Data and Metadata. 2025; 4:589
doi: 10.56294/dm2025589

ORIGINAL

A new hybrid approach based on machine learning for more efficient time series forecasting

Un nuevo enfoque híbrido basado en el aprendizaje automático para un pronóstico de series temporales más eficiente

Hassan Bousnguar[1], Lotfi NAJDI[2], Amal BATTOU[3]

[1]IMISR Laboratory, Faculty of Science AM, Ibn zohr University. Agadir, Morocco.
[2]ISIMA Laboratory, IBN ZOHR University Agadir. Morocco.
[3]IRF-SIC Laboratory, IBN ZOHR University Agadir. Morocco.

Cite as: Bousnguar H, NAJDI L, BATTOU A. A new hybrid approach based on machine learning for more efficient time series forecasting. Data and Metadata. 2025; 4:589. https://doi.org/10.56294/dm2025589

Submitted: 20-05-2024 **Revised:** 02-09-2024 **Accepted:** 12-04-2025 **Published:** 13-04-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding Author: Hassan Bousnguar

ABSTRACT

Introduction: forecasting new student enrollment in bachelor's degree programs has emerged as a critical need for higher education institutions. Accurate enrollment predictions are essential for improving the student-teacher ratio and optimizing resource allocation.
Method: a hybrid approach combining statistical and machine learning techniques was proposed to develop accurate forecasting models. The study utilized the historical enrollment database of Ibn Zohr University, which included data from over twenty institutions dating back to 2003. This dataset was used to train and validate the proposed models.
Results: the hybrid approach demonstrated superior accuracy compared to standalone statistical and machine learning models. The results indicated that the proposed method effectively captured enrollment trends and provided reliable forecasts.
Conclusions: the study concluded that the hybrid approach offers a robust solution for enrollment forecasting in higher education. It highlighted the potential of combining statistical and machine learning techniques to improve prediction accuracy, thereby aiding institutions in better planning and resource management.

Keywords: Forecasting Time Series; Hybrid Models; Machine Learning; ARIMA; GRU.

RESUMEN

Introducción: la previsión de la matrícula de nuevos estudiantes en programas de licenciatura se ha convertido en una necesidad crítica para las instituciones de educación superior. Las predicciones precisas son esenciales para mejorar la proporción de estudiantes por docente y optimizar la asignación de recursos.
Método: se propuso un enfoque híbrido que combina técnicas estadísticas y de aprendizaje automático para desarrollar modelos de predicción precisos. El estudio utilizó la base de datos histórica de matrículas de la Universidad Ibn Zohr, que incluye datos de más de veinte instituciones desde 2003. Este conjunto de datos se utilizó para entrenar y validar los modelos propuestos.
Resultados: el enfoque híbrido demostró una precisión superior en comparación con los modelos estadísticos y de aprendizaje automático por separado. Los resultados indicaron que el método propuesto capturó eficazmente las tendencias de matrícula y proporcionó pronósticos confiables.
Conclusiones: el estudio concluyó que el enfoque híbrido ofrece una solución robusta para la previsión de matrículas en la educación superior. Se destacó el potencial de combinar técnicas estadísticas y de aprendizaje automático para mejorar la precisión de las predicciones, ayudando así a las instituciones en una mejor planificación y gestión de recursos.

© 2025; Los autores. Este es un artículo en acceso abierto, distribuido bajo los términos de una licencia Creative Commons (https://creativecommons.org/licenses/by/4.0) que permite el uso, distribución y reproducción en cualquier medio siempre que la obra original sea correctamente citada

Palabras clave: Pronóstico de Series Temporales; Modelos Híbridos; Aprendizaje Automático; ARIMA; GRU.

INTRODUCTION

Forecasting student enrollment time series is a promising domain and has a major impact on all fields of activity, especially higher education. Hence the need to develop forecasting models that can anticipate the increase of student enrollment in higher education.

Several models have been developed in the literature for various areas of activity, including forecasting of wind speed and direction. A Study proposed a statistical model based on AutoRegression and Moving Average (ARMA) to forecast wind speed and direction.[1] A study presented a hybrid model based on machine learning for the forecasting of wind speed. Several research papers have also been done for financial time series.[2] A review of the state of the art of forecasting models applied to financial time series.[3,4,5] The health sector has also seen a multitude of works on forecasting, especially for COVID-19.[6,7,8] Education and higher education, in particular, have experienced a lack of forecasting models developed specifically for this context. Some researchers have worked in this area but their work has not yet met the needs of stakeholders.[9,10,11]

The time series data used in this paper were collected from all courses offered at IBN ZOHR University. The new approach developed in this work is used to train the data for the bachelor's degree program at IBN ZOHR University and evaluated to select the model suitable for our context.

The objective of our research is to develop a model that can produce accurate forecasts in an educational context to help top management make strategic decisions concerning the evolution of the number of enrollments in each institution, and consequently provide the necessary resources to monitor this evolution.

The remainder of this paper is organized as follows. Section 2 presents a literature review on enrollment forecasting. Section 3 describes the methodology of our approaches. Section 4 presents the dataset and the results obtained by the models used and the discussion of those results in section 5. A summary and direction for future work is presented in the section 6.

Related work

Today, higher education is facing a major challenge in turning raw data into meaningful knowledge that can support decision-making.[12] To overcome this, higher education institutions could adopt forecasting techniques to streamline resources and concentrate their efforts. Forecasting is a key step in the decision-making process and is widely employed in the business and organizational sectors.[13]

Multiple techniques and approaches are used to complete this task, and several discipline are engaged in the development of solution, Statistical scientists are the first to develop models for forecasting and one of the older model is appear in 1950s, this model is the exponential smoothing in witch[14] introduce the basic exponential smoothing model and its estimation procedure. Since then, various extensions and modifications of the basic model have been proposed, such as seasonal exponential smoothing (Holt-Winters method) and state space models with Exponential Smoothing (ES). Moreover, there have been efforts to incorporate additional features, such as trend and seasonality, into the basic model.

Recent studies have also explored the use of exponential smoothing for demand forecasting in various industries, such as healthcare and retail. AutoRegressive Integrated Moving Average (ARIMA) models are among the most commonly used time series forecasting techniques in both academic research and practical applications. The basic idea behind ARIMA is to model a time series as a combination of autoregressive (AR) and moving average (MA) components, with an optional differencing step to remove any trend or seasonality. Box et al.[15] first introduced the ARIMA methodology and proposed a systematic approach for model identification, parameter estimation, and diagnostic checking. Since then, numerous extensions and refinements of ARIMA have been proposed, such as seasonal ARIMA (SARIMA)[16] models, vector ARIMA (VARIMA) models, and transfer function models with ARIMA errors.

Recurrent Neural Networks (RNNs)[17] have gained popularity in recent years for their ability to model time series data. One type of RNN that has been particularly successful in time series forecasting is the Long Short-Term Memory (LSTM) network. LSTMs[18] are a type of RNN that can retain information over a long period of time, making them well-suited for modeling complex temporal dependencies. Another type of RNN that has shown promise in time series forecasting is the gated recurrent unit (GRU) network.[19] Like LSTMs, GRUs are able to capture long-term dependencies in time series data. However, GRUs have fewer parameters than LSTMs, making them more computationally efficient.

Hybrid models that combine different types of forecasting techniques have gained attention in recent years due to their ability to improve the accuracy and robustness of time series forecasting. One popular approach is the combination of statistical models and machine learning algorithms, such as ARIMA and Artificial Neural Networks (ANNs).

https://doi.org/10.56294/dm2025589

According to [5], hybrid models that combine statistical methods and machine learning techniques can provide more accurate and robust forecasts compared to individual models.

Another type of hybrid model that has gained attention in time series forecasting is the combination of multiple machine learning algorithms. For example, an ensemble of multiple ANN models can be used to reduce the variance in the predictions and improve the overall accuracy of the forecast. As [20] note, that Ensemble learning techniques have been widely used in time series forecasting, and can effectively improve the accuracy and robustness of the predictions by combining the outputs of multiple models.

Overall, hybrid models have shown great potential for improving the accuracy and robustness of time series forecasting. As [21] affirm that hybrid models can exploit the strengths of different methods and provide more accurate forecasts compared to individual models.

METHOD

In this section, we present the methodology for improving the forecasting process under higher education context from data collection to model validation including hybrid model development.

Data collection

Data collection is the first step in the data preprocessing. It involves gathering data from different sources. In our case, we use a relational database as the main data source.

A relational database provides us with structured, reliable, and secure data. The database we use, Oracle 11g, contains data from different sources:

- Operator input.
- Platforms of pre-enrollment.
- Flat file.

Data transformation is often necessary to prepare data before using it in a forecasting model. Here are some examples of transformations used in forecasting time series:

- Data normalization: this involves putting all variables on the same scale, which can be useful if some variables have very different scales of measurement. For example, if one variable is expressed in millions of euros and another in percent, it can be useful to normalize the data so that they are comparable, this type of transformation is often used for machine Learning models.
- Time series differencing: is used to eliminate any trends, and ensure that the series is constant over time, as well as the variance. The original series is replaced by the series of adjacent differences, this operation can be repeated several times on the series until a stationary series is obtained, this transformation is mandatory in the ARMA family models.

The database used in this work includes information on students from enrollment to graduation. We are most interested in the number of students enrolled by institution and by field of study, this indicator will be used to create our time series, as shown in figure 1.

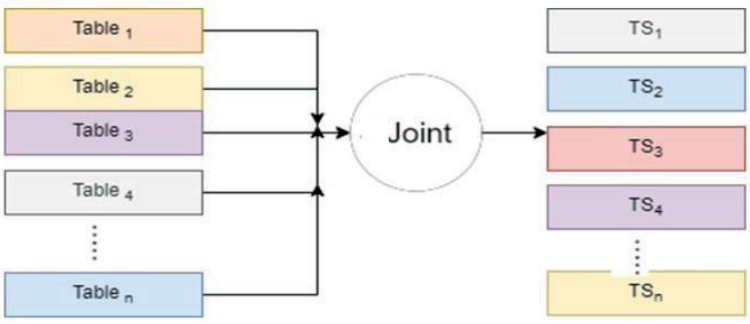

Figure 1. Generation of time series

The process of preparing time series is continuous and allows us to recover updated and reliable data. We use a synchronization module to recover the latest modifications, and then we clean the data and apply the necessary transformations before creating the time series.

Hybrid Model

In the context of forecasting, hybrid methods are defined as a sophisticated combination of statistical and machine learning methods that interact with each other. This combination is a new and different approach to merging several isolated forecasts, which separates hybrid methods from ensemble methods.

Statistical and machine learning methods have different strengths, as shown in table 1. Unfortunately, neither approach is perfect. The following table shows some advantages and disadvantages of the models in each family.

Table 1. Advantages and disadvantages of statistical approaches and machine learning algorithms

	Advantages	Disadvantages
Statistical approaches	Simple. understandable. Knowledge of existing data. Efficient with limited data.	Requires data linearity. No cross training.
Machine Learning Algorithms	No linearity precondition. Universal approximation. Cross training.	Very complex computation. Requires a very large amount of data.

The main idea is to transform or decompose the series, in order to isolate the different components of the time series. A time series yt, can be written as follows:

$$yt = mt + st + \varepsilon t \quad (1)$$

Three functions are obtained, the trend mt, the seasonal component st and the noise εt.
Also the time series can be written as two components, linear and non-linear.

$$yt = Mt + N \quad (2)$$

With Mt and N are the linear and non-linear components of the time series. The residuals are obtained from the processing of the linear modeling process of the ARIMA model:

$$rt = yt - \hat{Mt} \quad (3)$$

This residual is the entry point for the second model algorithm

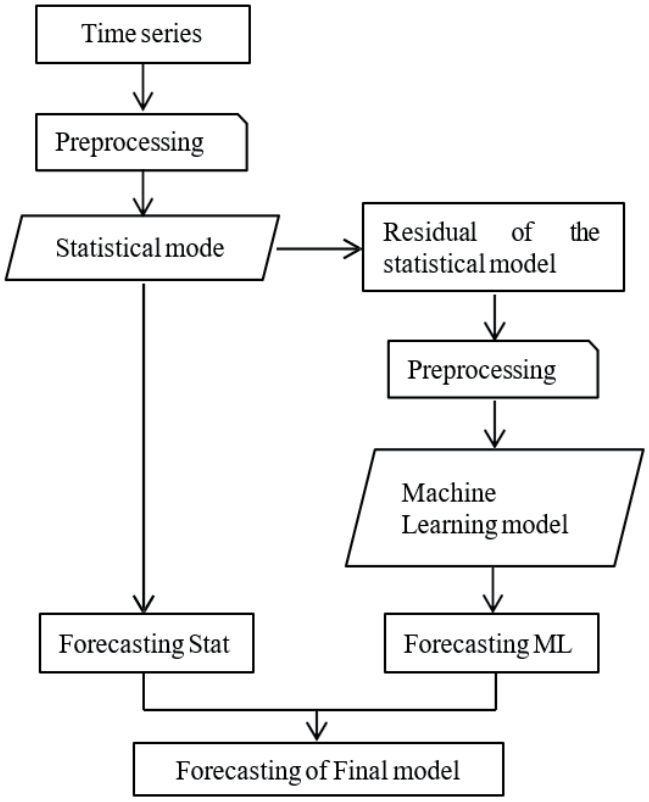

Figure 2. Hybrid forecasting model

The hybrid model used is a combination of two models that belong to two different families of approaches, statistical and machine Learning. Initially, the data is processed before being trained with a statistical model.

https://doi.org/10.56294/dm2025589

Then the residual is calculated, which is also pre-processed in order to make it adaptable to a supervised training problem, and then trained with a machine learning model, the two forecasts are then combined to produce the final forecast of the hybrid model, figure 2.

The hybrid model consists of two sub-models: a statistical model and a machine learning model. We have utilized the most popular and efficient algorithms within each respective family.

Statistical Model

In statistical models we have used ARIMA, which is widely known as the reference among forecasting models, the ARIMA model consists of three main components, namely, the "AR" for auto-regression, "I" for integration and "MA" for moving average, the main formula of an ARIMA is as follows:

$$Øp(B)(1 - B)dXt = θ0 + θq(B)et \quad (4)$$

Where:

Øp(B): is a characteristic polynomial of order p, θq(B) is a characteristic polynomial of order q, d is the order of the non-seasonal differencing, (1 − B) is the differencing operator.

xt: is the observation value at period t, θ_0 is a constant term.

et: is a white noise process.

Machine Learning Model

As the data that we are using is sequential and time dependent, we decided to use Recurrent Neural Networks (RNN). That choice was justified by the data type that we used, which is time series data, therefore RNN is the most adapted Machine Learning algorithm to our context.

Given the vanishing problem that presents RNN for data with long dependencies, we have decided to use the GRU algorithm, which is a simplified form of LSTM.[18] GRU is composed by two gates, figure 3, which allow it to memorize relevant information while removing non-relevant information.

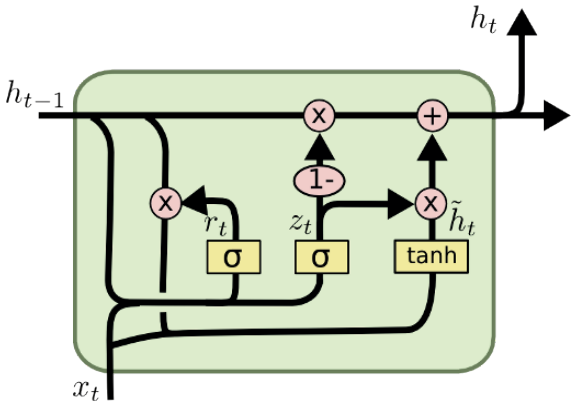

Figure 3. GRU Cell[22]

RESULTS

Dataset

For the experimentation of our approach, which we proposed for the forecasting of enrolments in higher education, 95 % of the data is extracted from the database used by APOGEE platform.

This application is based on an Oracle 11g database per university, each university has its own server, which manages the different institutions of the university. This architecture allows decision makers at the university level to have a global view on the management of enrolments and enrolment. The experimentation that we started in this study is based on the data of the university IBN ZOHR, but can be generalized for the other universities and also for the other types of courses existing in the higher education.

The notion of enrolled students is ambiguous, as it involves several types of enrollment:

New students: in a given academic year, these parameters refer to the number of students who apply for the first time at the institute and university.

Transfer students: in a given academic year, this parameter concerns the number of students enrolled for the first time at the institute but already enrolled at another institute the previous year.

Former students: is a student registered in the previous year in the institute and still registered in the current year.

We use the number of enrolled students registered on the system since 2003, which allows us to build

time series consisting of forty entries each. We have a history that covers twenty years with for each year the enrolments of the autumn and spring semester. In this experiment we focus on the basic degree in the public faculties of the university.

Finding

To simplify, we use only two time series, the first one is for Arabic Law in the Faculty of Legal, Economic and Social Sciences, the second one is for Hispanic Studies in the Faculty of Letters and Human Sciences.

FSJESAM_DrAr: Arab Law of the Faculty of Legal, Economic and Social Sciences.

FLSHA_ET_HISP: Hispanic Studies in the Faculty of Arts and Humanities.

We trained the time series using different approaches, including statistical, machine learning and hybrid.

The results we obtained for the ARIMA model using hyperparameter values obtained by the Gridsearch technique are as follows:

- p = 2.
- d = 1.
- q = 1.

Figure 4 and figure 5 show the results obtained for the two time series.

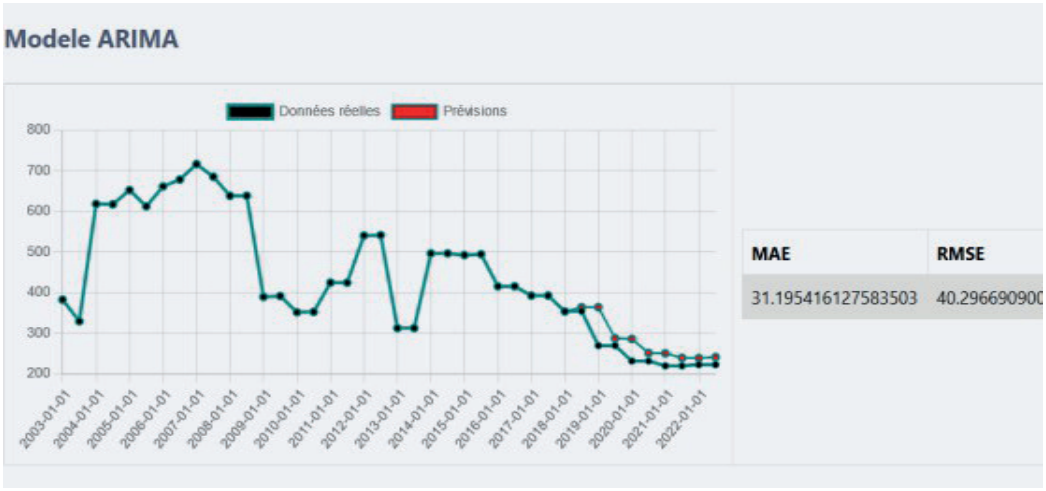

Figure 4. Forecasting model with ARIMA (FLSHA_ET_HISP)

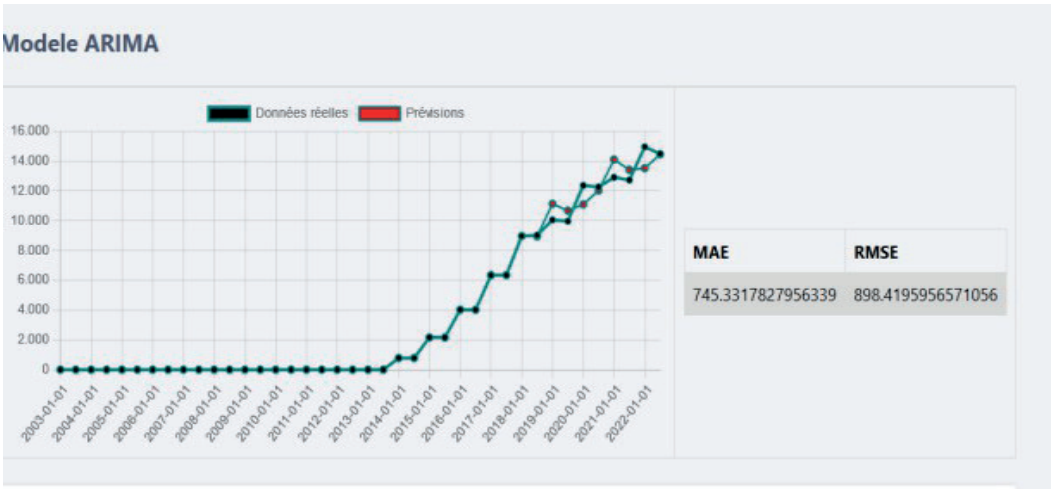

Figure 5. Forecasting model with ARIMA (FSJESAM_DrAr)

For GRU we use hyperparameter values that we have chosen by experimentation(Number of hidden layer, Number of units in a Dense Layer, etc).

Figure 6 and figure 7, show the results obtained for the two time series.

For hybrid model we used the two algorithms ARIMA and GRU to form the output model, so we used ARIMA to forecast the linear component and GRU to forecast the nonlinear component.

https://doi.org/10.56294/dm2025589

Figure 6. Forecasting model with GRU (FLSHA_ET_HISP)

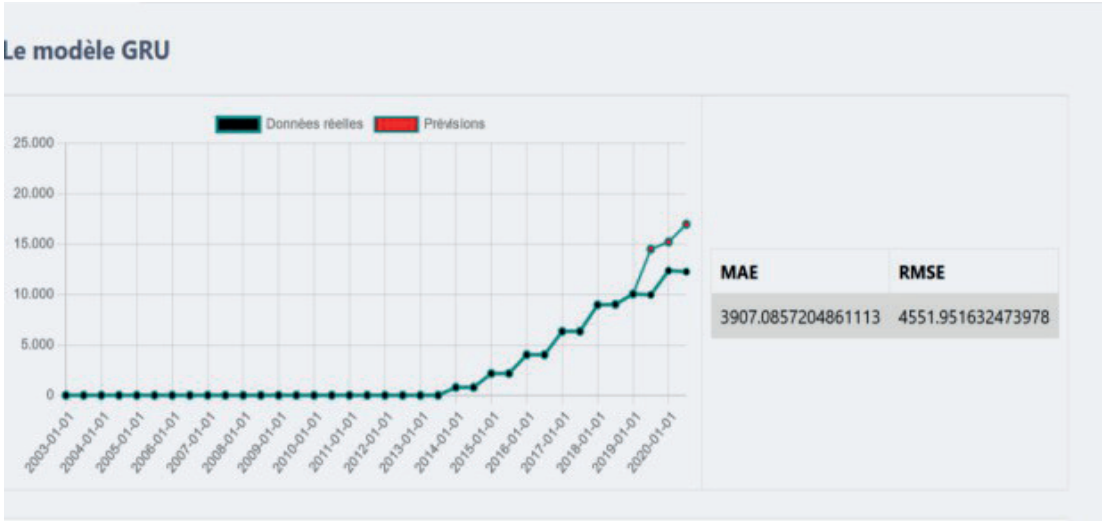

Figure 7. Forecasting model with GRU (FSJESAM_DrAr)

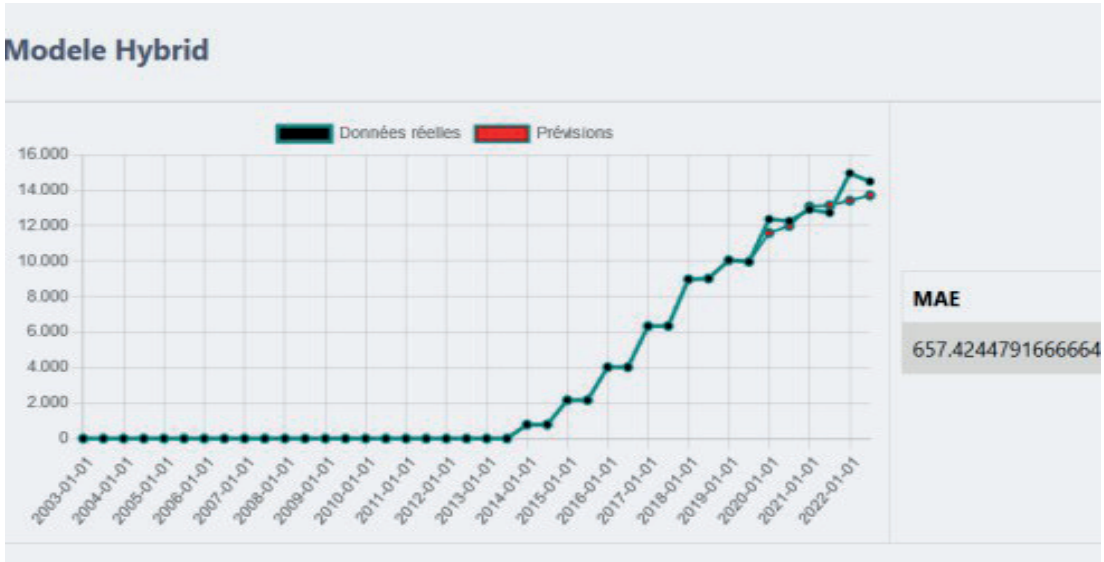

Figure 8. Forecasting with hybrid model ARIMA-GRU (FLSHA_ET_HISP)

We combined the two forecasts to obtain the final forecast of the hybrid model. Figure 8 and figure 9, show the results obtained for the time series.

https://doi.org/10.56294/dm2025589

Figure 9. Forecasting with hybrid model ARIMA-GRU (FSJESAM_DrAr)

Table 2. Model comparison

Models	Metric	FSJESAM_DrAr	FLSHA_ET_HISP
Hybrid(ARIMA_GRU)	MAE	657	12
	RMSE	799	15
Statistics (ARIMA)	MAE	745	31
	RMSE	898	40
Machine Learning(GRU)	MAE	3907	104
	RMSE	4551	119

DISCUSSION

For the time series FSJESAM_DrAr, the results obtained with ARIMA model are less accurate, but still acceptable for this type of course which has a large number of enrollees, one thing to consider for this last series is the poverty at the level of history, we only have a history dating from 2014.

For the GRU model, which belongs to the family of Machine Learning models and more specifically to the Deep Learning models, theoretically this model should provide better results than the Statistical models, but this is not the case, given the limited number of entries, even if the time series we use is one of the longest historical series available.

Regarding the GRU model training with the FSJESAM_DrAr time series, the accuracy of the results is very low, which is due to the very limited number of historical entries, about eight years (sixteen records). Machine Learning models are very strong with Big Data, but are very weak with small data, however we can combine them with statistical models to benefit from the strengths of both approaches and limit the weaknesses in both types of approach.

Contrary to simple approaches (statistical or Machine Learning), the hybrid approaches are composed of two simple models, generally a statistical model combined with a Machine Learning model.

We find that the accuracy for the hybrid ARIMA-GRU model is higher than the simple use of ARIMA or GRU

The hybrid model trained for the FLSHA_ET_HISP time series gives more accurate results compared to ARIMA, and GRU, which is normal, since the hybrid model is a combination of both models. The results obtained (see table 2) for both times series gives the best accuracy for hybrid model, followed by ARIMA and in the last place Machine Learning, which is logic, seen that series used in this paper are not larger time series.

Our approach can be further improved if trained on larger datasets from other universities. This limitation can be overcome in the future, as we currently only have an agreement with Ibn Zohr University. However, we are working on agreements with other universities and higher education institutions.

CONCLUSIONS

In this paper, a new hybrid model for time series is proposed to enhance accuracy in forecasting problems in higher education. We utilized a hybrid model composed of two algorithms: ARIMA for the statistical component

and GRU for the machine learning component. The approach involves combining the forecasts generated by these two algorithms. The hybrid model demonstrates superior results compared to both simple statistical and machine learning models. The findings obtained for Ibn Zohr University can potentially be extrapolated to other universities and institutions in higher education.

For future and perspective work, an interesting direction would be to extend the proposed hybrid model to incorporate multivariate time series data. By considering multiple variables simultaneously, such as enrollment numbers, course offerings, and student demographics, a more comprehensive analysis can be performed. This expansion would allow for a deeper understanding of the underlying patterns and relationships within the data, leading to improved forecasting accuracy.

BIBLIOGRAPHIC REFERENCES

1. E. Erdem y J. Shi, «ARMA based approaches for forecasting the tuple of wind speed and direction», Appl. Energy, vol. 88, n.o 4, pp. 1405-1414, abr. 2011, doi: 10.1016/j.apenergy.2010.10.031.

2. Z. Qian, Y. Pei, H. Zareipour, y N. Chen, «A review and discussion of decomposition-based hybrid models for wind energy forecasting applications», Appl. Energy, vol. 235, pp. 939-953, feb. 2019, doi: 10.1016/j.apenergy.2018.10.080.

3. P. J. Brockwell y R. A. Davis, «Forecasting Techniques», en Introduction to Time Series and Forecasting, en Springer Texts in Statistics. Cham: Springer International Publishing, 2016, pp. 309-321. doi: 10.1007/978-3-319-29854-2_10.

4. T.-C. Fu, F.-L. Chung, y C. Ng, «Financial Time Series Segmentation based on Specialized Binary Tree Representation.», ene. 2006, pp. 3-9.

5. L. Munkhdalai, M. Li, N. Theera-Umpon, S. Auephanwiriyakul, y K. Ryu, «VAR-GRU: A Hybrid Model for Multivariate Financial Time Series Prediction», 2020, pp. 322-332. doi: 10.1007/978-3-030-42058-1_27.

6. Z. Cui, J. Wu, W. Lian, y Y.-G. Wang, «A novel deep learning framework with a COVID-19 adjustment for electricity demand forecasting», Energy Rep., vol. 9, pp. 1887-1895, dic. 2023, doi: 10.1016/j.egyr.2023.01.019.

7. F. Kamalov, K. Rajab, A. K. Cherukuri, A. Elnagar, y M. Safaraliev, «Deep learning for Covid-19 forecasting: State-of-the-art review.», Neurocomputing, vol. 511, pp. 142-154, oct. 2022, doi: 10.1016/j.neucom.2022.09.005.

8. S. R. Srivastava, Y. K. Meena, y G. Singh, «Forecasting on Covid-19 infection waves using a rough set filter driven moving average models», Appl. Soft Comput., vol. 131, p. 109750, dic. 2022, doi: 10.1016/j.asoc.2022.109750.

9. H. Bousnguar, L. Najdi, y A. Battou, «Forecasting approaches in a higher education setting», Educ. Inf. Technol., pp. 1-19, ago. 2021, doi: 10.1007/s10639-021-10684-z.

10. A. Cruz et al., «Higher Education Institution (HEI) Enrollment Forecasting Using Data Mining Technique», Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, pp. 2060-2064, may 2020, doi: 10.30534/ijatcse/2020/179922020.

11. M. Edwards, «College Enrollment during Times of Economic Depression», J. High. Educ., vol. 3, n.o 1, pp. 11-16, ene. 1932, doi: 10.1080/00221546.1932.11775229.

12. Z. Ismail, «A decision support system for improving forecast using genetic algorithm and tabu search», ARPN J. Eng. Appl. Sci., vol. 3, n.o 3, Art. n.o 3, jun. 2008.

13. J. G. De Gooijer y R. J. Hyndman, «25 years of time series forecasting», Int. J. Forecast., vol. 22, n.o 3, pp. 443-473, ene. 2006, doi: 10.1016/j.ijforecast.2006.01.001.

14. R. G. Brown, R. F. Meyer, y D. A. D'Esopo, «The Fundamental Theorem of Exponential Smoothing», Oper. Res., vol. 9, n.o 5, pp. 673-687, 1961.A. Shadab, S. Said, y S. Ahmad, «Box–Jenkins multiplicative ARIMA modeling for prediction of solar radiation: a case study», Int. J. Energy Water Resour., vol. 3, n.o 4, pp. 305-318, dic. 2019, doi: 10.1007/s42108-019-00037-5.

15. Q. Mao, K. Zhang, W. Yan, y C. Cheng, «Forecasting the incidence of tuberculosis in China using the seasonal auto-regressive integrated moving average (SARIMA) model», J. Infect. Public Health, vol. 11, n.o 5, pp. 707-712, sep. 2018, doi: 10.1016/j.jiph.2018.04.009.

16. Z. C. Lipton, J. Berkowitz, y C. Elkan, «A Critical Review of Recurrent Neural Networks for Sequence Learning», ArXiv150600019 Cs, oct. 2015, Accedido: 11 de enero de 2021. [En línea]. Disponible en: http://arxiv.org/abs/1506.00019

17. A. H. Elsheikh, V. P. Katekar, O. L. Muskens, S. S. Deshmukh, M. Elaziz, y S. M. Dabour, «Utilization of LSTM neural network for water production forecasting of a stepped solar still with a corrugated absorber plate», Process Saf. Environ. Prot., vol. 148, pp. 273-282, abr. 2021, doi: 10.1016/j.psep.2020.09.068.

18. D. Lavrova, D. Zegzhda, y A. Yarmak, «Using GRU neural network for cyber-attack detection in automated process control systems», en 2019 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), jun. 2019, pp. 1-3. doi: 10.1109/BlackSeaCom.2019.8812818.

19. I. Livieris, E. Pintelas, S. Stavroyiannis, y P. Pintelas, «Ensemble Deep Learning Models for Forecasting Cryptocurrency Time-Series», Algorithms, vol. 13, p. 121, may 2020, doi: 10.3390/a13050121.

20. N. Kourentzes, F. Petropoulos, y J. R. Trapero, «Improving forecasting by estimating time series structural components across multiple frequencies», Int. J. Forecast., vol. 30, n.o 2, pp. 291-302, 2014.

21. H. Bousnguar, A. Battou, y L. Najdi, «Gated Recurrent units (GRU) for Time Series Forecasting in Higher Education», Int. J. Eng. Res. Technol., vol. 12, n.o 3, mar. 2023, doi: 10.17577/IJERTV12IS030091.

FINANCING

No financing.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Hassan BOUSNGUAR.
Data curation: Hassan BOUSNGUAR.
Research: Hassan BOUSNGUAR.
Methodology: Hassan BOUSNGUAR.
Project management: Hassan BOUSNGUAR, Lotfi NAJDI, Amal BATTOU.
Resources: Hassan BOUSNGUAR.
Software: Hassan BOUSNGUAR.
Supervision: Lotfi NAJDI, Amal BATTOU.
Validation: Lotfi NAJDI, Amal BATTOU.