applied sciences

MDPI

Article

Visual-Predictive Data Analysis Approach for the Academic Performance of Students from a Peruvian University

David Orrego Granados [1], Jonathan Ugalde [2,3], Rodrigo Salas [3,4] and Romina Torres [3,5] and Javier Linkolk López-Gonzales [6,*]

1 UPG Ingeniería y Arquitectura, Escuela de Posgrado, Universidad Peruana Unión, Lima 15464, Peru
2 Escuela de Ingeniería Informática, Universidad de Valparaíso, Valparaíso 2581967, Chile
3 Millennium Institute for Intelligent Healthcare Engineering (iHealth), Santiago 7820436, Chile;
4 Escuela de Ingeniería C. Biomédica, Universidad de Valparaíso, Valparaíso 2581967, Chile
5 Facultad de Ingeniería, Universidad Andres Bello, Viña del Mar 2531015, Chile
6 Facultad de Ingeniería y Arquitectura, Universidad Peruana Unión, Lima 15464, Peru
* Correspondence: javierlinkolk@gmail.com

Abstract: The academic success of university students is a problem that depends in a multi-factorial way on the aspects related to the student and the career itself. A problem with this level of complexity needs to be faced with integral approaches, which involves the complement of numerical quantitative analysis with other types of analysis. This study uses a novel visual-predictive data analysis approach to obtain relevant information regarding the academic performance of students from a Peruvian university. This approach joins together domain understanding and data-visualization analysis, with the construction of machine learning models in order to provide a visual-predictive model of the students' academic success. Specifically, a trained XGBoost Machine Learning model achieved a performance of up to 91.5% Accuracy. The results obtained alongside a visual data analysis allow us to identify the relevant variables associated with the students' academic performances. In this study, this novel approach was found to be a valuable tool for developing and targeting policies to support students with lower academic performance or to stimulate advanced students. Moreover, we were able to give some insight into the academic situation of the different careers of the university.

Keywords: students' performances; machine learning; learning analytics; educational data mining; business intelligence in education

Citation: Orrego Granados, D.; Ugalde, J.; Salas, R.; Torres, R.; López-Gonzales, J.L. Visual-Predictive Data Analysis Approach for the Academic Performance of Students from a Peruvian University. *Appl. Sci.* **2022**, *12*, 11251. https://doi.org/10.3390/app122111251

Academic Editor: Giancarlo Mauri

Received: 21 August 2022
Accepted: 3 November 2022
Published: 6 November 2022

1. Introduction

The educational system in Peru is one of the fundamental factors for the development and growth of the country [1]. Most countries consider improving their educational systems to provide a better learning environment for the students and give a public value to the society. For this reason, every year, the investment in public policies is constantly growing. Currently, universities need instruments and relevant information to help them understand the complex academic landscape and, thus, introduce targeted policies that allow helping those students with more significant difficulties.

In Peru, in 2006, the National System for Evaluation, Accreditation, and Certification of Education Quality was created in order to certify a quality education level for all students, evidencing essential quality aspects after a rigorous evaluation of specific standards and indicators [2]. Peru is a country that seeks to improve its educational system because it ranks 127th out of 137 countries studied in educational quality [3] in 2019.

In recent decades, universities around the world have made efforts to identify low-performing students to give them academic and vocational support in order to prevent them from dropping out or being eliminated from their careers or institutions [4]. In this sense, Peruvian universities have also paid attention to the student performance problem, and demand methods and systems to analyze and identify students at academic risk, in

order to guide the development of student retention policies and improvement of the academic performance of their students [1,2].

An important approach found in the literature to address student performance problem is the application of Data Mining to analyze students' context and factors related to their performance [4]. This approach is known as Educational Data Mining (EDM). EDM is the discipline in which Data Mining methods are explored, studied, and applied in educational settings with widely known problems such as desertion, admission, and student performance, among others [5]. Data Mining has a widely known method to extract information that is not easily recognizable from raw data, using statistical and machine learning techniques to identify relevant information and patterns that allow identifying the factors that influence the academic performance of students.

Machine learning has been used to assess student performance prediction in higher education for the last decade in the literature with promising results [4]. Most of these studies have been developed over two main type of datasets: student data from colleges/university databases and from online learning platforms called online Learning Management Systems (LMS). Authors have addressed the student performance prediction problem using classic machine learning algorithms, and recently, artificial neural networks and deep learning. In [6], data collected from a college database were used in order to predicts students' performance with a naive Bayes classifier, classification trees, k-nearest neighbors, and support vector machine, where k-nearest neighbors outperforms the other classifiers. In [7], authors used the Knowledge Discovery Database process to collect and prepare students' data through an LMS and applied logistic regression and support vector machine, achieving Accuracy of 73% and 79%, respectively.

On the other hand, Ref. [8] used student data from an online LMS from four Greek university courses and predicted student performance through a deep neural network applying transfer learning, from data of one course to another, achieving an Accuracy of 86%. In [9], a new deep-learning-based algorithm called GritNet was proposed. This algorithm is an evolution of bidirectional long short-term memory, for student performance classification. Likewise, authors in [10] trained a deep artificial neural network on data extracted from an online LMS to predict risky students for early intervention, achieving classification accuracies from 84% up to 93%.

For institutions that do not have an online LMS, such as the Peruvian university that is the subject of study in this work, student data stored in institutional databases can be used to build their own Learning Management System. Since the Peruvian university does not currently have a system of this type, a method to identify low-performing students becomes an important step in building a future Learning Management System for the institution.

The main goal of this work is to develop a visual-predictive data analysis scheme (VPDA) using machine learning techniques to support targeting policies in order to improve the academic performance of students. In this sense, the main contribution of this work is to provide a visual-predictive data analysis that allows estimating the academic projection of the student, since it constitutes a valuable tool for the development and targeting of policies to support students with lower academic performance or stimulate more advantaged students. Moreover, this study offers a novel approach for higher education institutions in the country and the greater South American region, allowing an analysis of the educational context, subject to particular conditions, different from institutions on other continents. The machine learning techniques allow classifying and predicting student academic performance based on the grades obtained. Likewise, it segments students according to their academic performance to follow them, taking advantage of their potential and academic abilities, leading to the support of policies in student accompaniment.

The rest of this study is structured as follows: Section 2 presents the different related works that precede this investigation. Section 3 details the visual-predictive methodology used to assess the student performance problem. The results obtained are presented in Section 4. Section 5 contrasts the discussion between the results obtained and other similar studies. Finally, Section 6 describes the conclusions and future work.

2. Related Work

Our visual-predictive data analysis approach can be situated in the Educational Data Mining domain, as it has been designed to address the well-known EDM problem of student academic performance prediction. In the literature, several studies about academic performance analysis using machine learning techniques have been successfully applied to predict the performance of higher education students [11–14]. Academic performance has different research approaches, including the evaluation of the effect of the use of social networks on academic performance using predictive models. For instance, the use of social networks in classes partially affects students' academic performance [15,16]. Likewise, Ref. [17] used machine learning algorithms to determine the academic performance of university students of the ultimate year and concluded that there is a relationship between the behavioral characteristics of students and academic performance. Moreover, Ref. [18] adopted machine learning algorithms to make an early prediction of students with a high risk of failure in their academic performance.

Student academic performance prediction and analysis are necessary to help educators identify student weaknesses and improve their scores, as well as help students improve their learning activities [19]. For instance, Ref. [20] used family background variables, and [21] used data from the interaction of the students to build predictive models of student learning performance. Moreover, Ref. [19] used machine learning methods to identify students with a high risk of dropping out of their careers and predict their future achievements. Several supervised and unsupervised machine learning techniques have been used to predict student performance by considering academic factors, highlighting pre-admission scores and previous qualifications, together with non-academic information such as emotional intelligence and resilience [4,22,23]. Different machine learning algorithms have been used more frequently to address this type of scenario and to automate different processes to evaluate academic performance [19,21,24,25].

Predicting student performance in higher education is challenging because of the large amount of data and variables that can be used for problem modeling and analysis. This situation generates a need to propose new models that could consider most of the available variables affecting students' academic performance and their learning process. In [26], the authors mentioned that the prediction of student performance and its analysis with current machine learning methods are valuable and necessary to help educators and educational institutions identify student weaknesses and design strategies to improve their scores as well as for students to improve their learning abilities.

Educational Data Mining works as the abovementioned often faced this problem from a quantitative perspective only, even when they used demographic or socioeconomic student data in their predictive model construction. In this work, our visual-predictive approach complements machine learning findings with data visualization insights, which could lead to additional findings that cannot be noticed with machine learning performance scores (e.g., findings related to careers or institution-specific context). In this context, this research seeks to develop a visual-predictive data analysis outline to provide a predictive model of student academic success. Likewise, it aims to learn about different profiles of students according to their academic performance, discover relevant factors and information on the aspects on which they should concentrate to increase academic performance, and consolidate the policies of academic performance of higher education students.

3. Visual-Predictive Data Analysis Scheme

In this work, we propose a visual-predictive data analysis scheme as a methodology to obtain relevant information for developing and targeting policies to support students with lower academic performance or to stimulate students with good performances. This scheme represents an initial step for the creation of a Learning Management System for the Peruvian university, and it has been designed as an adaptation of the well-known Knowledge Discovery Database and CRISP-DM methodologies for the Educational Data Mining field [27]. In this sense, the scheme becomes an extremely important tool for the

institution, since it synthesizes the path of each step involved in a visual and interactive way. As visualization is one of the pillars of this study, it is considered that the adaptation of the KDD process allows obtaining a broad and complete panorama following a common thread in the context of academic performance in Peruvian university students.

Figure 1 shows the VPDA scheme applied to extract the relevant information and patterns to gain knowledge about the student performance in a Peruvian university. The VPDA scheme consists of 6 sequential and recurrent stages: (i) Educational and Learning Understanding, (ii) Data Understanding, (iii) Data Preparation, (iv) Predictive Modeling, (v) Evaluation, and (vi) Selection. These stages are described below.

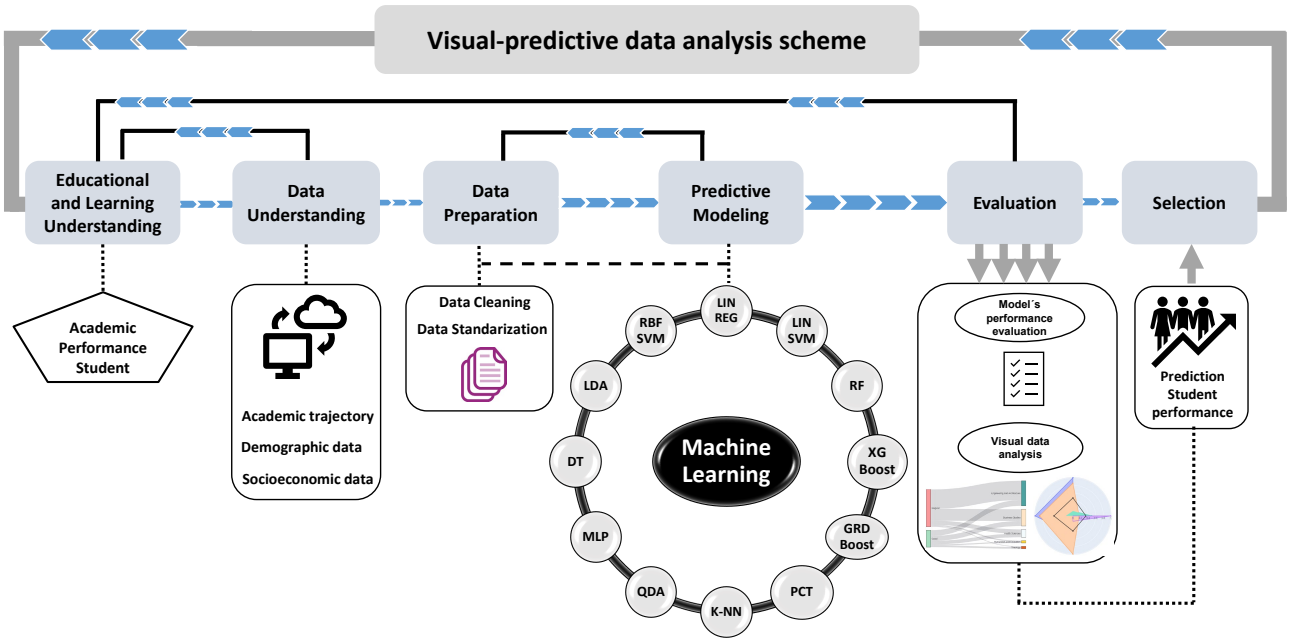

Figure 1. Visual-predictive data analysis scheme for the academic performance of the university students.

3.1. Educational and Learning Understanding

Academic performance is considered as an evaluation of the knowledge obtained by a university student by capturing the teaching received within the classroom [28]. This performance is the context of the interactions that occur between students, teachers, and the knowledge that circulates daily [29]. It is also pointed out that this variable is the one that determines how beneficial the study was within the University. In [30], the authors mention that academic performance fosters a double dimension of learning, both static and dynamic, with the dynamic aspect being the one that responds to learning processes, bringing out the student's ability and effort. In contrast, the static aspect comprises the product of the learning generated by the student and expresses an achievement behavior. The academic is visualized within the numerical qualifications on the student's capacity in front of a test. In most cases, the institutional, social, family, and personal aspects of students are not taken into account even though their impact on their academic responses has been demonstrated [31,32].

The academic performance is given through different factors that, together, generate the result of the expected academic performance of the students [29]. This performance is not only based on the degree of education. Moreover, it implies the economic-, social-, and family-related factors. For its part, the economic factor is of vital importance as a set of physical, psychological, and tangible conditions in student learning; economic deprivation forces the student to work and study, mostly neglecting studies. On the contrary, those students who have a stable economy demonstrate responsibility and a broad academic level within universities [1]. Likewise, Refs. [33,34] argue that the social factor involves

the relationship with society and is decisive for better use of education. In addition, it shows the degree of confidence in expressing themselves, including when investigating. Furthermore, Refs. [35,36] state that the social and family environment that surrounds the student plays a vital role in academic life, directly and indirectly, providing quality education and social integration, establishing discipline, rules, and routines.

The academic performance of undergraduate university students is analyzed, an indicator that can be measured with the academic grades presented during a given semester. These qualifications are the end result of the teaching process, which corresponds to the core business of any institution of higher education. In general, academic performance is a very important indicator for the decision-making processes of a university, allowing to assign benefits or incentives to students with good performance, as well as to intervene and accompany students whose academic performance is low.

3.2. Data Understanding

The objective of this stage is to carry out an initial analysis of the available data and select those attributes that are deemed pertinent based on different criteria. As an essential criterion for selecting attributes, the relationship between the data and the problem's nature is considered. In this sense, in the present study, attributes associated with the academic trajectory of the students were selected, among which are Time to Graduate, Failed Courses, Failed Courses 2 plus times, and Dropouts.

Attributes associated with academic performance were also selected, representing the grades obtained throughout the student's stay at the University, such as First-Year Score, Second-Year Score, and the classification of academic performance. These attributes were selected as they provide valuable information to the analysis and are directly related to the students' academic performance. It is important to note that attributes such as Third Year Score, Fourth Year Score, and Fifth Year Score were discarded since the later grades reflect the students' academic performance. Meanwhile, early grades, such as First-Year Score and Second Year Score, serve as prior antecedents for estimating the academic performance that a student can achieve at his final year, representing the objective of constructing a predictive model with machine learning methods.

On the other hand, demographic attributes such as gender, age, marital status, and religion were also selected since these types of data account for information underlying the students' environment and can influence their academic performance. Finally, the Payment Scheme attribute was considered as data that reflect the students' socioeconomic information. Table 1 presents all the selected attributes to perform the analysis and generate a predictive model of the academic performance of university students.

Table 1. Data dictionary.

Attributes	Data Type	Range-Values
Gender	category	F-M
Age	int	20-72
Marital Status	category	S-M-D-W-EC-O
Religion	category	A-C-E
Scholarship	category	B18-BC-BV-N
Time to Graduate	int	4-14
Failed Courses	int	0-35
Failed Courses 2 plus times	int	0-15
Dropouts	int	0-5
First Year Score	float	0-20
Second Year Score	float	0-20
Career	category	X
Payment Scheme	category	X
Performance	category	Good-Regular

3.3. Data Preparation

In this stage, the data are prepared for further analysis according to three criteria: completeness, consistency, and coherence. The first criterion focuses on the completeness

of the data, where the records with missing data are imputed with the mean values or the mode depending on whether they are numerical or categorical attributes, respectively. The second criterion corresponds to consistency, which stipulates that the values of all records must follow the same format and encoding, depending on the type of corresponding data. For example, for the categorical attribute Gender, the values "Female" and "F" are consolidated only with values "F", and the values "Male" and "M" only with values "M". The values of the numerical attribute Performance are limited to 4 decimal places. The third criterion corresponds to coherence, it indicates that all the values of an attribute must follow a specific probability distribution and the outliers must be handled. In this study, the outliers of the dataset were separated since exchange students who were passing through the institution were treated. Although one of the University's majors (Theology: Philosophy) has students much older than the average age, the rest of its attributes follow the trend of the rest of the students. Therefore, it is assumed that the dataset does not have clearly marked outliers, and all records are accepted for use in the experiments performed.

On the other hand, to effectively execute machine learning techniques, the data must have a specific and uniform format, and range of values. In this sense, categorical data transformation techniques are applied to numerical data so that the algorithms can interpret the data. The ranges of values are also standardized for all attributes so that the attributes have the same magnitude of values and the generated models do not have biases. The data standardization technique was used, which transforms the attribute values with a mean of 0 and standard deviation of 1. In Equation (1), for each feature, given a value X_i, a new value Z_i is calculated using the average $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the standard deviation $S_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ of the variable:

$$Z_i = \frac{(X_i - \overline{X})}{S_{n-1}} \quad i = 1, \ldots, n \tag{1}$$

3.4. Predictive Modeling

In this stage, the following machine learning models are implemented: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree (DT), Random Forest (RF), Regression Linear (LIN REG), SVM-Linear (LIN SVM), SVM-Radius-Basal (RBF SVM), Perceptron (PCT), Multi-Layer Perceptron (MLP), K-Nearest Neighbors (K-NN), Gradient Boost (GRD Boost), and eXtreme Gradient Boost (XGBoost).

Regarding a specific configuration of the execution parameters, the LDA used Singular Value Decomposition (SVD) as a solver. In contrast, the Decision Tree and Random Forest used a maximum branch depth of 4 levels and a Gini function to estimate the quality of the data division. On the other hand, the Linear Regression used a regularization parameter ($C = 100$). The Linear SVM and RBF SVM used a regularization parameter ($C = 1$) and a gamma kernel coefficient with value $1/(|X|\sigma)$, considering the data variance. An MLP with two layers of 5 neurons was used, with a parameter $\alpha = 0.1$ and a limit of 1000 iterations per epoch. The perceptron performed a maximum of 40 iterations. The KNN considered groups of 5 neighbors and a Euclidean distance in its execution. On the other hand, GRD Boost and XGBoost used 100 estimators and generated trees with a maximum depth of 4 levels.

3.5. Evaluation

3.5.1. Models' Performance Evaluations

The hold-out or cross-validation schemes can be used to evaluate the performances of the machine learning models. On the one hand, in the hold-out scheme, the dataset is separated into two subsets: the Training and the Testing sets. The training set is used to fit the machine learning models, while the testing set is used the evaluate the generalization performance. On the other hand, the cross-validation scheme separates the dataset into k subsets or folds, where $k - 1$ folds are used for training and the other fold is used for testing. The confusion matrix corresponds to a summary of the prediction results obtained with the

machine learning model. Given n samples, the TP is the number of true positives, the TN is the number of true negatives, FN is the number of false negatives, and FP is the number of false positives. To evaluate the performance, we use the classification metrics obtained from the confusion matrix. These metrics are *Accuracy*, *Precision*, *Recall*, and *F*1-score, following Equations (2)–(5), which are described below.

$$Accuracy = \frac{TP + TN}{n} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

Accuracy measures the proportion of samples that are correctly classified. Precision measures the proportion of predictive positives that are correct. *Recall* measures the sensitivity of the classifier to detect the positive cases. *F*1-score is the harmonic mean of the *Recall* and the *Precision* and gives a trade-off measure between the Recall and the Precision.

3.5.2. Visual Data Analysis

Different types of charts and visual representations of data allow seeing information patterns that are not visible with numerical indicators in a simple and explanatory way. This data representation and evaluation technique is part of a research line called Visual Data Analysis, which has been widely developed in the literature and in different fields of application, highlighting Business Intelligence, which has developed an industry with highly sophisticated and widely used products around the world such as Tableau, Google Data Studio, Microsoft Power BI, and QlikView, among others. These tools are utilized alongside many open-source libraries such as Matplotlib, Seaborn, and Pyplot, and a detailed visual analysis from simple charts to sophisticated and complex visual representations, such as georeferenced map charts and heat maps, bubble charts, Radar charts, or Sankey diagrams. The approach proposed in this research uses both sophisticated and straightforward charts to obtain insights related to the academic performance of university students.

3.6. Selection

As a product of the application of each algorithm over the data of university students, a model for student academic performance for new students was obtained. All the performance metrics need to be analyzed in order to select the best machine learning model. In this way, the generation of this predictive classification model means a mechanism for identifying students with high or low performance, and could help the Peruvian university to define pertinent policies and strategies for enhancing and supporting students according to their predicted academic performance.

4. Results

This section shows the results obtained by applying the VPDA approach over data from the Peruvian university students. In Section 4.1, the results of the visual data analysis are shown, while in Section 4.2, the results obtained with the predictive model are detailed.

4.1. Results of the Visual Data Analysis

Figure 2 shows the correlations between students' data attributes. A strong positive correlation can be observed between academic performance (Label attribute) and the score obtained during the first two years. On the other hand, a strong negative correlation with

respect to the Failed Courses was observed. However, this chart does not find a correlation between the student performances and the career or the department of origin.

Figure 3 shows that the Faculties of Theology and of Human Sciences and Education with their professional careers have a higher average academic performance: THEO_MSE with an average of 16.5, EDU_PIB with 16.4, THEO with 16.1, and EDU_MSA with 16.1. On the contrary, the FIA and FCE faculties with their professional careers have lower average academic performance: SYS_E with 14.8, MAR_IB with 14.8, ARCH with 14.4, ACC with 14.4, and CIV_E with 14.2.

There is also a relationship between academic performance and the average number of subjects failed by students. In Figure 4, it can be seen that the FIA and FCE faculties with their professional careers have a greater number of students who failed a course more than two times: CIV_E 14, MAN_IB 13, ENV_E 13, and ACC_TM 12. This is quite the opposite in the populations of TM_PHI 1, EDU_PIB 2, EDU_MSA 3, and THEO_MSE 4.

In terms of the academic performance of the students, in Figure 5, it can be seen that the main focuses of students with "Regular" performance are found in the careers Psychology, Book-keeping and Tax Management, Management and International Business, Environmental Engineering, and markedly in Civil Engineering. This indicates that there are certain careers in which students with "Regular" performance are grouped, either due to the difficulty of the curriculum or other types of factors, which must be explored and be the focus of attention for the university.

Meanwhile, Figure 6 shows the distribution of students according to their academic performance in the different faculties of the University. In the Engineering and Architecture and Business Studies faculties, it is observed that most of the student body has a "Regular" academic performance, which coincides with the careers that present the highest volume of regular students in Figure 5. Meanwhile, in the faculties of Humanities and Education and Theology, this trend is reversed, observing that the number of students with "Good" academic performance exceeds the number of students with "Regular" performance. However, in the Faculty of Health Sciences, the proportion of students is more balanced, with a slight tendency in favor of students with "Good" academic performance.

Figures 7 and 8 show the average values of the four attributes with the greatest weight in the selected predictive model, distributed in careers and faculties, respectively. It is important to mention that the values presented are scaled in the interval $[0,1] \in R$, considering the minimum value x_{min} and the maximum value x_{max} of the attribute, following Equations (6) and (7), where x_{std} corresponds to the standard deviation of the attribute and x_{scaled} to the scaled value.

$$x_{scaled} = x_{std} \cdot (x_{max} - x_{min}) + x_{min} \tag{6}$$

$$x_{std} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{7}$$

Considering these four attributes as the factors with the greatest influence on the classification of a student's academic performance, Figure 7 shows that, for each faculty, the classification factors change. For example, in the Humanities and Education Faculty and the Theology Faculty, the First Year Score and Second Year Score have a much higher average value than in the other faculties, which helps to explain the higher proportion of "Good" students compared with "Regular" students. In addition, in Theology, there is also an age rank much higher than the rest of the faculties.

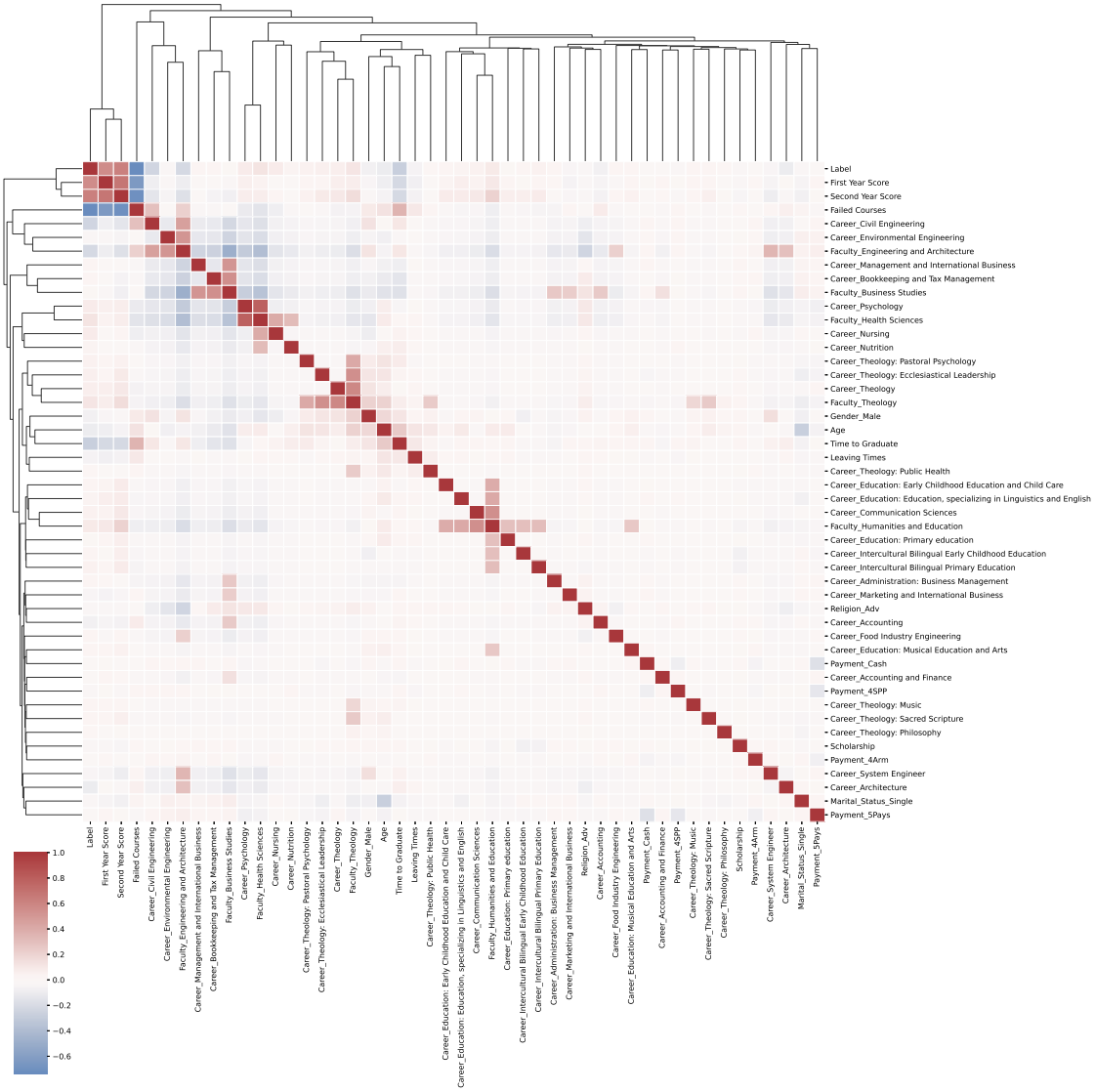

Figure 2. Student attribute correlation diagram.

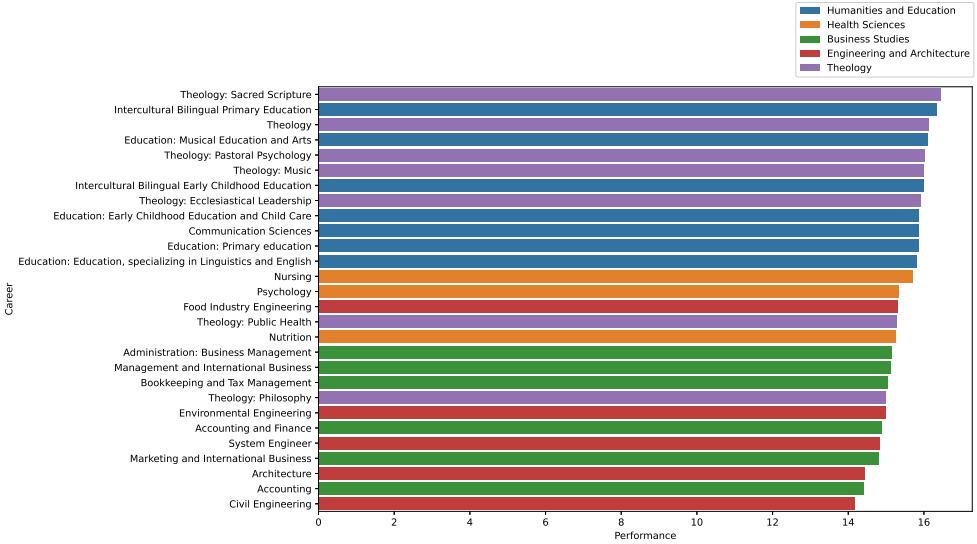

Figure 3. Average of qualifications according to professional career and faculty.

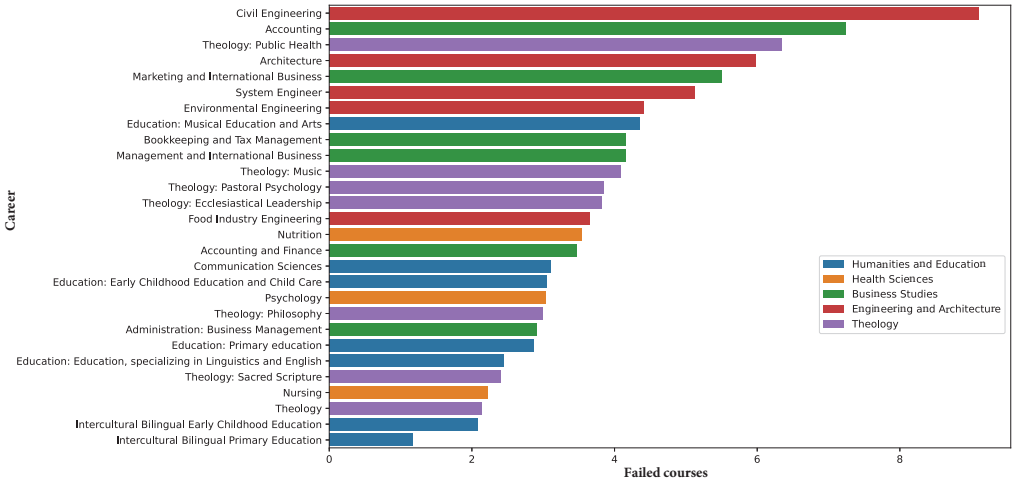

Figure 4. Average number of Failed Courses according to professional career and faculty.

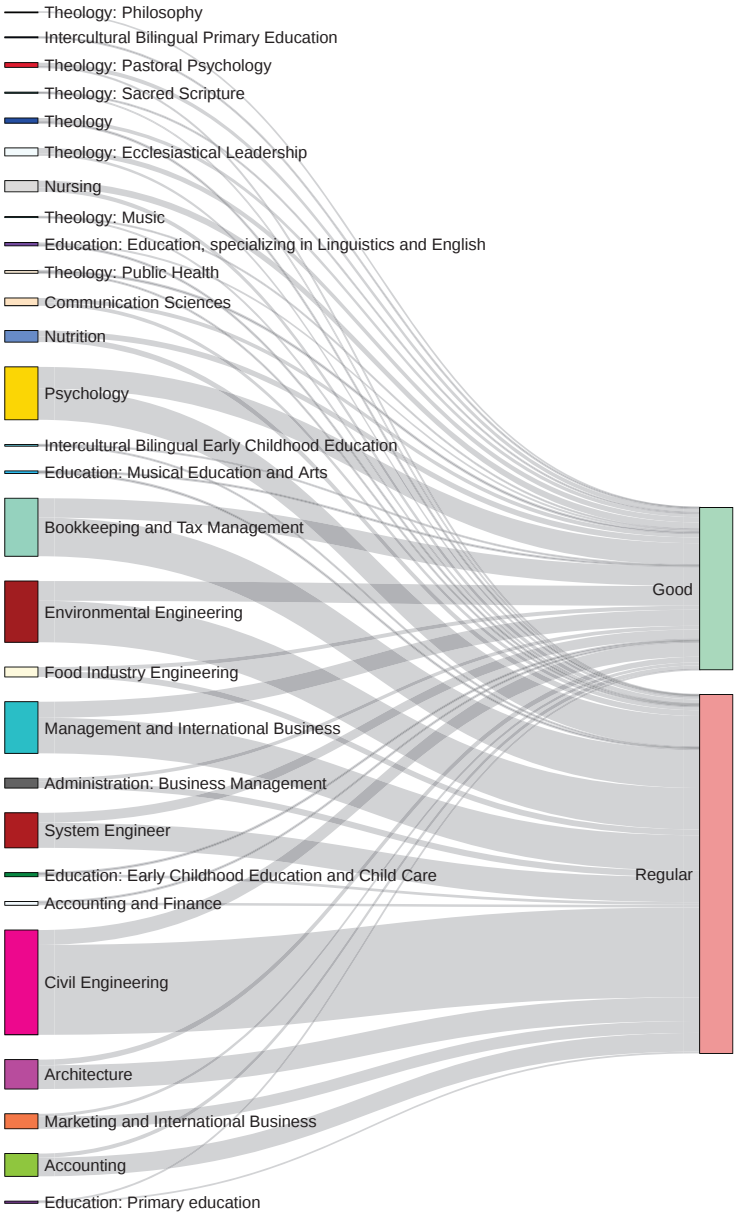

Figure 5. Student's Performance Classification distributed over career.

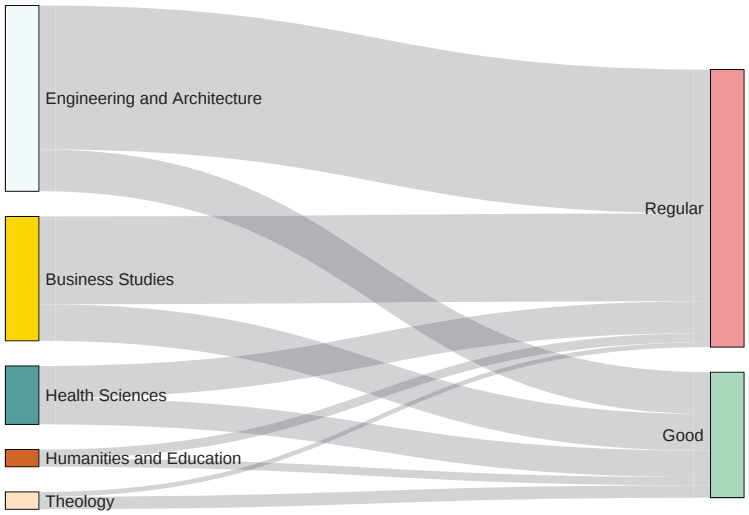

Figure 6. Student's Performance Classification distributed over faculty.

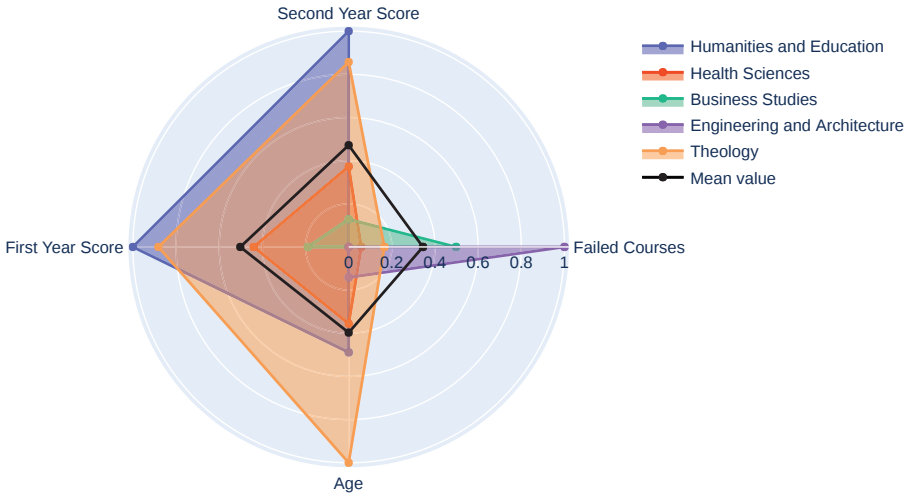

Figure 7. Principal factors by faculty. Distribution of the values of the most important attributes in the predictive model according to faculty.

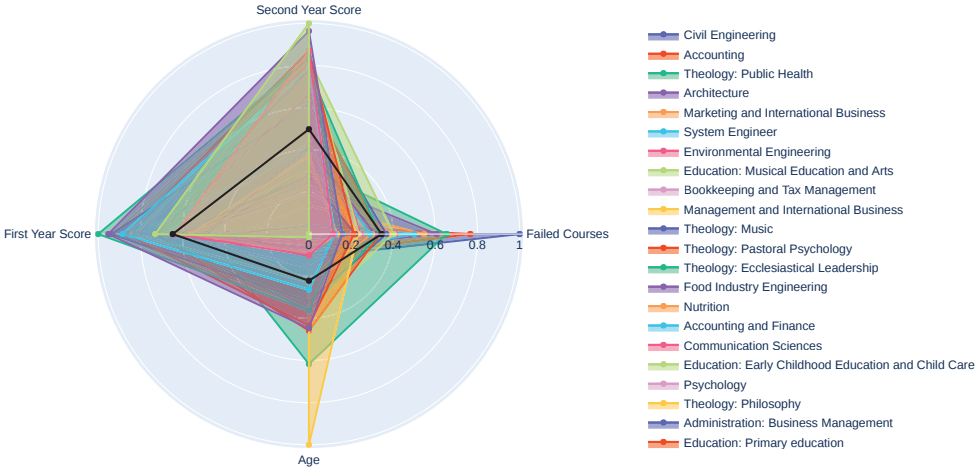

Figure 8. Principal factors by career. Distribution of the values of the most important attributes in the predictive model according to career.

In terms of Failed Courses, it is appreciated that the Faculties of Business Studies, and especially Engineering and Architecture, stand out for achieving the highest average values in this attribute. In addition, both schools have the minimum First Year Score and Second Year Score, which creates a notoriously complicated scenario in these schools. Consequently, these faculties demand more attention for the university, especially to address the high values in Failed Courses. For its part, the Faculty of Health Sciences has the average values of the most balanced attributes, approaching the Mean Value in both First Year Score and Second Year Score and age. However, it has the lowest value in Failed Courses.

Regarding the distribution of the most influential factors in the predictive model according to career, in Figure 8, it is observed that, in general, the average values of the Second Year Score are higher than the First Year Score, which indicates that almost transversally in all majors, there are students who after their first year of stay at the University manage to improve their grades in their second year. This phenomenon can generate a space for improvement in institutional policies for the accompaniment of "Regular" students, which could have some kind of effect on the reduction of the Failed Courses factor. In terms of age, it is observed that Theology: Philosophy has the highest average age, while other careers associated with theology have a much lower average age; this indicates a wide range of ages in these careers, which in turn, poses challenges in the teaching field to provide services to students of different ages. On the other hand, as can be seen in Figure 7, the degrees associated with engineering show the highest average values in Failed Courses.

4.2. Results of the Predictive Performance Evaluation

An experiment was carried out for each of the selected machine learning methods, based on 30 repetitions of the Hold-out validation scheme; 80% of the data were used for training and 20% for testing. In addition, the data separation was marked by the Shuffle Split technique, which performs a random permutation of the dataset records to form the training and test sets [37,38].

After the training and testing process of the selected classifiers, the results were obtained in Table 2. Almost all the machine learning models performed well, surpassing 86% Accuracy, with the exception of QDA, which obtained only 46.0% in that indicator. In this sense, the XGBoost classifier is recommended, as it shows the best performance indicators, mainly in Accuracy, reaching 0.912, with a good balance between Precision 0.9263 and Recall 0.9454, and an F1-score of 0.9358. On the other hand, although there are classifiers with similar performance to the XGBoost in terms of Precision, Recall, and F1-score, these differences are minimal and considered negligible. However, the LDA classifier stands out in Recall (0.9696), indicating that this classifier better identifies students with good academic performance. On the other hand, the MLP showed better Precision (0.9268).

Table 2. Results of the evaluation of the machine learning models with a cross-validation approach of 30 executions.

Classifier	Accuracy	F1-Score	Precision	Recall
XGBoost	0.912 ± 0.0072	0.9358	0.9263	0.9454
Linear SVM	0.909 ± 0.0071	0.9359	0.9243	0.9478
Linear Reg	0.908 ± 0.0075	0.9338	0.9180	0.9503
Decision Tree	0.905 ± 0.0089	0.9322	0.9228	0.9418
Random Forest	0.904 ± 0.0080	0.9280	0.9253	0.9309
LDA	0.901 ± 0.0091	0.9269	0.8879	0.9696
RBF SVM	0.900 ± 0.0077	0.9315	0.9147	0.9490
MLP	0.899 ± 0.0382	0.9312	0.9267	0.9357
Gradient Boost	0.893 ± 0.0097	0.9257	0.9218	0.9296
Perceptron	0.880 ± 0.0119	0.9121	0.9177	0.9066
K-NN	0.862 ± 0.0102	0.9062	0.8671	0.9490
QDA	0.460 ± 0.0657	0.2677	0.8904	0.1575

One of the possible causes that the considered classifiers obtain good performance is the separability of the Label attribute (classes), because as shown in Figure 9, if the Failed Courses and Second Year Score variables are considered, it is observed that the classes can

be separated even with a straight line. Although, the linear classifiers such as LDA, Linear SVM, and Linear Regression show good performance.

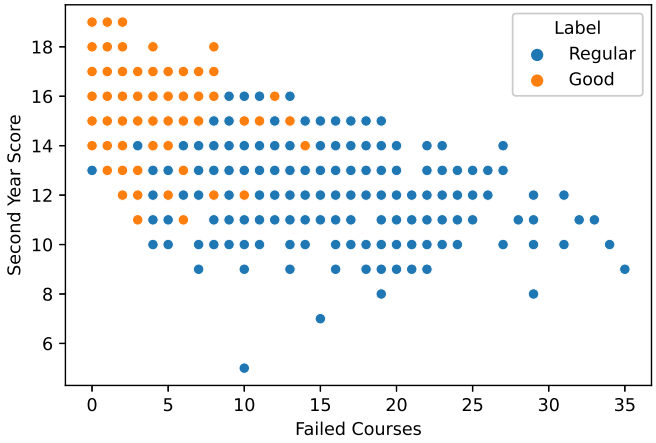

Figure 9. Scatter plot showing relationship between Failed Courses and Second Year Score; the colors that represent each value of the target indicate that they are linearly separable.

In terms of the importance of student attributes in predicting academic performance, the classifiers XGBoost, Random Forest, and Decision Tree assign coefficients to student features in such a way that the greater the value of the coefficient, the greater the weight or importance of the feature in the target. In order to interpret other types of classifiers, the shap-values [39] technique was used. This technique allows calculating coefficients, called shap-values, for any classifier and determining the features that have the greatest weight in the predictive model generated by it. In this way, it is possible to superficially understand the underlying model generated by the classifier and to know the features that are most important or most related to the value of the predictions. In this context, Figure 10 presents the shap-values associated with the predictive model generated by the XGBoost classifier, which achieved the highest performance indicators.

In Figure 10, a set of points is distributed according to their shap-value on a specific number line, for the 20 most important characteristics for the prediction of the model generated with XGBoost. On the other hand, each point is assigned a color defined in a color scale that represents the magnitude of the value of the characteristic in a register; this allows us to visually find patterns in the figure. For example, Failed Courses contains the highest shape-values magnitude (both in positive and negative magnitude).

Then, it can be seen that the highest Failed Courses values (with colors tending to red) are located towards the negative end of the shap-values, i.e., they have a more negative impact on the prediction value. Conversely, the lower Failed Courses values (with colors tending to blue), the higher the shap-values. This indicates that there is a negative correlation between the impact and the Failed Courses feature values—that is, the higher the feature value, the greater its negative impact on the negative axis. In addition, it can be interpreted that the greater the magnitude of Failed Courses, the lower the prediction value, which in the case of this study, tends to be classified as Regular student.

Finally, in Figure 10, it can be observed that the following characteristics of greater importance for the predictive model are Second Year Score and First Year Score with a positive correlation between their values and their impact (the higher the grade, the higher the value of the predicted performance tending to a Good student), followed by age and Career Architecture with negative correlation. Then, characteristics related to careers and faculties are found with considerably low magnitudes of importance in relation to the first three characteristics of greater importance. From the fourth place down, the importance list varies by the classifier, where places are swapped in the importance list features such as age; sex; Time to Graduate; features related to Payments Schemes; and some careers such

as Architecture, Civil Engineering, Psychology, and Nursing, among others. In addition, some faculties also appear, such as the Faculty of Theology, Faculty of Engineering and Architecture, and Faculty of Business Studies.

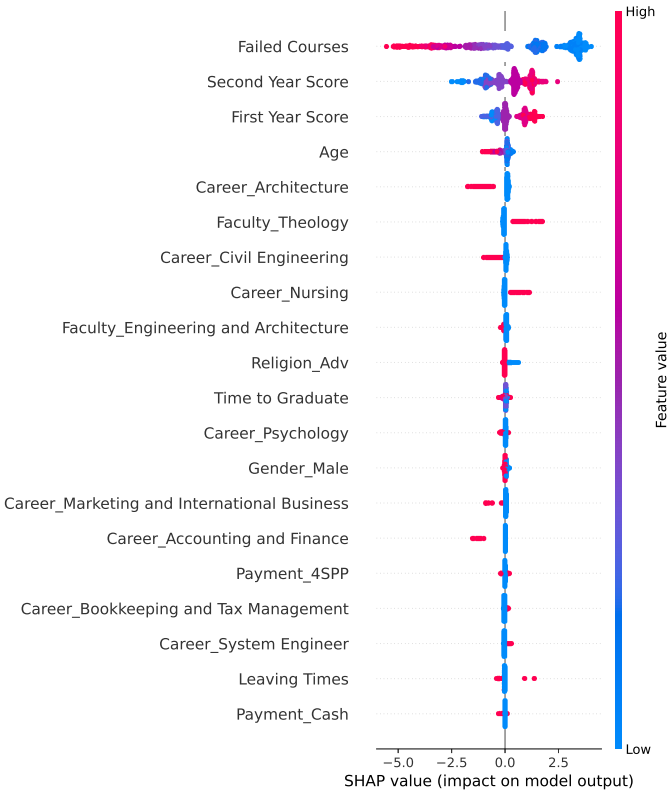

Figure 10. Shap-values for XGBoost, indicating the most important features of the predictive model generated by this classifier.

5. Discussion

There are several studies in the literature that address the prediction of academic performance using machine learning techniques [6,38,40–43], in which it has been possible to determine that there are various factors that influence the academic performance of a student. As has been seen in [40,42], the academic trajectory of students plays a fundamental role in predicting their academic performance. As it has been corroborated in this study, the number of Failed Courses and the First Year Score and Second Year score obtained by a student corresponds to a relevant factor in the prediction of their future academic performance, as it has also been corroborated in many studies such as [40].

In Table 2, it is observed that the XGBoost algorithm provides a high rate in Accuracy 0.91, Precision 0.94, and Sensitivity 0.94, developed with 30 executions.It is important to consider the shape-values technique to identify the features with greater importance for the prediction of academic success, due to the magnitude of the value, which indicates the strength with which the corresponding feature influences the decision-making process, according to success stories in [44,45]. In this sense, Figure 10 presents the coefficients associated with student attributes, where the higher the value of the coefficient, the greater magnitude of importance of the attribute in the model.

Consequently, it was identified that the attributes with the greatest magnitude of importance correspond to Failed Courses, First Year Score, and Second Year Score due to their relationship with the academic performance of students. Despite these well-known relationships, other important attributes are related to student performance such as student's professional career, where careers such as Career_Architecture, Career_Civil Engineering, Career_Psychology, and Career_Nursing are important to the model. This situation is

consistent with the distribution of academic performance according to professional career presented in Figure 3, where it is observed that the most important professional careers for the predictive models are those in which students have the worst grade point average. In the same way, the importance of the Failed Courses attribute is also linked to the distribution of Failed Courses shown in Figure 4. This leads to determine that the more difficult careers for students are more important for predictive models. All these findings allows us to use and validate with a study case the proposed visual-predictive approach, in order to give insights about academic performance in the Peruvian university, and the main factors related to regular-performing students and good-performing students.

Besides the relation between academics attributes and student performance, we analyzed the predictions made using XGBoost and found a particular scenario that has to be considered for the Peruvian university if they use our proposed visual-predictive data analysis approach. This scenario is related to Type II Error, which occurs when a student is predicted to have a good performance while he actually has a bad performance. From the testing dataset, those students that ended with poor performance (with a Fifth Year Score less than or equal to 12), 18% of them (8 of 44) were classified as good while having regular performance. This error is dangerous for the Peruvian university as they may not support some student that require attention in order to improve their performance, and in some cases, those students can drop their career or be academically eliminated.

Thus, it is important to verify that those students facing incorrect prediction do present low values on features related to their academic performance such as Failed Courses, Failed Courses 2 plus times, and Time to Graduate or Dropouts. Therefore, it is important to consider that these features could lead to Type II Error in this scenario, and it must be a critical consideration in a future LMS for the Peruvian university design. It may be presented as a special dashboard or list of students that have not been considered as Regular-performing students, and need to be considered in the application of an academic support strategy by the institution.

It is worth mentioning that this study has some limitations. It is the first time these data have been explored in a scientific report; there are no previous results in the literature. On the other hand, we have some restrictions on access to demographic, financial, and other non-academic-related variables. We believe that the analysis and approach would be more robust with the inclusion of these variables. In this sense, the scope would be more clarifying, and the findings could help decision-making-related problems in the teaching–learning process.

6. Conclusions and Future Scope

The results obtained allow us to affirm that the selected machine learning techniques presented an efficient predictive capacity, mainly due to the linear relationship between academic features and student performance. Eleven different models were used. However, among all these, the XGBoost is recommended because it shows better performance indicators, both in Accuracy as well as in Precision and Sensitivity. The key factors for classifying the academic performance of a student at the Peruvian university are the number of Failed Courses, together with the grades of the first and second year, which are decisive—that is, if we want the student to have a performance. Optimal action must be taken in the first two years and corrected so that in the following years, a good academic performance can be maintained according to the forecast of the model. This visual-predictive model represents a promising alternative for the identification of factors involved in academic performance for the Peruvian university, which does not have any identification method yet.

There are studies that have demonstrated the existence of factors external to the academic trajectory, influencing in an important way the academic performance of the students [46]. In this sense, another look at the contribution of this study corresponds to the fact of introducing additional factors, such as financial and demographic indicators, to the analysis. In this way, for future work, information on the academic curriculum, and demographic and socioeconomic data of the students can also be incorporated. The

incorporation of these factors enriches the mechanism for detecting students with academic problems, providing valuable information to develop strategies and activities for students who need to increase their academic performance. Analogously, the tools presented in this study make it possible to measure academic performance under categories, and this places universities at a higher level of maturity, as referred to in [47]. Finally, the proposed visual-predictive approach can be considered for an initial LMS for the Peruvian university in order to systematize the academic support to their students.

Author Contributions: Conceptualization, D.O.G. and J.L.L.-G.; methodology, D.O.G., J.L.L.-G. and R.S.; software, R.S. and J.U.; validation, J.L.L.-G. and R.S.; formal analysis, J.L.L.-G., R.S. and R.T.; investigation, D.O.G. and J.L.L.-G.; resources, D.O.G., J.L.L.-G., R.S. and R.T.; data curation, J.U.; writing—original draft preparation, D.O.G., J.L.L.-G., R.S., J.U. and R.T.; writing—review and editing, D.O.G., J.L.L.-G., R.S., J.U. and R.T.; visualization, D.O.G., J.L.L.-G. and J.U.; supervision, J.L.L.-G. and R.S.; project administration, J.L.L.-G. and R.S.; funding acquisition, J.L.L.-G. and R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially founded by Chilean ANID FONDECYT grant number 1221938 and ANID-Millennium Science Initiative Program ICN2021-004.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ramis, L.J.G. *Los Retos del Cambio Educativo*; Editorial Pueblo y Educación: Las Tunas, Cuba, 2021.
2. Rojas-Bujaico, R.W.; Huamán-Samaniego, H.; Medina-Castro, D.H.; Arauco-Esquivel, S. Modelo de la calidad de propósitos articulados de programas de estudios universitarios. *Ing. Ind.* **2021**, *42*, 1–19.
3. Pachas, M.; Castañeda, E.; Garro, L.; Aliaga, A.; Prado, H. La gestión institucional según los compromisos de desempeño: 2016-2018, Unidad de gestión educativa local 03–Lima. *Int. J. Inf. Res. Rev.* **2020**, *07*, 6714–6719.
4. Albreiki, B.; Zaki, N.; Alashwal, H. A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Educ. Sci.* **2021**, *11*, 552. [CrossRef]
5. Romero, C.; Ventura, S.; Pechenizkiy, M.; Baker, R.S. *Handbook of Educational Data Mining*; Chapman & Hall/CRC Data Mining and Knowledge Discovery Series; CRC Press: Boca Raton, FL, USA, 2010.
6. Vital, T.P.; Sangeeta, K.; Kumar, K.K. Student Classification Based on Cognitive Abilities and Predicting Learning Performances Using Machine Learning Models. *Int. J. Comput. Digit. Syst.* **2021**, *10*, 63–75. [CrossRef]
7. Bhutto, E.S.; Siddiqui, I.F.; Arain, Q.A.; Anwar, M. Predicting students' academic performance through supervised machine learning. In Proceedings of the 2020 International Conference on Information Science and Communication Technology (ICISCT), Karachi, Pakistan, 8–9 February 2020; pp. 1–6.
8. Tsiakmaki, M.; Kostopoulos, G.; Kotsiantis, S.; Ragos, O. Transfer learning from deep neural networks for predicting student performance. *Appl. Sci.* **2020**, *10*, 2145. [CrossRef]
9. Kim, B.H.; Vizitei, E.; Ganapathi, V. *GritNet: Student Performance Prediction with Deep Learning*; Cornell University: Ithaca, NY, USA, 2018.
10. Waheed, H.; Hassan, S.U.; Aljohani, N.R.; Hardman, J.; Alelyani, S.; Nawaz, R. Predicting academic performance of students from VLE big data using deep learning models. *Comput. Hum. Behav.* **2020**, *104*, 106189. [CrossRef]
11. Pérez-Suasnavas, A.L.; Cela, K.; Hasperué, W. Beneficios del uso de técnicas de minería de datos para extraer y analizar datos de twitter aplicados en la educación superior: Una revisión sistemática de la literatura. *Teoría Educ. Rev. Interuniv.* **2020**, *32*, 181–218. [CrossRef]
12. Mancilla-Vela, G.; Leal-Gatica, P.; Sánchez-Ortiz, A.; Vidal-Silva, C. Factores asociados al éxito de los estudiantes en modalidad de aprendizaje en línea: Un análisis en minería de datos. *Form. Univ.* **2020**, *13*, 23–35. [CrossRef]
13. Kanetaki, Z.; Stergiou, C.; Bekas, G.; Troussas, C.; Sgouropoulou, C. A Hybrid Machine Learning Model for Grade Prediction in Online Engineering Education. *Int. J. Eng. Pedagog.* **2022**, *12*, 4–24. [CrossRef]
14. Aluko, R.O.; Adenuga, O.A.; Kukoyi, P.O.; Soyingbe, A.A.; Oyedeji, J.O. Predicting the academic success of architecture students by pre-enrolment requirement: Using machine-learning techniques. *Constr. Econ. Build.* **2016**, *16*, 86–98. [CrossRef]
15. Nti, I.K.; Akyeramfo-Sam, S.; Bediako-Kyeremeh, B.; Agyemang, S. Prediction of social media effects on students' academic performance using Machine Learning Algorithms (MLAs). *J. Comput. Educ.* **2022**, *9*, 195–223. [CrossRef]

16. Alloghani, M.; Al-Jumeily, D.; Hussain, A.; Aljaaf, A.J.; Mustafina, J.; Petrov, E. Application of machine learning on student data for the appraisal of academic performance. In Proceedings of the 2018 11th International Conference on Developments in eSystems Engineering (DeSE), Cambridge, UK, 2–5 September 2018; pp. 157–162.
17. Dabhade, P.; Agarwal, R.; Alameen, K.; Fathima, A.; Sridharan, R.; Gopakumar, G. Educational data mining for predicting students' academic performance using machine learning algorithms. *Mater. Today Proc.* **2021**, *47*, 5260–5267. [CrossRef]
18. Yağcı, M. Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* **2022**, *9*, 1–19. [CrossRef]
19. Tomasevic, N.; Gvozdenovic, N.; Vranes, S. An overview and comparison of supervised data mining techniques for student exam performance prediction. *Comput. Educ.* **2020**, *143*, 103676. [CrossRef]
20. Huynh-Cam, T.T.; Chen, L.S.; Le, H. Using decision trees and random Forest algorithms to predict and determine factors contributing to first-Year University students' learning performance. *Algorithms* **2021**, *14*, 318. [CrossRef]
21. Abubakar, Y.; Ahmad, N.B.H. Prediction of Students' Performance in E-Learning Environment Using Random Forest. *Int. J. Innov. Comput.* **2017**, *7*, 1–5. [CrossRef]
22. Hellas, A.; Ihantola, P.; Petersen, A.; Ajanovski, V.V.; Gutica, M.; Hynninen, T.; Knutas, A.; Leinonen, J.; Messom, C.; Liao, S.N. Predicting academic performance: A systematic literature review. In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018.
23. Hamoud, A.; Hashim, A.S.; Awadh, W.A. Predicting student performance in higher education institutions using decision tree analysis. *Int. J. Interact. Multimed. Artif. Intell.* **2018**, *5*, 26–31. [CrossRef]
24. Chen, W.K.; Chen, L.S.; Pan, Y.T. A text mining-based framework to discover the important factors in text reviews for predicting the views of live streaming. *Appl. Soft Comput.* **2021**, *111*, 107704. [CrossRef]
25. Ahuja, R.; Sharma, S. Exploiting Machine Learning and Feature Selection Algorithms to Predict Instructor Performance in Higher Education. *J. Inf. Sci. Eng.* **2021**, *37*, 993–1009.
26. Baashar, Y.; Alkawsi, G.; Ali, N.; Alhussian, H.; Bahbouh, H.T. Predicting student's performance using machine learning methods: A systematic literature review. In Proceedings of the 2021 International Conference on Computer & Information Sciences (ICCOINS), Kuching, Malaysia, 13–15 July 2021; pp. 357–362.
27. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. *Knowledge Discovery and Data Mining: Towards a Unifying Framework*; KDD: Portland, OR, USA, 1996; Volume 96, pp. 82–88.
28. Dias Ledesma, S.K. Patrones de Consumo de drogas lícitas e ilícitas y su Influencia en el Rendimiento académico en una Comunidad Intercultural andina. Bachelor's Thesis, Universidad Estatal de Bolívar, Guanujo, Ecuador, 2020.
29. Marchesi, Á.; Tedesco, J.C.; Coll, C. *Calidad, Equidad y Reformas en la Enseñanza*; Fundación Santillana: Madrid, Spain, 2021.
30. Quintero, M.T.Q.; Vallejo, G.M.O. El desempeño académico: Una opción para la cualificación de las instituciones educativas. *Plumilla Educ.* **2013**, *12*, 93–115. [CrossRef]
31. Gueldner, B.A.; Feuerborn, L.L.; Merrell, K.W. *Social and Emotional Learning in the Classroom: Promoting Mental Health and Academic Success*; Guilford Publications: New York, NY, USA, 2020.
32. Walton, G.M.; Wilson, T.D. Wise interventions: Psychological remedies for social and personal problems. *Psychol. Rev.* **2018**, *125*, 617. [CrossRef] [PubMed]
33. Santos, B.; Yobany, H. Transición Demográfica en Honduras y su Incidencia en el Desarrollo. Ph.D. Thesis, Universidad Nacional Autónoma de Honduras, Tegucigalpa, Honduras, 2021.
34. Rodríguez Rodríguez, D.; Guzmán Rosquete, R. Rendimiento académico y factores sociofamiliares de riesgo. Variables personales que moderan su influencia. *Perfiles Educ.* **2019**, *41*, 118–134. [CrossRef]
35. Chang-Rodríguez, E. *Diásporas Chinas a las Américas*; Fondo Editorial de la PUCP: Lima, Peru, 2015.
36. Romagnoli, C.; Cortese, I. *¿Cómo la Familia Influye en el Aprendizaje y Rendimiento Escolar*; VALORAS: Santiago, Chile, 2015.
37. Helal, S.; Li, J.; Liu, L.; Ebrahimie, E.; Dawson, S.; Murray, D.J.; Long, Q. Predicting academic performance by considering student heterogeneity. *Knowl.-Based Syst.* **2018**, *161*, 134–146. [CrossRef]
38. Mueen, A.; Zafar, B.; Manzoor, U. Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *Int. J. Mod. Educ. Comput. Sci.* **2016**, *8*, 36–42. [CrossRef]
39. Cohen, S.; Ruppin, E.; Dror, G. Feature Selection Based on the Shapley Value. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, UK, 30 July–5 August 2005; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2005; pp. 665–670.
40. Kabakchieva, D. Predicting student performance by using data mining methods for classification. *Cybern. Inf. Technol.* **2013**, *13*, 61–72. [CrossRef]
41. Meedech, P.; Iam-On, N.; Boongoen, T. Prediction of student dropout using personal profile and data mining approach. In *Intelligent and Evolutionary Systems*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 143–155.
42. Patel, K.D.; Suthar, A.B. Recommendations for Student Performance Improvement Based on Result Data Using Educational Data Mining. In *Inventive Systems and Control*; Suma, V., Chen, J.I.Z., Baig, Z., Wang, H., Eds.; Springer: Singapore, 2021; pp. 403–411.
43. Cortez, P.; Silva, A. Using Data Mining to Predict Secondary School Student Performance. In Proceedings of the 5th Annual Future Business Technology Conference, Porto, Portugal, 5–12 April 2008.
44. Smith, M.; Alvarez, F. Identifying mortality factors from Machine Learning using Shapley values—A case of COVID19. *Expert Syst. Appl.* **2021**, *176*, 114832. [CrossRef]

45. Tideman, L.E.; Migas, L.G.; Djambazova, K.V.; Patterson, N.H.; Caprioli, R.M.; Spraggins, J.M.; Van de Plas, R. Automated Biomarker Candidate Discovery in Imaging Mass Spectrometry Data Through Spatially Localized Shapley Additive Explanations. *Anal. Chim. Acta* **2021**, *1177*, 338522. [CrossRef]
46. Saa, A.A. Educational data mining & students' performance prediction. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 212–220.
47. Tocto-Cano, E.; Paz Collado, S.; López-Gonzales, J.L.; Turpo-Chaparro, J.E. A Systematic Review of the Application of Maturity Models in Universities. *Information* **2020**, *11*, 466. [CrossRef]