UNIVERSIDAD PERUANA UNIÓN

ESCUELA DE POSGRADO

Unidad de Posgrado de Ingeniería y Arquitectura

Enfoque de análisis de datos visual - predictivo para el desempeño académico de los estudiantes de una universidad peruana

Tesis para obtener el Grado Académico de Maestro en Ingeniería de Sistemas con mención en Dirección y Gestión de Tecnologías de Información

Autor:

David Leandro Orrego Granados

Asesor:

M.Sc. Javier Linkolk López Gonzales

Lima, octubre del 2021

DECLARACIÓN JURADA DE AUTORÍA DE TESIS

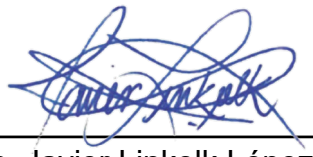
Javier Linkolk López Gonzales, de la Escuela de Posgrado, Unidad de Posgrado de ingeniería y arquitectura, de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: **"Enfoque de análisis de datos visual - predictivo para el desempeño académico de los estudiantes de una universidad"** constituye la memoria que presenta el Licenciado David Leandro Orrego Granados para aspirar al Grado Académico de Maestro(a) en Ingeniería de Sistemas con mención en Dirección y Gestión de Tecnologías de Información, cuya tesis ha sido realizada en la Universidad Peruana Unión bajo mi dirección.

Las opiniones y declaraciones en este informe son de entera responsabilidad del autor, sin comprometer a la institución.

Y estando de acuerdo, firmo la presente declaración en la ciudad de Lima, a los 13 días del mes de octubre del año 2021



M.Sc. Javier Linkolk López Gonzales

ACTA DE SUSTENTACIÓN DE TESIS DE MAESTRO(A)

En Lima, Ñaña, Villa Unión, a 12 días del mes de octubre del año 2021, siendo las 03:00 p.m, se reunieron en la modalidad online sincrónica, bajo la dirección del Señor Presidente del Jurado: Mg. Sergio Omar Valladares Castillo, el secretario: Mg. Nemias Saboya Rios, los demás miembros: Dra. Erika Inés Acuña Salinas y el Mg. Omar Leonel Loaiza Jara y el asesor: M.Sc. Javier Linkolk Lopez Gonzales, con el propósito de administrar el acto académico de sustentación de Tesis de Maestro(a) titulada: Enfoque de análisis de datos visual - Predictivo para el desempeño académico de los estudiantes de una universidad peruana del Bachiller/Licenciado(a) David Leandro Orrego Granados

Conducente a la obtención del Grado Académico de Maestro(a) en: Ingeniería de Sistemas

(Nomenclatura del Grado Académico)

con Mención en Dirección y Gestión de Tecnologías de Información El Presidente inició el acto académico de sustentación invitando al candidato hacer uso del tiempo determinado para su exposición. Concluida la exposición, el Presidente invitó a los demás miembros del Jurado a efectuar las preguntas, cuestionamientos y aclaraciones pertinentes, los cuales fueron absueltos por el candidato. Luego se produjo un receso para las deliberaciones y la emisión del dictamen del Jurado.

Posteriormente, el Jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Bachiller/Licenciado (a): David Leandro Orrego Granados

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	18	A-	Muy bueno	Sobresaliente

(*) Ver parte posterior

Finalmente, el Presidente del Jurado invitó al candidato a ponerse de pie, para recibir la evaluación final. Además, el Presidente del Jurado concluyó el acto académico de sustentación, procediéndose a registrar las firmas respectivas.

_____	_____
Presidente	Secretario

_____	_____	_____
Asesor	Miembro	Miembro

_____

Bachiller/Licenciado(a)

Visual-predictive data analysis approach for the academic performance of students from a Peruvian university

DAVID ORREGO GRANADOS[1,*], JONATHAN UGALDE[2,*], RODRIGO SALAS[3], (Senior Member, IEEE), ROMINA TORRES[4], AND JAVIER LINKOLK LÓPEZ-GONZALES[5,6], (Member, IEEE)

[1]UPG Ingeniería y Arquitectura, Escuela de Posgrado, Universidad Peruana Unión, Peru (e-mail: david.orrego@upeu.edu.pe)
[2]Escuela de Ingeniería Informática, Universidad de Valparaíso, Chile (e-mail: jonathan.ugalde@postgrado.uv.cl)
[3]Escuela de Ingeniería C. Biomédica, Universidad de Valparaíso, Chile (e-mail: rodrigo.salas@uv.cl)
[4]Facultad de Ingeniería, Universidad Andres Bello, Viña del Mar, Chile (e-mail: romina.torres@unab.cl)
[5]E.P. Ingeniería Ambiental, Facultad de Ingeniería y Arquitectura, Universidad Peruana Unión, Perú
[6]Instituto de Estadística, Universidad de Valparaíso, Valparaíso 2360102, Chile (e-mail: javier.lopez@postgrado.uv.cl)

Corresponding authors: Javier Linkolk López-Gonzales (e-mail: javier.lopez@postgrado.uv.cl) and Rodrigo Salas (e-mail: rodrigo.salas@uv.cl).

* This authors contributed equally in this paper.

ABSTRACT The academic success of university students is a result that depends in a multi-factorial way on the aspects related to the student and the career itself. In this work, we carry out a visual analysis of the data to obtain relevant information regarding the academic performance of students from a Peruvian university. This study was complemented with the construction of machine learning models to provide a predictive model of the students' academic success. In specific, the XGBoost Machine Learning method achieved a performance of up to 91.5% of Accuracy. In this sense, this study offers a novel visual-predictive data analysis approach as a valuable tool for developing and targeting policies to support students with lower academic performance or to stimulate advanced students. The results obtained allow us to identify the relevant variables associated with the students' academic performances. Moreover, we were able to give some insight into the academic situation of the different careers of the University.

INDEX TERMS Students' Performances, Machine Learning, Learning analytics, Educational Data Mining, Business Intelligence in Education

I. INTRODUCTION

The educational system in Peru is one of the fundamental factors for the development and growth of the country [1]–[4]. 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 of Evaluation, Accreditation, and Certification of Educational Quality was created in order to certify a quality education for all students, evidencing essential quality aspects after a rigorous evaluation of specific standards and indicators [5], [6]. Peru is a country that seeks to improve its educational system because it ranks 127th out of 137 countries studied in educational quality [7]. In [8] mentions that teaching in Peruvian society focuses on the fact that the teacher is everything and attributes many social functions to her/him.

Many authors have addressed the problem of predicting student performance using machine learning algorithms or Artificial Neural Networks. In [9] uses the KDD process to collect and prepare student's data through an online learning management system (LMS) and apply Logistic regression and Support Vector Machine (SVM), achieving an accuracy of 73% and 79% respectively. [10] uses data collected from a college database and predicts students' performance using Naive Bayes, Classification Trees (CTs), k-NN, C4.5, and SVM, where K-NN outperforms the other classifiers with an accuracy of 100%. On the other hand, [11] uses student's

data from an online LMS from four Greek university courses and predicts student's performance through a Deep Neural Network applying transfer Learning, from data of one course to another, achieving an accuracy of 86%. In [12] propose a new Deep Learning-based algorithm, called GritNet, an evolution of the bidirectional long short term memory (BiLSTM), for student's performance classification. Likewise, in [13] generated a Deep Artificial Neural Network on data extracted from an online LMS to predict risky students for early intervention, achieving classification accuracies from 84% to 93%.

The main contribution of this work is to provide a visual-predictive data analysis that allows identifying 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. In this sense, this study offers a novel approach for higher education institutions in the country and the South American region. It will allow 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 support policies in student accompaniment.

In what follows, the structure of the study consists of: Section II shows the theoretical foundations. Section III details the methodology used. The results obtained are presented in section IV. Section V contrasts the discussion between the results obtained and other similar studies. Finally, section VI describes the conclusions and future work.

II. MACHINE LEARNING TECHNIQUES

Machine Learning methods are data-driven computational models that are able to generalize and predict future values with high precision. This techniques have been successfully applied in areas such as education, commerce and marketing [14], web services [15], social network analysis [16], medical diagnostics [17], [18], finance [19], air quality [20], transportation [21], hydrology [22], and bioinformatics [23] among others.

In this article, we considered the machine learning techniques that are described below:

1) **K-Nearest Neighbors:**
   The KNN algorithm computes the distance between from the new sample to the all the data of the training set. The method selects the closest k points and it assigns the class using a majority voting rule.
   It is a non-parametric algorithm used for classification and regression. In both cases, it is based on choosing from a set of k elements, those found to catalog the object itself or make a regression. For this study, the output will be based on classification. In that sense, it will be a categorical variable. So, the object is going to be classified by the majority of its neighbors. The more neighbors the object has nearby, the vote will indicate that it belongs to that category [24], [25].
2) **Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA):**
   The LDA and QDA are statistical classification model that separate the observations of two or more classes of objects. This methods used the Bayes theorem combined with the assumption of normality distribution of the classes. The resulting models are linear (LDA) and quadratic (QDA) discriminant functions, where they induce a linear or quadratic decision boundary, respectively [26].
3) **Decision trees:**
   Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the characteristics of the data [27], [28].
   Its structure is based on a set of nodes, branches, and leaves. The ranges of x give the edges whose nodes are labeled by an attribute x, and a class labels each leaf [29]–[31]. These nodes are the points from which the branches come out where the tree branches out. The point of origin is called the root node, where the tree is born. On the contrary, a node that does not have any branch that follows the tree is called the leaf node. Typically, a decision tree starts with a single root node, then proceeds with a set of decision nodes. Ultimately, each ends at a terminal node, where the final decision rule is carried out, with classification being performed. All nodes represent a single variable, and each branch represents the possible categories that this variable takes. Finally, the terminal node represents the decision that the algorithm is going to make. Therefore, the value it returns at the end of the execution [28].
4) **Support Vector Machine (SVM):**
   It is a supervised learning algorithm for classification and regression, which is quite versatile as it adjusts to linear and non-linear models due to the availability of its kernel functions [32].
   SVMs maximize the margin of the decision boundary by finding the support vectors in the training set. The learning process consists in an optimization process based on quadratic programming, that is, finding a line of separation, called hyperplane, between data of two classes. This line seeks to maximize the distance between the closest points each of the classes. This distance between the boundary hyperplane and the first points of each class is called the margin [32], [33].
5) **Random Forest:**
   It is a supervised learning technique based on creating numerous decision trees on a group of training data. The final prediction is obtained by aggregating the output of all the models. The result is an strong model that is capable of working under adverse circumstances

such as high dimensionality, complex interactions, and non-linear data structures [34]–[37].

Each tree is obtained based on two phases. In the first phase, a large number of decision trees are created with the data set. Each contains a random subset of variables m (predictors) such that $m < M$ (where $M =$ total predictors). For its part, in the second phase, each tree ascends to its maximum extent [37].

Each tree created by the algorithm contains a set of random observations (chosen by bootstrapping). The observations not considered in the trees (also known as out of the bag) are used to validate the model. The outputs of all the trees are aggregated into a final output.

In [38] mentions the advantages of this algorithm, allowing to explore both classification and prediction, achieving very efficient performance in large amounts of data. It also handles hundreds of predictors without any exclusions, manages to estimate which are the most important predictors, and maintains its precision with large proportions of missing data.

6) **Extreme Gradient Boosting (XGBoost):**

The XGBoost operation is based on an ensemble of decision trees known as Classification and Regression Tree (CART), distributing them sequentially in a pipeline of CARTs that are trained based on the result of the previous CART (initializing with an arbitrary y_0). The philosophy behind the XGBoost operation is to minimize the L loss function by correcting prediction errors from a previous CART to the next one in the pipeline [37].

XGBoost is a gradient descent algorithm whose objective, in theoretical terms, is to minimize a loss function $L(y, f(x))$ into the hyperspace of possible solutions $y \in Y$, using the gradient ∇L to decrease as fast as possible (using a greedy function) the value of this function. To do this, XGBoost follows the negative direction of the gradients ∇L associated with the x_i components of the set of attributes of an input $x \in X$ record of the function $f(x)$, which produces a y value given a record x [39].

In the execution of the algorithm, an initial CART F_0 is defined to find an initial solution y_0, this result is synchronized with a residual error ($e_0 = y_0 - F_0$). Then, a CART h_1 is built that considers the prediction errors of the previous step (e_0) to build a new CART F_1 harmonizing its prediction error indicators e_i. Finally, in each iteration of the algorithm, the previously described process is executed iteratively, training all the CARTs $F_i \in F$ of the pipeline [40].

The use of CARTs allows this algorithm to address both classification and regression problems flexibly and efficiently. In this sense, due to the excellent performances achieved by XGBoost in a wide variety of applications [37], [41], and its great adoption by data science researchers, this algorithm has been extensively used in the last years.

The XGBoost algorithm is based on decision trees and uses ensemble methods, enabling multiple learning algorithms to improve predictive performance. In addition, XGBoost is a gradient augmentation library that focuses on the parallel tree model that solves data science problems quickly and accurately. The method was designed to be highly efficient, flexible, and portable, which means that within of XGBoost, there are two distinct parts: A model consisting of trees and hyperparameters, and settings used to build the model [40], [41].

Typically, the XGBoost has a sequential ensemble of decision trees called Classification and Regression Trees (CART). The trees are added sequentially by taking as a reference the result of the previous trees and correct existing errors.

7) **Perceptron and Multilayer Perceptron:**

The perceptron is the simplest type of artificial neuron that exists. Its operation is based on receiving several inputs and producing a single binary output. This output is determined by a weighted sum, which will be less than or greater than some threshold value. In addition to its inputs, it introduces different weights that express the importance of the respective inputs for the output. As well as the weights, or threshold is an actual number that becomes a parameter of the neuron. This builds it into a binary classifier [42], [43].

The Multilayer perceptron is an artificial neural network where the neurons are placed in layers, and they are connected with feed-forward links between neurons of two consecutive layers. This model is able to capture the non-linearity of the data. The learning process of the MLP is an iterative process where the training patterns are presented to the network, and based on the errors obtained, the adjustments of the synaptic weights are made in order to reduce the error in the next iterations. Training is carried out in a supervised way through the learning rule that minimizes error. The mechanism used for learning in MLP is known as the error backpropagation algorithm or backpropagation. Each of these signals is distributed as an input signal for all the neurons of the next layer, starting in the input layer to the output layer. That is, the MLP network works as a sequence of simply interconnected perceptrons to generate the propagation of input signals passing through all layers to the output of the network. [44], [45].

III. VISUAL-PREDICTIVE DATA ANALYSIS SCHEME

In this work, we propose a Visual-Predictive Data Analysis scheme (VPDA) 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 corresponds to an adaptation of the well-known [46] Knowledge Discov-

ery Database (KDD) and CRISP-DM methodologies for the educational data mining field.

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 methodology 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.

A. 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 [47]. This performance is the context of the interactions that occur between students, teachers, and the knowledge that circulates daily [48]. It is also pointed out that this variable is the one that determines how beneficial the study was within the University. In [49] mentions 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 [50], [51].

The academic performance is given through different factors that, together, generate the result of the expected academic performance of the students [48]. This performance is not only based on the degree of education. Moreover, it also implies the economic, social, and family 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 [4], [52]–[54]. Likewise, [55]–[57] 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. Also, [58], [59] 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.

B. DATA UNDERSTANDING

The objective of this stage is to carry out a first 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 Leaving Times.

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 and 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. While 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, representing the objective of constructing the predictive model in the present study.

On the other hand, demographic attributes such as gender, age, marital status, and religion were also selected since this type of data accounts for information underlying the students' environment and can influence their academic performance.

Finally, the Payment Scheme attribute was considered as data that reflects the students' socioeconomic information. In this way, TABLE 1 presents all the selected attributes to perform the analysis and generate a predictive model of the academic performance of university students.

C. DATA PREPARATION

In this stage, the data is 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 were imputed with the mean values or the mode depending on if they were 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 data that corresponds. For example, for the categorical attribute Gender, the values "Female" and "F" were consolidated only with values "F", and the values "Male" and "M" only with values "M". While the values of the numerical attribute Performance were limited to 4 decimal places. The third criterion corresponds to

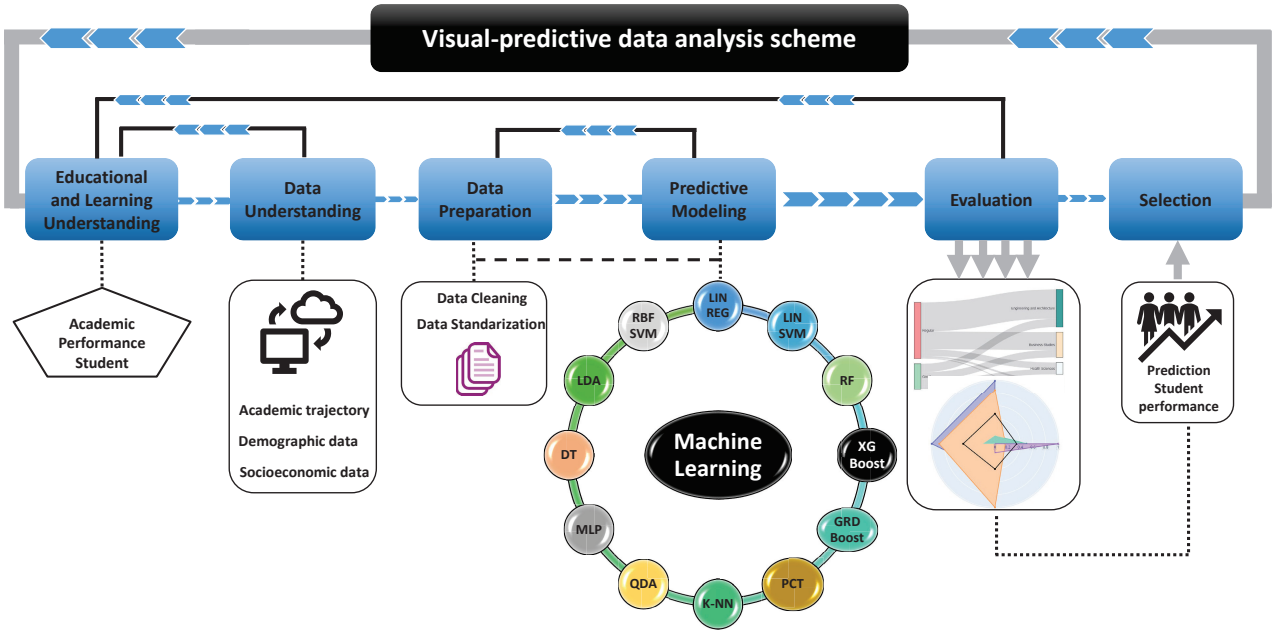


FIGURE 1. Visual-Predictive Data Analysis scheme for the academic performance of the university students.

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
Leaving Times	int	0 - 5
Career	category	X
First Year Score	float	8 - 19
Second Year Score	float	5 - 19
Performance	float	6 - 19
Performance Category	category	X
Payment Scheme	category	X

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 the 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:

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

D. PREDICTIVE MODELING

In this stage, the following machine learning models have been 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), Multilayer 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.

E. EVALUATION

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 F1-score, and they are described below.

$$Accuracy = \frac{TP + TN}{n}$$
$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$
$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

Accuracy measures the proportion of samples that are correctly classified. The precision measures the proportion of predictive positives that are correct. The Recall measures the sensitivity of the classifier to detect the positive cases. The F1-score is the harmonic mean of the Recall and the precision and gives a trade-off measure between the Recall and the precision.

2) Visual Data Analysis

Different types of charts and visual representations of data allows 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 or QlikView, among others. These tools, and many open-source libraries such as Matplotlib, Seaborn, Pyplot, a detailed visual analysis from simple charts to sophisticated and complex visual representations, such as georeferenced maps 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.

F. SELECTION

As a product of the application of each algorithm on the data of university students, a model is obtained that predicts the academic performance of new students. All the performance metrics are analyzed to select the best machine learning model. Finally, the performance of the model is evaluated in the student's classification as one of "Regular" or "Good" performance.

In this way, the generation of this predictive classification model means a mechanism for identifying students with high or low performance so that the University defines pertinent policies and strategies that enhance and accompany students according to the prediction of their academic performance.

IV. RESULTS

This section shows the results obtained when applying the VPDA approach to the performance data of the students obtained in a university in Peru. In section IV-A the results of the visual data analysis are shown, while in section IV-B the results obtained with the predictive model are detailed.

A. RESULTS OF THE VISUAL DATA ANALYSIS

FIGURE 2 shows the the correlations between attributes. As it is observed, there is a strong positive correlation between academic performance (target attribute) and the score obtained during the first two years, and a strong negative correlation with respect to the failed courses. However, this chart does not find a correlation between the student performances and the career of origin.

FIGURE 3 shows that the Faculty of Theology and 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. Quite the opposite in the race 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

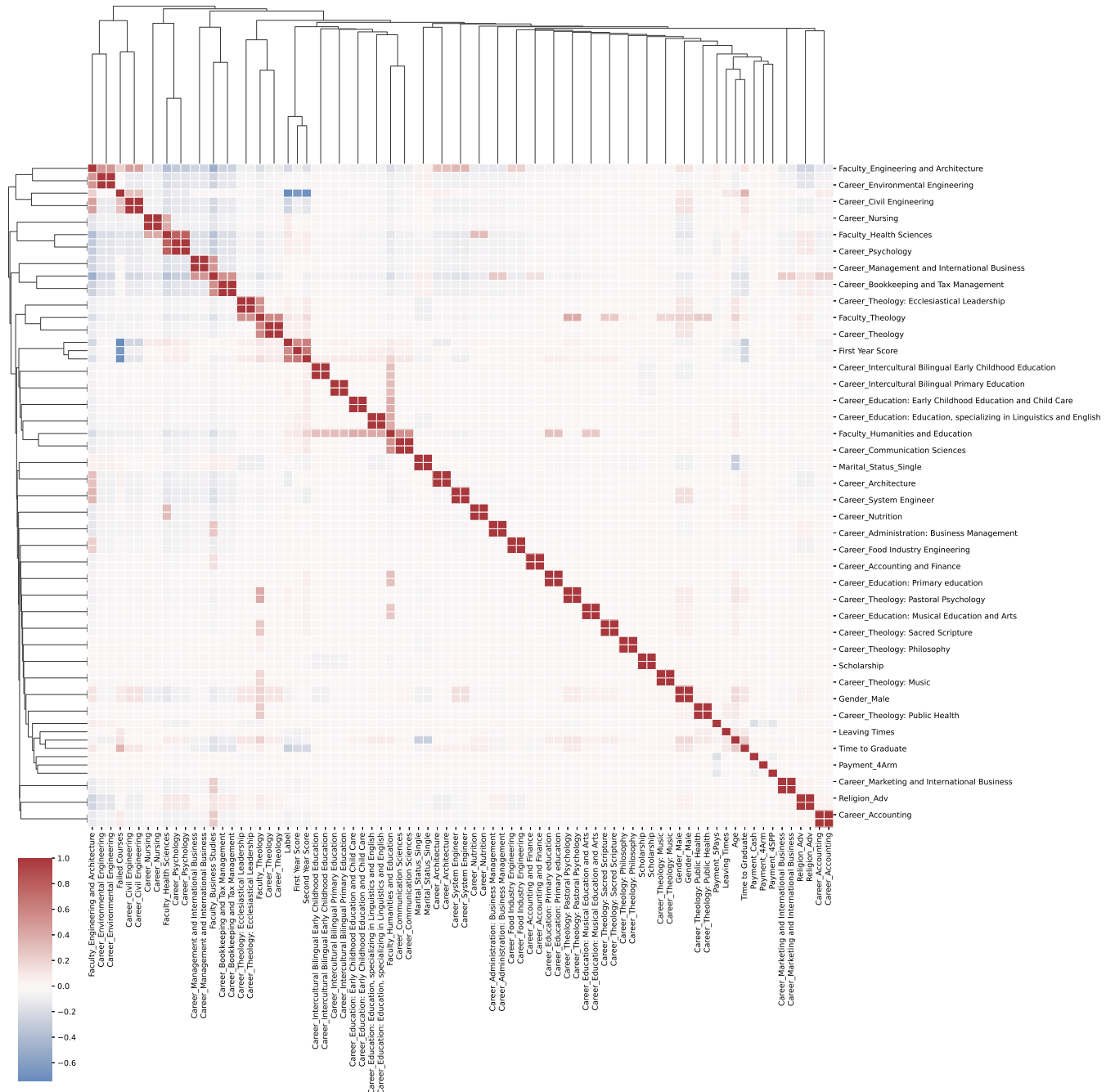


FIGURE 2. Student attribute correlation diagram.

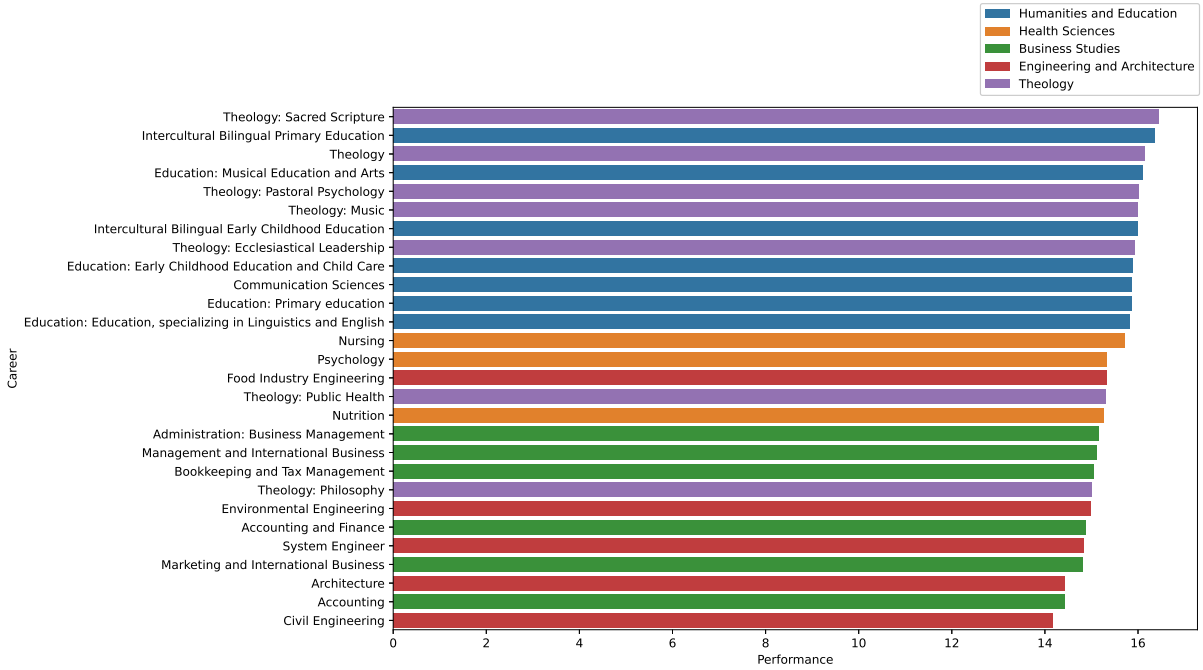


FIGURE 3. Average of qualifications according to Professional Career and Faculty.

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. While, 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 "Fair" academic performance.

FIGURE 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 the equations (2) and (3), where x_{std} corresponds to the standard deviation of the attribute and x_{scaled}, to the scaled value.

$$x_{scaled} = x_{std} \cdot (x_{max} - x_{min}) + x_{min} \tag{2}$$

$$x_{std} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{3}$$

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 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 by about "Regular" students. In addition, in Theology, there is also an Age rank much higher than the rest of the faculties.

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 would 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

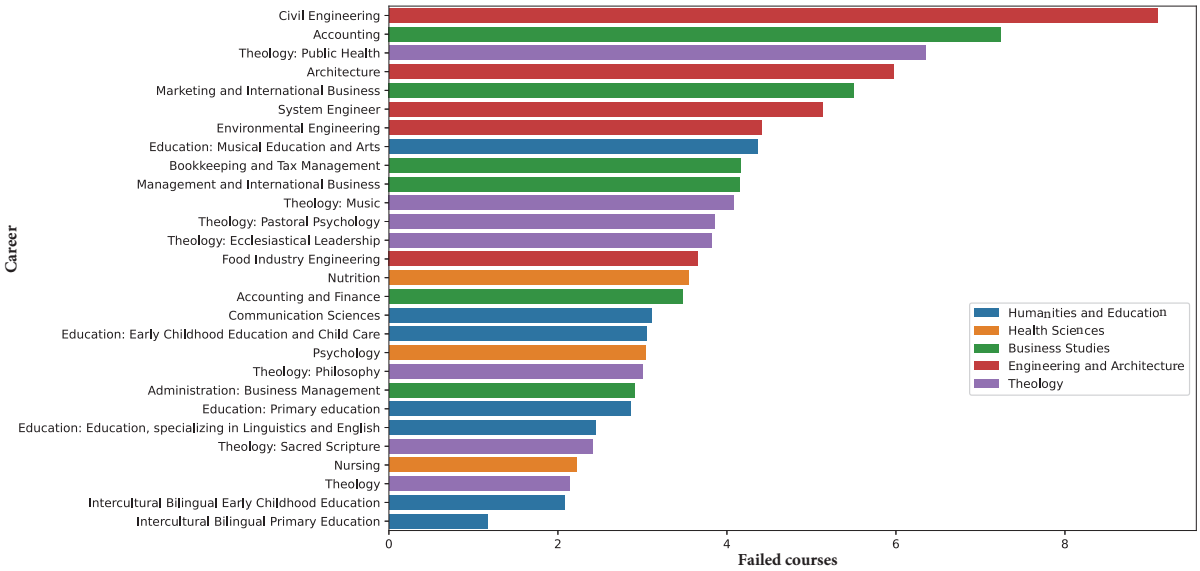


FIGURE 4. Average number of failed courses according to Professional Career and Faculty

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, 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.

B. 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 [60], [61].

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 the 86% of 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, where the F1 score was 0.9358. On the other hand, although there are classifiers with similar performance than 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 application 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 Regression	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 due to 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. Though, the linear classifiers such as LDA, Linear SVM, and Linear Regression show good performance.

In terms of the importance of student attributes in predicting academic performance, the classifiers XGBoost, Ran-

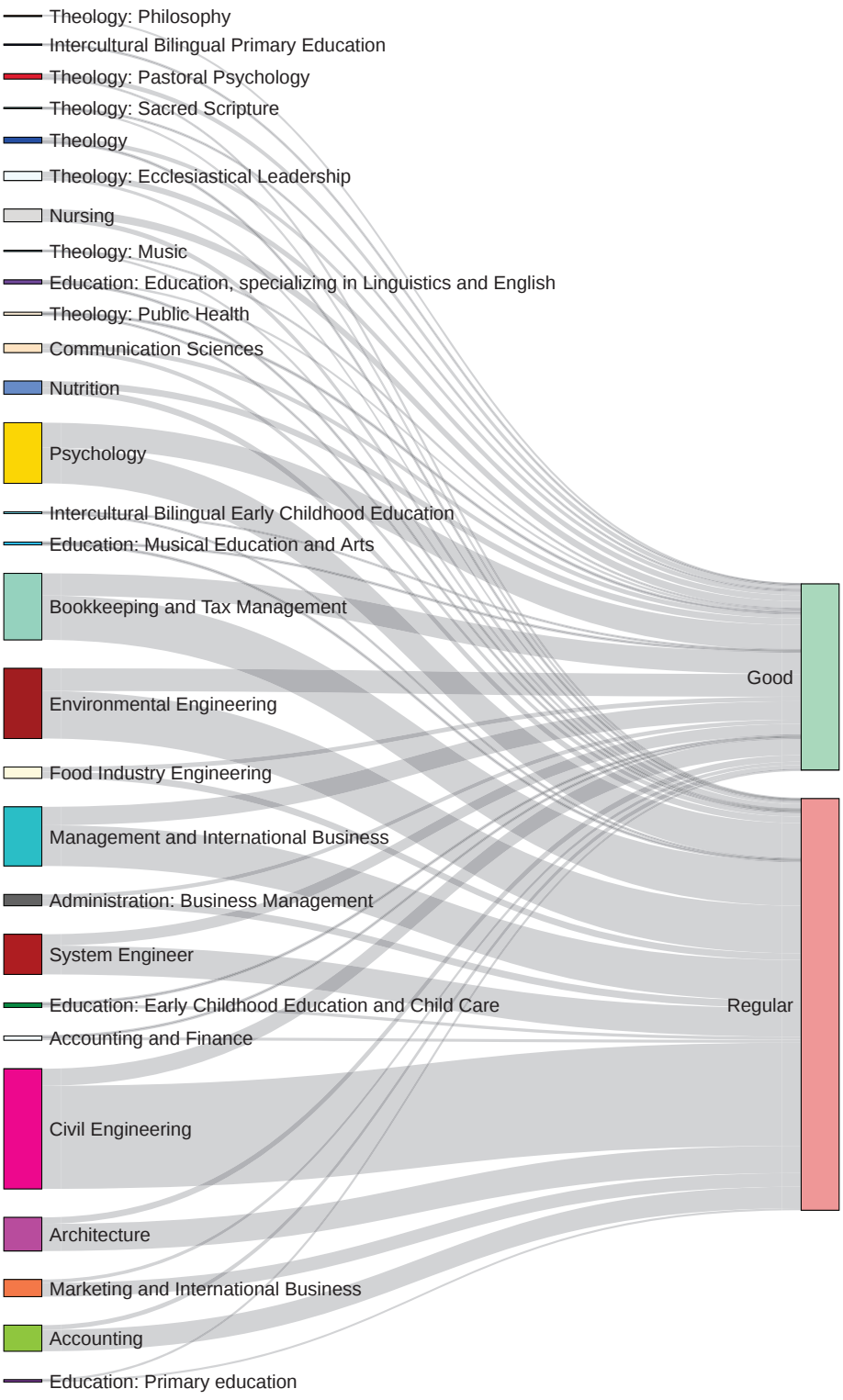


FIGURE 5. Student's Performance Classification distributed over career.

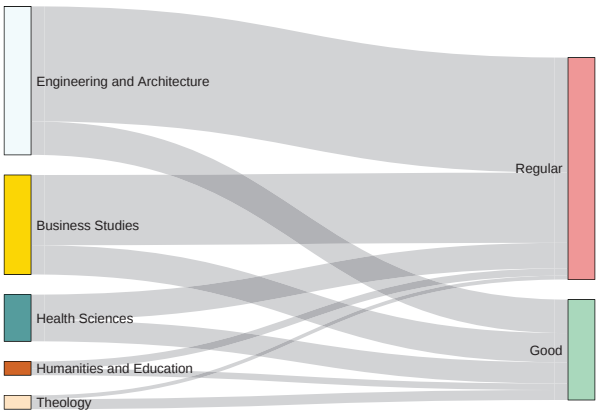


FIGURE 6. Student's Performance Classification distributed over faculty.

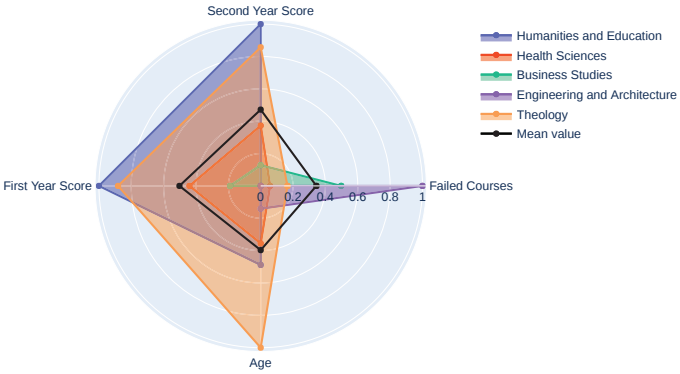


FIGURE 7. Principals factors by faculty. Distribution of the values of the most important attributes in the predictive model according to faculty.

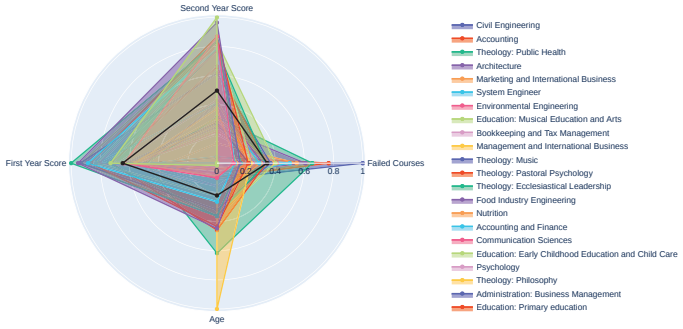


FIGURE 8. Principals factors by career. Distribution of the values of the most important attributes in the predictive model according to career.

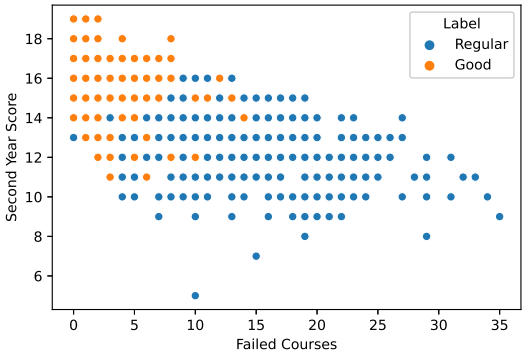


FIGURE 9. Scatter plot between Failed Courses y Second Year Score, the colors that represent each value of the target indicate that they are linearly separable.

dom 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 [62] 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 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, on another words, they have a more negative impact on the prediction value. While the lower Failed Courses values (with colors tending to blue) the higher 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, observing the 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

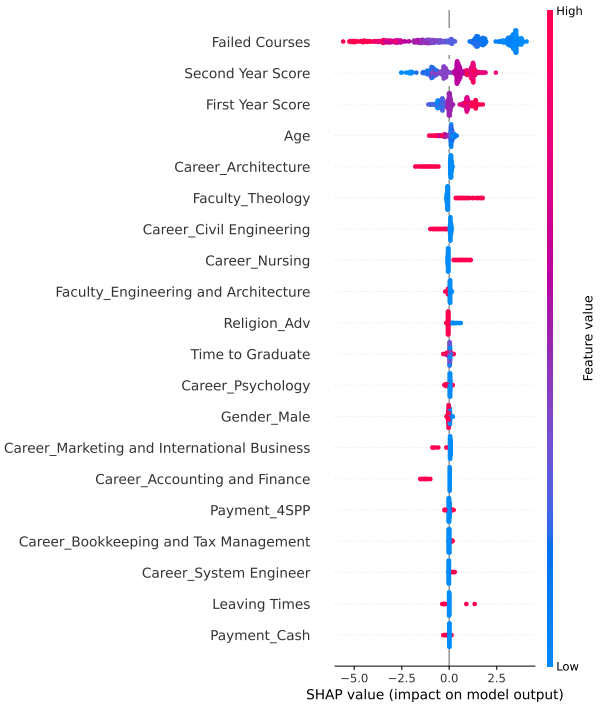


FIGURE 10. Shap values for XGBoost, indicating the most important features of the predictive model generated by this classifier.

features such as Age, Sex, Time to Graduate, features related to Payments Schemes, some careers such as Architecture, Civil Engineering, Psychology, Nursing among others. In addition, some faculties also appear, such as Faculty of Theology, Faculty of Engineering and Architecture, and Faculty of Business Studies.

V. DISCUSSION

There are several studies in the literature that address the prediction of academic performance using Machine Learning techniques [61], [63]–[67], in which it has been possible to determine that there are various factors that influence the performance academic of a student. As has been seen in [63], [66], the academic trajectory of students plays a fundamental role in predicting their academic performance. As it has been corroborated in this study, that the amount of Failed Courses and First Year Score and Second Year score obtained by a student corresponds to a relevant factor in the prediction of their future academic performance. It has also been corroborated in [63].

In TABLE 2 it is observed that the XGBoost algorithm provides a high rate both in accuracy 0.91, as in precision 0.94 and sensitivity 0.94, developed with 30 executions; this is contrasted with other studies that performed ten executions, showing the difference in the performance of each model [60], [68].

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 [69], [70]. In this sense, FIGURE 10 presents the coefficients associated with the attributes of said students, where the higher the value of the coefficient, the greater the 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. Other important attributes are related to the student's professional career, where careers such as Career_Arquitectura, Career_Civil Engineering, Career_Psicologia, Career_Nursing, are of great importance 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 in the same majors, as shown in FIGURE 4. This leads to determine that the more difficult careers for students are more important for predictive models.

VI. CONCLUSIONS

The results obtained allow us to affirm that the selected Machine Learning techniques presented an efficient predictive capacity, which represents a promising alternative for the identification of factors involved in academic 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 a 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.

There are studies that demonstrated the existence of factors external to the academic trajectory, influencing in an important way in the academic performance of the students [71]. 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, 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 and provides 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 [72].

REFERENCES

[1] A. B. Rubio, "Deserción universitaria en chile: incidencia del financiamiento y otros factores asociados," Revista CIS, vol. 9, no. 14, pp. 59–72, 2011.

[2] M. A. Miranda and J. Guzmán, "Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos," Formación universitaria, vol. 10, no. 3, pp. 61–68, 2017.

[3] S. C. Cáceres, P. Alvarez, M. L. Ortiz, and L. C. Collado, "Ci deserción universitaria: La epidemia que aqueja a los sistemas de educación superior," REVISTA PERSPECTIVA, vol. 20, no. 1, pp. 13–25, 2019.

[4] L. J. G. Ramis, Los retos del cambio educativo. Editorial Pueblo y Educación, 2021.

[5] R. W. Rojas-Bujaico, H. Huamán-Samaniego, D. H. Medina-Castro, and S. Arauco-Esquivel, "Modelo de la calidad de propósitos articulados de programas de estudios universitarios," Ingeniería Industrial, vol. 42, no. 1, pp. 1–19, 2021.

[6] OCCAA, "Marco de calidad de la educación superior universitaria," tech. rep., Universidad Nacional Mayor de San Marcos, Marzo 2018.

[7] M. Pachas, E. Castañeda, L. Garro, A. Aliaga, and H. Prado, "La gestión institucional según los compromisos de desempeño: 2016-2018, unidad de gestión educativa local 03–lima," International Journal of Information Research and Review, vol. 07, no. 2, pp. 6714–6719, 2020.

[8] F. Imbernón, Ser docente en una sociedad compleja: la difícil tarea de enseñar, vol. 50. Graó, 2017.

[9] E. S. Bhutto, I. F. Siddiqui, Q. A. Arain, and M. Anwar, "Predicting students' academic performance through supervised machine learning," in 2020 International Conference on Information Science and Communication Technology (ICISCT), pp. 1–6, IEEE, 2020.

[10] T. P. Vital, K. Sangeeta, and K. K. Kumar, "Student classification based on cognitive abilities and predicting learning performances using machine learning models," International Journal of Computing and Digital Systems, vol. 10, no. 1, pp. 63–75, 2021.

[11] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Transfer learning from deep neural networks for predicting student performance," Applied Sciences, vol. 10, no. 6, p. 2145, 2020.

[12] B.-H. Kim, E. Vizitei, and V. Ganapathi, GritNet: Student performance prediction with deep learning. Cornell University, 2018.

[13] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from vle big data using deep learning models," Computers in Human Behavior, vol. 104, p. 106189, 2020.

[14] J. Sánchez-Garcés, J. J Soria, J. E. Turpo-Chaparro, H. Avila-George, and J. L. López-Gonzales, "Implementing the reconac marketing strategy for the interaction and brand adoption of peruvian university students," Applied Sciences, vol. 11, no. 5, p. 2131, 2021.

[15] R. Torres, R. Salas, N. Bencomo, and H. Astudillo, "An architecture based on computing with words to support runtime reconfiguration decisions of service-based systems," Int. J. Comput. Intell. Syst., vol. 11, no. 1, pp. 272–281, 2018.

[16] E. Puraivan, E. Godoy, F. Riquelme, and R. Salas, "Fake news detection on twitter using a data mining framework based on explainable machine learning techniques," in 11th International Conference on Pattern Recognition Systems, pp. 1–6, 2021.

[17] J. S. Castro, S. Chabert, C. Saavedra, and R. Salas, "Convolutional neural networks for detection intracranial hemorrhage in ct images.," in CRoNe, pp. 37–43, 2019.

[18] S. Chabert, J. S. Castro, L. Muñoz, P. Cox, R. Riveros, J. Vielma, G. Huerta, M. Querales, C. Saavedra, A. Veloz, and R. Salas, "Image quality assessment to emulate experts' perception in lumbar mri using machine learning," Applied Sciences, vol. 11, no. 14, p. 15, 2021.

[19] R. Torres, M. A. Solis, R. Salas, and A. F. Bariviera, "A dynamic linguistic decision making approach for a cryptocurrency investment scenario," IEEE Access, vol. 8, pp. 228514–228524, 2020.

[20] A. A. Encalada-Malca, J. D. Cochachi-Bustamante, P. C. Rodrigues, R. Salas, and J. L. López-Gonzales, "A spatio-temporal visualization approach of pm10 concentration data in metropolitan lima," Atmosphere, vol. 12, no. 5, p. 609, 2021.

[21] C. Parra, C. Ponce, and R. Salas, "Evaluating the performance of explainable machine learning models in traffic accidents prediction in california," in 2020 39th International Conference of the Chilean Computer Science Society (SCCC), pp. 1–8, IEEE, 2020.

[22] Y. Morales, M. Querales, H. Rosas, H. Allende-Cid, and R. Salas, "A self-identification neuro-fuzzy inference framework for modeling rainfall-runoff in a chilean watershed," Journal of Hydrology, vol. 594, p. 125910, 2021.

[23] E. Cantor, R. Salas, H. Rosas, and S. Guauque-Olarte, "Biological knowledge-slanted random forest approach for the classification of calcified aortic valve stenosis," BioData Mining, vol. 14, p. 11, 2021.

[24] P. Horton and K. Nakai, "Better prediction of protein cellular localization sites with the it k nearest neighbors classifier.," in Ismb, vol. 5, pp. 147–152, 1997.

[25] K. Fukunaga and P. M. Narendra, "A branch and bound algorithm for computing k-nearest neighbors," IEEE transactions on computers, vol. 100, no. 7, pp. 750–753, 1975.

[26] M. J. Zaki, W. Meira Jr, and W. Meira, Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.

[27] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in icml, vol. 99, pp. 124–133, Citeseer, 1999.

[28] A. Priyam, G. Abhijeeta, A. Rathee, and S. Srivastava, "Comparative analysis of decision tree classification algorithms," International Journal of current engineering and technology, vol. 3, no. 2, pp. 334–337, 2013.

[29] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Wadsworth International Group, 1984.

[30] W. Buntine and T. Niblett, "A further comparison of splitting rules for decision-tree induction," Machine Learning, vol. 8, no. 1, pp. 75–85, 1992.

[31] S. Maliah and G. Shani, "Using pomdps for learning cost sensitive decision trees," Artificial Intelligence, vol. 292, p. 103400, 2021.

[32] W. S. Noble, "What is a support vector machine?," Nature biotechnology, vol. 24, no. 12, pp. 1565–1567, 2006.

[33] J. Zhou, Y. Qiu, S. Zhu, D. J. Armaghani, C. Li, H. Nguyen, and S. Yagiz, "Optimization of support vector machine through the use of metaheuristic algorithms in forecasting tbm advance rate," Engineering Applications of Artificial Intelligence, vol. 97, p. 104015, 2021.

[34] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[35] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, "A statistical approach to predict flight delay using gradient boosted decision tree," in 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1–5, IEEE, 2017.

[36] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, "Prediction of undrained shear strength using extreme gradient boosting and random forest based on bayesian optimization," Geoscience Frontiers, vol. 12, no. 1, pp. 469–477, 2021.

[37] J. J. Espinosa-Zúñiga, "Aplicación de algoritmos random forest y xgboost en una base de solicitudes de tarjetas de crédito," Ingeniería, investigación y tecnología, vol. 21, no. 3, pp. 1–16, 2020.

[38] N. J. Vickers, "Animal communication: when i'm calling you, will you answer too?," Current biology, vol. 27, no. 14, pp. R713–R715, 2017.

[39] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," The Annals of Statistics, vol. 29, no. 5, pp. 1189 – 1232, 2001.

[40] M. Luckner, B. Topolski, and M. Mazurek, "Application of xgboost algorithm in fingerprinting localisation task," in IFIP International Conference on Computer Information Systems and Industrial Management, pp. 661–671, Springer, 2017.

[41] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, 2016.

[42] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," Journal of microbiological methods, vol. 43, no. 1, pp. 3–31, 2000.

[43] N. K. Bose and P. Liang, Neural network fundamentals with graphs, algorithms, and applications. McGraw-Hill, Inc., 1996.

[44] F. Murtagh, "Multilayer perceptrons for classification and regression," Neurocomputing, vol. 2, no. 5-6, pp. 183–197, 1991.

[45] B. Chaudhuri and U. Bhattacharya, "Efficient training and improved performance of multilayer perceptron in pattern classification," Neurocomputing, vol. 34, no. 1-4, pp. 11–27, 2000.

[46] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, et al., "Knowledge discovery and data mining: Towards a unifying framework.," in KDD, vol. 96, pp. 82–88, 1996.

[47] S. K. Dias Ledesma, "Patrones de consumo de drogas lícitas e ilícitas y su influencia en el rendimiento académico en una comunidad intercultural andina.," B.S. thesis, Universidad Estatal de Bolívar, 2020.

[48] Á. Marchesi, J. C. Tedesco, and C. Coll, Calidad, equidad y reformas en la enseñanza. Fundación Santillana, 2021.

[49] M. T. Q. Quintero and G. M. O. Vallejo, "El desempeño académico: una opción para la cualificación de las instituciones educativas.," Plumilla educativa, vol. 12, no. 2, pp. 93–115, 2013.

[50] B. A. Gueldner, L. L. Feuerborn, and K. W. Merrell, Social and emotional learning in the classroom: Promoting mental health and academic success. Guilford Publications, 2020.

[51] G. M. Walton and T. D. Wilson, "Wise interventions: Psychological remedies for social and personal problems.," Psychological review, vol. 125, no. 5, p. 617, 2018.

[52] D. Gómez-Sánchez, E. I. Martínez-López, and R. Oviedo-Marín, "Factores que influyen en el rendimiento académico del estudiante universitario," Tecnociencia Chihuahua, vol. 5, no. 2, pp. 90–97, 2011.

[53] L. P. Garzón and A. M. C. Pérez, "Revisión de algunos estudios sobre la deserción estudiantil universitaria en colombia y latinoamérica (review of some studies on university student desertion in colombia and latin america)," Theoria, vol. 21, no. 1, pp. 9–20, 2012.

[54] R. T. M. García, "Factores que intervienen en el rendimiento académico universitario: Un estudio de caso," Opción, vol. 31, no. 6, pp. 1041–1063, 2015.

[55] M. A. Cano Celestino and R. Robles Rivera, "Factores asociados al rendimiento académico en estudiantes universitarios.," Revista Mexicana de Orientación Educativa, vol. 15, no. 35, pp. 1–25, 2018.

[56] B. Santos and H. Yobany, Transición demográfica en Honduras y su Incidencia en el desarrollo. PhD thesis, Tegucigalpa, Honduras, 2021.

[57] D. Rodríguez Rodríguez and R. Guzmán Rosquete, "Rendimiento académico y factores sociofamiliares de riesgo. variables personales que moderan su influencia," Perfiles educativos, vol. 41, no. 164, pp. 118–134, 2019.

[58] E. Chang-Rodríguez, Diásporas chinas a las Américas. Fondo Editorial de la PUCP, 2015.

[59] C. Romagnoli and I. Cortese, ¿ Cómo la familia influye en el aprendizaje y rendimiento escolar. VALORAS, 2015.

[60] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long, "Predicting academic performance by considering student heterogeneity," Knowledge-Based Systems, vol. 161, pp. 134–146, 2018.

[61] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques.," International journal of modern education & computer science, vol. 8, no. 11, pp. 36–42, 2016.

[62] S. Cohen, E. Ruppin, and G. Dror, "Feature selection based on the shapley value," in Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI'05, (San Francisco, CA, USA), p. 665–670, Morgan Kaufmann Publishers Inc., 2005.

[63] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," Cybernetics and information technologies, vol. 13, no. 1, pp. 61–72, 2013.

[64] T. P. Vital, K. Sangeeta, and K. K. Kumar, "Student classification based on cognitive abilities and predicting learning performances using machine learning models," International Journal of Computing and Digital Systems, vol. 10, no. 1, pp. 63–75, 2021.

[65] P. Meedech, N. Iam-On, and T. Boongoen, "Prediction of student dropout using personal profile and data mining approach," in Intelligent and Evolutionary Systems, pp. 143–155, Springer, 2016.

[66] K. D. Patel and A. B. Suthar, "Recommendations for student performance improvement based on result data using educational data mining," in Inventive Systems and Control (V. Suma, J. I.-Z. Chen, Z. Baig, and H. Wang, eds.), pp. 403–411, Springer Singapore, 2021.

[67] Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008, Using data mining to predict secondary school student performance, EUROSIS-ETI, 4 2008.

[68] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict covid-19 infection," Chaos, Solitons & Fractals, vol. 140, p. 110120, 2020.

[69] M. Smith and F. Alvarez, "Identifying mortality factors from machine learning using shapley values–a case of covid19," Expert Systems with Applications, vol. 176, p. 114832, 2021.

[70] L. E. Tideman, L. G. Migas, K. V. Djambazova, N. H. Patterson, R. M. Caprioli, J. M. Spraggins, and R. Van de Plas, "Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized shapley additive explanations," Analytica Chimica Acta, vol. 1177, p. 338522, 2021.

[71] A. A. Saa et al., "Educational data mining & students' performance prediction," International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, pp. 212–220, 2016.

[72] E. Tocto-Cano, S. Paz Collado, J. L. López-Gonzales, and J. E. Turpo-Chaparro, "A systematic review of the application of maturity models in universities," Information, vol. 11, no. 10, p. 466, 2020.

DAVID ORREGO GRANADOS received the B.S. degree in systems Engineering from Universidad Peruana Unión (UPeU, Perú). Currently, he is a MSc. candidate in Systems Engineering with mention in IT direction and management of Universidad Peruana Unión (UPeU, Perú). His main research interests include Academic performance with Machine Learning, IT and data science.

JONATHAN UGALDE received the the B.S. degree in Informatics Engineering from Universidad de Valparaíso, Chile. He was a lecturer at International Business and Information and Management Control Engineering Departments in Universidad de Valparaíso between 2019 and 2020. Currently, he is a Ph.D. student in Applied Informatics Engineering at Universidad de Valparaíso. His main research interests include Artificial Intelligence, Data Science, Visual Data Mining, Fairness and Explainability in Machine Learning.

RODRIGO SALAS (Senior Member, IEEE) received the B.S. and MSc. degrees in Informatics Engineering and the Dr. Eng. degree in informatics from the Federico Santa María Technical University (UTFSM) in Chile, in 2001, 2002 and 2010, respectively. From 2002 to 2004, he was a research assistant with the Informatics Department, UTFSM. Since 2004, has been with the Universidad de Valparaíso, where he is currently a Full Professor of the Biomedical Engineering School and teaches in Data Mining, Probability and Statistics, and Machine Learning. Dr. Salas is main researcher at *Millennium Institute for Intelligent Healthcare Engineering*, and main researcher at the *Center of Research and Development in Health Engineering (CINGS-UV)*. His research interests include Artificial Intelligence, Data Science, Computational Statistics, Decision Support Systems, Intelligent Systems and their applications to finance, air pollution, healthcare and medicine.

ROMINA TORRES (Member, IEEE) received the B.S. and M.Sc. degrees in Informatics engineering and the Dr. Eng. degree in informatics from Federico Santa María Technical University (UTFSM), Valparaíso, Chile, in 2001, 2003, and 2014, respectively. From 2001 to 2008, she worked with major international companies, such as Motorola, and Software AG. Since 2013, she has been a Professor with the Engineering Faculty, Andres Bello University (UNAB), Viña del Mar, Chile. Currently, she is the Director of the computer science master programs. Her research interests include intelligent systems and their applications for decision making in several interdisciplinary areas.

JAVIER LINKOLK LÓPEZ GONZALES (Member, IEEE) received the B.S. degree in Statistical and Informatics Engineering from Universidad Peruana Unión (UPeU, Perú) and the M.Sc. degree in Metrology from Pontifical Catholic University of Rio de Janeiro (PUC-Rio, Brazil). Currently, He is a Ph.D. candidate in Statistics of Universidad de Valparaíso (UV, Chile). His main research interests include pattern recognition in Machine Learning, air pollution with Deep Learning techniques, and Time Series with Singular Spectrum Analysis. He is an associate professor of Facultad de Ingeniería y Arquitectura, Universidad Peruana Unión.