Data Mining Educacional: Uma Revisão da Literatura

Educational Data Mining: A Literature Review

Maria P. G. Martins
Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança
Bragança, Portugal
CISE - Centro de Investigação em Sistemas Electromecatrónicos, Universidade da Beira Interior
Covilhã, Portugal

Vera L. Miguéis
Faculdade de Engenharia da Universidade do Porto, INESC TEC
Porto, Portugal

D. S. B. Fonseca, *Member*, *IEEE*
CISE - Centro de Investigação em Sistemas Electromecatrónicos, Universidade da Beira Interior
Covilhã, Portugal

Resumo — **Com o objetivo de divulgar o potencial e a aptidão de Data Mining Educacional (EDM), como um instrumento de análise e de investigação, no apoio à gestão de instituições dedicadas ao ensino, apresenta-se, no presente artigo, uma sucinta descrição de alguns dos estudos mais relevantes da área. A análise efetuada permite evidenciar as inovações que o EDM tem vindo a promover, bem como as tendências de investigação atuais e futuras.**

Palavras Chave - data mining educacional, data mining, eficiência institucional.

Abstract — **With the aim of disseminating the potential and the capacity of Educational Data Mining (EDM) as an instrument of investigation and analysis in the support to the management of Higher Education Institutions, this paper presents a brief description of some of the most relevant studies in the area. The analysis carried out allows to highlight the innovations that EDM has been promoting, as well as current and future research trends.**

Keywords - educational data mining, data mining, institutional efficiency.

I. INTRODUÇÃO

O uso crescente da tecnologia nos processos de ensino e a evolução permanente dos sistemas informáticos propiciaram que se gerassem e armazenassem grandes volumes de dados associados aos sistemas educacionais. No entanto, as instituições deparam-se com a grande dificuldade que é tratar todos esses dados de forma a transformá-los em informação verdadeiramente útil. A grande volumetria de dados, provenientes de múltiplas fontes e em configurações heterogéneas, excede a capacidade humana de analisar e extrair informação útil, sem a ajuda de técnicas de análise automatizada ([33]). Nos últimos anos tem havido um interesse crescente no uso do *data mining* para investigar questões científicas na área educacional, a qual é designada de Educacional Data Mining (EDM) ([1]).

De acordo com [2] [3] [4] o EDM está ancorado em várias disciplinas de referência, como é o caso dos sistemas de informação, dos sistemas de recomendação pedagógica, da análise visual de dados, do *data mining*, da psicopedagogia e da psicometria. De facto, pode ser desenhada como o resultado de três áreas principais (ver Fig.1): as ciências da computação, a educação e a estatística [3]. A interseção destas três áreas também forma outras subáreas estreitamente relacionadas com o EDM como a educação por computador, *data mining* e aprendizagem automática e aprendizagem analítica (idem).

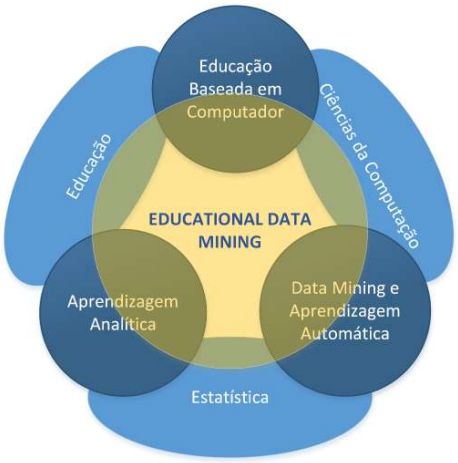

Figura 1. Principais áreas relacionadas com o *data mining* educacional (adaptada de Romero and Ventura [3]).

De acordo a Comunidade Internacional do EDM[1], o EDM é uma disciplina emergente que visa o desenvolvimento de métodos que explorem os dados provenientes dos contextos educacionais para melhor entender os alunos e os ambientes em que eles aprendem. Mas o EDM não se limita a dar resposta a este tipo de questões. Pode vir a ser aproveitado, num futuro próximo, como instrumento de análise e investigação, na identificação de necessidades de aprendizagem prioritárias para os estudantes, que permitam aumentar as taxas de graduação, a avaliação do desempenho institucional, a maximização dos recursos dos *campus* e a otimização dos planos de estudo dos cursos ([4]). De facto, o EDM na atualidade já não se restringe apenas ao uso ou benefício por parte de professores e estudantes, uma vez que tem utilidade para as próprias instituições educativas e até mesmo para os Estados ([5]).

Com o intuito de afirmar e valorizar o EDM como um instrumento de análise e construção de estratégias nos processos de tomada de decisão das instituições académicas, apresenta-se, no presente artigo, uma breve revisão da literatura sobre alguns dos estudos científicos mais citados na área e que são mais orientados para a divulgação das principais contribuições e inovações do EDM. O conjunto de artigos referenciados e analisados foram obtidos através de pesquisas efetuadas nas bases de dados *Web Science of Knowledge, Scopus, Google Scholar, IEEEXplore e Journal of Educational Data Mining*, usando, para o efeito, combinações das palavras-chave "*Educational Data Mining", "Survey", "Review", "State of art", "Dropout Prevision"* e *"Predicting Student's Performance"*. A análise efetuada permitiu evidenciar a utilidade do EDM nos sistemas de apoio à decisão em instituições do ensino superior, bem como identificar algumas oportunidades de investigação futura.

Além desta introdução, na secção que se segue, apresenta-se uma síntese das revisões sistemáticas da literatura; na terceira seção indicam-se quais os métodos de *data mining* mais usados na implementação das tarefas de EDM; por fim, na quarta e última secção apresentam-se as principais conclusões e identificam-se oportunidades de investigação futura.

II. Revisões Sistemáticas da Literatura

Com a tabela que se segue, pretende-se sintetizar as revisões sistemáticas da literatura selecionadas, com uma breve descrição do seu âmbito, objetivos e conclusões respetivas.

TABELA 1. Revisões Sistemáticas

Obra	Objetivo	Conclusões
Romero e Ventura, 2007 [10] (Analisam 81 estudos, publicados entre 1995 e 2005)	Identificar: tendências de investigação da área; sistemas educacionais; métodos e algoritmos de data mining mais usados.	Notória predominância dos estudos em contexto e-learning face à adoção do EDM no ensino tradicional presencial. Enfatizam a necessidade de desenvolvimento de trabalhos relacionados, até à consolidação do EDM como área de investigação.
Baker and Yacef, 2009 [11] (Analisam 48 estudos, publicados entre 2000 e 2009)	Identificar as tendências de evolução do EDM. Determinar quais os métodos e algoritmos de DM mais usados em tarefas de EDM.	São quatro as principais tarefas do EDM: a melhoria dos modelos de ensino; a melhoria dos modelos dominantes, o estudo do suporte pedagógico que os softwares proporcionam e a investigação científica com vista ao desenvolvimento do ensino e dos estudantes. O EDM socorre-se de sete métodos diferentes: a estatística; a visualização; a *web mining*; o *clustering*; a classificação; a associação e o *text mining*.
Romero, Ventura, Pechenizkiy, and Baker, 2010 [5] (Analisam 306 estudos, publicados entre 1993 a 2009)	Aferir a importância do DM no contexto atual da educação. Evidenciar as inovações que o EDM tem vindo a promover. Identificar as funcionalidades onde o DM pode ser útil na área de educação	Reforçam a necessidade de aprofundar estudos, mais unificados e colaborativos, por forma ao EDM consolidar-se como área de investigação madura. Funcionalidades: a criação de feedbacks; o favorecimento da criação de recomendações; a previsão do desempenho; a análise e a visualização de dados; a construção de modelos sobre os estudantes; a possibilidade de agrupar estudantes em função de determinadas características; a possibilidade de detetar tipos de comportamentos dos estudantes; a análise do comportamento na rede social onde o estudante se insere; o planeamento a construção de ferramentas eletrónicas direcionadas para a educação e o desenvolvimento de conceitos. Para concretizar todas estas tarefas de EDM os métodos mais usados continuam a ser a regressão, a classificação, o clustering e a associação. Os algoritmos mais usados são as árvores de decisão, as redes neuronais e as redes bayesianas. O EDM não se restringe apenas ao uso ou benefício de professores e estudantes, uma vez que tem utilidade para as próprias instituições educativas, para os profissionais responsáveis pela criação e desenvolvimento de planos de estudos e até mesmo para os Estados.
Romero e Ventura, 2013 [3] (Analisam 67 estudos, publicados entre 2001 e 2012)	Construir um guia orientador para quem pretende desenvolver estudos na área de EDM. Fornecer uma visão atualizada do estado dos conhecimentos em EDM. Identificação dos tópicos de investigação.	Os principais tópicos de investigação na área de EDM: - (Re)Organização das aulas ou da avaliação, a colocação de materiais com base no uso e dados de desempenho; a identificação daqueles que poderão beneficiar de comentários, conselhos de estudo ou outros géneros de ajuda; a delineação de estratégias para ajudar os alunos a encontrar e pesquisar materiais e bibliografia útil. - A criação de grupos de estudantes de acordo com as suas características de aprendizagem; modelação de alunos para desenvolver e ajustar os seus modelos cognitivos; construção de material didático para ajudar os instrutores e administradores no planeamento de cursos futuros. Indicam como linhas de orientação futura: - A necessidade de desenvolver uma cultura baseada em dados que permita

[1] www.educationaldatamining.org.

Obra	Objetivo	Conclusões
		tomar decisões destinadas a promover a eficiência das instituições. Entre possíveis soluções para este problema, apontam o uso de sistemas de suporte à decisão, mecanismos de recomendação e algoritmos de DM que autonomizem e facilitem aos instrutores todo o processo de EDM. - Reforçam uma vez mais a necessidade crescente de estudos de replicação para testar generalizações mais amplas.
Huebnar, 2013 [2] (Analisa 36 estudos publicados entre 2002 e 2012)	Identificar as formas como o data mining tem sido usado quando se pretende melhorar o sucesso dos aprendizes e os processos diretamente ligados à aprendizagem. Identificar oportunidades de Investigação.	Abarca exclusivamente tópicos como a retenção e o abandono escolar, os sistemas pessoais de recomendação em contextos educativos e as formas como o data mining tem sido usado quando se pretende melhorar o sucesso dos aprendizes ou otimizar os processos diretamente relacionados com a aprendizagem. Os métodos de previsão, de *clustering*, de classificação e de associação, são o paradigma dominante nos modelos analíticos desenvolvidos. Realçam três necessidades: - Estudar formas de tornar os resultados de data mining mais generalizáveis, providenciando o desenvolvimento de modelos que possam ser usados em múltiplos contextos; - O desenvolvimento de sistemas de apoio à decisão e de sistemas de recomendação que minimizem a intervenção dos educadores; - O desenvolvimento de ferramentas que protejam a privacidade individual dos intervenientes, ao mesmo tempo que possibilitam a EDM.
Papamitsiou and Economides, 2014 [4] (Analisam 40 estudos, publicados entre 2008 e 2013)	Identificar quais os métodos mais usados na literatura existente, para determinar a eficácia da implementação do EDM. Verificar até que ponto os métodos têm contribuído para a implementação do EDM como instrumento de análise e investigação.	Entre os métodos mais adotados surge em primeiro lugar o método de classificação, seguindo-se o método de clustering e de associação. Mais recentemente, já foram identificados estudos que comportavam novos métodos como, o *text mining*, o *association rule mining*, o *social network analysis*, o *discovery with models* e a visualização. Todos estes instrumentos terão sido aplicados com mais incidência em estudos de modelação comportamental dos estudantes e também na determinação de formas mais eficazes de prever o seu desempenho.
Peña-Ayala, 2014 [12] (Analisam 240 estudos, publicados entre 2010 e o 1º período de 2013)	Identificar sistemas educacionais; tópicos de investigação; e métodos e algoritmos usados.	No conjunto dos 222 artigos analisados na perspetiva da caracterização e das funcionalidades do EDM, quase 82% dos estudos estão relacionados com as três versões de modelação de alunos: comportamental, desempenho e geral. O complemento (18%) foi distribuído por abordagens no âmbito do suporte pedagógico e feedback dos alunos,

Obra	Objetivo	Conclusões
		domínio de conhecimento (habilidades a serem treinadas) e suporte a professores. Nos 18 artigos que deram ênfase quer à funcionalidade quer aos instrumentos do EDM, os autores identificaram que 8 (a maioria) apresentaram o EDM como meio de análise, 6 justificam-no como ferramenta de visualização de dados e os restantes consideraram esta metodologia como forma de extrair informação de bases de dados. É também realçada a necessidade de aprofundamento dos estudos que visem a procura de modelos práticos de aplicação.
Sukhija, Jindal, Aggarwal, 2015 [9] (Analisam 19 estudos, publicados entre 2001 e 2015)	Promover e valorizar o EDM enquanto ferramenta de análise e construção de estratégias nos processos de tomada de decisão nas instituições académicas.	O EDM é ainda uma disciplina em expansão, a que está associado um vasto conjunto de métodos e algoritmos, e que continua a exigir atenção por parte dos investigadores, sobretudo ao nível do alargamento do conjunto de algoritmos que permitam a hibridização das técnicas de análise e agrupamento. Concluem que o EDM pode vir a constituir-se num instrumento que permita aos professores, estudantes e administrações educativas beneficiar do melhor que eles próprios tenham para oferecer.
Shahiri, Husain, and Rashid, 2015 [8] (Analisam 30 estudos, publicados entre 2002 e janeiro de 2015)	Revisão sobre os preditores de desempenho académico. Determinação dos algoritmos mais usados na mesma tarefa.	Concluem que são 6 os atributos usados com mais frequência: CGPA (Cumulative Grade Point Average); indicadores de avaliação interna pós ingresso no ensino superior; características demográficas; indicadores de avaliação externa pré-ingresso; características relacionadas com a interação social dos estudantes. Os algoritmos *Decision Tree* (DT), *Artificial Neural Networks* (ANN), *Naive Bayes* (NB), *K-Nearest Neighbor* (K-NN) e o *Support Vector Machine* (SVM) foram, por ordem decrescente de importância, os mais usados nos trabalhos analisados.
Del Río and Insuasti, 2016 [7] (Analisam 51 estudos, publicados entre 2011 e agosto de 2016)	Revisão sobre os preditores de desempenho académico, mas delimitado ao sistema presencial tradicional. Determinação dos métodos e algoritmos de data mining usados na tarefa que deu enfase à revisão literária.	Para inferir a média final de curso, os investigadores usaram apenas indicadores de desempenho académico após o ingresso no ensino superior, em 37.5% dos estudos, e em combinação com mais um outro tipo de atributo, em 51.8% dos casos. Entre os métodos de *data mining* mais usados, os autores destacaram o de classificação, dado que foi reportado em 71.4% das investigações referenciadas. Os métodos de agrupamento e de regras de associação foram os que se seguiram, tendo incidido, respetivamente, em 8.9% e 7.1% dos estudos.
Kumar, Singh, and Handa, 2017 [6] (Analisam 14 estudos, publicados entre 2009 a 2016)	Revisão sobre os preditores de desempenho académico; Determinação dos métodos e algoritmos usados na mesma tarefa.	Na maioria dos estudos revistos, a média de acesso, o nível educacional e ocupação dos pais, e uma metodologia de ensino pobre são os principais fatores que afetam o resultado dos alunos. Os investigadores recorreram, essencialmente, aos métodos de classificação e associação, com cerca de

Obra	Objetivo	Conclusões
	Identificar as funcionalidades onde o EDM pode ser útil.	50% dos trabalhos referenciados a reportarem estes dois métodos. Identificam as seguintes funcionalidades do EDM: com a análise de padrões comportamentais de estudo em cursos online, com a previsão dos resultados académicos dos alunos, com a previsão do ranking do aluno, com a análise dos hábitos de aprendizagem online, com a análise de cursos MOOC, com previsões de progressos ou retrocessos dos estudantes e predominantemente com previsão de abandono escolar. Sublinharam ainda que a previsão do abandono escolar é tida como uma tarefa importante e desafiadora, para os investigadores, professores, instituições de ensino e até para os decisores políticos, confirmando-se a elevada contribuição do EDM em tarefas desta tipologia, nomeadamente quando se pretende identificar as características dos estudantes propensos ao abandono.

Apesar da relevância que o EDM tem adquirido, como é percetível a partir das conclusões dos trabalhos destacados, bem como do crescente número de estudos que têm vindo a surgir, alguns autores destacam algumas lacunas que dificultam o processo de consolidação do EDM. Por exemplo, Sukhija, Jindal and Aggarwal [9] identificam: a indisponibilidade de bases de dados consistentes e suficientemente abrangentes; a falta de versatilidade das bases de dados que sustentam o funcionamento do data mining na educação; as próprias ferramentas do EDM e, em particular, os seus algoritmos, por se traduzirem normalmente em instrumentos inflexíveis e ainda pouco aptos à utilização conjunta; e a falta de confiança que as entidades responsáveis, nomeadamente os governos, demonstram ainda ter, face aos resultados obtidos através do processo de EDM.

Também Huebner [2] adverte que a investigação nesta área desenvolve-se de forma isolada e não se conhece com exatidão de que forma as instituições têm implementado as metodologias que visam melhorar a aprendizagem dos alunos ou os respetivos processos educacionais.

Peña-Ayala [12] aponta igualmente, como aspeto penalizador, a falta de reconhecimento daquelas que são as verdadeiras capacidades do EDM em ampliar e melhorar as conquistas tradicionais dos sistemas educacionais. Adverte ainda que a maioria das abordagens EDM relacionam-se essencialmente com a implementação do DM para explorar assuntos educacionais, não dando propriamente contributos para a área de data mining. Ainda que nos primeiros anos abrangidos pelo estudo de revisão [12], os investigadores referenciados tenham recorrido essencialmente aos métodos preditivos, rapidamente começaram a ser usados com alguma frequência os métodos de *clustering*, de classificação e de regressão.

III. Principais Métodos de Data Mining Usados em EDM

A literatura revista demonstra que os métodos mais populares e transversais a todos os domínios do data mining, como a previsão, que se consegue através da classificação e regressão, o *clustering*, análise de associação e visualização, têm sido aplicados com sucesso no domínio educacional. Na classificação a variável a prever é uma variável binária ou categórica. O objetivo da classificação é encontrar um modelo que identifique a qual classe um determinado registo pertence. A cada uma das classes identificadas corresponde um rótulo, que pode ter vários valores discretos, como por exemplo, um aluno pertencer à classe "sucesso" ou "insucesso".

Os modelos de regressão têm como objetivo prever os valores futuros (ou desconhecidos) de uma ou mais variáveis numéricas contínuas, a partir de outros atributos presentes no conjunto de dados.

Os métodos de previsão têm-se revelado de grande utilidade na previsão e compreensão do desempenho académico e comportamentos futuros dos estudantes, em geral. Por exemplo, foi esse tipo de métodos que os autores Natek and Zwilling [13], Aluko, Adenuga, Kukoyi, Soyingbe and Oyedeji [14], e Rubiano and Garcia [15], usaram para inferir o (in)sucesso académico global dos estudantes – expresso, em todo eles, pelo indicador de sucesso média final de curso. Os mesmos métodos têm sido igualmente úteis em previsões precoces de reprovação ou desistência em disciplinas específicas (ver, por exemplo, [16] [17] [18] [19]) e em previsões de abandono escolar (ver, por exemplo, [15] [17] [18] [20] [21] [22] [23] [24] [25] [26] [27]).

Também o método de *clustering* tem sido muito popular no âmbito do EDM. De acordo com [3] [4] [5] [33] os procedimentos de agrupamento tem tido aplicações diversas: para formar grupos de estudantes de acordo com suas características pessoais e respetivos dados de aprendizagem; para encontrar padrões de resolução eficaz de problemas em ambientes de aprendizagem baseados em computador; para encontrar semelhanças ou diferenças entre estudantes, ou mesmo entre escolas; para categorizar um estudante recém ingressado de acordo com as suas características; para fornecer recomendações personalizadas baseadas na capacidade do aluno; para indicar materiais de estudo personalizados com base na habilidade e conhecimentos do aluno.

Bydžovská [19], Cerezo, Santillán, Ruiz and Núñez [28] Tiwari, Singh and Vimal [30] também recorreram ao método de *clustering* com o objetivo de segmentar os alunos com comportamentos semelhantes e diferentes níveis de desempenho, para assim preverem o seu nível de desempenho. Num outro estudo foi usado para investigar a estrutura de redes sociais e avaliar a sua influência na aprendizagem dos estudantes (Marcos et al. [29]). Também para agrupar os estudantes em função das suas interações com os conteúdos de aprendizagem dispostos em plataformas informáticas (Kizilcec, Piech and Schneider [31]). No estudo [32], através do algoritmo de K-means, foram segmentados os alunos de acordo com o seu desempenho com o objetivo de se construir um modelo que permita comparar os estudantes em geral com o padrão de estudante ideal. O método permitiu identificar estratégias para melhorar o desempenho dos estudantes (Campagni, Merlini, Sprugnoli and Verri [32]).

O método de associação, ou de afinidade de grupos, usa-se para descobrir relações interessantes entre os atributos presentes em

grandes repositórios de dados. Também no contexto da educação o método de associação tem sido usado com êxito no suporte à decisão em instituições dedicadas ao ensino. De acordo com Algarni [33], o método é útil sempre que se pretende encontrar a relação entre o nível de educação dos pais e o risco de abandono escolar dos estudantes, na descoberta de associações curriculares, no desenvolvimento de estratégias pedagógicas para uma aprendizagem mais eficaz, e também na descoberta de relações comportamentais entre estudantes. No estudo [17] através de sequência de padrões, foram revelados padrões de relacionamento entre o desempenho académico e a data e ordem da realização dos exames. O método de associação é igualmente útil quando se pretende relacionar a pontuação obtida nas unidades curriculares dos alunos com o seu tempo de estudo (Tiwari, Singh and Vimal [30]), ou recomendar atividades em contexto de ensino *e-learning,* como por exemplo, recomendar materiais de estudo aos alunos (Romero and Ventura [34]). As regras de associação permitem ainda relacionar o sucesso ou insucesso numa disciplina com o sucesso ou insucesso numa outra disciplina.

Porque "uma imagem vale por mil palavras", a identificação de padrões de aprendizagem e diferenças individuais dos estudantes, a partir da visualização, é um método chave para exploração de bases de dados educativas (Baker [35]). De acordo com Romero and Ventura [3], a deteção de valores atípicos através da visualização também pode ser usada para identificar desvios no comportamento dos alunos, ou até dos educadores, para localizar estudantes com dificuldades de aprendizagem e para detetar processos de aprendizagem irregulares.

Outros métodos, como por exemplo, a destilação de dados para julgamento humano, a descoberta com modelos e a factorização da matriz não negativa também têm demonstrado nos últimos tempos uma proeminência particular na análise e exploração de dados gerados no contexto educativo (Romero and Ventura [3]).

IV. Conclusões e Oportunidades de Investigação Futura

Sendo o EDM uma tendência de investigação emergente, que alguns autores consideram estar ainda na fase de "infância", a primeira evidência que transparece da abordagem global à literatura existente, é o facto de haver ainda muito a estudar e a considerar acerca das metodologias a aplicar nesta área de intervenção, da sua real capacidade de abrangência e também das possibilidades de aplicação concreta do EDM. A maioria dos autores das revisões sistemáticas analisadas realçaram a necessidade de aprofundamento dos estudos, providenciando o desenvolvimento de modelos práticos e generalizáveis de aplicação em múltiplos contextos.

Foi igualmente salientado que a maioria das abordagens de EDM prende-se, essencialmente, com a implementação de técnicas de data mining com o objetivo de explorar assuntos educacionais, não contribuindo propriamente para o desenvolvimento da área.

Pretendendo-se demonstrar o potencial de data mining no contexto dos sistemas de apoio à decisão de instituições de ensino, referiram-se as funcionalidades de EDM identificadas no conjunto das revisões sistemáticas apresentadas.

Face às funcionalidades de EDM identificadas sobressai que o alvo preferido dos investigadores é a modelação de estudantes, principalmente, a antecipação do seu desempenho académico, as análises comportamentais e de conhecimento de domínio. Mas, tal como foi referido pela generalidade dos autores, outras funcionalidades, como por exemplo, o suporte aos alunos e aos professores, os ambientes pessoais de aprendizagem e até as próprias ferramentas de data mining, reivindicam o foco e o interesse da comunidade de EDM.

Quanto aos instrumentos mais usados pelos investigadores, é notória a predominância da previsão, seguindo-se o clustering, a associação e a visualização. De entre os métodos de classificação sobressaí que há relativamente menos artigos que reportam o uso de métodos conjuntos, o que abre perspetivas para que se investigue a pertinência da utilização destes métodos.

A maioria dos estudos sobre previsão de desempenho académico identifica os fatores que o possam influenciar, fundamentais para a definição de estratégias de gestão centradas na promoção do sucesso e na prevenção do abandono escolar. Neste contexto sobressaí que aqueles que estão relacionados com indicadores de desempenho académico, pré e pós-ingresso no ensino superior, têm sido, globalmente, os mais explicativos para as previsões efetuadas. De salientar, no entanto, que há um conjunto de outros fatores que não têm sido reportados, mas que poderão, igualmente, desempenhar um papel significativo nas previsões efetuadas. Por exemplo, informação sobre o círculo de amigos, que podem influenciar os hábitos de sono, os métodos de estudo, a assiduidade às aulas, a frequência de horário de atendimento e outras dimensões com um vínculo direto com o desempenho acadêmico, poderão vir a revelar-se importantes para a precisão das previsões efetuadas.

Perspetiva-se que no futuro as investigações sobre o EDM aumentem de forma significativa, pelo facto de abarcar metodologias muito promissoras que têm permitido a descoberta de conhecimento oculto de padrões dos alunos em ambientes educacionais. Acredita-se que o EDM deva ser visto como uma ferramenta a ser adotada pelas instituições de ensino, pois poderá vir a ser uma mais-valia na construção de estratégias de promoção educacional, não obstante a necessidade firmada de desenvolvimento e melhoria dos estudos a serem realizadas acerca desta temática.

Agradecimentos

Este trabalho foi suportado pela Fundação para a Ciência e Tecnologia (FCT) através do Projeto UID/EEA/04131/2013.

Referências Bibliográfica

[1] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions", JEDM-Journal of Educational Data Mining, vol. 1(1), 2009, pp. 3–17.

[2] R. A. Huebner, "A survey of educational data-mining research", Research in higher education journal, vol. 19, 2013.

[3] C. Romero and S. Ventura, "Data mining in education", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3(1), 2013, pp. 12–27.

[4] Z. K. Papamitsiou and A. A. Economides, “Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence”, Educational Technology & Society, vol. 17(4), 2014, pp. 49–64.

[5] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. D. Baker, “Handbook of educational data mining”, CRC Press, 2010.

[6] M. Kumar, A. J. Singh, and D. Handa, “Literature survey on educational dropout prediction”, International Journal of Education and Management Engineering (IJEME), vol. 7(2), 2017, pp. 8–19.

[7] C. A. Del Río and J. A. P. Insuasti, “Predicting academic performance in traditional environments at higher-education institutions using data mining: A review”, Ecos de la Academia, vol. 2016(7), 2016.

[8] A. M. Shahiri, W. Husain, and N. A. Rashid, “A review on predicting student’s performance using data mining techniques”, Procedia Computer Science, vol. 72, 2015, pp. 414–422.

[9] K. Sukhija, M. Jindal, and N. Aggarwal, “The recent state of educational data mining: A survey and future visions”, In MOOCs, IEEE 3rd International Conference on Innovation and Technology in Education (MITE), 2015, pp. 354–359.

[10] C. Romero and S. Ventura, “Educational data mining: A survey from 1995 to 2005”, Expert systems with applications, vol. 33(1), 2007, pp. 135–146.

[11] R. S. J. D. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions”, JEDM-Journal of Educational Data Mining, vol. 1(1), 2009, pp. 3–17.

[12] A. Peña-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works”, Expert systems with applications, vol. 41(4), 2014, pp. 1432–1462.

[13] S. Natek and M. Zwilling, “Student data mining solution-knowledge management system related to higher education institutions”, Expert systems with applications, vol. 41(14), 2014, pp. 6400–6407.

[14] R. O. Aluko, O. A. Adenuga, P. O. Kukoyi, A. A. Soyingbe, and J. O. Oyedeji, “Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques”, Construction Economics and Building, vol. 16(4), 2016, pp. 86–98.

[15] S. M. M. Rubiano and J. A. D. Garcia, “Analysis of data mining techniques for constructing a predictive model for academic performance”, IEEE Latin America Transactions, vol. 14(6), 2016 pp. 2783–2788.

[16] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, “Data mining algorithms to classify students”, In Educational Data Mining 2008.

[17] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers”, Predicting students drop out: A case study”, In Educational Data Mining 2009.

[18] T. A. Pascoal, D. M. Brito, and T. G. Rêgo, “Uma abordagem para a previsão de desempenho de alunos de computação em disciplinas de programação”, Nuevas Ideas en Informática Educativa TISE, 2015, pp. 454–458.

[19] H. Bydžovská, “A comparative analysis of techniques for predicting student performance”, In Proceedings of the 9th International Conference on Educational Data Mining 2016.

[20] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, “Preventing student dropout in distance learning using machine learning techniques”, In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2003, pp. 267–274.

[21] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, “Dropout prediction in e-learning courses through the combination of machine learning techniques”, Computers & Education, vol. 53(3), 2009, pp. 950–965.

[22] L. M. B. Manhães, “Predicting Academic Performance of Undergraduate Students Using Educational Data Mining” (*Predição do Desempenho Académico de Graduandos Utilizando Mineração de Dados Educacionais*), PhD thesis (*Tese Doutorado*), Universidade Federal do Rio de Janeiro, 2015.

[23] C. M. Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. M. Fardoun, and S. Ventura, “Early dropout prediction using data mining: a case study with high school students”, Expert Systems, vol. 33(1), 2016, pp. 107–124.

[24] D. Delen, “A comparative analysis of machine learning techniques for student retention management”, Decision Support Systems, vol. 49(4), 2010, pp. 498–506.

[25] A. Nandeshwar, T. Menzies, and A. Nelson, “Learning patterns of university student retention”, Expert Systems with Applications, vol. 38(12), 2011, pp.14984–14996.

[26] M. Sweeney, H. Rangwala, J. Lester, and A. Johri, “Next-term student performance prediction: A recommender systems approach”, Journal of Educational Data Mining, vol. 8(1), 2016, pp. 22–51.

[27] S. Lehr, H. Liu, S. Kinglesmith, A. Konyha, N. Robaszewska, and J. Medinilla, “Use educational data mining to predict undergraduate retention”, In 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 2016, pp. 428– 430.

[28] R. Cerezo, M. S. Santillán, M. P. P. Ruiz, and J. C. Núñez, “Students’ lms interaction patterns and their relationship with achievement: A case study in higher education”, Computers & Education, vol. 96, 2016, pp. 42–54.

[29] L. Marcos, E. G. López, A. G. Cabot, J. A. M. Merodio, A. Domínguez, J. J. M. Herráiz, and T. D. Folledo, “Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance”, Computers in Human Behavior, vol. 60, 2016, pp. 312–321.

[30] M. Tiwari, R. Singh, and N. Vimal, “An empirical study of applications of data mining techniques for predicting student performance in higher education”, International Journal of Computer Sciences and mobile Computing, vol. 2(2), 2013, pp. 53–57.

[31] R. F. Kizilcec, C. Piech, and E. Schneider, “Deconstructing disengagement: analyzing learner subpopulations in massive open online courses”, In Proceedings of the third international conference on learning analytics and knowledge, ACM, 2013, pp. 170–179.

[32] R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, “Data mining models for student careers”, Expert Systems with Applications, vol. 42(13), 2015, pp. 5508–5521.

[33] A. Algarni, “Data mining in education”, International Journal of Advanced Computer Science and Applications, vol. 7, 2016, pp. 456–461.

[34] C. Romero and S. Ventura, “Educational data mining: a review of the state of the art”, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40(6), 2010, pp. 601–618.

[35] R. S. J. D. Baker, “Data mining for education”, International encyclopedia of education, vol. 7 (3), 2010, pp. 112–118.