Predicting undergraduate academic performance in a leading Peruvian university: A machine learning approach

FABIO SALAS*
JOSUÉ CALDAS**
Pontificia Universidad Católica del Perú - Perú

Recibido el 29-10-23; primera evaluación el 16-01-24;
aceptado el 17-02-24

Abstract

Despite improved higher education accessibility in low and middle-income countries (LMICs), challenges persist in student drop-out, especially for socio-economically disadvantaged students. While machine learning models have enhanced our understanding of this challenge by predicting academic performance, many studies overlook LMIC-specific institutional factors or focus on specific courses, limiting their generalizability and policy uses. To address these issues, the authors compiled a comprehensive database using administrative and census data to predict undergraduate academic performance at the Pontifical Catholic University of Peru (PUCP). The study found that the most effective models were tree-based ensembles, particularly Random Forest, with key predictors including prior secondary school performance and university admission test scores. They present a high-performing model using only ten features that can predict future academic performance and potentially aid in reducing student drop-out at PUCP.

Keywords: Academic performance, Machine Learning, Higher Education, Peru.

* I am a Peruvian economist specialized in the fields of development and education. Currently, I work as a junior researcher at the Institute of Human Development of Latin America (IDHAL-PUCP), focusing on studying poverty in the region through the lens of the capabilities approach. Additionally, I serve as data analyst at the Central Admission Office (OCAI-PUCP), leveraging data insights to shape and inform admission policies at PUCP. E-mail: fabio.salas@pucp.edu.pe ORCID: https://orcid.org/0000-0003-1667-6577

** I hold a Bachelor's Degree in Political Science and Government at Pontifical Catholic University of Peru. I specialize in data analysis with a focus on public policy and development economics. Currently, I am engaged as a Research Assistant at the Artificial Intelligence and Computational Methods Laboratory (QLAB-PUCP), where I leverage data science methodologies to inform and enhance public policy applications. E-mail: josue.caldas@pucp.edu.pe ORCID: https://orcid.org/0009-0002-2056-7664

Prediciendo el rendimiento académico de estudiantes de pregrado en una universidad destacada de Perú: Una aproximación con herramientas de Machine Learning

Resumen

Aunque la accesibilidad a la educación superior ha mejorado en países de renta baja y media (PRMB), persiste el abandono, especialmente entre estudiantes socioeconómicamente desfavorecidos. A pesar de los avances en modelos de *Machine Learning* para entender este desafío, muchos estudios descuidan factores institucionales específicos de los PRMB o se centran en cursos específicos, limitando su aplicabilidad y relevancia política. Para abordar esto, creamos una base de datos usando registros administrativos y censales para predecir el rendimiento académico en la Pontificia Universidad Católica del Perú (PUCP). Los modelos más efectivos, entre ellos *Random Forest*, destacaron predictores como el rendimiento previo y puntuaciones en pruebas de admisión. Presentamos un modelo eficiente con diez características que puede predecir el rendimiento futuro y así aportar a la reducción de la deserción en PUCP.

Palabras clave: Rendimiento Académico, Machine Learning, Educación Superior, Perú

Prevendo o desempenho acadêmico de estudantes de graduação em uma universidade destacada do Peru: Uma abordagem com ferramentas de Machine Learning

Resumo

Embora a acessibilidade ao ensino superior tenha melhorado em países de baixa e média renda (PBMR), a evasão persiste, especialmente entre estudantes socioeconomicamente desfavorecidos. Apesar dos avanços em modelos de *Machine Learning* para compreender esse desafio, muitos estudos negligenciam fatores institucionais específicos dos PBMR ou se concentram em cursos específicos, limitando sua aplicabilidade e relevância política. Para abordar isso, criamos uma base de dados usando registros administrativos e censitários para prever o desempenho acadêmico na Pontifícia Universidade Católica do Peru (PUCP). Os modelos mais eficazes, incluindo o *Random Forest*, destacaram preditores como desempenho prévio e pontuações em testes de admissão. Apresentamos um modelo eficiente com dez características que pode prever o desempenho futuro e assim contribuir para a redução da evasão na PUCP.

Palavras-chave: Desempenho Acadêmico, Aprendizado de Máquina, Ensino Superior, Peru

1. Introduction

Access to higher education in low- and middle-income countries (LMICs) has increased in the past decades following a combination of a rising number of secondary school leavers and a renewed family demand for social progress and opportunities (Romero, 2021). Yet increasing access to higher education institutions has not necessarily translated into higher retention and graduation of university students (Schendel & McCowan, 2016; Salas-Pilco & Yang, 2022), especially among students with a low socio-economic status (Balán, 2020). At the same time, the educational landscape in LMICs has been affected by an increasing adoption of artificial intelligence tools to learn about patterns in students' behavior (Salas-Pilco & Yang, 2022). Driven by the adoption of these tools and the availability of new educational datasets, an applied literature has also developed to predict the academic performance of undergraduate students based on machine learning (ML) algorithms (Albreiki et al., 2021).

Predicting undergraduate students' academic performance is a relevant academic and administrative task on three grounds. First, by knowing which students face the highest risk of failing a course, faculty and administrative workers can take specific preventive measures and improve their students' academic achievements (Moreno-Ger & Burgos, 2021). Second, predicting achievement is a task tightly bound to awarding scholarships or delivering tuition discounts (Hajar et al., 2022). Universities often select scholarship recipients based on their expectation of future academic performance in university. Third, predicting academic performance also sheds light on the most important factors for academic success, and can therefore inform selection criteria for the admission of new student cohorts to university (Liu et al., 2023; Niri, 2021).

However, higher education systems in LMICs encounter challenges when attempting to predict undergraduate academic performance. These challenges primarily stem from the limited information available, which often rely heavily on entrance tests that emphasize rote memorization (Romero, 2021; De los Rios, 2023), and from the problem of significant secondary school quality heterogeneity (Andrabi et al., 2022; LBDEAC, 2020). In this context, applied research predicting academic performance in LMICs has one of the following limitations: (1) it focuses on specific programs or courses, reducing the number of observations and affecting generalizability; (2) it lacks historical academic performance data or uses non-comparable information due to secondary school quality heterogeneity; (3) it predicts performance at a single point in time (using cross-sectional data), potentially biasing the view of university performance.

The main objective of this study is twofold: (i) to identify the most accurate ML algorithm for predicting the academic performance of undergraduate students at the Pontifical Catholic University of Peru (PUCP), and (ii) to determine the key predictors of students' academic performance. To address the limitations mentioned above, the authors analyzed three cohorts of students encompassing all academic programs and the two primary admission channels at PUCP. They ensured the comparability of academic information during the secondary school stage by leveraging the results of a national standardized assessment. Additionally, they employed a panel database that incorporates academic and sociodemographic information spanning students' secondary school years, the admission period, and their first two years in university.

2. Theoretical framework

This section first briefly describes the key characteristics of the Peruvian context that permeate the nature of the study. It goes on to discuss the common definition of academic performance in other studies within the field of learning analytics and points out why it is relevant to ground the definition on institutional and contextual aspects. It then presents an enhanced framework for predicting undergraduate academic performance in LMICs with compulsory university admission tests.

2.1. The Peruvian higher education system

In Peru, around 80 percent of higher education enrollment is in private institutions, a result both of Law No.882 (passed in the 1990s) geared to promoting private investment in education, and of the historical neglect of public higher education. In 2014, the Peruvian government became more involved in the higher education system and introduced a licensing process for institutions to meet basic quality standards. Only 32.7 percent of for-profit private institutions secured a license (Benites, 2021).

In this context, a distinct group of Peruvian universities stands out, due to their commitment to teaching standards, quality infrastructure, and strict admission criteria. PUCP is a private university, is well-regarded in various rankings (QS WUR; 2023; SIR, 2023) and offers a diverse range of academic programs. PUCP employs multiple admission channels, notably the Upper Third (ITS) and First Option (PO), which require admission tests in mathematics, language, and writing, together with five years of secondary education grades.

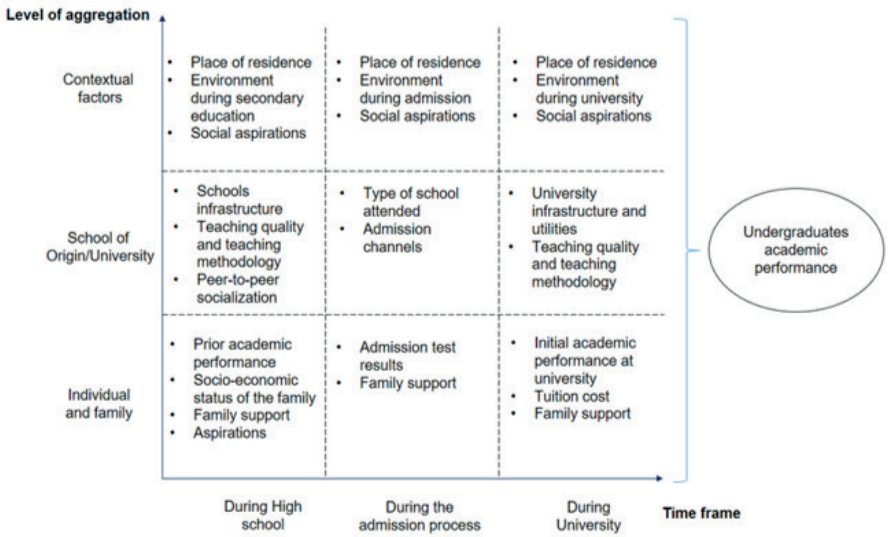
2.2. Learning analytics and academic performance prediction

Learning analytics (LA) aims to provide practical insights for educators and institutions to improve teaching and learning processes, often emphasizing efficient resource allocation and enhancing students' academic achievements in higher education (Leitner et al., 2017; Peña-Ayala et al., 2017). Employing ML techniques, LA focuses on predicting academic performance or drop-out risks, and classifying units based on available data (Susnjak, 2023; Athey & Imbens, 2019). The primary concern when employing a ML model centers on its out-of-sample performance, which involves gauging the model's ability to predict a target variable accurately when presented with new datasets (Athey & Imbens, 2019).

In applied LA research, the precise definition of "academic performance" is often overlooked, leading to its interchangeable use with "academic achievement" (Alyahyan & Düştegör, 2020). Although these concepts are interconnected, it is necessary to have a clear understanding of academic performance to address academic achievement (York et al., 2015). The former – depending on the data available – can be measured throughout the educational experience, whereas the latter usually correspond to specific educational attainments, such as graduating from university with relatively high grades, or passing a final test. This study defines undergraduate academic performance as an individual-based outcome of the process of undergraduate education that is aligned to educational goals set by the interplay of institutions, faculty and students (Kumar et al., 2021).

The definition of academic performance mentioned above highlights the need to factor in particular institutional arrangements when making predictions. In particular, within a context characterized by high information frictions, where university admission is no longer contingent upon school grades, it becomes crucial to take into account information derived from admission tests and other supplementary sources, such as standardized school assessments administered by a public entity. To understand the various factors that influence academic performance in LMICs, the authors propose a classification framework that is grounded in two fundamental dimensions of the most commonly employed variables: (i) the level of aggregation and (ii) the time frame.

Figure 1. *Enhanced framework for predicting undergraduate academic performance in a LMIC with a compulsory university admission test.*



Source: Authors' own formulation.

In this framework for predicting undergraduate academic performance at PUCP, the study considers individual and family-level data, such as prior academic records, in addition to information from the school and broader contextual factors. This framework recognizes the dynamic nature of this information across various educational phases, encompassing two key theoretical perspectives: cognitive theories and sociocultural theories (Lavin, 1965). Cognitive theories emphasize mental processes and intelligence, which have been found to strongly correlate with academic achievement in previous research (Roth et al., 2015; Lemos et al., 2014). These abilities are typically assessed through standardized tests on mathematical skills, reading comprehension, and vocabulary knowledge (Fonteyne et al., 2017), highlighting the importance of both innate capacities and learned knowledge and skills (Kuncel & Hezlett, 2010). Socio-cultural theories highlight the influence of cultural context, societal norms, and socio-economic background on academic performance (Rodriguez et al., 2020; Coleman, 1968).

3. Literature review

The applied literature on the prediction of undergraduate academic performance using ML models has been prolific (Rastrollo-Guerrero, 2020; Contreras et al., 2022). Despite the diversity of approaches to this prediction task, every study must face three main choices (Alyahyan & Düştegör, 2020): How to define academic performance? Which student attributes should be taken into account? Which is the best performing ML model for this task? This study used these choices as a basis to review the general literature on undergraduate academic performance prediction using ML models with an emphasis on studies in LMICs.

Many studies on the above-mentioned literature discretize the target variable, typically using Grade Point Average (GPA) or cumulative GPA, treating the prediction task as classification (Mueen et al., 2016; Al-Barrak & Al-Razgan, 2016; Almasri et al., 2018; Rifat et al., 2019). They often focus on predicting academic success or failure in specific courses or programs (Sekeroglu et al., 2021). However, this approach has drawbacks. Firstly, academic success varies across stakeholders (York et al., 2015; Cachia et al., 2018), as succeeding in a single course might not guarantee obtaining a degree, and the same degree could reflect different prior levels of academic performance. Secondly, imbalanced data affects predicting minority classes (Rastrollo-Guerrero, 2020), such as drop-outs and top performers. Thirdly, many studies use arbitrary thresholds for classification, neglecting alternative methods.

The literature indicates that prior academic performance, including high school performance, significantly influences initial university-level performance (Gil et al., 2021; Contreras et al., 2022). Standardized admission test results are also relevant factors for predicting academic success (Contreras et al., 2022). Earlier studies employing classic econometric methods in academic performance prediction emphasized the synergy between high-stakes assessments and teacher evaluations (Silva et al., 2020). Additionally, students' socio-economic backgrounds play a critical role in shaping their academic trajectory (Albreiki et al., 2021). Research in Peru has highlighted the impact of socio-economic factors like parental education, household income, and access to educational resources on educational aspirations, access, and academic performance (Guerrero et al., 2016; Benites, 2021). Economically disadvantaged students often face barriers limiting their educational pursuits, from restricted access to academic support services to financial constraints (Sánchez et al., 2021).

Some of the most common models used to predict undergraduate academic performance in Latin America are Tree-Based Algorithms –such as Gradient Boosted Tree and Random Forest– and the Multilayer Perceptron (Salas-Pilco & Yang, 2022). Accordingly, studies within the reference literature have also shown that ensemble methods tend to outperform individual algorithms (Contreras et al., 2022). Regarding prediction accuracy, there are certain differences depending on the specification of the target variable, namely if academic performance is defined at the course level –rather than the degree or year level– accuracy tends to be higher (Alyahyan & Düştegör, 2020).

In Peru, only research at the primary education level integrates administrative and census data (MINEDU, 2022; Infante & Rojas, 2021). University research relies on university academic records, limiting use for admissions and scholarships. When background data is considered, it often focuses on specific academic programs (Saire, 2023; García, 2021). Furthermore, most studies predict course or year success, not overall academic performance (Puga & Torres, 2023; Incio et al., 2023; Menacho, 2017), limiting policy applications. Studies that consider data at a single point in time also face the problem of "concept drift", namely, a high risk of disconnection between the training data and new rounds of real-life data in a changing environment (Mathrani et al., 2021). Some studies incorporate ad hoc student surveys with small datasets (<100 observations), affecting model generalizability (Incio et al., 2023; García, 2021).

In the literature, ML studies predicting undergraduate academic performance in LMICs often face one out of three limitations: (i) they focus on specific programs or courses, limiting their applicability for university policies; (ii) they lack access to students' prior academic performance and background information; and (iii) they rely on achievement indicators, causing conceptual inconsistencies in measuring academic performance. This study contributes to this literature by predicting the academic performance of all undergraduate students admitted through the two primary admission channels at PUCP. This prediction relies on a novel database that integrates comprehensive prior and current academic information from both administrative and census data. The study has developed a tailored model to address the key tasks associated with predicting undergraduate academic performance.

4. Method

This research uses a qualitative variable, membership in an academic performance group, to predict students' academic performance. The selection of

ML models was guided by two criteria: choosing models with the best performance metrics in previous studies on predicting undergraduate academic performance in LMICs (Salas-Pilco and Yang, 2022; Sekeroglu et al., 2021; Infante & Rojas., 2021) and prioritizing models with higher prediction accuracy for smaller datasets. The selected models include Logistic Regression, Ridge, Lasso, Random Forest, and Extreme Gradient Boosting[1].

Ridge regression is a linear regression technique that mitigates overfitting by introducing regularization, which is particularly beneficial for addressing multicollinearity (high predictor correlation). It appends an L2 regularization term to the linear regression cost function, discouraging large weights on predictors, effectively "shrinking" coefficients towards zero, reducing sensitivity to outliers and multicollinearity. In contrast, Lasso, another regularized linear regression technique, employs L1 regularization, which possesses feature selection capabilities. It can drive irrelevant predictor coefficients to exactly zero, resulting in a sparse and interpretable model (James et al., 2013).

Random Forest is a tree-based ensemble ML model that uses bootstrap training samples. It employs a random subset of predictors when considering splits, addressing high correlation between individual trees. By drawing different training sets, the Random Forest model ensures diverse feature sets for the base trees. With a small "m" value, correlation between trees decreases, useful when dealing with highly correlated predictors. During prediction, Random Forest combines tree results via majority voting (classification) or averaging (regression) (James et al., 2013). Finally, the Extreme Gradient Boosting model builds an ensemble of decision trees sequentially, where each tree corrects the errors made by the previous ones. Like the regularized linear regression techniques, Extreme Gradient Boosting incorporates L1 and L2 regularization terms into the objective function to control the complexity of individual trees and avoid overfitting.

We evaluate the selected models using key classification performance metrics: Accuracy, AUC ROC, and F1. Firstly, Accuracy, representing the proportion of correctly classified observations, is straightforward but not ideal for imbalanced learning. Imbalanced learning occurs when models are trained on datasets with unevenly distributed prediction categories, leading Accuracy to overestimate the capability of models predicting the most common class. Secondly, AUC ROC (Area Under the Curve ROC) is more suitable for imbalanced learning. It assesses the models' ability to make class predictions across

[1] For hyperparameter tuning this study uses an exhaustive cross-validated grid search algorithm named *gridsearchcv* from Python Library Scikit-Learn (Pedregosa et al., 2011). For details about considered hyperparameters, see Appendix 2. This study uses Scikit-Learn version 1.2.2.

all possible thresholds considering probabilistic predictions. Probabilistic predictions range between 0 and 1, and a 0.5 threshold is commonly used to convert probabilities into classes. Thirdly, F1, the standard metric for imbalanced learning, is the harmonic mean between precision and recall. Precision measures correct positive predictions among all positive predictions, while recall represents the number of correctly identified positive observations. Consequently, F1 penalizes false positive and false negative instances.

Furthermore, for tree-based classifier models, assessing the contribution of each predictor is crucial. Gini impurity-based importance is widely used for this purpose, measuring how effectively observations are split by class after each tree node (Disha & Waheed, 2022). This approach calculates feature importance based on the mean decrease in impurity resulting from the splits (Koh & Blum, 2021), offering insights into the relevance of predictors for model explainability and feature selection.

5. Data

In this study, the authors created the PUCP Academic Performance Database (BRA-PUCP) by merging data from four sources. PUCP's Central Admission Office administrative records (OCAI-PUCP) provided two sources, while the other two were datasets from Peruvian Ministry of Education (MINEDU) censuses. The study initially linked PUCP's administrative records with MINEDU's records using matching identification codes: the modular code (MINEDU) and administrative code (PUCP). Yet not all PUCP-registered schools were matched. To enhance matching, the study developed an algorithm to pair school names, with two key conditions: a minimum similarity score of 77 (out of 100) and schools located in the same district, province, and region. Establishing direct correspondence between school names was challenging due to significant variations, as shown in Table 1.

Table 1. *Example of school names matched by the algorithm's similarity score and common location in both records*

School's name registered by PUCP	School's name registered by MINEDU	Similarity by matching algorithm	Region	Province	District
Juana Larco de Dammert	6050 Juana Larco de Dammert	90	Lima	Lima	Miraflores
Fe y Alegría Nro. 5	Fe y Alegría 05	81	Lima	Lima	San Juan de Lurigancho
Científico Nikola Tesla	Colegio Científico Nikola Tesla	83	Lima	Lima	Villa María del Triunfo
School Thales de Mileto	School Thales de Mileto	100	Tumbes	Tumbes	Tumbes
Pamer La Libertad	Colegio Pamer La Libertad	81	Lima	Lima	San Miguel
Innova Schools - Chincha	Innova Schools	79	Ica	Chincha	Alto Larán

Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

The latter dataset can be described as the List of Identification Codes (LIC). Corresponding identification codes or matching school names facilitate linkage between PUCP's administrative databases and MINEDU databases, as shown in Table 2. The Reduced Academic Performance Database (BAR) includes academic, administrative, demographic, and socio-economic data for students admitted via ITS and PO in 2018-1, 2019-1, and 2020-1. BAR encompasses secondary education grades, admission test scores, academic performance measures during the first two years, home school pensions, university pension scale, gender, age, and pre-admission residence. It also records academic programs, faculty, and admission channels for each student. The Historical Performance Database (BHR) covers academic performance from 2010-1 to 2021-2 for all channels and programs, sharing all BAR variables except secondary education performance. Both datasets use students as the unit of observation, with all administrative records anonymized to protect personal data.

Table 2. *Keys and geographic information employed to merge PUCP administrative datasets with MINEDU datasets*

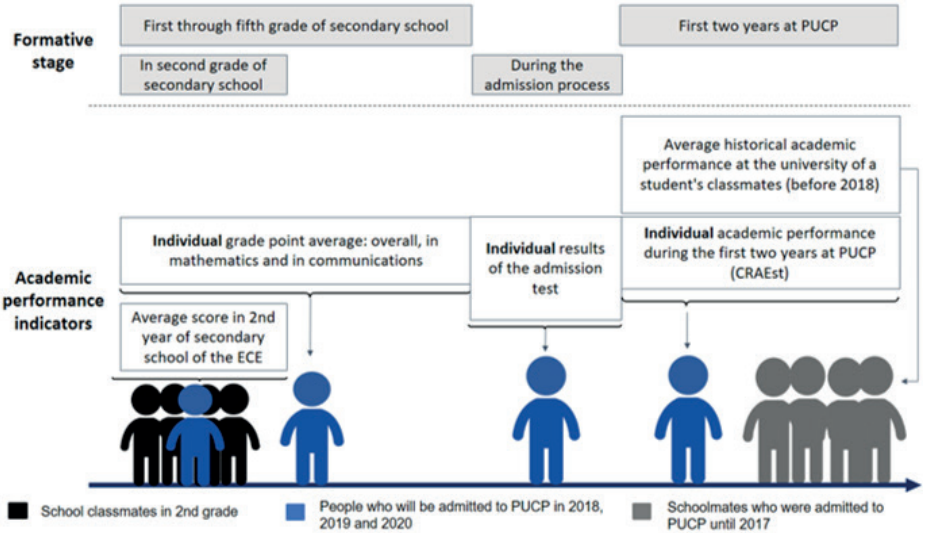
		Source: MINEDU		Source: PUCP		Geographic information		
		Modular code (cod_modular)	School's name registed by MINEDU	Administrative code (cod_pucp)	School's name registed by PUCP	Region	Province	District
Administrative datasets	BAR							
	BHR						*Province where the student lived prior to admission	*District where the student lived prior to admission
Census data	ECE							
	E-Census							
List of Identificacion Codes (LIC)	Original	*Only a subset of all schools		*Only a subset of all schools				
	Expanded by matching algorithm							

Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

MINEDU oversees two additional datasets: the annual Educational Census (E-Census) and the Student Census Evaluation (ECE). The E-Census includes all registered schools in Peru, providing information on infrastructure, services, enrollment, type, location, staff, and materials. Schools, identified by a modular code, serve as the unit of observation. Data was obtained on high school student numbers, teachers, school modular codes, and geographic locations from these datasets. The ECE comprises national standardized assessments in subjects like mathematics and reading comprehension, occasionally including natural sciences and history. The study employed ECE data for 2015, 2016, and 2018, with proxy variables for students' family socioeconomic status in 2015 and 2016. ECE data is aggregated at the school level, presenting average academic performance indicators.

The final database includes 3513 observations and 132 variables, focusing on the academic performance of students enrolled at PUCP from 2018-1 to 2020-1. It incorporates data on their university performance, admission test results, high-school performance, and historical performance indicators for their respective schools. Additionally, the database encompasses demographic, geographic, and socio-economic information, capturing the students' backgrounds and context. It tracks students' academic progress up to their second year at the university (see Figure 2).

Figure 2. *Type of academic performance information available in the Academic Performance Database (BRA-PUCP)*



Source: Authors' own formulation.

The study uses the Coefficient of Standardized Academic Achievement (CRAEst) as its target variable. The CRAEst represents an average of standardized grades, weighted by the number of credits of each course *c* that a student *i* has completed in their academic history up to the academic cycle *T*. It is calculated using the following formula:

$$CRAEst_i^T = \frac{\sum_C (s\_grade_{iC} * credits_C)}{\sum_C credits_C}$$

The standardization of grades is determined by the following expression:

$$s\_grade_{iC} = \left(\frac{grade_{iC} - \mu_C}{\sigma_C} \times 10\right) + 50$$

As a cumulative measure, CRAEst covers all courses from the beginning of the undergraduate program up to cycle *T*. For example, $CRAEst^4$ reflects the cumulative record up to the fourth cycle, $CRAEst^3$ up to the third, and so on, ensuring it captures the entire academic history. The study chose $CRAEst^3$

for two reasons: First, to account for students' adaptation to university life beyond just the first semester, and second, to reduce potential bias from drop-out data exclusion, as *CRAEst*[3] has fewer missing values than *CRAEst*[4].

6. Results

This section presents the study's empirical findings on undergraduate academic performance prediction at PUCP. Although its target variable, *CRAEst*[3], is initially continuous, the study categorizes it into two classes using various quantiles: 50 percent (median), 33 percent (tertile), 20 percent (quintile), 10 percent (decile), 5 percent, and 1 percent (percentile). For example, two classes are established for the 33 percent quantile (tertile): one for CRAEst values below the lowest tertile and another for values above it. The same process is followed for all quantiles.

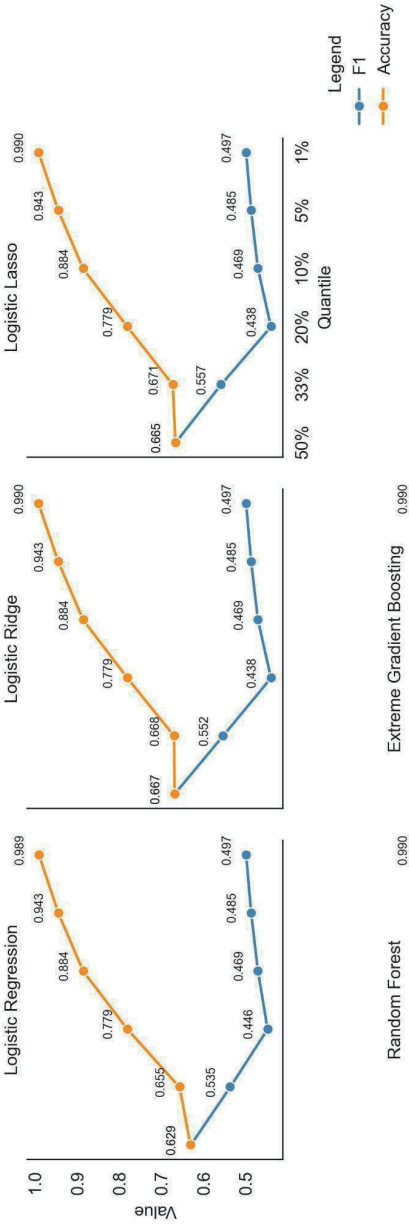
Table 3. *Number of observations per category of prediction at different quantile widths*

	Quantile widths					
Number of observations	**50%**	**33%**	**20%**	**10%**	**5%**	**1%**
Belonging to the lowest value of the quantile	1757	1173	703	352	176	36
Not belonging to the lowest value of the quantile	1756	2340	2810	3161	3337	3477

Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

The study explores various quantile widths to set a binary outcome threshold. Smaller widths offer deeper insights: for instance, a narrower quantile reveals higher drop-out risk when a student's *CRAEst*[3] falls below its lowest value. However, narrowing quantiles can harm ML model performance due to increased data imbalance. Figure 3 displays model F1 and accuracy metrics at different quantile widths, showing that reducing the width from the median (50 percent) to the percentile (1 percent) substantially lowers F1 while boosting accuracy. In the presence of imbalanced data, F1 demonstrates robustness, while accuracy tends to overestimate the predictive capability of models. The study therefore gives priority to F1 as its primary metric. Figure 3 illustrates that narrower quantiles generally result in reduced model performance.

Figure 3. *Quantiles versus accuracy and F1 for trained models*



Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

In the trade-off between narrowing the quantile range and model performance, the study found that the 33 percent quantile (tertile) is the optimal binary threshold for the Random Forest model. This choice is based on the Random Forest's more gradual performance decline beyond the 33 percent quantile compared with other models. At the 33 percent quantile, Random Forest models achieve an F1 score of 0.613, higher than other models, particularly linear models. The study hence proceeds with its analysis using the lower tertile as the binary threshold.

Table 4. *Classification models' results when considering the lowest tertile as a threshold*

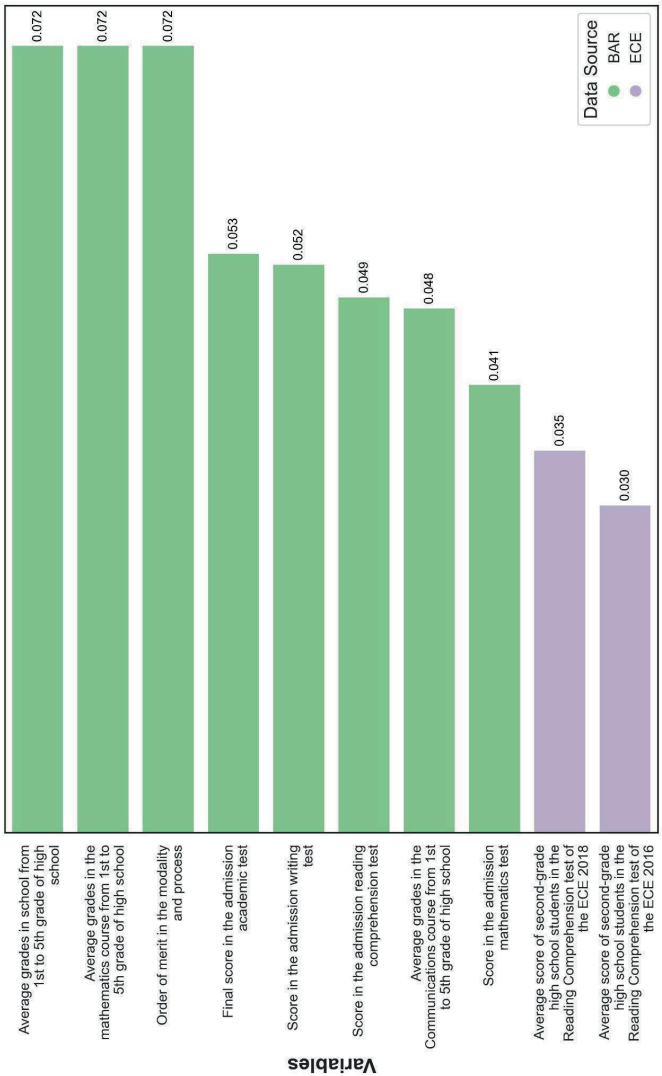
Performance metric	Logistic Regression	Ridge	Lasso	Random Forest	Extreme Gradient Boosting
Accuracy	0.655	0.668	0.671	0.692	0.692
AUC ROC	0.661	0.699	0.697	0.718	0.717
F1	0.535	0.552	0.557	0.613	0.606

Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

Table 4 presents results for classification models using the lowest tertile as a threshold. As discussed in Chapter 4, model performance is assessed using three metrics: Accuracy, Area Under the Curve ROC (AUC ROC), and F1. Notably, Random Forest and Extreme Gradient Boosting achieve the highest Accuracy at 0.692. In terms of AUC ROC, Random Forest leads with a substantial value of 0.718, followed closely by Extreme Gradient Boosting at 0.717. For F1, Random Forest excels with a score of 0.613, making it the top-performing model. The second-best is the Extreme Gradient Boosting model with an F1 score of 0.606. Given the slightly higher F1 score, the study selected the Random Forest model as the optimal choice, despite both models demonstrating comparable performance.

The study also evaluates the importance of predictors according to the Gini impurity-based features importance criterion. Figure 4 shows the top ten input features. These predictors mainly refer to two dimensions: school grades and university admission results. School-related information includes general average grades, average grades for mathematics and communications courses, and average scores for second graders from the same school in the reading comprehension test of the ECE 2016 and 2018. Information from the university admission process includes general order of merit, and scores in the admission writing, reading tests, as well as the final score in the admission academic test. Data sources for the top ten predictors are BAR and ECE.

Figure 4. *Top ten predictors according to Gini impurity-based features importance criteria for the Random Forest model*



Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

7. Policy applications

This section shows that ML models can be used as tools for predicting the likelihood of students belonging to PUCP's lowest academic performance tertile. To achieve this, it constructs classification models by utilizing the top ten predictors determined through the Gini impurity-based features importance criterion. It limits the input features to ten for practicality. Performance metrics in Table 5 indicate that Random Forest is the best-performing model, as anticipated. Notably, reducing the predictors to ten has not substantially impacted classification capability. A comparison with Table 4 reveals a slight decrease in F1 (from 0.613 to 0.604), a modest reduction in accuracy (from 0.692 to 0.684), and unchanged AUC ROC (0.718). Given the reduction from 132 to 10 input features, this minor performance decline is reasonable.

Table 5. *Classification models results for top ten predictors*

Performance Metric	Logistic Regression	Ridge	Lasso	Random Forest	Extreme Gradient Boosting
Accuracy	0.676	0.679	0.678	0.684	0.683
AUC ROC	0.703	0.698	0.695	0.718	0.716
F1	0.568	0.556	0.551	0.604	0.595

Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

To enhance transparency, the authors aimed to demystify the optimal model (Random Forest trained with the top ten features). To achieve this, the study employed SHAP (SHapely Additive exPlanations) values for in-depth analysis of each predictor's impact on the model's output. SHAP values blend additive feature attribution methods with the Shapley framework from game theory. Additive feature attribution methods dissect the model's outcome into the individual contributions of each feature, adhering to principles of local accuracy, missingness, and consistency (Lundberg et al., 2018). Meanwhile, the Shapley framework, rooted in cooperative game theory, ensures fairness by examining all possible player combinations (Michalak et al., 2013). In the context of binary classification, SHAP values reveal the direct contribution of input features to the outcome for each prediction category.

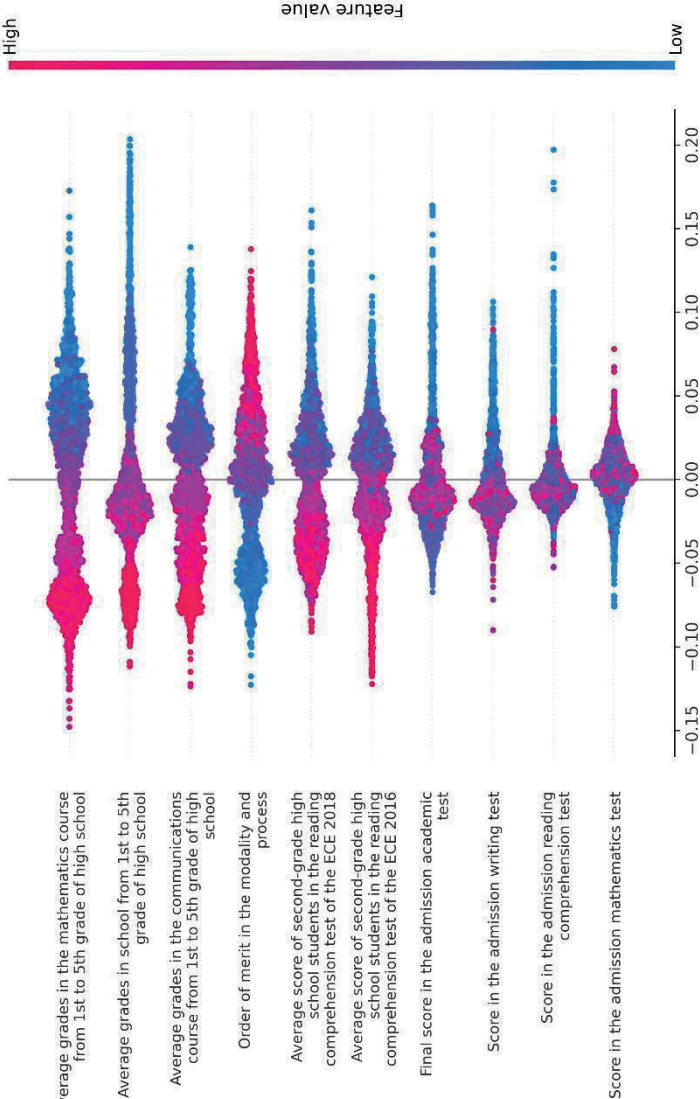


Figure 5. *SHAP values for Random Forest model trained with top ten variables.*

Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

Figure 3 presents SHAP (SHapley Additive exPlanations) values for the Random Forest model trained with ten first predictors. In the figure, colors correspond to predictor values, with redder tones indicating higher predictor values and bluer tones indicating the opposite. The horizontal axis represents the SHAP values. Larger positive SHAP values for a predictor indicate a greater positive contribution to the output, while larger negative SHAP values signify the opposite.

According to SHAP values, input features fall into two categories. The first category includes features where higher values positively predict belonging to the lowest tertile. Notably, this group comprises variables like the order of merit in the admission process and scores in the admission mathematics test. Thus, students admitted at the bottom of the merit order, and those with higher math test scores are more likely to be classified in the lowest tertile by the Random Forest model.

The second category encompasses features where higher values negatively predict classification within the lowest tertile. These factors relate to high school academic performance, including overall grade point averages, math and communication course grades, and average scores of second-grade students from the same school on reading comprehension tests conducted in 2016 and 2018. Additionally, final admission scores and math and communication test scores are included. Consequently, students from schools with enhanced reading comprehension scores, those with strong high school academic performance, and those with high university admission scores tend to be classified outside the lowest tertile by the Random Forest model.

Conclusion

The expansion of higher education access in low- and middle-income countries coexists with disparities in institutional quality and low graduation rates, especially for low-income students. This study utilizes machine learning models to predict the academic outcomes of undergraduate students at PUCP. The main objectives of this research were: (I) to identify the most effective ML algorithm for predicting undergraduate students' academic performance, and (II) to determine the key predictors that influence this performance.

In addressing its first objective, the study discovered that ensemble tree-based models, specifically Random Forest and Extreme Gradient Boosting, are highly effective in predicting undergraduate academic performance. Among these, the Random Forest model slightly outperforms the Extreme Gradient Boosting model. The study also found that a Random Forest model, when

trained with the top ten features identified through Gini impurity criteria, accurately predicts students' academic outcomes. This model can serve as an effective preventative tool to mitigate student attrition. However, it is important to underscore the need for transparency, explicability, and accountability when applying machine learning tools. This study, employed SHAP values to explain the decision-making process of the model, detailing how each input feature contributes to the predicted outcome.

In pursuing the second objective, the study found that variables related to prior academic achievement and admission criteria are crucial predictors of student performance, as pointed out by previous research. The analysis utilized SHAP values and indicated that students with stronger results in these areas are less likely to be classified into the lowest tertile of academic performance. Currently, the burden of admission criteria in many Peruvian universities lies only in the results of an entrance test. Yet the study findings suggest that to predict potential academic performance, authorities can consider prior academic performance more accurately in the form of both average school grades and ECE average scores. Moreover, a composite index of the two variables mentioned can be developed to inform admission decisions.

This study has three main limitations. First, it does not consider variables referring to students' motivation or teaching pedagogy. Although difficult to measure, these variables can arguably serve as relevant predictors of students' performance. Second, the model was trained using data from 2018 to 2020, during the pandemic. Future iterations should incorporate post-COVID-19 data to avoid the problem of "concept drift". Third, while providing insights into student attrition, it does not offer program-specific recommendations.

Acknowledgments

We would like to acknowledge OCAI-PUCP for providing us with access to administrative data that greatly contributed to the completion of this work. In particular, we extend our gratitude to José Rodriguez, former OCAI"s director and current Economics professor at PUCP, for his unwavering support from the beginning of this research. We also thank Abel Camacho, Nadja Florian, Daniel Calderón and Esteban Cabrera for their valuable comments on a previous version of this work.

References

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552-579. https://doi.org/10.3390/educsci11090552

Almasri, A., Celebi, E., & Alkhawaldeh, R. (2019). EMT: Ensemble meta-based tree model for predicting student performance. *Scientific Programming, 2019*. https://doi.org/10.1155/2019/3610248

Al-Barrak, M., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: a case study. *International journal of information and education technology*, *6*(7), 528-533. https://doi.org/10.7763/ijiet.2016.v6.745

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(3), 1-21. https://doi.org/10.1186/s41239-020-0177-7

Andrabi, T., Bau, N., Das, J., & Khwaja, A. (2022, November). *Heterogeneity in School Value-Added and the Private Premium* (Working Paper No. 30627). National Bureau of Economic Research. https://doi.org/10.3386/w30627

Athey, S., & Imbens, G. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, *31*(2), 3-32. https://doi.org/10.1257/jep.31.2.3

Balán, J. (2020). Expanding access and improving equity in higher education: the national systems perspective. In S. Schwartzman (Ed.), *Higher education in Latin America and the challenges of the 21st century* (pp. 59-75). Springer. https://doi.org/10.1007/978-3-030-44263-7

Beck, H., & Davidson, W. (2001). Establishing an Early Warning System: Predicting Low Grades in College Students from Survey of Academic Orientations Scores. *Research in Higher Education*, 42, 709-723. https://doi.org/10.1023/A:1012253527960

Benites, R. (2021, April). La educación superior universitaria en el Perú post-pandemia (Policy Document No.1). Pontificia Universidad Católica del Perú. https://repositorio.pucp.edu.pe/index/handle/123456789/176597

Cachia, M., Lynam, S., & Stock, R. (2018). Academic success: Is it just about the grades? *Higher Education Pedagogies*, *3*(1), 434-439. https://doi.org/10.1080/23752696.2018.1462096

Coleman, J. S. (1968). Equality of educational opportunity. *Integrated education*, *6*(5), 19-28. https://doi.org/10.1080/0020486680060504

Contreras, L., Caro, J., & Morales, D. (2022). A review on the prediction of students' academic performance using ensemble methods. *Ingeniería Solidaria, 18*(2), 1-28. https://doi.org/10.16925/2357-6014.2022.02.01

Daud, A., Radi, N., Abbasi, R., Lytras, M., Abbas, F. & Alowbdi, J. (2017). Predicting Student Performance using Advanced Learning Analytics. *Proceedings of the 26th international conference on world wide web companion*, 415-421. https://doi.org/10.1145/3041021.3054164

De Los Rios, F. (2023, April). ¿Es el enfoque correcto?: El problema de la modalidad de ingreso por examen de admisión a las universidades nacionales del Perú. Estudios Generales Letras - Pontificia Universidad Católica del Perú. https://files.pucp.education/facultad/generales-letras/wp-content/uploads/2022/06/15113956/%C2%BFEs-el-enfoque-correcto_-El-problema-de-la-modalidad-de-ingreso-por-examen-de-admision-a-las-universidades-nacionales-del-Peru.docx.pdf

Disha, R., & Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. *Cybersecurity, 5*(1), 1-22. https://doi.org/10.1186/s42400-021-00103-8

Fonteyne, L., Duyck, W., & De Fruyt, F. (2017). Program-specific prediction of academic achievement on the basis of cognitive and non-cognitive factors. *Learning and Individual Differences*, 56, 34-48. https://doi.org/10.1016/j.lindif.2017.05.003

Garcia, J. (2021). *Machine learning para predecir el rendimiento académico de los estudiantes universitarios* [Bachelor thesis, Universidad César Vallejo]. Universidad César Vallejo. https://repositorio.ucv.edu.pe/handle/20.500.12692/83442

Gil, P., Da Cruz Martins, S., Moro, S., & Costa, J. (2021). A data-driven approach to predict first-year students' academic success in higher education institutions. *Education and Information Technologies*, *26*(2), 2165-2190. https://doi.org/10.1007/s10639-020-10346-6

Guerrero, G., Sugimaru, C., Cussianovich, A., De Fraine, B., & Cueto, S. (2016, March). *Education aspirations among young people in Peru and their perceptions of barriers to higher education* (Working Paper No. 148). https://www.grade.org.pe/en/publicaciones/education-aspirations-among-young-people-in-peru-and-their-perceptions-of-barriers-to-higher-education/

Hajar, M., Adil, J., Ali, Y., & Khalid, A. (2022). Predicting Student Success in a Scholarship Program: A Comparative Study of Classification Learning Models. In S. Motahhir & B. Bossoufi (Eds.) *Digital Technologies and Applications: Proceedings of ICDTA'22, Fez, Morocco, Volume 2*, 333-341. Springer. https://doi.org/10.1007/978-3-031-02447-4_35

Incio, F., Capuñay, D., & Estela, R. (2023). Modelo de red neuronal artificial para predecir resultados académicos en la asignatura Matemática II. *Revista Electrónica Educare*, *27*(1), 1-19. https://doi.org/10.15359/ree.27-1.14516

Infante, L. & Rojas, J. (2021). Identification of factors that affect the academic performance of high school students in Peru through a machine learning algorithm. *Proceedings of the 19th LACCEI International Multi-Conference for Engineering, Education and Technology*. https://www.laccei.org/LACCEI2021-VirtualEdition/full_papers/FP68.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (1st ed.). Springer.

Kumar, S., Agarwal, M., & Agarwal, N. (2021). Defining and measuring academic performance of Hei students-a critical review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(6), 3091-3105.

Kuncel, N. R., & Hezlett, S. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*, *19*(6), 339-345. https://doi.org/10.1177/0963721410389459

Lavin, D. E. (1965). *The prediction of academic performance*. Russel Sage Found.

LBDEAC - Local Burden of Disease Educational Attainment Collaborators. (2020). Mapping disparities in education across low-and middle-income countries. *Nature, 577*(7789), 235-238. https://doi.org/10.1038/s41586-019-1872-1

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning analytics in higher education—a literature review. In Peña-Ayala, A. (eds.), *Learning Analytics: Fundaments, Applications, and Trends. Studies in Systems, Decision and Control*, 1-23, Springer. https://doi.org/10.1007/978-3-319-52977-6_1

Lemos, G., Abad, F., Almeida, L., & Colom, R. (2014). Past and future academic experiences are related with present scholastic achievement when intelligence is controlled. *Learning and Individual Differences*, *32*, 148-155. https://doi.org/10.1016/j.lindif.2014.01.004

Lundberg, S., Erion, G., & Lee, S. (2018). Consistent individualized feature attribution for tree ensembles. https://doi.org/10.48550/arXiv.1802.03888

Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open, 2*, https://doi.org/10.1016/j.caeo.2021.100060.

Menacho, C. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos, 78*(1), 26-33. http://doi.org/10.21704/ac.v78i1.811

Michalak, T., Aadithya, K., Szczepanski, P., Ravindran, B., & Jennings, N. (2013). Efficient computation of the Shapley value for game-theoretic network centrality. *Journal of Artificial Intelligence Research*, *46*, 607-650. https://doi.org/10.1613/jair.3806

MINEDU - Ministerio de Educación del Perú. (2018). Desafíos en la medición y el análisis del estatus socioeconómico de los estudiantes peruanos. Lima. https://hdl.handle.net/20.500.12799/5862

MINEDU - Ministerio de Educación del Perú. (2022). Alerta Escuela: Machine Learning para el cálculo del riesgo de interrupción de estudios en el Perú. https://repositorio.minedu.gob.pe/handle/20.500.12799/8668

Moreno-Ger, P., & Burgos, D. (2021). Machine Learning and Student Activity to Predict Academic Grades in Online Settings in Latam. In Burgos, D., Branch, J.W. (eds), Radical Solutions for Digital Transformation in Latin American Universities. Lecture Notes in Educational Technology, 243-257. Springer, https://doi.org/10.1007/978-981-16-3941-8_13

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. I*nternational Journal of Modern Education and Computer Science, 8*(11), 36-42. https://doi.org/10.5815/ijmecs.2016.11.05

Niri, O. (2021). Using Machine Learning for University Admission: Mapping the Socio-Technical Issue. Delft University of Technology [Bachelor Thesis, Delft University of Technology]. Research repository. http://resolver.tudelft.nl/uuid:be135436-2a52-483a-b3bb-cebbe2ed8b6a

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, MPrettenhofer, P., Weiss, R., Duborh, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of machine Learning research*, *12*, 2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Peña-Ayala, A., Cárdenas-Robledo, L., & Sossa, H. (2017). A landscape of learning analytics: An exercise to highlight the nature of an emergent field. In Peña-Ayala, A. (eds.). *Learning Analytics: Fundaments, Applications, and Trends. Studies in Systems, Decision and Control*, 65-112. Springer

Puga, J. & Torres, R. (2023). Redes neuronales artificiales para pronosticar el rendimiento académico de alumnos de ingeniería de sistemas e informática de la Universidad Nacional de la Amazonía Peruana. [Master thesis, Universidad Nacional de la Amazonía Peruana]. Repositorio Institucional Digital UNAP. https://repositorio.unapiquitos.edu.pe/bitstream/handle/20.500.12737/9204/Jorge_TrabajoDeInvestigacion_Maestria_2023.pdf?sequence=1&isAllowed=y

QS WUR - Quacquarelli Symonds World University Ranking. (2023, october, 29). QS World University Rankings 2023. https://www.topuniversities.com/university-rankings/world-university-rankings/2023

Rastrollo-Guerrero, J., Gómez-Pulido, J., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied sciences, 10*(3), 1042. https://doi.org/10.3390/app10031042

Rifat, M. R. I., Al Imran, A., & Badrudduza, A. S. M. (2019). Educational performance analytics of undergraduate business students. *International Journal of Modern Education and Computer Science, 11*(7), 44. https://doi.org/10.5815/ijmecs.2019.07.05

Rodríguez, C., Cascallar, E. and Kyndt, E. (2020). Socio-economic status and academic performance in higher education: A systematic review. *Educational Research Review, 29*, 100305. https://doi.org/10.1016/j.edurev.2019.100305

Romero, R. (2021). La formación académica de los jóvenes y las pruebas de admisión a la educación superior. Una experiencia peruana. *Horizontes. Revista de Investigación en Ciencias de la Educación, 5*(19), pp.714-752. https://doi.org/10.33996/revistahorizontes.v5i19.234

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118-137. https://psycnet.apa.org/doi/10.1016/j.intell.2015.09.002

Sahlaoui, H., Nayyar, A., Agoujil, S., & Jaber, M. M. (2021). Predicting and interpreting student performance using ensemble models and shapley additive explanations. *IEEE Access*, *9*, 152688-152703. https://doi.org/10.1109/ACCESS.2021.3124270

Saire, E. (2023). Predicción de la ruta de rendimiento académico con algoritmos de clasificación. [Doctoral thesis, Universidad Nacional San Agustín de Arequipa]. Repositorio Institucional UNSA. https://hdl.handle.net/20.500.12773/16154

Salas-Pilco, S. Z., & Yang, Y. (2022). Artificial intelligence applications in Latin American higher education: a systematic review. *International Journal of Educational Technology in Higher Education, 19*(1), 1-20. https://doi.org/10.1186/s41239-022-00326-w

Sánchez, A., Favara, M., & Porter, C. (2021). *Stratification of returns to higher education in Peru: the role of education quality and major choices* (Working Paper No. 14339). IZA Institute of Labor Economics. https://www.iza.org/publications/dp/14339/stratification-of-returns-to-higher-education-in-peru-the-role-of-education-quality-and-major-choices

SIR - Scimago Institutions Ranking. (2023, october 29). Scimago Institutions Ranking in Latinamerica 2023. https://www.scimagoir.com/rankings.php?sector=Higher+educ.&country=Latin%20America

Schendel, R., & McCowan, T. (2016). Expanding higher education systems in low- and middle-income countries: the challenges of equity and quality. *Higher education, 72*(4), 407-411. https://doi.org/10.1007/s10734-016-0028-6

Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. *Applied Sciences*, *11*(22), 10907. https://doi.org/10.3390/app112210907

Silva, L., Catela, L., Seabra, C., Balcao, A. and Alves, M. (2020). Student selection and performance in higher education: admission exam vs. high school scores. *Education Economics, 28*(5), 437-454. https://doi.org/10.1080/09645292.2020.1782846

Susnjak, T. (2023). Beyond Predictive Learning Analytics Modelling and onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT. *International Journal of Artificial Intelligence in Education*, 1-31. https://doi.org/10.1007/s40593-023-00336-3

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical assessment, research, and evaluation*, *20*(1), 5. https://doi.org/10.7275/hz5x-tx03

APPENDICES

Appendix 1. *Descriptive statistics for main features.*

Variable	Mean	Standard Deviation	First quartile	Second quartile	Third quartile
Third cycle CRAEst	51.021	5.770	47.483	50.870	54.630
Socioeconomic status of the student family	1.165	0.179	1.072	1.173	1.288
Percentage of second-grade high school students who achieved the "satisfactory" level in the language test of the ECE 2016	0.482	0.150	0.379	0.482	0.576
Percentage of second-grade high school students who achieved the "at the beginning" level in the language test of the ECE 2016	0.137	0.090	0.080	0.128	0.182
Percentage of second-grade high school students who achieved the "satisfactory" level in the mathematics test of the ECE 2016	0.358	0.139	0.267	0.354	0.436
Percentage of second-grade high school students who achieved the "at the beginning" level in the mathematics test of the ECE 2016	0.285	0.114	0.212	0.279	0.349
Average grades in school from 1st to 5th grade of high school	16.390	1.240	16.000	16.000	17.000
Average grades in the mathematics course from 1st to 5th grade of high school	15.960	1.950	15.000	16.000	17.000
Average grades in the Communications course from 1st to 5th grade of high school	15.804	1.562	15.000	16.000	17.000
Age	20.436	0.949	20.000	20.000	21.000
Order of merit in the university admission process	234.904	154.025	107.000	222.000	344.000

Source: Authors' own formulation based on OCAI (2023) and MINEDU (2018).

Appendix 2. *Considered hyperparameter values for trained models*

Models	Number of trees	Max depth of trees	Percentage of predictors used when looking for best split	Regularization strength
Logistic Regression	–	–	–	–
Ridge	–	–	–	0.001, 0.01, 0.1, 1, 10 and 100
Lasso	–	–	–	0.001, 0.01, 0.1, 1, 10 and 100
Random Forest	250, 500 and 1000	10, 20 and 30	20%, 30% and 40%	–
Gradient Boosting Trees	250, 500 and 1000	1 and 2	20%, 30% and 40%	–

Source: Authors' own formulation based on documentation from Python Scikit-Learn library (Pedregosa et al., 2011).

Appendix 3. *Optimal hyperparameters for models trained with all input features.*

Models	Number of trees	Max depth of trees	Percentage of predictors used when looking for best split	Regularization strength
Logistic Regression	–	–	–	–
Ridge	–	–	–	100
Lasso	–	–	–	100
Random Forest	500	10	40%	–
Gradient Boosting Trees	250	1	20%	–

Source: Authors' own formulation.

Appendix 4. *Optimal hyperparameters for models trained with top ten input features*

Models	Number of trees	Max depth of trees	Percentage of predictors used when looking for best split	Regularization strength
Logistic Regression	–	–	–	–
Ridge	–	–	–	10
Lasso	–	–	–	10
Random Forest	250	10	20%	–
Gradient Boosting Trees	250	1	20%	–

Source: Authors' own formulation.

Appendix 5. *Literature on the prediction of undergraduate academic performance using machine learning models in LMICs (Part I)*

Authors	Input data	Dependent variable	Trained models	Best model	Performance metrics
Daud et al. (2017)	Students from different universities of Pakistan (690 obs.)	Discretized academic performance categories	Bayesian Network, Naïve Bayesian, Support Vector Machine, C4.5 and & CART	Support Machine	F1 Score 0.867
Al-Barrak & Al-Razgan (2016)	Transcript data for female students at King Saud University in year 2012 (236 obs.)	Discretized five GPA categories (excellent, very Good, Good, average, and poor)	Tree based models	J48 tree	Accuracy: 87%
Sahlaoui et al. (2017)	Public college data with 17 features (480 obs.)	Discretized three GPA categories (low, average, and high performance)	K – Neighbor Classifier, Decision Trees, Random Forest, Baggin & ExtraTree Classifier	Bagging	Accuray: 98%

Source: Authors' formulation based on the referenced studies.

Appendix 6. *Literature on the prediction of undergraduate academic performance using machine learning models in LMICs (Part II)*

Authors	Input data	Dependent variable	Trained models	Best model	Performance metrics
Mueen et al. (2016)	Undergraduate students who had taken Programming Fundamental and Advanced Operating System (38 features)	Discretized academic performance categories	Decision Trees (C4.5), Artificial Neural Networks and Naïve Bayes	Naïve Bayes	Accuracy: 86%
Almasri et al. (2018)	Students records with 13 attributes (400 obs.)	Discretized academic performance categories (high, middle, and low performance)	Learning techniques families: Bayes, Function, Lazy and Trees	Ensemble Meta-Based Tree Model	Accuracy: 98.5%
Rifat et al. (2019)	Bussiness students from University of Bangladesh (398 obs.)	Discretized academic performance categories on final CGPA (Honors, First Class, Second Class)	Gradient Boosted Tree, Random Forest, Tree Ensemble, Decision Tree, SVM y KNN	Random Forest	Accuray: 94.1%

Source: Authors' formulation based on the referenced studies.

Roles de autor: **Salas, F.**: Conceptualización, Software, Validación, Investigación, Escritura – Borrador original, Escritura – Revisión y edición, Supervisión. **Caldas, J.**: Metodología, Software, Validación, Análisis formal, Investigación, Curación de datos, Visualización, Escritura – Borrador original, Escritura – Revisión y edición.

Cómo citar este artículo: Salas, F., & Caldas, J. (2024). Predicting undergraduate academic performance in a leading Peruvian university: A machine learning approach, *Educación, XXXIII*(64), 55-85. https://doi.org/10.18800/educacion.202401.M003

Primera publicación: 8 de marzo de 2024.