IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 22, 2019, accepted May 22, 2019, date of publication May 27, 2019, date of current version June 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2919343

Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions

YURI NIETO[1], VICENTE GACÍA-DÍAZ[1], CARLOS MONTENEGRO[2], CLAUDIO CAMILO GONZÁLEZ[3], AND RUBÉN GONZÁLEZ CRESPO[4]

[1]Department of Computer Science, University of Oviedo, 33007 Oviedo, Spain
[2]Engineering Department, Francisco Jose de Caldas District University, Bogotá 1100100, Colombia
[3]Engineering Department, National Open and Distance University UNAD, Bogotá 1100100, Colombia
[4]School of Engineering, Universidad Internacional de La Rioja (UNIR), 26002 Madrid, Spain

Corresponding author: Rubén González Crespo (ruben.gonzalez@unir.net)

This work was supported by the Corporation Unified National Higher Education (CUN).

ABSTRACT Decisions made at the strategic level of Higher Educational Institutions (HEIs) affect policies, strategies, and actions that the institutions make as a whole. Decision's structures at HEIs are depicted in this paper and their effectiveness in supporting the institutions' governance. The disengagement of the stakeholders and the lack of using efficient computational algorithms lead to 1) the decision process takes longer; 2) the ''whole picture'' is not involved along with all data necessary; and 3) small academic impact is produced by the decision, among others. Machine learning is an emerging field of artificial intelligence that using various algorithms analyzes information and provides a richer understanding of the data contained in a specific context. Based on the author's previous works, we focus on supporting decision-making at a strategic level, being deans' concerns the preeminent mission to bolster. In this paper, three supervised classification algorithms are deployed to predict graduation rates from real data about undergraduate engineering students in South America. The analysis of receiver operating characteristic (ROC) curve and accuracy are executed as measures of effectiveness to compare and evaluate decision tree, logistic regression, and random forest, where this last one demonstrates the best outcomes.

INDEX TERMS Decision trees, random forest, logistic regressions, machine learning, strategic decisions, Higher Educational Institutions.

I. INTRODUCTION

Notoriously a ''disengagement'' occurs regarding Higher Educational Institutions and education policymakers, students, managers and their subordinates [1]. Many barriers including technological conditions, rigid governance structure or vulnerability to continuous changes in government rules, may impede the support needed by managers and university's directors when making a decision.

In regard to technological conditions, face-to-face model education still addresses significant obstacles. For example, their administrative, as well as their academic information, is stored in various silos making formats employed vary significantly [2]. Additionally, as we observe in our case study university, some of the transactional processes such as the record of students' attendance or registered graduate documents, to mention a few examples, are still done manually in paper notebooks.

The associate editor coordinating the review of this manuscript and approving it for publication was Hazrat Ali.

Nevertheless, in distance and blended educational models, the arena offers more resources. The amount of systematized information is natural because computers are highly suitable and practical for this work [3].

Stored data is not enough when directors and managers are deciding. Educational data, whether it is systematically or manually stored, should be analyzed to provide a proper presentation of valuable information to support these complex processes.

Therefore, the usage of efficient computational algorithms is vital to enhance this process. Through the latest years, Machine Learning has shown its outstanding capacity for pattern recognition and predicting outcomes for diverse datasets despite the field. Most of the work done in Machine Learning has focused on supervised algorithms. Their main strength is that they produce models that we can incorporate in the decision-making process [4]. In order to choose the most suitable learning algorithm, a clear objective is required, and an analysis of previous data must be performing. Thus, the feasibility of using a supervised algorithm

over a not-supervised algorithm can be determined. Afterwards, depending on the viability of each, a choice between classification or regression algorithms needs to be made. Even though many studies have used ML algorithms to identify students; our research differs from the existing ones. (I) the data comes from a face-to-face educational model. (II) more diverse and numerous features from data collection are included on the algorithms architecture leading to achieve a higher overall accuracy that we analyze. (III) neither the stakeholders nor the objective goals have been evaluated before. (IV) we investigate deans and directors concerns when making academic decisions [5] as our work driver.

This article seeks to classify the decision's structure at HEIs and the influence of the institutional governance among them. This section is developed with the aim of depicting the impact and responsibility of strategic decisions not just in the academic context but the complete environment where the Higher Educational Institution is located.

An extensive literature review looks for the classification of the uses of Information and Communication Technologies (ICT) at HEIs and ongoing applications that used ML in the education field. The primary goal focuses on the comparison of three supervised Machine Learning (ML) algorithms that, used as predictors, would enhance decision at the strategic level. Specifically, we applied Decision Trees, Random Forests and Logistic Regression to predict graduation rates using real data from a face-to-face model education university in South America. Analysis of the ROC curve and Accuracy are executed as measures of effectiveness to compare and evaluate the three algorithms.

The rest of the paper is laid out as follows: In Section 2, a review of the current literature is discussed. In Section 3, we introduce the classification of decisions at Higher Educational Institutions. In Section 4, details of the method to compare Decision Tree, Random Forest and Logistic Regression in a real case study are illustrated. Results are presented in Section 5. Ultimately conclusions and discussions are reported in Section 6.

II. LITERATURE REVIEW

Information and Communication Technologies (ICT) have transformed the academic field, not just what we teach (curriculum) or how we teach it (pedagogy), but how the institutions respond and manage these changes [6]. Researchers set out investigating the ICT impact on education in the past, in particular in the e-learning arena.

The first subsection seeks to isolate the different uses of ICT at HEIs with the aim of highlight the small research held for the academic decision-making process. Hence, the focus was on classifying the ICT used in the educational field into the following categories; e-learning, academic research, quality measurement and decision- making process.

Afterward, we provide an overview of selected works that uses Machine Learning at Higher Educational Institutions for solving academic problems. We set their stakeholder, goals, and algorithms used. Moreover, some of the ongoing applications are highlighted to establish the reliability of using these algorithms in a face to face educational model with the dataset obtained.

A. ICT USES AT HIGHER EDUCATIONAL INSTITUTIONS

In the current dynamic environmental economy sectors, in particular, the tertiary sector (service sector) has to keep track with ICT and align these technologies to satisfy stakeholder's needs and expectations. In these contexts, universities have to adapt the services they provide to develop, improve, and enhance the quality of the provided services [7].

Generally speaking, the existing studies and application of ICT at Higher Educational Institutions have primarily focused on four major streams: e-learning, academic research, a quality measurement, and decision-making process.

1) E-LEARNING

Transforming the conventional face-to-face education model through technological platforms have set up blended-learning and distance learning models [8], [9]. Having virtual classrooms has changed communication and interaction between teachers and students, education resources, and others. Although e-learning has developed new educational models, researchers have become aware of the need to sustain the development of abilities and competencies to promote intellectual capital [10]–[12].

Thus, the new advances in e-learning have created life-long learning/teaching, the transfer of knowledge [13] and introduced new concepts like Mobile Learning (m-learning) with the increase of technology [14]. Moreover, the raising of public awareness about environmental problems demands new competencies in sustainability. Hence, a few years ago e-learning had been used to promote and improve the quality of life-long education through the acquirement of knowledge, skills, and values for Sustainable Development [15].

2) ACADEMIC RESEARCH

Academic research has been transformed by information technology due to the rapid, widespread diffusion of electronic papers, digitalization of libraries and journals, web access to information and repositories among other facilities [6]. worldwide, researchers and enterprise leaders collaborate from different perspectives in diverse projects thanks to the capability for remote exchange and communication [16]. the innovative advances on use of data (technical use) and elaboration and presentation of projects (academic use) enhance teachers' curriculum and university's visibility [17]. furthermore, information technologies serve as a control mechanism for academic misconduct like plagiarism, self-plagiarism, coercive citations, and questionable reviewing [18].

3) QUALITY MEASURE

E-learning service providers have improved their mechanism to assure the quality of their products and services [19].

Quality in education has been discussed in one of the following themes: measure the impact of knowledge [3], [20]–[22], teaching quality [23], [24], assessment quality [25] and timetabling quality [26]. Higher Educational Institutions must guarantee quality for all stakeholders: students, teachers, directors, government and society. In computer assisted-learning, the possibilities of rendering information are more numerous [27], which improves the outcomes and allows that the data insight the educational field supports quality measurement. Besides it leads us the third ICT usage at Higher Education Institutions [3], [28].

4) DECISION-MAKING PROCESS

In order to improve the decision-making process, the information first needed to be automatized, even small transactional operations such as attendance lists. Decisions made at HEIs have an administrative and academic nature. Thus, universities have computational systems to aid mostly administrative operations. Most of these are separated by departments, such as the Accounting Information System or the Academic Information System, where data can be extracted from the different silos to support the decision-making process. ICT at universities serves to help management (supporter) but also serves to improve (enabler) the decision-making process. In the organizational context some examples using Information and Communication Technologies are applied in Accounting Systems [29] and Enterprise Resource Planning [30], as well as the academic management [31].

The impact of Information and Communication Technologies depends on its infrastructure, accessibility, and the intensity of use. Although the computational advancement in processing speed and algorithms designed, has shown significant progress, more efficient and user-friendlier applications are needed when it comes to the decision-making process. Opportunely, Machine Learning arises with different algorithms that learn from data to support various task in this field.

B. MACHINE LEARNING ALGORITHMS IN THE EDUCATIONAL FIELD

Applications of Machine Learning (ML) in the academic field focused mainly on using supervised algorithms to predict students' behaviors with the possibility of early intervention. Some authors have covered different educational problems using ML such as course planning, institutions' and teachers' quality, intervention and prediction, and learning product selection. Although we focus on Higher Educational Institutions applications, other works using and comparing Machine Learning algorithms in public and private schools have been developed [3].

Educational activities datasets, such as web-log files traced from Leaning Management Systems (LMS) or Massive Open Online Couse (MOOC's) are increasingly being used to analyze students' learning behavior. Interesting examples come from various universities around the world including Massachusetts Institute of Technology [32], [33], University of Vigo [34], University of Liége [35], Open University of China [36], University of Alagoas [37] along with others, which used information from LMS and other repositories and applied different Machine Learning algorithms to predict students' performances. Table 1 highlights some successful ML works at some of the mentioned Higher Educational Institutions.

TABLE 1. Machine learning currently projects at higher educational institutions.

HEI	Goals	Stakeholder	ML Algorithm Used
University of Liége [35]	To predict freshmen´s failure. They used data from school records and environmental factors to efficient remediation and or study mediation	Students and Advisors	RF, LR, ANN
University of Alagoas [37]	To predict student's failure rates in introductory programming courses to solve educator's questions regarding these aspects	Teachers and tutors	DT, SVM, ANN, NB
Babcock University [38]	To predict students' performance and recommend necessary action through a framework of an intelligent recommender system, based on background factors	Advisors and Intelligent tutors	DT, RF, ANN
University of Jordan [22]	To predict and analyze knowledge level of learners in order to adapt content presentation and evaluation	Teachers and curriculum designers	RF, ANN, SVM, NB
Massachusetts Institute of Technology [32]	To predict and intervene in final course scores using process-level information. They analyze the overall prediction accuracy as well as the chronological progression of prediction accuracy	Teachers and intelligent tutors	LIR, LR, ANN, SVM
Amrita School of Engineering [39]	To predict the placement performance of students and propose a recommendation system to identify and pay attention to specific students' skills	Students and Teachers	DT, LR

RF = Random Forest, LR= Logistic Regression, ANN = Artificial Neural Network, DT = Decision Tree, SVM = Support Vector Machine, NB = Naïve Bayes, LIR = Linear Regression

Table 1 shows authors that successfully used various Machine Learning algorithms to compare the accuracy of them. As illustrated in Figure 1, the most popular algorithms used were Artificial Neural Network (ANN) and Support Vector Machine (SVM), followed by Decision Trees (DT), Random Forest (RF) and Logistic Regression (LR). Due to in our previous worked [5], we have already compared the use of SVM and ANN when predicting graduation rates in a public University in Colombia, with the aim of continuing that research, in this work we analyze DT, RF, and LR. The choose of these last three algorithms is done because it suits our prediction objective and have shown excellent outcomes

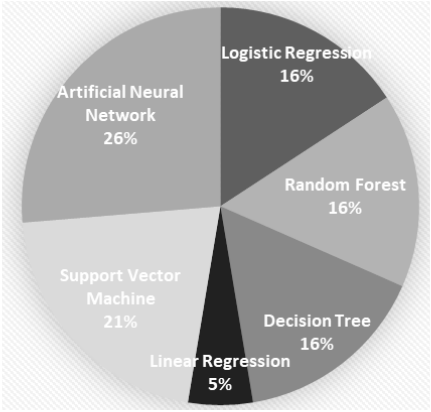

FIGURE 1. Machine learning algorithms used in related works.

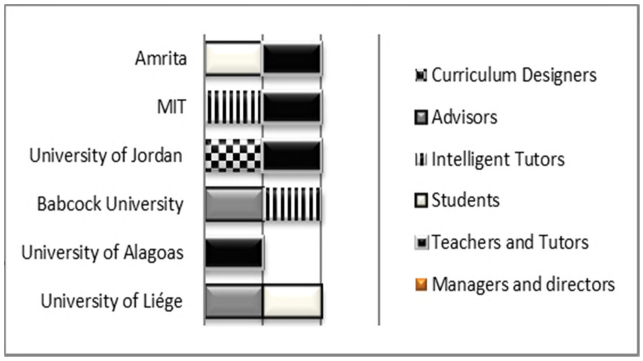

FIGURE 2. Stakeholders at reviewed works using ML executed projects.

on related works, where it shows are in the trend of usage. Furthermore, it is observed that recent studies are not focused on directors or administrators as most of the research was limited to specific stakeholders, mainly students and teachers.

As illustrated in Figure 2., although the same classification algorithms as related works are used, they differ from our research. (I) Our data comes from a face-to-face educational model. (II) Due to more features are included, algorithms architecture differ. (III) The stakeholders are directors and deans from HEIs which have particular visualization results needs, and have not been adressed before. (IV) Strategic decisions are supported when the right information is given to the high chain management as would be exposed in Section 3.

In a public university in Brazil [37] they used four prediction techniques: Support Vector Machine, Decision Tree, Neural Network and Naive Bayes to predict students' failure rates in introductory programming courses. To this aim, data was extracted from distance education. After applying data preprocessing and algorithm fine-tuning, the effectiveness of these algorithms was improved. First, they reduce the number of attributes and balance the information by applying the Synthetic Minority Over-sampling Technique. Then they fine-tune the data according to the parameters in each algorithm. Like them, we compare Machine Learning algorithms using the effectiveness metric to predict student's failure rates. Besides, our research is allocated on predictive models from educational data. However, the focus in our research is on graduation rates, involving the whole curriculum rather than a particular subject. Moreover, stakeholders and educational model are also different. Preprocessing data considering the number attributes is held to this aim.

This analysis of the related work also corroborated what was stated in our prior work [2], [5], [40], [41]; reviewed researches haven't solved Directors and Managers necessities when it comes to making a decision. Their primary goal is not focused on supporting the strategic level in universities. Although few of the works involved Machine Learning algorithms in their development, the stakeholders are mainly students and teachers and seek to fulfill their requirements and not the academic as a whole.

Therefore, in the next sections, we focus on decisions' classification at HEIs, and we use Machine Learning algorithms in a real case study to set a baseline to support directors at higher educational institutions during the decision-making process regarding issues of graduation rates.

III. DECISIONS' CLASSIFICATION AT HIGHER EDUCATIONAL INSTITUTIONS

Higher Educational Institutions are a particular type of organization of the tertiary sector. They currently hold primary responsibility for the governance and management of their finances, activities, and personnel by retaining the autonomy to decide their organizational operations. In higher education how decisions are made about institutional priorities, strategies, goals, and resource allocations, and who is held accountable for these decisions, are all functions of institutional governance [42].

Three types of governance is observed at HEIs, which influenced the operational and managerial behavior: Academic, Bureaucratic, and Corporate [43].

A. ACADEMIC

Faculty members work to retain authority and decision-making powers in areas such as teaching, curriculum, academics, and administration.

B. BUREAUCRATIC

University retains hierarchy layers with divisions of labor characterized by procedures, fixed administrations, and direct orders by higher leaders.

C. CORPORATE

Viewing the education as a service makes the students the core costumers, which is a natural consequence of taking marketing in higher education? These main marketing activities are in support of recruiting and retention efforts [44]. University follows the practice of enterprise to highlight customer needs and market competition.

Although in real world governance exhibits variations a mixture between these three categories, it is quite usual universities to prefer and implement the academic and bureaucracy styles [43]. Among HEIs are specific differences in the mission and management strategies, for instance, private

universities are more market-oriented and action-oriented while public universities focus on the roles of students and alumni in the society. However, all of them work on behalf of students' success, and HEIs make their decisions to ensure it. Therefore, in order to classify the kind of decisions at HEIs, and considering their primary goal between the types of HEIs, we follow the hierarchic structure to divide work vertically according to decisions' responsibilities, which resemble an organizational pyramid structure [45].

FIGURE 3. Decisions' structure levels at HEIs.

1) STRATEGIC

The uppermost level defines the policies and strategies for the organization integrating the primary goals and actions into a cohesive whole. The higher level of the institutions is the more ambitious in their strategic planning [36]. The managing positions are frequently represented by the governing board, rector and deans [46].

The managing positions are frequently represented by the governing board, rector and deans. They discuss the critical factors in strategic planning and provide guidelines for its execution. Decision-making at this level impacts the entire university. For instance, one strategic decision at this level is the number of freshmen accepted each semester. The decision affects the university's resource allocation (e.g., Budget, teachers, and facilities), as well as the society as more people, might access professional programs. At this level, according to relative works analysis, machine learning algorithms have not been used to aid this corporate stage.

2) TACTIC

The purpose of the tactic level is to identify and execute the detailed plans made at the strategic level. Generally, deans work together with the head of departments or programs directors to achieve the above planning. Intermediate directors coordinate resources usage efficiently, providing management and planning at specific times. Once the strategic planning is accepted, the tactic level is in charge of its implementation and control. Thus, quality assurance is an essential task at this level. This middle management performs decisions such as the number of students per teacher or curriculum changes. Algorithms such as Naïve Bayes and Artificial Neural Network have been used to ensure teaching, assessment and timetabling quality.

3) OPERATIONAL

The lowest level is in charge of everyday processes and through their job they sustain the whole structure. Specific tasks and transactional activities are performed to support the operations of the institution. This level holds the majority of Information Technology requires by HEIs. IT governance in this stage works as an instrument to control and manage the IT resources such as infrastructure technology and people [47]. Collaborators as teachers, advisors, tutors, programs assistants, and secretaries, among others, execute their task according to the guidelines provided by the strategic and tactic level. Although decision-making at this level affects a smaller population within the university, it might impact students' success (i.e., schedule and timetabling evaluation) and operational mechanisms (i.e., subject registration process). At the state-of-the-art examination, we found that most of the works done using machine learning on the educational field are a focus on the operational stage. Some of the prominent algorithms used are Artificial Neural Networks and Support Vector Machine.

We observed that each of the levels in the pyramid have a decision-making process that, from top to bottom, affect a more significant portion of the community. Although some software applications can support decision-making processes, higher levels generally work with the information provided by the operational stage and data is not analyzed and visualized easily to support decisions at high stages. Norman and Ahmed identified the central software misfit in Higher Education planning software. Some of the cases they stated are poor consultant effectiveness, poor-reliance on heavy customization, reduced IT infrastructure, poor project management effectiveness, poor management support, too tight project schedule and poor knowledge transfer [30].

Two global statistics are frequently cited as measures of student success: the cohort graduation rate and the freshman-to-sophomore retention rate. Thus, Faculty decisions should focus on their enhancement. Students' persistence to complete their educational goals are a key gauge of student success, and therefore institutional success [44]. Hence, in the next section, we propose the usage of Machine Learning algorithms to predict graduation rates and collaborate with academic decisions on behalf of student's success. Moreover, we set a baseline to support decision making at the strategic level at HEIs according to directors' needs analyzed in our previous work.

IV. METHOD

One of the main goals of this study is to compare the effectiveness of existing Machine Learning algorithms in predicting graduate rates that will support decision making at the strategic level. Thus, we will be classifying student's academy performance to predict the number of graduated and not graduated students, being this our objective variable.

First Subsection presents data characterization, cleaning and preparation. Subsequently, subsection B contextualize the three Machine Learning algorithms used in this work. We present their basics, method, architecture, and configuration used. The tools and metrics used are respectively indicated in subsection C and D.

A. DATA COLLECTION AND PREPARATION

The experiment was conducted with real data from a public university in Colombia. The data source contains information from 6100 engineer students. We analyzed students from five different engineer programs such as industrial, cadastral, systems, electrical and electronical engineering, enrolled during the ten years period of 2004-2014. This number of students is assumed after disregarding cases of missing data such as students who leave during the first semester and students who enter the university after 2009. Students who enter after 2009 will not graduate before 2014 because engineering careers take ten semesters to graduate and therefore will not address the supervised algorithms needs. More than 55200 records were available to analyze.

As our case study is a public university, data policies are strict. Although our research was restricted by their data-protection policies, and we lack information about students' gender or age among other socio-demographic data, for the most part we use students' academic records to held the graduation rates prediction. We believe the inclusion of socio-demographic and socio-economic data would be worth to analyze in the future. However, in this study the academic information obtained is efficient to analyze the insightful outcomes.

The students' academic features include in all three algorithms are stated in Table 2.

Once the classification objective was set (i.e., graduated and not graduated students) and data was acquired, we conducted the following steps to build every algorithm model:

i. Using the scaling method, we transformed data by giving them values from a range [0, 1] 0 as a minimum and a maximum of 1.

ii. Set initial hyperparameters for each algorithm.

iii. Using a stratified sampling technique, split the dataset into two subsets 70% for training and 30% for prediction, to keep the data distribution. The sampling method alleviate the effect of class imbalance problem as one of the most employed method [48].

iv. From the training subset in step three, we use 5-fold cross validation technique to tune the hyperparameters in each algorithm.

v. We execute the algorithms with the initial settings. In each k-fold we save the accuracy obtained as well as the values entered in each hyper parameter, to adjust them in each run and encounter the most suitable values for them until the accuracy reached the expectations.

vi. Finally the algorithms are executed to train the whole training set using the best values obtained for the hyperparameters in each algorithm. Hyperparameters set in each algorithm as well as the architecture and contextualization of each algorithm is exposed in the next subsection.

TABLE 2. Variable definitions and type of measurement.

Variable	Measurement
Arithmetic mean of the final grade of all subjects coursed	Grade from 0.0 to 5.0
Subjects enrolled in each semester	Quantity
Subjects satisfactory approved in each semester	Quantity
Subjects failed in each semester	Quantity
Subjects that were failed but after an extra summary exam become approved (validated)	Quantity
Subjects enrolled in each semester	Quantity
Median grade	Grade from 0.0 to 5.0
Maximum grade	Grade from 0.0 to 5.0
Minimum grade	Grade from 0.0 to 5.0
The range of the grades	Quantity
The middle grade between the smallest grade and the median of the data set	Quantity
Median grade	Quantity
The middle grade between the median and the highest grade of the data set	Quantity
The square root of the average of the squared deviations of the values subtracted from their average value	Quantity
Summary of the subjects enrolled during the whole career	Quantity
Summary of the satisfactory approved subjects during the whole career	Quantity
Summary of the failed approved subjects during the whole career	Quantity
Summary of the subjects that become approved after the validation exam during the whole career	Quantity
The middle grade between the smallest grade and the median of the data set	Quantity
Median grade	Quantity
The middle grade between the median and the highest grade of the data set	Quantity
The square root of the average of the squared deviations of the values subtracted from their average value	Quantity
Summary of the subjects enrolled during the whole career	Quantity
Summary of the satisfactory approved subjects during the whole career	Quantity
Summary of the failed approved subjects during the whole career	Quantity
Summary of the subjects that become approved after the validation exam during the whole career	Quantity
Socio-economic status according to the students' residence address	{1,2,3,4,5,6}

B. MACHINE LEARNING ALGORITHM ARCHITECTURE

1) DECISION TREE (dt)

It is a highly used classifiers due to its simplicity for understanding and interpretation. It requires little data preparation, handles numerical and categorical data, and performs very well with large data set in a short time [49]. Additionally, the hierarchical tree structure resembles a human way of decision-making, providing extending information about the sequence to classify and individually into a class, discovering rules in a more comprehensible manner [50]. In our

case study, the classification falls either into ''graduated'' or ''not graduated.'' DT is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. Although there are many specific decision tree algorithms (e.g., ID3, C4.4, C5.0, CART, and CHAID), we worked with the most popular developed by Quinlan. C5.0 is significantly faster and more efficient than its predecessors C4.5 and ID3. C5.0 supports boosting, which gives the trees more accuracy.

Furthermore, it allows the weighting of different attributes and misclassification types and separates the data automatically to help reduce noise [20]. When constructing decision trees, it is essential to find the best splitting point measurement (i.e., information gain, gain ratio, Gini index, and entropy measure). The selection of the split attribute should directly decide the learning trend [51]. Gini index is used as a split measure for choosing the most appropriate splitting attribute for each node. The split function has this form:

$$I_g(p) = 1 - \sum_{i=1}^{J} p_i^2 \tag{1}$$

Gini index is mathematically shown above; whit J classes suppose I $\in$ {1, 2, 3, ...J} and pi is the fraction of items labeled with class 'i' in the set. The data set is portioned into two subsets. The split can be made at the highest value of the lower partition, at the lowest value of the lower partition, or at the average of the two. We choose these last split point to force the model only binary splits on nominal columns. On the other hand, the pruning method used is the Minimum Description Length (MDL) which reduces the rate of misclassification, provides high accuracy and fast execution times.

2) RANDOM FOREST

RF is a classifier that combines the performances of numerous Decision Trees algorithms to predict the value of a variable [52].

Each tree in the forest gives a classification and ''votes'' for that class. The forest chooses the classification having the most votes (over all the trees in the forest). RF Regression predictor has the form:

$$f_{rf}^{K}(x) = \frac{1}{K} \sum_{K=1}^{k} T(x) \tag{2}$$

When RF receives and (x) input vector, made up of the values of different evidential features analyzed for a given training area, RF builds a number K of regression trees and averages the results. After K such trees $T(x)_1^K$ are grown. About the architecture used, the split attribute chosen was Gain Radio since it did not represent a significant difference among the other options provided by knime (Information Gain, Information Gain Radio y Gini Index). Moreover, the number of models used was one hundred mainly because of the dataset size and testing results. The number of decision trees to learn of number of models set is 100.

Additionally, the use of static random seed is required y knime to start the prediction, and the one automatically generated by the software was used (i.e., 1508210392822).

3) LOGISTIC REGRESSION

This classification algorithm could be confusing by its name. It is used to estimate discrete values (e.g., binary values) based on a given set of independent variables.

It is known as logit regression because it predicts the probability of an occurrence, of any event, by fitting data into a logit function [53]. Logistic regression gives linear class boundaries. Due to the fact it uses an 'S'-shaped curve instead of a straight line it is a natural fit for dividing data into groups.

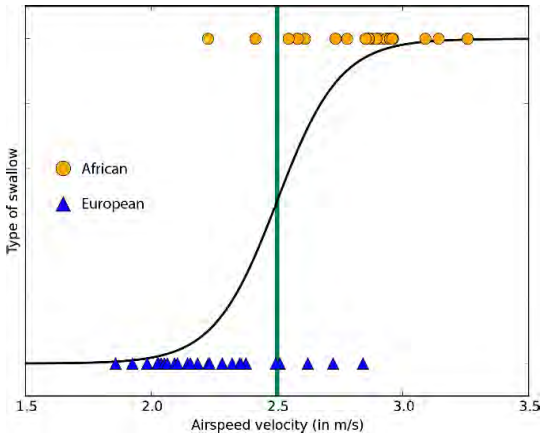

FIGURE 4. Logistic regression representation of two-class data with just one feature [54].

Figure 4 shows an example took from Azure of logistic regression to two-class data with just one feature. The class boundary is the point at which the logistic curve is just as close to both classes [54].

Target column used was Objective (i.e., graduated or not graduated) with true as the reference category. Stochastic Average Gradient (SAG) was used as the solver mainly because it minimizes the negative of the Log Likelihood function and supports regularization. It relies on the idea of the gradient descent method meaning that in each interaction the algorithm moves into a descending direction of the negative of Log Likelihood function with step size Δ., which is called the learning rate. The learning rate strategy was fixed to a step size of 0.2. Finally, the termination conditions were set at a maximal number of epochs (i.e., 200) and Epsilon (i.e., $\varepsilon = 0.001$).

C. INSTRUMENTATION

To perform the pre-processing of data and all the Machine Learning algorithms we used KNIME 3.4.0 (Konstanz Information Miner) analytic platform [55]. KNIME is open-source software, developed in Java which allows ETL processes (Extraction, Transformation, and Loading) in addition to various modular components for machine learning and data mining.

D. EFFECTIVENESS METRICS

To evaluate the performance of the compared Machine Learning algorithms, we use the area under the curve (AUC) as the evaluation criteria. AUC is a popular measure for ranking class performance of the learned classifiers [56] AUC is calculated as follows:

$$E = \frac{P_0 - \frac{t_0(t_0+1)}{2}}{t_0 t_1} \tag{3}$$

where t0 and t1 are numbers of negative and positive instances respectively. P0 = ∑ri denoting the rank of the nth negative instance in the ranked list. This equation can only handle a two-level class problem corresponding to the prediction sought (graduated or not graduated).

Moreover, the use of a confusion matrix is useful to present the prediction results of the test. If the data set contains 'n' distinct classes, the confusion matrix is an (n×n) matrix [57]. Since this case examines only two types of data (graduated or not graduated), we have a (2×2) confusion matrix indicated in Table 3.

TABLE 3. Confusion matrix.

Prediction / Actual	Graduated	Not graduated
Graduated	A	C
Not Graduated	D	B

From the confusion matrix data, we defined the overall accuracy, the precision rate and the recall rate as follows [58]: The precision rate of the graduated class = $A/(A + D)$ the precision rate of the not graduated class = $B/(B + C)$. The recall rate of the graduated class = $A/(A + C)$ and the recall rate of the not graduated class $B/(B + D)$ and the overall accuracy $(A + B)/(A + B + C + D)$.

V. RESULTS

Once the 55200 records were divided into the training set (i.e., 70%) and test set (i.e., 30%) three machine learning methods were used to process the test set (i.e., Decision Trees, Random Forests and Logistic Regression). First, we used the Receiver Operating Characteristic Curve (ROC) as a standard metric for the binary outcome expected (graduated or not graduated.) AUC helps to reduce the ROC curve to a single value, representing the expected performance of the classifier.

The x-axis indicates the false positive rate, while the y-axis indicates the true positive rate. It is clear that the AUC for Logistic Regression (i.e., 0.9028) represented in Fig. 5 is slightly higher than Random Forest's AUC (i.e., 0.8994) and Decision Trees (i.e., 0.8830). One of the goals of this study is to identify the potential graduated students. Comparatively, RF was the most effective and more precise in predicting student graduates.

Table 4 shows the results of the methods mentioned in section 4.3. In terms of the precision rate and the recall rate of the graduated class, there was little difference in the performance of the three prediction models: Random Forest has the highest recall rate (91.93%), followed by Decision Tree (91.38%) and Logistic Regression with the lowest recall rate (90.935 %). However, regarding the precision rates of the same class RF is the lowest (86.44%), followed by DT (86.61%) and LR (87.04%) with the highest precision rate.

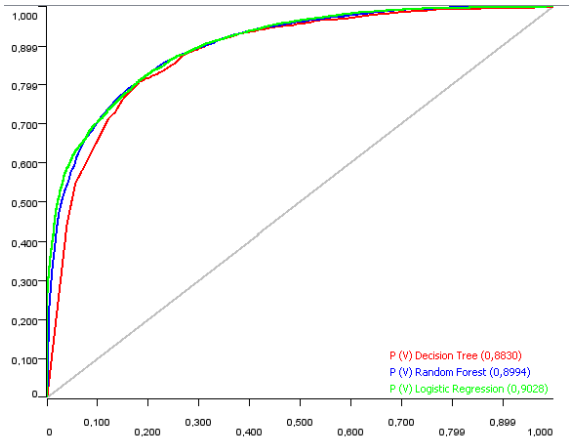

FIGURE 5. Shows ROC curves for the three algorithms compared.

TABLE 4. Evaluation of prediction results.

Evaluation index	DT	RF	LR
Recall rate of graduated class	91.38%	91.93%	90.93%
Recall rate of not graduated class	65.91%	65.21%	67.33%
Precision rate of graduated class	86.61%	86.44%	87.04%
Precision rate of not graduated class	76.01%	77.00%	75.46%
Overall Effectiveness	83.92%	84.11%	84.02%

Among the three prediction models, RF had the highest overall accuracy (84.11%), followed by LR (84.02%), while the DT had the lowest accuracy rate (83.92%). The three models had a relatively high overall accuracy rate that exceeded 83%. With the aim of comparing the impact of the first features include in each training algorithm, we test them using 6, 15 and 19 features as shown in Table 5. We observe that the overall accuracy increases as the number of features tested increase. Revealing more features leads to obtaining greater accuracy. For instance, Random forest accomplishes an accuracy enhancement of 4.4%. Considering 10000 students and increase from 79.71% to 81.11% would represent more than 400 students correctly classified. Algorithm's improvement also relies on data distribution and data set.

Compare to related works mentioned in Section 2 [22], [32], [35], [37]–[39]; our results indicate an accuracy rate above of the 83%, which is more than four percentage point higher accuracy from other researchers. Moreover, an additional comparison of including more features in the initial training set was developed to evaluate whether it is possible to obtain higher accuracy results. Table 5., shows the augment of the overall accuracy when more data features are involved.

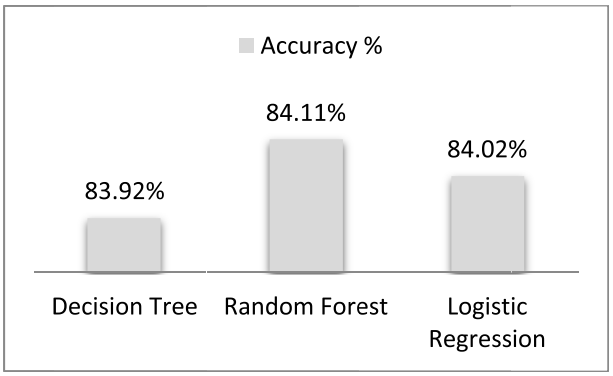

FIGURE 6. Comparative overall accuracy of the algorithms tested. First.

TABLE 5. Overall accuracy comparison by initial features tested.

	6 Features	9 Features	16 Features
Decision Tree	79,57	81,34	83,92
Random Forest	79,71	81,34	84,11
Logistic Regression	79,55	81,14	84,02

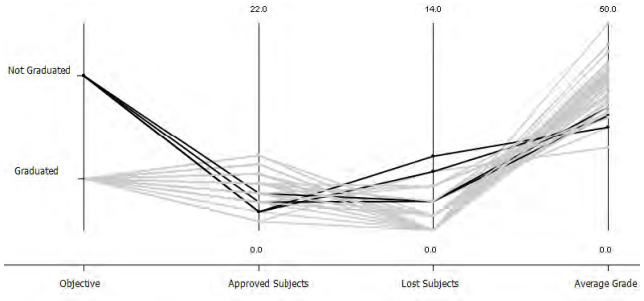

FIGURE 7. Relevant features to Graduate. The figure shows three relevant academic features: student's graduation: approved subjects, lost subjects, and average grade.

Visualization of the information is an imperative task in any analytics process. Through KNIME data visualization is available, including scatter plots, pie charts, box plots, histograms, and others. From the data collected at District University Francisco Jose de Caldas (JFC), a parallel coordinate are plotted to analyze multivariate data: features tested regarding grades.

The Parallel Coordinates plots just 50 records were taken randomly to this plot. Black color represents not graduated students, while gray color graduated students. First vertical parallel line from left to right, represents the objective sought (graduated or not graduated), subsequently is approved subjects, followed by lost subjects and average grade. We observe that graduated students (gray lines) reach a better average grade and more approved subjects. By contrast, not graduated students (black lines) approve fewer subjects and reach lower average grades. These patterns are easily recognized thanks to machine learning algorithm.

Depict insightful information through an intentional graphic, improve not just the data' understanding through the recognition of trends and hidden patterns, but enhance the actions taken from the data processed. For instance, Figure 7, allows HEIs Administrators to observe how the average grade and the number of lost subjects impact the fact that students get graduated or not, to mention some of the features that are analyzed from the data set. Although some graphics provide by KNIME are useful, we propose in the next section significant improvements to succeed in the information launch and discuss potential efficient metrics that could be added.

VI. CONCLUSION AND DISCUSSIONS

Even though at educational field Machine Learning is still emerging, its effectiveness to analyze information is notorious. Through the analysis, predictions, and visualizations of information, HEIS' directors obtain a greater understanding of the different variables involved when making a decision. Machine Learning supports this process providing various algorithms suitable to the different kinds of data and the different kinds of predictions required.

We employ three supervised classification algorithms: Decision Trees, Random Forests and Logistic Regression, where Random Forest performs the best outcomes.

From a recent literature review [3], [34], we observed not just their efficiency but also their acceptance on the research field when contributing to decision processes. Results indicate that the three tested Machine Learning algorithms can identify at least an 83% accuracy rate, which is a slightly higher rate reached compare to related works (i.e., $\overline{X} = 79\%$). Using these experiments the early identification of students likely to not graduate is highly effective, although from this prediction other aspects such as students' academic performance and dropout rates could be analyzed.

Continuing with the research done in previous work [5] we compare different Machine Learning algorithms in this paper as well as we analyze decision's structure at HEIs and how they are managed according to the institutional governance. Having these students recognized early can allow HEIs governance strategic planning abilities are respecting students' exclusion policies, students' dropout rates, retention rates, strengthen programs, and a whole host of others.

The comparison of ROC, accuracy, precision rates and recall rates were conducted. We observed that the overall accuracy is prevailing in Random Forest although the area under the curve (AUC) is slightly superior in Logistic Regression. However, as effectiveness metric, accuracy is more significance than ROC curve due to ROC being insensitive to data sets with unbalanced proportion classes [59]. Regarding precision rate and recall rate, the three algorithms are similar.

Therefore, a more in-depth analysis of the number of features tested, or data normalization, will be presented in a forthcoming paper. Future research will also include the analysis of other effectiveness metrics such as F-Measure or Specificity as well as the comparison with other classification algorithms, that would be worth to analyze in other Higher Educational Institutions. Moreover, would be worth to include socio-demographic and socio-economical information about the students when analyzing the variables

that might affect the graduation rates at Higher Educational Institutions,

Performing this recognition has managerial implications not just for reducing processing time but increasing reliability on its prediction.

This last is essential due to help during the decision-making process with insightful information that cooperates improving the decision effectiveness even in the different educational field such as resource planning, teacher's management, curriculum design, and related factors. Although the visualization nodes provided by KNIME are efficient to represent the results for this study, a future goal for this research includes the development of a computational tool for deans and university administrators. Specific technician knowledge is required to understand the reports and graphics on KNIME, and a more straightforward interpretation of the academic data and predictions using Human Computer Interaction techniques would significantly support the decisions making process.

REFERENCES

[1] A. Lašáková, L. Bajzíková, and I. Dedze, "Barriers and drivers of innovation in higher education: Case study-based evidence across ten European Universities," *Int. J. Educ. Develop.*, vol. 55, pp. 69–79, May 2017.

[2] Y. V. N. Acevedo and C. E. M. Marín, "System architecture based on learning analytics to educational decision makers toolkit," *Adv. Comput. Sci. Eng.*, vol. 13, no. 2, pp. 89–105, 2014.

[3] A. R. T. Góes, M. T. A. Steiner, and P. J. S. Neto, "Education quality measured by the classification of school performance using quality labels," *Appl. Mech. Mater.*, vols. 670–671, pp. 1675–1683, Oct. 2014.

[4] H. Lounis and T. Fares, "Using efficient machine-learning models to assess two important quality factors: Maintainability and reusability," in *Proc. Joint Conf. 21st Int. Workshop Softw. Meas. 6th Int. Conf. Softw. Process Product Meas.*, Nov. 2011, pp. 170–177.

[5] Y. Nieto, V. García-Díaz, C. Montenegro, and R. G. Crespo, "Supporting academic decision making at higher educational institutions using machine learning-based algorithms," *Soft Comput.*, vol. 23, no. 12, pp. 4145–4153, 2018.

[6] K. V. Pincus, D. E. Stout, J. E. Sorensen, K. D. Stocks, and R. A. Lawson, "Forces for change in higher education and implications for the accounting academy," *J. Accounting Educ.*, vol. 40, pp. 1–18, Sep. 2016.

[7] A.-P. Pavel, A. Fruth, and M.-N. Neacsu, "ICT and e-learning—Catalysts for innovation and quality in higher education," *Procedia Econ. Finance*, vol. 23, pp. 704–711, Jan. 2015.

[8] F. V. Elena, A. M. Manuel, and G. G. S. Carina, "Which use give teachers at La Laguna University to ICTs?" *Procedia-Social Behav. Sci.*, vol. 93, pp. 1646–1651, Oct. 2013.

[9] H. M. Vo, C. Zhu, and N. A. Diep, "The effect of blended learning on student performance at course-level in higher education: A meta-analysis," *Stud. Educ. Eval.*, vol. 53, pp. 17–28, Jun. 2017.

[10] I. M. Taucean and M. Tamasila, "Research challenges for eLearning support in engineering and management training," *Procedia-Social Behav. Sci.*, vol. 124, pp. 210–218, Mar. 2014.

[11] E. Faham, A. Rezvanfar, S. H. M. Mohammadi, and M. R. Nohooji, "Using system dynamics to develop education for sustainable development in higher education with the emphasis on the sustainability competencies of students," *Technol. Forecast. Social Change*, vol. 123, pp. 307–326, Oct. 2017.

[12] F. A. Goni, A. G. Chofreh, M. Mukhtar, S. Sahran, S. A. Shukor, and J. J. Klemeš, "Strategic alignment between sustainability and information systems: A case analysis in Malaysian public higher education Institutions," *J. Clean. Prod.*, vol. 168, pp. 263–270, Dec. 2017.

[13] I. González-González and A. I. Jiménez-Zarco, "Using learning methodologies and resources in the development of critical thinking competency: An exploratory study in a virtual learning environment," *Comput. Hum. Behav.*, vol. 51, pp. 1359–1366, Oct. 2015.

[14] F. Moreira, M. J. Ferreira, C. P. Santos, and N. Durão, "Evolution and use of mobile devices in higher education: A case study in Portuguese higher education Institutions between 2009/2010 and 2014/2015," *Telematics Inform.*, vol. 34, no. 6, pp. 838–852, Sep. 2017.

[15] U. M. Azeiteiro, P. Bacelar-Nicolau, F. J. P. Caetano, and S. Caeiro, "Education for sustainable development through e-learning in higher education: Experiences from Portugal," *J. Clean. Prod.*, vol. 106, pp. 308–319, Nov. 2015.

[16] K. H. Wang, N. J. Ray, D. N. Berg, A. T. Greene, G. Lucas, K. Harris, A. Carroll-Scott, B. Tinney, and M. S. Rosenthal, "Using community-based participatory research and organizational diagnosis to characterize relationships between community leaders and academic researchers," *Preventive Med. Rep.*, vol. 7, pp. 180–186, Sep. 2017.

[17] A. Valentín, P. M. Mateos, M. M. González-Tablas, L. Pérez, E. López, and I. García, "Motivation and learning strategies in the use of ICTs among University students," *Comput. Educ.*, vol. 61, pp. 52–58, Feb. 2013.

[18] C. Hopp and G. A. Hoover, "How prevalent is academic misconduct in management research?" *J. Bus. Res.*, vol. 80, pp. 73–81, Nov. 2017.

[19] S. E. Alptekin and E. E. Karsak, "An integrated decision framework for evaluating and selecting e-learning products," *Appl. Soft Comput.*, vol. 11, pp. 2990–2998, Apr. 2011.

[20] D. Delen, H. Zaim, C. Kusey, and S. Zaim, "A comparative analysis of machine learning systems for measuring the impact of knowledge management practices," *Decis. Support Syst.*, vol. 54, pp. 1150–1160, Jan. 2013.

[21] N. Cohen, J. Gattuso, and K. MacLennan-Brown, *CCTV Operational Requirements Manual*, no. 28. London, U.K.: Home Office Scientific Develpoment Branch, 2009.

[22] N. Ghatasheh, "Knowledge level assessment in e-learning systems using machine learning and user activity analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 4, pp. 107–113, 2015.

[23] R. Rodríguez and G. Rubio, "Teaching quality and academic research," *Int. Rev. Econ. Educ.*, vol. 129, pp. 10–27, Sep. 2016.

[24] X.-Y. Liu, "Private colleges teachers evaluation system based on support vector machine (SVM)," in *Proc. Int. Conf. Appl. Sci. Eng. Innov. (ASEI)*, 2015, pp. 1918–1921.

[25] K. J. Gerritsen-van Leeuwenkamp, D. Joosten-ten Brinke, and L. Kester, "Assessment quality in tertiary education: An integrative literature review," *Stud. Educ. Eval.*, vol. 55, pp. 94–116, Dec. 2017.

[26] A. Muklason, A. J. Parkes, E. Özcan, B. McCollum, and P. McMullan, "Fairness in examination timetabling: Student preferences and extended formulations," *Appl. Soft Comput.*, vol. 55, pp. 302–318, Jun. 2017.

[27] I. Smeureanu and N. Isaila, "Information technology, support for innovation in education sciences," *Procedia-Social Behav. Sci.*, vol. 15, pp. 751–755, Jan. 2011.

[28] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, "A machine learning framework to identify students at risk of adverse academic outcomes," in *Proc. Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 1909–1918.

[29] S. M. Shuhidan, N. Mastuki, and W. M. N. W. M. Nori, "Accounting information system and decision useful information fit towards cost conscious strategy in Malaysian higher education institutions," *Procedia Econ. Finance*, vol. 31, no. 15, pp. 885–895, 2015.

[30] A. Y. Noaman and F. F. Ahmed, "ERP systems functionalities in higher education," in *Proc. Int. Conf. Commun., Manage. Inf. Technol.*, vol. 65, pp. 385–395, Jan. 2015.

[31] T. Anastasios, S. Cleo, P. Effie, T. Olivier, and M. George, "Institutional research management using an integrated information system," *Procedia-Social Behav. Sci.*, vol. 73, pp. 518–525, Feb. 2013.

[32] A. J. Stimpson and M. L. Cummings, "Assessing intervention timing in computer-based education using machine learning algorithms," *IEEE Access*, vol. 2, pp. 78–87, 2014.

[33] S. Boyer, B. U. Gelman, B. Schreck, and K. Veeramachaneni, "Data science foundry for MOOCs," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal.*, Oct. 2015, pp. 1–10.

[34] C. Gonzalez, E. Elhariri, N. El-Bendary, and A. Fernandez, "Machine learning based classification approach for predicting students performance in blended learning," in *Proc. 1st Int. Conf. Adv. Intell. Syst. Inform. (AISI)* (Advances in Intelligent Systems and Computing), vol. 407. Switzerland: Springer, 2016, pp. 47–56.

[35] A.-S. Hoffait and M. Schyns, "Early detection of University students with potential difficulties," *Decis. Support Syst.*, vol. 101, pp. 1–11, Sep. 2017.

[36] M. Tan and P. Shao, "Prediction of student dropout in e-Learning program through the use of machine learning method," *Int. J. Emerg. Technol. Learn.*, vol. 10, no. 1, pp. 11–17, 2015.

[37] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' Academic failure in introductory programming courses," *Comput. Hum. Behav.*, vol. 73, pp. 247–256, Aug. 2017.
[38] M. Goga, S. Kuyoro, and N. Goga, "A recommender for improving the student academic performance," *Procedia-Social Behav. Sci.*, vol. 180, pp. 1481–1488, May 2015.
[39] S. K. Thangavel, P. D. Bkaratki, and A. Sankar, "Student placement analyzer: A recommendation system using machine learning," in *Proc. 4th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Jan. 2017, pp. 1–5.
[40] Y. V. Nieto, "Modelo de un sistema de software basado en las técnicas de learning analytics como herramienta de apoyo en la toma de decisiones Académico-administrativas en las Instituciones públicas de educación superior," Univ. Distrital Francisco José de Caldas, Bogotá, Colombia, Tech. Rep. 131115, 2015.
[41] Y. V. Nieto, V. G. Diaz, and C. E. Montenegro, "Academic decision making model for higher education institutions using learning analytics," in *Proc. 4th Int. Symp. Comput. Bus. Intell. (ISCBI)*, Sep. 2016, pp. 27–32.
[42] A. Clark, "IT Governance: Determining who decides," *EDUCAUSE Center Anal. Res. Bull.*, vol. 2005, no. 24, pp. 1–13, Nov. 2005.
[43] S. J. Chan and C. Y. Yang, "Governance styles in Taiwanese universities: Features and effects," *Int. J. Educ. Develop.*, vol. 63, pp. 29–35, Nov. 2018.
[44] M. Guilbault, "Students as customers in higher education: The (controversial) debate needs to end," *J. Retail. Consum. Services*, vol. 40, pp. 295–298, Jan. 2018.
[45] G. R. Jones, *Organizational Theory, Design, and Change*, 7th ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2011.
[46] J. Hu, H. Liu, Y. Chen, and J. Qin, "Strategic planning and the stratification of Chinese higher education institutions," *Int. J. Educ. Develop.*, vol. 63, pp. 36–43, Nov. 2018.
[47] I. S. Bianchi and R. D. Sousa, "IT Governance mechanisms in higher education," *Procedia Comput. Sci.*, vol. 100, pp. 941–946, Jan. 2016.
[48] L. Nanni, C. Fantozzi, and N. Lazzarani, "Coupling different methods for overcoming the class imbalance problem," *Neurocomputing*, vol. 158, pp. 48–61, Jun. 2015.
[49] A. Trabesli, Z. Elouedi, and E. Lefevre, "Decision tree classifiers for evidential attribute values and class labels," *Fuzzy Sets Syst.*, vol. 366, pp. 46–62, Jul. 2019.
[50] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. J. Murray, and Q. Long, "Predicting academic performance by considering student heterogeneity," *Knowl.-Based Syst.*, vol. 161, pp. 134–146, Dec. 2018.
[51] H. Sun and X. Hu, "Attribute selection for decision tree learning with class constraint," *Chemom. Intell. Lab. Syst.*, vol. 163, pp. 123–129, Apr. 2017.
[52] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, pp. 804–818, Dec. 2015.
[53] S. K. Singh, R. W. Taylor, M. M. Rahman, and B. Pradha, "Developing robust arsenic awareness prediction models using machine learning algorithms," *J. Environ. Manage.*, vol. 211, pp. 125–137, Apr. 2018.
[54] M. Azure. (2017). *How to Choose Machine Learning Algorithms*. Accessed: Jan. 17, 2018. [Online]. Available: https://docs.microsoft.com/es-es/azure/machine-learning/studio/algorithm-choice
[55] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz information miner," in *Data Analysis, Machine Learning and Applications* (Studies in Classification, Data Analysis, and Knowledge Organization). Germany: Springer, 2007.
[56] Y. Zhang, J. Wu, C. Zhou, and Z. Cai, "Instance cloned extreme learning machine," *Pattern Recognit.*, vol. 68, pp. 52–65, Aug. 2017.
[57] D. J. Yu, J. Hu, Q. M. Li, Z. M. Tang, J. Y. Yang, and H. B. Shen, "Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction," *IEEE Trans. Nanobiosci.*, vol. 14, no. 1, pp. 45–58, Jan. 2015.
[58] M. R. K. Darzi, S. T. A. Niaki, and M. Khedmati, "Binary classification of imbalanced datasets: The case of CoIL challenge 2000," *Expert Syst. Appl.*, vol. 128, pp. 169–186, Aug. 2019.
[59] G. Hackeling, *Mastering Machine Learning With Scikit-Learn*. Birmingham, U.K.: Packt Publishing, 2014.

YURI NIETO received the M.Sc. degree in computer science and communications from Francisco Jose de Caldas District University. She is currently pursuing the Ph.D. degree with the Computer Science Program, University of Oviedo. In 2012, she was an Industrial Engineer with the AXON and GIIRA Investigation Groups, where she is currently a member. Her research interests include machine learning, decision support systems, learning analytics, distributed systems, and virtualization.

VICENTE GACÍA-DÍAZ received the Ph.D. degree in computer science from the University of Oviedo, in 2011, where he is currently a Software Engineer and an Associate Professor with the Department of Computer Science. He has supervised over 60 academic projects and published over 70 research papers in journals, conferences, and books. His research interests include machine learning, natural language processing, model-driven engineering, and domain-specific languages. He is also part of the Editorial and Advisory Board of several journals and has been an Editor of several special issues in books and journals.

CARLOS MONTENEGRO received the Diploma of Advanced Studies degree from the Pontifical University of Salamanca, in 2008, the M.Sc. degree in information and communication systems from the Universidad Distrital Francisco José de Caldas, and the Ph.D. degree in systems and computer services for the Internet from the University of Oviedo, Asturias, Spain, in 2012. He is currently a Systems Engineer. His research interest includes object-oriented.

CLAUDIO CAMILO GONZÁLEZ received the M.Sc. degree in urban development from San Buenaventura University, Colombia. He is currently pursuing the Ph.D. degree in educational technology with Lleida University. He is also a System Engineer and a member of the Byte and Desing Research Group, UNAD University, Colombia. His research interests include education, e-learning, and decision support systems.

RUBÉN GONZÁLEZ CRESPO received the Ph.D. degree in computer science engineering. He is currently the Dean of the Higher School of Engineering, UNIR, and the Director of the AENOR (Spanish Association for Standardization and Certification) Chair in Certification, Quality and Technology Standards. He is also a member of different committees at the ISO Organization. He is also an Advisory Board Member of the Ministry of Education at Colombia and an Evaluator of the National Agency for Quality Evaluation and Accreditation of Spain (ANECA).