

## **Ciência de Dados Educacionais: definições e convergências entre as áreas de pesquisa**

**Leandro A. Silva<sup>1</sup>, Ismar Frango Silveira<sup>1</sup>, Luciano Silva<sup>1</sup>,  
Jorge Luis Cavalcanti Ramos<sup>2</sup>, Rodrigo Lins Rodrigues<sup>3</sup>.**

<sup>1</sup>Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie  
Rua da Consolação, 930 - 01302-907 – São Paulo – SP – Brasil  
{leandroaugusto.silva, ismar.silveira, luciano.silva}@mackenzie.br

<sup>2</sup>Colegiado de Engenharia de Computação - Universidade Federal do Vale do São Francisco  
Av. ACM, 510, Santo Antônio - CEP 48.902-300 - Juazeiro - BA – Brasil  
jorge.cavalcanti@univasf.edu.br

<sup>3</sup>Departamento de Educação - Universidade Federal Rural de Pernambuco  
Rua Dom Manoel de Medeiros, s/n, Dois Irmãos - CEP 52171-900 - Recife - PE – Brasil  
rodrigo.linsrodrigues@ufrpe.br

***Abstract.** The growing interest in the application of Big Data-based techniques in problems belonging to the context of Computers in Education stimulated the emergence of research initiatives, gathered under different names: Educational Data Mining, Learning Analytics and, more recently, Academic Analytics. Considering the existence of several common points between these areas and also a certain lack of definition about the boundaries between them, this article intends to contribute to a greater conceptual clarity in this field, when adopting the term Educational Data Science to broadly analyze the main topics covered for the articles of six national and international events dedicated to the subject.*

***Resumo.** O crescente interesse na aplicação de técnicas oriundas da área de Big Data em problemas pertencentes ao contexto da Informática na Educação estimulou o surgimento de pesquisas reunidas sob diferentes nomes: Mineração de Dados Educacionais, Analíticas de Aprendizagem e, mais recentemente, Analíticas Acadêmicas. Considerando os diversos pontos em comum entre essas áreas e também uma certa indefinição da fronteira entre elas, este artigo pretende contribuir para uma maior clareza conceitual nesse campo, ao adotar o termo Ciência de Dados Educacionais para analisar, de maneira ampla, os principais tópicos trabalhados pelos artigos de seis eventos nacionais e internacionais dedicados ao tema.*

### **1. Introdução**

Com o crescimento exponencial da geração de dados por usuários, dispositivos e sistemas, as tecnologias associadas ao *Big Data* apresentam-se como novas oportunidades para análise, entendimento, modelagem e predição de diversas variáveis presentes em grande volume de dados.

Assim como nas demais áreas afetadas pelo *Big Data*, o campo educacional vem incorporando cenários dessas tecnologias, em virtude das diversas abordagens

educacionais gerarem cada vez mais dados e também demandarem análises detalhadas e voltadas para um melhor planejamento e execução de ações na área da educação.

A análise de dados educacionais, de uma maneira geral, representa uma área de pesquisa emergente em Informática em Educação para o desenvolvimento de métodos que exploram dados oriundos de ambientes educacionais e também administrativos com a finalidade de entender melhor os estudantes e os cenários em que eles aprendem [Daniel 2016]

Como pesquisa, a análise de dados educacionais se desdobra em temáticas como *Educational Data Mining* (EDM) (ou Mineração de Dados Educacionais - MDE), *Learning Analytics* (LA) e *Academic Analytics* (AA).

Estes temas são conflitantes em sua definição, por terem como ponto em comum a maneira com que os dados educacionais são analisados, diferenciando-se basicamente na abordagem em que se coloca cada tipo de problema. Por exemplo, em EDM (ou MDE), o objetivo é analisar dados gerados em ambientes de ensino-aprendizagem, a partir da aplicação de tarefas de mineração de dados como predição (regressão, séries temporais e classificação), agrupamento ou associação de dados, a fim de realizar descobertas de conhecimento intrínseco nos dados.

O LA, por outro lado, implica no uso de técnicas de análises de dados, como análise estatística exploratória e até mesmo as tarefas de mineração de dados, para confirmar hipóteses colocadas em atividades que envolvem a aprendizagem do aluno, fomentando assim recursos analíticos para entendimento e aprimoramento do ensino-aprendizagem. Portanto, devido a origem dos dados para essas duas áreas ser de sistemas de aprendizagem dos alunos e as técnicas oriundas de abordagens como estatística e tarefas de mineração de dados, a interseção de EDM e LA é evidente.

Com pouco menos de confusão ao que tange análise de dados educacionais, emerge a temática *Academic Analytics* (AA), que tem na essência as mesmas abordagens de análise de dados aplicadas em EDM e LA, porém com alteração na origem dos dados, nesse caso advindo de sistemas educacionais administrativos e de gestão acadêmica.

O fato que decorre das relações comuns entre as áreas de pesquisa acima expostas (EDM, LA e AA) é uma indefinição de fronteiras entre as propostas de trabalhos que se confundem no decorrer do desenvolvimento de pesquisas científicas. Esse fato vem sendo observado nas três últimas edições do Workshop de Mineração de Dados Educacionais (WMDE), evento satélite ao CBIE dos últimos três anos. Uma parcela significativa dos trabalhos submetidos apresenta conflito conceitual, principalmente entre EDM e LA. Esse fato foi ainda mais nítido na edição de 2015, quando se propôs o *Latin American Workshop on Learning Analytics* (LALA) como um Workshop do LACLO e a sessão de trabalhos feita de maneira conjunta ao WMDE (WMDE e LALA). O que se pôde perceber foi uma nítida sobreposição das duas áreas de pesquisa, ou seja, trabalhos de mineração de dados educacionais no LALA e de *Learning Analytics* no WMDE. Entretanto, em relação ao AA, ainda não há um evento com este debate nacionalmente e que acaba - quando algum trabalho nesta temática aparece - por recar no tema de EDM.

Diante dos fatos expostos, o objetivo deste trabalho é apresentar um estudo por meio da análise de trabalhos publicados nos principais eventos da área, a fim de evidenciar a possibilidade do tratamento único das áreas de pesquisas MDE, LA e AA.

Como objetivo específico, o trabalho propõe uma definição de pesquisa única a estas áreas, possibilitando uma discussão ampla do tema educação no contexto de análise de dados.

A proposta é aqui chamada como Ciência de Dados Educacionais (CDE), que se propõe a explorar dados educacionais para o entendimento de situações advindas de ambientes acadêmicos, assim como, propor uma melhor interação entre as áreas de MDE, LA e AA, pois entende-se que muitos problemas e soluções possam ser descobertas considerando todas as dimensões educacionais e não apenas uma área específica.

## 2. Fundamentos de Análise de Dados Educacionais

A análise de dados é realidade em muitas áreas do conhecimento, com motivações claras e reais de utilização e de estratégia de uso com a finalidade de descobrir relações não óbvias ou ainda não experimentadas, mas que residem de forma implícita nos volumosos bancos de dados [Silva, Peres, Boscaroli, 2016].

Na área educacional, o interesse em usar e analisar dados já foi despertado, porém como tratamento feito de maneira segregada em três linhas de pesquisa: *Educational Data Mining*, *Learning Analytics* e *Academic Analytics*.

Conceituando cada uma destas linhas, a Mineração de Dados Educacionais (do inglês *Educational Data Mining*, EDM) é uma área de pesquisa que utiliza as tarefas da Mineração de Dados como Análise Preditiva, Agrupamento e Associação de Dados aplicados a problemas de contexto educacional [Romero e Ventura, 2007; Siemens e Baker, 2012; Costa et al., 2013]. A EDM tem como objetivos fazer descobertas sobre o comportamento dos estudantes e o ambiente no qual a aprendizagem ocorre, fornecendo insumos para o professor ou aluno investigar eventuais padrões descobertos [Romero e Ventura, 2007; Romero et al., 2016; Ducange et al., 2016].

Os trabalhos na área de EDM incluem, basicamente, dados oriundos de Sistemas de Gerenciamento de Aprendizagem (do inglês *Learning Management Systems*, LMS), Sistemas Tutoriais Inteligentes (do inglês *Intelligent Tutorial Systems*, ITS), e-learning, repositórios de objetos de aprendizagem e aplicações web utilizadas na educação.

*Learning Analytics*, por outro lado, foi definido, de acordo com Souza et al (2016), como um “processo para a medição, coleta, análise e comunicação de dados sobre os alunos e os seus contextos, para fins de compreensão e otimização da aprendizagem nos ambientes em que esse processo ocorre” [Siemens et. al., 2011]. O termo *Learning Analytics* (LA) ainda não possui uma tradução de consenso para o português.

Alguns autores nacionais traduzem como “Analítica da Aprendizagem” outros como “Análise da Aprendizagem”. Parece-nos que essa última definição é mais adequada, uma vez que essa área de pesquisa envolve o uso de ferramentas de análise de dados para avaliar processos de aprendizagem estabelecidos por educadores aos seus educandos conforme mostram os trabalhos de Siemens e Baker, (2012), Daniel (2016) e Romero et al. (2016).

Nessa mesma linha, qualquer tipo de estratégia de aprendizado, seja com uso de recursos de aprendizagem (objetos de aprendizagem) elaborados pelos educadores ou por uma equipe de TI, ao final espera-se que os resultados sejam analisados e, portanto, é nesse momento em que se insere os estudos de LA [Papamitsiou e Economides, 2014; Martinez-Maldonado et al., 2016; Knight e Littleton, 2016; Quigley et al., 2017].

As ferramentas de análise de dados usadas no LA são as mais amplas possíveis, incluindo as tarefas da Mineração de Dados. E nesse ponto que inicia-se a geração de conflito conceitual entre a EDM e LA.

Para um melhor entendimento das diferenças entre EDM e LA, Siemens e Baker (2012) propõem um contraste entre ambas, conforme apresentado no Quadro 1, adaptado neste trabalho. Enquanto MDE e LA focam na aprendizagem do aluno, há também uma série de outras informações dos mesmos que saem do escopo do aprendizado, mas que pode estar intrínseca ao aprendizado do aluno ou mesmo ser incorporadas de alguma maneira nas análises. Portanto, trata-se então dos dados acadêmicos.

A Análise de Dados Acadêmicos, tradução nossa à *Academic Analytics* ou AA, tem como foco o uso dos dados oriundos dos sistemas de informação da Instituição de Ensino (IE) para tentar entender os dados cadastrais dos alunos e outros que se relacionam com a vivência acadêmica do aluno na instituição [Campbell e Oblinger, 2007; Baepler e Murdoch, 2010]. Exemplo é o uso de dados demográficos, desempenho acadêmico, histórico escolar, censo da instituição, uso dos recursos computacionais, financeiro (para IE particulares) e uma série de outros dados que podem implicar de alguma maneira no desempenho do aluno [Campbel, 2007].

No entanto, como colocado por Baepler e Murdoch (2010), a análise deve ter um olhar mais amplo aos instrutores e alunos, mas também aos gestores. E nesse aspecto, pode-se considerar a AA como apoio aos gestores, a partir do momento em que se usam as análises para avaliar, por exemplo, projetos pedagógicos, processos administrativos, uso dos recursos da biblioteca, entre outros.

**Quadro 1: Contraste entre definições de MDE e LA.**

	<b>EDM</b>	<b>LA</b>
<b>Descoberta</b>	O objetivo é usar técnicas inteligentes para fazer descobertas depois confirmadas com base na aprendizagem do aluno	Prioriza a aprendizagem do aluno e usa sistemas inteligentes para análises confirmatórias
<b>Adaptação e personalização</b>	A ênfase é na criação de sistemas automáticos sem alunos e professores no processo	Foco na informação e empoderamento de alunos e professores
<b>Técnicas e Métodos</b>	Classificação, Agrupamento, Associação, Visualização de Dados e etc.	Análise de redes sociais, análise de sentimentos, análise de discurso, análise de conceitos e etc.

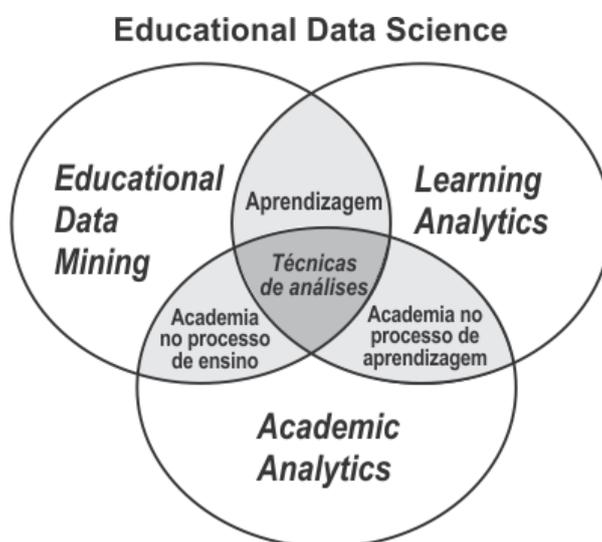
**Fonte: Adaptado de (SIEMENS, G.; BAKER, 2012).**

Essas áreas de pesquisa começam a se unir em torno de questões e problemas educacionais. Destacando a forma como a comunidade de pesquisadores começou a convergir em torno da mineração de dados educacionais (EDM), Análise de Dados Acadêmicos (AA) e, mais recentemente, juntar-se à comunidade de analítica da aprendizagem (LA) para formar um campo atual de pesquisa intitulado Ciência de Dados Educacionais.

## 2.1. Ciência de Dados Educacionais

De acordo com as definições anteriores, é possível notar que existe uma sobreposição nas linhas de pesquisa sobre Ciência de Dados Educacionais. A Figura 1 propõe uma ilustração de como essas sobreposições acontecem. A intersecção maior deve-se às técnicas de análise. Em todas as áreas se usa, basicamente, análise exploratória de dados (*Exploratory Data Analysis* - EDA), estatística descritiva, algoritmos de inteligência artificial e visualização de dados. E isso acaba gerando certo conflito conceitual entre a EDM e LA, como já se observa no trabalho de Siemens e Baker (2012).

Além dessa intersecção geral, há as intersecções específicas, como é o caso EDM e AA, sendo que se pode agregar dados acadêmicos no entendimento do processo do ensino ou, na mesma linha entre LA e AA, abordagem semelhante, porém com foco na aprendizagem. E, a relação comum entre MDE e LA é o aprendizado do aluno, independente se o interesse é o processo do ensino ou o conhecimento adquirido.



**Figura 1: Diagrama de Relacionamento entre linhas de pesquisa de análise de dados acadêmicos (Fonte: autores)**

Contudo, neste trabalho, entende-se que estas áreas não devem ser tratadas de maneira isoladas. Como proposta, coloca-se a origem dos dados como dimensões que implicam no resultado do ensino-aprendizagem. Ou seja, deve-se considerar os dados deixados em ambientes de educação, usados como apoio aos alunos na aula; os dados advindos de sistemas acadêmicos, em que tem a informação do aluno durante a aula; e, por fim, os dados do sistema de gestão, que mantém os dados sobre o cadastro do aluno.

Nesta proposta, entende-se que a combinação das várias dimensões resulta na definição do termo Ciência de Dados Educacionais, como sendo a exploração de dados para aumentar a compreensão e a qualidade das experiências de aprendizagem. Por meio da combinação de técnicas provindas da estatística, computação e educação.

Embora esta definição não seja consensual, a definição de Ciência de Dados Educacionais, assim como o papel assumido pelo cientista de dados educacionais, ainda padece de maiores formalismos, como já apontavam Buckingham Shum et al. (2013), assim como muitas de suas práticas e aportes teóricos ainda são emergentes, como afirmavam Piety et al. (2014). De acordo com Buckingham Shum et al. (2013), as

habilidades de Cientista de Dados Educacionais incluiriam “explorar o mundo real, propor medidas significativas, modelar os dados, visualizar a saída, compartilhar a técnica e automatizar os processos de ensino e aprendizagem”.

### 3. Metodologia

Para se estabelecer um panorama das temáticas publicadas em Mineração de Dados Educacionais e *Learning Analytics*, os títulos das publicações realizadas nos principais eventos foram selecionados e analisados sob a abordagem de *Text Mining* (Silva, Peres e Boscaroli, 2016).

Dos eventos internacionais em Mineração de Dados, os dois principais eventos da *International Educational Data Mining Society* foram consultados: o *International Conference on Educational Data Mining* (EDM) e o *Journal of Educational Data Mining* (JEDM).

No caso dos eventos sobre Learning Analytics, os dois eventos da *Society for Learning Analytics Research* (SOLAR): o *International Conference on Learning Analytics & Knowledge* (LAK) e o *Journal of Learning Analytics* (JLA).

Sob a perspectiva de discussões no âmbito nacional, apenas a MDE tem sido discutida e que não está associada a sociedades específicas, mas sim a um Workshop que leva o mesmo nome da temática, portanto WMDE e que faz parte do Congresso Brasileiro de Informática da Educação (CBIE), evento anual da Sociedade Brasileira da Computação (SBC). No caso do LA no país, um único evento foi oferecido com o nome de *Latin American Workshop on Learning Analytics* (LALA) como um Workshop do *Latin American Conference on Learning Technologies* (LACLO), quando este foi realizado no Brasil de forma conjunta ao CBIE no ano de 2015.

O objetivo da análise que será apresentada na seção seguinte é ter uma visão geral do volume de trabalhos aceitos em cada uma das áreas e, também, dos títulos dos trabalhos publicados nos anos de 2015 e 2016. Portanto, deseja-se fazer uma sumarização dos trabalhos publicados.

Para o caso dos títulos, foi aplicada técnicas de Text Mining para a geração de um corpus representado e a visualização do mesmo por nuvens de palavras [Silva, Peres e Boscaroli 2016].

### 4. Resultados e discussões

Os números de publicações em cada um dos eventos apresentados acima estão representados no gráfico da Figura 1. O que se pode notar é que o número de publicações na área de análise de dados tem aumentado exceto para o JEDM e LALA (que ocorreu em um único ano). E ainda, o Congresso de MDE é o evento que tem mais aceito trabalhos no mundo.

Por fim, como última análise, o evento nacional WMDE, embora tenha mais que dobrado o número de publicações de um ano para outro, ainda está muito abaixo dos congressos internacionais (MDE e LAK).

Com a finalidade de se ter uma ideia das temáticas publicadas nos eventos, as análises discutidas a seguir serão combinadas sobre os trabalhos relativos a EDM e LA, de âmbito nacional ou internacional, sem a distinção de anais ou revistas.

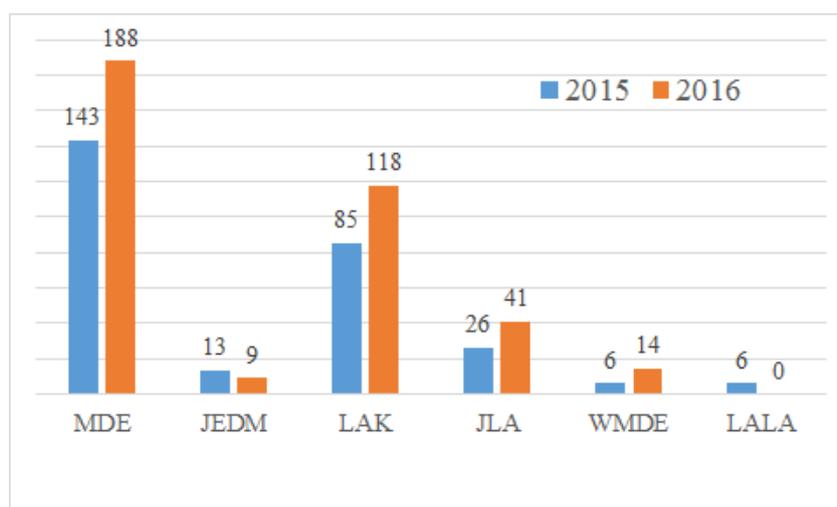
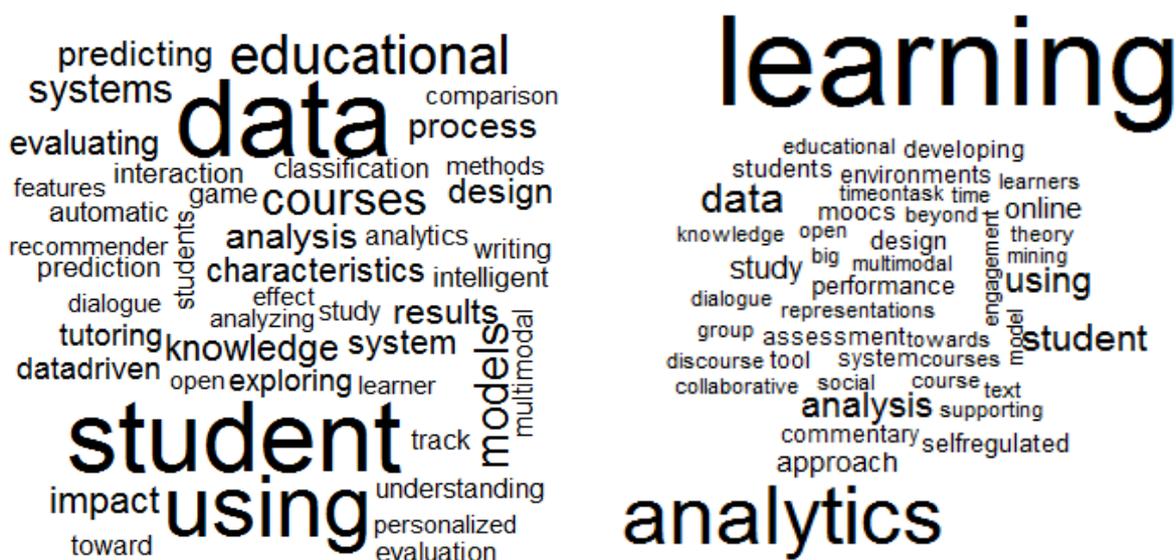


Figura 2: Contagem do número de publicações em MDE e LA

Na Figura 3, está a nuvem de palavras relativa aos eventos internacionais. Para facilitar a análise, foi aplicado um filtro na nuvem para representar 500 palavras extraídas dos títulos dos artigos e com frequência mínima de 10 ocorrências.

Analisando, por hora, de maneira separada, note que aparentemente pelas maiores frequências, há diferença nas palavras e, talvez assim, na caracterização de cada evento. Nos eventos de EDM, além do reforço ao próprio termo, há uma predominância de termos relacionados aos estudantes (*impact, models, predicting, evaluating, characteristics...*). Nos trabalhos de LA, as ocorrências de palavras distintas do termo principal têm uma distribuição mais próxima, possivelmente indicando focos mais diversificados dos trabalhos analisados.



a) Resultados de análise para EDM e JEDM    b) Resultados de análise do LAK e JLA

Figura 3: Análises realizadas a partir dos títulos dos artigos publicados nos anos de 2015 e 2016 nos congressos e revistas de EDM e LA, respectivamente



grandes volumes de dados, o *Big Data*, mas dentro de um contexto educacional que aponte novos caminhos, tecnologias e métodos para aperfeiçoar os processos de ensino-aprendizagem assim como possibilitar melhores tomadas de decisão pelos gestores.

Estudos analisados apontam o perfil do cientista de dados educacionais, como um profissional versátil, multidisciplinar e conectado com problemas do mundo real, capaz de gerar e compartilhar conhecimentos importantes para o domínio educacional.

A evolução das pesquisas, com a consolidação dos eventos e periódicos da área apontam caminhos promissores para a comunidade de ciências dos dados educacionais.

## Referências

- Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, 4(2), 17.
- Buckingham Shum, S., Hawksey, M., Baker, R. S., Jeffery, N., Behrens, J. T., & Pea, R. (2013, April). Educational data scientists: a scarce breed. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 278-281). ACM.
- Campbell, J. P., & Oblinger, D. G. (2007). Academic analytics. *EDUCAUSE review*, 42(4), 40-57.
- Costa, E., Baker, R. S., Amorim, L., Magalhães, J., & Marinho, T. (2013). Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. *Jornada de Atualização em Informática na Educação*, 1(1), 1-29.
- Daniel, B. K. (Ed.). (2016). *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*. Springer.
- Dede, C. (2015). Data-intensive research in education: Current work and next steps. Computer Research Association. Retrieved from <http://cra.org/cra-releases-report-on-data-intensive-research-in-education>.
- Ducange, P., Pecori, R., Sarti, L., & Vecchio, M. (2016, October). Educational Big Data Mining: How to Enhance Virtual Learning Environments. In *International Conference on European Transnational Education* (pp. 681-690). Springer International Publishing.
- Ifenthaler, D., & Tracey, M. W. (2016). Exploring the relationship of ethics and privacy in learning analytics and design: implications for the field of educational technology. *Educational Technology Research and Development*, 64(5), 877-880.
- Knight, S., & Littleton, K. (2016). Dialogue as Data in Learning Analytics for Productive Educational Dialogue. *Journal of Learning Analytics*, 2(3), 111-143.
- Martinez-Maldonado, R., Schneider, B., Charleer, S., Shum, S. B., Klerkx, J., & Duval, E. (2016, April). Interactive surfaces and learning analytics: data, orchestration aspects, pedagogical uses and challenges. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 124-133). ACM.

- Papamitsiou, Z. K., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4), 49-64.
- Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014, March). Educational data sciences: framing emergent practices for analytics of learning, organizations, and systems. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (pp. 193-202). ACM.
- Quigley, D., Ostwald, J., & Sumner, T. (2017, March). Scientific modeling: using learning analytics to examine student practices and classroom variation. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 329-338). ACM.
- Roberts, L. D., Chang, V., & Gibson, D. (2017). Ethical considerations in adopting a university-and system-wide approach to data and learning analytics. In *Big Data and Learning Analytics in Higher Education* (pp. 89-108). Springer International Publishing.
- Romero, C.; Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, Elsevier, 33(1), 135–146.
- Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, M. (2016). EDUCATIONAL PROCESS MINING. *Data Mining and Learning Analytics: Applications in Educational Research*, 1-28.
- Silva, L. A., Peres, S. M., Boscarioli, C (2016). *Introdução à mineração de dados: com aplicações em R*. 1. ed. Elsevier, Rio de Janeiro, 2016.
- Siemens G. LAK'11 1st International Conference on Learning Analytics and Knowledge. Disponível em < <https://tekri.athabascau.ca/analytics/>> Acesso em: 09 de abril de 2017.
- Siemens, G.; Baker, R. S. (2012). Learning analytics and educational data mining: towards communication and collaboration. In: ACM. *Proceedings of the 2nd international conference on learning analytics and knowledge* (p. 252–254).
- Souza, R., Neto, F. M., Santos, A., Fontes, L., Naassom, E., & Valentim, R. (2016, November). Um Ambiente Inteligente de Avaliação de Comportamentos de Tutores e Turmas no Ambiente Virtual de Aprendizagem Moodle. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* (Vol. 5, No. 1, p. 417).
- Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and ‘real-time’ policy instruments. *Journal of Education Policy*, 31(2), 123-141.