Data Mining in Education

by Carlotta Amaduzzi

16th August 2021

Prepared for:

Prof. Donghwa Kim

Data Science 603

University of Maryland Baltimore County

Big Data Analysis, the collection and analysis of large quantities of information, is now part of our modern lives. Technologies are being developed to allow greater numbers of individuals to learn, quickly, how to interact and use large quantities of information with the objective of learning from them. What used to be known as statistical descriptive and inferential tools are evolving into processes and techniques openly declared to be used to identify patterns and generalizations, from which to learn and extract the foundations for better decisions and policies. Data Mining in the field of education has no smaller aspirations. In the field of Higher Education Data Mining is being viewed, and is being used more and more, as a tool that can aid in solving a variety of challenges, inherent to the learners themselves, but also the institutions involved in their preparation, the individual teachers responsible for their students' success, and beyond. With this paper I intend to offer a brief overview of how data mining in education is being used, why it has come to the forefront of discussion, and what challenges it poses. I will stive to conclude with a personal perspective on the matter.

Data Mining in education

Data Mining applied in the field of education has been receiving greater attention in more recent years and is defined on The Educational Data Mining community website as "an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in."

There are primarily two great areas of focus, on the one hand, the students (their learning and potential for success) and, on the other hand, the institutions (their ability to support and retain

students, to predict and intervene to support students' success, and to develop learning strategies and technology better able to respond to today's rapidly changing world).

While these are the two primary areas of interest it is important to keep in mind that the impact of data mining based policies in the field of education spill over well beyond having an impact on students and educational institutions alone, given the central role education has for our society and the increased demands being placed on education with generally declining support from government and private funding.

At the foundation of data mining in education is the realization that students leave behind a trail of information as members of education communities and when interacting with technology based learning systems (whether Learning Managements Systems or not) that can reveal a lot about their emotions, social connections, intentions, goals, habits, and abilities, and that this information, if properly analyzed can reveal the best strategies possible to ensure their long term success, while opening big opportunities for profit recovery and profit creation for the institutions themselves.

According to an OECD report from 2013 on Education, adequate analysis of this data could be the avenue through which institutions of higher education can find answers to their current business models and find ways to respond to the challenges they are facing in recent times. The data collected within higher education institutions can lead to greater insight at the institutional level which in turn can lead to interventions across departments and academic programs to reduce organizations' structural inefficiencies, but it can also lead to greater insight on institutions' IT systems bringing about greater integration between divisions within the institutions themselves and, more simply, greater organizational efficiencies. It can enhance the understanding of academic programs at the same time, helping institutions address program

specific challenges, and, finally, it can help develop learning analytics to explicitly focus on learning systems' effectiveness and ideally develop individualized learning experiences for students. (Daniel 2015)

Big data, in other words, can be utilized to disclose information on a variety of aspects that ultimately allow institutions to better respond to their "customers'" needs (the students) and improve their internal structures. Since higher education institutions are more and more evolving into business entities with "(1) the need to enhance prestige and market share; (2) the need to embrace an entrepreneurial mindset; and (3) the need to expand interactions and value co-creation with key stakeholders." (page 311, Pucciarelli and Kaplan, 2016) the potential benefits of data mining are clear.

The type of data collected and the data analysis performed, greatly depends on the objectives' pursued which have changed over time, increasing in complexity as technological tools have also advanced, making the analysis easier to accomplish. The focus of the studies have also evolved from association rule analysis (1995-2005) with a primarily descriptive function, towards clustering and classification-type studies (Romero et al, 2008), with, to a lesser extent and more recently, studies involving sequential pattern analysis whose primary emphasis is predictive in nature. (2013, Mohamad and Tasir)[1]

[1] Clustering and data segmentation approaches focus on identifying commonalities among the data collected to identify the most salient features upon which to focus interventions.
Classification, similarly, aggregates data according to predetermined categories and constitute the foundation for machine learning algorithms.
Association approaches focus on highlighting the interactions existing between characteristics of the data with the objective of building predictive models, such as linear or multi-regression models.
Factor analysis on the other hand attempts to find ways to naturally group variables together as sets and in education-based data mining is an approach used to achieve dimensionality reduction (Daniel 2015).

As the data collected is expanded and becomes more rich so does the potential analyses possible, echoing a trend that is true in general in data mining, with the underlying reasoning being: collect the data first then the analysis will tell its story.

This is however, currently only still partially true in education. In fact, and especially with respect to predictive analysis of students performance, researchers have mostly used and still use more limited and targeted datasets that have led to precise but limited portraits of students and their learning patterns, leading to studies with limited generalizability. (Romero et al, 2008; Huang and Fang, 2013; Hamsa et al. 2016; Yang and Li, 2017; Migueis et al, 2018; Burgos et al. 2018)

As data mining techniques, however, are becoming better understood and their potential better known, greater interest is being placed on using larger and larger datasets in this arena as well. Yet, there is still some resistance on the part of institutions and educational organizations to open their data up to analysis. Even mandatory reported data is muddied to try to prevent disclosure of sensitive information when data mining analysis is applied to predict (students' and programs') performance.

In part, this resistance is connected to the general and growing awareness that data collection is challenging individual privacy and possibly individual rights and, at the same time, laying bare the institutions themselves, even though "institutions are now aware that the early inference of students potential academic performance may enable them to foster higher levels of academic achievement" (Migueis et al. 2017)

The right to privacy when analyzing data is so fundamental that it is defended by the UN Fundamental Principles of Official Statistics (principle 6) which states " Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons,

are to be strictly confidential and used exclusively for statistical purposes'. However, implementing regulations in the field of data mining, or even a framework under which to guide this widespread collection and analysis of data, is not simple, and generates conflicting interests, within and across borders.

So far, many regulative proposals in the USA have been focused on giving consumers the choice (and thus the responsibility) of giving access to their data, fundamentally treating the issue as a consumer protection challenge. Consumers in other words are asked to freely offer their data up for analysis. One could argue that these pseudo-informed choices consumers are asked to make, are not that informed at all, and thus ultimately fail to really protect them. Most consumers (in our case, students and institutions) in fact, have still limited or no understanding of the extent to which their data can be used and the extent to which their data can reveal information about them. Furthermore, their consent is often too broad and gives "data collectors" wide discretion on what data is collected and its use.

At the other end of the spectrum there are attempts to regulate the field, such as in the case of the European Union, or incremental approaches, such as China's, where instead of attempting to regulate the field directly (for the moment), the decision was to adopt a framework "*to establish non-binding standards for privacy protection"* (page 22 McMillan 2018)

In Europe, privacy concerns have led to the relatively recent (2016) approval of the General Data Protection Regulation (GDPR) by the European Parliament, which came into effect in 2018 and highlights an individual's "right to be forgotten" (i.e. the right to see your data deleted from data warehouses and not stored indefinitely over time) and raises the bar in terms of the responsibilities associated with data collection. (MacFeely 2018; McMillan 2018) In China the 2018 National Standards on Information Security Technology – Personal Information Security

Specification GB/T 35273-201 and appears to be the first step in a series of measures still to be drafted, sets standards inspired by the EU's GDPR, and places greater burden on the companies on protecting personal information.

The European Union and China, in other words, chose to take a stance that is quite different than the one prevailing in the USA, where individuals ultimately remain less protected. It is worth noting in fact that the data is "Big Data" in the sense that the volume of data collected is big but also in the sense that often the scale of the analysis and conclusions that can be drawn from it are big and can have significant consequences.

To have a sense of the magnitude of the data collected in the field of education alone and of the potential consequences, we can consider that "in 2017, the top five MOOC [Massive Open Online Courses] providers reported over 60 million registered users" (Dhawal, 2018 as quoted in Johanes and Thille 2018). If we simply imagine all the learning analytics data potentially collected from this wide array of users, without even considering the potential for cross references between large data sets (collected for example through social media or other IoT devices), we can start having a sense of the extent of the problem and how serious the challenge is in terms of privacy. (Manca et al. 2016)

However, "Big data and AI require consideration of human rights concepts beyond individual privacy, extending assessment to ethical implications of data use not only on individuals but also on groups of people." (page 5 UN Global Pulse and IAPP, 2017) In education, for example, data has already started being used with life-changing effects in the admissions' process of higher education institutions. Limited discussion over the extent of the acceptable use of such information has taken place so far. This is concerning. Biases may easily taint data mining-based decisions in such situations and their consequences can be significant for individuals (and groups

of individuals with some common traits), but even when biases may not be at play, the consequences of the collection and analysis of all this data can be significant as some examples in the news have already demonstrated[2].

One might think that the greater the variability in the data collected, the greater the protection against biases and the wider the number of avenues through which to analyze trends and patterns potentially protecting from misrepresentation. This may be true, and standards in data collection could facilitate the use of the data and its manipulations more in ethical and equitable ways, but there is no guarantee and we are in unchartered territory. Furthermore, technology or our ability to fully exploit it, has not quite reached the necessary development stage. (Ifenthaler 2016; Ma et al. 2015)

It is undoubtedly true, that elaboration of learners' data may lead to positive and beneficial outcomes, for the institutions, but also for the individuals involved since "Algorithms analyzing big data, it turns out, are extremely successful predictors of designated attributes, such as dropping out, educational success, and effectiveness of educational programs for individual students." . (page 309 Harel Ben Shahar 2017) and thus, when appropriately used, they can effectively support teachers' pedagogy and lead to positive outcomes all around. It is also true that, in general, the data collected per se does not necessarily create a challenge or pose threats to the individual student whose information is being aggregated and then analyzed (unless biases seep into the analysis and other limited situations such as admission processes). Then, what is the concern? The challenge rests on the consideration that it is the ex-post manipulation of the data itself, and especially the conclusions drawn from it (based on underlying assumptions set by the researcher) that may lead to the generalization of trends and the identification of features on

[2] https://www.nytimes.com/2020/07/02/us/racism-social-media-college-aData Mining issions.html

which to leverage interventions. The use of such generalizations to draw conclusions and especially make decisions can lead to impacts on the wider society and cause long term changes when in fact they are based on the "preferences" of a few. When based on historical biases or more generally when based on "traditions", then societal transformations and generally speaking, evolutions, can become more challenging over time, thus limiting one of the advocated impacts of data mining and one of the founding characteristics of democratic societies.

Furthermore, while data analytics has the potential to help students grow and help them find better avenues to maximize their potential for example suggesting students educational paths they might otherwise either not consider or not be aware of. (Hamsa et al. 2016) the models may risk producing exactly the opposite results. What we mean to refer to here is the concern that some have raised about the potential homogenization that could occur "by guiding a large number of people to a limited number of popular resources, rather than by increasing the range of resources and ideas that all users are exposed to" which ultimately would stifle innovation under the guise of creating tailored learning environments (Lee and Hosanagar as quoted by Wilson 2017)

All of this notwithstanding, there is still little emphasis placed on how to depict let alone address the consequences that may arise from partial, flawed, or biased data analysis. The general concept of developing for example, more human Artificial Intelligence machines/robots a more recent area of application of data mining, is in infancy but is a prominent example of the initial steps taken to wrestle with these challenges and make sure that AI is developed with "humanness" in mind even when applied to the field of education. (Yang et al. 2021; Zhang and Aslan 2021)

In addition, data mining techniques to promote and recommend policy and decision making are founded on the assumption that the data collected is authentic. Data authenticity problems (reliability of the data collected) are not a trivial matter and do not exclusively refer to keeping the sources of the data free from hackers. It is true that with data mining techniques we are moving away from traditional data collection mechanisms founded on surveys and direct questioning techniques more readily prone to false reporting or inaccuracies, however the authenticity of the data may still be undermined. Even capturing students' reactions and engagement for example through automatic capturing mechanisms does not ensure automatic authenticity of the data, as any instructor who has proctored a live exam knows well, further distorting the potential results of the analysis and the decisions founded upon them.

Furthermore, there does not seem to be sufficient responsibility associated with the outcomes of Big Data processing, even when there is an increased awareness – slowly – of the potential risks associated with decision making based exclusively on machine learning models, given the fact that they have the potential to recommend solutions which could be counterproductive or, more simply, wrong. (United Nations Global Pulse and the International Association of Privacy Professionals May 2017) Increased attention and debate is emerging over the consequences that automatic decision making based on machine learning processes can have from a societal perspective, but we are still exploring this theme. (Johnson, 2014) To offer another concrete example, this time in terms of educational-justice, it is sufficient to think about limited resources' allocation. Researchers' underlying assumptions and beliefs, in parallel with their data analysis, could answer the question of whether to assign limited resources to enhance the probability of success of struggling students or to ensure greater achievements for those already performing. Their conclusions and recommendations might not be founded on a shared and generally

embraced perspective. Their recommendations could both be supported and justified by their data analysis. Yet, we, as a society or even more simply the institutions the researchers belong to, may never have had a chance to offer input (or understand) the implications of such decisions. (Harel Ben Shahar 2017)

Researchers have attempted to fill in this void of lack of guidelines in the field. PPDM - Privacy Preserving Data Mining – as the term implies, is an approach to data mining intended to supersede at least the challenge of privacy protection, albeit from a pragmatic perspective rather than a regulatory one. (Upadhyay et al. 2018) Here, it is worth noting that privacy in this context includes individuals' right to privacy from disclosing personal identification information (content) and privacy from being associated with specific internet content (interaction). (M. BinJubier et al. 2019; Florea and Florea, 2020; Milan et al, 2018) It is also worth noting that privacy is challenged immediately as data is created, even before it is collected from the source, due to the fact that often data collectors, data storers, and data manipulators are not one and the same, and that this can lead to data duplication and fundamentally a loss of control over the data created. This is not necessarily clear even to regulators who still believe that "To mitigate some of the risks, the privacy and ethical challenges associated with big data use should be addressed during the data collection and project development phase, as opposed to reactively, after the fact." (page 4 UN Global Pulse and IAPP, 2017)

From the technical perspective, some of the techniques proposed to maintain data privacy by researchers have been data exchanges (which imply the safe and trustworthy exchange of data sources among trusted collectors, so as to obfuscate the data's true origins), data cryptographic (which imply the application of complex mathematical formulas to encode the data without even the parties involved in submitting the data being aware of the underlying rules of these

transformations), and data manipulation (i.e. "data perturbation and anonymization-based techniques"). (page 20070, M. BinJubier *et al.* 2019; Qi and Zong, 2011; Upadhyay et al. 2018)

Data Anonymity techniques include K- Anonimity, L- Diversity, and T closeness. (M. BinJubier *et al.* 2019; Qi and Zong, 2011) Their primary objective is to make sure that the identity of the individual whose data is being collected and shared is not identifiable by obfuscating the identity of the data source. Data Perturbation approaches can be sub divided into two groups value based approaches (uniform versus probability based approaches) and dimension based approaches (Random Rotation Transformation and Random Projection) and their primary objective is to have data providers modify the data prior to sharing it with collectors.

Recent studies have shown however, that most of these methods fall short of achieving their goal for one reason or another (either they are easily reversible, or they distort the data to the point that it then becomes less usable for analysis which is clearly not desirable). (M. BinJubier *et al.* 2019)

More promising have been dimension based approaches such as 3D-Rotation which are based on more elaborate data transformation that have the added benefit of not distorting the underlying data trends. (Upadhyay et al. 2018) However, as the report by PCAST – the President's Council of Advisors on Science and Technology [3]- stated in 2014 in the paper by the title "*Big Data and Privacy: A Technological Perspective"*, "Privacy protection cannot be achieved by technical measures alone." (page XIII PCAST, 2014)

At the same time as attention is placed on privacy protection, growing concerns are being voiced over the oversimplified nature of many data mining studies due to the fact that often these studies

[3] PCAST is a council of experts in science and technology put together to offer policy support directly to the President from within the White House.

do not include variables intended to measure proxies more traditionally associated with education theory. It is underlined that "What seems to be missing in current research is that the real aspects of collaborative learning approach, which involve joint intellectual effort among students or between students and teachers and how they really engage and connect with the educational learning theories and strategies in learning." (Mohamad and Tasir 2013) Yet the use of graph mining and network analysis studies promising tools to start overcoming these limitations.

Some authors have associated this narrow focus to the limited data collected and used in data mining studies on students' performance. Narrowly collected data is argued, can lead to precise but very limited portraits of students and their learning patterns, with limited generalizability of the studies' results and thus limited significance. A study intended to evaluate the effectiveness of different data analysis techniques founded in data mining pattern identification objectives and applied to a specific high stakes engineering class, is an interesting example. The study suggests that traditional multi-regression techniques are effective in the aggregate prediction of whole-class performance, while more elaborate models, based on machine learning techniques (such as support vector machine models), because of their inherent ability to take into consideration a wider array of variables and their interactions, were more successful in predicting individual students' success in the class. (Huang and Fang 2013) In spite of the interesting results, the authors of this study raised serious concerns around the possibility of generalizing their results. They underlined that many "soft" variables, "such as learning styles, self-efficacy, achievement goals, motivation, interest, and teaching and learning environment" that should have been part of their analysis were not. They concluded that greater work needed to be done to extend data mining to include harder to measure/ less tangible variables. (Huang and Fang 2013) They

underlined explicitly, in other words, how even their significant data mining analysis based on a relatively large data set ,should be interpreted and used with care.

Lastly, it is important to remember that data mining in education is still an expensive endeavor, expensive in terms of costs and risks associated with the data collection and in terms of the time required for the data analysis itself.

Concluding remarks

As I have shown, data mining is an important tool that can be effectively used to pursue education improvement objectives, leading to advancements of students' performance but also to institutional improvements in terms of efficiency and efficacy. It can help us move closer to, for example, Obama's 2020 objectives for higher education which fundamentally had set goals in terms of overall higher education graduation rates and are still to be met according to Pew research[4].

Using data mining to identify the best approaches to intervene reliably would imply a significant step forward in addressing these goals. However, as we have seen, data mining analysis is not as simple a solution to adopt as may appear at first glance. It requires policy discussion.

Whenever there are social impacts at stake, mathematical models, should be carefully built and implemented, especially since the results of any recommended interventions may be visible with a significant delay (probably no sooner than a twenty year term).

[4] http://pewrsr.ch/2jn2PdH

The new Administration set forth on January 15, 2021, just before taking office, a series of broad priorities to the President's Council of Advisors on Science and Technology[5] (PCAST) Missing from the list in specific terms is any reference to Big Data and its management, even though, just the previous administration had made Big Data and understanding its use and challenges, a priority. This seems to be a move in the wrong direction.

The time for modern analytics based on large data collection is here to stay and has already demonstrated in various areas its ability to move forward discussion and problem solve, having ultimately positive effects, at least so far and at least from the perspective of revenue recovery and revenue creation for private organizations. It no doubt has the potential to have positive effects in education as well. Many however, as we saw, are the aspects that should be taken into consideration when proposing the implementation of such tools, and all depends on the goal intended to be solved. Avoiding policy discussion around the matter does not seem to be helpful. As some researchers have said "'we must understand the underlying structure of the phenomenon we seek to explore with analytics prior to digging into the data'" (Knight et al as cited by Wilson 2017)

There are those who criticize data mining in the field of education relegating it to a system to embrace social-based education theories by counting interactions between students on learning management systems. As the techniques become more sophisticated, as we have seen with more recent developments, for example in AI, it is hard to be able to defend these positions and really maintain their merit. However, they do raise awareness – again – on the need to raise the

[5] https://www.whitehouse.gov/ostp/news-updates/2021/01/15/a-letter-to-dr-eric-s-lander-the-presidents-science-advisor-and-director-of-the-office-of-science-and-technology-policy/

sophistication of such analytical methods and/or to keep into consideration the limitations of the models and not try to expand their application beyond their limits.

It is emblematic that the discussion and development of models to interpret education-based data has not seen greater participation on the part of education thinkers. This is resolvable. Actually, we could even think of promoting – like some countries already are – the possibility to develop within Schools of Education, branches specialized in addressing and developing sounder approaches to analytics, founded on educational theory and not just leave the model development in education to the devise of data scientists alone. (Wilson 2017)

Either way it is essential to expand, improve, and protect the data collected in order to reassure all contributors, us, that any recommendations proposed will not be tainted by partial or skewed perceptions. The first step in the right direction is to widely propose and apply effective privacy protection techniques such as 3D Rotations and, at the same time, have a serious and open discussion inherent to the goals we might want to choose as a society when proposing interventions that may affect access and outcomes in education.

It may seem absurd to conclude a paper on big data mining wondering if we are collecting enough data to really perform effective and systematic, informative data mining in education. However, if we reference for example the data collected across the world to pursue the Sustainable Development Goals set forth by the UN as world goals for 2030 as an example, we already know that only about 40% of the goals set, are grounded on data currently being collected by most countries around the world. It is not surprising then to wonder whether the data collected also in the field of education is sufficient. (MacFeely 2019)

Still more data will not be the solution if we will not be able, at the same time, to enhance our modeling capabilities and our overall understanding of their multifaceted implications. As the

UN Global Pulse and International Association of Privacy Professionals put in May 2017: “In today’s world, a comprehensive solution for realizing big data and AI benefits for the greater good requires a combination of technical, governance, legal, and ethical responses. This calls for a multidisciplinary approach that draws on the expertise of major players in these distinct, yet complementary fields.” (page 4) Only if we can continue to pursue this goal can the impacts can be significant for the students, who could see better short and long term outcomes, for the institutions, who could perform better and be more readily able to respond to structural changes, but also for society at large given the impact education has on society as a whole.

References

Altujjar Y., Altamimi W., Al-Turaiki I., Al-Razgan M. (2016). Predicting Critical Courses Affecting Students Performance: A Case Study. Procedia Computer Science, Vol. 82, 65-71. https://doi.org/10.1016/j.procs.2016.04.010

Anwar, M. J., Gill, A. Q., Hussain, F. K., & Imran, M. (2021). Secure big data ecosystem architecture: challenges and solutions. *EURASIP Journal on Wireless Communications & Networking*, *2021*(1), 1–30. https://doi.org/10.1186/s13638-021-01996-2

Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3-17. https://doi.org/10.5281/zenodo.3554657

Behnam, N., & Crabtree, K. (2019). Big data, little ethics: confidentiality and consent. *Revista Migraciones Forzadas*, *61*, 4–6.

Ben Shahar, T. H. (2017). Educational justice and big data. *Theory & Research in Education*, *15*(3), 306–320.

Ben-Porath, Sigal & Shahar, Tammy. (2017). Introduction: Big data and education: ethical and moral challenges. Theory and Research in Education. 15. 243-248. 10.1177/1477878517737201.

Binjubeir M., Ali Ahmed A., Arfian Bin Ismail M., Safaa Sadiq A., & Khurram Khan M. (2020). Comprehensive Survey on Big Data Privacy Protection. *IEEE Access*, *8*, 20067–20079. https://doi.org/10.1109/ACCESS.2019.2962368

Bound, J., Braga, B., Khanna, G., & Turner, S. (2021). The Globalization of Postsecondary Education: The Role of International Students in the US Higher Education System. *Journal of Economic Perspectives*, *35*(1), 163–184. https://doi.org/10.1257/jep.35.1.163

Burgos C., Campanario M.L., de la Peña D., Lara J.A., Lizcano D., & Martínez M.A. (2018) Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, Vol. 66, 541-556. https://doi.org/10.1016/j.compeleceng.2017.03.005

Castilla, E. J., & Rissing, B. A. (2019). Best in Class: The Returns on Application Endorsements in Higher Education*. *Administrative Science Quarterly*, *64*(1), 230–270. https://doi.org/10.1177/0001839218759965

Chamikara, M. A. P., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2020). Efficient privacy preservation of big data for accurate data mining. *Information Sciences*, *527*, 420–443. https://doi.org/10.1016/j.ins.2019.05.053

Chen J, & Yang L.. (2020). Special Section on Privacy Computing: Principles and Applications, Information Sciences, Volume 527, 293. https://doi.org/10.1016/j.ins.2020.04.047

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, *46*(5), 904–920. https://doi.org/10.1111/bjet.12230

Dule, C. S., & H. A., G. (2017). Content an Insight to Security Paradigm for BigData on Cloud: Current Trend and Research. *International Journal of Electrical & Computer Engineering (2088-8708)*, *7*(5), 2873–2882. https://doi-org.montgomerycollege.idm.oclc.org/10.11591/ijece.v7i5.pp2873-2882

Duncum, P. (2018). Responding to Big Data in the Art Education Classroom: Affordances and Problematics. *International Journal of Art & Design Education*, *37*(2), 325–332. https://doi.org/10.1111/jade.12129

Fernandes E., Holanda M., Victorino M., Borges V., Carvalho R., & Van Erven G..(2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, Vol. 94, 335-343. https://doi.org/10.1016/j.jbusres.2018.02.012

González Canché, M. S. (2018). Geographical network analysis and spatial econometrics as tools to enhance our understanding of student migration patterns and benefits in the U S higher education network. *Review of Higher Education: Journal of the Association for the Study of Higher Education*, *41*(2), 169–216. https://doi.org/10.1353/rhe.2018.0001

Hackl, P. (2016). Big Data: What can official statistics expect? *Statistical Journal of the IAOS*, *32*(1), 43–52. https://doi.org/10.3233/SJI-160965

Helal S., Li J., Liu L., Ebrahimie E., Dawson S., Murray D.J., & Long Q. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems*, Vol. 161, 134-146. https://doi.org/10.1016/j.knosys.2018.07.042

Holloway, K. (2020). Big Data and learning analytics in higher education: Legal and ethical considerations. *Journal of Electronic Resources Librarianship*, *32*(4), 276–285. https://doi.org/10.1080/1941126X.2020.1821992

Ifenthaler, D. (2017). Are Higher Education Institutions Prepared for Learning Analytics? *TechTrends: Linking Research & Practice to Improve Learning*, *61*(4), 366–371. https://doi-org.montgomerycollege.idm.oclc.org/10.1007/s11528-016-0154-0

Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. *Educational Technology Research & Development*, *68*(4), 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

Injadat M.N., Moubayed A., Bou Nassif A., & Shami A. (2020). Systematic ensemble model selection approach for educational data mining, Knowledge-Based Systems. Volume 200. https://doi.org/10.1016/j.knosys.2020.105992

Johanes, P., & Thille, C. (2019). The heart of educational data infrastructures = Conscious humanity and scientific responsibility, not infinite data and limitless experimentation. *British Journal of Educational Technology*, *50*(6), 2959–2973.

Kyritsi K. H., Zorkadis V., Stavropoulos E. C, & Verykios V. S. (2019). The Pursuit of Patterns in Educational Data Mining as a Threat to Student Privacy. *Journal of Interactive Media in Education*, *2019*(1). https://doi.org/10.5334/jime.502

Lemay D. J., Baek C., & Doleck T.. (2021) Comparison of learning analytics and educational data mining: A topic modeling approach. *Computers and Education: Artificial Intelligence*, Vol. 2, https://doi.org/10.1016/j.caeai.2021.100016

Logeswaran, K., Suresh, P., Ponselvakumar, A. P., & Savitha, S. (2020). A study on data driven technologies involved in the development of viable anticipated smart cities. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2020.12.1067

Ma S., Li J., Liu L., & Duy Le T. (2016). Mining combined causes in large data sets. *Knowledge-Based Systems*, Vol. 92, 104-111. https://doi.org/10.1016/j.knosys.2015.10.018

Manca S., Caviglione L., & Raffaghelli J.E. (2016). Big data for social media learning analytics: potentials and challenges. *Je-LKS : Journal of e-Learning and Knowledge Society*, *12*(2). https://doi.org/10.20368/1971-8829/1139

Menon, S., & Sarkar, S. (2016). Privacy and Big Data: Scalable Approaches to Sanitize Large Transactional Databases for Sharing. *MIS Quarterly*, *40*(4), 963–982.

Miguéis V.L., Freitas A., Garcia P. J.V., & Silva A.. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, Vol. 115, 36-51. https://doi.org/10.1016/j.dss.2018.09.001

Miotto G., Del-Castillo-Feito C., Blanco-González A. (2020). Reputation and legitimacy: Key factors for Higher Education Institutions' sustained competitive advantage. *Journal of Business Research*, Vol. 112, 342-353. https://doi.org/10.1016/j.jbusres.2019.11.076

Mohamed H. S. & Al-Razgan M. S. (2016). Pre-University Exams Effect on Students GPA: A Case Study in IT Department. *Procedia Computer Science*, Vol. 82, 127-131. https://doi.org/10.1016/j.procs.2016.04.018

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLoS ONE*, *13*(6), 1–21. https://doi.org/10.1371/journal.pone.0198746

Prinsloo, P., Archer, E., Barnes, G., Chetty, Y., & van Zyl, D. (2015). Big(ger) Data as Better Data in Open Distance Learning. *International Review of Research in Open and Distributed Learning*, *16*(1), 284–306

Priyank Jain, Manasi Gyanchandani, & Nilay Khare. (2018). Differential privacy: its technological prescriptive using big data. *Journal of Big Data*, *5*(1), 1–24. https://doi.org/10.1186/s40537-018-0124-9

Qi X. and Zong M. 2012, An Overview of Privacy Preserving Data Mining, 2011 International Conference on Environmental Science and Engineering (ICESE 2011), Procedia Environmental Sciences 12, 1341-1347

R. A., A., Hegadi, R. S., & T. N., M. (2018). A Study on Big Data Privacy Protection Models using Data Masking Methods. *International Journal of Electrical & Computer Engineering (2088-8708)*, *8*(5), 3976–3983

Rabeea Mahdi O., Nassar I. A., & Almsafir M.K. (2019). Knowledge management processes and sustainable competitive advantage: An empirical examination in private universities. *Journal of Business Research*, Vol. 94, 320-334. https://doi.org/10.1016/j.jbusres.2018.02.013

Radhika, D., & Aruna Kumari, D. (2018). Misusability Measure Based Sanitization of Big Data for Privacy Preserving MapReduce Programming. *International Journal of Electrical & Computer Engineering (2088-8708)*, *8*(6), 4524–4532. https://doi.org/10.11591/ijece.v8i6.pp4524-4532

Rafferty, A. N., Whitehill, J., Romero, C., & Cavalli-Sforza, V., International Educational Data Mining Society. (2020). Proceedings of the International Conference on Educational Data Mining (EDM) (13th, Online, July 10-13, 2020). *International Educational Data Mining Society*.

Ram Mohan Rao, P., Murali Krishna, S., & Siva Kumar, A. P. (2018). Privacy preservation techniques in big data analytics: a survey. *Journal of Big Data*, *5*(1). https://link.gale.com/apps/doc/A555336216/AONE?u=rock77357&sid=bookmark-AONE&xid=50dd5b04

Reidenberg, J. R., & Schaub, F. (2018). Achieving big data privacy in education. *Theory & Research in Education*, *16*(3), 263–279.

Rhahla, M., Allegue, S., & Abdellatif, T. (2021). Guidelines for GDPR compliance in Big Data systems. *Journal of Information Security and Applications*, *61*. https://doi.org/10.1016/j.jisa.2021.102896

Riahi, A., Savadi, A., & Naghibzadeh, M. (2020). Comparison of analytical and ML-based models for predicting CPU–GPU data transfer time. *Computing*, *102*(9), 2099–2116. https://doi.org/10.1007/s00607-019-00780-x

Romero C, Ventura S., Espejo P. G., & Hervás C. (2008). Data Mining to Classify Students. Initially published at the 1st International Conference on Educational Data Mining (Montreal, Canada)

Schwieger, D., & Ladwig, C. (2016). Protecting Privacy in Big Data: A Layered Approach for Curriculum Integration. *Information Systems Education Journal*, *14*(3), 45–54.

Siti Khadijah M. & Zaidatun T. (2013). Educational Data Mining: A Review, *Procedia - Social and Behavioral Sciences,* Vol. 97, 320-324. https://doi.org/10.1016/j.sbspro.2013.10.240

Terziev, V., & Lyubcheva, M. (2020). Internal and External Challenges Facing Higher Education. *Business Management / Biznes Upravlenie*, *4*, 19–33.

Travis, T. A., & Ramirez, C. (2020). Big Data and Academic Libraries: The Quest for Informed Decision-Making. *Portal: Libraries & the Academy*, *20*(1), 33–47. https://doi.org/10.1353/pla.2020.0003

Troisi, O., Grimaldi, M., Loia, F., & Maione, G. (2018). Big data and sentiment analysis to highlight decision behaviours: a case study for student population. *Behaviour & Information Technology*, *37*(10/11), 1111–1128. https://doi.org/10.1080/0144929X.2018.1502355

Upadhyay S., Sharma C., Sharma P., Bharadwaj P, & Seeja K.R. (2018). Privacy preserving data mining with 3-D rotation transformation. *Journal of King Saud University*, *Computer and Information Sciences*, Vol. 30, Issue 4, 524-530, https://doi.org/10.1016/j.jksuci.2016.11.009

Viloria, A., Padilla, J. G., Vargas-Mercado, C., Hernández-Palma, H., Llinas, N. O., & David, M. A. (2019). Integration of Data Technology for Analyzing University Dropout. *Procedia Computer Science*, *155*, 569–574. https://doi.org/10.1016/j.procs.2019.08.079

Wang Y.. (2017). Education policy research in the big data era: Methodological frontiers, misconceptions, and challenges. *Education Policy Analysis Archives*, *25*(0). https://doi.org/10.14507/epaa.25.3037

Wieczorkowski, J., & Polak, P. (2017). Big data and privacy: The study of privacy invasion acceptance in the world of big data. *Online Journal of Applied Knowledge Management*, *5*(1), 57–71. https://doi.org/10.36965/ojakm.2017.5(1)57-71

Winer, A., & Geri, N. (2019). Learning analytics performance improvement design (LAPID) in higher education: Framework and concerns. *Online Journal of Applied Knowledge Management*, *7*(2), 41–55. https://doi.org/10.36965/ojakm.2019.7(2)41-55

Yang F., & Li F. W.B.(2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, Vol. 123, 97-108. https://doi.org/10.1016/j.compedu.2018.04.006

Yang S. J.H., Ogata H., Matsui T., & Chen N.S. (2021) Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, Vol. 2. https://doi.org/10.1016/j.caeai.2021.100008

Yin, H. H. S., Langenheldt, K., Harlev, M., Mukkamala, R. R., & Vatrapu, R. (2019). Regulating Cryptocurrencies: A Supervised Machine Learning Approach to De-Anonymizing the Bitcoin

Blockchain. *Journal of Management Information Systems*, *36*(1), 37–73. https://doi.org/10.1080/07421222.2018.1550550

Zhang K. & Begum Aslan A.. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2021.100025

Additional Resources:

International Educational Data Mining Society. e-mail: admin@educationaldatamining.org; Web site: http://www.educationaldatamining.org

OECD Report education at a glance 2013 https://www.oecd.org/education/eag2013%20(eng)--FINAL%2020%20June%202013.pdf

Pew research http://pewrsr.ch/2jn2PdH

The New York Times https://www.nytimes.com/2020/07/02/us/racism-social-media-college-aData Mining issions.html

The White House https://www.whitehouse.gov/ostp/news-updates/2021/01/15/a-letter-to-dr-eric-s-lander-the-presidents-science-advisor-and-director-of-the-office-of-science-and-technology-policy/

UN Fundamental Principles of Official Statistics https://unece.org/statistics/fundamental-principles-official-statistics

United Nations Global Pulse and the International Association of Privacy Professionals May 2017, Building Ethics into Privacy Frameworks for Big Data and AI