Ingeniería Solidaria

Educational data integration and machine learning for academic performance prediction

Integración de datos educativos y aprendizaje automático para la predicción del rendimiento académico

Integração de dados educacionais e aprendizado de máquina para previsão de desempenho acadêmico

Leonardo Emiro Contreras Bravo [1]
Héctor Javier Fuentes López [2]
Nayive Nieves Pimiento [3]

Received: February 2nd, 2025
Accepted: April 10th, 2025
Available: May 11th, 2025

How to cite this article:
L. E. Contreras Bravo, H. J. Fuentes López, and N. Nieves Pimiento, "Educational data integration and machine learning for academic performance prediction / Integración de datos educativos y aprendizaje automático para la predicción del rendimiento académico," *Revista Ingeniería Solidaria*, vol. 21, no. 2, 2025.
https://doi.org/10.16925/2357-6014.2025.02.07

Research article. https://doi.org/10.16925/2357-6014.2025.02.07

[1] Docente Titular, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
E-mail: lecontrerasb@udistrital.edu.co. Orcid: https://orcid.org/0000-0003-4625-8835
[2] Docente Titular, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
E-mail: hjfuentesl@udistrital.edu.co. Orcid: https://orcid.org/0000-0001-6899-4564
[3] Docente TCO, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
E-mail: nnievesp@udistrital.edu.co. Orcid: https://orcid.org/0000-0003-2914-4836

Abstract

Introduction: This article is the result of the research project "Integration of Educational Data and Machine Learning Techniques for the Prediction of Student Academic Performance", conducted at Universidad Distrital Francisco José de Caldas between 2021 and 2023.

Problem: When processed with appropriate tools, educational data can be used to predict, prevent, and take action to improve students' academic performance.

Objective: The aim of this study is to predict academic performance in three engineering programs using machine learning techniques. The dataset comprises a total of 7,000 student records.

Methodology: Approximately 325 variables were analyzed in each run. The most influential variables were selected for each academic semester. Feature selection methods revealed standard variables that consistently influence academic performance, regardless of the type of engineering program.

Results: The prediction models were evaluated using supervised learning algorithms (SVC, KNN, Decision Tree, LDA) and ensemble methods, including Bagging techniques (RandomForest, ExtraTreesClassifier), Boosting techniques (AdaBoost, GBM, XGBoost, CatBoost, LightGBM), and Voting techniques (Blending, Stacking).

Conclusion: The proposed model, which uses a super learner algorithm (in one- and two-stage configurations), yielded the highest prediction accuracy for academic performance, followed by Stacking and Blending algorithms. The models achieved average accuracy scores of 85% for training and 75% for testing.

Originality of the Study: This study stands out for its integration of heterogeneous sources of academic, administrative, and socioeconomic data, combined with statistical analysis and advanced machine learning techniques, to generate predictive models tailored to the Colombian educational context.

Limitations: The study identified limitations in the availability and quality of historical data, as well as the absence of certain contextual variables not captured by institutional information systems. These factors may affect the generalizability of the models to other educational contexts.

Keywords: Educational data analysis; machine learning; higher education; model; academic performance.

Resumen

El artículo es producto de la investigación "Integración de Datos Educativos y Técnicas de Aprendizaje de Máquina para la Predicción del Desempeño Académico Estudiantil", desarrollada en la Universidad Distrital Francisco José de Caldas entre los años 2021 y 2023. Esta investigación parte del reconocimiento de que la información académica, cuando es procesada con las herramientas adecuadas, puede permitir predecir, prevenir y actuar para mejorar el rendimiento de los estudiantes.

El objetivo fue predecir el rendimiento académico en tres programas de ingeniería mediante técnicas de aprendizaje automático, utilizando un total de 7.000 registros estudiantiles. Para ello, se analizaron aproximadamente 325 variables en cada corrida, seleccionando cada semestre aquellas con mayor influencia. Los métodos de selección de características permitieron identificar variables comunes que impactan en el desempeño estudiantil, independientemente del tipo de ingeniería. La predicción se evaluó con algoritmos supervisados (SVC, KNN, árbol de decisión, LDA) y con algoritmos de ensamblado que incluyeron métodos de Bagging (RandomForest, ExtraTreesClassifier), Boosting (AdaBoost, GBM, XGBoost, CatBoost, LightBoost) y Voting (Blending, Stacking). El modelo propuesto, basado en un algoritmo superaprendiz de una y dos etapas, alcanzó los mejores resultados, seguido por Stacking y Blending, con valores promedio de precisión del 85% en entrenamiento y 75% en prueba.

El estudio aporta originalidad al integrar fuentes heterogéneas de datos académicos, administrativos y socioeconómicos, combinadas con análisis estadístico y técnicas avanzadas de aprendizaje automático, para generar modelos predictivos ajustados al contexto educativo colombiano. No obstante, se identificaron limitaciones relacionadas con la disponibilidad y calidad de los datos históricos, así como la ausencia de variables contextuales no capturadas por los sistemas institucionales, lo cual puede restringir la generalización de los resultados a otros entornos.

Palabras clave: análisis de datos educativos, aprendizaje automático, educación superior, modelos predictivos, rendimiento académico.

Resumo

Este artigo é produto do projeto de pesquisa “Integração de Dados Educacionais e Técnicas de Aprendizado de Máquina para Prever o Desempenho Acadêmico de Estudantes", realizado na Universidade Distrital Francisco José de Caldas entre 2021 e 2023. Esta pesquisa parte do reconhecimento de que informações acadêmicas, quando processadas com as ferramentas apropriadas, podem permitir a previsão, a prevenção e a intervenção para melhorar o desempenho dos estudantes.

O objetivo foi prever o desempenho acadêmico em três cursos de engenharia utilizando técnicas de aprendizado de máquina, com base em um total de 7.000 registros de estudantes. Para tanto, aproximadamente 325 variáveis foram analisadas em cada execução, selecionando-se aquelas com maior influência em cada semestre. Os métodos de seleção de variáveis permitiram a identificação de variáveis comuns que impactam o desempenho dos estudantes, independentemente da área da engenharia.

A previsão foi avaliada usando algoritmos supervisionados (SVC, KNN, árvore de decisão, LDA) e algoritmos de montagem que incluíram Bagging (RandomForest, ExtraTreesClassifier), Boosting (AdaBoost, GBM, XGBoost, CatBoost, LightBoost) e métodos de Votação (Blending, Stacking). O modelo proposto, baseado em um algoritmo de superaprendizagem de um e dois estágios, obteve os melhores resultados, seguido por Stacking e Blending, com valores médios de precisão de 85% no treinamento e 75% no teste.

O estudo é original por integrar fontes heterogêneas de dados acadêmicos, administrativos e socioeconômicos, combinadas com análise estatística e técnicas avançadas de aprendizado de máquina, para gerar modelos preditivos adaptados ao contexto educacional colombiano. No entanto, foram identificadas limitações relacionadas à disponibilidade e qualidade dos dados históricos, bem como à ausência de variáveis contextuais não capturadas pelos sistemas institucionais, o que pode restringir a generalização dos resultados para outros ambientes.

Palavras-chave: análise de dados educacionais, aprendizado de máquina, ensino superior, modelos preditivos, desempenho acadêmico.

1. INTRODUCTION

Education plays a vital role in achieving sustainable development, as it provides the foundation for building a better future. To fulfill this role effectively, education must not only equip students with the knowledge and skills needed to address social, environmental, and economic challenges but must also ensure equitable access and support for all members of the community to complete their studies successfully [1]. High-quality education enables individuals to understand complex global issues and contribute meaningfully to development processes.

Various indicators are commonly used to measure the quality and efficiency of educational systems, including retention rates, dropout rates, graduation rates, and cohort-based dropout statistics. However, in the context of Latin America, where dropout rates can range from 40% to 75%, these metrics may not fully capture the complexities of the region [2]. Several scholars [3], [4], [5] have argued that academic performance is strongly linked to these indicators and that it is essential to develop tools and strategies that help students enhance their academic outcomes—thus improving the overall efficiency of the education system.

In recent years, academic performance in higher education has become a focal point for research, especially as universities operate within societies characterized by rapid technological advancements and increased access to information [6]. This evolving landscape has led to transformations in education, particularly through the integration of technologies that allow for the analysis of institutional data. When properly processed, this information can be used to predict, prevent, and respond to academic performance issues [7].

Numerous studies have sought to determine which variables most significantly influence student success [8]. These have been grouped into categories such as academic, sociodemographic, online learning, academic management, psychosocial, and academic environment factors [9]–[14]. Variables considered include pre-admission factors [15] and performance-related data during university, such as grades in specific subjects, which indicate the degree to which learning objectives are achieved [16], [17].

Other influential variables include sociodemographic characteristics (e.g., age, gender, marital status, nationality) [18], [19] and socioeconomic factors, such as socioeconomic strata, household income, parental education levels, employment, and commuting distance to the university [20], [21]. Research has also examined institutional characteristics, including course types, program duration, class formats, and faculty influences [22], [23]. With the rise of online education, especially in response to dynamic global conditions, digital learning variables have become increasingly relevant [24], [13]. Moreover, studies have focused on tracking student progress throughout their university trajectory [25], [26].

The psychosocial dimension is another crucial factor, as explored in works such as [21], [27], which examine how social relationships and peer dynamics shape student behavior and performance.

Despite extensive research, academic performance remains a vast field for exploration. Although many techniques and variables have been proposed, there is still significant potential to develop predictive models that can inform institutional decision-making. Effective use of these models could not only save time and resources but also support targeted interventions, leading to improved outcomes in student success and educational quality indicators, such as reduced dropout rates and increased graduation rates. Furthermore, students themselves could benefit from data-driven insights into their academic progress, enabling them to make informed decisions about their studies.

The present study aims to identify the variables that most influence academic performance in three engineering disciplines—Industrial, Electrical, and Systems Engineering—and to assess whether these variables hold consistent importance across different curricula. Drawing on variables identified in the literature as influential, the research proposes the development of a computational model capable of predicting academic performance using data from consecutive semesters or academic years. This model is designed to be replicable across other disciplines and aims to detect low-performing students early, enabling institutions to implement policies and strategies that support academic success, increase retention, and reduce dropout rates—especially during the critical first semesters.

1.1 Literature review or research background

Regarding the areas of knowledge that have studied the multidimensional nature of academic performance, psychology has addressed this variable by analyzing factors such as interest, motivation, attendance, integration, stress, and anxiety [28], [29]. These studies often use tools such as questionnaires to gather student perceptions regarding academic performance, and are typically followed by statistical analyses that place greater emphasis on demographic data and their influence on the variable of interest [30], [13], [31].

In contrast, the field of machine learning, a subfield of data science, offers a powerful framework for building mathematical models capable of making predictions from unknown inputs. Machine learning is generally divided into several categories depending on the nature of the data: supervised learning, unsupervised learning, ensemble learning, and deep learning.

Supervised learning focuses on estimating an output (dependent) variable based on known input variables. Among the most commonly used algorithms is the decision tree, which has been applied to predict university academic performance. However, some studies report accuracy metrics below 60% for this approach [32], [33], [34]. Another frequently used algorithm is k-Nearest Neighbor (KNN), which classifies instances based on distance between them. KNN has been employed to estimate student performance in virtual environments [24], to assess prior academic preparation, and to predict graduation outcomes [25]. Reported prediction accuracy for KNN is around 68% [26], [20]. The Naive Bayes classifier, which relies on prior and posterior probabilities, has also been used to predict student performance based on demographic, educational, and psychological variables [35], [36], [37]. Similarly, the Support Vector Machine (SVM) algorithm has been applied using demographic and academic variables. For example, [38] used SVM to classify students according to academic scores, while [21] and [27] used it to predict student success leading to graduation.

Unsupervised learning, on the other hand, works without labeled outputs and performs clustering or association rule discovery based solely on input data [39]. In traditional unsupervised learning, various studies have used clustering techniques to group students based on behavioral patterns in both face-to-face [40], [41], [42] and virtual environments [43]. Some of these techniques have also been used as a preliminary step for building new supervised classifier models tailored to each cluster [44]. These approaches have helped classify students based on their performance in examinations [42].

Contemporary machine learning includes ensemble methods and deep learning. Ensemble learning combines multiple algorithms to minimize the prediction errors of individual models. Ensemble methods have been applied to predict academic performance using online interaction data with algorithms from the Bagging family [45], [46], [47], Boosting methods [48], [49], and Voting classifiers [50], [51], as well as in traditional learning environments. Academic performance prediction using ensemble methods has been conducted at various stages of the university experience, including within individual subjects [52], at the end of the first academic year [53], [54], in advanced courses using previous academic data, at the end of specific semesters or groups of subjects [55], [16], at the completion of four-year programs [56], and at the end of five-year programs using data from the first three years [18].

2. MATERIALS AND METHODS

The methodology used is developed in the following six steps: (1) data and tools, (2) data processing and transformation, (3) feature analysis, (4) prediction algorithms, (5) evaluation measures, and (6) computational model and optimization.

2.1 Data and tools

The study is initially based on a dataset of 7,000 students from the Industrial, Electronic, and Systems Engineering programs at the Universidad Distrital (Colombia), covering the period from 2012 to 2022. A total of 325 variables will be analyzed using Python.

2.2 Processing and transformation

The initial dataset underwent a data cleaning process, resulting in a total of 6,554 valid records, which, according to [72], can lead to improved algorithm performance in the modelling stage. The variable transformation phase is considered a crucial step for enhancing the predictive capacity of a model [73]. This phase aims to transform the distribution of variables in the dataset into a quasi-Gaussian distribution, thereby reducing the influence of independent variables with disproportionately large numerical values on the output variable. In this study, several transformation techniques were applied, including Standardization, Normalization, Rescaling, Box-Cox, Robust Standardization, and Yeo-Johnson.

2.3 Feature Analysis

To determine which variables are most relevant at each stage of students' academic trajectories, various feature selection methods were implemented. These methods are grouped into four categories: filter, wrapper, embedded, and ensemble methods.

Filter methods apply statistical measures to assign a relevance score to each feature independently of any learning algorithm. Wrapper methods use a predictive model to evaluate combinations of features and assign scores to estimate which subsets yield the best performance. Embedded methods differ in that they perform feature selection as part of the model training process, identifying which features contribute most to the model's accuracy during its construction. Lastly, ensemble methods assess the importance of input features using ensemble learning models, allowing for the classification of features based on their contribution to the output variable.

2.4 Ensemble and Supervised Prediction Algorithms

The output variable, referred to as academic performance, originally consists of numerical values. To align with the guidelines established by the Colombian Ministry of Education, an ordinal coding process was applied to convert these numerical values into nominal categories. Four performance levels were defined: Superior performance (scores between 50 and 45), High (44 to 40), Fundamental (39 to 30), and Low (29 to 0).

For the modeling phase, this study employed supervised learning algorithms, including K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, and Linear Discriminant Analysis (LDA). Additionally, ensemble methods from the three main families—Bagging, Boosting, and Voting—were also implemented.

2.5 Evaluation and optimization measures

According to the literature, various evaluation metrics are used to assess how well an algorithm predicts the dependent variable in classification tasks involving discrete labels. Commonly used metrics include accuracy, precision (also referred to as specificity), recall (sensitivity), and the F1 score [74].

2.6 Computational Model

It is important to note that, in order to develop a model capable of predicting academic performance, it was necessary to experiment with various feature selection techniques and machine learning algorithms. These algorithms required hyperparameter optimization to determine which methods—and how many variables—contributed added value to the model. Optimizing hyperparameters is a critical step when aiming to enhance the performance of a selected algorithm. In this study, optimization was performed using the Python programming language, specifically through the GridSearch method available in the Scikit-learn library. It is also worth noting that while optimization significantly improves model performance, it involves a high computational and time cost.

3. RESULTS

After carrying out the sequence of steps of the described methodology, the following results were obtained:

3.1 Processing and transformation

In the pursuit of improved performance metrics for prediction algorithms, various transformation methods were applied to the independent variables to bring them into a standard range, approximating a quasi-normal distribution. This preprocessing step was necessary to ensure compatibility with different prediction algorithms. The results indicate that the performance of both supervised and ensemble learning algorithms is significantly influenced by the use—or absence—of these transformation methods when predicting university academic performance (Figure 1).

On average, model performance improved by approximately 5% when transformation methods were applied compared to when they were not. Among the techniques tested, the Yeo-Johnson transformation method yielded the best results, followed closely by Box-Cox, both of which enhanced the predictive accuracy of the implemented algorithms.

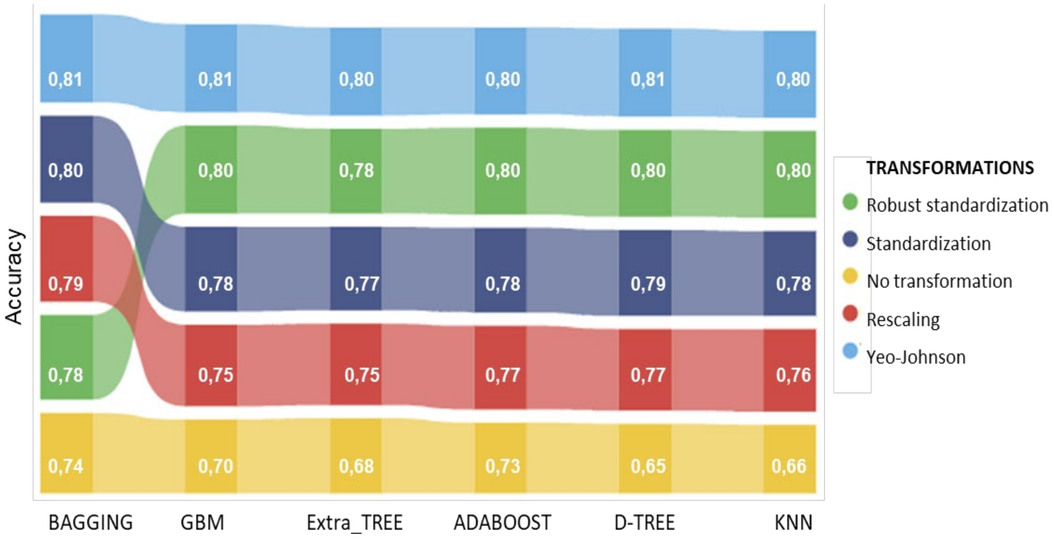


Figure 1. Improved Performance of Supervised and Ensemble Algorithms with Various Transformation Methods

3.2 Feature Analysis

This step is essential, as failing to identify and provide only the most influential variables, and instead supplying the algorithms with the entire set—including irrelevant or noisy variables—can result in model performance values falling below 60% on average. To address this, a comprehensive feature selection process was carried out using a combination of techniques across four methodological categories.

Filter methods included Pearson correlation, Chi-square, ANOVA, and mutual information. Wrapper methods involved recursive feature elimination (RFE) with different estimators such as logistic regression, linear regression, and decision trees, along with bi-directional elimination, forward selection, and backward selection. Embedded methods utilized linear regression, Ridge regularization, and Lasso regularization. Finally, ensemble methods employed algorithms such as CART, Random Forest, LightGBM, ExtraTreesClassifier, CatBoost, and XGBoost.

Variable selection was conducted independently for each of the three engineering programs to identify the most influential features in predicting academic performance, whether across consecutive semesters or over yearly periods (i.e., every two semesters) (Figure 2).

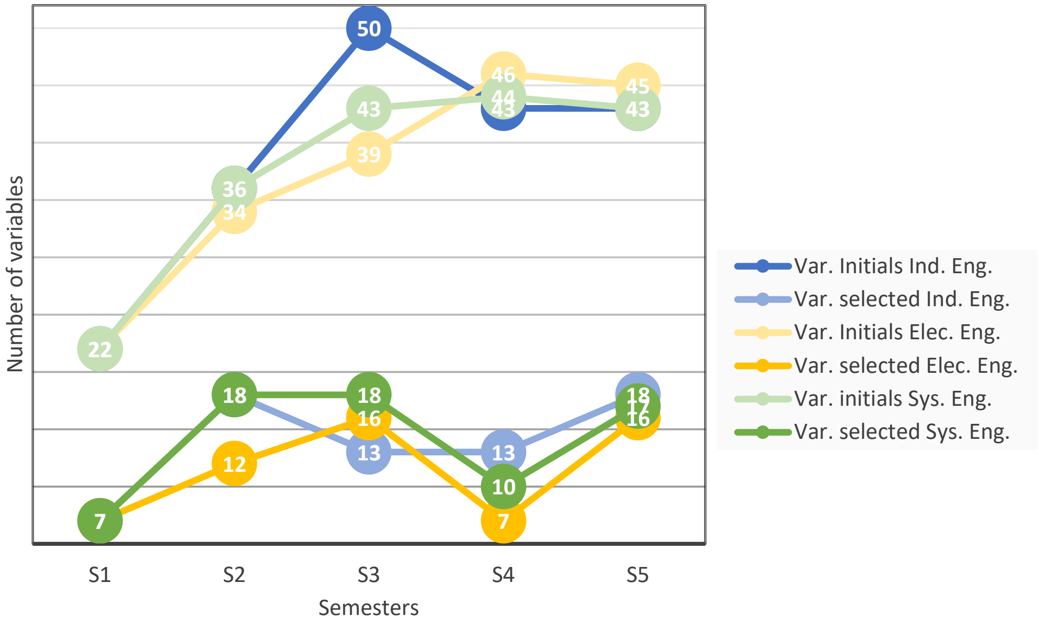


Figure 2. Number of best features to predict academic performance in each major during the initial four semesters

Based on pre-university variables—that is, information available about students prior to their entry into university—a preliminary analysis was conducted. The results indicated that, regardless of the engineering discipline pursued (Industrial, Electronic, or Systems Engineering), the variables that most influence first-semester academic performance fall into two key categories.

The first category is the academic background factor, composed of variables derived from Colombia's national standardized exams. These variables correspond to the performance areas of mathematics, physics, chemistry, and languages. The second category involves demographic factors, including variables such as the geographic location of the student's high school and the place of residence of the incoming student.

An additional noteworthy aspect analyzed was that engineering programs typically share a common core of foundational courses during the initial semesters. As such, the study sought to identify which of these core subjects had the strongest influence on academic performance across the first four semesters, when evaluated consecutively (Table 1).

It is important to highlight that the ensemble methods—specifically XGBoost, LightGBM, and CatBoost—not only improved the predictive performance of the models but also facilitated a more accurate comparison of the influential core subjects across the different programs.

Table 1. Influential variables in determining academic performance within the basic cycle of industrial, electronic, and systems engineering.

Semester	Common variables	Other influencing variables for Industrial Engineering	Other influencing variables for Electronic Engineering	Other influencing variables for Systems Engineering
1	ICFES Global Score ICFES Math Area ICFES Chemistry Area ICFES Language Area ICFES Physics Area School Location Residence Location			
2	Residence Location ICFES Global Score ICFES Math Area ICFES Chemistry Area ICFES Language Area Student Average (1 Semester) Note_Differential Calculation Note_Catedra Fjc Note_Seminar Note_Text Note_Basic Programming	Number of Subjects Repeated (1 Semester) Number of Credits Studied (1 Semester) Number of Subjects Approved (1 Semes.) Note_Drawing Note_Chemistry	Note_Physical_ one	Note_Cathedra Note_Logic
3	ICFES Global Score ICFES Math Area ICFES Language Area Student Average (1 Semester) Student Average (2 Semester) School Location Note_Basic Programming Note_Algebra Note_Integral Calculation Note_Text	Note_Differential Calculation Note_Drawing Number of Credits Studied (2 Semester) Number of Credits Approved (2 Semes.) Note_Materials	ICFES Chemistry Area Number of Credits Studied (2 Semester) Number Of Subjects Repeated (2 Semester) Note_History Note_Physical_ two Note_Circuits_one	Note_Cathedra Democracy Note_Oriented Programming Note_Ethics
4	ICFES Global Score Student Average (1 Semester) Student Average (2 Semester) Student Average (3 Semester) Note_Integral Calculation Note_General systems theory	Note_Algebra Note_Materials Number of Subjects Approved (3 Semester) Note_Multivariate Calculation Note_Estadistic_One Note_Thermodynamic	ICFES Chemistry Area Note_Multivariate Calculation Note_Differential Equations	Note_Basic Programming Note_Ethics Note_Physical_ two

3.3 Prediction algorithms

This study aimed to identify the most effective algorithm for predicting academic performance at the university level on a semester- or year-based basis, using cleaned, filtered, and transformed data alongside the most influential variables. To ensure model validity, the dataset was divided into training and test sets, as recommended by [74]. In order to mitigate issues related to variance, under-sampling, or over-sampling, a 10-fold cross-validation method was implemented in all cases [75].

The initial phase focused on identifying the best supervised learning algorithm to predict academic performance, based on the evaluation of key performance metrics such as accuracy and precision. The results obtained from predicting academic performance semester by semester using these algorithms are presented in Table 2.

It is important to note that for each algorithm evaluated, hyperparameter optimization was carried out to enhance model performance. Among the classifiers tested, the decision tree algorithm yielded the highest average performance, achieving an accuracy of 75.02%, outperforming the other models.

Table 2. Accuracy of Supervised algorithms in predicting University academic performance in the first four consecutive Semesters

Algorithms	Industrial Engineering	Electronic Engineering	Systems Engineering	Industrial Engineering	Electronic Engineering	Systems Engineering
SEMESTER 1				SEMESTER 2		
KNN	0,615	0,743	0,647	0,836	0,762	0,686
SVC	0,534	0,680	0,641	0,773	0,669	0,622
ARBOL	0,639	0,748	0,666	0,805	0,779	0,687
NAIVE	0,582	0,742	0,593	0,685	0,732	0,637
LDA	0,628	0,745	0,651	0,812	0,759	0,688
SEMESTER 3				SEMESTER 4		
KNN	0,815	0,795	0,585	0,804	0,835	0,792
SVC	0,591	0,708	0,461	0,703	0,773	0,650
ARBOL	0,817	0,806	0,615	0,820	0,844	0,774
NAIVE	0,670	0,689	0,573	0,669	0,803	0,692
LDA	0,805	0,786	0,602	0,776	0,805	0,778

Subsequently, ensemble algorithms were implemented to determine whether they could provide better performance metrics than the supervised algorithms previously evaluated. To this end, each ensemble algorithm was optimized using the GridSearch method, an approach that systematically tests combinations within a predefined hyperparameter space.

Table 3 presents some of the hyperparameters that were optimized using 10-fold cross-validation. It is important to note that each optimization process involved a high level of computational complexity and incurred a significant computational cost due to the number of iterations required to identify the most effective configuration.

Table 3. Optimized hyperparameters for the assembly algorithms

Algorithm	Optimized hyperparameters
Bagging – Decision Tree	base_estimator, n_estimators, bootstrap, bootstrap_features, max_features, max_samples, criterion, max_depth max_features, min_samples_leaf, min_samples_split, random_state, splitter
Random forest	n_estimators, max_features, min_samples_split, max_dept, min_samples_leaf, bootstrap, criterion
ExtraTreesClassifier	n_estimators, max_features, bootstrap, criterion, max_depth, min_samples_split, min_samples_leaf
AdaBoost	criterion, max_depth max_features, min_samples_leaf, min_samples_split, splitter, base_estimator,n_estimator, algorithm, learning_rate
(GBM) Gradient Boosting Machine	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, subsample, learning_rate, criterion
XGBoost	n_estimators, max_depth, min_child_weight, gamma, learning_rate, colsample_bytree, subsample, reg_alpha, reg_lambda
CatBoost	n_estimators, border_count, depth,l2_leaf_reg, learning_rate
LightGBM	n_estimators, class_weigh, colsample_bytree, num_leaves, max_depth, min_child_sample, reg_alpha, reg_lambda, min_split_gain, subsample, learning_rate, boosting_type

3.4 Evaluation measures

A second phase in the search for an algorithm, capable of more accurately predicting academic performance, involved the implementation of ensemble algorithms using both Bagging and Boosting approaches. Figure 3 displays the accuracy values, where a value close to 1 indicates a highly successful model, while values closer to 0 reflect poor performance.

Among the algorithms tested, XGBoost, CatBoost, and LightGBM consistently returned higher performance metrics—including accuracy, precision, and recall—compared to any of the previously evaluated supervised algorithms. This trend held true whether academic performance was predicted across consecutive semesters or over alternate (yearly) periods.

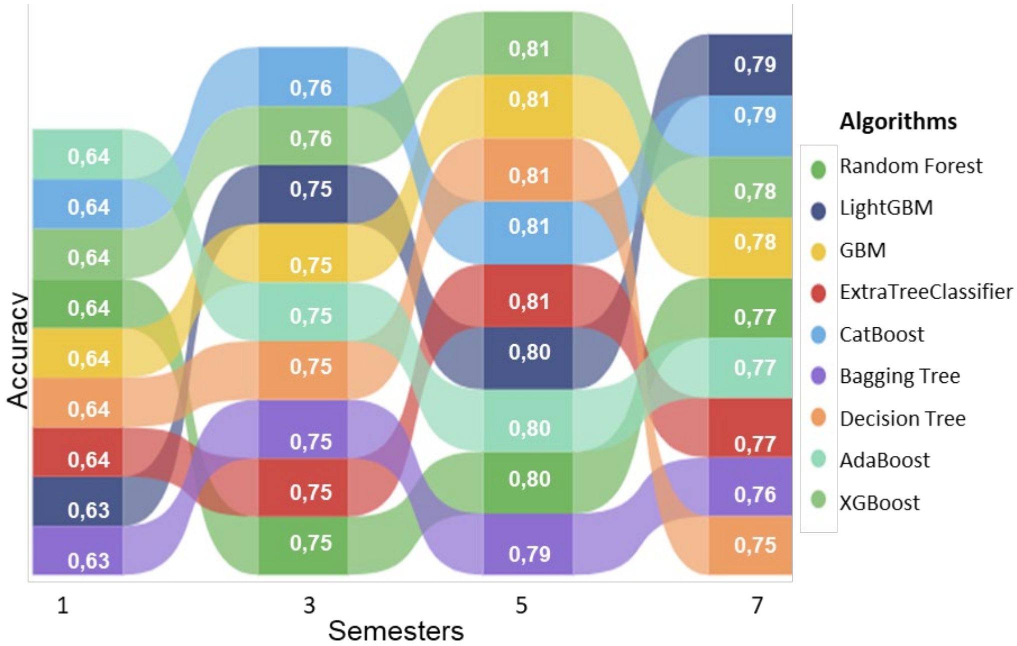


Figure 3. Example of Supervised ML and Assembly Algorithm Evaluation Metric Values for Industrial Engineering by Year

A third stage in the search for improved predictive performance involved implementing the Voting ensemble algorithm and its various variants. The results, presented in Figure 4, show that variants such as Soft Voting with different types of base (weak) algorithms, Soft Voting with weighted variations, and Blending (using base models and a linear meta-model) often produced performance metrics comparable to those achieved by the previously tested supervised and ensemble algorithms.

However, the Stacking variants (involving base models and a non-linear meta-model) and the Super Learner approach demonstrated significantly stronger performance. These methods achieved accuracy values approaching 90% on both the training data and the test data (unseen by the model), highlighting their robustness and predictive power.

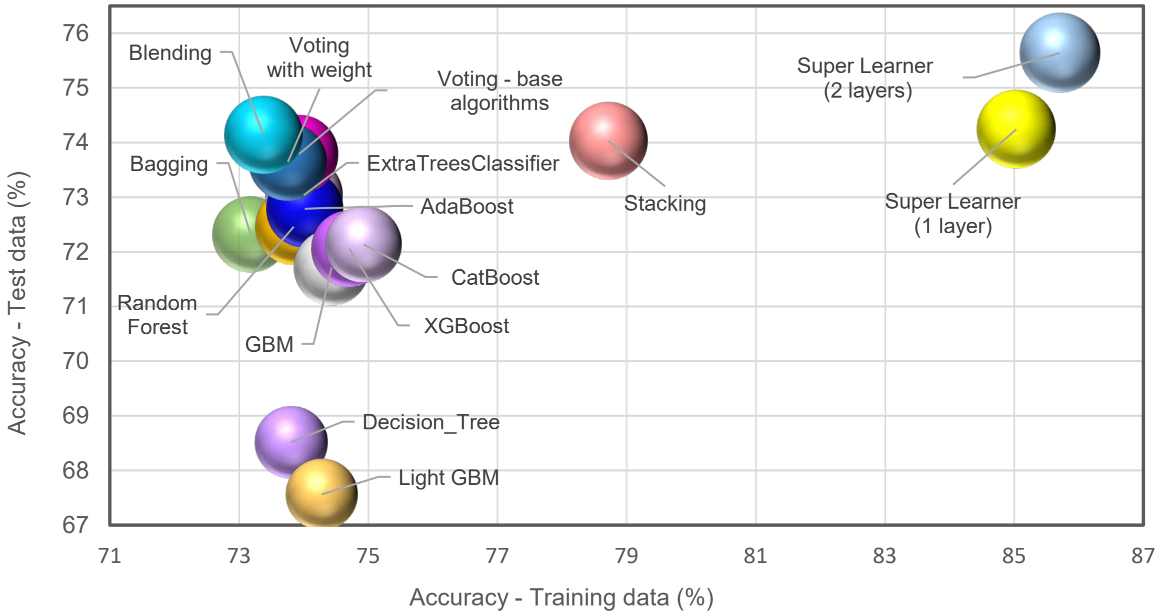


Figure 4. Machine Learning algorithms for prediction of academic performance in Industrial Engineering

3.5 Computational Model

The computational model for integrating data from the educational environment with machine learning tools to predict academic performance in higher education is presented in Figure 5.

In the first stage, it is essential to capture information from various data sources, including academic, pre-university, and admission records. These data are grouped into five key factors: sociodemographic, learning, academic, institutional academic environment, and psychological. The most influential factors—as determined by the model—are highlighted in the figure. Additionally, from the existing data, two derived variables that significantly influence performance are constructed: the distance between the student's home and the university, and the time spent in daily transportation.

The second stage involves data integration and cleaning, given that academic records often contain duplicated information across different institutional databases. This stage also addresses the presence of outliers, which may distort analysis and modeling outcomes.

The third stage focuses on data transformation. Due to the skewed and non-Gaussian nature of academic data—often exhibiting exponential distributions (e.g., the number of students failing a subject)—the Yeo-Johnson transformation method is identified as the most suitable option for normalization.

For feature selection, the model recommends the use of ensemble methods, particularly XGBoost and CatBoost. Among the algorithms implemented in this study, these methods not only identified the most influential variables but also led to improvements in predictive accuracy, making them the most effective for determining academic performance.

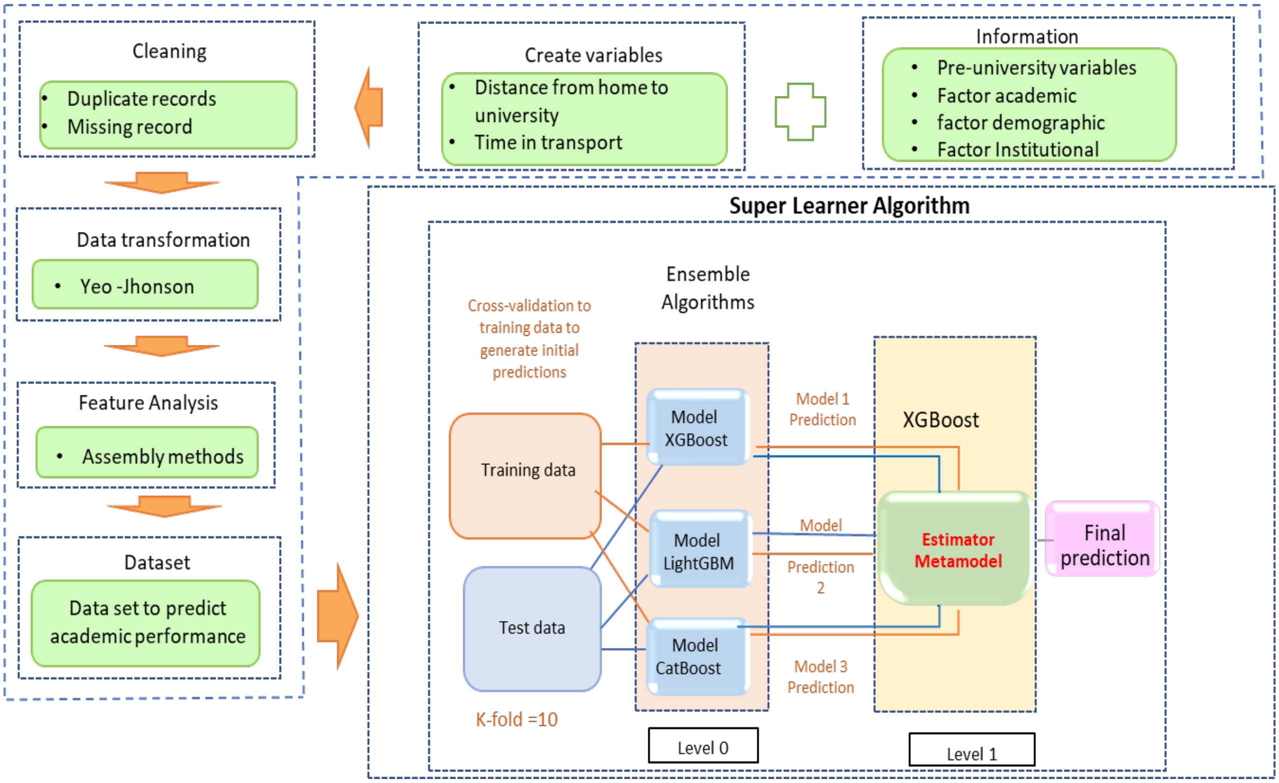


Figure 5. A computational model for integrating data from the educational field and Machine Learning tools to determine the academic performance

At this stage, the dataset is well-conditioned, and the process advances to the final stage of the model: the implementation of the prediction algorithm. The modeling procedure consists of selecting a dataset, cleaning and transforming it, and dividing it into training and testing sets. Subsequently, k-fold cross-validation is applied to all base models (level 0) to prevent overfitting, and an estimator metamodel is used to generate out-of-fold predictions from each level 0 model.

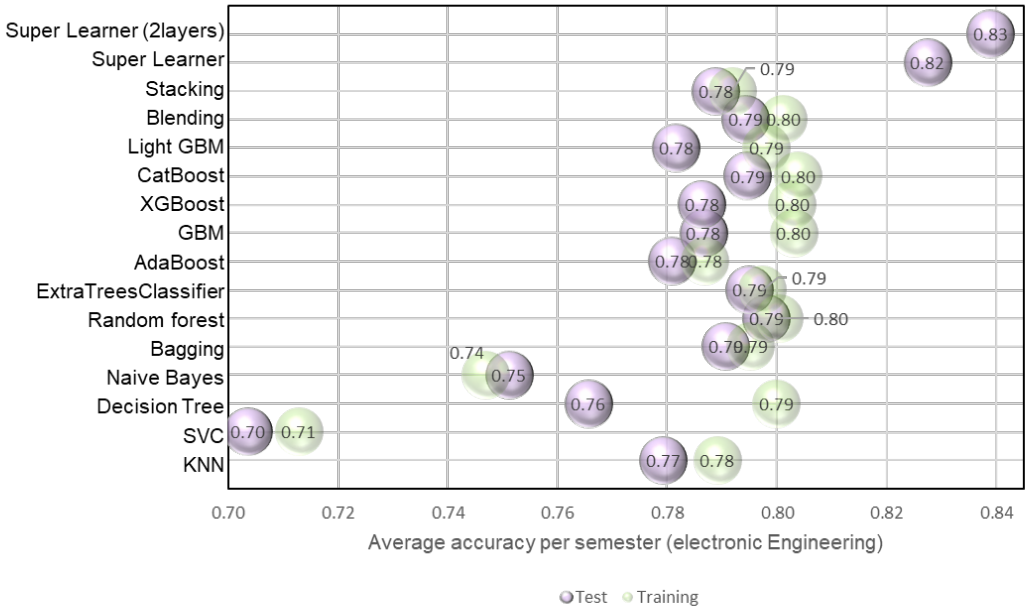
The implementation of a Super Learner algorithm enables the use of a meta-learner or meta-estimator, along with various algorithms—both supervised and ensemble-based—as level 0 models responsible for producing initial predictions. These first-level predictions are generated based on optimized hyperparameter configurations, and their outputs can be combined to enhance overall model performance.

Following experimentation with 18 algorithms, the recommended level 0 algorithms for this model are XGBoost, CatBoost, and LightGBM, as they demonstrated the best predictive performance.

The meta-estimator or meta-metamodel (when applied in two layers) tested in this study included both linear models (logistic regression, linear regression) and non-linear models (XGBoost, CatBoost). Among these, XGBoost emerged as the recommended meta-estimator, as it consistently produced results that generalized well, achieving test accuracy values close to training accuracy.

Additionally, the base algorithms achieved higher scores on evaluation metrics compared to other methods. The Super Learner algorithm offers the advantage of combining and weighting individual base models according to how well each minimizes a specific loss function. It also allows the development of a metamodel that corrects the predictions of weaker base models. Moreover, by creating multiple layers of prediction (as done in this study with two layers), the model effectively proposes a meta-metamodel estimator capable of enhancing prediction accuracy.

Figure 6 presents the average evaluation metrics of the final model applied to two of the engineering programs analyzed.



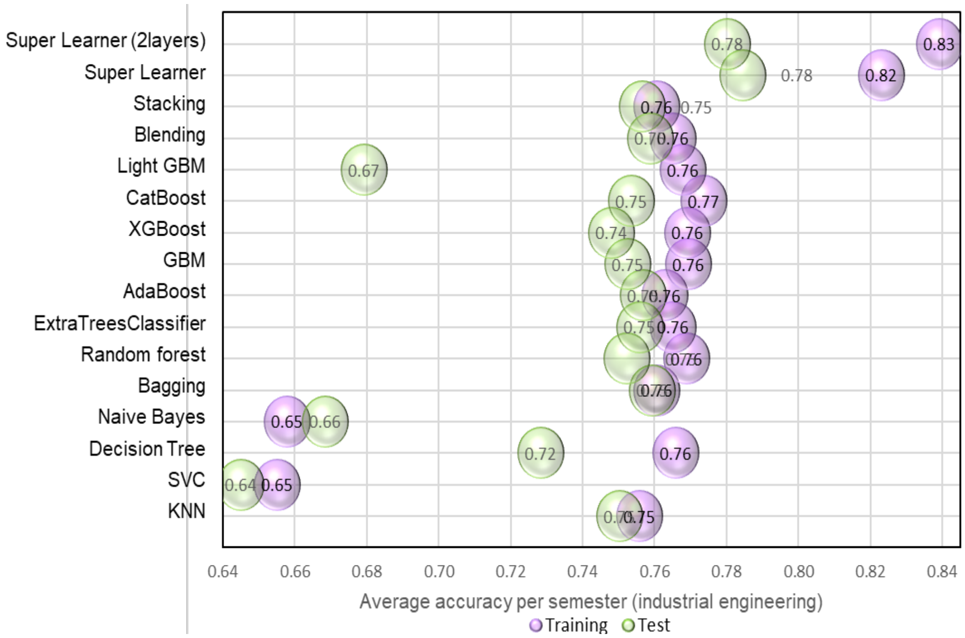


Figure 6. Comparison of model results for performance prediction in various engineering

4. DISCUSSIONS

Previous research has indicated that the gender variable may directly influence students' academic performance. Studies such as those by [76], [77] reported significant differences between genders, suggesting a measurable impact on academic outcomes. However, in alignment with the findings of this study, [78] concluded that gender differences were not statistically significant, implying that gender has no consistent or invariant effect on academic performance in higher education.

Regarding the feasibility of predicting first-semester performance using pre-university variables, it is important to highlight that while the Colombian ICFES state exam—taken at the end of high school—has predictive value for student performance in subsequent semesters, it is not a reliable indicator of performance in the first semester. According to [79], more effective results could be obtained through career-specific entrance exams, which have demonstrated stronger predictive power for early academic success. Although pre-university data contributes to understanding student trajectories, this research found that transformation methods, variable characteristics, and selected algorithms produced less favorable metrics for first-semester predictions compared to those in later semesters or academic years.

No strong correlation was found between academic year or semester (also referred to as study generation) and student performance, a finding that aligns with [77], [80], [81]. Instead, the most influential factor was the prior academic record, particularly grades obtained in specific subjects since university entry. Additional influential variables included the number of courses taken and the number of courses passed, as also noted by [81].

The study also confirmed that the choice of feature selection method significantly affects algorithm performance. For instance, while [53] utilized ensemble methods to predict performance at the end of the first year—applying linear regression for variable selection—this study implemented 20 feature selection methods and found that ensemble-based approaches (XGBoost, CatBoost, and LightGBM) outperformed filter (Pearson, ANOVA, Chi-Square), wrapper (logistic and linear regression), and embedded methods (Lasso, Ridge) in identifying the most influential features for classification tasks.

Regarding algorithm performance, supervised algorithms achieved accuracy and precision scores ranging from 50% to 70% on average when used to predict performance across the three engineering disciplines analyzed—well below the ideal of 100%. These results are comparable to those reported in [82], who achieved 86% accuracy using AdaBoost with academic, demographic, and behavioral features. In contrast, ensemble algorithms implemented in this study delivered higher metrics, with Bagging techniques averaging 73% and Boosting techniques averaging 75% in the analyzed engineering programs.

Further experimentation with Voting-based ensemble methods revealed that Blending (with a linear meta-model) yielded average accuracy metrics of approximately 73%, while Stacking performed slightly better, with averages around 79%. However, the Super Learner variant showed superior performance, achieving an average accuracy of 86% when used as a meta-meta estimator for predicting university academic performance. Other performance metrics, such as precision and recall, followed a similar trend. The Super Learner algorithm stands out for its ability to combine and weight multiple optimized base models, such as XGBoost, CatBoost, and LightGBM, minimizing the loss function through cross-validation. Unlike Blending and Stacking, the Super Learner trains a meta-model using out-of-fold predictions from the optimized base models, improving its generalization ability.

Among all tested algorithms, the Super Learner, under the Voting ensemble category, provided the best performance metrics across semester and yearly evaluations for all three engineering disciplines. However, its effectiveness depends heavily on the optimization of hyperparameters for each base algorithm—an inherently computationally intensive process that must be tailored to each specific dataset.

This study proposes a comprehensive methodology for estimating academic performance in higher education by integrating data from multiple sources, selecting influential variables, and applying machine learning techniques—both supervised and ensemble-based. The resulting Super Learner model, configured as either a one-stage metamodel or a two-stage meta-meta model, consistently outperformed all individual and ensemble models in predicting student performance.

The findings also highlight that when predicting performance across academic years, data quality and quantity significantly impact evaluation metrics such as accuracy and precision. This supports the notion that early identification of at-risk students—through educational data mining—can enable institutions to design timely interventions, training programs, and support strategies that enhance student success rates.

Despite concerns over the privacy of academic data, this research demonstrates that institutions often possess valuable information stored in disparate repositories. With appropriate data governance, this information can be leveraged not only to support students and faculty but also to inform institutional policies, such as implementing academic leveling programs for students in their first semesters. This approach fosters an integration between engineering and educational research and enhances the university's social impact by reducing dropout, attrition, and academic mortality in the early stages of higher education.

Finally, the implementation of machine learning-based predictive models carries important pedagogical implications. It strengthens data-driven academic decision-making by anticipating student performance, thereby facilitating the planning of preventive interventions, tutoring programs, and targeted support for underperforming students. These models promote a preventive rather than reactive approach to academic management, enabling institutions to reduce repetition and dropout rates, particularly in the first semesters.

By identifying key performance-influencing variables, institutions can also develop evidence-based policies, revise admissions criteria, and adapt curricula to emphasize subjects most critical to academic success. Moreover, the ability to offer personalized academic guidance and curriculum adaptation based on each student's profile represents a step toward more inclusive and effective education. However, achieving this requires institutional transformation, including the development of analytical capacities among staff and a strong commitment to the ethical and responsible use of data.

5. CONCLUSIONS

According to the results obtained, several main conclusions can be drawn. The effectiveness of machine learning algorithms in predicting student academic performance has been clearly demonstrated, particularly through the use of ensemble methods, which reduce the bias and variance commonly associated with supervised predictors such as SVC, KNN, decision trees, and Naive Bayes. The proposed model—which integrates data transformation, feature selection, and algorithm optimization—can be applied across various engineering fields and potentially extrapolated to other disciplines. However, it is important to note that the hyperparameters of the base algorithms required for the implementation of the Super Learner must be optimized according to the specific dataset being used. Furthermore, the multidimensional variable known as academic performance has been addressed from multiple disciplinary perspectives using diverse sets of variables. This study offers an approach from the field of data analytics that identifies influential factors and the most effective predictive algorithms, with the purpose of providing a model that supports both educators and students in making informed decisions throughout the academic journey.

REFERENCES

[1] D. Buenaño, D. Gil, and S. Luján, “Application of machine learning in predicting performance for computer engineering students: A case study,” *Sustainability*, vol. 11, no. 2833, May 2019. https://doi.org/10.3390/su11102833

[2] M. Ferreyra, C. Avitabile, J. Botero Álvarez, F. Haimovich Paz, and S. Urzúa, *At a Crossroads Higher Education in Latin America and the Caribbean Human Development*. Washington, 2017.

[3] V. Tinto, “Limits of theory and practice in student attrition,” *J. Higher Educ.*, vol. 53, no. 6, pp. 687–700, 1982, Accessed: Jun. 26, 2018. [Online]. Available: http://www.jstor.org/stable/1981525

[4] D. Rico, *Caracterización de la deserción estudiantil en la Universidad Nacional de Colombia*, 1st ed. Medellín, 2006.

[5] M. Khalil and M. Ebner, “Learning analytics: Principles and constraints,” *EdMedia World Conf. Educ. Media Technol.*, vol. 2015, no. 1, pp. 1789–1799, Jun. 2015, Accessed: Oct. 14, 2017. [Online]. Available: https://www.learntechlib.org/p/151455/

[6] C. Daza and P. Parra, “La gobernanza y su incidencia en los procesos de calidad en las instituciones de educación superior,” *Rev. Boletín Redipe*, vol. 8, no. 10, pp. 111–124, Oct. 2019. https://doi.org/10.36260/rbr.v8i10.838

[7] A.-B. Mashael and A.-R. Muna, “Predicting students’ final GPA using decision trees: A case study,” *Int. J. Inf. Educ. Technol.*, vol. 6–7, pp. 528–533, 2016. https://doi.org/10.7763/ijiet.2016.v6.745

[8] T. York, C. Gibson, and S. Rankin, “Defining and measuring academic success,” *Pract. Assessment, Res. Eval.*, vol. 20, no. 1, p. 5, Nov. 2019. https://doi.org/10.7275/hz5x-tx03

[9] M. Oladejo, “A path-analytic study of socio-psychological variables and academic performance of distance learners in Nigerian universities,” *Thesis*, p. 197, 2010. https://doi.org/10.13140/RG.2.2.19443.73762

[10] J. Tomás, M. Expósito, and S. Sempere, “Determinantes del rendimiento académico en los estudiantes de grado. Un estudio en administración y dirección de empresas,” *Rev. Investig. Educ.*, vol. 32, no. 2, pp. 379–392, 2014. https://doi.org/10.6018/rie.32.2.177581

[11] J. Estrada and R. Quintero, *Bajo rendimiento académico en la Universidad Distrital Francisco José de Caldas*. Bogotá: Editorial UD, 2015.

[12] D. García, “Construcción de un modelo para determinar el rendimiento académico de los estudiantes basado en learning analytics (análisis del aprendizaje), mediante el uso de técnicas multivariantes,” Ph.D. dissertation, Univ. Sevilla, 2016.

[13] D. García, J. Pino, and J. Muñoz, “Learning analytics as an analysis factor of university academic performance,” in *CEUR Workshop Proc.*, 2019, pp. 42–50, Accessed: May 14, 2020. [Online]. Available: http://ceur-ws.org/Vol-2231/

[14] E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: Literature review and best practices,” *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, pp. 1–21, Dec. 2020. https://doi.org/10.1186/s41239-020-00177-7

[15] R. Garg, “Predicting student performance of different regions of Punjab using classification techniques,” *Int. J. Adv. Res. Comput. Sci.*, vol. 9, no. 1, pp. 236–240, Feb. 2018. https://doi.org/10.26483/ijarcs.v9i1.5234

[16] J. Vega, “Modelo de pronóstico de rendimiento académico de alumnos en los cursos del programa de estudios básicos de la Universidad Ricardo Palma usando algoritmos de machine learning,” *Thesis*, pp. 1–358, 2019.

[17] H. A. Mengash, “Using data mining techniques to predict student performance to support decision making in university admission systems,” *IEEE Access*, vol. 8, pp. 55462–55470, 2020. https://doi.org/10.1109/access.2020.2981905

[18] A. I. Adekitan and E. Noma-Osaghae, “Data mining approach to predicting the performance of first year student in a university using the admission requirements,” *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1527–1543, 2019. https://doi.org/10.1007/s10639-018-9839-7

[19] N. Tomasevic, N. Gvozdenovic, and S. Vranes, “An overview and comparison of supervised data mining techniques for student exam performance prediction,” *Comput. Educ.*, vol. 143, p. 103676, 2020. https://doi.org/10.1016/j.compedu.2019.103676

[20] M. Mohammadi, M. Dawodi, W. Tomohisa, and N. Ahmadi, “Comparative study of supervised learning algorithms for student performance prediction,” in *Proc. 1st Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, 2019, pp. 124–127. https://doi.org/10.1109/icaiic.2019.8669085

[21] I. Burman and S. Som, “Predicting students academic performance using support vector machine,” in *Proc. Amity Int. Conf. Artif. Intell. (AICAI)*, Apr. 2019, pp. 756–759. https://doi.org/10.1109/aicai.2019.8701260

[22] S. Wiyono, T. Abidin, P. H. Bersama, P. Lor, and K. Tegal, “Comparative study of machine learning KNN, SVM, and decision tree algorithm to predict student's,” *Int. J. Comput. Appl.*, vol. 7, no. 1, pp. 190–196, 2019. https://doi.org/10.5281/zenodo.2550651

[23] J. Hou and Y. Wen, “Prediction of learners' academic performance using factorization machine and decision tree,” in *Proc. IEEE Int. Congr. Cybermatics*, 2019, pp. 1–8. https://doi.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00024

[24] E. J. De La Hoz and T. J. Fontalvo, “Methodology of machine learning for the classification and prediction of users in virtual education environments,” *Inf. Tecnol.*, vol. 30, no. 1, pp. 247–254, Feb. 2019. https://doi.org/10.4067/s0718-07642019000100247

[25] A. Das and E. Rodriguez, “A predictive analytics system for forecasting student academic performance: Insights from a pilot project at Eastern

Washington University," in *Proc. 8th Int. Conf. Informatics, Electron. Vision (ICIEV)*, 2019, pp. 255–262. https://doi.org/10.1109/ICIEV.2019.8858523

[26] S. S. M. Ajibade, N. B. B. Ahmad, and S. M. Shamsuddin, "Educational data mining: Enhancement of student performance model using ensemble methods," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 551, no. 1, 2019. https://doi.org/10.1088/1757-899x/551/1/012061

[27] H. Anderson, B. Afshan, and R. Baker, "Predicting graduation at a public R1 university," 2019. [Online]. Available: https://www.researchgate.net/publication/331230878_Predicting_Graduation_at_a_Public_R1_University

[28] T. C. Hakyemez and S. Mardikyan, "The interplay between institutional integration and self-efficacy in the academic performance of first-year university students: A multigroup approach," *Int. J. Manag. Educ.*, vol. 19, no. 1, 2021. https://doi.org/10.1016/j.ijme.2020.100430

[29] T. Icekson, O. Kaplan, and O. Slobodin, "Does optimism predict academic performance? Exploring the moderating roles of conscientiousness and gender," *Stud. High. Educ.*, vol. 45, no. 3, pp. 635–647, 2020. https://doi.org/10.1080/03075079.2018.1564257

[30] A. Lenskiy, R. Shariat, and S. Seol, "The effect of academic breaks on undergraduate academic performance," *Int. J. Electr. Eng. Educ.*, pp. 1–12, 2020. https://doi.org/10.1177/0020720920922518

[31] A. M. Pavelea and O. Moldovan, "Why some fail and others succeed: Explaining the academic performance of PA undergraduate students," *NISPAcee J. Public Adm. Policy*, vol. 13, no. 1, pp. 109–132, 2020. https://doi.org/10.2478/nispa-2020-0005

[32] R. Patil, S. Salunke, M. Kalbhor, and R. Lomte, "Prediction system for student performance using data mining classification," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, 2018, pp. 2018–2021. https://doi.org/10.1109/iccubea.2018.8697770

[33] F. J. Kaunang and R. Rotikan, "Students' academic performance prediction using data mining," in *Proc. 3rd Int. Conf. Informatics Comput. (ICIC)*, 2018, pp. 1–5. https://doi.org/10.1109/iac.2018.8780547

[34] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Bus. Res.*, vol. 94, pp. 335–343, 2019. https://doi.org/10.1016/j.jbusres.2018.02.012

[35] A. Rico and D. Sánchez, “Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN / Design of a model to automate the prediction of academic performance in students of IPN,” *RIDE Rev. Iberoam. Investig. Desarro. Educ.*, vol. 8, no. 16, pp. 246–266, 2018. https://doi.org/10.23913/ride.v8i16.340

[36] C. Jalota and R. Agrawal, “Analysis of educational data mining using classification,” in *Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Perspectives Prospect. (ComITCon)*, 2019, pp. 243–247. https://doi.org/10.1109/comitcon.2019.8862214

[37] W. Wei, J. Han, J. Kong, and H. Xia, “Prediction of the scholarship using comprehensive development,” in *Proc. 4th Int. Conf. Enterp. Syst. Adv. Enterp. Syst. (ES)*, 2016, pp. 183–188. https://doi.org/10.1109/es.2016.30

[38] S. Bhutto, I. F. Siddiqui, Q. A. Arain, and M. Anwar, “Predicting students' academic performance through supervised machine learning,” in *Proc. 2nd Int. Conf. Inf. Sci. Commun. Technol. (ICISCT)*, 2020. https://doi.org/10.1109/icisct49550.2020.9080033

[39] L. M. Crivei, G. Czibula, G. Ciubotariu, and M. Dindelegan, “Unsupervised learning based mining of academic data sets for students' performance analysis,” in *Proc. IEEE 14th Int. Symp. Appl. Comput. Intell. Informatics (SACI)*, 2020, pp. 11–16. https://doi.org/10.1109/saci49304.2020.9118835

[40] O. Iatrellis, I. Savvas, P. Fitsilis, and V. Gerogiannis, “A two-phase machine learning approach for predicting student outcomes,” *Educ. Inf. Technol.*, 2020. https://doi.org/10.1007/s10639-020-10260-x

[41] L. Santoso and W. Yulia, *The Analysis of Student Performance Using Data Mining*, vol. 924. Singapore: Springer, 2019.

[42] V. Kumar, A. Krishna, P. Neelakanteswara, and C. Basha, “Advanced prediction of performance of a student in a university using machine learning techniques,” in *Proc. Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, 2020, pp. 121–126. https://doi.org/10.1109/icesc48915.2020.9155557

[43] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, “Student engagement level in an e-learning environment: Clustering using K-means,” *Am. J. Distance Educ.*, vol. 34, no. 2, pp. 137–156, 2020. https://doi.org/10.1080/08923647.2020.1696140

[44] A. Almasri, R. S. Alkhawaldeh, and E. Çelebi, “Clustering-based EMT model for predicting student performance,” *Arab. J. Sci. Eng.*, vol. 45, no. 12, pp. 10067–10078, 2020. https://doi.org/10.1007/s13369-020-04578-4

[45] R. Trakunphutthirak and V. C. S. Lee, *Application of Educational Data Mining Approach for Student Academic Performance Prediction Using Progressive Temporal Data*, 2021.

[46] M. Kumar, G. Mehta, N. Nayar, and A. Sharma, “EMT: Ensemble meta-based tree model for predicting student performance in academics,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021. https://doi.org/10.1088/1757-899x/1022/1/012062

[47] S. Sakri and A. S. Alluhaidan, “RHEM: A robust hybrid ensemble model for students' performance assessment on cloud computing course,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 388–396, 2020. https://doi.org/10.14569/ijacsa.2020.0111150

[48] M. Ragab, A. M. K. Abdel Aal, A. O. Jifri, and N. F. Omran, “Enhancement of predicting students performance model using ensemble approaches and educational data mining techniques,” *Wirel. Commun. Mob. Comput.*, vol. 2021, 2021. https://doi.org/10.1155/2021/6241676

[49] M. Jawthari and V. Stoffova, “Effect of encoding categorical data on student's academic performance using data mining methods,” *eLearning Softw. Educ.*, vol. 1, pp. 521–526, 2020. [Online]. Available: http://10.0.49.209/2066-026X-20-068

[50] W. Chango, R. Cerezo, and C. Romero, “Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses,” *Comput. Electr. Eng.*, vol. 89, 2021. https://doi.org/10.1016/j.compeleceng.2020.106908

[51] H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” *Comput. Electr. Eng.*, vol. 89, p. 106903, 2021. https://doi.org/10.1016/j.compeleceng.2020.106903

[52] M. Bucos and B. Drăgulescu, “Predicting student success using data generated in traditional educational environments,” *TEM J.*, vol. 7, no. 3, pp. 617–625, 2018. https://doi.org/10.18421/TEM73-19

[53] E. Yamao, L. Saavedra, R. Campos, and V. Huancas, “Prediction of academic performance using data mining in first year students of Peruvian university,” *Rev. USMP – Campus*, vol. 23, no. 26, pp. 151–160, 2018.

[54] H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” *Comput. Electr. Eng.*, vol. 89, p. 106903, 2021. https://doi.org/10.1016/j.compeleceng.2020.106903

[55] A. C. Lagman, L. P. Alfonso, M. L. I. Goh, J. A. P. Lalata, J. P. H. Magcuyao, and H. N. Vicente, “Classification algorithm accuracy improvement for student graduation prediction using ensemble model,” *Int. J. Inf. Educ. Technol.*, vol. 10, no. 10, pp. 723–727, 2020. https://doi.org/10.18178/ijiet.2020.10.10.1449

[56] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining,” *Comput. Educ.*, vol. 113, pp. 177–194, 2017. https://doi.org/10.1016/j.compedu.2017.05.007

[57] G. Kostopoulos, S. Kotsiantis, C. Pierrakeas, G. Koutsonikos, and G. A. Gravvanis, “Forecasting students’ success in an open university,” *Int. J. Learn. Technol.*, vol. 13, no. 1, pp. 26–43, 2018. https://doi.org/10.1504/IJLT.2018.091630

[58] J. F. Vega García, “Modelo de pronóstico de rendimiento académico de alumnos en los cursos del programa de estudios básicos de la Universidad Ricardo Palma usando algoritmos de Machine Learning,” *Thesis*, 2019.

[59] J. Del Campo, G. Ramos, R. Morales, and M. Baena, “Minería de datos educativos para la predicción personalizada del rendimiento académico,” 2018. [Online]. Available: https://riuma.uma.es/xmlui/handle/10630/15477

[60] L. L. Ochoa, K. Rosas Paredes, and C. Baluarte Araya, “Evaluación de técnicas de minería de datos para la predicción del rendimiento académico,” in *Proc. LACCEI Int. Multi-Conference Eng. Educ. Technol.*, vol. 2017-July, 2017. https://doi.org/10.18687/LACCEI2017.1.1.368

[61] L. Contreras, G. Tarazona, and J. Rodríguez, “Tecnología y analítica del aprendizaje: una revisión a la literatura,” *Rev. Científica*, vol. 41, no. 2, pp. 150–168, May 2021. https://doi.org/10.14483/23448350.17547

[62] L. Contreras, H. Fuentes, and J. Molano, “Analítica académica: nuevas herramientas aplicadas a la educación,” *Rev. Bol. Redipe*, vol. 10, no. 3, pp. 137–158, 2021.

[63] H. Fuentes, L. Contreras, and J. Rodríguez, “Academic interruption model using automatic learning algorithms,” *Int. J. Mech. Prod. Eng. Res. Dev.*, vol. 10, no. 3, pp. 16075–16086, 2020. [Online]. Available: http://www.tjprc.org

[64] L. Contreras, H. Fuentes, and J. Rodríguez, “Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático,” *Form. Univ.*, vol. 13, no. 5, pp. 233–246, 2020. https://doi.org/10.4067/S0718-50062020000500233

[65] K. Gonzalez, J. Rodríguez, and L. Contreras, “Academic performance and alternatives with prediction-oriented machine learning: A review of the state of the art,” *Int. J. Mech. Prod. Eng. Res. Dev.*, vol. 10, no. 3, pp. 16329–16340, 2020.

[66] G. Tarazona, L. Contreras, and H. Fuentes, “Machine learning variables and algorithms that influence academic performance: A review,” *Int. J. Mech. Prod. Eng. Res. Dev.*, vol. 10, no. 3, pp. 16011–16028, 2020.

[67] E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: Literature review and best practices,” *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, pp. 1–21, Dec. 2020. https://doi.org/10.1186/s41239-020-00177-7

[68] A. Namoun and A. Alshanqiti, “Predicting student performance using data mining and learning analytics techniques: A systematic literature review,” *Appl. Sci.*, vol. 11, no. 1, pp. 1–28, Jan. 2021. https://doi.org/10.3390/app11010237

[69] J. Rastrollo, J. Gómez, and A. Durán, “Analyzing and predicting students' performance by means of machine learning: A review,” *Appl. Sci.*, vol. 10, no. 3, Feb. 2020. https://doi.org/10.3390/app10031042

[70] S. Salas and Y. Yang, “Artificial intelligence applications in Latin American higher education: A systematic review,” *Int. J. Educ. Technol. High. Educ.*, vol. 19, no. 21, 2022. https://doi.org/10.1186/s41239-022-00326-w

[71] M. van der Laan, E. Polley, and A. Hubbard, “Super learner,” *Stat. Appl. Genet. Mol. Biol.*, vol. 6, no. 1, Sep. 2007. https://doi.org/10.2202/1544-6115.1309

[72] E. Costa, B. Fonseca, M. Almeida, and F. Ferreira, “Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses,” *Comput. Human Behav.*, vol. 73, pp. 247–256, Aug. 2017. https://doi.org/10.1016/j.chb.2017.01.047

[73] P. Sánchez, J. García, M. Orozco, and S. Obredor, “Knowledge capture for the prediction and analysis of results of the quality test of higher education in Colombia,” *Rev. Form. Univ.*, vol. 12, no. 4, pp. 55–62, 2019. https://doi.org/10.4067/S0718-50062019000400055

[74] O. Castrillón, W. Sarache, and S. Ruiz, “Predicción del rendimiento académico por medio de técnicas de inteligencia artificial,” *Rev. Form. Univ.*, vol. 13, no. 1, pp. 93–102, 2020. https://doi.org/10.4067/S0718-50062020000100093

[75] R. M. Aguilar, J. M. Torres, and C. A. Martín, “Automatic learning for the system identification. A case study in the prediction of power generation in a wind farm,” *RIAI – Rev. Iberoam. Autom. e Inform. Ind.*, vol. 16, no. 1, pp. 114–127, 2019. https://doi.org/10.4995/riai.2018.9421

[76] X. J. Lin et al., “Stress and its association with academic performance among dental undergraduate students in Fujian, China: A cross-sectional online questionnaire survey,” *BMC Med. Educ.*, vol. 20, no. 181, p. 181, 2020. https://doi.org/10.1186/s12909-020-02095-4

[77] Q. Mazumder, S. Sultana, and F. Mazumder, “Correlation between classroom engagement and academic performance of engineering students,” *Int. J. High. Educ.*, vol. 9, no. 3, pp. 240–247, 2020. https://doi.org/10.5430/ijhe.v9n3p240

[78] A. Furnham, S. Nuygards, and T. Chamorro, “Personality, assessment methods and academic performance,” *Instr. Sci.*, vol. 41, no. 5, pp. 975–987, 2013. [Online]. Available: https://www.jstor.org/stable/43575408

[79] T. González and A. García, “El rendimiento académico en matemáticas discretas: un estudio predictivo / Academic performance in discrete mathematics: a predictive study,” *Atenas Rev. Científico Pedagógica*, vol. 14, no. 11, pp. 4573–4578, 2020. [Online]. Available: https://cache.1science.com/3b/18/3b18065b3a38adb07839c6bf07bd00c87e01147b.pdf

[80] S. Awadalla, E. B. Davies, and C. Glazebrook, “A longitudinal cohort study to explore the relationship between depression, anxiety and academic performance among Emirati university students,” *BMC Psychiatry*, vol. 20, no. 448, 2020. [Online]. Available: https://bmcpsychiatry.biomedcentral.com/track/pdf/10.1186/s12888-020-02854-z.pdf

[81] G. Bautista and F. Gatica, “Factores relacionados con el rendimiento académico en una carrera técnica en salud impartida en línea,” *Investig. en Educ. Médica*, vol. 9, no. 33, pp. 89–97, 2020. [Online]. Available: http://www.scielo.org.mx/pdf/iem/v9n33/2007-5057-iem-9-33-89.pdf

[82] A. I. Adekitan and E. Noma-Osaghae, “Data mining approach to predicting the performance of first year student in a university using the admission requirements,” *Educ. Inf. Technol.*, vol. 24, pp. 1527–1543, 2018. https://doi.org/10.1007/s10639-018-9839-7