of these. First, it is analyzed whether there is an imbalance between the 0 and 1 registers of the target variable and, given that this is indeed the case, the dataset is split into three parts: training, validation and testing. As a first step, the general dataset is split into an initial training set and a testing set; subsequently, the initial training set is balanced applying a data resampling technique and then split into a training set and a validation set. According to this approach, the model will be trained on a balanced set (training set) and the metrics will be validated on both a balanced set (validation set) and an unbalanced set (testing set), which will allow evaluating the performance of the model in these two scenarios.

TABLE I
DICTIONARY OF VARIABLES USED FOR THE CLASSIFICATION MODEL

Variable	Description
GROUP_M_BEFORE_BEGINNING	Dummy variable that represents whether the level of achievement of the student is qualified as Before beginning
GROUP_M_BEGINNING	Dummy variable that represents whether the level of achievement of the student is qualified as Beginning
GROUP_M_IN_PROCESS	Dummy variable that represents whether the level of achievement of the student is qualified as In process
GROUP_M_SATISFACTORY	Dummy variable that represents whether the level of achievement of the student is qualified as Satisfactory
AREA_RURAL	Dummy variable that represents whether the educational institution is located in a rural area
AREA_URBAN	Dummy variable that represents whether the educational institution is located in an urban area
COD_DISTRICT	Variable that identifies the district where the educational institution is located
COD_NAME_UGEL	Variable that identifies the UGEL to which the educational institution belongs
MANAGEMENT_STATE	Dummy variable that represents whether the educational institution is managed by the State
MANAGEMENT_NON_STATE	Dummy variable that represents whether the educational institution is not managed by the State
SES	Variable that represents the socioeconomic index of the student
PERC_INTERNET_ACCESS	Percentage of schools with internet access - high school (% of total per district)
PERC_STUDENTS_MOTHERS_UNIVERSITY	Percentage of students whose mothers have completed higher university education - high school (% of total per district)
PERC_STUDENTS_SAME_PROVINCE_BIRTHPLACE	Percentage of students who study in the same province in which they were born - high school (% of total per district)
PERC_PASSED_SECOND_GRADE	Percentage of students who passed - second grade of high school (% of final enrollment per district)
PERC_SCHOOL_BACKWARDNESS_SECOND_GRADE	Percentage of students with school backwardness - second grade of high school (% of initial enrollment per district)
PERC_FAILED_SECOND_GRADE	Percentage of students who failed - second grade of high school (% of final enrollment per district)
PERC_STUDENTS_SEN	Percentage of IE EBR with at least one student with SEN that receives SAANEE - second grade of high school (% of the total number of IE EBR with at least one student with SEN per district)
PERC_CONNECTIVITY_HOME	Percentage of high school students with connectivity at home (% of total per district)
PERC_REPEATED_SECOND_GRADE	Percentage of students who repeated - second grade of high school (% of initial enrollment per district)
PERC_WITHDREW_SECOND_GRADE	Percentage of students who withdrew - second grade of high school (% of initial enrollment per district)
PERC_ INTERANNUAL_TRANSFER_PRIVATE_PUBLIC	Percentage of interannual transfer from private to public education service - high school (% of final enrollment per district)
PERC _INTERANNUAL_TRANSFER_HIGH_SCHOOL	Percentage of interannual transfer of educational service - high school (% of final enrollment per district)
CERTFIIED_TEACHERS	Certified teachers - high school (% of total per district)
RATIO_STUDENTS_COMPUTER	Ratio of students per computer - high school (number of students per district)
RATIO_STUDENTS_TEACHER	Ratio of students per teacher - high school (number of students per district)
SEX_MALE	Dummy variable that represents whether the student is male or not
SEX_FEMALE	Dummy variable that represents whether the student is female or not
AVERAGE_CLASS_SIZE	Average class size - high school (number of students per district)
RATE_INTERANNUAL_DROPOUT_HIGH_SCHOOL	Interannual dropout rate in high school education (% of final enrollment per district)
RATE_ PERMANENT_DROPOUT_HIGH_SCHOOL	Permanent dropout rate in high school education (% of final enrollment per district)
DISTRICT_AREA_LIMA_CENTRO	Dummy variable that represents whether the educational institution is located in the district area of Lima Centro
DISTRICT_AREA_LIMA_ESTE	Dummy variable that represents whether the educational institution is located in the district area of Lima Este
DISTRICT_AREA_LIMA_MODERNA	Dummy variable that represents whether the educational institution is located in the district area of Lima Moderna
DISTRICT_AREA_LIMA_NORTE	Dummy variable that represents whether the educational institution is located in the district area of Lima Norte
DISTRICT_AREA_LIMA_SUR	Dummy variable that represents whether the educational institution is located in the district area of Lima Sur

As a third stage of the modeling, a multivariate analysis is performed in order to eliminate those variables that are highly correlated with each other and thus eliminate multicollinearity. For this purpose, the correlation between the potential explanatory variables is evaluated and, as a result, those pairs of variables that present a high correlation will be evaluated to determine which of them will be finally excluded from the modeling. To achieve this, the correlation between each variable and the target is determined, and it is finally decided to keep the variable with the highest correlation. It is important to mention that this stage is worked with the joint of the training set and the validation set and, once the final variables to be included in the modeling have been identified, they are selected in each of the set (training set, validation set and testing set).

As a fourth stage of the modeling, the baseline metrics against which the performance of each of the models will be compared are determined. For this purpose, the target variable of the initial training set (unbalanced) is taken as a reference and a random classification is simulated. This is done by means of a loop of 1000 iterations and the results are obtained for each of the metrics detailed below.

Metrics

Table II presents the confusion matrix and the related metrics used to evaluate how well the classification is performed through the trained model and thus determine whether the selected variables are indeed the ones that best explain academic performance.

TABLE II
CONFUSION MATRIX - PREDICTED VALUES VERSUS ACTUAL VALUES

		Actual Values	
		Positive (1)	**Negative (0)**
Predicted Values	**Positive (1)**	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

- **Accuracy**:

$$\frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision**:

$$\frac{TP}{TP + FP}$$

- **Recall**:

$$\frac{TP}{TP + FN}$$

- **F1 Score**:

$$2 * (\frac{Recall * Precision}{Recall + Precision})$$

Finally, as the last stage of the modeling, the training of the classification model is performed. For this, in the first instance, tests are performed with a basic model such as Logistic Regression, with which an overview of the importance of the explanatory variables and their performance in the model to explain the target is obtained. Likewise, a random forest model is considered to review the importance of the variables through an ensemble model; however, it is noted that the importance of the variables is not adequately distributed and only focuses on one variable (SES). Therefore, boosting algorithms such as the Gradient Boosting Classifier and the XGBoost are used, and it is observed that the latter presents a better performance in the evaluation metrics, as well as a better distribution of the importance of the variables. Based on the mentioned criteria, it is finally decided to use the XGBoost algorithm for the training of the models for each of the four levels of learning achievements.

Fig. 2 illustrates the intuition behind the boosting algorithm and Table III details the sequence of steps followed for the building of the classification models.

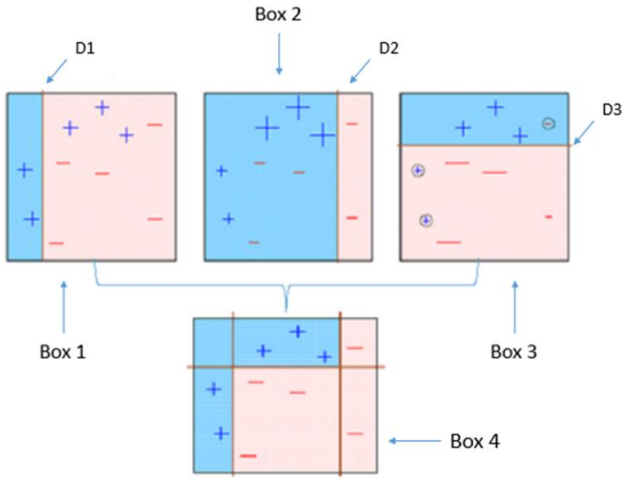

Fig. 2 Graphical representation of the boosting algorithm
Source: DataCamp (2019)
https://www.datacamp.com/community/tutorials/xgboost-in-python

TABLE III
PROCEDURE FOR BUILDING THE CLASSIFICATION MODEL FOR EACH LEVEL OF LEARNING ACHIEVEMENT

Step	Description	Operation
1	Reading of datasets	Read the files of the ECE_2019 dataset and the high school characteristics dataset
2	Formatting of variable names	Convert all variable names in uppercase letters Rename some of the variables, according to the convenience of the analysis
3	Formatting of variable contents	Convert all categorical variable observations to uppercase letters Convert variables that contain identification codes to string type Convert the socioeconomic index variable to float type Convert blank observations to NaN (Not a Number) values
4	Merging of datasets	Merge the ECE_2019 and high school characteristics datasets
5	Creation of the dataset for the geographic area of analysis	Select the observations whose value in the variable PROVINCE is equal to LIMA
6	Creation of additional variables for analysis	Extract the code of the variables DISTRICT and UGEL, and register them as new variables Create the variable DISTRICT_AREA based on the area where the district is located
7	Creation of the dataset for the subject of analysis	Select the observations of academic performance in mathematics for the Lima Metropolitan Area
8	Elimination of missing values	Delete rows (observations) that have at least one missing value in any column (variable)
9	Creation of dummy variables	Convert categorical variables into dummy variables Convert the variable GROUP_M into a dummy variable for each level of achievement
10	Creation of a dataset for each target variable	Create a dataset for each level of achievement Eliminate GROUP_M variables referring to achievement levels that do not correspond to the analysis model
11	Evaluation of dataset balancing	Verify if the dataset of the level of achievement analyzed is balanced
12	Splitting of the dataset into training and testing sets	Split the dataset into 80% for the initial training set and 20% for the testing set
13	Obtaining baseline metrics	Randomly classify each observation of the initial training set Evaluate the accuracy, precision, recall and f1 score between actual and random classification Iterate 1000 times the previous steps Obtain the average of the results obtained for each metric
14	Rebalancing of the training set	Split the initial training set into two sets according to the value designated in the target variable: 0 or 1 Obtain the number of observations for each set and identify the largest of these Balance the set that has the smallest number of observations through a resampling technique Join the training set after resampling
15	Application of the multivariate analysis	Obtain the correlation among potential explanatory variables Identify those variables that have a high correlation with each other Eliminate the variable that has the lowest correlation with the target variable
16	Identification of the final variables	Generate a list of final variables for each level of achievement after applying the multivariate analysis
17	Updating of datasets	Maintain in the initial training set and testing set only the final variables and the target variable
18	Splitting of the dataset into training and validation sets	Split the dataset into 80% for the initial training set and 20% for the validation set
19	Training and evaluation of classification models	Train the classification model (Logistic Regression, Random Forest, Gradient Boosting and XGBoost) Tune the parameters that correspond to each model Train the model with the training set Obtain the prediction of the classification for the validation set (balanced) Obtain the prediction of the classification for the testing set (unbalanced) Evaluate the classification results Evaluate the metrics between the actual classification and the one obtained with the model Compare results with baseline metrics Evaluate the distribution of the importance of the explanatory variables Iterate until finding optimal values for the metrics and for the distribution of the importance of the variables
20	Selection of classification model	Choose the model and parameters that generate the optimal values for classification Generate the final report of the metrics obtained Generate the report of the importance of the explanatory variables

Parameters

In order to obtain the model that best fits the needs of the object of analysis, the following parameters available for the XGBoost algorithm have been used, whose descriptions are detailed according to the documentation available for the application of the model with the Scikit-Learn package [13].

- **n_estimators**: Number of gradient boosted trees. Equivalent to number of boosting rounds
- **max_depth**: Maximum tree depth for base learners
- **learning_rate**: Boosting learning rate
- **objective**: Specify the learning task and the corresponding learning objective or a custom objective function to be used
- **booster**: Specify which booster to use (gbtree, gblinear or dart)
- **colsample_bytree**: Subsample ratio of columns when constructing each tree
- **reg_alpha**: L1 regularization term on weights
- **reg_lambda**: L2 regularization term on weights
- **subsample**: Subsample ratio of the training instance
- **num_parallel_tree**: Used for boosting random forest
- **random_state**: Random number seed

The optimization of the model through the adjustment of these parameters provides two main advantages: avoiding over-fitting in the training of the model and adequately distributing the importance of the variables. The first of these is achieved through parameters of regularization such as reg_alpha and reg_lambda, while the second one is achieved through the adjustment of parameters such as n_estimators, max_depth, colsample_bytree and subsample, which guarantees that all the variables have the option of being considered in the building of the decision trees. In addition to the improvement in model performance, there is also an improvement in computational performance during the training of the model, given that the algorithm allows the building of decision trees in parallel and the pruning of them using a depth-first approach.

V. Results

The classification models for the analysis of the four levels of learning achievement have been trained using the XGBoost algorithm, for which the parameters related to the model have been tuned through the iterative training process with different values. Table IV presents the final values of the parameters that have allowed to build the classification models with the best performance in the evaluation metrics.

TABLE IV
OPTIMAL VALUES OF THE PARAMETERS USED FOR TRAINING THE MODELS

Before beginning	
Parameter	**Value**
n_estimators	500
max_depth	20
learning_rate	0.01
objective	reg:logistic
booster	gbtree
colsample_bytree	0.3
reg_alpha	400
reg_lambda	200
subsample	0.5
num_parallel_tree	2
random_state	0

Beginning	
Parameter	**Value**
n_estimators	400
max_depth	20
learning_rate	0.01
objective	reg:logistic
booster	gbtree
colsample_bytree	0.5
reg_alpha	300
reg_lambda	500
subsample	0.5
num_parallel_tree	2
random_state	0

In process	
Parameter	**Value**
n_estimators	500
max_depth	10
learning_rate	0.01
objective	reg:logistic
booster	gbtree
colsample_bytree	0.3
reg_alpha	50
reg_lambda	25
subsample	0.5
num_parallel_tree	2
random_state	0

Satisfactory	
Parameter	**Value**
n_estimators	500
max_depth	10
learning_rate	0.03
objective	reg:logistic
booster	gbtree
colsample_bytree	0.5
reg_alpha	2100
reg_lambda	1200
subsample	0.5
num_parallel_tree	2
random_state	0

Based on the training of the models and the classifications obtained from them, a comparison was made between the actual classifications and those obtained from the model, both for the validation (balanced) and the testing (unbalanced) bases.

Table V shows the values obtained for the metrics evaluated. It is important to mention that the results to which most attention should be paid are those obtained from the unbalanced set, since this simulates a real scenario; moreover, given this particularity, the analysis is mainly focused on reviewing the performance obtained in the recall metric, since this indicator allows evaluating how many observations have been correctly classified with respect to the total number of observations that actually belong to that classification.

TABLE V
PERFORMANCE OF THE CLASSIFICATION MODELS

Before beginning					
Baseline (unbalanced)		**Validation (balanced)**		**Testing (unbalanced)**	
accuracy	0.670244	accuracy	0.590594	accuracy	0.532282
precision	0.208111	precision	0.576273	precision	0.260617
recall	0.208208	recall	0.691987	recall	0.685769
f1 score	0.208158	f1 score	0.628851	f1 score	0.377696

Beginning					
Baseline (unbalanced)		**Validation (balanced)**		**Testing (unbalanced)**	
accuracy	0.562625	accuracy	0.546855	accuracy	0.501838
precision	0.323184	precision	0.537275	precision	0.355769
recall	0.323169	recall	0.676546	recall	0.681163
f1 score	0.323175	f1 score	0.598921	f1 score	0.467411

In process					
Baseline (unbalanced)		**Validation (balanced)**		**Testing (unbalanced)**	
accuracy	0.662723	accuracy	0.529377	accuracy	0.505239
precision	0.214850	precision	0.528530	precision	0.232236
recall	0.214849	recall	0.577278	recall	0.564466
f1 score	0.214847	f1 score	0.551830	f1 score	0.329080

Satisfactory					
Baseline (unbalanced)		**Validation (balanced)**		**Testing (unbalanced)**	
accuracy	0.621145	accuracy	0.609115	accuracy	0.624600
precision	0.253979	precision	0.615792	precision	0.356354
recall	0.253951	recall	0.578560	recall	0.570418
f1 score	0.253963	f1 score	0.596596	f1 score	0.438664

According to the results, it is noted that there is a good performance of the models when comparing the baseline value of the metrics with the value obtained from the classification generated with the models. Mainly an outstanding result is observed in the recall metric, since it is in this metric where the tuning of the parameter values has been focused in order to obtain the best performance; however, in general, it is also noticed that the other metrics are close and better, in certain cases, to the baseline values, which is convenient given the nature of an unbalanced database.

Table VI shows the importance of the variables in each classification model, obtained through the XGBoost algorithm. This ranking of the variables is used to identify those factors that have the greatest impact on academic performance and that are relevant for a student to accomplish a specific level of learning achievement.

In the case of the two lowest levels of the evaluation scale (Before beginning and Beginning), it is observed that a relevant aspect associated with academic performance is the management of the educational institution, specifically if it belongs to the State. In addition, other variables that also play an important role are the socioeconomic status and whether the school belongs to the district area of Lima Moderna, both of which are closely related to the socioeconomic status of the student. Furthermore, the availability of connectivity at home can also be included, since internet access is related to the household economic status. With regard to continuity of studies, two noteworthy variables are the interannual dropout rate and the percentage of students who repeated, which illustrate that the interruption of the school year or the discontinuity in the next school year affects the learning achievement. Finally, a variable that stands out for its inherent characteristic is the sex of the student, given that being a woman would presumably increase the probability of being classified in one of the lowest levels.

In the case of the In process level, the non-state management of an educational institution represents an important attribute to be classified in this category. Similar to the lower levels, the socioeconomic status of the student is relevant through aspects such as the socioeconomic index, the district area where the school is located and the connectivity at home. An attribute that becomes relevant is the percentage of students whose mothers have completed higher university education, which is justified since it is the mother who is mainly in charge of the organization of household chores [14]. Finally, it is noted that the sex of the student is a determining factor in this level of achievement, which would be associated with the idea that being male increases the probability of belonging to this level of achievement.

In the case of the Satisfactory level, the most important variable in the model is the non-state management associated with an educational institution. This attribute is also related to the UGEL, another of the most important variables, given that it implements educational policies within its jurisdiction [15]. In line with these two factors, another indicator also associated with academic performance is the ratio of students per teacher, which is managed and planned by each educational institution. Finally, the socioeconomic status continues to be a relevant attribute in the level of learning achieved, through variables such as the socioeconomic index, the district area and the connectivity at home.

TABLE VI
IMPORTANCE OF THE VARIABLES FOR EACH CLASSIFICATION MODEL

Before beginning	
Variable	**Importance**
MANAGEMENT_STATE	0.248356
SES	0.154625
DISTRICT_AREA_LIMA_MODERNA	0.114626
RATE_INTERANNUAL_DROPOUT_HIGH_SCHOOL	0.090515
PERC_CONNECTIVITY_HOME	0.065721
SEX_FEMALE	0.060000
PERC_FAILED_SECOND_GRADE	0.037383
PERC_PASSED_SECOND_GRADE	0.034928
PERC_WITHDREW_SECOND_GRADE	0.025863
PERC_INTERNET_ACCESS	0.021316
PERC_ INTERANNUAL_TRANSFER_PRIVATE_PUBLIC	0.019816
DISTRICT_AREA_LIMA_NORTE	0.019646
DISTRICT_AREA_LIMA_ESTE	0.016459
PERC_SCHOOL_BACKWARDNESS_SECOND_GRADE	0.014925
PERC_REPEATED_SECOND_GRADE	0.014193

Beginning	
Variable	**Importance**
MANAGEMENT_STATE	0.300015
DISTRICT_AREA_LIMA MODERNA	0.120928
SES	0.079138
PERC_CONNECTIVITY_HOME	0.077270
PERC_REPEATED_SECOND_GRADE	0.043631
SEX_FEMALE	0.041815
PERC_FAILED_SECOND_GRADE	0.040587
RATE_INTERANNUAL_DROPOUT_HIGH_SCHOOL	0.035132
PERC_ INTERANNUAL_TRANSFER_PRIVATE_PUBLIC	0.035008
COD_NAME_UGEL	0.032131
PERC _INTERANNUAL_TRANSFER_HIGH_SCHOOL	0.027981
DISTRICT_AREA_LIMA_ESTE	0.020167
RATIO_STUDENTS_TEACHER	0.019399
PERC_SCHOOL_BACKWARDNESS_SECOND_GRADE	0.019383
DISTRICT_AREA_LIMA_NORTE	0.018404

In process	
Variable	**Importance**
MANAGEMENT_NON_STATE	0.118774
SEX_MALE	0.114677
PERC_STUDENTS_MOTHERS_UNIVERSITY	0.056283
DISTRICT_AREA_LIMA_MODERNA	0.051359
SES	0.050588
PERC_CONNECTIVITY_HOME	0.046635
PERC_FAILED_SECOND_GRADE	0.044220
COD_NAME_UGEL	0.041320
PERC_PASSED_SECOND_GRADE	0.039840
PERC_SCHOOL_BACKWARDNESS_SECOND_GRADE	0.038092
PERC_WITHDREW_SECOND_GRADE	0.034513
PERC_INTERNET_ACCESS	0.033240
PERC_REPEATED_SECOND_GRADE	0.032863
PERC _INTERANNUAL_TRANSFER_HIGH_SCHOOL	0.031291
DISTRICT_AREA_LIMA_SUR	0.028541

Satisfactory	
Variable	**Importance**
MANAGEMENT_NON_STATE	0.245174
RATIO_STUDENTS_TEACHER	0.118520
COD_NAME_UGEL	0.103869
PERC _INTERANNUAL_TRANSFER_HIGH_SCHOOL	0.097490
SES	0.088697
DISTRICT_AREA_LIMA_MODERNA	0.072122
PERC_CONNECTIVITY_HOME	0.042798
PERC_STUDENTS_MOTHERS_UNIVERSITY	0.038808
PERC_ INTERANNUAL_TRANSFER_PRIVATE_PUBLIC	0.035496
PERC_REPEATED_SECOND_GRADE	0.034431
COD_DISTRICT	0.025960
PERC_STUDENTS_SEN	0.019947
RATE_ PERMANENT_DROPOUT_HIGH_SCHOOL	0.018753
PERC_FAILED_SECOND_GRADE	0.012467
PERC_SCHOOL_BACKWARDNESS_SECOND_GRADE	0.011875

VI. Conclusions and recommendations

Through the analysis of the factors that influence the academic performance in the subject of mathematics of high school students residing in the Lima Metropolitan Area, two main characteristics have been identified that represent the most important variables in the level of learning achievement: management and socioeconomic status.

On the one hand, state management has a high impact on academic performance and it is associated with the lowest levels of the scale (Before beginning and Beginning), while non-state management registers a greater correspondence for the highest levels (In process and Satisfactory). Given that this is the most important variable in the four classification models and that low performance is mainly associated with those managed by the State, it is important to promote equal conditions with public education policies aimed at reducing the gaps in aspects related to access to education, infrastructure, access to information and communication technologies, and quality of teaching.

On the other hand, socioeconomic status is another variable that is present as a feature of great importance in the models. Its influence is observed in variables such as the socioeconomic index, the district area, internet access and the percentage of students whose mothers have completed higher university education. It is inferred that a deficiency in these attributes most likely causes a negative impact on the academic performance of students, which is mainly reflected in the first three levels (Before beginning, Beginning and In process). In conclusion, the socioeconomic status is one of the own characteristics of the student which is highly relevant to academic performance.

The scope of this research has been limited to the subject of mathematics and for students of the Lima Metropolitan Area; however, it is important to extend this analysis to the other areas of knowledge (reading and science and technology), since this will allow the identification of specificities according to each case and the evaluation of which variables have a greater impact in each of these subjects. It is also considered relevant to analyze other geographical areas, given that aspects such as management and socioeconomic status could play an even more fundamental role in the academic performance of students.

References

[1] The World Bank. (2020). Education - Overview. https://www.worldbank.org/en/topic/education/overview

[2] Martínez, R. (2008). La evaluación de aprendizajes en América Latina. México: Instituto Nacional para la Evaluación de la Educación. https://www.inee.edu.mx/wp-content/uploads/2019/01/P1C140.pdf

[3] Ministerio de Educación del Perú. (2020). Evaluaciones censales. http://umc.minedu.gob.pe/evaluaciones-censales/

[4] Ministerio de Educación del Perú. (2019). ¿Qué aprendizajes logran nuestros estudiantes? - Resultados de las evaluaciones nacionales de logros de aprendizaje 2019. http://umc.minedu.gob.pe/wp-content/uploads/2020/06/Reporte-Nacional-2019.pdf

[5] Beltrán, A. & Seinfeld, J. (2013). La trampa educativa en el Perú: Cuando la educación llega a muchos pero sirve a pocos. Lima: Universidad del Pacífico. https://repositorio.up.edu.pe/bitstream/handle/11354/1419/TrampaeducativaBeltranArlette2013.pdf

[6] Asencios, R. (2016). Rendimiento escolar en el Perú: Análisis secuencial de los resultados de la Evaluación Censal de Estudiantes. Lima: Banco Central de Reserva del Perú. https://www.bcrp.gob.pe/docs/Publicaciones/Documentos-de-Trabajo/2016/documento-de-trabajo-05-2016.pdf

[7] United Nations Educational, Scientific and Cultural Organization. (2017). Revisión de las políticas educativas 2000-2015. Continuidades en las políticas públicas en educación en Perú: aprendizajes, docentes y gestión descentralizada. https://unesdoc.unesco.org/ark:/48223/pf0000249171

[8] Uskov, V., Bakken, J., Byerly, A. & Shah, A. (2019). Machine Learning-based Predictive Analytics of Student Academic Performance in STEM Education. 2019 IEEE Global Engineering Education Conference (EDUCON), Dubai, United Arab Emirates, pages 1370-1376. https://ieeexplore.ieee.org/document/8725237

[9] Ferreira, A. & Figueiredo, M. (2012). Boosting Algorithms: A Review of Methods, Theory, and Applications. In: Zhang C., Ma Y. (eds) Ensemble Machine Learning. Springer, Boston, MA., pages 35-85. https://link.springer.com/chapter/10.1007/978-1-4419-9326-7_2

[10] Bisong, E. (2019). Optimization for Machine Learning: Gradient Descent. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA., pages 203-207. https://link.springer.com/chapter/10.1007/978-1-4842-4470-8_16

[11] Ministerio de Educación del Perú. (2019). Evaluaciones de logros de aprendizaje - Resultados 2019. http://umc.minedu.gob.pe/wp-content/uploads/2020/06/PPT-web-2019-15.06.19.pdf

[12] Ministerio de Educación del Perú. (2020). ¿Qué es el Censo Educativo? http://escale.minedu.gob.pe/censo-escolar-eol/

[13] XGBoost Developers. (2020). XGBoost Python Package - Python API Reference. https://xgboost.readthedocs.io/en/latest/python/python_api.html#module-xgboost.sklearn

[14] Inter-American Development Bank. (2007). Outsiders? The Changing Patterns of Exclusion in Latin America and the Caribbean. Washington: Inter-American Development Bank. https://publications.iadb.org/publications/english/document/Outsiders-The-Changing-Patterns-of-Exclusion-in-Latin-America-and-the-Caribbean.pdf

[15] Dirección Regional de Educación de Lima Metropolitana. (n.d.). Funciones. https://www.drelm.gob.pe/drelm/funciones/