A review on the prediction of students' academic performance using ensemble methods

Una revisión sobre la predicción del rendimiento académico mediante métodos de ensamble

Uma revisão sobre a previsão do desempenho acadêmico usando métodos de ensemble

Leonardo Emiro Contreras Bravo[1]
Joan Alejandro Caro Silva[2]
Danna Lorena Morales Rodríguez[3]

Received: January 15th, 2022
Accepted: March 25th, 2022
Available: May 10th, 2022

How to cite this article:
L.E. Contreras Bravo, J.A. Caro Silva, D.L. Morales Rodríguez, "A review on the prediction of students' academic performance using ensemble methods," *Revista Ingeniería Solidaria,* vol. 18, no. 2, 2022.
doi: https://doi.org/10.16925/2357-6014.2022.02.01

Research article. https://doi.org/10.16925/2357-6014.2022.02.01

[1] Ingeniero, Estudiante de doctorado en ingeniería, Docente de planta. Facultad de ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia.

Email: lecontrerasb@udistrital.edu.co

ORCID: https://orcid.org/0000-0003-4625-8835

CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000675024

[2] Estudiante de Ingeniería Industrial, Facultad de ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia.

Email: jacaros@correo.udistrital.edu.co

ORCID: https://orcid.org/0000-0003-0373-0757

CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001999698

[3] Estudiante de Ingeniería Industrial, Facultad de ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia.

Email: dlmoralesr@correo.udistrital.edu.co

ORCID: https://orcid.org/0000-0001-5816-9499

CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001999697

Abstract

Introduction: This article is a product of the research "Ensemble methods to estimate the academic performance of higher education students", developed at the Universidad Distrital Francisco José de Caldas in the year 2021, focusing on the review of research work developed in the last five years related to the prediction of academic performance using ensemble algorithms.

Objective: The literature review aims to identify the most used algorithms and the most relevant variables in the prediction of academic performance.

Methodology: A systematic review of the literature was carried out in different academic databases (Science Direct, Scopus, SAGE Journals, EBSCO, ResearchGate, Google Scholar), using search equations built with keywords.

Results: 54 related articles were found that meet the inclusion criteria of the review. Additionally, benefits were found in the application of ensemble methods in the prediction of academic performance.

Conclusion: It was found that the most influential variables in academic performance correspond to the academic factor. The algorithm used that presents the best results is Random Forest; in addition to being the most used. The use of these algorithms is an accurate tool to predict academic performance at any stage of university life, and at the same time provide information to generate strategies to improve dropout and academic retention indicators.

Keywords: Educational Data Mining (EDM), Academic Performance, Machine Learning, Ensemble Methods.

Resumen

Introducción: El presente artículo es producto de la investigación "Métodos de ensamble para estimar el rendimiento académico de estudiantes de educación superior", desarrollado en la Universidad Distrital Francisco José de Caldas en el año 2021 y se centra en la revisión de trabajos de investigación desarrollados en los últimos cinco años relacionados a la predicción del rendimiento académico utilizando algoritmos de ensamble.

Objetivo: La revisión de la literatura tiene como objetivo identificar los algoritmos más utilizados y las variables más relevantes en la predicción del rendimiento académico.

Metodología: Se realizó una revisión sistemática de la literatura en distintas bases de datos académicas (Science Direct, Scopus, SAGE Journals, EBSCO, ResearchGate, Google Scholar), utilizando ecuaciones de búsqueda construidas con palabras claves.

Resultados: Se encontraron 54 artículos relacionados que cumplen con los criterios de inclusión de la revisión. Además, se encontraron beneficios en la aplicación de métodos de ensamble en la predicción del rendimiento académico.

Conclusión: Se encontró que las variables más influyentes en el rendimiento académico corresponden al factor académico, el algoritmo utilizado que presenta mejores resultados es Random Forest, además de que fue el más utilizado, y que el uso de estos algoritmos es una herramienta precisa para predecir el rendimiento académico en cualquier etapa de la vida universitaria, y a su vez brindar la información para generar estrategias que permitan mejorar los indicadores de deserción y retención académica.

Palabras clave: Educational Data Mining (EDM), Rendimiento académico, Machine Learning, Métodos de ensamble.

Resumo

Introdução: Este artigo é produto da pesquisa "Métodos de montagem para estimar o desempenho acadêmico de estudantes do ensino superior", desenvolvida na Universidade Distrital Francisco José de Caldas no ano de 2021 e tem como foco a revisão de trabalhos de pesquisa desenvolvidos nos últimos cinco anos. anos relacionados à previsão de desempenho acadêmico usando algoritmos de conjunto.

Objetivo: A revisão de literatura visa identificar os algoritmos mais utilizados e as variáveis mais relevantes na previsão do desempenho acadêmico.

Metodologia: Foi realizada revisão sistemática da literatura em diferentes bases acadêmicas (Science Direct, Scopus, SAGE Journals, EBSCO, ResearchGate, Google Scholar), utilizando equações de busca construídas com palavras-chave.

Resultados: foram encontrados 54 artigos relacionados que atendem aos critérios de inclusão da revisão. Além disso, foram encontrados benefícios na aplicação de métodos de ensemble na previsão do desempenho acadêmico.

Conclusão: Constatou-se que as variáveis mais influentes no desempenho acadêmico correspondem ao fator acadêmico, o algoritmo utilizado que apresenta os melhores resultados é o Random Forest, além de ser o mais utilizado, e que o uso desses algoritmos é uma ferramenta precisa prever o desempenho acadêmico em qualquer fase da vida universitária e, por sua vez, fornecer informações para gerar estratégias para melhorar os indicadores de evasão e retenção acadêmica.

Palavras-chave: Mineração de Dados Educacionais (EDM), Desempenho Acadêmico, Aprendizado de Máquina, Métodos de Montagem.

1. INTRODUCTION

The university educational field faces the constant challenge of maintaining and improving academic quality day by day, since today's society requires it. Due to this, strategies are proposed in search of being able to guarantee adequate quality standards, which result in improving the performance and retention of students [1].

It is a difficult task to be able to define the appropriate actions and decisions that maximize student performance since there are several influencing factors [1] such as: social, economic, historical, individual, macroeconomic, state educational policy and institutional factors, among others [2]; it is a complex and multidimensional theoretical construct. For this reason, academic performance has been represented in different ways in the various studies that have addressed the subject, and it also requires an integration of the different techniques and methodologies in order to predict it [3], [4].

Low academic performance is often associated with a high dropout rate [5]. In most middle and higher educational institutions around the world, failure and dropout rates show very high values when compared to the results of basic education [6]–[8]. This has become a problem of growing interest not only for higher education institutions, but also for educational authorities due to its socio-economic consequences [9], [10].

Dropout can be defined as an individual event of interruption or disassociation from the institutional academic trajectory as a result of one or several processes at the personal, institutional or social level [9]. In order to reduce the dropout rate, it is necessary to have a mechanism that allows students to determine possible academic risk situations [11], [12].

That is why, in this document, the cases of application of data mining and assembly methods in education are compiled and presented. It is important to mention that the objective is to provide a systematic review that guarantees transparency in the methodology; gray literature (such as government reports and policy documents) is also omitted since it can bias perspectives [13].

To contextualize the topics addressed in the literature review, the following concepts are addressed: analytics in education, machine learning and ensemble methods.

1.1 Analytics in education

The main objective of educational systems is to provide knowledge, tools and skills for students. The way in which educational systems effectively meet this objective is a determining factor for social and economic progress [14]. This is why the need arises to have management systems that help to make the right decisions, since this not only affects academic departments and internal issues, but also activities such as accreditations [6].

Due to the great advances within ICT, it is common for educational institutions to have enough information about their students that can be easily accessed. It is certain that potentially useful information can be found in these data [7] that can benefit the teaching and learning processes of various educational institutions. This type of analysis is performed through data mining techniques [15] which have the purpose of extracting significant knowledge from the data [16]. The application of data mining methods to educational data is known as Educational Data Mining (EDM) [17], [18], which, together with Machine Learning, are responsible for the collection, analysis and dissemination of educational data with the purpose of understanding and optimizing related aspects of the teaching-learning process [19].

In the same way, there are several ways to measure the efficiency of the educational process, as is the case of the retention phenomenon, defined as the difference between 'the number of students who enter the first semester' and 'the number of graduates' per year [20] and the academic performance, defined as the primary indicator of student success or failure.

It is here where areas such as data mining and engineering add value by proposing solutions to multiple aspects of an academic-administrative nature [3], providing tools that generate timely analyses to establish adequate strategies to improve academic performance and reduce student desertion.

1.2 Machine learning and ensemble methods

Machine learning, also called automatic learning, is a discipline of artificial intelligence whose objective is to develop techniques so that machines can learn automatically through experience, which is obtained through the analysis of millions of data and the identification of their behavior patterns [21]. This can be divided into two large groups: classic and contemporary, as shown in Figure 1.

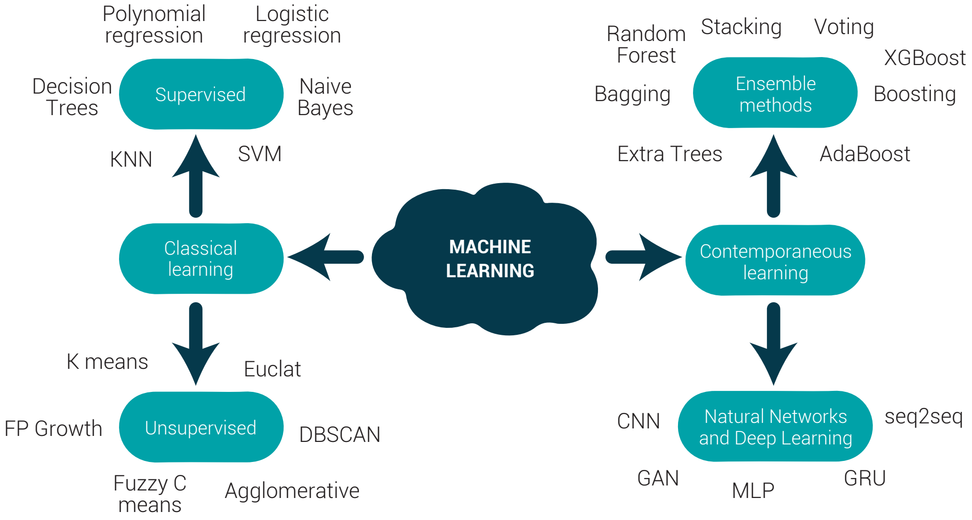

Figure 1. Overview of machine learning.
Source: Own work.

In classical learning there is supervised learning, which takes a set of training data and the algorithm tries to build patterns from these to predict an output value [22]. These supervised learning algorithms are frequently used for regression (prediction of numerical variables) and regression (prediction of categorical variables) tasks [22], [23]. On the other hand, we have unsupervised learning, in which a set of training data is taken, and the algorithm itself identifies characteristics, regularities, correlations or categories of the set, but does not predict a response value [21]–[23].

Within contemporary learning are neural networks, which simulate the biological central nervous system to predict a response based on a reward system [22], these are very efficient to solve regression and classification problems [24]. On the other hand, there are the so-called aggregation or assembly methods, which consist of building several predictors, from the same set of data, and combining them in some way to obtain a more stable predictor with higher performance than from other predictors, had they worked alone [25]–[27]. The main techniques that are developed within the assembly methods in Machine Learning are Bagging, Boosting and Stacking.

Bagging or Bootstrap aggregation [28]–[30] is a method that aggregates or assembles independent predictions and forms a final prediction [31]. This is effective even with unstable learning algorithms [32]. This algorithm partitions the dataset into 'm' random data samples of size 'n' [33] and almost always runs the same model for each subsample. The independence of each algorithm used in parallel is exploited, and the final prediction of the combination is obtained by voting or the average of all responses [25], [31].

Boosting is a method used to solve classification and regression tasks [34]. This algorithm takes a set of data and executes sequential models in a cascade fashion. Each model starts from the result of a previous model [25], [31], [35], and each one learns to correct the prediction errors of the previous model [36]. The final prediction is produced by combining the predictions of the previous models by means of voting or weighted sum, in this way a more efficient final prediction is obtained, which reduces the variance and also the bias [25], [31].

Stacking is an algorithm that has the particularity of combining the predictions from different learning algorithms, and the final model results from the combination of the predictions of the simple models, called base-learners [22], [33], [37]. To reduce the risk of overfitting, simple models such as meta-learners [38] are usually chosen. Training a stacking model is computationally expensive as it is done through Cross Validation [36].

2. METHODOLOGY

A systematic review of the literature must guarantee the integrity and quality of the available information related to the field of study. The methodology used for the review is described below:

2.1 Information sources

The search was carried out in six academic databases (Science Direct, Scopus, SAGE Journals, EBSCO, ResearchGate, Google Scholar) during the 2016-2021 period, with the last review date in July. The search equations were formed using combinations of terms such as: Academic Performance, Machine Learning, Ensemble Techniques, Ensemble Methods, Predicting, Bagging, Boosting, Stacking, Educational Data Mining.

145 articles were found, of which works related to: prediction of academic desertion, studies that are a bibliographic review of other works, studies applied in primary or secondary education, and studies that do not apply ensemble algorithms in the prediction of academic performance were excluded. As a result, 54 papers were obtained for review.

2.2 Objectives of the literature review

It is intended to identify the methods, tools, variables and findings in the application of Machine Learning, specifically those authors who use assembly methods in the prediction of academic performance. According to this objective, the following questions are raised: What are the variables that researchers take into account for the case studies? And of these, which ones have an impact on the prediction of student desertion? What is the application efficiency of ensemble methods for prediction in this field? What tools are the most used in investigations?

3. RESULTS

Based on the review of the literature, this section provides the results obtained to answer the research questions.

3.1 Factors that influence academic performance

Multiple authors have developed studies to determine the most influential variables in academic performance, and for this purpose they have applied different statistical and machine learning techniques. In the works consulted following the review methodology, various factors evaluated for the prediction of academic performance were found, which were grouped into: academic, sociodemographic, online learning, academic management, psychosocial and academic environment.

3.2 Academic factor

It refers to the characteristics related to the learning and teaching process of students in the educational field. Variables that allow the institution to qualify the students' learning level are found in this group, such as: cumulative average grades, subject grades, among others. Variables that measure the student's academic performance before entering university are: scores on state tests (taken in some countries), admission scores, high school grade point averages, among others. Table 1 shows the variables found that belong to the factor and the authors who used some of these for the development of their models.

Table 1. Variables academic factor and related authors.

Factor variables	Authors who consider variables of this factor
Admission score, Course scores, Course submitted, Courses scores, Cumulative GPA, English level, GPA of the course, High School GPA, High School location, Math level, Semester scores, State test score, Study time	[22], [23], [26], [28]–[31], [37]–[70]

Source: Own work.

3.3 Sociodemographic factor

It refers to the general characteristics and size of a population group. Among the variables included in this factor are aspects inherent to the student such as: age, gender, nationality, ethnicity, marital status, among others; and variables related to socioeconomic conditions such as: economic dependence, economic status of the parents, type of housing, level of education achieved by the parents, employment status, among others. Table 2 shows the variables found that belong to the factor and the authors who used some of these for the development of their models.

Table 2. Sociodemographic factor variables and related authors.

Factor variables	Authors who consider variables of this factor
Access to internet, Age, Economic dependence, Employment status, Ethnicity, Family Highest Education Level, Family income, Family size, Gender, Health status, House to university distance, Lives in town or village, Marital Status, Nationality, Parents job, Parents occupation, Parents qualification, Parent status, Total income	[17], [22], [23], [26], [28]–[31], [35], [38], [41]–[43], [46], [48]–[55], [57]–[60], [63]–[65], [67], [68], [70]–[77]

Source: Own work.

3.4 E-learning factor

The E-learning factor refers to the way in which the student acquires knowledge of a subject through digital pedagogical tools. This factor includes variables related to the interaction of students with the virtual educational platforms of the universities, such as: access to the platform, resources visited, Moodle questionnaires, among others. Table 3 shows the variables found that belong to the factor and the authors who used some of these for the development of their models.

Table 3. E-learning factor variables and related authors.

Factor variables	Authors who consider variables of this factor
Computational Knowledge, Content Read, Discussion groups, Forum viewed, Internet time, Internet usage activity, Moodle forum, Moodle quiz, Moodle task, Moodle time, Raises hand on class, Resource viewed, Visited resources	[28], [35], [45], [47], [49], [50], [55], [57], [59], [65], [70], [72]–[79]

Source: Own work.

3.5 Academic management factor

The academic management factor refers to institutional educational and pedagogical processes in order to respond to educational needs. This factor includes variables related to the evolution of the student during their university stage, such as: scholarships, credits taken, study plan, student status, among others. Table 4 shows the variables found that belong to the factor and the authors who used some of these for the development of their models.

Table 4. Academic management factor variables and related authors.

Factor variables	Authors who consider variables of this factor
Academic year, Admission category, Career, Course name, Credits in current term, Credits taken, Educational stage, Enrolment average grade, Enrolment stage, Failures, Program, Scholarship, Semester, Student Entry Age, Student status, Subject category, Topic, Transfer status, Type of course, Year of entry into university	[22], [23], [26], [28]–[30], [35], [38]–[40], [42]–[44], [46], [49]–[52], [55], [57], [58], [60], [61], [63]–[68], [70]–[72], [74]–[77], [79], [80]

Source: Own work.

3.6 Psychosocial factor

The psychosocial factor refers to human behavior and its insertion in society. This factor takes into account the variables that measure certain personality traits of students and may be linked to academic performance, such as: discipline, personality, attendance, habits, social integration, among others. Table 5 shows the variables found that belong to the factor and the authors who used some of these for the development of their models.

Table 5. Psychosocial factor variables and related authors.

Factor variables	Authors who consider variables of this factor
Absence rate, Adaptation, Aptitude, Attendance, Consumption of alcohol, Consumption of tobacco, Discipline, Focus in the class, Impact of friend circle, Institutional commitment, Interaction, Interest towards courses, Number of friends, Persistence, Personality, Social integration, Study method	[23], [28], [30], [35], [42]–[45], [48]–[55], [57]–[59], [63], [70], [72]–[79], [81]

Source: Own work.

3.7 Academic environment factor

The academic environment factor refers to the facilities, contexts and cultures in which students develop their knowledge acquisition process. Within this factor, aspects of the functioning of the institution where the student develops are taken into account, such as: study campus, professor's rank, subjects, study group, among others. Table 6 shows the variables found that belong to the factor and the authors who used some of these for the development of their models.

Table 6. Factor variables of academic environment and related authors.

Factor variables	Authors who consider variables of this factor
Assignments, Campus of study, Class, College/university integration, Course size, Extra paid classes, Extracurricular activities, Rank teacher, School's environment, Students in the course, Subjects, University support	[28], [30], [35], [42], [43], [48]–[50], [53]–[56], [58], [59], [64], [68], [81]

Source: Own work.

3.8 Ensemble methods used for predicting academic performance

Below is a detailed description of each of the works carried out during the last six years, consulted according to the aforementioned review methodology. The works are divided according to their year of publication; one group covers the period 2016 - 2018 while the other covers the period 2019 - 2021. In each one, the author(s), algorithms used, sample size, model and associated accuracy, and software used for data analysis and model development are shown.

The convention for the algorithms is shown below. ABO: AdaBoost, ANN: Artificial Neural Network, BA: Bagging, BAT: Bagged Trees, BN: Bayesian Network, BNB: Bernoulli Naive Bayes, BO: Boosting, CART: Classification and Regression Tree, CDB: Bayesian Discriminant Classifier, CL: Clustering , DISC: Discriminant Analysis, DT: Decision Tree, EPP: Ensemble-based Progressive Prediction, EXT: Extra Trees, GBT: Gradient Boosting, GLM: Generalized Linear Model, KNN: K Nearest Neighbors, LGBO: LogitBoost, LIR: Linear Regression, LL: Lazy Learning; LOR: Logistic Regression, MLP: Multilayer Perceptron, MR: Multiple Regression, NB: Naive Bayes, NN: Neural Network, PMLP: Personalized Multi-Linear Regression, PNN: Probabilistic Neural Network, PWC: Pairwise Coupling, RF: Random Forest, RT : Random Tree, RTF: Rotation Forest, SGD: Stochastic Gradient Descent, SMO: Sequential Minimal Optimization, STK: Stacking, SVM: Support Vector Machine, SVR: Support Vector Regression, VOT: Voting, XGB: XGBoost.

3.8.1 Period 2016 – 2018

19 of the 54 articles taken into account in the review fall within the period 2016-2018; these are summarized in Table 7.

Table 7. Results period 2016 - 2018.

Year	Author	Algorithms used	Sample size	Model Accuracy	Software used
2016	[79]	ANN, DT, NB, **BA, BO, ABO, RF**	500	**BO**, Accuracy 79.1%	WEKA
2016	[64]	SGD, KNN, PMLR, **RF**	33000	**RF**, RMSE 0.7381	Python
2016	[48]	KNN, GLM, NN, BN, **RF**	15519	**RF**, Accuracy 85.87%	-
2017	[17]	DT, ANN, KNN, NB, **RF**	210	NB, Accuracy 89.65%	RapidMiner
2017	[61]	DT, NB, KNN, SVM, **RF**	69	SVM, Accuracy 100%	R

(continúa)

(viene)

Year	Author	Algorithms used	Sample size	Model Accuracy	Software used
2017	[66]	**EPP**	367	**EPP**, Accuracy 75%	-
2018	[77]	**SVM, LOR, RF, VOT**	500	**VOT**, Accuracy 80.9%	-
2018	[76]	ID3, NB, KNN, SVM, **BA, BO, VOT**	500	**VOT**, Accuracy 89%	WEKA
2018	[62]	NB, KNN, DT, **BAG, VOT, STK**	1000	**BAT**, Accuracy 86.47%	WEKA
2018	[60]	DT, SVM, NB, **RF, BAT, ABO**	2495	**RF**, Accuracy 96.1%	RapidMiner
2018	[51]	J48, PART, BN, **RF**	300	**RF**, Accuracy 99%	WEKA
2018	[57]	C4.5, REPTree, KNN, NB, SMO, M5, **ABO, RTF, LGBO**	3882	Ensemble REPTree - M5 Rules, MAE 0.55	Java
2018	[47]	KNN, DT, **RF**	124	**RF**, Accuracy 71,57%	WEKA
2018	[35]	NB, ANN, DT, KNN, **BA, BO, ABO, RF**	480	**ABO** using ANN, Accuracy 78.6%	-
2018	[28]	ANN, SVM, DT, **STK**	141	**STK**, Precision 79.62%	SPSS, RapidMiner
2018	[43]	DT, LOR, SVM, **RF, EXT**	1077	**RF**, Accuracy 83%	Python
2018	[67]	SVM, **BO**	1304	**BO**, Accuracy 82.87%	-
2018	[59]	NB, LOR, NN, LMT, **RF, STK, XGB**	31029	**STK**, AUC 0.939	R
2018	[42]	J48, **RF, RT, STK**	28991	**STK**, Accuracy 96.11%	-

Source: Own work.

3.8.2 Period 2019 – 2021

35 of the 54 articles taken into account in the review fall within the period 2019-2021; these are summarized in Table 8. The referential research was carried out until December 2021.

Table 8. Results period 2019 - 2021.

Year	Author	Algorithms used	Sample size	Model Accuracy	Software used
2019	[71]	NB, DT, KNN, DISC, **BO, ABO**	500	**ABO** using KNN, Accuracy 86.01%	-
2019	[74]	NB, C5.0, CART, KNN, SGD, **RF**	480	**RF**, Accuracy 0.7959	R
2019	[69]	NN, **GBT, STK**	500	**STK**, MAE 3.0856	-
2019	[55]	NB, DT, SVM, NN, CL, **GBT, RF**	588	Ensemble Model, Accuracy 98.5%	RapidMiner
2019	[26]	LOR, **RF**	11637	**RF**, Precision 84%	Python

(continúa)

(viene)

Year	Author	Algorithms used	Sample size	Model Accuracy	Software used
2019	[31]	CART, CDB, GLM, LIR, SVM, **RF, VOT**	2029	GLM, AUROC 0.704	SQL, R
2019	[39]	DT, NB, LOR, MLP, NN, **RF, ABO**	1841	NN, Accuracy 51.9%	KNIME
2019	[78]	SVM, KNN, DT, **BA, BO, RF**	500	**RF**, Accuracy 0.89	-
2019	[81]	PART, FURIA, NNge, OneR, NB, MLP, SMO, LL, KNN, J48, LMT, REPTree, CART, RT, DT, **RF, BO, BA, VOT, ABO**	400	**ABO** using NB - J48, Accuracy 0.985	-
2019	[22]	ANN, **GBT, XGB, STK**	9118	**STK**, Accuracy 0.94	Excel, R, Python
2019	[23]	J48, KNN, LOR, MLP, **RF**	12698	**RF**, Accuracy 69.35%	WEKA
2019	[73]	ANN, LOR, NB, SVM, DT, **RF, BA. VOT, XGB**	480	**RF**, Accuracy 77.08%	-
2019	[52]	J48, NNge, MLP, **ABO**	1044	**ABO** using J48, Accuracy 95.78%	-
2019	[50]	DT, ANN, SVM, **RF, BA, BO, ABO, STK, VOT**	4413	**VOT**, Accuracy 75.56%	Python, MySQL
2019	[40]	PNN, NB, LOR, DT, **RF, BAT**	1841	LOR, Accuracy 89.15%	KNIME, MATLAB
2020	[72]	NB, DT, KNN, DISC, PWC, **ABO, BA, BO**	500	**ABO**, Accuracy 86%	MATLAB
2020	[38]	SVM, LOR, DT, BNB, **RF, ABO, GBT, XGB**	486	**STK**, AUROC 0.9138	Python
2020	[41]	J48, NB, KNN, **BO**	-	NB, Accuracy 97.15%	GAMS
2020	[29]	DT, NB, LOR, MLP, NN, **BA**	129	**BA** using LOR, Accuracy 86.82%	WEKA
2020	[37]	LIR, DT, KNN, **RF, BA, STK**	2200	**STK** using LIR - DT, MAE 7.5989	WEKA
2020	[75]	**RF, GBT**	480	**RF**, Accuracy 81%	Python
2020	[56]	DT, KNN, **BA, BO, RF, STK**	121	**STK**, Accuracy 0.78	-
2020	[53]	SVM, RBF, LOR, NB, KNN, NN, **RF, BA**	601	**BA** using LOR, Accuracy 93.1%	R, MATLAB
2020	[54]	NB, ANN, DT, LOR, **BA, LGBO, ABO, RF**	887	**RF**, Accuracy 96%	WEKA
2020	[63]	NB, MLP, KNN, DT, **BA, RTF**	319	Ensemble **RTF -** MLP, Accuracy 91.70%	MATLAB
2020	[49]	NN, **RF, GBT**	450	NN, Accuracy 0.782	RapidMiner
2020	[80]	SVM, MLP, KNN NB, LOR, **RF**	601	Ensemble using **RF -** KNN - SVM, Accuracy 0.35	R
2021	[46]	SVM, DT, C5.0, NN, **BO, XGB, RF**	2761	**BO** using C5.0, Accuracy 0.9312	R
2021	[65]	DT, LOR, NN, SVM, KNN, **RF, GBT**	1708	DT, Accuracy 0.942	-

(continúa)

(viene)

Year	Author	Algorithms used	Sample size	Model Accuracy	Software used
2021	[58]	NB, MLP, DT, J48, **BA, ABO**	480	**ABO** using MLP, Accuracy 80.33%	WEKA
2021	[70]	DT, NB, MLP, **BA, BO, VOT**	480	**BA** using DT, Accuracy 91.39%	-
2021	[68]	ANN, KNN, CL, NB, SVM, LOR, DT, **VOT**	1491	**VOT**, Accuracy 83%	WEKA
2021	[30]	MLP, **BA, BO**	649	**BA**, Accuracy 88%	WEKA
2021	[44]	NB, DT, LOR, **RF**	11312	**RF**, Accuracy 97%	Python
2021	[45]	J48, REPTree, JRip, Nnge, PART, **VOT**	57	**VOT**, Accuracy 87.47%	Excel, WEKA

Source: Own work.

4. DISCUSSION AND CONCLUSIONS

Below is a brief discussion about the main aspects found in the literature review.

4.1 As for the factors

As shown in Figure 2, the most influential variables in academic performance are those related to academic, academic management and sociodemographic factors, since the authors used them to develop their models 43, 43 and 42 times, respectively. The most used variables within these factors were cumulative grade point average, gender, course scores, semester, and age. We found 31 authors who used psychosocial factor variables for the development of their models, among which assistance and satisfaction stand out. Finally, it can be seen that the authors used variables of online learning and academic environment factors in 21 and 17 articles, respectively, finding that the most used variables within these factors were activity in discussion groups, resources visited, hands raised in class, extracurricular activities and subjects.

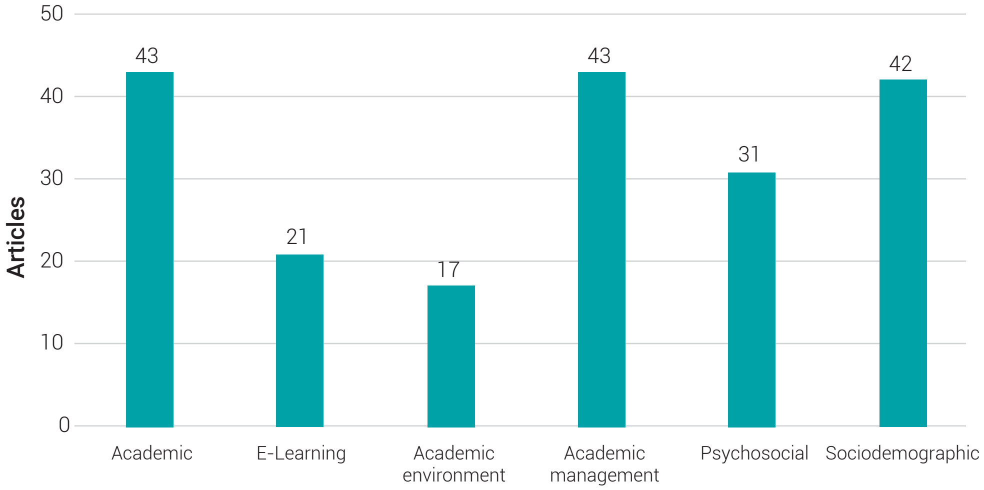

Figure 2. Articles related to each factor.
Source: Own work.

It is worth mentioning some documents that have been cited in different works that support the variables grouped in the previous factors. Such is the case of Landa [82], who indicates that the most influential variables within his model are related to the academic environment factor; highlighting pedagogy, class schedules and the relationship between the student and the teacher. In the study of Kamal [55], it is evident that there are variables that positively affect performance, such as: the percentage of the appropriate course, high test scores and the location of residence near the university; and others that negatively affect you, such as: alcohol and/or tobacco use, health issues and family difficulties.

Regarding the academic factor, Asif [17] shows that it is possible to predict the graduation performance in a four-year program from variables related to the school and with the grades of the first and second year of university. Miguéis [60] shows that the most influential variable in their study for predicting academic performance was the average enrollment grade. Hussain [51] applies various variable selection techniques, and these conclude that the grades obtained and the percentage achieved in the class are the most relevant variables in student performance. Lauria [59] seeks to predict academic performance using different datasets, and concludes that models that include academic variables, especially course grades, present better results in their evaluation metrics. Jayaprakash [54] classifies the variables according to their importance, stating that the grades obtained in the courses are the most relevant within his entire data set.

Regarding the academic management factor, Adekitan [39] applies statistical techniques to determine the relationship between academic variables of admission to exams and academic performance in their first year, which results in a weak relationship between these types of variables. The study concludes that admission requirements are vital to admit students with a variety of knowledge and unique qualities, since performance cannot be predicted solely with variables of the academic management factor. On the other hand, Bucos [43] tries to predict the academic performance of students in different weeks of the object-oriented programming course, and through the statistical test of chi square affirms that there is a relationship between some psychosocial (attendance), academic management (credits earned the last year) and academic variables (note of the cuts). Orihuela [26], in his thesis identifies that the most influential variables are related to academic and academic management factors, among which are the teachers, the semester, the student's condition, the year and course credits. Hassan [50] obtains the best results by including online learning, academic and academic management variables, highlighting the importance of the courses viewed variable.

In relation to the sociodemographic factor, Yamao [67] seeks to predict the performance of students at the end of their first year of university and, through statistical techniques, manages to determine that the most influential variables in performance are gender, age, type of income and distance from their home to the study center. On the other hand, Brohi [74] shows that the most influential variables of the sociodemographic factor are those related to the relative responsible for the student and the nationality. Jayaprakash [54] classifies the variables according to their importance, and finds that the most relevant variables related to this factor are parental status, gender, family size and parental education. Jawthari [75] evaluates the importance of the variables, and determines that the most influential and related to this factor are nationality and gender.

Regarding the psychosocial factor, Rahman [35] finds that these variables improve the performance of predictive models, especially the variable related to attendance (which is stated by Kostokopoulos [57] in his study, in which there is a progressive addition of variables to the training data to determine the level of improvement in the accuracy of their models), and indicates that the most relevant variables are attendance at face-to-face activities and delivery of written tasks. Brohi [74] and Jawthari [75] also show that the most influential variable in academic performance within the psychosocial factor is the one related to attendance.

Regarding the online learning factor, Brohi [74] states that including more variables does not mean better results in the models, since the accuracy achieved with 11 variables is higher than the models that add 10 or 12 variables. In addition, it indicates that the online learning variables that are related to the resources visited and the hands raised in class are the most influential in the performance of the students. Amrieh [79] finds that the psychosocial and online learning variables, especially resources visited, are the most decisive in their predictive model, since it manages to increase the accuracy by almost six percent, which is also concluded by Adejo [28], since it shows that including online learning variables and resources visited in the training data of their models improves their performance. Campo [47] demonstrates the usefulness of data mining in the educational field, since it manages to predict the final grade of the course with the activity in Moodle and the grades achieved in the intermediate controls. In addition, he highlights the importance of other variables, such as the average number of hours of study and support from the university. Kumari [76] says in his work that the rise in the use of technology in education makes it necessary to take into account the variables of online learning, and shows that these variables help to classify students in a better way. Trakunphutthirak [65] states that the semester and the use of the internet significantly influence performance.

4.2 As for the algorithms

Machine learning algorithms have turned out to be useful and accurate in predicting academic performance, most authors share the idea that ensemble algorithms present better results than supervised and unsupervised learning algorithms and neural networks if applied in isolation. This can be evidenced in the "model precision" column of Tables 7 and 8, where the models with the best performance developed by the authors are shown; in 46 of 54 articles the best result was obtained by a model developed with assembly algorithms. Such is the case of Almasri [81], who develops algorithms based on rules, Bayes theorum, functions, lazy learning and assembly, and it is the latter that obtains the best results in terms of the evaluation metrics established in the work.

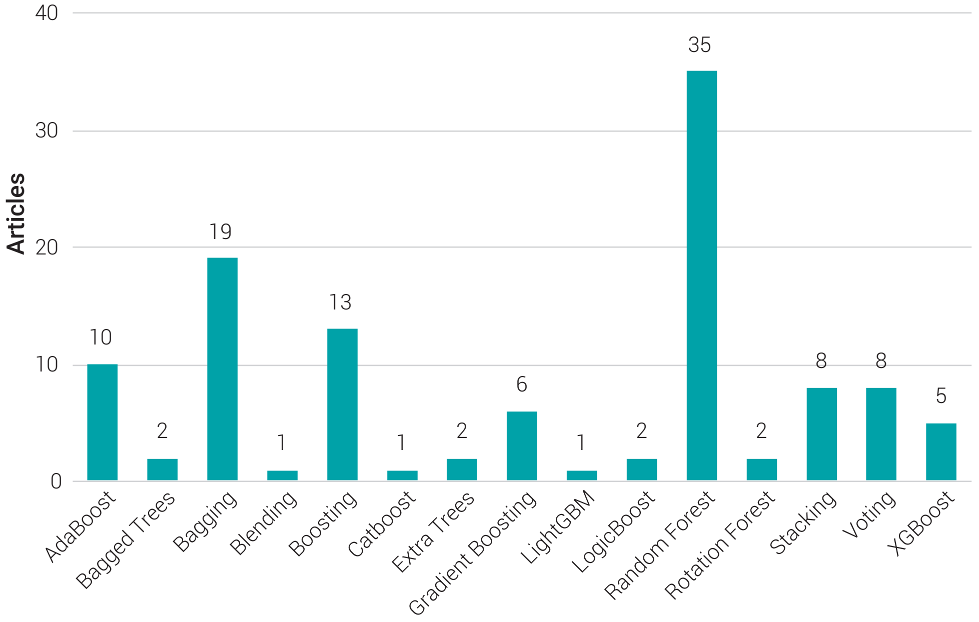

Figure 3. Articles related to assembly algorithms.
Source: Own work.

Figure 3 shows that ensemble algorithms have been widely used to predict academic performance, with Random Forest being the most popular among the works consulted, followed by Bagging and Boosting. However, more recent algorithms are being applied, as is the case of Kostopoulos [57], which develops the so-called Rotation Forest and LogitBoost, obtaining these two better results than the classic algorithms and the other applied ensemble algorithms. Sakri [63] also uses Rotation Forest for the development of models, and obtains the highest accuracy within the algorithms applied in isolation. On the other hand, Jayaprakash [54] uses LogitBoost, with which it obtains the highest accuracy within the algorithms developed for predictions. Ajibade [71] also proposes a model which he calls ADDE, which is based on a combination of the AdaBoost M2 algorithm with an optimization metaheuristic called differential evolution.

4.3 As for the approach

The prediction of academic performance can be carried out at different stages of the students' university life. For example, Yamao [67] aims to predict the performance of students at the end of their first year of university, unlike Asif [17], who seeks to predict the academic performance of students at the end of a four-year university program.

Some of the approaches the authors took in developing their predictive models are briefly described below.

Kostopoulos [57] develops a model that seeks to predict the academic performance of a student in the "introduction to computer science" module. Bucos [43] seeks to predict the performance of students in the course "Object Oriented Programming", a subject taught in the second year of university, and Campo [47] seeks to predict the performance of students in the subject "automata theory and formal languages" through the application of intermediate controls. Pandey [62] and Injadat [53] also develop predictions at different stages in a course, usually in academic cuts, and these are called multilevel predictions.

Sweeney [64] develops a system that aims to predict the grades of the students in the courses they enroll in the following semester, something similar to what Adekitan [39] proposes, which makes predictions of the academic performance of the students in the 13 courses of the basic study program of the university.

Ochoa [61] predicts, through a machine learning model, the academic performance of students at the end of the semester, as well as Adekitan [39], Candia [23] and Zeineddine [68], since they develop their models seeking to predict the academic performance of newly admitted students at the end of their first year of college. Miguéis [60] seeks to predict the final general average of the students through the information of their first-year university grades.

Adekitan [40] seeks to predict the academic performance and the general average of students at the end of the fifth year of university, using the information available from the first three years. Trakunphutthirak [65] evaluates different characteristics and their effect on academic performance in the first ten weeks of the semester.

4.4 Conclusions

After the detailed review and referential analysis of the algorithms developed, it is possible to affirm that the ensemble methods are a more effective tool than the classical algorithms alone for the prediction of academic performance.

The factors with the greatest impact on the prediction of academic performance, according to the review of each of the works, are related to: academic, academic management and sociodemographic factors; however, the authors who included psychosocial type variables presented more precise results in their evaluation metrics. In addition, the application of variable selection techniques allows the models to obtain better performance, since information that can generate noise in the training process is omitted.

The most common approach, in the models developed by the authors, seeks to predict academic performance in first-year university students, since several agree that the early identification of students at possible risk of falling into low academic performance allows teachers of educational institutions to design strategies to mitigate this problem.

Currently, most educational institutions do not take full advantage of digital resources, such as learning management platforms, as these can become automated student data acquisition systems. This can translate into data with less bias and fewer resources in the process of cleaning and conditioning the information. The data that can be obtained from these platforms is related to variables such as access and duration in e-learning systems, activities in discussion forums, resources visited, virtual exam notes, among others.

It was evidenced that only a few studies proposed corrective solutions to provide timely feedback to students and educators to address situations of poor academic performance; they were limited to defining the impact variables, predicting performance and calculating the effectiveness of what was proposed.

5. REFERENCES

[1] O. D. Castrillón, W. Sarache, and S. Ruiz-Herrera, "Predicción del rendimiento académico por medio de técnicas de inteligencia artificial," *Formación universitaria*, vol. 13, pp. 93–102, 2020. doi: http://doi.org/10.4067/S0718-50062020000100093.

[2] M. Rodríguez Urrego, "La investigación sobre deserción universitaria en Colombia 2006-2016. Tendencias y resultados," *Pedagogía y Saberes*, vol. 51, pp. 49–66, 2019. doi: 10.17227/pys.num51-8664.

[3] L. E. Contreras, H. J. Fuentes, and J. I. Rodríguez, "Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático," *Formación universitaria*, vol. 13, no. 5, pp. 233–246, 2020. doi: http://doi.org/10.4067/S0718-50062020000500233.

[4] E. Porcel, G. N. Dapozo, and M. V. López, "Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios," in *XI Workshop de Investigadores en Ciencias de la Computación*, 2009, pp. 635–639.

[5] D. L. la Red Martínez, M. J. Karanik, M. E. Giovannini, and N. Pinto, "Perfiles de rendimiento académico: un modelo basado en minería de datos," *Campus virtuales: revista científica iberoamericana de tecnología educativa.*, vol. IV, pp. 12–30, 2015.

[6] K. Eckert and R. Suénaga, “Aplicación de técnicas de minería de datos al análisis de situación y comportamiento académico de alumnos de la UGD,” in *XV Workshop de Investigadores en Ciencias de la Computación*, 2013, pp. 92–96.

[7] C. Marqués, “Predicción del fracaso y el abandono escolar mediante técnicas de minería de datos,” Universidad de Córdoba, 2015, pp. 1 - 92.

[8] Y. Zhang, S. Oussena, T. Clark, and H. Kim, “Use Data Mining to Improve Student Retention in Higher Education - A Case Study.,” in *Proceedings of the 12th International Conference on Enterprise Information Systems*, Jan. 2010, pp. 190–197. doi: 10.5220/0002894101900197.

[9] E. Castaño, S. Gallón, K. Gómez, and J. Vásquez, “Análisis de los factores asociados a la deserción y graduación estudiantil universitaria,” *Lecturas de economía*, vol. 65, pp. 11–35, 2006.

[10] G. Psacharopoulos and H. A. Patrinos, “Returns to investment in education: a further update,” *Education Economics*, vol. 12, no. 2, pp. 111–134, 2004. doi: http://doi.org/10.1080/0964529042000239140.

[11] N. Bedregal-Alpaca, D. Aruquipa-Velazco, and V. Cornejo-Aparicio, “Técnicas de data mining para extraer perfiles comportamiento académico y predecir la deserción universitaria,” *Revista Ibérica de Sistemas e Tecnologias de Informaçao*, no. E27, pp. 592–604, 2020.

[12] J. I. R. Molano, L. D. F. Zea, and Y. F. P. Reina, “Proposal of Architecture and Application of Machine Learning (Ml) as A Strategy for the Reduction of University Desertion Levels Due to Academic Factors,” *Ingeniería Solidaria*, vol. 15, no. 3, pp. 1–23, 2019. doi: https://doi.org/10.16925/2357-6014.2019.03.06.

[13] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student’performance prediction using machine learning techniques,” *Education Sciences*, vol. 11, no. 9, p. 552, 2021. doi: https://doi.org/10.3390/educsci11090552.

[14] F. Alshareef, H. Alhakami, T. Alsubait, and A. Baz, “Educational Data Mining Applications and Techniques,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 4, pp. 729–734, 2020.

[15] A. Rico Páez, N. D. Gaytán Ramírez, and D. Sánchez Guzmán, “Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes,” *Diálogos sobre educación. Temas actuales en investigación educativa*, vol. 10, no. 19, pp. 1–18, 2019. doi: https://doi.org/10.32870/dse.v0i19.509.

[16] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011, pp. 327 – 439.

[17] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, pp. 177–194, 2017. doi: https://doi.org/10.1016/j.compedu.2017.05.007.

[18] R. S. J. D. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–16, 2009. doi: https://doi.org/10.5281/zenodo.3554657.

[19] A. L. Dyckhoff, D. Zielke, M. Bültmann, M. A. Chatti, and U. Schroeder, "Design and implementation of a learning analytics toolkit for teachers," *Journal of Educational Technology & Society*, vol. 15, no. 3, pp. 58–76, 2012.

[20] A. Salcedo, "Desertion in Colombian Universities," *Revista Academia y Virtualidad*, vol. 3, no. 1, pp. 50–60, 2010.

[21] J. G. López Martínez and Ó. A. Méndez Aguirre, "Técnicas de Machine Learning para la predicción de desempeño académico en el desarrollo del espacio proyectivo del pensamiento espacial," Universidad Pedagógica Nacional, 2019, pp. 82 – 127.

[22] 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," Universidad Ricardo Palma, 2019, pp. 1 – 147.

[23] D. I. Candia Oviedo, "Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático," Universidad Nacional de San Antonio Abad del Cusco, 2019, pp. 1 – 129.

[24] R. J. Rojas Pari, "Modelo de Aprendizaje Automático Supervisado para Identificar Patrones de Bajo Rendimiento Académico en los Ingresantes al Instituto de Educación Superior Pedagógico Público–Juliaca," Universidad Peruana Unión, 2021, pp. 19 – 96.

[25] M. Bourel, "Model aggregation methods and applications," *Memoria de Trabajos de Difusión Científica y Técnica*, no. 10, pp. 19–32, 2012.

[26] G. Y. Orihuela Maita, "Aplicación de Data Science para la Predicción del Rendimiento Académico de los Estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú," Universidad Nacional del Centro del Perú, 2019, pp. 1 – 83.

[27] F. F. Patacsil, “Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models,” *Universal Journal of Educational Research*, vol. 8, no. 9, pp. 4036–4047, 2020. doi: http://doi.org/10.13189/ujer.2020.080929.

[28] O. W. Adejo and T. Connolly, “Predicting student academic performance using multi-model heterogeneous ensemble approach,” *Journal of Applied Research in Higher Education*, vol. 10, no. 1, pp. 61–75, 2018. doi: 10.1108/JARHE-09-2017-0113.

[29] 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,” *International Journal of Information and Education Technology*, vol. 10, no. 10, pp. 723–727, 2020. doi: 10.18178/ijiet.2020.10.10.1449.

[30] J. Malini and Y. Kalpana, “Investigation of factors affecting student performance evaluation using education materials data mining technique,” *Materials Today: Proceedings*, vol. 47, pp. 6105–6110, 2021. doi: https://doi.org/10.1016/j.matpr.2021.05.026.

[31] D. E. Alessandrini López, “Aprendizaje estadístico en educación: Una propuesta de modelización para carreras de grado en Ingeniería.,” Universidad de la República, 2019, pp. 1 – 95.

[32] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996. doi: https://doi.org/10.1007/bf00058655.

[33] J. J. Campo Yepes and D. L. Cruz Castro, “Modelos Apilados y factores que pueden afectar la eficiencia,” Universidad Santo Tomás, 2017, pp. 1 – 12.

[34] M. H. D. M. Ribeiro and L. dos Santos Coelho, “Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series,” *Applied Soft Computing*, vol. 86, pp. 1–30, 2020. doi: https://doi.org/10.1016/j.asoc.2019.105837.

[35] M. Hasibur Rahman and M. Rabiul Islam, “Predict Student’s Academic Performance and Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques,” *2nd International Conference on Electrical and Electronic Engineering, ICEEE 2017*, no. September 2020, pp. 1–4, 2018. doi: 10.1109/CEEE.2017.8412892.

[36] “Ensemble methods — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/ensemble.html#bagging-meta-estimator

[37] E. J. Phua and N. K. Batcha, “Comparative analysis of ensemble algorithms’prediction accuracies in education data mining,” *Journal of Critical Reviews*, vol. 7, no. 3, pp. 37–40, 2020. doi: 10.31838/jcr.07.03.06.

[38] L. Yan and Y. Liu, "An ensemble prediction model for potential student recommendation using machine learning," *Symmetry*, vol. 12, no. 5, p. 728, 2020. doi: https://doi.org/10.3390/SYM12050728.

[39] 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," *Education and Information Technologies*, vol. 24, no. 2, pp. 1527–1543, 2019. doi: 10.1007/s10639-018-9839-7.

[40] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2 e01250, pp. 1–16, 2019. doi: 10.1016/j.heliyon.2019.e01250.

[41] M. Ashraf, M. Zaman, and M. Ahmed, "An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches," *Procedia Computer Science*, vol. 167, no. 2019, pp. 1471–1483, 2020. doi: 10.1016/j.procs.2020.03.358.

[42] M. Ashraf, M. Zaman, and M. Ahmed, "Using Ensemble StackingC Method and Base Classifiers to Ameliorate Prediction Accuracy of Pedagogical Data," *Procedia Computer Science*, vol. 132, no. Iccids, pp. 1021–1040, 2018. doi: 10.1016/j.procs.2018.05.018.

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

[44] D. L. Bustamante Peña, "Modelo predictivo de rendimiento académico para el apoyo, prevención y disminución de la tasa de deserción universitaria," Universidad de Bogotá Jorge Tadeo Lozano, 2021, pp. 1 – 40.

[45] W. Chango, R. Cerezo, and C. Romero, "Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses," *Computers and Electrical Engineering*, vol. 89, no. November 2020, pp. 1–11, 2021. doi: 10.1016/j.compeleceng.2020.106908.

[46] J. E. Chaparro Mesa and J. Cuatindioy Imbachi, "Análisis comparativo de técnicas de clasificación para determinar la deserción estudiantil de la facultad de ingeniería de la Universidad de Antioquia, Colombia," *Espacios*, vol. 42, no. 7, pp. 63–81, 2021. doi: http://doi.org/10.48082/espacios-a21v42n07p05.

[47] D. Campo-Ávila, G. P. Ramos-Jimenez, R. Morales-Bueno, and M. Baena-García, "Minería de datos educativos para la predicción personalizada del rendimiento académico," in *Conferencia Internacional de Procesamiento de la Informacion*, 2018, pp. 1–10.

[48] H. Dissanayake, D. Robinson, and O. Al-Azzam, “Predictive modeling for student retention at St. Cloud state university,” in *Proceedings of the International Conference on Data Science (ICDATA)*, 2016, pp. 215–221.

[49] S. F. Aziz, “Students’ Performance Evaluation Using Machine Learning Algorithms.,” *College of Basic Education Researchers Journal*, vol. 16, no. 3, pp. 976–986, Jul. 2020.

[50] H. Hassan, S. Anuar, and N. B. Ahmad, “Students’ performance prediction model using meta-classifier approach,” *Communications in Computer and Information Science*, pp. 221–231, 2019, doi: 10.1007/978-3-030-20257-6_19.

[51] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, “Educational data mining and analysis of students’ academic performance using WEKA,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, pp. 447–459, 2018. doi: 10.11591/ijeecs.v9.i2.pp447-459.

[52] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, “Student academic performance prediction using supervised learning techniques,” *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, pp. 92–104, 2019. doi: 10.3991/ijet.v14i14.10310.

[53] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, “Multi-split optimized bagging ensemble model selection for multi-class educational data mining,” *Applied Intelligence*, vol. 50, no. 12, pp. 4506–4528, 2020. doi: 10.1007/s10489-020-01776-3.

[54] S. Jayaprakash, S. Krishnan, and J. Jaiganesh, “Predicting Students Academic Performance using an Improved Random Forest Classifier,” in *2nd IEEE International Conference on Emerging Smart Computing and Informatics, ESCI 2020*, 2020, pp. 238–243. doi: 10.1109/ESCI48226.2020.9167547.

[55] P. Kamal and S. Ahuja, “An ensemble-based model for prediction of academic performance of students in undergrad professional course,” *Journal of Engineering, Design and Technology*, vol. 17, no. 4, pp. 769–781, 2019. doi: 10.1108/JEDT-11-2018-0204.

[56] S. Kausar *et al.*, “Mining Smart Learning Analytics Data Using Ensemble Classifiers.,” *International Journal of Emerging Technologies in Learning*, vol. 15, no. 12, pp. 81–102, Dec. 2020.

[57] G. Kostopoulos, S. Kotsiantis, C. Pierrakeas, G. Koutsonikos, and G. A. Gravvanis, “Forecasting students’ success in an open university,” *International Journal of Learning Technology*, vol. 13, no. 1, pp. 26–43, 2018. doi: 10.1504/IJLT.2018.091630.

[58] M. Kumar, G. Mehta, N. Nayar, and A. Sharma, “EMT: Ensemble meta-based tree model for predicting student performance in academics,” *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, pp. 0–10, 2021. doi: 10.1088/1757-899X/1022/1/012062.

[59] E. J. M. Lauría, E. Presutti, M. Kapogiannis, and A. Kamath, “Stacking classifiers for early detection of students at risk,” *CSEDU 2018 - Proceedings of the 10th International Conference on Computer Supported Education*, vol. 1, no. Csedu 2018, pp. 390–397, 2018. doi: 10.5220/0006781203900397.

[60] V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, “Early segmentation of students according to their academic performance: A predictive modelling approach,” *Decision Support Systems*, vol. 115, no. September, pp. 36–51, 2018. doi: 10.1016/j.dss.2018.09.001.

[61] L. L. Ochoa, K. Rosas Paredes, and C. Baluarte Araya, “Evaluación de técnicas de minería de datos para la predicción del rendimiento académico,” *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, vol. 2017-July, no. January, 2017, pp. 1 - 8. doi: 10.18687/LACCEI2017.1.1.368.

[62] M. Pandey and S. Taruna, “An ensemble-based decision support system for the students' academic performance prediction,” in *Advances in Intelligent Systems and Computing*, 2018, pp. 163–169. doi: https://doi.org/10.1007/978-981-10-6602-3_16.

[63] S. Sakri and A. S. Alluhaidan, “RHEM: A Robust Hybrid Ensemble Model for Students' Performance Assessment on Cloud Computing Course,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, pp. 388–396, 2020. doi: 10.14569/IJACSA.2020.0111150.

[64] M. Sweeney, H. Rangwala, J. Lester, and A. Johri, “Next-Term Student Performance Prediction: A Recommender Systems Approach,” pp. 1–27, 2016. doi: 10.5281/zenodo.3554603.

[65] R. Trakunphutthirak and V. C. S. Lee, “Application of Educational Data Mining Approach for Student Academic Performance Prediction Using Progressive Temporal Data,” *Journal of Educational Computing Research*, pp. 1–29, 2021. doi: 10.1177/07356331211048777.

[66] J. Xu, Y. Han, D. Marcu, and M. Van Der Schaar, “Progressive prediction of student performance in college programs,” in *31st AAAI Conference on Artificial Intelligence*, 2017, pp. 1604–1610.

[67] E. Yamao, L. C. Saavedra, R. Campos Pérez, and V. de J. Huancas Hurtado, “Prediction of academic performance using data mining in first year students of peruvian university,” *Revista USMP - Campus*, vol. 23, no. 26, pp. 151–160, 2018. doi: https://doi.org/10.24265/campus.2018.v23n26.05.

[68] H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” *Computers and Electrical Engineering*, vol. 89, pp. 1–9, 2021. doi: 10.1016/j.compeleceng.2020.106903.

[69] S. Zhang, M. Liu, and J. Zhang, “An Academic Achievement Prediction Model Enhanced by Stacking Network,” in *International Forum on Digital TV and Wireless Multimedia Communications*, 2019, pp. 235–245. doi: https://doi.org/10.1007/978-981-15-3341-9_20.

[70] 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,” *Wireless Communications and Mobile Computing*, pp. 1–8, 2021. doi: 10.1155/2021/6241676.

[71] S. S. M. Ajibade, N. B. Binti Ahmad, and S. M. Shamsuddin, “A Novel Hybrid Approach Of Adaboostm2 Algorithm And Differential Evolution For Prediction Of Student Performance,” *International Journal of Scientific and Technology Research*, vol. 8, no. 7, pp. 65–70, 2019.

[72] S.-S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, “A data mining approach to predict academic performance of students using ensemble techniques,” *Joint Conferences on 18th International Conference on Intelligent Systems Design and Applications, ISDA 2018 and 10th World Congress on Nature and Biologically Inspired Computing , NaBIC 2018*, vol. 940. Springer Verlag, Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia, pp. 749–760, 2020. doi: 10.1007/978-3-030-16657-1_70.

[73] S. Almutairi, H. Shaiba, and M. Bezbradica, “Predicting Students’ Academic Performance and Main Behavioral Features Using Data Mining Techniques,” *1st International Conference on Intelligent Cloud Computing, ICC 2019*, vol. 1097 CCIS. Springer, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, pp. 245–259, 2019. doi: 10.1007/978-3-030-36365-9_21.

[74] S. N. Brohi, T. R. Pillai, S. Kaur, H. Kaur, S. Sukumaran, and D. Asirvatham, “Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education,” *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 285, no. August, pp. 254–261, 2019. doi: 10.1007/978-3-030-23943-5_19.

[75] M. Jawthari and V. Stoffova, “Effect of encoding categorical data on student’s academic performance using data mining methods.,” *eLearning & Software for Education*, vol. 1, pp. 521–526, Jan. 2020.

[76] P. Kumari, P. K. Jain, and R. Pamula, “An efficient use of ensemble methods to predict students academic performance,” in *4th IEEE International Conference on Recent Advances in Information Technology, RAIT 2018*, 2018, pp. 1–6. doi: 10.1109/RAIT.2018.8389056.

[77] A. Salini, U. Jeyapriya, S. M. College, and S. M. College, "A Majority Vote Based Ensemble Classifier for Predicting Students Academic Performance," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 24, pp. 1–11, 2018.

[78] S. S. M. Ajibade, N. Bahiah Binti Ahmad, and S. Mariyam Shamsuddin, "Educational Data Mining: Enhancement of Student Performance model using Ensemble Methods," *IOP Conference Series: Materials Science and Engineering*, vol. 551, no. 1, pp. 1–5, 2019. doi: 10.1088/1757-899X/551/1/012061.

[79] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *International journal of database theory and application*, vol. 9, no. 8, pp. 119–136, 2016. doi: https://doi.org/10.1007/978-3-319-21024-7_28.

[80] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Systems*, pp. 1–15, 2020. doi: 10.1016/j.knosys.2020.105992.

[81] A. Almasri, E. Celebi, and R. S. Alkhawaldeh, "EMT: Ensemble meta-based tree model for predicting student performance," *Scientific Programming*, pp. 1–12, 2019. doi: 10.1155/2019/3610248.

[82] B. D. Landa, R. M. Romero, and W. J. M. Rodriguez, "Rendimiento académico de estudiantes en Educación Superior: predicciones de factores influyentes a partir de árboles de decisión," *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales*, vol. 23, no. 3, pp. 616–639, 2021. doi: https://doi.org/10.36390/telos233.08.