



Logistic regression and learning personalization in higher education

Regresión logística y personalización del aprendizaje en la educación superior

**Karla Karina Ruiz Mendoza**  
<https://orcid.org/0000-0001-8978-8364>   
Universidad Autónoma de Baja California, Mexico  
[ruiz.karla32@uabc.edu.mx](mailto:ruiz.karla32@uabc.edu.mx) (correspondence)

**Luis Horacio Pedroza Zúñiga**  
<https://orcid.org/0000-0002-5256-2967>   
Universidad Autónoma de Baja California, Mexico

ABSTRACT

This study focuses on the early prediction of academic performance in higher education students to personalize the learning process and enable timely interventions. Logistic Regression, widely used for its interpretability and effectiveness, serves as a starting point to assess its validity and utility in the context of higher education. A dataset of 10,184 students from the Universidad Autónoma de Baja California was analyzed. Three variable configurations (Basic, Complete, and Exam) and three classification algorithms (Logistic Regression, Naive Bayes, and Decision Tree) were compared using five-fold cross-validation and random sampling (90% training, 10% testing). Accuracy, Recall, F1 Score, and AUC-ROC were employed as evaluation metrics. Logistic Regression (Basic configuration) achieved the best metrics, yielding a Recall near 0.88 and an AUC-ROC around 0.72–0.76, outperforming Naive Bayes and Decision Tree. High school GPA emerged as the most influential variable, followed by Writing scores. These findings highlight the potential of Logistic Regression for early risk detection and learning personalization, although further investigation is warranted to address predictive fairness and incorporate socio-emotional factors that ensure a more inclusive and effective educational approach.

**Keywords:** logistic regression, extrapolation, personalized learning, higher education.

RESUMEN

Este estudio aborda la predicción temprana del rendimiento académico (a través de examen y promedio de bachillerato) de estudiantes universitarios con el fin de personalizar el proceso educativo y facilitar intervenciones oportunas. La regresión logística sirve como punto de partida para analizar su validez (según el Enfoque Basado en Argumentos) y utilidad en educación superior. Se trabajó con una muestra de 10,184 estudiantes de la Universidad Autónoma de Baja California. Se aplicaron tres configuraciones de variables (Básico, Completo y Examen) y tres algoritmos de clasificación (Regresión Logística, Naive Bayes y Árbol de Decisión), evaluados mediante validación cruzada (cinco pliegues) y muestreo aleatorio (90% de entrenamiento, 10% de prueba). La Regresión Logística (Básico) mostró los mejores indicadores, con un Recall cercano a 0.88 y un AUC-ROC cercano a 0.72–0.76, superando a Naive Bayes y Árbol de Decisión. El promedio de bachillerato emergió como la variable más influyente. Estos hallazgos refuerzan la pertinencia de la regresión logística como herramienta para la detección temprana de riesgo académico y la personalización educativa, aunque se requiere profundizar en la equidad de sus predicciones y en la incorporación de factores socioemocionales que garanticen una educación inclusiva y eficaz.

**Palabras clave:** regresión logística, extrapolación, personalización del aprendizaje, educación superior.

ARTICLE HISTORY

**Received:** 2025-03-11  
**Revised Version:** 2025-04-23  
**Accepted:** 2025-05-30  
**Published:** 2025-06-05  
**Copyright:** © 2025 by the authors  
**License:** CC BY-NC-ND 4.0  
**Document type:** Article

ARTICLE INFORMATION

**Main topic:**  
Empirical methods and academic performance  
**Main practical implications:**  
This study shows that Logistic Regression can effectively predict academic risk in higher education using basic data like high school GPA. It enables early identification of at-risk students, supports personalized learning, and guides efficient resource allocation. Its simplicity and interpretability make it practical for institutions, though future work should address fairness and include socio-emotional factors for broader inclusivity.

## INTRODUCTION

The prediction of academic performance in university students is an area of research of growing relevance, since it allows early identification of those who could face difficulties in their performance and, thus, facilitates the implementation of timely and personalized interventions (Cuji et al., 2020). Within this field, logistic regression has established itself as a widely used predictive technique due to its interpretability, simplicity and effectiveness in educational contexts (Villar, 2024; Zerkouk et al., 2024). However, the value of this technique lies not only in its statistical capacity to classify or predict results, but also in the ethical and practical implications of its use, especially in the search for a more just and equitable education.

### Argument-based validity and its relationship with logistic regression

To understand in depth how the use of predictive models in educational decision making is justified, it is essential to frame the study on the Argument-Based Approach (ABA) (Chapelle, 2021; Chapelle et al., 2010; Kane, 2006). According to this approach, validity is not limited to determining whether an instrument measures what it should measure but requires the construction of a logical argument that shows how and why the interpretations and uses of the results (or scores) are appropriate for the intended purpose.

In this sense, logistic regression provides empirical evidence to support the move from *observation* (predictor variables such as prior grades or test scores) to *interpretation* (probability of academic success). Such evidence contributes in a particular way to the inference of extrapolation or prediction, since, if the model demonstrates significant correlations with external performance criteria (e.g., performance in the first year of university), the argument that these variables measure relevant aspects of the construct to be evaluated is reinforced (Chapelle, 2021). However, as the same author warns, it is insufficient to base validity only on the predictive capacity: multiple layers of evidence are required that include, among other aspects, the fairness and usefulness of the results (Messick, 1989).

Equity refers to the fact that the interpretation and use of the predictions fairly benefit all groups of participants, without incurring in systematic biases that favor or harm any specific group (Chapelle, 2021). To guarantee this, it must be verified whether the predictive model works in a comparable way in different subgroups (according to gender, ethnicity or socioeconomic level, among others) and whether the decisions made on the basis of the predictions preserve equitable treatment (Chapelle, 2021; Serrano & Moreno-García, 2024). The evidence of fairness represents a fundamental layer within the validity argument; otherwise, its absence would undermine the legitimacy of the instrument or model in educational contexts.

Utility, on the other hand, involves examining the practical consequence and benefit of using the predictive model. A model may be statistically sound, but its true validity in educational practice is reflected in how well it promotes decisions that generate a positive impact (Chapelle, 2021; Messick, 1989). In the field of higher education, the usefulness of a logistic regression model is manifested when, for example, it allows early identification of at-risk students and, consequently, more effective tutoring, counseling or academic reinforcement (Paterson & Guerrero, 2022).

### Personalization and fairer education through artificial intelligence

The use of machine learning techniques, including logistic regression and other more complex techniques, is currently extending towards the personalization of learning and the search for a more equitable education. Forero-Corba and Negre Bennasar (2024) show, for example, how artificial intelligence can predict the level of digital competence of teachers and suggest individualized training itineraries. This type of solution provides additional evidence of usefulness: not only to select or diagnose, but also to provide feedback and propose concrete improvements.

Nevertheless, the debate on whether AI-mediated personalization represents a novel educational transformation or a *recycled promise* of previous approaches persists (Serrano & Moreno-García, 2024). From a historical and theoretical-practical perspective, studies such as that of Sánchez Sordo (2019) integrated pedagogical principles-in his case, Connectivism-with machine learning algorithms to personalize the learning experience and monitor the evolution of the student body in digital environments. Such research sets precedents for understanding that the true innovative value of AI depends more on its reflective and critical implementation than on its technological novelty per se (Serrano & Moreno-García, 2024).

### Purpose of the study

In the context of the Autonomous University of Baja California (UABC), the present study aims to evaluate the predictive capacity of logistic regression to identify first-year students at risk of low academic performance, analyzing as main variables the scores of the Higher Education Entrance Exam (ExIES), high school GPA, and geographic location of the campus. Aligned with EBA, aspects of equity and utility will be considered when interpreting the findings, to substantiate not only the statistical effectiveness of the model, but also its ethical and practical impact on decision-making for a fairer and more personalized education. The following are the guiding questions for this research:

- To what extent does a logistic regression model - based on ExIES scores, high school GPA, and campus location - succeed in predicting the risk of academic underachievement during the first year of college?
- How does the performance of logistic regression compare with that of other classification algorithms (Naive Bayes and Decision Tree) in identifying at-risk students?
- Which variables contribute most to prediction (e.g., high school GPA, specific ExIES scores)?
- How can the classification threshold be adjusted to balance sensitivity (detecting the greatest number of at-risk students) and accuracy (avoiding false positives) according to the needs of the institutions?

## METHODOLOGY

The methodological strategy was designed to examine the predictive capacity of several algorithms - with emphasis on logistic regression - with respect to the academic performance of first-year students at the Autonomous University of Baja California (UABC). The following is a detailed description of the steps followed in the study: population, variables, data preprocessing and validation of the models.

### Research design

A quantitative non-experimental design was adopted, with an explanatory scope (Hernández-Sampieri et al., 2018). The main objective is to determine the effectiveness of different statistical and machine learning models to predict the variable "risk of academic underachievement". Although the approach is primarily quantitative, the argumentative construction of validity (Chapelle, 2021) was contemplated to discuss the usefulness and fairness of the predictions in educational decision making.

### Population and selection criteria

The sample comes from a registry of 10,184 UABC students in their first year, whose data were obtained from internal institutional sources (IIDE, Vice Rector's Office). Records were included only when:

- There was complete information on ExIES scores (Language, Mathematics and Writing).
- They had a documented baccalaureate average.
- Campus location data was available (Ensenada, Mexicali or Tijuana).
- The academic performance variable was available (classified as "Passed" or "Did not pass" based on a threshold of 0.60).

Those records with incomplete or inconsistent information were excluded, following data cleaning guidelines from previous research (Burkov, 2019).

### Study variables

The analysis was developed using a database provided by the UABC IIDE and the UABC Vice Chancellor's Office, which includes information on 10,184 students after preprocessing. The table includes ExIES scores (Language, Mathematics and Writing), high school average, and campus (categorized as ENSENADA, MEXICALI or TIJUANA). Table 1 specifies each of the variables, as well as the label to be considered. Two types of variables were used:

- Predictor variables (features):
  - *ScoreL*, *ScoreM*, *ScoreE*: Scores obtained in the ExIES entrance exam: Reading Comprehension (*ScoreL*), Mathematics (*ScoreM*) and Written Language (*ScoreE*).
  - *Prom\_Prep*: Average obtained in high school.
  - *Campus\_ENSENADA*, *Campus\_MEXICALI*, *Campus\_TIJUANA*: Categorical variables representing the location of the campus.
- Label (target):

*Passed*: Binary variable created from the annual average academic performance (*Prom\_Año*), where it is classified as "1" if the student obtained an average higher than 0.60, and "0" otherwise.

**Table 1.** Possible variables

Table of Variables	Description	Data Type	Use in the Model
ScoreL	Language Score	Numeric	Predictor
ScoreM	Mathematics Score	Numerical	Predictor
ScoreE	Writing Score	Numerical	Predictor
High_School_average	High School Average	Numerical	Predictor
Ensenada_Campus	Dummy for Ensenada campus	Categorical (1/0)	Predictor
Campus_MEXICALI	Dummy for Mexicali campus	Categorical (1/0)	Predictor
Campus_TIJUANA	Dummy for campus Tijuana	Categorical (1/0)	Predictor
Passed	Annual average > 0.60?	Binary (1/0)	Label (target variable)

**Note.** Authors' development

## Libraries used in Python

To implement the analysis, the following libraries were used:

- *pandas*: For data manipulation and cleaning.
- *numpy*: For mathematical and linear algebra operations.
- *scikit-learn*: For the development of predictive models (Logistic Regression, Naive Bayes and Decision Tree), cross-validation and evaluation of metrics (*accuracy*, *recall*, *F1 Score*, *AUC-ROC*).
- *scipy.stats*: For the detection and elimination of *outliers* using the Z-score.
- *statsmodels*: For multicollinearity analysis through the *Variance Inflation Factor* (VIF).

## Data preprocessing

The preprocessing included several steps to ensure the quality and consistency of the information before training the models (Géron, 2019). Table 2 visualizes each of the steps per variable.

1. *Conversion of data types*: Some columns (e.g., ScoreL, ScoreM, ScoreE) were classified as *object* instead of numeric, so they were converted to *float* using `pd.to_numeric`.
2. *Handling of missing values*: Null values were imputed with the mean of the corresponding column in order to preserve the original distribution of the data.
3. *Coding of categorical variables*: The *One-Hot Encoding* method was applied (creating dummies such as Campus\_ENSENADA, Campus\_MEXICALI, Campus\_TIJUANA) so that the algorithms could process them without imposing an artificial order.
4. *Normalization and scaling*: To prevent higher magnitude values from dominating the training, we scaled to a range [0, 1] or applied a *StandardScaler*, depending on the approach.
5. *Outlier detection and elimination*: Outliers beyond  $\pm 3$  standard deviations were identified using Z-scores and eliminated when they were considered to distort the overall distribution.
6. *Multicollinearity check*: The *Variance Inflation Factor* (VIF) was calculated to detect whether there was high redundant correlation between predictor variables. In case of very high levels, the possibility of discarding or transforming any variable was considered.

Table 2. Data preprocessing

Stage	Variable(s)	Description of the problem	Action taken	Justification
Conversion of data types	L-Score, M-Score, E-Score, EndScore	Columns were sorted as <i>object</i> instead of <i>numeric</i> .	Conversion to float type using <code>pd.to_numeric</code> .	Ensures that numeric variables can be used in mathematical operations and supervised learning algorithms.
Missing values	L_Score, M_Score, E_Score, Prep_average	Presence of null values in key predictor columns.	Imputation with the mean for each column.	Imputation with the mean preserves the original distribution of the data, avoiding biases by eliminating complete records (Géron, 2019).
Coding of variables	Campus	The categorical variable <i>Campus</i> includes nominal values not processable by numerical algorithms.	Application of One-Hot Encoding to create dummy variables: <code>Campus_ENSENADA</code> , <code>Campus_MEXICALI</code> and <code>Campus_TIJUANA</code> .	It allows the models to correctly interpret the categories without imposing an artificial order, avoiding erroneous information in the prediction (Burkov, 2019).
Normalization and scaling	ScoreL, ScoreM, ScoreE, ScoreFinal, Average_Prep.	The scales of the numerical variables were different, which could bias the weights of the algorithms.	Scaled in the range [0, 1] using <code>MinMaxScaler</code> from <code>sklearn.preprocessing</code> .	Ensures that all variables have a proportional impact on the model, preventing larger values from dominating the optimization of the algorithm (Géron, 2019).
Detection and handling of outliers.	ScoreL, ScoreM, ScoreE, Average_Prep.	Presence of outliers outside 3 standard deviations, which distort the overall distribution.	Elimination of records with outliers using the Z-score method.	Extreme outliers can negatively affect model performance, especially in models sensitive to data distributions, such as logistic regression (Burkov, 2019).
Multicollinearity check.	L-score, M-score, E-score.	High correlation between predictor variables (according to correlation matrix), which may cause redundancy.	Calculation of Variance Inflation Factor (VIF) to detect collinearity.	Elimination or transformation of redundant variables improves model stability and avoids numerical problems (Géron, 2019).

**Note.** Authors' development

### Variable configurations and analysis models.

To test the individual and joint relevance of the variables, three configurations (Basic, Full and Examination) were analyzed. From them, three algorithms were trained and evaluated (Hastie et al., 2009):

1. Logistic Regression
2. Naive Bayes
3. Decision Tree

This multi-algorithmic strategy allows us to identify not only the overall effectiveness, but also the interpretability and robustness of each method (Dawar et al., 2024). Although logistic regression received greater emphasis for its ability to adjust classification thresholds and for its ease in providing indicators of variable relevance (Chapelle et al., 2010).

To compare the relevance of different combinations of variables, three sets of *features* were defined to train the models:

#### 1. Basic:

- Includes only the ExIES scores (ScoreL, ScoreM, ScoreE) and the bacculaureate average (Prom\_Prep).
- It is considered the minimum configuration necessary to capture the influence of prior performance and skills measured on the entrance exam.

#### 2. Complete:

- Includes all the variables of the Basic set plus the categorical campus indicators (Campus\_ENSENADA, Campus\_MEXICALI, Campus\_TIJUANA) and the variable indicating whether the student took a remedial course (Curso\_Si).

- It seeks to analyze whether campus location and participation in additional courses add predictive value.

### 3. **Exam:**

- It is composed solely of the ExIES scores (ScoreL, ScoreM, ScoreE).
- It allows determining whether the baccalaureate information is indispensable or whether the entry scores alone achieve competitive predictive performance.

These three subsets were scaled and used to train and evaluate the models independently, both in preliminary experimentation (Orange) and in more detailed programming (Python).

### **Evaluation and data partitioning**

1. Data partitioning: The dataset was partitioned into 90% training and 10% testing, following usual evaluation practices in machine learning (Kohavi, 1995). Also, stratified five-fold cross-validation (StratifiedKFold) was implemented to ensure that the minority class ("Did not pass" cases) was balanced in each partition (Géron, 2019).
2. Performance metrics:
  - Accuracy: overall hit ratio.
  - Recall: ability to identify students who actually passed ("Passed").
  - F1 Score: harmonic average between accuracy and recall.
  - AUC-ROC: area under the ROC curve, which measures the overall discrimination of the model (Chapelle, 2021).
3. Preliminary tests: These were performed in Orange, a visual analysis platform that facilitates exploratory experimentation. Subsequently, the results were refined in Python, adjusting hyperparameters and optimizing the models.
4. Threshold adjustment (Logistic Regression): In order to balance sensitivity (detect the most at-risk cases) and accuracy (reduce false positives), alternative thresholds (e.g., 0.3, 0.5, and 0.7) were explored (Paterson & Guerrero, 2022).

Discussion of findings: Finally, metrics were compared and analyzed under the lens of the Argument-Based Approach, assessing the fairness of predictions in different subgroups (Serrano & Moreno-Garcia, 2024) and the potential usefulness for the implementation of early warning systems (Messick, 1989).

## **RESULTS AND DISCUSSION**

### **Results processed with Python**

In this section, we show the findings derived from training three classification algorithms (Naive Bayes, Decision Tree and Logistic Regression) under different variable configurations -Basic, Full and Examination- both in Cross Validation and Random Sampling. Logistic regression maintains, according to the literature, an adequate balance between performance and ease of interpretation, a particularly valuable aspect in the educational context (Bishop, 2006; Breiman, 2001). In contrast, other more complex algorithms such as decision trees or ensemble methods may slightly improve accuracy, but sacrifice the clarity of their decisions, which complicates their practical adoption in institutions with limited resources.

In this regard, the comparison of different machine learning algorithms has been widely addressed in the literature (Contreras et al., 2020; Dawar et al., 2024). While some approaches prioritize simplicity and interpretability-characteristics associated with logistic regression and linear methods (Montgomery et al., 2012)-others emphasize maximizing accuracy even with more complex structures (Hastie et al., 2009). In our study, the inclusion of Naive Bayes and Decision Tree aligned with previous work that demonstrated their applicability in predicting academic success or failure (Contreras et al., 2020), as well as their effectiveness in varied scenarios when compared to more sophisticated algorithms (Dawar et al., 2024).

The results indicate that Logistic Regression (Basic), see Table 3, obtained the highest Recall (0.881) and F1 Score (0.798) in Cross Validation, outperforming Naive Bayes (Recall of 0.803 and F1 of 0.777) and Decision Tree (with lower scores in all metrics). Likewise, the AUC-ROC of the Logistic Regression (Basic) stands at 0.724, showing a better ability to distinguish between "Passed" and "Did not pass" classes compared to the other models. In Random Sampling (90% training, 10% test), the performance pattern was maintained, reflecting the robustness of the model to generalize to new data.

When analyzing the variable configurations (Basic, Complete and Exam), it was corroborated that the most efficient is the Basic configuration, since adding information about the campus or a booster course (Complete configuration) did not show relevant increases in the metrics, and restricting the variables to only the ExIES scores (Exam configuration) decreased the precision and Recall. These findings point to baccaulaureate GPA being a strong predictor of accurately predicting academic success in the first year, reinforcing the idea that prior performance is a strong indicator of student persistence and overall skills.

**Table 3** Results of the different by validation type

Model	Accuracy (VC)	Recall (VC)	F1 Score (VC)	AUC-ROC (VC)	Accuracy (RS)	Recall (RS)	F1 Score (RS)	AUC-ROC (RS)
Naive Bayes (Basic)	0.693617	0.803666	0.777451	0.713613	0.716389	0.787234	0.793424	0.752835
Decision Tree (Basic)	0.616258	0.698874	0.708182	0.575262	0.611384	0.668085	0.704036	0.576054
Logistic Regression (Basic)	0.702128	0.881119	0.797616	0.724512	0.740922	0.880851	0.824701	0.763383
Naive Bayes (Full)	0.695472	0.804976	0.778793	0.71368	0.720314	0.790071	0.796283	0.754068
Decision Tree (Complete)	0.604801	0.692157	0.700037	0.561475	0.609421	0.685106	0.708211	0.562109
Logistic Regression (Full)	0.7018	0.8803	0.797311	0.724226	0.737978	0.883688	0.823529	0.765831
Naive Bayes (Test)	0.662193	0.819221	0.763683	0.648096	0.689892	0.829787	0.787349	0.677185
Decision Tree (Test)	0.571195	0.646639	0.667683	0.53914	0.605496	0.68227	0.705279	0.563091
Logistic Regression (Test)	0.672995	0.941378	0.793215	0.6489	0.687929	0.93617	0.805861	0.68449

**Note.** VC = Cross Validation, RS = Random Sampling.

To deepen the relevance of each predictor, **Table 4** was analyzed, where indexes such as Information Gain, Gain Ratio and Gini Decrease are presented. The results confirm that *AverageBach\_Systems* is the most influential variable, followed by *E-Score* (Writing). This finding is consistent with previous literature, suggesting that sustained performance in high school better captures the competencies needed to successfully meet the challenges of the first year of college. In contrast, variables based on the entrance exam provide additional but less determinant information, reinforcing the importance of historical academic performance rather than point measurement on a single exam.

**Table 4.** Importance of Predictor Variables for the Prediction of At-Risk Students

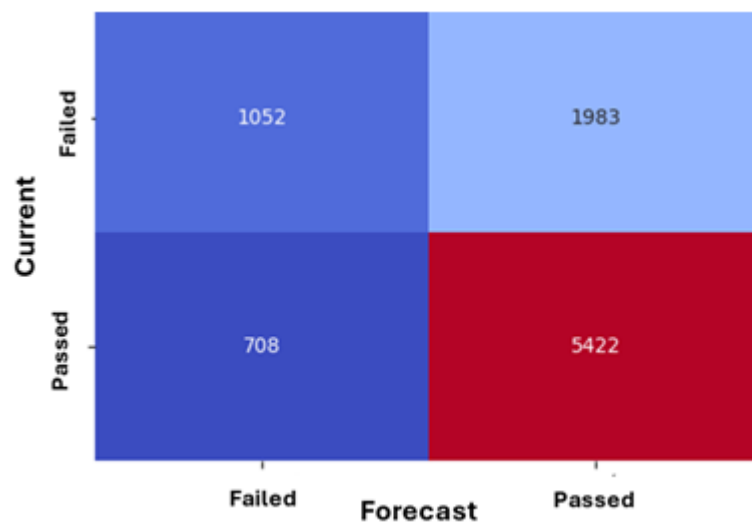
Variable	Information Gain	Gain Ratio	Gini Decrease
<b>AverageBach_Systems</b>	<b>0.087</b>	<b>0.044</b>	<b>0.051</b>
<b>E-score</b>	0.032	0.016	0.019
<b>ScoreL</b>	0.018	0.009	0.011
<b>ScoreM</b>	0.025	0.012	0.015

**Note.** Authors' development

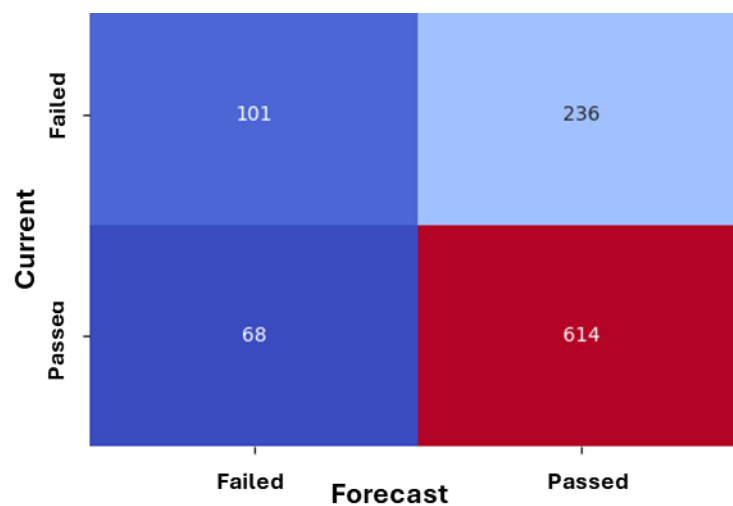
Figure 1 and Figure 2 summarize the average confusion matrices in the training set (90%) and the test set (10%). It was observed that, despite a slight drop in sensitivity during the test phase (common in supervised learning tasks), Logistic Regression retained an acceptable balance between Recall and Accuracy.

In addition, different classification thresholds (Table 5 and Table 6) were examined to adjust the balance between detection of at-risk students and reduction of false positives, with the result that a threshold of 0.5 provides an ideal balance point for most institutions, while lower (0.3) or higher (0.7) thresholds ( ) could be selected according to the specific needs of each program or academic context.

Overall, the results suggest that Logistic Regression (Basic) is the most robust option for predicting underachievement, as it combines high levels of Recall and F1 Score with good discriminative ability according to the AUC-ROC. This translates into an early detection of at-risk students, as well as a relative reduction in false positives, facilitating a more focused intervention by tutoring and academic accompaniment services.

**Figure 1.** Logistic Regression Confusion Matrix (90% training)

**Note.** Authors' development

**Figure 2.** Logistic Regression Confusion Matrix (10% test)

**Note.** Authors' development

**Table 5.** Threshold Setting in Logistic Regression of the Basic Model

Threshold	Accuracy	Recall (Sensitivity)
0.3	0.682	0.986
0.5	0.732	0.882
0.7	0.815	0.618

**Note.** Authors' development

**Table 6.** Threshold score on 10% of the test

Threshold	Accuracy	Sensitivity (Recall)	Accuracy	F1 Score	AUC-ROC
0.3	0.677	0.987	0.676	0.803	0.744
0.5	0.722	0.900	0.702	0.802	0.744
0.7	0.827	0.651	0.675	0.728	0.744

**Note.** Authors' development

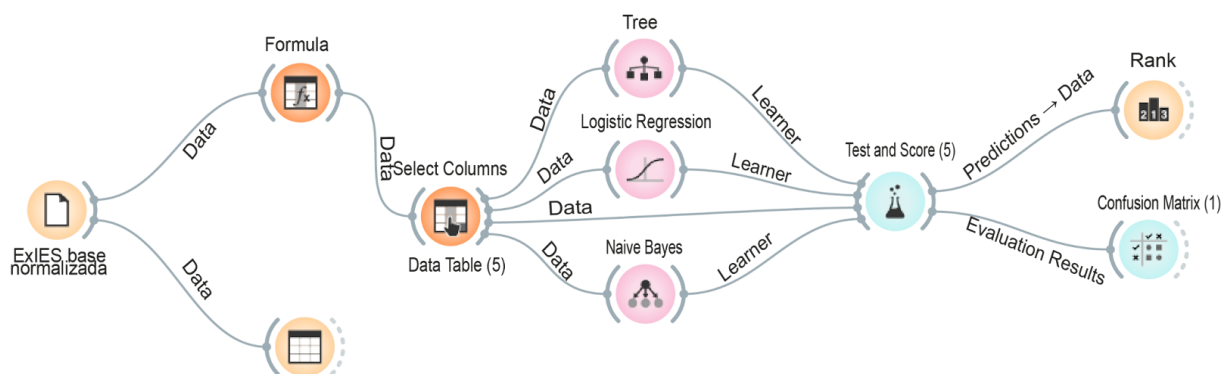


## Results in Orange

To complement the analysis, the same datasets and variable configurations were used within the Orange visual platform, where a workflow was designed (see Figure 3) that integrates feature selection with different classification algorithms and evaluation modules. Logistic Regression, Naive Bayes and Decision Tree were tested, both with Cross Validation (5 folds) and Random Sampling (10 replicates), to estimate the robustness of each model.

The Orange flowchart also allowed us to interactively observe how the classification varied when the predictor columns or the evaluation approach were modified. This exercise corroborated the relevance of high school GPA and scores in Written Language, Mathematics, and Language Arts as leading indicators of future performance. Also, the limited contribution of campus or extra-curricular course data, previously evidenced in Python, was evident in the Orange experiments.

**Figure 3.** Model Organization in Orange



**Note.** Authors' development

Table 6 shows the key metrics obtained. As in the Python experiments, Logistic Regression proved to be the most consistent model, achieving AUC values between 0.72 and 0.73, an Accuracy above 0.69 and a high F1 Score. Furthermore, the Matthews Correlation Coefficient (MCC) reinforces the interpretation that the model achieves a considerable balance between the hits on the positive and negative classifications. Naive Bayes ranked second, with an AUC around 0.72, while the Decision Tree exhibited the lowest discrimination ability, ranking below 0.58 in AUC under cross-validation.

**Table 6.** Results of the different Models according to Orange

Evaluation Method	Model	AUC	CA (Accuracy)	F1 Score	Accuracy	Recall	MCC
Cross Validation (5 folds)	Logistic Regression	<b>0.732</b>	<b>0.699</b>	<b>0.654</b>	<b>0.679</b>	<b>0.699</b>	<b>0.235</b>
	Naive Bayes	0.722	0.690	0.685	0.682	0.690	0.280
	Decision Tree	0.567	0.611	0.616	0.623	0.611	0.147
Random Sampling (10 repeats)	Logistic Regression	<b>0.728</b>	<b>0.702</b>	<b>0.662</b>	<b>0.683</b>	<b>0.702</b>	<b>0.248</b>
	Naive Bayes	0.717	0.689	0.684	0.680	0.689	0.277
	Decision Tree	0.576	0.611	0.617	0.624	0.611	0.150

**Note.** Authors' development

In this sense, the results in Orange fully coincided with those obtained in Python, which reinforces the validity and reliability of Logistic Regression as a predictive model suitable for the educational context analyzed. Its interpretability, accompanied by high Recall and AUC-ROC metrics, endorses its practical implementation in early warning systems and academic tracking strategies, so that institutions can anticipate and address the needs of students at high risk of underachievement in their first year of college. Two additional benefits are the scalability and simplicity of this model, which facilitates its adoption in resource-constrained environments, and its integration with other academic information management systems.

## DISCUSSION AND CONCLUSIONS

The results obtained provide evidence that logistic regression-with a basic model composed of bacculaureate average and ExIES scores-constitutes a robust and highly interpretable method for predicting the academic performance of first-year college students. The performance consistently outperformed Naive Bayes and Decision Tree on indicators such as Accuracy, Recall, F1 Score and AUC-ROC, which is consistent with previous studies highlighting the simplicity and effectiveness of logistic regression in educational settings (Cuji et al., 2020; Villar, 2024; Zerkouk et al., 2024). Likewise, the finding that bacculaureate GPA is the main predictor reinforces the idea that historical performance more stably reflects the academic competencies required in higher education (Paterson & Guerrero, 2022; Reyes Rocabado et al., 2007). Although the campus location variable (Complete) or the exclusive use of entrance scores (Exam) did not offer significant improvements, their analysis suggests that contextual information may not be determinative in this specific population, or else, that additional factors are required to capture potential inequities in the sample (Chapelle et al., 2010).

These findings can be framed as validity from the EBA (Chapelle, 2021; Kane, 2006), by providing evidence of extrapolation (the significant correlation between predictor variables and actual performance) and use (the potential application of the model for early identification of at-risk students). In terms of equity, although the study did not detect substantial differences between campuses, there is still a need to investigate other subgroups or variables related to gender, ethnicity or socioeconomic status that could reveal inadvertent biases (Serrano & Moreno-García, 2024). This approach is essential to ensure that the adoption of predictive tools does not deepen inequalities but rather promotes fairer and more tailored interventions (Chapelle, 2021). From a utility perspective, the model's ability to flag those with a high probability of underachievement provides clear support for academic decision making, such as the allocation of tutoring, remedial courses or support scholarships (Paterson & Guerrero, 2022). By adjusting the classification threshold, institutions can prioritize either coverage of the at-risk population (maximizing sensitivity) or accuracy to avoid unnecessary interventions, which emphasizes the operational flexibility of logistic regression. Beyond its predictive efficacy, this type of modeling opens opportunities for personalization and improvement of educational practice by aligning with emerging machine learning initiatives that stimulate continuous teacher training and pedagogical adaptation (Forero-Corba & Negre Bennasar, 2024; Sánchez, 2019).

Taken together, these results confirm that logistic regression, employed within a framework of argumentative validity and with attention to equity, can provide a solid basis for an early warning system with tangible benefits for students' academic trajectories. Finally, while statistical robustness and practical relevance are appreciated, future research should deepen the analysis of subpopulations, incorporate socioemotional and economic factors, and investigate the impact of interventions derived from the predictions. A reflective and critical implementation of these predictive strategies can strengthen the search for a more equitable and higher quality higher education, ensuring that artificial intelligence and machine learning become effective and ethically sustainable instruments for educational improvement (Serrano & Moreno-García, 2024; Messick, 1989).

## REFERENCES

- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
- Burkov, A. (2019). The hundred-page machine learning book. Andriy Burkov.
- Cecenardo-Galiano, C., Sumaran-Pedraza, C., Obregon-Palomino, L., Iparraguirre-Villanueva, O., & Cabanillas-Carbonell, M. (2024). Predictive model with machine learning for academic performance. En X. S. Yang, R. S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of Eighth International Congress on Information and Communication Technology (ICICT 2023)* (Vol. 695, pp. 955–967). Springer. [https://doi.org/10.1007/978-981-99-3043-2\\_81](https://doi.org/10.1007/978-981-99-3043-2_81)
- Chapelle, C. A. (2021). Argument-based validity in testing: Building and evaluating the case for test use. *Language Testing*, 38(3), 361–377. <https://doi.org/10.4135/9781071878811>
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13. <https://doi.org/10.1111/j.1745-3992.2009.00165.x>
- Contreras, L. E., Fuentes, H. J., & Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria*, 13(5), 233–246. <https://doi.org/10.4067/S0718-50062020000500233>
- Cuji Chacha, B. R., Gavilanes López, W. L., Vicente Guerrero, V. X., & Villacis Villacis, W. G. (2020). Student dropout model based on logistic regression. *Applied Technologies* (pp. 321–333). Springer. [https://doi.org/10.1007/978-3-030-42520-3\\_26](https://doi.org/10.1007/978-3-030-42520-3_26)
- Dawar, I., Negi, S., Lamba, S., & Kumar, A. (2024). Enhancing student academic performance forecasting: A comparative analysis of machine learning algorithms. *SN Computer Science*, 5, artículo 758. <https://doi.org/10.1007/s42979-024-03118-3>
- Forero-Corba, W., & Negre Bennasar, F. (2024). Diseño y simulación de un modelo de predicción para la evaluación de la competencia digital docente usando técnicas de Machine Learning. *EduTec, Revista Electrónica de Tecnología Educativa*, (89), 18–43. <https://doi.org/10.21556/edutec.2024.89.3201>
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2a ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>

- Kane, M. T. (2006). Validation. En R. L. Brennan (Ed.), *Educational Measurement* (4.a ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2, 1137–1143.
- Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (3a ed., pp. 13–103). American Council on Education.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5a ed.). John Wiley & Sons. <https://doi.org/10.1002/9781118097281>
- Paterson, K., & Guerrero, A. (2022). Predictive analytics in education: Considerations in predicting versus explaining college student retention. *Research in Higher Education Journal*, 44. <https://files.eric.ed.gov/fulltext/EJ1401369.pdf>
- Reyes Rocabado, J., Escobar Flores, C., Duarte Vargas, J., & Ramírez Peradotto, P. (2007). Una aplicación del modelo de regresión logística en la predicción del rendimiento estudiantil. *Estudios Pedagógicos*, 33(2), 101–120. <https://doi.org/10.4067/S0718-07052007000200008>
- Sánchez Sordo, J. M. (2019). Desarrollo de un entorno digital de aprendizaje desde el Conectivismo y su posterior análisis utilizando algoritmos de machine learning. *EduTec, Revista Electrónica de Tecnología Educativa*, (69), 1–22. <https://doi.org/10.21556/edutec.2019.69.1355>
- Serrano, J. L., & Moreno-García, J. (2024). Inteligencia artificial y personalización del aprendizaje: ¿innovación educativa o promesas recicladas? *EduTec, Revista Electrónica de Tecnología Educativa*, (89), 1–17. <https://doi.org/10.21556/edutec.2024.89.3577>
- Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, 4(2). <https://doi.org/10.1007/s44163-023-00079-z>
- Zerkouk, M., Mihoubi, M., & Chikhaoui, B. (2024). A machine learning-based model for student dropout prediction in online training. *Education and Information Technologies*, 29, 15793–15812. <https://doi.org/10.1007/s10639-024-12500-w>

### Contribution of each author to the manuscript:

Task	% of contribution of each author	
	A1	A2
A. theoretical and conceptual foundations and problematization:	50%	50%
B. data research and statistical analysis:	50%	50%
C. elaboration of figures and tables:	50%	50%
D. drafting, reviewing and writing of the text:	50%	50%
E. selection of bibliographical references	50%	50%
F. Other (please indicate)	-	-

### Indication of conflict of interest:

There is no conflict of interest

### Source of funding

There is no source of funding

### Acknowledgments

There is no acknowledgment