PROPOSAL OF A TIME SERIES-BASED MODEL FOR THE CHARACTERIZATION AND PREDICTION OF DROPOUT RATES AT THE NATIONAL OPEN AND DISTANCE UNIVERSITY*

*Gabriel Elías Chanchí G.***
*Luis Fernando Monroy Gómez****
*Dayana Alejandra Barrera Buitrago*****

Received: 02/11/2023 • Accepted: 06/03/2024
https://doi.org/10.22395/rium.v23n44a7

ABSTRACT

Dropout rates are a key indicator of educational quality, making it imperative for educational institutions to design strategies to reduce them, thereby contributing to improved student retention and the achievement of academic objectives. While dropout research has primarily focused on machine learning methods applied to in-person education datasets, this article introduces a novel approach based on time series models for dropout rates analysis at the National Open and Distance University (UNAD). Methodologically, an adaptation of the CRISP-DM methodology was undertaken in four phases, namely: F1. Business and data understanding, F2. Data preparation, F3. Model building and evaluation, and F4. Model deployment. In terms of results, an open dataset on UNAD dropout, obtained from the SPADIES system between 1999 and 2021, was employed. Using Python libraries statsmodels and pandas, an ARIMA model was implemented, displaying optimal error metrics. This ARIMA model was utilized to forecast future dropout rates at UNAD, projecting a future dropout rate fluctuating around 23%. In conclusion, the ARIMA model developed for UNAD stands as an innovative and essential tool in

* This article has been derived from the thesis titled: 'Software System for the Characterization and Prediction of Dropout Rates at UNAD Using Time Series Models,' corresponding to the Data Science and Analytics postgraduate diploma at UNAD.

** DaToS (Desarrollo Tecnológico para la Sociedad) Research Group, Facultad de Ingeniería de la Universidad de Cartagena, Cartagena de Indias – Colombia (E-mail: gchanchig@unicartagena.edu.co), Orcid: https://orcid.org/0000-0002-0257-1988

*** Student of the Specialization in Data Science and Analytics of the UNAD, Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI de la Universidad Nacional Abierta y a Distancia, Tunja – Colombia (E-mail: lfmonroyg@unadvirtual.edu.co), Orcid: https://orcid.org/0009-0005-0128-6754

**** SIGICIENTIC Research Group, Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI de la Universidad Loyola, España (E-mail: dayana.barrera@unad.edu.co), Orcid: https://orcid.org/0000-0001-8867-9705

the educational realm, capable of accurately anticipating dropout rates for upcoming semesters. This provides UNAD with a unique advantage in strategic decision-making.

Keywords: ARIMA model, dropout, predictive model, retention, time series, time series model.

PROPUESTA DE UN MODELO BASADO EN SERIES TEMPORALES PARA LA CARACTERIZACIÓN Y PREDICCIÓN DE LAS TASAS DE DESERCIÓN EN LA UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA

RESUMEN

La tasa de deserción es un indicador clave de la calidad educativa, por lo que es imperativo que las instituciones educativas diseñen estrategias para reducirla y así aumentar la retención estudiantil y alcanzar los logros académicos. Mientras que la investigación sobre la deserción se ha concentrado principalmente en métodos de aprendizaje automático aplicados a conjuntos de datos sobre educación presencial, este artículo introduce un enfoque novedoso al utilizar modelos de series temporales para analizar la tasa de deserción de la Universidad Nacional Abierta y a Distancia (UNAD). En cuanto a la metodología, se adaptó el proceso CRISP-DM en cuatro fases, a saber: F1. Comprensión del negocio y de los datos, F2. Preparación de los datos. F3. Modelado y evaluación y F4. Despliegue del modelo. Respecto a los resultados, se empleó un conjunto de datos abiertos sobre la deserción en la UNAD que abarca desde 1999 hasta 2021, el cual se obtuvo del sistema SPADIES. Mediante el uso de las bibliotecas de Python statsmodels y pandas, se implementó un modelo ARIMA, el cual arrojó excelentes resultados en las medidas de error. Este modelo ARIMA se utilizó para predecir la tasa de deserción futura de la UNAD, la cual se proyecta que oscilará alrededor del 23 %. En conclusión, el modelo ARIMA desarrollado para la UNAD se destaca como una herramienta innovadora y fundamental en el ámbito educativo, capaz de predecir de forma precisa la tasa de deserción de semestres futuros, lo cual le otorga a la UNAD una ventaja única en la toma decisiones estratégicas.

Palabras clave: modelo ARIMA, deserción, modelo predictivo, retención, series temporales, modelo de series temporales

INTRODUCTION

Student dropout is a phenomenon of great significance in educational institutions, as strategic decisions focused on its characterization and reduction contribute to the expansion of education coverage, as well as the enhancement of quality, relevance, and efficiency in education [1]. Student dropout refers to the situation of a student who, whether voluntarily or involuntarily, remains unenrolled for two or more consecutive academic periods in their initially registered program, without being formally recognized as having graduated or withdrawn due to disciplinary actions [2], [3]. Furthermore, in accordance with the presentation in [4]–[6], dropout can be understood as the abandonment of educational activities before completing a certain grade or educational level. Although academic dropout is directly related to other factors such as course repetition and academic lag, this phenomenon may have explanations beyond the academic realm, being potentially associated with demographic, economic, and social aspects [7]–[10].

Given the widespread integration of artificial intelligence (AI) and the diverse range of tools that support its application, there has been a proliferation of studies aimed at characterizing and forecasting dropout rates across different educational institutions.

In [11], a literature review is conducted to examine the application of machine learning techniques for dropout rate characterization. The study reveals the utilization of various supervised and unsupervised learning models, emphasizing the significant influence of demographic factors on dropout rates. In [12], the evaluation and determination of the best machine learning model for predicting dropping out in the UNITELS of Peru are discussed. The study utilizes a dataset with socio-economic and academic variables, achieving a 91% accuracy with the KNN model. In [13], unsupervised learning techniques and principal component analysis are utilized for dropout characterization in a higher education institution in Bogotá. The study reveals that economic difficulties represent the primary cause of dropping out among males, contributing to 67.6% of cases. In [14], a predictive model based on neural networks is employed for the characterization and prediction of dropping out in the Faculty of Economic Sciences at the University of Buenos Aires between 2011 and 2013, achieving accuracy ranging between 90% and 95%. In [15], various supervised learning models are evaluated to predict dropout at the National Intercultural University of the Amazon. The study utilizes a dataset with variables related to academic profiles, economic aspects, and academic performance, with the KNN model obtaining the highest accuracy at 88.84%. In [16], supervised learning models are applied to the characterization of dropping among systems engineering students at the University of Cundinamarca, with the logistic regression model demonstrating the best accuracy at 71%.

The study utilizes a dataset comprising socio-economic and family-related aspects. In [17], various supervised learning models are evaluated on a dataset incorporating social, demographic, financial, geographic, and familial factors associated with a private university in Mexico, with the decision tree model achieving the highest accuracy at 93.33%. In [18], different machine learning models and neural networks are compared for predicting dropping out in Chilean school-level students, revealing that the neural network-based model had the best accuracy at 93.80%. In [19], various machine learning and neural network models are explored to characterize dropping out and associated economic losses in extension courses at the Technological University of Argentina. The evaluated models demonstrated an accuracy exceeding 90%. In [20], the optimal attributes for constructing a dataset to apply predictive models characterizing dropping out at the National University of Santa were identified, resulting in a total of 18 demographic and academic variables.

After reviewing the previous studies, two important aspects can be observed: a) the majority of the presented works have focused on characterizing dropping out in traditional education using supervised learning models and/or neural networks; b) the predictive models used are label-based, where the objective is to determine whether a student is likely to drop out or not by analyzing various types of variables including academic, demographic, and social factors. Thus, the application of time series-based models predicting the future dropout rate in distance higher education institutions has not been evidenced.

In this article, we propose an innovative time series-based model for characterizing and predicting dropout rates at the National Open and Distance University (UNAD), specifically employing ARIMA models. The focus of this study is on the National Open and Distance University, chosen due to its particular model of distance education and its leadership in said model in Colombia. This university requires students to be autonomous and possess self-learning capabilities, making it an ideal setting for the development and application of a time series-based model aimed at characterizing and predicting dropout rates. Given the distinctive nature of its educational approach and its prominent position in the Colombian educational landscape, understanding and addressing dropout phenomena in this context is crucial for improving retention strategies and enhancing student success in distance learning environments. The model was fitted and implemented using dropout data obtained from the SPADIES platform (System for Dropout Prevention in Higher Education) and utilizing the functionalities provided by the pandas and statsmodels libraries in Python. These libraries enable data manipulation, stationarity tests, model implementation, and validation. The time series-based model, designed to analyze dropout rates at UNAD, emerges as an innovative tool in the educational domain, capable of accurately anticipating future dropout rates.

In this manner, equipped with this information, the institution can promptly respond to emerging trends and proactively devise and implement policies and strategies to enhance student retention. Likewise, this model aims to serve as a reference that can be extrapolated to other educational institutions, offering guidance for characterizing and implementing dropout prediction models elsewhere.

The rest of the article is organized as follows: Section 2 presents the description of the methodological phases considered for the development of this work. In Section 3, the results obtained in this research studies are presented, encompassing the understanding of the dataset, the separation of data into training and testing sets, the determination of the ARIMA model parameters (p,d,q), the model fitting, and the predictions obtained for future semesters. Finally, Section 4 provides the conclusions and future work derived from this research study.

1. MATERIALS AND METHODS

For the development of the present research study, the 6 phases of the CRISP-DM methodology [21]–[23] were adapted into 4, thus defining the following phases: P1. Understanding the business model and data, P2. Data preparation, P3. Model building and evaluation, and P4. Model deployment (see Figure 1).

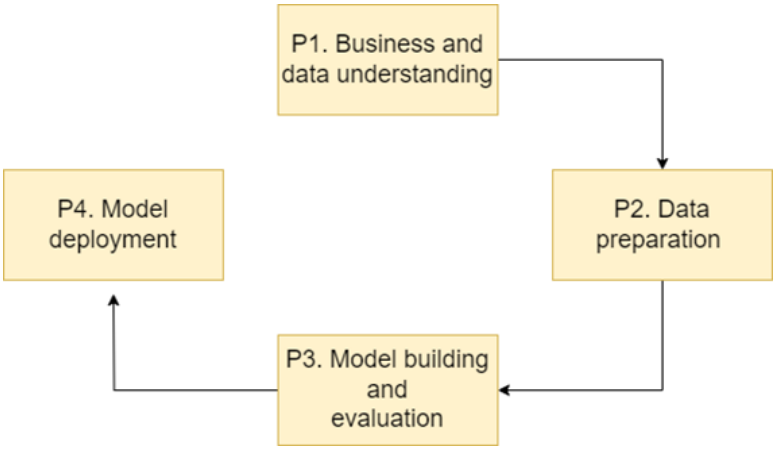
Phase 1. Understanding the Business Model and Data. In the first phase of the CRISP-DM methodology, the focus was on understanding the business and data. This step was essential for identifying and understanding the student dropout problem. Considering that the SPADIES platform of the National Ministry of Education is responsible for gathering and monitoring dropout rates in public and private universities in Colombia, data on dropout rates for UNAD were downloaded from this platform. These data encompass the collective dropout rates across various programs at UNAD and include historical values of dropout and retention rates. Therefore, for the present study, the dropout data were selected for model fitting purposes.

Phase 2. Data Preparation. The second phase focused on data preparation, a crucial step for any analysis project. Data cleaning tasks were carried out, and the dataset was divided into two parts: 85% for training and 15% for testing, which is a common practice in predictive modeling. In addition, the parameters p, d, and q of the ARIMA model, essential for time series analysis, were determined [24]. For the d parameter, the Dickey-Fuller stationarity test was used, and for the p and q parameters, correlation and partial autocorrelation were analyzed. All this was carried out using Python's statsmodels library, which offers robust tools for this type of analysis.

Phase 3. Modeling and Model Evaluation. In the third phase, the ARIMA model was fitted and implemented. Using the previously defined parameters, the model was fitted based on the training set. This phase was crucial to ensure that the model could adequately capture the trends and patterns in the dropout data. Subsequently, the effectiveness of the model was evaluated using the test dataset, and error metrics such as MSE, MAE, and RMSE were used. These metrics provided a clear view of the model's performance and its accuracy in predicting dropout data.

Phase 4. Model Deployment. In this phase, the model is evaluated and used to forecast dropout rates. At the evaluation level, the performance of the ARIMA model was extensively assessed to gauge its effectiveness in predicting student dropout, utilizing metrics such as MSE, MAE, and RMSE on both the training and test sets. These metrics provided an objective and quantitative evaluation, allowing for adjustments and improvements. The rigorous evaluation ensured that the model not only fit well with historical data but was also capable of making reliable and relevant predictions for the issue at hand. Subsequently, the deployment phase put the developed model into practice. In this stage, the ARIMA model was used to forecast dropout in future semesters, becoming a strategic tool for decision-making at the Open and Distance National University (UNAD). The ability to anticipate dropout trends allowed the Vice-rectorate of Services for Applicants, Students, and Graduates (VISAE) to implement proactive measures and strategies focused on student retention. This deployment of the model not only highlighted the practical applicability of the analysis, but also underscored the importance of data analytics to improve educational processes and evidence-based decision-making.

Figure 1: Methodology selected for the development of the research project

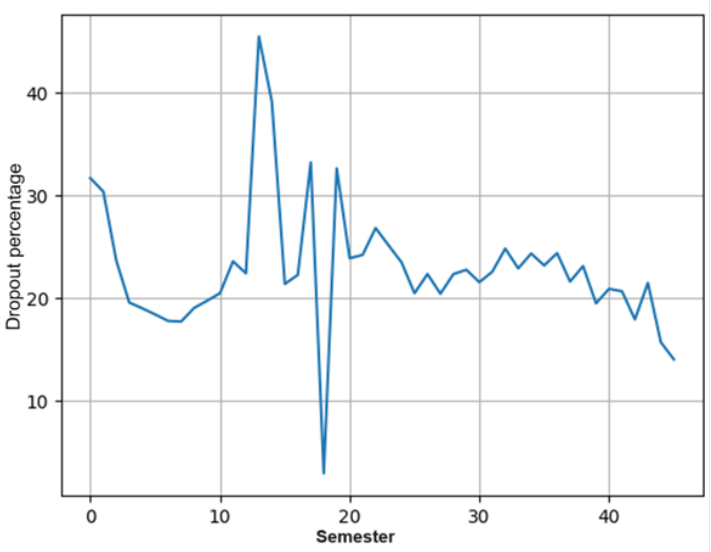


Source: own elaboration

2. RESULTS

First and foremost, it is important to note that the dropout data from UNAD between 1999 and 2021 were downloaded from the SPADIES web portal. These data encompass the historical dropout and retention rates at UNAD over the specified period. From these two variables, the dropout variable was considered, comprising a total of 46 samples corresponding to different academic semesters, as illustrated in Figure 2. The y-axis of the graph presented in Figure 2 represents the percentage of students dropping out associated with each of the 46 academic semesters.

Figure 2: Dropout rates at UNAD between 1999 and 2021



Source: own elaboration

After identifying dropping out data, each dropout percentage was indexed with a date stamp in Python to format it into a time series structure, facilitating data manipulation using the statsmodels and pandas libraries. Subsequently, the data series was split into a training set (85%) and a test set (15%) to fit the time series model with the training set and validate its effectiveness by comparing predictions on the test set with the real dropout data.

After separating the data into training and test sets, the ARIMA model parameters d, p, and q were determined using the training set. Regarding the d parameter, the Dickey-Fuller test was applied, using the statsmodels library, to the undifferenced series, as well as to the first and second differences, obtaining the results presented in Table 1. It is worth mentioning that the ADF statistical variable represents how far

the series is from being stationary, while the p-value variable indicates whether that difference is statistically significant. If the p-value variable is less than 0.05, the null hypothesis is rejected, and it is concluded that the series is stationary.

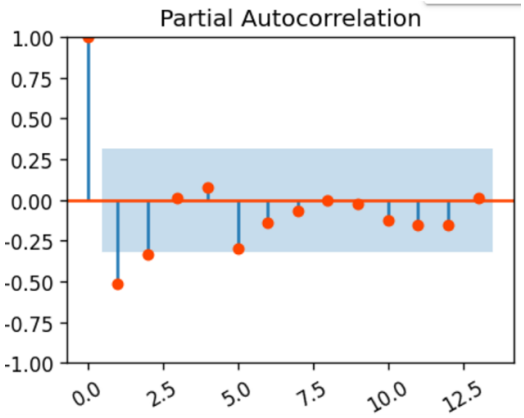
Table 1: Results of the Dickey-Fuller Test

Serie	Results obtained
Undifferenced series	ADF statistic = -5.679 p-value = 8.57e-07
First differencing	ADF statistic = -7.207 p-value = 2.277e-10
Second differencing	ADF statistic = -7.651 p-value = 1.782e-11

Source: own elaboration

According to the results of the Dickey-Fuller test presented in Table 1 for the series, it is possible to conclude that the p-value, both in the undifferenced series and in the first and second differences, is less than 0.05. Thus, the value of the variable d can be 0, 1, or 2. Regarding the determination of the p parameter of the model, the partial autocorrelation function (PACF) plot of the first differencing of the series (taking the value of d=1) was utilized. In this plot, the horizontal axis represents different lags, while the vertical axis displays the autocorrelation values (see Figure 3).

Figure 3: Analysis of the p parameter through partial autocorrelation

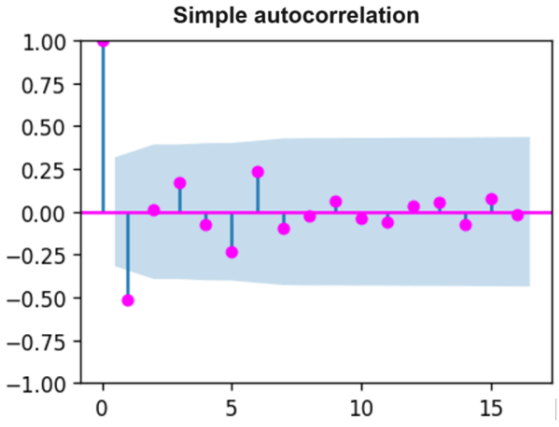


Source: own elaboration

According to Figure 3, both in the first and second lag, the autocorrelation value falls outside the confidence band. Consequently, the value of p can be either 1 or 2. However, these p values must be contrasted when evaluating the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) metrics of the ARIMA

model. In this regard, the autocorrelation plot of the first differencing of the series (d=1) was used to determine the q parameter of the model. The x-axis in the plot represents different lags, while the y-axis displays the autocorrelation values, as depicted in Figure 4, similar to the PACF plot.

Figure 4: Analysis of the q parameter through autocorrelation



Source: own elaboration

According to Figure 4, autocorrelation is significant only at the first lag, as subsequent lags fall within the confidence zone. Thus, the potential value that the q parameter can take is 1; however, it is necessary to contrast this value based on the AIC and BIC metrics. In conclusion, it is essential to evaluate ARIMA models with combinations of d=1,2, p=1,2, and q=1, while also considering the value of 0 for each of these variables. Based on these considerations, a comparative analysis was conducted among different ARIMA models, taking into account that the p-values of their parameters were significant (p-value < 0.05) and that the values of the AIC and BIC metrics were as low as possible. These comparative results for the different models are presented in Table 2.

Table 2: Results obtained for the ARIMA models evaluated

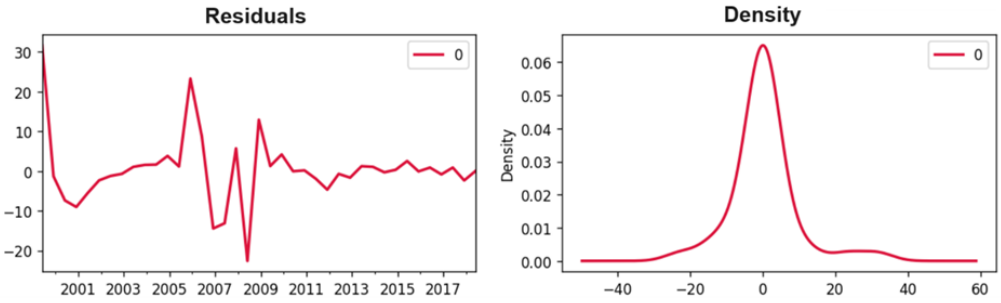
Model (p,q,d)	Results
(1,1,0)	AIC 266.778 BIC 270.053 p-values: ar. L1 1.987279e-06 sigma2 3.227014e-12
(1,1,1)	AIC 260.746 BIC 265.659 p-values: ar.L1 0.429077 ma.L1 0.993025 sigma2 0.993024

Model (p,q,d)	Results
(2,1,0)	AIC 264.322 BIC 269.235 p-values: ar.L1 1.583406e-11 ar.L2 1.236391e-02 sigma2 3.534466e-12
(2,1,1)	AIC 266.322 BIC 272.872 p-values: ar.L1 6.963156e-02 ar.L2 2.023228e-01 ma.L1 9.921796e-01 sigma2 3.893482e-11
(1,2,0)	AIC 290.709 BIC 293.93 p-values: ar.L1 2.070149e-19 sigma2 1.710740e-11
(1,2,1)	AIC 267.205 BIC 272.037 p-values: ar.L1 0.000003 ma.L1 0.899357 sigma2 0.899101
(2,2,0)	AIC 276.086 BIC 280.919 ar.L1 6.609809e-41 ar.L2 1.453771e-06 sigma2 1.322955e-08
(2,2,1)	AIC 265.394 BIC 271.837 ar.L1 3.372957e-09 ar.L2 4.626693e-02 ma.L1 9.623340e-01 sigma2 9.621830e-01

Source: own elaboration

According to the results obtained in Table 2, it is possible to conclude that in four of the eight compared models ((1,1,0), (2,1,0), (1,2,0), and (2,2,0)), the p-values are significant. However, the AIC and BIC metrics are superior for the model (2,1,0). Therefore, this model was chosen for the adjustment and validation processes. Additionally, it can be observed that, although the value of q was empirically estimated as 1, the AIC and BIC metrics determined that the most appropriate value for q is 0. Similarly, the values of p and d were in line with the empirical analysis conducted through partial autocorrelation and the Dickey-Fuller test. Prior to the adjustment process, residuals for the ARIMA model (2,1,0) were graphed (see Figure 5), revealing that these residuals follow a distribution close to normal, which is suitable for the selected model.

Figure 5: Residuals of the selected model



Source: own elaboration

It is crucial for ARIMA model residuals to adhere to a normal distribution to ensure accurate data representation and unbiased parameter estimation. Normality facilitates valid statistical inference and hypothesis testing, enabling reliable conclusions about model performance. Deviations from normality may indicate misspecification or unmodeled factors, necessitating further investigation. Ensuring normality of residuals is essential for a robust and valid interpretation of ARIMA model results.

Once the residuals corresponding to the selected ARIMA model were verified, the model was fitted with parameters (2,1,0) using the training set. Subsequently, the error metrics MSE and MAE of the fitted model were compared with respect to both the training and test sets. The results of the ARIMA model regarding the MSE and MAE error metrics are presented in Table 3. Notably, the coefficient of determination (R^2) is excluded from the analysis due to its suitability for regression models but not for time series models, given the temporal nature of the data, where R^2 may yield misleading results.

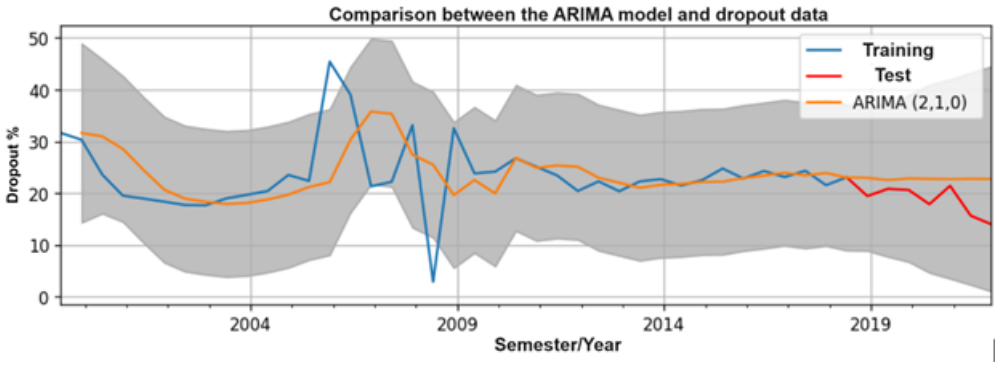
Table 3: Results obtained in the error metric of the model

Series	Error metrics obtained
Training	MSE = 76.267
	MAE = 4.985
	RMSE = 8.733
Tests	MSE = 24.837
	MAE = 4.222
	RMSE = 4.984

In Table 3, it can be observed that, although error metrics do not have a defined range, values in both the training and test sets are close and range between 4.9 and 76.3 for the training set and between 4.22 and 24.84 for the test set. This indicates that the model performs better in the case of the test set. This observation is further clarified

in Figure 6, where the curves of the training and test data are compared with the curve of the ARIMA model.

Figure 6: Comparison between the ARIMA model and dropout data



Source: own elaboration

Finally, once the ARIMA (2,1,0) model was evaluated, a set of predictions for the 12 semesters following the first semester of 2018 was generated. The first semester of 2018 served as the endpoint for the training dataset. Consequently, the model predicts dropout values up to the first semester of 2024 (See Table 4). Table 4 presents only 5 new predictions compared to the test set, covering the period from 2022-I to 2024-I. It is important to clarify that these 5 predictions were obtained through the ARIMA model and were not present in the original dataset. Additionally, it is worth highlighting that, instead of using a specific time window for the prediction, the complete dropout curve characterized by the ARIMA (2,1,0) model was considered, as it best captures the dropout data pattern for UNAD. Based on the results presented in Table 4, it is evident that dropout predictions for UNAD fluctuate around the mean value of 22.788%, with a standard deviation of 0.0034. This suggests that the various predictions are close to each other. Consequently, according to the model, approximately 23 out of every 100 students in any of the distance programs are predicted to drop out.

Table 4: Predictions obtained from 2022-I to 2024-I

Year	Prediction
2022-I	22.783284
2022-II	22.793902
2023-I	22.788193
2023-II	22.788546
2024-I	22.790183

3. RESULT DISCUSSION AND ANALYSIS

As part of the discussion, it is important to highlight that this study contributes by introducing an ARIMA time series model for the characterization and estimation of the UNAD's dropout rate in future semesters. This contribution stands in contrast to various state-of-the-art investigations, as documented in references [11]–[20], which leverage diverse economic, social, demographic, and academic factors to train various machine learning models aiming to predict whether a student is at risk of dropping out. Thus, the approach of the proposed model differs from those found in the state-of-the-art, both in terms of results and the employed technique.

Similarly, while machine learning allows for the use of approaches such as linear regression, an examination of the UNAD dropout curve reveals a behavior that is not entirely linear but rather stationary. This is evident when confirming that the data passes the Dickey-Fuller test without the need for time series differencing. Therefore, conventional machine learning models, which rely on diverse factors but do not explicitly consider the temporal structure of the data, may overlook the seasonal variability that could be present in dropout patterns. In this regard, time series-based models prove more relevant and better align with the inherent characteristics of the studied dropout data. These findings aim to serve as a reference for other institutions to implement time series models for characterizing and estimating future dropout rates within their own institutions.

Broadening the discussion, it is essential to recognize the advantages of using an ARIMA time series model to analyze dropout rates at UNAD. Unlike many contemporary studies that incorporate various economic, social, demographic, and academic factors into machine learning models to predict the risk of student dropout, this study's focus is on the temporal patterns of dropout data. This distinction is significant because it addresses a gap in current research methodologies. Although the factors used in traditional machine learning models are undoubtedly important, they often fail to capture the subtleties of time-dependent patterns in dropout rates.

The ARIMA model's ability to handle stationary data, as confirmed by the Dickey-Fuller test's results without the need for time series differencing, is particularly notable. This indicates that the dropout rate at UNAD exhibits consistent patterns over time, a characteristic that might be overlooked by conventional machine learning models. These models, sophisticated at handling various types of data, might not adequately account for the seasonal or cyclical trends inherent in educational data. The stationary nature of the dropout rates suggests that historical patterns are likely to repeat, which is crucial knowledge for predictive modeling.

Moreover, the use of a time series model like ARIMA allows for a more nuanced exploration of these patterns. It acknowledges that dropout rates are not just a product of individual factors but also of broader temporal trends, potentially influenced by institutional policies, economic cycles, or social changes. By focusing on these aspects, the study offers a more holistic understanding of dropout rates.

This approach also has practical implications for other educational institutions. Demonstrating the effectiveness of a time series model in predicting dropout rates paves the way for other institutions to consider similar methodologies. Institutions could benefit from understanding the temporal dynamics of their dropout rates, which might be influenced by unique institutional policies, academic calendars, or student support services.

In conclusion, the use of an ARIMA model for predicting dropout rates at UNAD not only offers a methodological alternative to conventional machine learning approaches but also highlights the importance of considering the temporal aspects of educational data. This perspective is valuable for educational institutions aiming to understand and mitigate dropout rates effectively, ensuring that their strategies are informed by the full scope of data available, including its time-dependent characteristics.

CONCLUSIONS

Considering the findings from the state-of-the-art review conducted, it was revealed that many studies concentrate on evaluating machine learning models to characterize dropout in traditional higher education settings. These models typically revolve around labeling a student as a dropout or non-dropout based on specific economic, social, and academic attributes. In this study, we propose a novel addition: an ARIMA-type time series model to estimate dropout rates at UNAD in future semesters. The aim of this model is to serve as a reference for other educational institutions or higher education entities intending to implement time series models that characterize dropout rates and support decision-making concerning student retention.

The proposed time series-based model emerges as an innovative tool in the educational domain that supports the work of the Vice-Rectorate for Services to Prospective Students, Current Students, and Graduates at UNAD. It possesses the capability to accurately forecast future dropout rates. In this manner, armed with this information, the institution not only has the opportunity to react promptly to impending trends, but also to proactively formulate and implement policies and strategies aimed at student retention. Furthermore, the model can be fine-tuned with new data derived from the University's strategic plans, thereby serving as a monitoring tool for student dropout at UNAD.

One of the significant advantages of the Python programming language lies in its suite of tools facilitating the implementation of various artificial intelligence models. While machine learning libraries are the most widely disseminated, this study has demonstrated the relevance and effectiveness of open-source libraries (specifically, pandas, matplotlib, and statsmodels) in the preprocessing, determination of parameters (p, d, and q), fitting, and evaluation of the proposed ARIMA model. In this context, these tools are intended to serve as a reference for the development of data science research focused on time series implementation.

Utilizing dropout data from UNAD obtained from the SPADIES platform of the Ministry of National Education, this research determined that the time series model with the best fit is the (2,1,0). Upon evaluating this model concerning the training set, error metric values (MSE, MAE, and RMSE) ranging from 4.9 to 76.3 were obtained, while concerning the test set, error metrics varied from 4.22 to 24.84. The conclusion drawn is that the error level of the ARIMA model regarding dropout data is low. Furthermore, predictions generated by the adjusted model for future semesters (up to 2024-I) indicated that future dropout predictions hover around 22%, with an average value of 22.788% and a standard deviation of 0.034. This suggests that, according to the model, out of every 100 students, 23 will drop out in the coming semesters at UNAD.

As a future extension of present research, the intention is to enhance the proposed ARIMA model by initially incorporating academic retention and subsequently incorporating other academic factors as exogenous variables within a SARIMAX model. The objective is to compare, in the future, the error metrics obtained in the SARIMAX model with those of the proposed ARIMA model in this article.

ACKNOWLEDGEMENTS

The authors express their gratitude to the National Open and Distance University (UNAD) and the University of Cartagena for the support received during the development of the current research study.

REFERENCES

[1] Ministerio de Educación Nacional, *Deserción estudiantil en la educación superior Colombiana*. Ministerio de Educación Nacional, 2009. [Online]. Available: https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_libro_desercion.pdf

[2] K. Y. Romero-Contreras, D. Castillo-Gil, D. J. Higuera-Hurtado, and C. E. Villalba-Gómez, "Factores influyentes de la deserción estudiantil en la Universidad de La Salle (2018-2020)," *Virtu@lmente*, vol. 9, no. 2, Apr. 2022, doi: 10.21158/2357514x.v9.n2.2021.3196.

[3] MEN (Ministerio de Educación Nacional), "Educación Superior."

[4] R. Vega-García, M. Vázquez-Alamilla, R. Flores-Jiménez, I. Flores-Jiménez, and R. Rodríguez-Moreno, "Propuesta para analizar la calidad educativa y deserción escolar a nivel superior en el estado de Hidalgo. Caso de un Instituto Tecnológico Superior en el occidente del estado de Hidalgo," *Boletín Científico la Esc. Super. Tlahuelilpan*, vol. 2, no. 3, 2014, [Online]. Available: https://www.uaeh.edu.mx/scige/boletin/tlahuelilpan/n3/e6.html

[5] D. Velez, A.; López, "Estrategias para vencer la deserción universitaria," *Educ. y Educ.*, vol. 7, 2004.

[6] F. Sáez, Y. López, R. Cobo, and J. Mella, "Revisión sistemática sobre intención de abandono en educacion superior," *IX Conf. Latinoam. sobre el Abandon. en la Educ. Super.*, vol. 500, pp. 91–100, 2020.

[7] Ó. Espinoza, L. E. González Fiegehen, and J. Loyola Campos, "Factores determinantes de la deserción escolar y expectativas de estudiantes que asisten a escuelas alternativas," *Educ. y Educ.*, vol. 24, no. 1, pp. 113–134, May 2021, doi: 10.5294/educ.2021.24.1.6.

[8] M. Aldeman and M. Szekely, "An Overview of School Dropout in Central America: Unresolved Issues and New Challenges for Education Progress," *Eur. J. Educ. Res.*, vol. 6, no. 3, pp. 235–259, Jul. 2017, doi: 10.12973/eu-jer.6.3.235.

[9] A. L. Noltemeyer, R. M. Ward, and C. Mcloughlin, "Relationship between school suspension and student outcomes: A meta-analysis," *School Psych. Rev.*, vol. 44, no. 2, pp. 224–240, 2015.

[10] M. Alban and D. Mauricio, "Neural networks to predict dropout at the universities," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 149–153, 2019.

[11] E. Cruz, M. González, and J. C. Rangel, "Técnicas de machine learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión.," *Prism. Tecnológico*, vol. 13, no. 1, pp. 77–87, Feb. 2022, doi: 10.33412/pri.v13.1.3039.

[12] J. E. Valero Cajahuanca, Á. F. Navarro Raymundo, A. C. Larios Franco, and J. D. Julca Flores, "Deserción universitaria: evaluación de diferentes algoritmos de Machine Learning para su predicción," *Rev. ciencias Soc. ISSN-e 1315-9518, Vol. 28, Nº. 3, 2022, págs. 362-375*, vol. 28, no. 3, pp. 362–375, 2022, Accessed: Sep. 11, 2023. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=8526463&info=resumen&idioma=ENG

[13] J. O. Gutierrez-Villareal, L. R. Fonseca-Gómez, and W. Pineda-Ríos, "Estimación de las principales causas de la deserción universitaria mediante el uso de técnicas de machine learning," *Aglala*, vol. 12, no. 2, pp. 293–311, 2021, [Online]. Available: https://revistas.curn.edu.co/index.php/aglala/article/view/2105

[14] E. Chinkes, "Pronósticos y data mining para la toma de decisiones. Pronóstico sobre la deserción de alumnos de una Facultad," *Cuad. del Cimbage*, vol. 1, no. 20, pp. 107–132, 2018, [Online]. Available: https://ojs.econ.uba.ar/index.php/CIMBAGE/article/view/1184/1793

[15] K. Rivera Vergaray, "Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica," *Innovación y Softw.*, vol. 2, no. 2, pp. 6–13, Sep. 2021, doi: 10.48168/innosoft.s6.a40.

[16] H. Y. Ayala-Yaguara, G. M. Valenzuela-Sabogal, and A. Espinosa-García, "Obtención de un modelo de minería de datos aplicado a la deserción universitaria del programa de Ingeniería de Sistemas de la Universidad de Cundinamarca," *Rev. ONTARE*, no. 7, pp. 134–150, 2019, [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=8705565

[17] A. Kuz and R. Morales, "Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México," *Educ. Knowl. Soc.*, vol. 24, p. e30080, Jun. 2023, doi: 10.14201/eks.30080.

[18] J. Smith Uldall and C. Gutiérrez Rojas, "An Application of Machine Learning in Public Policy: Early Warning Prediction of School Dropout in the Chilean Public Education System," *Multidiscip. Bus. Rev.*, vol. 15, no. 1, pp. 20–35, Jun. 2022, doi: 10.35692/07183992.15.1.4.

[19] I. Urteaga, L. Siri, and G. Garófalo, "Predicción temprana de deserción mediante aprendizaje automático en cursos profesionales en línea," *Rev. Iberoam. Educ. a distancia*, vol. 23, no. 2, pp. 169–182, 2020, [Online]. Available: https://redined.educacion.gob.es/xmlui/handle/11162/201572

[20] H. E. Caselli Gismondi and L. V. Urrelo Huiman, "Características para un modelo de predicción de la deserción académica universitaria. Caso Universidad Nacional de Santa," *LLamkasun Rev. Investig. Científica y Tecnológica*, vol. 2, no. 4, pp. 2–22, 2021, [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=8152475

[21] J. J. Espinosa Zúñiga, "Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública," *Ing. Investig. y Tecnol.*, vol. 21, no. 1, pp. 1–13, Jan. 2020, doi: 10.22201/fi.25940732e.2020.21n1.008.

[22] C. A. Dorado Bastidas, E. Y. Córdoba Campos, and G. E. Chanchí Golondrino, "Dashboard apoyado en inteligencia de negocios para toma de decisiones en el sector salud," *Rev. Gestión y Desarro. Libr.*, vol. 8, no. 16, pp. 1–13, 2023, doi: https://doi.org/10.18041/2539-3669/gestionlibre.16.2023.10226.

[23] U. Shafique and H. Qaiser, "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)," *Int. J. Innov. Sci. Res.*, vol. 12, no. 1, pp. 217–222, 2014.

[24] R. F. Ayala Castrejon and C. Bucio Pacheco, "Modelo ARIMA aplicado al tipo de cambio peso-dólar en el periodo 2016-2017 mediante ventanas temporales deslizantes," *Rev. Mex. Econ. y Finanz.*, vol. 15, no. 3, pp. 331–354, Jul. 2020, doi: 10.21919/remef.v15i3.466.