www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(2), April 2025, 71-83. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i2.1075

Application of Time Series Models in the Characterization of Dropout at the University of Cartagena, Colombia

Gabriel Elías Chanchí-Golondrino[1,2], Manuel Alejandro Ospina-Alarcón[1], Yasmín Moya-Villa[1]

[1] Facultad de Ingeniería, Programa de Ingeniería de Sistemas, Universidad de Cartagena, Cartagena de Indias, Colombia
[2] Grupo de Investigación LOGICIEL, Fundación Universitaria de Popayán, Popayán, Colombia

E-mails: gchanchig@unicartagena.edu.co, mospinaa@unicartagena.edu.co, ymoyav@unicartagena.edu.co

Abstract. One of the key factors affecting educational quality in Higher Education Institutions is student dropout. A high dropout rate can indicate student dissatisfaction with the relevance and quality of the education they are receiving. Therefore, a pressing challenge for these institutions is to characterize the dropout rates of academic programs, as well as overall dropout rates, with the aim of identifying trends or potential future variations that could support strategic decision-making to mitigate student dropout. In this regard, it has been generally observed that predictive models have predominantly employed machine learning techniques to predict whether a student will drop out based on social, economic, academic, and demographic variables, without focusing on characterizing percentage variations or future dropout trends. Thus, this article proposes an ARIMA-based time series model as a contribution to the characterization of historical dropout rates at the University of Cartagena, Colombia, from 1998 to 2022, with the goal of forecasting dropout rates for future years. This study was developed through four methodological phases: P1. Creation of training and testing datasets, P2. Identification of the model parameters p, q, and d, P3. Adjustment of potential models, P4. assessment of the ARIMA model and P5. Forecasting future dropout rates. The results showed that the ARIMA model (p=1, d=2, q=0) provided the best fit, enabling predictions to be made up to the second semester of 2028. A key conclusion of the study is that the dropout rate at the University of Cartagena over the next four years is expected to hover around 6%, meaning that for every 100 students entering the university, approximately 6 will drop out.

Keywords: ARIMA, student dropout, autoregressive models, time series, dropout rate.

Article Info
Received Jan 2, 2025
Accepted March 20, 2025

1 Introduction

Considering that education enhances productivity and facilitates technological advancements, thereby driving economic growth and expanding key sectors such as agriculture, manufacturing, and services, it is essential to pay attention to phenomena such as dropout in the university context, as this stage is when individuals have the ability to make crucial decisions and take control of their professional lives [1]–[3]. In this regard, student dropout represents a highly relevant issue for educational institutions, as any effort aimed at its characterization and reduction contributes to increasing coverage, as well as improving the quality, relevance, and efficiency of education [4]. Accordingly, failing to characterize and monitor dropout can result in direct economic losses for educational institutions, while also imposing significant social costs on both students and the broader community [5], [6]. In the same vein, academic dropout poses a significant challenge to economic growth, employment, competitiveness, and productivity, generating adverse effects not only for students and their families but also for society as a

whole. By interrupting educational training, human capital development is limited, which impacts the capacity for innovation and the production of skilled labor, both of which are key factors for the progress of any nation [7,8].

Student dropout is defined as the situation in which a student, either voluntarily or involuntarily, fails to enroll in their academic program for two or more consecutive periods, without being listed in the records as graduated or withdrawn for disciplinary reasons [9], [10]. Similarly, dropout can be defined as leaving school activities before completing a certain grade or educational level [11]–[13]. Dropout is directly related to factors such as grade repetition and academic delay; however, this phenomenon has a background that extends beyond the academic sphere [7]–[17]. Despite this, the factors contributing to dropout can vary significantly across different national and institutional contexts, highlighting the importance of adopting tailored approaches that are adapted to the specific circumstances of each environment in order to develop more effective and appropriate strategies to mitigate dropout [18], [19]. Although university dropout is often expressed in numerical terms, the reality for students from diverse backgrounds is complex and multidimensional. This means that the interruption of an educational process in a university career is not solely determined by economic factors, but also by other aspects, such as psychological factors, including motivation, course satisfaction, self-regulation, and self-efficacy expectations [20], [21]. Additionally, cultural and physical factors also play a significant role in this phenomenon [2], [9]. According to [14], the factors influencing school dropout can be grouped into two categories: extra-school variables, such as poverty, vulnerability, socioeconomic status, unemployment, ethnic background, family disintegration, and limited family educational expectations [22]–[26]; and intra-school variables, which include teacher authoritarianism, behavioral problems, adult-centrism, and poor academic performance [27].

With the widespread dissemination of artificial intelligence across various application contexts, these tools and models have increasingly been applied in the educational field. Specifically in the case of dropout, predictive models can provide insights that guide the mitigation of dropout risks at different stages of the academic process [28]–[31]. In this sense, predictive models can explain the different reasons or factors that influence a student's decision to drop out academically [30], [32], [33]. It is important to note that machine learning models have been widely adopted in this context, with models such as decision trees, logistic regression, and neural networks proving effective in predicting dropout by analyzing demographic, socioeconomic, and macroeconomic factors [34]–[36]. Additionally, research has shown that machine learning models such as gradient boosting and decision trees provide good results in predicting school dropout [29], [33]. Furthermore, as reported in [37]–[39], the ensemble model Random Forests has demonstrated high accuracy in predicting student dropout, often outperforming other machine learning models. Similarly, models like DeepFM, which combines Deep Learning and Factorization Machines, have achieved accuracy rates exceeding 99% in dropout prediction [29]. Moreover, these machine learning models have also been integrated into early warning systems focused on detecting students at risk of dropping out, enabling teachers and school administrators to intervene in a timely manner [37], [40].

Despite the widespread adoption and accuracy of these machine learning models, challenges such as data imbalance and the need to develop models that are generalizable across different educational institutions still persist [38], [41]. Similarly, it is worth mentioning that, in general, these studies have effectively focused on predicting whether a student will drop out based on various social, economic, and demographic factors. However, there has been little evidence of using these models to characterize and predict dropout rates through the use of time series models, for example. A time series can be defined as an ordered sequence of observations with a temporal component, which are often captured or recorded at regular time intervals [42]–[44]. These models allow for the prediction of future values of one or more variables by using only the information contained in the historical values of the series, enabling the analysis and measurement of the evolution of the studied variables over time [45]. Among the most widely used time series models are autoregressive integrated models, also known as ARIMA, which consist of three components: autoregressive (AR), differencing (I), and moving average (MA). These models are defined by three parameters: p (AR order), d (degree of differencing), and q (MA order) [46]. Due to their linear structure, ARIMA models have proven effective in capturing temporal patterns and making accurate predictions based on historical values of the analyzed variables, which has led to their extensive use in time series forecasting, particularly in economic and hydrological applications [46]–[48]. Additionally, these autoregressive integrated models are particularly useful when the time series being modeled is non-stationary due to a pronounced trend [49].

This article proposes an ARIMA-based time series model as a contribution to the characterization of student dropout data at the University of Cartagena, covering the period from the first semester of 1998 to the second semester of 2002. The primary goal is to generate predictions of future dropout rates, providing academic administrators with a key tool for strategic decision-making aimed at mitigating dropout within the program. The ability to anticipate dropout trends offers a significant advantage for higher education institutions, as it facilitates the early implementation of corrective measures, optimizes resource allocation, and promotes student retention, thereby contributing to the improvement of educational quality. For the implementation of the model, open-source tools were used, including Python libraries such as statsmodels, pmdarima, pandas, and matplotlib, ensuring

the reproducibility and accessibility of the approach in both academic and business settings. This work aims to serve as a methodological reference that can be extrapolated to other educational institutions, encouraging the adoption of predictive models to support continuous improvement in the quality of higher education.

The rest of the article is organized as follows: Sections 2 and 3 respectively present a review of the state of the art and a theoretical framework describing the mathematical foundations of the ARIMA model. Section 4 outlines the methodological phases that guided this research. Section 5 describes the results obtained in this study, including the determination of the optimal parameters for the ARIMA model, as well as the fitting and evaluation of various models using error metrics to identify the most optimal model. Additionally, this section presents dropout predictions for future semesters and a comparison between the ARIMA model and a linear regression model. Finally, Section 6 provides the conclusions and future work derived from this research.

2 State of the Art

The study of student dropouts has gained relevance in recent decades due to its implications for educational quality and institutional performance. Several predictive approaches have been explored to characterize and mitigate this phenomenon. Among them, machine learning models such as Random Forest, logistic regression and neural networks have shown efficacy in identifying at-risk students based on academic, economic and social factors [9, 19]. However, these approaches are limited to classifying whether or not a student will drop out, lacking temporal projections of dropout rates.

In the context of long-term prediction, models based on time series, such as ARIMA models, present significant advantages. These models have been widely used in fields such as economics, health and meteorology due to their ability to analyze non-stationary data and generate predictions based solely on historical values [8,50,51]. Although its application in education is less frequent, recent research has demonstrated its potential to project attrition trends in educational institutions, providing useful tools for strategic decision making [23].

For example, studies such as [8] have explored autoregressive models to analyze student dropout, while research in Latin America highlights the use of hybrid techniques to integrate predictions based on historical data and exogenous variables [9]. The present research distinguishes itself by employing a pure ARIMA model to characterize and predict dropout rates, establishing a clear contrast with methodologies focused on machine learning. This approach allows not only to understand historical patterns but also to project future scenarios with greater accuracy, an approach that still requires further development in educational literature.

3 Theoretical Background

ime series analysis, such as the one applied in this study, is based on ARIMA (Autoregressive Integrated Moving Average) models, which combine autoregressive, moving average and differencing components. These models have been used in various fields, such as economics and hydrology, to capture temporal patterns and make predictions based solely on historical data [54]-[57]. The ARIMA methodology differs from other predictive approaches in its ability to model non-stationary data and capture trends over time. In the educational context, student dropout is a multidimensional phenomenon that includes academic, economic and psychological factors. Models such as those described by [8] and [9] have highlighted these variables in ranking analyses, while this study employs ARIMA to project the long-term evolution of dropout. The ARIMA model is a tool for analyzing non-stationary time series. This model combines three fundamental components [51]:

1. Autoregressive (AR): Relates a current value to its past values by means of a linear model. This component uses several lags, represented as p, to capture autoregressive patterns in the series. Mathematically, it is expressed as [52]:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t \quad (1)$$

where Y_t Represent the value of the dependent variable (dropout rate) at the current time t (in the current semester). It is the data being predicted or analyzed at a specific time instant; $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ represents the past values of the time series (dropout rate) at instants t-1,t-2,...,t-p (in the past semesters). These values are used in the autoregressive (AR) component to model the relationship between the current value and its own previous values; $\phi_1, \phi_2, \ldots, \phi_p$, are the autoregressive coefficients that quantify the influence of each past value ($Y_{t-1}, Y_{t-2}, \ldots$) on the current value (Y_t). These coefficients are estimated during model fitting and

ϵt is the random error or noise term at instant t, which represents variations not explained by the past values of the series. It is assumed that ϵ_t follows a normal distribution with zero mean and constant variance.

2. Moving average (MA): Captures dependencies between past observations of the random errors (ϵ). This component is represented by the parameter q and is defined as [52]:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \phi_q Y_{t-q} \quad (2)$$

where μ is the base level of the mean around which past errors and fluctuations of Y_t are distributed; $\epsilon_{t-1},\epsilon_{t-2},...,\epsilon_{t-q}$ represent past errors at instants t-1,t-2,...,t-q. These terms are used in the moving average (MA) component to model the dependence between the current value and past errors, and $\theta_1, \theta_2,...,\theta_q$ are the moving average coefficients that quantify the influence of past errors ($\epsilon_{t-1},\epsilon_{t-2},...$) on the current value ($Y_t$).

3. Integrated (I): It is used to transform a non-stationary series into a stationary series by differentiation. Differentiation is represented by the parameter d, which indicates the number of times the series is differentiated [53]:

$$Y_t' = Y_t - Y_{t-1} \quad (3)$$

the series is repeatedly differentiated until stationarity is reached.

The full ARIMA model is denoted as ARIMA(p,d,q), where p is the autoregressive order, d is the degree of integration (differencing) and q is the order of the moving average. The general equation of the model can be expressed as [53]:

$$\Phi(B)(1-B)^d Y_t = \Theta(B)\epsilon_t \quad (4)$$

where B is the delay operator ($BY_t = Y_{t-1}$), Φ(B) and Θ(B) are polynomials of the coefficients ϕ and θ, respectively.

Student dropout is a critical problem that affects the quality and sustainability of higher education institutions. This study focuses on the University of Cartagena, where the analysis of dropout rates is essential to formulate strategies to increase retention. Unlike previous studies based on machine learning, this work employs ARIMA models to provide a time projection tool to anticipate future scenarios. The expected results not only benefit academic program administrators, but also contribute to the improvement of educational policies at the national level, optimizing the allocation of resources and strengthening the educational system.

4 Methodology

This study employed a methodology divided into five phases (see Fig. 1), which allowed the characterization and prediction of student dropout through the use of ARIMA time series models. The phases are described in detail below.

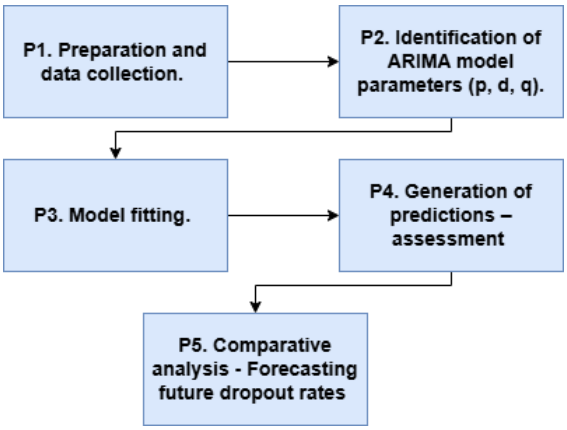

Fig. 1. Methodology Considered

Phase 1: Preparation and data collection. In this phase, historical data on student attrition from the Systems Engineering program at the University of Cartagena were collected, covering the period between the first semester of 1998 and the second

semester of 2022. The data were obtained from the SPADIES platform of the Ministry of National Education, consolidating a total of 49 semester records (see Fig.2). A sequential partition of the data was made in two sets: 85% for training (42 records) and 15% for testing (7 records), ensuring that the temporal order was maintained to avoid leakage of information.

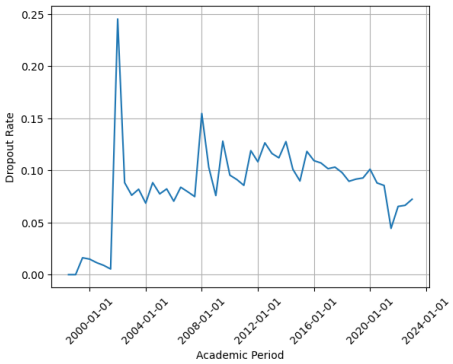

Fig. 2. Historical dropout data at the University of Cartagena

Phase 2: Identification of ARIMA model parameters (p, d, q). The Augmented Dickey-Fuller (ADF) test was used to evaluate the stationarity of the time series. The p-value obtained determined whether differencing was necessary to stabilize the series. The partial autocorrelation function (PACF) was used to determine the p parameter (autoregressive order), while the autocorrelation function (ACF) was used to identify the q parameter (moving average order). The parameter d was set according to the number of differentiations required to achieve stationarity.

Phase 3: Model fitting. Based on the parameters obtained, several ARIMA models were fitted using the Python statsmodels library. The models were evaluated using the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) criteria to select the one with the best balance between fit and complexity. Additionally, error metrics such as MSE (Mean Squared Error) and MAE (Mean Absolute Error) were calculated on both the training and test sets to assess model accuracy.

Phase 4: Generation of predictions – assessment. Predictions for the next 12 semesters were generated from the last observation of the training set. These predictions make it possible to anticipate possible dropout trends, facilitating strategic decision making to mitigate student dropout.

Phase 5: Comparative analysis - Forecasting future dropout rates. In this phase, a comparative analysis of the fitting capacity of a supervised learning model based on linear regression was conducted to evaluate its performance on the training and test sets. This aimed to determine the fitting and predictive capacity of this model compared to the ARIMA model adjusted in phase 3. This was done considering that the objective of this research is to characterize the curve of the dropout rate percentage as a function of academic semesters. Thus, in this study, it is necessary to perform the comparison with regression models rather than classification models.

In line with the above, previous studies have employed techniques such as neural networks [8]-[9], [19] and regression models [8], [11], [23] to predict student dropout, typically providing a binary classification of whether a student will drop out or not. However, these models fail to project the temporal evolution of the dropout rate. In contrast, the use of ARIMA models in this study enables the generation of continuous predictions over time, offering a valuable tool for long-term academic planning and management. This methodological distinction underscores the rationale behind the chosen approach and highlights its applicability in characterizing specific dropout patterns.

5 Results and discussion

This section presents the results obtained in this study, including the determination of the feasible parameters (p, q, and d) of the ARIMA model, the comparative evaluation of the models, the selection of the best-fitting model, the calculation of the error metrics for the selected model, and the generation of predictions for future semesters.

Based on the division of the time series data into training (85%) and testing (15%) sets, the first step was to determine the p, q, and d parameters of the ARIMA model. To determine the d parameter, the advantages of Python's statsmodels library were used

to apply the Dickey-Fuller stationarity test to the original series, as well as to its first and second differences, with the results shown in Table 1. The ADF statistic reflects the extent to which the series deviates from stationarity, indicating how far it diverges from stable behavior over time. On the other hand, the p-value assesses whether this deviation is statistically significant. If the p-value is less than 0.05, the null hypothesis that the series contains a unit root can be rejected, allowing us to conclude that the series is stationary.

Table 1. Results of the ARIMA model d-parameter analysis test

Time series	Statistical value
Original series	ADF statistic = -2.974 p-value = 0.037
First differentiation	ADF statistic = -7.378 p-value = 8.62×10^{-11}
Second differentiation	ADF statistic = -4.369 p-value = 0.00033

Based on the results of the Dickey-Fuller test, it can be concluded that the original series, as well as the first and second differences, exhibit p-values lower than 0.05, indicating that the null hypothesis of non-stationarity is rejected. This suggests that the series may be stationary in any of the three transformations (original series, first difference, and second difference). Regarding the determination of the p parameter of the ARIMA model, the partial autocorrelation function (PACF) plot of the original series was used, where the horizontal axis represents the order of the different lags, and the vertical axis displays the autocorrelation values (see Figure 3).

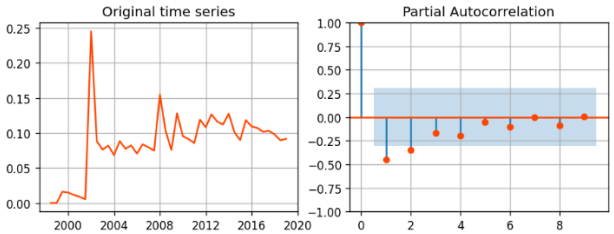

Fig. 3. Partial autocorrelation of the time series

Upon examining the partial autocorrelation function (PACF) plot presented in Figure 3, it is evident that the first two lags have values that significantly exceed the confidence bands, indicating significant correlation at these points. This suggests that the series exhibits dependence on the first two lags. In the context of ARIMA modeling, where the parameter p represents the number of lags in the autoregressive (AR) component, this evidence suggests that the value of p could be 1 or 2, as the PACF shows a rapid decline after these points. In other words, it is reasonable to assume that the autoregressive process does not extend significantly beyond the second lag.

Similarly, to determine the q parameter of the model, the autocorrelation function (ACF) plot of the original series was used, taking into account the lags that exceed the confidence bands. It should also be noted that in the ACF plot, the x-axis represents the lags of the series, while the y-axis shows the autocorrelation values (see Figure 4).

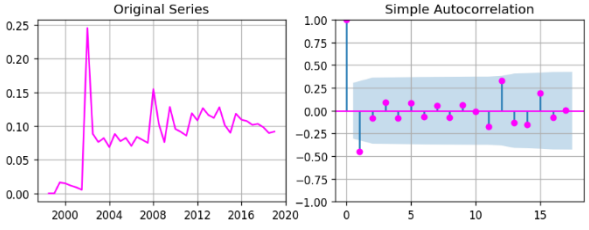

Fig. 4. Autocorrelation of the time series

According to the results shown in Figure 4, it can be observed that the first lag is significantly outside the confidence bands, indicating substantial autocorrelation at this point. The subsequent lags fall within the confidence bands, suggesting that there are no significant autocorrelations beyond the first lag. In this context, for the ARIMA model, the parameter q represents the number of lags in the moving average (MA) component of the model. Thus, since the first lag shows significant correlation

while the subsequent lags do not, it can be inferred that the moving average component of the time series has a possible order of q=1.

After identifying the possible parameters for the model, a comparison was made between the different ARIMA models that met these conditions. This evaluation focused on ensuring that the p-values associated with the model coefficients were statistically significant (p-values < 0.05), indicating the rejection of the null hypothesis that the coefficients are equal to zero, thus confirming that the autoregressive (AR), integrated (I), and moving average (MA) components contribute meaningfully to the model. Additionally, the AIC and BIC information criteria were considered, prioritizing models with the lowest values of these indicators, as they reflect a better balance between model accuracy and complexity. The comparative results of the different models that satisfy the possible values of p, q, and d are presented in Table 2. Thus, for each model, the p-values of the model components according to the order, as well as the AIC and BIC criteria, are included.

Table 2. Comparison of ARIMA models

ARIMA Model (p,d,q)	Results
(1,0,0)	AIC -141.502 BIC -136.288 Log Likelihood 74 P_values: const 1.091946e-10 ar.L1 3.639954e-03 sigma2 4.277948e-12
(1,0,1)	AIC -142.526 BIC -135.575 Log Likelihood 75 P_values: const 0.011047 ar.L1 0.005191 ma.L1 0.407341 sigma2 0.000007
(2,0,0)	AIC -140.905 BIC -133.954 Log Likelihood 74 P_values: const 3.386613e-06 ar.L1 8.911722e-02 ar.L2 3.023547e-01 sigma2 5.342688e-10
(2,0,1)	AIC -140.527 BIC -131.838 Log Likelihood 75 P_values: const 0.017175 ar.L1 0.431888 ar.L2 0.994302 ma.L1 0.521294 sigma2 0.000010
(1,1,0)	AIC -134.725 BIC -131.298 Log Likelihood 69 P_values: ar.L1 5.831536e-05 sigma2 4.452478e-46
(1,1,1)	AIC -139.699 BIC -134.558 Log Likelihood 73

	P_values: ar.L1 9.790465e-01 ma.L1 2.591839e-01 sigma2 4.099563e-39
(2,1,0)	AIC -137.67 BIC -132.529 Log Likelihood 72 P_values: ar.L1 4.896678e-02 ar.L2 5.532685e-02 sigma2 6.469281e-56
(2,1,1)	AIC -137.828 BIC -130.974 Log Likelihood 73 P_values: ar.L1 9.668829e-01 ar.L2 9.247272e-01 ma.L1 4.508271e-01 sigma2 1.289283e-42
(1,2,0)	AIC -98.776 BIC -95.398 Log Likelihood 51 P_values: ar.L1 6.279762e-19 sigma2 8.199401e-16
(1,2,1)	AIC -124.141 BIC -119.075 Log Likelihood 65 P_values: ar.L1 1.060571e-07 ma.L1 9.799869e-01 sigma2 9.799541e-01
(2,2,0)	AIC -109.668 BIC -104.601 Log Likelihood 58 P_values: ar.L1 7.047349e-42 ar.L2 1.484203e-05 sigma2 5.514287e-22
(2,2,1)	AIC -126.592 BIC -119.836 Log Likelihood 67 P_values: ar.L1 0.000002 ar.L2 0.002577 ma.L1 0.466914 sigma2 0.446234

According to the results presented in Table 2, it can be observed that the only models for which the p-values of the model components are significant are (1,0,0), (1,1,0), (1,2,0), and (2,2,0). This indicates that, for these models, the null hypothesis that the autoregressive and moving average component coefficients are equal to zero is rejected, confirming that these components contribute significantly to the model's fit. It is worth noting that when comparing the models, the AIC and BIC information criteria are lower for the (1,2,0) model. These criteria assess the balance between goodness of fit and model complexity, penalizing more complex models to avoid overfitting. Therefore, selecting the model with the lowest AIC and BIC values implies choosing the model that offers the best fit with the least complexity, ensuring better generalization capability. As a result, the (1,2,0) model is chosen for the adjustment and validation of the time series.

Before proceeding with the adjustment and validation of the selected (1,2,0) model, a thorough analysis of the residuals was conducted, as shown in Figure 4. In the left-hand plot, it can be observed that the residuals fluctuate around zero, suggesting that no structured patterns remain, which is an essential condition for considering the model appropriate. Additionally, the right-hand plot shows the distribution of the residuals, which approximates a normal distribution, although slight asymmetries are present. The closeness of the residuals to a normal distribution indicates that the model's errors are independent and approximately normally distributed, further supporting the validity of the ARIMA (1,2,0) model for predicting the time series in question.

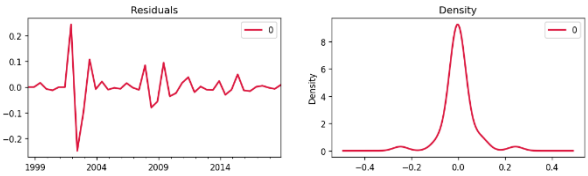

Fig. 4. Residuals corresponding to the selected ARIMA Model

Following the validation of the residuals for the selected ARIMA model, the ARIMA (1,2,0) model was fitted using the training set to evaluate its performance. To achieve this, it was essential to compare error metrics such as MSE (Mean Squared Error) and MAE (Mean Absolute Error) across both the training and testing sets. These metrics are crucial for quantifying the model's accuracy, as they measure the deviations between the model's predictions and the actual values of the time series. By analyzing these metrics, it is possible to assess the model's generalization capability and its effectiveness in predicting unseen data.

Table 3. Error metrics obtained from the ARIMA model

Dataset	Metrics obtained
Training	MSE: 0.0042
	MAE: 0.0345
	RMSE: 0.0649
Test	MSE: 0.00023
	MAE: 0.0111
	RMSE: 0.0152

As shown in Table 3, the error metrics MSE, MAE, and RMSE exhibit low values for both the training and testing sets, suggesting that the ARIMA (1,2,0) model has adequately captured the dynamics of the time series. Although these metrics do not have specific reference ranges, the values for the testing set (MSE = 0.00023, MAE = 0.0111, RMSE = 0.0152) are comparable to those of the training set (MSE = 0.0042, MAE = 0.0345, RMSE = 0.0649), indicating good model generalization and a low risk of overfitting. The similarity between the errors in both sets confirms that the model can accurately predict unseen data, validating its robustness in contexts beyond the sample used for fitting.

This can be seen more clearly in Figure 5, where the curve of the original time series (training and testing sets) is contrasted with the curve generated by the ARIMA (1,2,0) model, showing the dropout rates for the University of Cartagena over the 49 semesters studied. The graph shows that the ARIMA model closely follows the trend of the original series in both the training and testing sets, indicating a strong predictive capability. Despite some minor discrepancies, particularly in the early years where fluctuations are more pronounced, the model fits well with the overall evolution of the data over time. This is reflected in the proximity of the curves, especially in the more recent periods, where the model accurately predicts the series' behavior with a low margin of error, validating its use for future predictions. The confidence band shown in the figure further reinforces the model's robustness, as it encompasses most of the original series data points, suggesting low variability in the predictions.

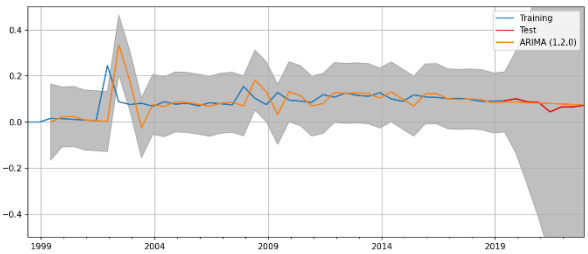

Fig. 5. Comparison of the original series vs the obtained ARIMA model

Similarly, considering that the objective of this research was to predict the percentage of academic dropout at the University of Cartagena, a supervised learning model based on regression was used, specifically linear regression. For this purpose, a workflow was modeled using the data mining tool Orange, as shown in Fig. 6.

From Fig. 6, it can be observed that the two variables considered in the model were the semester and the dropout rate, yielding a correlation of 0.329. This indicates a weak relationship between the two variables, which could significantly affect the predictive capacity of the linear regression model.

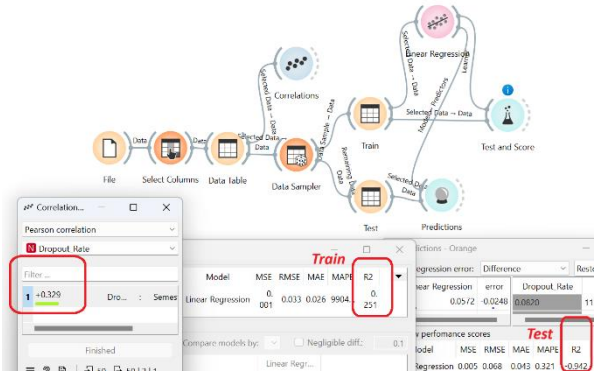

Fig. 6. Application of linear regression model on dropout data

On the other hand, as shown in Fig. 6, the dropout data were divided into training (80%) and test (20%) sets to obtain a coefficient of determination (R^2) of 0.251 for training and -0.942 for testing in the linear regression model. This indicates that the model exhibits poor performance in both sets, suggesting that the target variable (percentage dropout) cannot be explained by the academic semester variable. In this regard, the linear regression model is not considered suitable for modeling the series. This can be explained by the fact that, upon analyzing the original series and its differences, it exhibits stationary behavior, which makes it suitable for time series to be characterized using models such as ARIMA.

Finally, after evaluating the ARIMA (1,2,0) model, a set of 20 predictions was generated for the 20 semesters following the last period considered in the training set (second semester of 2018). These predictions provide a projection of the dropout rate at the University of Cartagena up to the second semester of 2028 (see Table 4). Since the initial predictions cover periods prior to 2024, Table 4 only presents the predictions for the semesters between the first semester of 2025 and the second semester of 2028. According to the results in Table 4, the ARIMA model suggests that the dropout rate for the 8 projected semesters hovers around an average of 5.892, with a standard deviation of 0.434. This indicates minimal variability in the predictions, further reinforcing the model's stability in its estimates. In practical terms, these predictions suggest that, on average, 6 out of every 100 students entering the University of Cartagena would not complete their studies, indicating relatively consistent dropout behavior throughout the projected period.

Table 4. Predictions based on the ARIMA model.

Academic semester	Predictions (%)
2025-I	6.555
2025-II	6.366
2026-I	6.176
2026-II	5.987
2027-I	5.797
2027-II	5.608
2028-I	5.418
2028-II	5.229

In discussing the results obtained, it is important to highlight that the main contribution of this study was the proposal of an ARIMA-based time series model for the characterization and prediction of student dropout at the University of Cartagena. This approach allows for the projection of future dropout rates, providing a valuable tool for academic authorities in decision-making. By having medium- and long-term predictions, institutions can proactively design and adjust retention strategies based on the trends anticipated by the model, thus reinforcing the potential of this work to improve educational management in higher education institutions.

This approach offers a significant advantage compared to previous proposals based on supervised learning techniques [54], [55] or neural networks [17], [56], [57], where the primary goal is to predict whether a particular student will drop out based on academic, social, or economic attributes. While these machine learning models have proven effective in identifying risk factors at the individual level, they do not allow for estimating the overall dropout behavior in future periods. According to the above, this article evaluated the effectiveness of the proposed model compared to a supervised learning model, finding that the linear regression model exhibits poor fit in both the training and test sets. This can be explained by the fact that both the original series and its differences exhibit stationary behavior. In this regard, the use of an ARIMA model provides added value by offering continuous temporal projections, enabling institutions to monitor the impact of their retention policies over time and adjust their interventions based on reliable predictions of dropout behavior.

6 Conclusions and future work

This study presented a model based on ARIMA time series to characterize and predict student dropout rates at the University of Cartagena. Unlike traditional methodologies based on machine learning, this research focused on analyzing and projecting temporal trends of dropout, providing a predictive tool that contributes to strategic decision making in educational management. Among the most significant findings, it was identified that the ARIMA (1,2,0) model presented the best fit for historical college dropout data, achieving low error metrics in both the training and test sets (MSE = 0.00023, MAE = 0.0111). These metrics evidence the model's ability to capture the dynamics of the time series, highlighting its usefulness in anticipating future trends with an adequate level of accuracy.

As an additional contribution, the time series data were evaluated using a supervised learning model based on linear regression, yielding poor fits in both the training and test sets. This suggests that the series is more accurately modeled by the ARIMA model, given the stationarity observed in the original series as well as in the first and second differences. This reinforces the idea that the ARIMA model effectively characterizes and provides solid forecasts regarding the academic dropout behavior at the University of Cartagena.

Fluctuations in dropout rates are sensitive to external factors such as retention policies and structural changes in the university. These observations highlight the relevance of using time series models not only as predictive tools, but also as inputs to monitor the impact of institutional interventions.

Finally, this work contributes to the advancement of the state of the art in the application of ARIMA models in the educational context, serving as a methodological reference for other institutions interested in addressing the dropout phenomenon. In addition, it opens the possibility of integrating more complex models, such as ARIMAX or hybrids, to consider external factors in future analyses.

As a line of future research, it is proposed to carry out a comparative analysis between the ARIMA model used in this study and other advanced methodologies, such as recurrent neural networks (RNN) and hybrid models. Likewise, the incorporation of exogenous variables through the use of ARIMAX models is suggested, which would make it possible to evaluate the impact of socioeconomic and academic factors on dropout rates. Finally, the implementation of early warning systems based on time series predictions could facilitate a timelier intervention by academic managers, optimizing student retention efforts.

Acknowledgements
The authors of this article would like to thank the University of Cartagena for the support provided during this research, and the Ministry of National Education for the data supplied through the SPADIES platform.

References

1. Tu, Z., Zheng, S., & Yuille, A. (2008). Shape matching and registration by data-driven EM. Computer Vision and Image Understanding, 109(3), 290–304.
2. Yin, P. Y. (2000). A tabu search approach to polygonal approximation of digital curves. International Journal of Pattern Recognition and Artificial Intelligence, 14(2), 243–255.
3. Pal, L. C. (2023). Impact of Education on Economic Development. Khazanah Pendidikan Islam, 5(1), 10–19. doi: 10.15575/kp.v5i1.25199.
4. Dávila-Morán, R. C., Aguero-Corzo, E. del C., Portillo-Ríos, H., & Quimbita-Chiluisa, O.-R. (2022). Deserción universitaria de los estudiantes de una universidad peruana. Revista Universidad y Sociedad, 14(2), 421–427. Available: http://scielo.sld.cu/pdf/rus/v14n2/2218-3620-rus-14-02-421.pdf.

5. Ozturk, I. (2008). The Role of Education in Economic Development: A Theoretical Perspective. SSRN Electronic Journal. doi: 10.2139/ssrn.1137541.
6. Ministerio de Educación Nacional. (2009). Deserción estudiantil en la educación superior Colombiana. Ministerio de Educación Nacional. Available: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf.
7. Véliz Palomino, J. C., & Ortega, A. M. (2023). Dropout Intentions in Higher Education: Systematic Literature Review. Journal on Efficiency and Responsibility in Education and Science, 16(2), 149–158. doi: 10.7160/eriesj.2023.160206.
8. Barroso, P. C. F., et al. (2022). DROPOUT FACTORS IN HIGHER EDUCATION: A LITERATURE REVIEW. Psicologia Escolar e Educacional, 26. doi: 10.1590/2175-35392022228736t.
9. Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. Data, 7(11), 146. doi: 10.3390/data7110146.
10. Silva, J., et al. (2020). Prediction of Academic Dropout in University Students Using Data Mining: Engineering Case. In Proceedings of the 2020 Conference on Innovative Data Science. doi: 10.1007/978-981-15-3125-5_49.
11. Romero-Contreras, K. Y., Castillo-Gil, D., Higuera-Hurtado, D. J., & Villalba-Gómez, C. E. (2022). Factores influyentes de la deserción estudiantil en la Universidad de La Salle (2018-2020). Virtu@lmente, 9(2). doi: 10.21158/2357514x.v9.n2.2021.3196.
12. MEN (Ministerio de Educación Nacional). Educación Superior.
13. Vega-García, R., Vázquez-Alamilla, M., Flores-Jiménez, R., & Rodríguez-Moreno, R. (2014). Propuesta para analizar la calidad educativa y deserción escolar a nivel superior en el estado de Hidalgo. Boletín Científico de la Escuela Superior de Tlahuelilpan, 2(3). Available: https://www.uaeh.edu.mx/scige/boletin/tlahuelilpan/n3/e6.html.
14. Velez, A., & López, D. (2004). Estrategias para vencer la deserción universitaria. Educación y Educación, 7.
15. Sáez, F., López, Y., Cobo, R., & Mella, J. (2020). Revisión sistemática sobre intención de abandono en educación superior. IX Conferencia Latinoamericana sobre el Abandono en la Educación Superior, 500, 91–100.
16. Espinoza, Ó., González Fiegehen, L. E., & Loyola Campos, J. (2021). Factores determinantes de la deserción escolar y expectativas de estudiantes que asisten a escuelas alternativas. Educación y Educación, 24(1), 113–134. doi: 10.5294/educ.2021.24.1.6.
17. Aldeman, M., & Szekely, M. (2017). An Overview of School Dropout in Central America: Unresolved Issues and New Challenges for Education Progress. European Journal of Educational Research, 6(3), 235–259. doi: 10.12973/eu-jer.6.3.235.
18. Noltemeyer, A. L., Ward, R. M., & Mcloughlin, C. (2015). Relationship between school suspension and student outcomes: A meta-analysis. School Psychology Review, 44(2), 224–240.
19. Alban, M., & Mauricio, D. (2019). Neural networks to predict dropout at the universities. International Journal of Machine Learning and Computing, 9(2), 149–153.
20. Diaz Lema, M., Vooren, M., Cannistrà, M., van Klaveren, C., Agasisti, T., & Cornelisz, I. (2024). Predicting dropout in Higher Education across borders. Studies in Higher Education, 49(1), 141–156. doi: 10.1080/03075079.2023.2224818.
21. Guzmán, A., Barragán, S., & Cala Vitery, F. (2021). Dropout in Rural Higher Education: A Systematic Review. Frontiers in Education, 6. doi: 10.3389/feduc.2021.727833.
22. Díaz-Mujica, A., Pérez-Villalobos, M. V., Bernardo-Gutiérrez, A., Fernández-Castañón, A., & González-Pienda, J. (2019). Affective and cognitive variables involved in structural prediction of university dropout. Psicothema, 31(4), 429–436. doi: 10.7334/psicothema2019.124.
23. Sommer, M., & Dumont, K. (2011). Psychosocial factors predicting academic performance of students at a historically disadvantaged university. South African Journal of Psychology, 41(3), 386–395.
24. Ingrum, A. (2007). High School Dropout Determinants: The Effect of Poverty and Learning Disabilities. Park Place Economist, 14, 73–79. Available: https://digitalcommons.iwu.edu/parkplace/vol14/iss1/16.
25. De Witte, K., & Rogge, N. (2013). Dropout from Secondary Education: all's well that begins well. European Journal of Education, 48(1), 131–149. doi: 10.1111/ejed.12001.
26. Peña, M., & Toledo, C. (2017). Ser pobre en el Chile Neoliberal: Estudio discursivo en una escuela vulnerable. Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 25(1), 207–218. doi: 10.11600/1692715x.1511225012016.
27. Foley, K., Gallipoli, G., & Green, D. A. (2014). Ability, Parental Valuation of Education, and the High School Dropout Decision. Journal of Human Resources, 49(4), 906–944. doi: 10.1353/jhr.2014.0027.
28. Espinoza, O., González, L. E., McGinn, N., & Castillo, D. (2021). Engaging dropouts with differentiated practices: some evidence from Chile. Research Papers in Education, 36(6), 637–656. doi: 10.1080/02671522.2020.1736615.
29. Darling-Hammond, L., & Cook-Harvey, C. (2018). Educating the whole child: Improving school climate to support student success. Available: https://learningpolicyinstitute.org/product/educating-whole-child-report.
30. Jeno, L. M., Danielsen, A. G., & Raaheim, A. (2018). A prospective investigation of students' academic achievement and dropout in higher education: a Self-Determination Theory approach. Educational Psychology, 38(9), 1163–1184. doi: 10.1080/01443410.2018.1502412.

31. Alruwais, N. M. (2023). Deep FM-Based Predictive Model for Student Dropout in Online Classes. IEEE Access, 11, 96954–96970. doi: 10.1109/ACCESS.2023.3312150.
32. López-Angulo, Y., Sáez-Delgado, F., Mella-Norambuena, J., Bernardo, A. B., & Díaz-Mujica, A. (2023). Predictive model of the dropout intention of Chilean university students. Frontiers in Psychology, 13. doi: 10.3389/fpsyg.2022.893894.
33. Song, Z., Sung, S.-H., Park, D.-M., & Park, B.-K. (2023). All-Year Dropout Prediction Modeling and Analysis for University Students. Applied Sciences, 13(2), 1143. doi: 10.3390/app13021143.
34. Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data. IEEE Access, 9, 133076–133090. doi: 10.1109/ACCESS.2021.3115851.
35. Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. European Journal of Higher Education, 10(1), 28–47. doi: 10.1080/21568235.2020.1718520.
36. Kim, S., Choi, E., Jun, Y.-K., & Lee, S. (2023). Student Dropout Prediction for University with High Precision and Recall. Applied Sciences, 13(10), 6275. doi: 10.3390/app13106275.
37. Kim, S., Yoo, E., & Kim, S. (2023). Why Do Students Drop Out? University Dropout Prediction and Associated Factor Analysis Using Machine Learning Techniques. Available: https://doi.org/10.48550/arXiv.2310.10987.
38. Barros, T. M., Souza Neto, P. A., Silva, I., & Guedes, L. A. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective. Education Sciences, 9(4), 275. doi: 10.3390/educsci9040275.
39. Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review, 96, 346–353. doi: 10.1016/j.childyouth.2018.11.030.
40. Cho, C. H., Yu, Y. W., & Kim, H. G. (2023). A Study on Dropout Prediction for University Students Using Machine Learning. Applied Sciences, 13(21), 12004. doi: 10.3390/app132112004.
41. Andrade-Girón, D., et al. (2023). Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review. ICST Transactions on Scalable Information Systems. doi: 10.4108/eetsis.3586.
42. Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., & Paganoni, A. M. (2022). Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques. Studies in Higher Education, 47(9), 1935–1956. doi: 10.1080/03075079.2021.2018415.
43. Oqaidi, K., Aouhassi, S., & Mansouri, K. (2022). Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. International Journal of Emerging Technologies in Learning, 17(18), 103–117. doi: 10.3991/ijet.v17i18.25567.
44. Gerolimetto, M. (2010). Introduction to time series analysis. Wiley-Blackwell. doi: 10.1002/9781118856406.ch5.
45. Aljandali, A., & Tatahi, M. (2018). Time Series Analysis. In Advanced Statistics for the Social and Behavioral Sciences (pp. 37–55). doi: 10.1007/978-3-319-92985-9_3.
46. Chattopadhyay, A. K., & Chattopadhyay, T. (2014). Time Series Analysis. In Statistical Tools for Managers (pp. 217–240). doi: 10.1007/978-1-4939-1507-1_9.
47. González, J., Amado, N., & Serrano, A. (2018). Análisis de Probabilidad de Devaluación de la Tasa de Cambio y Pronóstico de la Inflación en Colombia a través de un Modelo de Regresión Logística y Serie de Tiempo Estacionaria 2005 - 2016. Revista Espacios, 39(8), 8.
48. Liu, T., Liu, S., & Shi, L. (2020). ARIMA Modelling and Forecasting. In Business Analytics and Statistical Analysis Using R (pp. 61–85). doi: 10.1007/978-981-15-0321-4_4.
49. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175. doi: 10.1016/S0925-2312(01)00702-0.
50. Wang, L., Zou, H., Su, J., Li, L., & Chaudhry, S. (2013). An ARIMA-ANN Hybrid Model for Time Series Forecasting. Systems Research and Behavioral Science, 30(3), 244–259. doi: 10.1002/sres.2179.
51. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.
52. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). Time Series Analysis: Forecasting and Control. Wiley.
53. Shumway, R. H., & Stoffer, D. S. (2017). Time Series Analysis and Its Applications: With R Examples. Springer.
54. Guerrero-Quintana, M. J., & Medina-Jiménez, S. A. (2016). Modelos de series de tiempo aplicados a los expedientes de la Comisión de Derechos Humanos del Distrito Federal. Economía Informal, 398, 89–99. doi: 10.1016/j.ecin.2016.04.007.
55. Sharma, R. R., Kumar, M., Maheshwari, S., & Ray, K. P. (2021). EVDHM-ARIMA-Based Time Series Forecasting Model and Its Application for COVID-19 Cases. IEEE Transactions on Instrumentation and Measurement, 70, 1–10. doi: 10.1109/TIM.2020.3041833.
56. Sandoval-Palis, I., Naranjo, D., Vidal, J., & Gilar-Corbi, R. (2020). Early dropout prediction model: A case study of university leveling course students. Sustainability, 12(22), 1–17.
57. Zárate-Valderrama, J., Bedregal-Alpaca, N., & Cornejo-Aparicio, V. (2021). Classification models to recognize patterns of desertion in university students. Ingenieare, 29(1), 168–177.