Journal of Computing in Higher Education (2021) 33:545–550
https://doi.org/10.1007/s12528-021-09299-7

Data science for analyzing and improving educational processes

Shadi Aljawarneh[1] · Juan A. Lara[2]

Accepted: 20 October 2021 / Published online: 29 October 2021

Abstract
In this full review paper, the recent emerging trends in Educational Data Science have been reviewed and explored to address the recent topics and contributions in the era of Smart Education. This includes a set of rigorously reviewed world-class manuscripts addressing and detailing state-of-the-art, frameworks and techniques research projects in the area of Data Science applied to Education, using different approaches such as Information Fusion, Soft Computing, Machine Learning, and Internet of Things, among others. Based on this systematic review, we have put some recommendations and suggestions for researchers, practitioners and scholars to improve their research quality in this area.

Keywords Educational data science · Learning analytics · Educational data mining · Educational processes

Introduction to the domain of research

The term Data Science (DS) refers to an interdisciplinary field that involves a series of methods, processes and systems, with the aim of extracting knowledge from data. DS, which is a discipline very related to Computing, has proved to be of great application in very different domains, particularly Education (Klašnja-Milićević et al., 2017). In Educational environment, there are lots of learning-related processes involved, and great amounts of potential rich data are generated in educational institutions continuously. In order to extract knowledge from those data for a better

✉ Shadi Aljawarneh
saaljawarneh@just.edu.jo

Juan A. Lara
juanalfonso.lara@udima.es

[1] Software Engineering Department, Jordan University of Science and Technology, Irbid, Jordan

[2] Escuela de Ciencias Técnicas e Ingeniería, Madrid Open University, UDIMA, Ctra. de la Coruña, km 38.500 – Vía de Servicio, 15 - 28400, Collado Villalba, Madrid, Spain

understanding or learning-related processes, the use of DS approach seems to be useful and necessary (Mitrofanova et al., 2019).

The application of DS in the field of Education may result of great interests for involved stakeholders (students, instructors, institutions, …) since the extracted knowledge from educational data would be useful to deal with educational problems such as students' performance improvement, high churning rates in educational institutions, learning delays, and so on. There are a series of disciplines related to Educational Data Science, such as Educational Data Mining and Learning Analytics (Romero & Ventura, 2020), and all of them are of importance for this special issue.

In this introductory paper, Sect. 2 includes the summaries of the selected papers. Section 3 includes a set of recommendations for researchers, practitioners and scholars to improve their research quality in this area. In Sect. 4, the conclusions have been dawn.

Related work: the selected papers

The purpose of this special issue is to present original contributions of studies on the application of DS techniques in order to extract knowledge of interest for educational stakeholders as long as the analysed data represent a particular educational process and the knowledge extracted is used to improve that process in some way. We have considered papers that include discussions of the implementation of software and/or hardware approaches that also focus on the implications for the improvement of any learning process. Priority has been be given to papers that demonstrate a strong grounding in learning theory and/or rigorous educational research design. We have considered studies focused on tertiary and further education of any type (e-learning, blended and traditional education). All accepted works include an exhaustive validation and include extraordinarily new ideas in the area.

The special issue includes 10 papers, which have been subject to a rigorous peer-review process. Each paper has been reviewed by three independent experts. The rest of this section includes a summary of the selected papers.

The research presented in "Multilayered-Quality Education Ecosystem (MQEE): An Intelligent Education Modal for Sustainable Quality Education", by Verma et al., intends to unfold some hidden parameters that are affecting the quality education ecosystem (QEE). Academic loafing, unawareness, non-participation, dissatisfaction, and incomprehensibility are the main parameters under this study. A set of hypothesis and surveys are exhibited to study the behaviour of these parameters on quality education at the institution level. The bidirectional weighted sum method is deployed for precise and accurate results regarding boundary value analysis of the survey. The association between parameters understudy and quality education is illustrated with correlation and scatter diagrams. Academic loafing, the hidden and unintended rudiment that affects the QEE is also defined, intended and explored in this work.

In the paper "Improving prediction of students' performance in Intelligent Tutoring Systems using attribute selection and ensembles of different multimodal data sources", Chango et al. intend to predict university students' learning performance using different sources of performance and multimodal data from an Intelligent Tutoring System. They collected and preprocessed data from 40 students from different multimodal sources: learning strategies from system logs, emotions from videos of facial expressions, allocation and fixations of attention from eye tracking, and performance on post-tests of domain knowledge. Their objective is to test whether the prediction could be improved by using attribute selection and classification ensembles.

In “Automated Text Detection from Big Data Scene Videos in Higher Education”, Manasa et al. employed a novel approach to clean up the video frames to feed a neural network model based on region proposal network (RPN) with convolutional neural networks by finding appropriate anchor ratios to extract the text candidates. The trained their model with extracted frames to predict for the test videos. The proposed method is evaluated on ICDAR Video text benchmark datasets and few publicly available test datasets to achieve high recall.

In the paper “Improve teaching with modalities and collaborative groups in an LMS: an analysis of monitoring using visualisation techniques”, by Sáiz-Manzanares et al., the main objective is to test the effectiveness of three teaching modalities (all using Online Project-based Learning -OPBL- and Flipped Classroom experiences and differing in the use of virtual laboratories and Intelligent Personal Assistant -IPA-) on Moodle behaviour and student performance taking into account the covariate "collaborative group". Both quantitative and qualitative research methods were used. With regard to the quantitative analysis, differences were found in student behaviour in Moodle and in learning outcomes, with respect to teaching modalities that included virtual laboratories. Similarly, the qualitative study also analysed the behaviour patterns found in each collaborative group in the three teaching modalities studied.

The study titled “Fuzzy-based Active Learning for Predicting Student Academic Performance using autoML: a step-wise approach”, by Tsiakmaki et al., introduces a fuzzy-based active learning method for predicting students’ academic performance which combines, in a modular way, autoML practices. A lot of experiments were carried out, revealing the efficiency of the proposed method for the accurate prediction of students at risk of failure. These insights may have the potential to support the learning experience and be useful the wider science of learning.

In the paper “Peer Assessment Using Soft Computing Techniques”, by Pinargote-Ortega et al., a peer assessment scenario was applied at the Universidad Técnica de Manabí (Ecuador). Students and professors evaluate some works through rubrics, assign a numerical score, and textual feedback grounding the reasons why such numerical score is determined. Interesting scenario to detect inaccuracy between both assessments. It is proposed a model with soft computing techniques to detect inaccuracy and reduce the workload of the professor in the correction process.

In “A Novel Automated Essay Scoring Approach for Reliable Higher Educational Assessments”, Beseiso et al. present a transformer-based neural network model for improved Automatic Essay Scoring performance using Bi-LSTM (Bidirectional Long Short-Term Memory) and RoBERTa language model based on Kaggle’s ASAP (Automated Student Assessment Prize) dataset. The proposed model uses

Bi-LSTM model over pre-trained RoBERTa language model to address the coherency issue in essays that is ignored by traditional essay scoring methods, including traditional Natural Language Processing pipelines, deep learning-based methods, a mixture of both. The comparison of the experimental results on essay scoring with human raters concludes that the proposed model outperforms the existing methods in essay scoring in terms of QWK (Quadratic Weighted Kappa) score.

The main goal of the research presented in the paper “Personalized training model for organizing blended and lifelong distance learning courses and its effectiveness in Higher Education”, by Bekmanova et al., is to improve the personification of learning in higher education. The proposed flexible model for organizing blended and distance learning in higher education involves the creation of an individual learning path through testing students before the start of training. Based on the learning outcomes, the student is credited to the learning path. The training path consists of mandatory and additional modules for training; additional modules can be skipped after successfully passing the test, without studying these modules. The paper examines the composition of intelligent learning systems: student model, learning model and interface model.

In the paper “IoT Text Analytics in Smart Education and Beyond”, Mohammed et al. highlight the main components of IoT analytics, along with a comprehensive background of text analytics used techniques and applications. This paper provides a comprehensive survey and comparison of the leveraged IoT Text Analytics models and methods in Smart Education and many other applications.

Finally, in “A Framework to Capture the Dependency between prerequisite and Advanced Courses in Higher Education”, Hriez & Al-Naymat propose a new graph mining algorithm combined with statistical analysis to reveal the dependency relationships between Course Learning Outcomes (CLOs) of prerequisite and advanced courses. In addition, a new model is built to predict students’ performance in the advanced courses based on prerequisites. The evaluation proves that the proposed algorithm is accurate, efficient, effective, and applicable to real-world graphs more than the traditional algorithm.

Discussions and recommendations

A number of recommendations have been suggested to improve the research in this field as follows:

- The papers selected for inclusion in this special issue have described a number of data science techniques for extracting knowledge for educational data. However, the knowledge extracted is only applicable to the problem addressed. It is desirable to obtain general models that can be applied in other scenarios (López-Zambrano et al., 2021).
- Most research is focused on analyzing only one source of educational data. However, in current smart classrooms, a lot of different multi-source and multi-modal data are recorded and it can be very interesting to fuse those data in order to obtain richer and more accurate models.

- Many DS approaches generate models that are hard to interpret, in spite of the fact that they can obtain very accurate results. However, interpretability is a requirement in Education sometimes, since it helps understand the learning processes and, therefore, improve them by interventions.
- Current educational models are designed on the premise of ubiquity, particularly in the event of emergencies such as the one caused by the Covid-19 pandemic (Maatuk et al., 2021). In this scenario, the student needs to be able to self-regulate his or her learning, which is hard sometimes. It is very important to count on tools for personalized learning that adapt to each student depending on his or her emotions at a certain moment. The use of virtual affective agents is a promising line nowadays.

Conclusions

In this special issue, 10 selected papers have been included that present important advancements in the area of Educational Data Science. The selected papers include interesting studies about the development of this area, works about promising existing technologies and outstanding research about theories and methods that will play a crucial role in the future of this discipline.

As guest editors, we are aware of the fact that this issue cannot completely cover all the advancements in this area, but we expect that this special issue can stimulate further research in the domain of Educational Data Science.

Acknowledgements We would like to thank the referees who have reviewed the papers for providing a valuable feedback to authors. We would also like to thank the authors for their manuscripts that represent an important contribution to the existing knowledge in the area. Finally, we would like to thank Professor Stephanie L. Moore, Editor-in-Chief, and editorial assistants of Journal of Computing in Higher Education for their support during the preparation of this issue.

References

Klašnja-Milićević, A., Ivanović, M., & Budimac, Z. (2017). Data science in education: Big data and learning analytics. *Computer Applications in Engineering Education, 25*, 1066–1078. https://doi.org/10.1002/cae.21844

López-Zambrano, J., Lara, J. A., & Romero, C. (2021). Improving the portability of predicting students' performance models by using ontologies. *Journal of Computing in Higher Education*. https://doi.org/10.1007/s12528-021-09273-3

Maatuk, A. M., Elberkawi, E. K., Aljawarneh, S., Rashaideh, H., & Alharbi, H. (2021). The COVID-19 pandemic and E-learning: Challenges and opportunities from the perspective of students and instructors. *Journal of Computing in Higher Education*. https://doi.org/10.1007/s12528-021-09274-2

Mitrofanova, Y. S., Sherstobitova, A. A., & Filippova, O. A. (2019). Modeling smart learning processes based on educational data mining tools. In V. Uskov, R. Howlett, & L. Jain (Eds.), *Smart Education and e-Learning 2019. Smart Innovation, Systems and Technologies.* (Vol. 144). Springer. https://doi.org/10.1007/978-981-13-8260-4_49

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wires Data Mining and Knowledge Discovery, 10*, e1355. https://doi.org/10.1002/widm.1355

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Shadi Aljawarneh is a full professor, Software Engineering, at the Jordan University of Science and Technology, Jordan; visiting professor, Concordia University, Montreal, Canada. His research is centered in software engineering, web and network security, e-learning, AI, Machine Learning, and Cloud Computing. Aljawarneh has presented at and been on the organizing committees for a high number of international conferences and esteemed Journals and is a board member of the International Community for ACM, Jordan ACM Chapter, ACS, and IEEE. A number of his papers have been selected as "Best Papers" in conferences and journals.

Juan A. Lara is Associate Professor and Research Scientist at Madrid Open University, UDIMA, Spain. He is currently member of Department of Computer Science. He holds a Ph.D. in Computer Science and two Post Graduate Masters in Information Technologies and Emerging Technologies to Develop Complex Software Systems from Technical University of Madrid, Spain. He is author of more than 30 papers published in international impact journals. His research interests in computer science include data mining, knowledge discovery in databases, data fusion, artificial intelligence and e-learning.