Machines and Algorithms
http://www.knovell.org/mna

Research Article

Personalized Education Enhanced by AI and Predictive Analytics

Qazi Mudassar Ilyas[1], Sheikh Abdul Hannan[2] and Sadia Aziz[3]

[1]Department of Information Systems, College of Computer Sciences and Information Technology, King Faisal University, Al-Hasa, 31982, Saudi Arabia

[2]Virtual University of Pakistan, Lahore, 54000, Pakistan

[3]Department of Computer Science and Information Technology, La Trobe University, Melbourne, 3086, Australia

[*]Corresponding Author: Qazi Mudassar Ilyas. Email: qilyas@kfu.edu.sa

Received: 22 February 2024; Revised: 1 March, 2024; Accepted: 4 March 2024; Published: 14 March 2024

AID: 003-01-000035

Abstract: Unlike conventional learning, where an instructor delivers a set of topics in a predefined sequence, e-learning allows students to learn an arbitrary series of topics. With the availability of many learning profiles containing sequences of learning items followed by various learners, educational institutions might generate recommendations for future learners. Predictive analytics techniques can be used to analyze existing sequences of students to recommend a new arrangement to a new student. This study presents a time series analysis framework to generate such recommendations. The proposed framework uses clustering for dimensionality reduction. The clusters are passed through moving window transformations and fed into time series analysis models. Two models are used for time series forecasting: the Vector Auto-regression Model (VAR) and Auto-Regressive Integrated Moving Average (ARIMA) model. The output of such time series analysis models can be used to propose a sequence of learning items to a new learner. We used several evaluation metrics to compare the performance of the two models. The VAR model achieved better performance for median absolute error (0.008), prediction of change in direction (28.4), and coefficient of determination (-0.01). The respective values for the ARIMA model were 0.009, 27.1, and -2.984. The ARIMA model outperformed the VAR model for root mean squared error (0.120), mean absolute percent error (0.010), akaike information criterion (-3499.2), and Bayesian information criterion (-3435.1). The respective values for the VAR model were 0.163, 0.019, -108.8, and -108.3. These results suggest that the ARIMA model achieved a higher accuracy and better model fit. In contrast, the VAR model captured improved directional changes for model features and explained a larger portion of the variance in data.

Keywords: e-learning recommendations; predictive analytics; time series analysis; VAR and ARIMA models; clustering and dimensionality reduction;

1. Introduction

Education is considered to be a fundamental human right by UNESCO [1]. The COVID-19 pandemic has affected all walks of life, and educational institutions were among the most affected entities due to the fear of spreading novel coronavirus through students, especially the younger ones [2]. A UNESCO report estimates that about 70% of students are globally affected by the COVID-19 pandemic [3]. A natural response to this emergency was to exploit e-learning systems and continue education in distance learning mode [4], [5]. Several institutions have used e-learning systems primarily to augment conventional teaching strategies [6]. Notably, a report by Syngene Research predicted the e-learning market to reach $336.98 billion by 2026 in the world [7]. The COVID-19 pandemic acted as a catalyst for educational institutions to adopt a distance learning mode of education. E-learning offers several benefits to the students. First and foremost, it frees teachers and learners from time and space constraints [8]. Other services include higher scalability, reduced costs, and a richer experience [9].

Predictive analytics can be defined as the process of applying statistical, machine learning, and data mining techniques to identify hidden and useful patterns from large amounts of data and make predictions about future events [10]. Predictive analytics has recently gained much traction because of cheaper storage space and the ubiquity of information sources. Predictive analytics techniques are being used in countless domains today. The predictive analytics solutions encompass personal [11], business [12], government [13], and even defense [14] applications. Such techniques are widely used for decision-making, prediction, analysis, and unsupervised learning.

Asynchronous e-learning offers the freedom to repeat a lecture or other content as often as a learner needs. A learner also enjoys self-paced and self-organized learning experiences in asynchronous mode [15]. As content organization is critical in learning, institutions offering e-learning services wish to analyze various aspects of content usage by learners, such as the order in which different learning items were accessed and the number of times a la learner accessed a learning item. An institution can use predictive analytics techniques to perform such analysis and organize the learning content in a better way that can benefit the other learners [16]. This knowledge can enable an institution to personalize the e-learning systems according to a learner's needs.

The rest of the paper is organized as follows. We give a formal problem statement in Section 2. Section 3 discusses recent related works on predictive analytics for e-learning systems. The proposed predictive analytics approach is presented in Section 4 with details of the dataset used in the study, exploratory data analytics, data pre-processing, and model building. Section 5 gives results and discussion, and the paper is concluded in Section 6.

2. Problem Statement

The problem of predicting the next learning item for a learner who has already accessed some learning items in a given sequence can be formally stated as follows.

Assume the set L represents a set of all learners in a module.

$$L = \{L1, L2, L3, \dots Ln\}$$

The set I represents learning items in a learning module.

$$I = \{I1, I2, I3, \dots . . Iz\}$$

Assume a learner Li has accessed the learning items in the sequence S given below:

$$S = \{Is1, Is2, Is3, \dots . Ist\};\ S \subset I$$

The next learning item Lst+1 for the learner Li, is predicted from Y, a set of sequences of other learners, as given below.

$$Y = \{S1, S2, S3, \dots . St\}$$

3. Related Work

Predictive analytics techniques have been used successfully by several researchers in the domain of e-learning. This section briefly reviews some applications of predictive analytics in e-learning systems.

Stapel et al. reduced constraints for accurately predicting student performance factors by leveraging domain knowledge and a combination of representing the knowledge graph and event scopes [17]. It proceeds with particular scope classifiers combined with the ensemble to predict student performance learning objectives early. Koprinska et al. presented temporal predictions of students' performance metrics by depicting the effectiveness of data and its performance [18]. The authors analyzed datasets that included student submissions, assessment information, and activity data collected from various forums and online sources associated with campus program courses. They also declare their problem a multiclass classification problem, further divided into multiple examination performance-based levels. Arsad et al. used the artificial neural network-based model to predict individual program students' educational performance by taking the Grade Point Average of preparatory courses based on demographics; the Cumulative Grade Point Average is produced as output [19]. Yan et al. proposed partial multi-label learning with mutual teaching, which gives prediction networks and the corresponding teacher networks when assumed to study in collaboration and mutual learning and training procedure [20]. It repetitively declares labels of confidence matrix using multiple self-ensemble teacher-networks. Lin et al. presented multi-label learning for a sample and multiple labels using multiple support vector machines to determine the relationship, along with convergence analysis, examining computational complexity for performance metrics [21].

Essa & Ayad argue that students and their teachers' independence and openness give an edge to their diverse nature and behavior in predicting performance challenges [22]. They proposed a domain-specific decomposition of several web-based and online learning systems. Gómez et al. featured gender differences in students' aptitude and found that female students mainly attain a positive knowledge-seeking smartness compared to male students [23].

Berry presented a broad predictive analytics building model as an iterative process with several steps for student performance analysis. Considering the selected academia, they used this method to establish student success ratios[24]. Devasia et al. stated the system as a web-based application utilizing a Naïve-Bayesian mining algorithm to extract knowledge nuggets, experimenting with over 700 students and 19 attributes [25].

Phillips analyzed server tools in learning management systems (LMS), which offer online learning, including course content, quizzes, assignments, and online forums [26]. LMS provides easy-to-use for faculty members while easy-to-learn for students. Hooshyar et al. proposed an automated evaluation method comparing specific clustering methodologies with multiple internal/external performance metrics on different academic datasets varying in size and based on the University of Tartu Moodle system [27]. It extended the work by presenting the effects of the normalizing performance of clustering the methodologies and employed a multiple-criteria decision-making method. Educational predictive analytics provides a useful understanding of pedagogy among students and teachers by adapting rare academic datasets to helpful knowledge. Although a higher predicting accuracy model could be obtained by supervised learning, they are frequently inapplicable compared to educational data without class labels. [28], [29]. Tomasevic et al. presented a comparative analysis of supervised machine learning approaches to solve the task of student examination and predict their performance [30].

Shapiro et al. considered three categories of supervised machine-learning techniques: similarity-based, model-based, and probabilistic approaches [31]. The similarity-based method was used to predict exam performance, which is leveraged by discovering students with similar past performances. A second approach is a model-based approach driven by estimating implicit correlation among input learning data comprising the underlying model. The supervised probabilistic method was used to fit probability distribution features and their representation methods to find students at high risk of dropping out of courses. They were also evaluated for examination performance classification and regression activities.

4. Proposed Predictive Analytics Approach

As stated earlier, asynchronous e-learning offers self-paced and self-organized learning in which learners can define their sequence of learning items. The institution can analyze the usage of learning objects to improve their predefined sequence [15]. This learning can also be coupled with learners' analytics to provide a personalized learning experience for each learner [32]. A new learner's learning profile may be matched with past learners' learning profiles, and their learning sequence can be used to provide an enhanced learning experience for the new learner [9]. The technique used for this kind of analysis is called time series analysis. As the name implies, time series analysis learns from a sequence of temporal events to predict such sequential outcomes in the future. A time-series analysis requires a sequence of panel data to learn the sequence. Several models have been developed for performing time series analysis. These models fall into three main categories: autoregressive, moving average, and integrated models. These three classes of models have also been combined to propose hybrid models like Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), and Autoregressive Fractionally Integrated Moving Average (ARFIMA) models. An interested user may refer to several resources related to the topic [33], [34], [35], [36].

4.1. Dataset and Exploratory Data Analytics

We have used the Open University Learning Analytics dataset for this case study, a public dataset available for download from *https://analyse.kmi.open.ac.uk/open_dataset*. It consists of academic records and the students' personal information. The following data tables are available in this dataset:

1. Student Info
2. Courses
3. Student Registration
4. VLE
5. StudentVLE
6. Assessments
7. Student Assessments

We have used student info, VLE, and Student VLE tables for this case study, which are briefly described below.

The "Student Info" table contains the students' personal information. There are 12 attributes in this data table namely code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, and final_result. The attributes "code_module" and "code_presentation" represent a course in a module. Code_presentation consists of the year when the course is presented while appending B or J for course offerings in February and October, respectively. The rest of the attributes are obvious by their names.

The VLE table contains information about items in the virtual learning environment. The attributes in this table are id_site, code_module, code_presentation, activity_type, week_from, and week_to. While the other characteristics are apparent, activity_type needs a little more elaboration. It is used to categorize course material into one of 20 activities such as homepage, subpage, content, resource, forum, HTML activity, or external quiz.

The StudentVLE table is our main table, with over ten million records. It stores information about student interaction with the items in the virtual learning environment. The table contains code_module, code_presentation, id_student, id_site, date and sum_click attributes. The id_site attribute is the unique ID for every VLE item, while sum_click represents how many times a student accessed a given item.

4.2. Data pre-processing

The following pre-processing prepares the dataset for the time series analysis task.

First of all, the three tables are merged into one table. StudentVLE is considered to be the master table. The information about students and VLE is extracted from respective tables and added to this master table. Every student's sequence in which they interacted with the items is preserved using the data attribute in the StudenVLE table.

Data can be prepared for time series analysis either in long or wide format. Long format holds one item accessed by a student in one record, while wide format appends all items accessed by a student one after the other in a single record. We have used a long format as several student interactions are not uniform for all students.

Close observation of the data reveals that the range of values for different attributes is not uniform. This may have the undesirable consequence of a variable with large values dictating a learning algorithm's output. All variables have been normalized to overcome this issue.

Finally, the number of records was reduced due to a limitation of the tool for calculating the Silhouette coefficient in clustering.

4.3. Model Building

Figure 1 below shows the model used to perform a time-series analysis on the sequence of learning items in the virtual learning environment. The workflow of the model is described below:

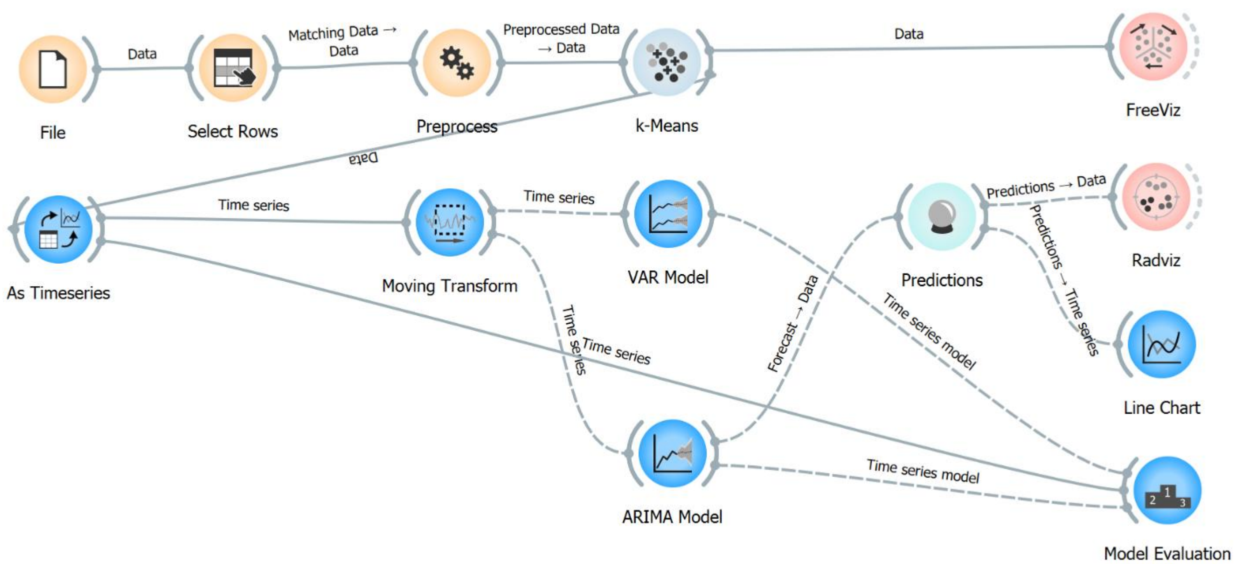

Figure 1: Proposed framework

1. First, data is imported, and the number of records is reduced, as described earlier.
2. The pre-processing step is used to normalize the attributes as described above.
3. Clustering is used as a dimensionality reduction technique, and the clusters are used to improve the performance of the time series process. The algorithm used for clustering is the k-mean clustering algorithm.
4. The output of clustering is visualized using the FreeViz chart.
5. Data, now in the form of clusters, is passed on to the time series process.
6. "Moving transform" is used to perform aggregation operations by applying rolling window functions.
7. The transformed data is ready for time series analysis. This data is fed into two time-series models: the Vector Autoregression Model (VAR) and the Auto-Regressive Integrated Moving Average (ARIMA) model.
8. The predictions of both models are visualized using RadViz and line charts.
9. Model performances are compared, and the results are exported in the final step.

5. Results and discussion

As stated above, clustering is used as a dimensionality reduction process. The k-means clustering algorithm is used to form data clusters. Figure 2 presents the output of the clustering process.

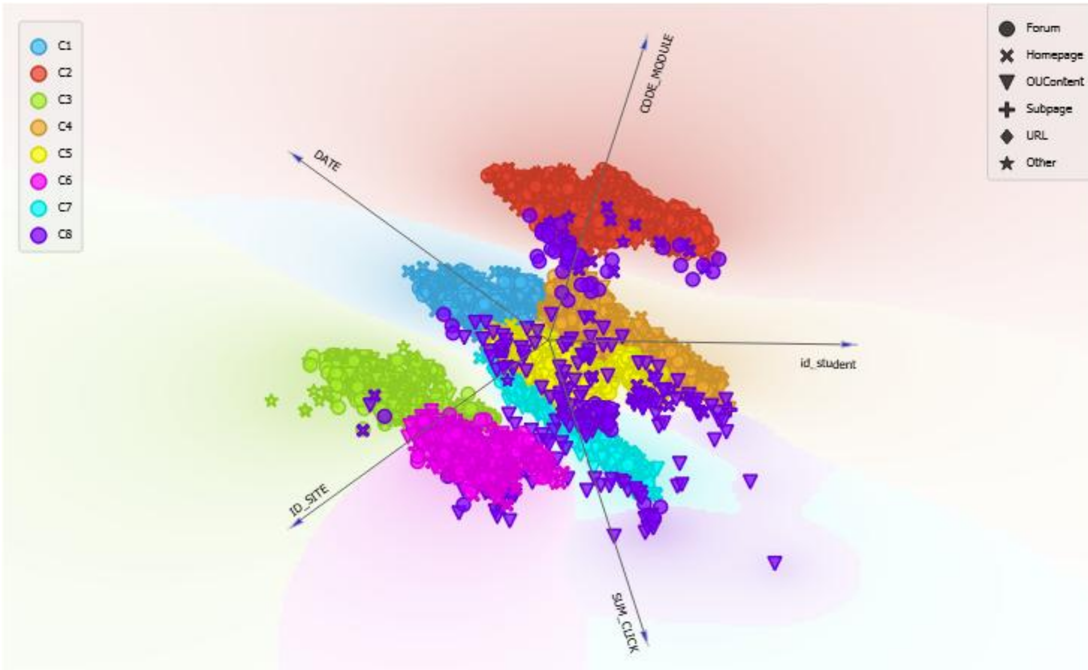

Figure 2: Data set divided into eight clusters

A total of eight clusters are produced. Further cluster analysis reveals a substantial similarity between URL and OUContent activity types, which suggests that these activity types share similar characteristics. A justification for this high similarity is the use of descriptive URL identifiers for content. Similarly, there is a high overlap between Forum and OUContent because of the discussion of topics on the forum.

The sequence predictions produced by the VAR model can be visualized in Figures 3 and 4. As "id_site" has been used as the sequential attribute, and this attribute's values are very close to each other, the sequential output also looks like a cluster. An exploded version of the chart may provide better visualization. It can also be noted in Figure 4 that the predictions converge after some time.

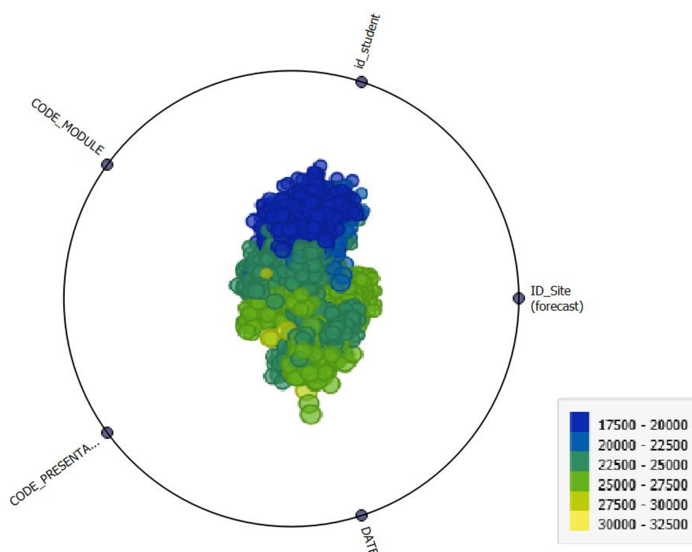

Figure 3: A visualization of predictions by the VAR model

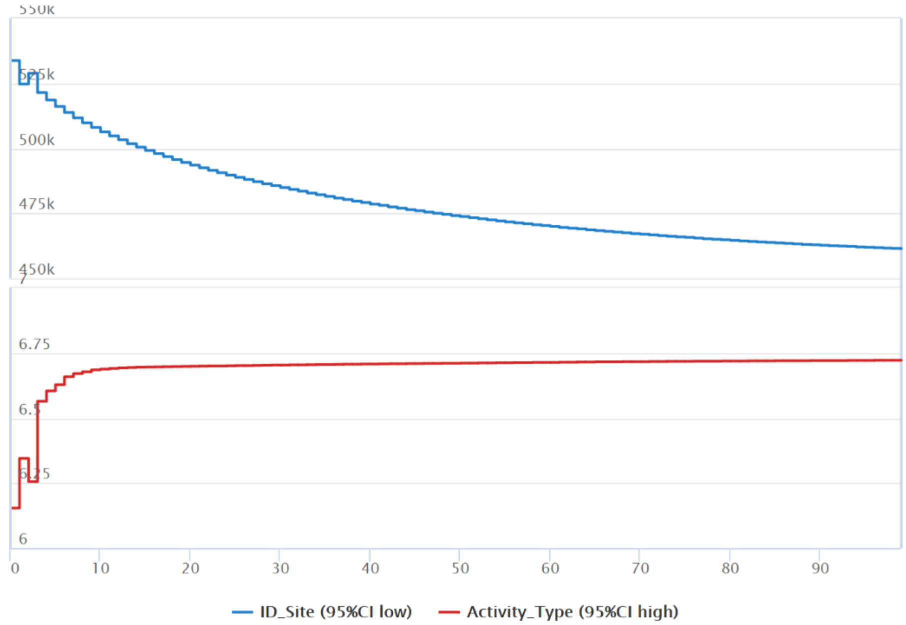

Figure 4: Step line charts showing predictions for ID_Site and Activity type with 95% confidence interval

A comparison of the VAR and ARIMA models is presented in Table 1. The evaluation measure used for comparison includes Root Mean Squared Error (RMSE), Median Absolute Error (MAE), Mean Absolute Percent Error (MAPE), Prediction of Change in Direction (POCID), Coefficient of Determination (R^2), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). As shown in the results, the performance of VAR and ARIMA is comparable for some measures, while each model achieved better results for some metrics and performed poorly for others. The ARIMA model outperformed the VAR model in terms of RMSE (0.120 vs 0.163) and MAPE (0.010 vs 0.019). VAR achieved slightly better MAE with 0.008 compared to 0.009 for ARIMA. The VAR model also achieved slightly better performance regarding POCID (28.4 vs 27.1), indicating slightly better performance in capturing directional changes for model features. VAR also achieved significantly better performance for R^2 (-0.012 vs. -2.984), which shows that the VAR model explains a larger portion of the variance in data compared to the ARIMA model. The ARIMA model achieved a very low score for AIC (-3499.2) compared to the VAR model (-108.8), showing a better-fit model that balances goodness of fit and complexity. The ARIMA model also outperformed the VAR for BIC (-3435.1 vs. -108.3), indicating a better-fit model.

Table 1: Results of sequence prediction by VAR and ARIMA models

Model	RMSE	MAE	MAPE	POCID	R^2	AIC	BIC
VAR	0.163	0.008	0.019	28.4	-0.012	-108.8	-108.3
ARIMA	0.120	0.009	0.010	27.1	-2.984	-3499.2	-3435.1

6. Conclusion

In today's age of personalized e-services, it is natural to offer e-learning services to learners according to their specific needs. One possible way of personalizing e-learning systems is to recommend a sequence of learning items. This study presents a case study to perform a time-series analysis of previous learners' learning object sequences to recommend a personalized arrangement to a new learner. Institutions can use it to provide a better quality of service, an improved learning experience, and a higher satisfaction rate among students. One may think of further enhancing the proposed framework by customizing the learning content. Other possible extensions include personalized tests, personalized assignments, and course recommendations.

References

[1] UNESCO, "United Nations Decade of Education for Sustainable Development (2005-2014): International Implementation Scheme," *Sustainable Development*, 2005.

[2] Andersen, Kristian G., Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes, and Robert F. Garry. "The proximal origin of SARS-CoV-2." *Nature medicine* 26, no. 4 (2020): 450-452.

[3] UNESCO, "COVID-19 Educational Disruption and Response," *Unesco.Org*, 2020.

[4] C. for S. N. CoSN, "COVID-19 Response : Preparing to Take School Online," *(Consortium for School Networking).*, no. March, 2020.

[5] G. Tam and D. El-Azar, "3 Ways the Coronavirus Pandemic Could Reshape Education," *World Economic Forum*, 2020.

[6] Halkiopoulos, Constantinos, and Evgenia Gkintoni. "Leveraging AI in e-learning: Personalized learning and adaptive assessment through cognitive neuropsychology—A systematic analysis." *Electronics* 13, no. 18 (2024): 3762.

[7] Syngene Research LLP, "Global E-Learning Market Analysis 2019," 2019.

[8] Gligorea, Ilie, Marius Cioca, Romana Oancea, Andra-Teodora Gorski, Hortensia Gorski, and Paul Tudorache. "Adaptive learning using artificial intelligence in e-learning: a literature review." *Education Sciences* 13, no. 12 (2023): 1216.

[9] Ozyurt, Ozcan, Hacer Ozyurt, and Deepti Mishra. "Uncovering the educational data mining landscape and future perspective: A comprehensive analysis." *Ieee Access* 11 (2023): 120192-120208.

[10] Dada, Michael Ayorinde, Johnson Sunday Oliha, Michael Tega Majemite, Alexander Obaigbena, and Preye Winston Biu. "A review of predictive analytics in the exploration and management of us geological resources." *Engineering Science & Technology Journal* 5, no. 2 (2024): 313-337.

[11] Garett, Renee, and Sean D. Young. "The role of artificial intelligence and predictive analytics in social audio and broader behavioral research." *Decision Analytics Journal* 6 (2023): 100187.

[12] Dehankar, Pooja, A. Amudha, S. Jayasudha, R. Pallavi, Devendra Kumar Doda, and Nelson Mandela. "Predictive Analytics Powered by Artificial Intelligence." In *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, pp. 1-5. IEEE, 2023.

[13] Qadadeh, Wafa, and Sherief Abdallah. "Governmental data analytics: an agile framework development and a real world data analytics case study." *International Journal of Agile Systems and Management* 16, no. 3 (2023): 289-316.

[14] Khan, Fahad Ali, Gang Li, Anam Nawaz Khan, Qazi Waqas Khan, Myriam Hadjouni, and Hela Elmannai. "AI-Driven Counter-Terrorism: Enhancing Global Security Through Advanced Predictive Analytics." *IEEE Access* 11 (2023): 135864-135879.

[15] Bayly-Castaneda, Karla, María Soledad Ramirez-Montoya, and Adelina Morita-Alexander. "Crafting personalized learning paths with AI for lifelong learning: a systematic literature review." In *Frontiers in Education*, vol. 9, p. 1424386. Frontiers Media SA, 2024.

[16] Ayeni, Oyebola Olusola, Nancy Mohd Al Hamad, Onyebuchi Nneamaka Chisom, Blessing Osawaru, and Ololade Elizabeth Adewusi. "AI in education: A review of personalized learning and educational technology." *GSC Advanced Research and Reviews* 18, no. 2 (2024): 261-271.

[17] Stapel, Martin, Zhilin Zheng, and Niels Pinkwart. "An Ensemble Method to Predict Student Performance in an Online Math Learning Environment." *International Educational Data Mining Society* (2016).

[18] Koprinska, Irena, Joshua Stretton, and Kalina Yacef. "Predicting student performance from multiple data sources." In *Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings 17*, pp. 678-681. Springer International Publishing, 2015.

[19] Arsad, Pauziah Mohd, and Norlida Buniyamin. "A neural network students' performance prediction model (NNSPPM)." In *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, pp. 1-5. IEEE, 2013.

[20] Yan, Yan, Shining Li, and Lei Feng. "Partial multi-label learning with mutual teaching." *Knowledge-Based Systems* 212 (2021): 106624.

[21] Lin, Luyue, Bo Liu, Xin Zheng, Yanshan Xiao, Zhijing Liu, and Hao Cai. "An efficient multi-label learning method with label projection." *Knowledge-Based Systems* 207 (2020): 106298.

[22] Essa, Alfred, and Hanan Ayad. "Student success system: risk analytics and data visualization using ensembles of predictive models." In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pp. 158-161. 2012.

[23] González-Gómez, Francisco, Jorge Guardiola, Óscar Martín Rodríguez, and Miguel Ángel Montero Alonso. "Gender differences in e-learning satisfaction." *Computers & Education* 58, no. 1 (2012): 283-290.

[24] Berry, Michael A., and Gordon S. Linoff. "Mastering data mining: The art and science of customer relationship management." *Industrial Management & Data Systems* 100, no. 5 (2000): 245-246.

[25] Devasia, Tismy, T. P. Vinushree, and Vinayak Hegde. "Prediction of students performance using Educational Data Mining." In *2016 international conference on data mining and advanced computing (SAPIENCE)*, pp. 91-95. IEEE, 2016.

[26] Phillips, Rob. "Tools used in Learning Management Systems: analysis of WebCT usage logs." In *Proceedings of the 23rd Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education. Sydney University Press*, pp. 663-673. 2006.

[27] Hooshyar, Danial, Yeongwook Yang, Margus Pedaste, and Yueh-Min Huang. "Clustering algorithms in an educational context: An automatic comparative approach." *IEEE Access* 8 (2020): 146994-147014.

[28] Dutt, Ashish, Maizatul Akmar Ismail, and Tutut Herawan. "A systematic review on educational data mining." *Ieee Access* 5 (2017): 15991-16005.

[29] C. Anuradha, T. Velmurugan, R. Anandavally, and A. Professor, "Clustering Algorithms in Educational Data Mining: A Review…C.Anuradha et al., CLUSTERING ALGORITHMS IN EDUCATIONAL DATA MINING: A REVIEW," *International Journal of Power Control and Computation(IJPCSC)*, 2015.

[30] Tomasevic, Nikola, Nikola Gvozdenovic, and Sanja Vranes. "An overview and comparison of supervised data mining techniques for student exam performance prediction." *Computers & education* 143 (2020): 103676.

[31] Shapiro, Heather B., Clara H. Lee, Noelle E. Wyman Roth, Kun Li, Mine Çetinkaya-Rundel, and Dorian A. Canelas. "Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers." *Computers & Education* 110 (2017): 35-50.

[32] Sahu, Sourav, Neelamadhab Padhy, Satyam Mohapatra, Amrutansu Patra, Anurag Kumar, and Rajiv Kumar Choudhary. "Educational Data Mining for Personalized Learning: A Sentiment Analysis and Process Control Perspective." In *Proceedings*, vol. 105, no. 1, p. 77. MDPI, 2024.

[33] Kirchgässner, Gebhard, Jürgen Wolters, and Uwe Hassler. *Introduction to modern time series analysis*. Springer Science & Business Media, 2012.

[34] Wong, Chun Shan, and Wai Keung Li. "On a mixture autoregressive model." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 62, no. 1 (2000): 95-115.

[35] Weiß, Christian H. *An introduction to discrete-valued time series*. John Wiley & Sons, 2018.

[36] Koosha, Mohaddeseh, Ghazaleh Khodabandelou, and Mohammad Mehdi Ebadzadeh. "A hierarchical estimation of multi-modal distribution programming for regression problems." *Knowledge-Based Systems* 260 (2023): 110129.