Revista EaD &
tecnologias digitais na educação

Mineração de dados educacionais em um *mooc* brasileiro

Vanessa Faria Souza, UFRGS

vanessa.souza@ibiruba.ifrs.edu.br

***Resumo**: No contexto atual da educação a distância, os Learning Management System (LMS) permitem o armazenamento de grande volume de dados sobre as atividades realizadas e para compreender a respeito do padrão de comportamento dos alunos nesse ambiente é preciso que os educadores e gestores repensem as abordagens tradicionais de análise desses dados, sendo essencial a utilização de soluções computacionais apropriadas, como a Mineração de Dados Educacionais (MDE). Este tem como objetivo a aplicação de algoritmos de MDE e análise dos resultados de um MOOC brasileiro com 702 alunos. Como resultados apresenta-se o tipo de atributo que contribuiu de maneira mais significativa para conclusão dos alunos e o padrão de comportamento de grupos de alunos que desistem.*

***Palavras-chave**: Mineração de Dados Educacionais, MOOCs.*

***Abstract**: In the current context of distance education, the Learning Management System (LMS) allows the storage of a large volume of data on the activities carried out and in order to understand about the behavior pattern of students in this environment, educators and managers must rethink the approaches traditional methods of analyzing these data, it is essential to use appropriate computational solutions, such as Educational Data Mining (MDE). This aims to apply MDE algorithms and analyze the results of a Brazilian MOOC with 702 students. The results show the type of attribute that contributed most significantly to the completion of students and the pattern of behavior of groups of students who drop out.*

***Keywords**: Educational Data Mining, MOOCs.*

1. Introdução

No atual cenário do ensino superior, a modalidade a distância tem apresentado um expressivo crescimento em relação ao número de alunos matriculados nos últimos

anos (Allen e Seaman, 2015). A partir do surgimento dos MOOCs (Massive Open Online Course), ocorreu uma mudança de dimensão a respeito da quantidade de alunos inscritos em um único curso, pois em razão de serem totalmente online, sem pré-requisitos e também por não exigirem pagamento inicial de taxas, tais cursos têm atraído, em geral, expressivo volume de alunos (Hyman, 2012; Cooper e Sahami, 2013).

O aspecto massivo presente em seu acrônimo pode ser destacado, por exemplo, com o curso, Introduction to Computer Science I, oferecido pela Universidade de Harvard com parceria com a provedora edX. Esse é um MOOC que chegou a 150.349 alunos matriculados. Não é comum cursos com mais de 100.000 alunos, e um MOOC típico apresenta em média 25.000 alunos matriculados (Jordan, 2015).

Nos MOOCs, os Ambientes Virtuais de Aprendizagem (AVA) ou LMS (Learning Management System) comerciais e de código aberto, assim como os ambientes virtuais utilizados pelas grandes provedoras como Cousera e edX são o elemento central de qualquer projeto. Esses cursos são ministrados de forma "automática", pois são baseados em videoaulas, atividades com correção automática e projetos com avaliação pelos pares, sem o contato do aluno com um tutor.

Os fóruns de discussão são importantes para apoiar a colaboração entre os pares, permitindo aos alunos obterem informações e também interação social com os outros alunos. Apesar de existir uma trilha de aprendizagem previamente definida, os próprios alunos podem gerenciar sua aprendizagem (Nanfito, 2014; You, 2016). Uma enorme quantidade de dados sobre a navegação, atividades realizadas, interação com o material didático e com os outros alunos é registrada e coletada possibilitando que sejam elaboradas análises relacionadas ao padrão de comportamento dos alunos no ambiente, e atualmente os LMSs incluem módulos que registram automaticamente cada evento ocorrido no ambiente.

Essas análises permitem direcionar melhor a relação do aluno com o curso e podem predizer suas dificuldades e oferecer também reforço quando for necessário, sendo portanto um material rico para permitir o autogerenciamento do curso (Pardo e Kloos, 2011; Hu, Lo, e Shih, 2014; Nanfito, 2014; You, 2016). Geralmente esses dados gerados pelos LMSs não podem ser analisados adequadamente por softwares aplicativos básicos como uma planilha eletrônica ou por mecanismos tradicionais de análise estatística ou ferramentas para acesso a banco de dados transacionais, em razão de fatores como, a enorme quantidade de registros, elevado número de atributos, valores ausentes, presença de dados qualitativos e não quantitativos, entre outros.

Os dados coletados de cursos massivos possibilitam que os educadores e gestores repensem as abordagens tradicionais de análise, e a utilização de soluções computacionais vem se consolidando como o caminho mais apropriado (Bala e Ojha, 2012; Romero e Ventura, 2013; Muñoz-Merino, Ruipérez-Valiente, Alario-Hoyos, e Perez Sanagustin, 2014; Crossley, Paquette, Dascalu, Mcnamara, e Baker, 2016). O desenvolvimento e uso de ferramentas computacionais para a análise de dados, como Data Mining e Learning Analytics, no campo da educação, foi bastante tardio, em comparação com as áreas de ciências, como biologia e física, além de outras como marketing, manufatura e finanças. A aplicação de tais técnicas tem enorme potencial de transformação, para, por exemplo, prever o desempenho dos alunos e também compreender o comportamento deles no processo de ensino e aprendizagem. (Siemens e Long, 2011; Bala e Ojha, 2012; Romero e Ventura, 2013; Baker, 2014; Natek e Zwilling, 2014).

ISSN 2318-4051

Há uma área de pesquisa, relativamente recente, conhecida como "mineração de dados educacionais" (Educational Data Mining - EDM), que possibilita a compreensão do desempenho e padrão de comportamento dos alunos analisando os dados do LMS. (Romero e Ventura, 2010; Chatti, Dyckhoff, Schroeder, e Thüs, 2012; Calders e Pechenizkiy, 2012; Campagni, Merlini, Sprugnoli, e Verri, 2015). O objetivo do presente artigo é analisar as contribuições e restrições da aplicação de métodos de mineração de dados educacionais em um conjunto de dados de um curso massivo.

No caso desta pesquisa, a contribuição principal reside na aplicação de MDE sob os dados gerados no Curso Esportes e Atividades ao Ar Livre, disponível na plataforma Lúmina, LMSs especializada em cursos MOOCS da Universidade Federal do Rio Grande do Sul (UFRGS) com 702 alunos matriculados. Esse curso apresenta uma temática bastante atual, pois nos últimos anos tem se tornado perceptível a ampliação do número de praticantes de canoagem, trekking, surfe, skate, stand up paddle, slackline, acampamento, entre outros esportes e atividades ao ar livre.

O curso tem chamado a atenção dos gestores do Lúmina, pois tem obtido uma alta taxa de conclusão, se comparado a outros MOOCs ofertados pela plataforma, na sua primeira edição, que é avaliada no decorrer desse artigo, obteve 49% de alunos concluintes. Como comparação elenca-se três MOOCs disponíveis na mesma plataforma: (1) O setor de games no Brasil: panorama, carreiras e oportunidades, que obteve uma taxa de finalização de apenas 15%, (2) Análise de Sentimentos em Computação com 16%, (3) Avaliação de Usabilidade 2ª Edição, que atingiu 24% de alunos concluintes.

Dessa forma, espera-se contribuir para melhorar o processo de análise e tomada de decisão por parte dos professores e gestores de MOOCs, para melhorar o processo de aprendizagem e aumentar o nível de permanência dos alunos nos cursos. Além desta introdução, este trabalho foi dividido em mais cinco partes. Na segunda parte é apresentada uma fundamentação teórica sobre o processo de KDD e mineração de dados educacionais. Na sequência, apresentam-se a metodologia da pesquisa, os dados do curso Esportes e Atividades ao Ar Livre e a aplicação dos algoritmos de mineração de dados, a discussão e considerações finais e, por último, as referências bibliográficas.

1.1 Trabalhos na área de mineração de dados educacionais

Na literatura é possível encontrar trabalhos correlatos, ou seja, sobre a utilização de mineração de dados em diversos contextos educacionais, predominantemente com grupos reduzidos de alunos. O trabalho de Yadav, Bharadwaj & Pal (2012) utiliza a técnica de árvore de decisão com a aplicação de três diferentes algoritmos para analisar dados de 48 estudantes de turmas que já concluíram seus estudos, com o objetivo de gerar um modelo para previsão de desempenho dos estudantes da turma atual, possibilitando que os professores consigam identificar aqueles alunos que necessitam de maior grau de atenção durante as atividades do semestre, visando aumentar a taxa de aprovação e também avaliar medidas a serem adotadas para os próximos semestres.

Outro trabalho de pesquisa é o de Romero, Zafra, Luna, e Ventura (2013), aplicando algoritmos de regras de associação como Apriori e FP-Growth para descobrir associações entre os atributos de 104 alunos que realizaram testes (quizzes) no LMS Moodle. A partir da descoberta de regras, foi possível fornecer aos professores informações para melhorar os testes.

Em outra pesquisa, Natek e Zwilling (2014) concentram-se na mineração de dados para pequenos conjuntos de dados (máximo de 106 alunos), utilizando diferentes algoritmos de árvore de decisão para prever a taxa de sucesso dos alunos da turma em

curso, com base no desempenho de turmas anteriores da disciplina de Informática de um curso de Economia. A conclusão da pesquisa indica que o uso dessas técnicas em ambiente real pode ser útil e promissor, podendo fornecer aos administradores ferramentas preciosas para a tomada de decisão.

A pesquisa de Campagni, Merlini, Sprugnoli, e Verri (2015) utiliza mineração de dados educacionais para também analisar pequenos conjuntos de dados, no caso, os percursos acadêmicos de 141 alunos de Ciência da Computação da Universidade de Florença na Itália. No trabalho, foram utilizadas diferentes abordagens baseadas em técnicas de agrupamento e padrões sequenciais para identificar estratégias para melhorar o desempenho dos alunos e a programação dos exames. Como resultado, os gestores puderam inserir alterações no curso, como a inclusão de professores tutores para orientar os alunos na sua vida acadêmica, como, por exemplo, na escolha de disciplinas.

Em relação especificamente aos algoritmos de mineração de dados educacionais, outras pesquisas podem ser mencionadas, como o trabalho de Shahiri, Husain, e Rashid (2015), que apresenta por meio de uma revisão da literatura, quais algoritmos de predição seriam os mais utilizados para identificar os atributos mais importantes para a performance em um determinado conjunto de dados de estudantes. Após a pesquisa, os autores concluiram que os principais algoritmos citados para predição da performace de alunos são os de Árvore de Decisão (decision tree) e Redes Neurais (neural network).

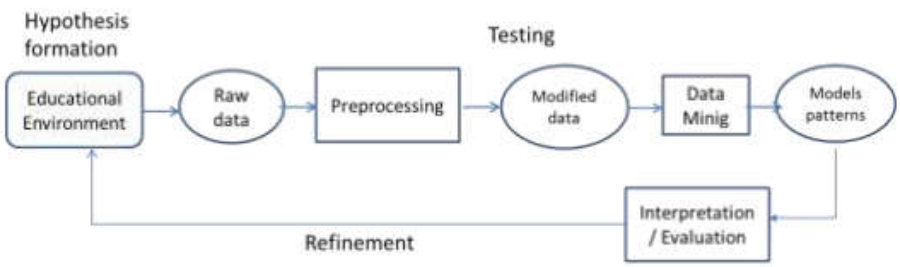
Finalmente, o trabalho de Dutt, Aghabozrgi, Ismail, & Mahroeian (2015) apresenta uma revisão da literatura a respeito dos principais algoritmos de agrupamento (clustering), identificando que K-means é o mais utilizado em trabalhos envolvendo MDE.

2. Fundamentação teórica

2.1. KDD e mineração de dados

No cenário da educação superior com a grande difusão de sistemas informatizados, cresce a cada dia o volume de dados gerados e armazenados em bases de dados (Rigo, Cambruzzi, Barbosa, e Cazella, 2014). Este grande volume de dados tem propiciado a utilização em contextos educacionais de uma área denominada Descoberta de Conhecimento em Banco de Dados ou Knowledge Discovery in Databases (KDD). Uma das definições mais aceitas para KDD foi a proposta inicialmente por Fayyad, Piatetsky-Shapiro, & Smyth (1996), conforme pode ser observado na figura 1, que corresponde a um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis, a partir de grandes conjuntos de dados.

Figura 01: O Processo de KDD.



Fonte: Adaptado de Romero e Ventura (2013)

O processo de KDD depende inicialmente do ambiente educacional (educational environment), pois diferentes tipos de dados podem ser coletados, por exemplo, a partir

de um ambiente de educação presencial ou educação a distância, além do tipo de LMS utilizado e também das fontes de dados disponíveis (dados administrativos, do LMS, questionários,dentre outros).

Coletar e integrar esses dados brutos não é uma tarefa trivial. A etapa seguinte de pré-processamento é essencial nesse processo (Romero e Ventura, 2013). A etapa de pré-processamento (preprocessing) consiste no tratamento e na preparação dos dados. Nessa etapa devem-se identificar, corrigir e retirar valores inválidos, inconsistentes ou redundantes. Por exemplo, a limpeza dos dados trataria da definição de um possível intervalo de valores (domínio) para um determinado atributo.

Caso surgisse algum valor diferente do definido no domínio, esse valor deve ser corrigido ou mesmo eliminado da base de dados. Na sequência, a etapa de transformação (modified data), abrange, quando necessário, alguma transformação linear ou mesmo não linear nos dados, de forma a encontrar aqueles mais relevantes para o problema em estudo. Geralmente são aplicadas técnicas de redução de dimensionalidade e de projeção dos dados (Elmasri e Navathe, 2011).

A etapa seguinte de mineração de dados (data mining) deve ser entendida como uma das etapas do processo mais amplo de KDD e utiliza algoritmos específicos para a extração de padrões dessas bases de dados (Rigo, Cambruzzi, Barbosa, e Cazella, 2014) A etapa final de interpretação consiste na análise dos resultados da mineração e na geração de conhecimento pela interpretação e utilização dos resultados em benefício da aplicação em questão. Etapa complexa, em que são identificados os padrões pelo sistema, estes são interpretados em conhecimentos e validados, para em seguida proporcionarem suporte a tomada de decisões humanas (Elmasri e Navathe, 2011).

A Mineração de Dados Educacionais (MDE) ou Educational Data Mining (EDM) trata da aplicação das técnicas da Mineração de Dados junto aos novos conjuntos de dados obtidos nos diversos ambientes educacionais. A MDE utiliza predominantemente as técnicas de classificação (classification), regras de associação (association rules) e agrupamento (clustering). (Romero e Ventura, 2013; Hu, Lo, e Shih, 2014; Campagni, Merlini, Sprugnoli, e Verri, 2015).

2.2. Principais técnicas para MDE

A Mineração de Dados Educacionais emprega técnicas comuns de mineração de dados, e as principais são as seguintes: Na descoberta de Regras de Associação, o banco de dados é considerado um conjunto de transações. Cada transação é composta por um conjunto de itens que frequentemente ocorrem de forma simultânea em transações do conjunto de dados.

Uma regra de associação tem a forma X => Y, onde X = {x1, x2, ..., xn} e Y = { y1, y2, ..., yn} são conjuntos de itens , com xi e yj, sendo itens distintos para todo i e j. Essa associação indica que, se um cliente compra X, provavelmente comprará Y. Pode ser aplicado, por exemplo, na área de marketing para se descobrir pessoas que compram de forma associada dois produtos diferentes. Algoritmos como Apriori, GSP e DHP são exemplos da implementação da tarefa de Descoberta Regras de Associação (Elmasri & Navathe, 2011).

A classificação é uma forma de análise de dados que extrai modelos que descrevem classes de dados importantes. A tarefa de classificação consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram (Elmasri e Navathe, 2011).

A árvore de decisão é bastante representativa em relação à técnica de classificação, sendo um método adequado quando o objetivo da mineração é a classificação de dados ou predição de saídas. Uma árvore de decisão possui uma estrutura de árvore, em que cada nó interno (não-folha), pode ser entendido como um atributo de teste, e cada nó-folha (nó-terminal) possui um rótulo de classe.

O nó de mais alto nível numa árvore de decisão é chamado de nó-raiz. (Han, Pei, e Kamber, 2011). O agrupamento (clustering) tem como objetivo principal descobrir dados que se agrupam naturalmente, classificando os dados em diferentes grupos e/ou categorias, e os registros em um grupo devem ser semelhantes uns aos outros e diferentes dos registros em outros grupos. Esses grupos e categorias não são conhecidos inicialmente.

Em MDE é possível, por exemplo, descobrir grupos de escolas (para investigar as diferenças e similaridades entre escolas), ou achar grupos de alunos (para investigar as diferenças e similaridades entre alunos). Uma vez que os grupos são formados, é possível fazer uma análise dos elementos que compõem cada um deles, identificando as características comuns aos seus elementos. (Elmasri e Navathe, 2011; Han, Pei, e Kamber, 2011).

3. Metodologia

Em relação às metodologias utilizadas em MDE é possível citar duas com maior destaque. A primeira é a conhecida como CRISP-DM (Cross Industry Standard Process for Data Mining), que propõe um modelo de processo para projetos de mineração de dados, apresentando seis fases de maneira cíclica, e são as seguintes: a) entendimento do negócio; b) entendimento dos dados; c) preparação dos dados; d) modelagem; e) avaliação e f) aplicação.

Tal metodologia é apresentada como um padrão desenvolvido por empresas de software como SPSS e NCR, além de indústrias como a Daimler-Benz. A segunda é a metodologia já apresentada no item 2.1, conhecida como KDD, que será utilizada na presente pesquisa, por ser um modelo voltado para pesquisas acadêmicas e serviu como base para um bom número de trabalhos publicados, como, por exemplo, Ramamohan, Vasantharao, Chakravarti, e Ratnam (2012), Romero e Ventura (2013), Asif, Merceron, e Pathan, (2014), Jeevalatha, Ananthi, e Kumar (2014) e Selvan, Beleya, Muniandy, Heng, e Remendran (2015) e Shaleena e Shaiju (2015), que aplicaram as etapas do processo de Descoberta de Conhecimento em Banco de Dados ou Knowledge Discovery in Databases (KDD) em suas pesquisas.

Seguindo o processo de KDD, para a implantação do objetivo principal da presente pesquisa, em uma primeira etapa, foi feita a importação dos dados do Lúmina e a seleção dos atributos que serão utilizados, posteriormente, na etapa de mineração de dados. Em seguida, tais dados foram submetidos à etapa de pré-processamento, e foram eliminados os dados com inconsistência ou redundância.

Na etapa seguinte, de transformação, foram criadas novas colunas calculadas, como resultado e número de atividades. As duas últimas etapas foram a da mineração propriamente dita, que consistiu na busca por padrões, por meio da aplicação de algoritmos para árvore de decisão (decision tree) e o agrupamento (clustering) com o uso da ferramenta Rapidminer. Por fim, foi efetuada a interpretação dos resultados da etapa anterior.

ISSN 2318-4051

4. Esportes e atividades ao ar livre

O curso foi criado e ofertado pela Plataforma Lúmina, esta é uma instalação do Moodle, com um tema customizado. O formato empregado nos cursos segue um modelo padrão: o conteúdo é transmitido prioritariamente na forma de vídeos, mas também são usados textos, imagens e outros materiais que possam ser inseridos no Moodle. Na plataforma todos os MOOCs têm um vídeo de apresentação que fica disponível, mesmo sem o cadastramento do participante; os cursos possuem blocos com informações sobre o curso e sobre os professores, e as avaliações se dão na forma de testes de múltipla escolha (com a atividade "questionário", do Moodle).

É necessário ao menos um teste de múltipla escolha para que o certificado de participação, emitido pela plataforma e impresso pelo participante, seja liberado. Os cursos são auto formativos e não existe interação com professores ou tutores. Desta forma, qualquer ferramenta disponível no Moodle, que não exija obrigatoriamente a presença de um professor ou tutor acompanhando o curso, pode ser utilizada.

O Curso foi ofertado devido a identificação que os esportes e atividades ao ar livre, além de despertar o interesse na população, possuem potencialidades ou benefícios que podem ser explorados e se referem a questões pessoais, ambientais, econômicas e socioculturais. O curso expressa que em paralelo a expressividade, interesse e potencialidades dos esportes e atividades ao ar livre, torna-se relevante considerar alguns aspectos éticos, como os relacionados ao meio ambiente e a presença de riscos nas práticas.

Os esportes e atividades ao ar livre podem contribuir para o desenvolvimento de competências relevantes à melhoria da qualidade do meio ambiente, ao mesmo tempo em que podem gerar impactos ambientais. Além disso, a exposição dos praticantes a riscos, que não deve ser considerada impeditiva ou negligenciada, gera a demanda por propostas estruturadas de gestão.

Pretende-se, com a realização deste curso, sensibilizar os participantes em relação à abrangência dos esportes e atividades ao ar livre; relevância da educação ambiental ao ar livre; importância da gestão de riscos para a prática de esportes e atividades ao ar livre e às possibilidades de práticas, inclusive de docência, centradas em esportes e atividades ao ar livre.

4.1. Características do curso

O curso teve duração de 40h, nas quais os alunos tem a liberdade para organizar sus tempos de estudos, este foi ofertado em 2018, de forma gratuita para estudantes que fizessem seu cadastro na Plataforma Lúmina. O curso foi composto por 04 unidades de aprendizagem, conforme apresentado no Quadro 01.

Quadro 01: Organização do Curso

Unidade de aprendizagem	Materiais disponíveis	Atividades para os alunos
Módulo 1 - Caracterização dos Esportes e Atividades ao Ar Livre	Vídeos online Referências Vídeo aula Gravada pelo Professor	Fórum Questionário
Módulo 2 - Educação Ambiental ao Ar Livre	Vídeos online Referências Vídeo aula Gravada pelo Profes-	Fórum Questionário

	sor	
Módulo 3 - Gestão de Riscos em Esportes e Atividades ao Ar Livre	Vídeos online Referências Vídeo aula Gravada pelo Professor	Fórum Questionário
Módulo 4 - Relato de Experiência com Esportes e Atividades ao Ar Livre	Referências	Fórum

O processo de avaliação foi composto de testes de múltipla escolha, e cada unidade correspondeu a 25% da nota final. Para poder emitir seu certificado, o aluno deveria obter nota final igual ou maior que 7,0 (sete). Além dos quatro módulos enunciados acima o curso ainda possui um quinto de finalização do curso o qual contém um questionário avaliativo sobre o curso e um link para gerar os certificados.

4.2. Análise do material do curso

Vídeos – Havia dois tipos de materiais em vídeo, alguns selecionados do Youtube, como formato de motivação para participação nos Fóruns, todos sobre a temática do Curso e também uma vídeo aula com conteúdo preparado pelo professor que ofertou o curso, e apresentada também por ele, com gravação e edição feita pela Equipe do NEAPED (produção multimídia para educação) da UFRGS. Em relação ao tempo de duração dos vídeos é possível encontrar, na literatura, pesquisas que apontam a média de tempo para reter melhor a atenção dos estudantes. Na visão de Khan (2012), o tempo ideal para melhorar o engajamento dos estudantes fica entre 10 a 15 minutos. A pesquisa de Khalil & Ebner (2017) foi direcionada para um MOOC denominado "Social Aspects of Information Technology" ofertado pela provedora iMooX na Áustria, que contou com 21 vídeos de duração média de 17 minutos. Os dados de pesquisa da empresa Kaltura (2016) com 1.500 respondentes (educadores, profissionais especializados em vídeo e alunos) apontam o intervalo de 5 a 10 minutos como o mais indicado para a duração de um vídeo. Contudo, é possível encontrar valor inferior como ideal para a duração de vídeo. Por exemplo, o trabalho de Guo, Kim, & Rubin (2014) analisou os dados de quatro MOOCs da provedora edX e chegou a conclusão que vídeos de até 06 minutos são muito mais envolventes para reter a atenção dos alunos. No caso do Curso Esportes e Atividades ao ar livre as vídeo aulas tinham uma média de 12 minutos, o que está bem coerente com as pesquisas na área.

Referências – São indicações de livros, artigos e textos básicos e complementares para elaboração dos questionários e fóruns que compõe as atividades dos alunos.

Fórum de discussão – o fórum foi um diferencial observado nesse curso, pois o professor ofertante teve participação ativa nas discussões o que não é uma característica comum aos MOOCs. Para cada unidade, foi lançado um tema para que alunos pudessem se manifestar e debater a respeito do assunto, usando um modelo de discussão entre os pares para a construção coletiva do conhecimento intermediado pelo professor. Os fóruns tinham como objetivo responder algumas questões propostas pelo professor como por exemplo, para o primeiro fórum foram sugeridas as seguintes questões: Que características dos esportes e atividades ao ar livre podem estar relacionadas ao crescente interesse da população nessas práticas? Quais os benefícios que podem ser gerados pela prática desses esportes e atividades ao ar livre? Quais aspectos devem ser levados em consideração para a promoção de iniciativas focadas em esportes e atividades

ISSN 2318-4051

des ao ar livre? Ao final dos 04 fóruns foram 2.145 publicações, com 753 no primeiro, 438 no segundo, 529 no terceiro e 425 no quarto.

Questionário – Os questionários compostos de cinco a sete questões de múltipla escolha elaborados com conteúdo presentes em especial nas vídeo aulas apresentadas pelo professor, os alunos tinham a possibilidade de realizar três tentativas para responder, das quais a maior nota é a validada pela plataforma, ao final com uma média acima de 7,0 o aluno poderia gerar seu certificado.

5. Processo de mineração de dados educacionais

A presente etapa contemplará o processo para obtenção do padrão de comportamento e desempenho dos alunos e será inspirado no modelo de KDD. Na etapa inicial, os dados foram extraídos do Lúmina gerando 3 planilhas em formato Excel para cada um dos questionários respondidos pelos alunos, assim como uma contagem foi realizada para verificar quais alunos responderam a avaliação do Curso. Na primeira geração das planilhas dos questionários respondidos obteve-se: Questionário 1 – 739 respostas, Questionário 2 – 504 respostas, Questionário 3 – 554 respostas, nessas planilhas estavam todos os alunos assim como havia todas as tentativas realizadas.

Essa base inicial foi submetida à fase de pré-processamento, em que foram eliminadas todas as tentativas duplicadas, permanecendo apenas a resposta com nota mais alta de cada aluno. Dessa forma foram obtidas planilhas com a seguinte quantidade de dados: Questionário 1 – 427 respostas, Questionário 2 – 361 respostas, Questionário 3 – 342 respostas, contagem de respostas na avaliação do curso 342 respostas.

A taxa de desistência foi de 51% geral desde os alunos que não efetuaram nenhuma atividade até aqueles que fizerem três, faltou apenas uma, para efeito de comparação, nos MOOCs a taxa de evasão ou desistência em média é de 90%. (Sandeen, 2013; Hew e Cheung, 2014; Alraimi, Zo, & Ciganek, 2015).

Depois dessa primeira análise, a etapa seguinte foi a de transformação, em que foram criadas novas colunas calculadas, como o número de atividades e o resultado (Concluinte ou Desistente), além da criação de uma coluna para medir a frequência de entrega das atividades. Para os alunos que enviaram apenas 01 atividade foi atribuída a classificação “ruim”, para aqueles enviaram 02 atividades atribuiu-se a classificação “regular”, para 03 atividades a classificação atribuída foi “bom” e, finalmente, para aqueles que fizeram todos os questionários e responderam a avaliação final a classificação foi “excelente”.

Cabe ainda salientar que 275 alunos não realizaram nenhuma atividade quase 40% dos inscritos, o que leva a questionamentos sobre qual o motivo da inscrição inicial e a descontinuidade ao ponto de não realizar nenhuma das tarefas propostas.

A presente classificação foi inspirada nos trabalhos de Clow (2013), Coffrin, Barba, Corrin e Kennedy (2014) e Wilkowski, Deutsch e Russell (2014), que criaram categorias para classificar os estudantes em função do modo como eles interagem com o curso e pelo desempenho nas atividades. O resultado dessa classificação foi o seguinte:

- ✓ 18 alunos classificados com o conceito “ruim” – 4,2%;
- ✓ 23 alunos classificados com o conceito “regular” – 5,3%;
- ✓ 50 alunos classificados com o conceito “bom” – 11,7%
- ✓ 342 alunos classificados com o conceito “excelente” – 80%

ISSN 2318-4051

Tal ação teve por objetivo melhorar a qualidade dos atributos e aumentar o nível de detalhamento do estudo. Em relação ao desempenho dos alunos, todos que fizeram as 3 atividades e responderam a avaliação final puderam gerar seus certificados. Esse alto índice aprovação está relacionado com o curso ter caráter informativo, direcionado para atender a um grande número de alunos de diferentes áreas do conhecimento e não ter o nível de exigência de uma disciplina integrante da matriz curricular de um curso da graduação.

Para a mineração foi utilizada a ferramenta RapidMiner em sua versão acadêmica que permite trabalhar com número ilimitado de registros. Inicialmente ocorreu a importação da planilha Excel gerada pelo Lúmina, com os seguintes atributos:

1. Instituição/Departamento;
2. Nome;
3. AS_I até AS_II onde AS significa Atividade de Sistematização, com a nota "média" cada atividade. Cada uma teve 33% na participação das notas;
4. Total (nota final) – 0 a 10,0;
5. Resultado – considerando 0 para desistente e 1 para concluinte.

Na etapa de transformação foram adicionadas as seguintes colunas:

1. Num_ativ – número de atividades entregues pelos alunos;
2. Freq_atividades – classificados em ruim, regular, bom ou excelente;
3. Condição – Concluinte ou Desistente

No processo de MDE, a primeira etapa foi realizada com a importação da planilha em formato XLSX com 427 linhas pela ferramenta RapidMiner. A partir desse momento, a ferramenta faz um processo de verificação com o objetivo de detectar algum tipo de erro nos dados. Em seguida, foram utilizados os recursos para mineração de dados da ferramenta RapidMiner, com os algoritmos de árvore de decisão (decision tree) e agrupamento (clustering) Tais algoritmos foram selecionados, pois são aplicados com sucesso em contextos educacionais (Baker, 2010; Romero e Ventura, 2013). Os experimentos e as análises estão descritos a seguir.

5.1 Experimento A

– Árvore de Decisão (Decision Tree): A árvore de decisão é representativa em relação à técnica de classificação, sendo um método adequado quando o objetivo da mineração é a classificação de dados ou predição de saídas. Para esse primeiro experimento foi utilizado o operador Retrieve para importar os dados da planilha gerada ao final das etapas de pré-processamento e transformação, e na sequência utilizou-se o operador Set Role para definir o atributo que será utilizado como classe, no caso o atributo Condição (concluinte ou desistente).

Em seguida, foi utilizado o operador Select Attributes para determinar quais atributos seriam utilizados no processo, sendo desconsiderados atributos como, por exemplo, "Nome" e "Instituição/Departamento" que não serão utilizados no processo de classificação da árvore de decisão. Por fim, foi inserido no processo o operador Decision Tree, com a função de gerar a árvore de decisão. O algoritmo analisa os diversos campos de forma interativa, buscando identificar aquele com maior influência no resultado das classes (concluinte ou desistente), nesse caso indicado pelo atributo Condição. O atributo de maior influência é colocado no topo da arvore (raiz) e, então, o algoritmo continua buscando novos campos significativos. Nesse caso, o atributo mais significativo foi Freq_Ativ. A Figura 03 representa a árvore de decisão gerada.

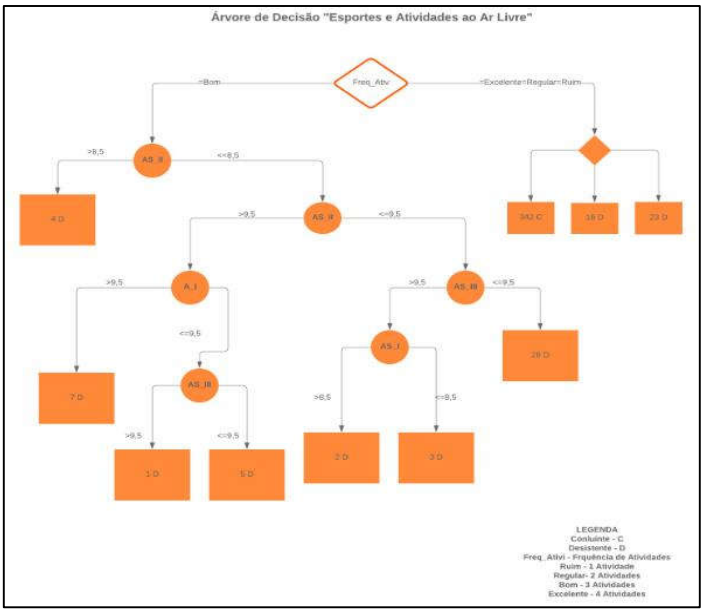
ISSN 2318-4051

O atributo mais significativo para o sucesso dos alunos foi Freq_Ativ (ruim, regular, bom e excelente). No gráfico pode-se visualizar isso por ser ele o nó raiz, colocado no topo da árvore e separando os alunos classificados como “Bom”, dos demais, ou seja, “Excelente”, “Regular” e “Ruim”. Para o lado direito da árvore, os alunos que fizeram todas as atividades, classificação “Excelente”, são 342. Como se tratou de um curso atípico, com objetivo de atender alunos das mais diversas áreas e também não apresentou nível de exigência significativo, o alto índice de aprovações influenciou na análise e geração da árvore.

O foco principal da análise a partir desse momento se concentra no conjunto de alunos não concluintes, mas considerados “bom”, 85 alunos. Ainda do lado direito da árvore é possível verificar que 18 alunos, realizaram apenas 01 atividade. Provavelmente são os alunos que fizeram apenas a primeira atividade e desistiram do curso, assim como 23 alunos que avançaram um pouco mais, fazendo 02 atividades, mas também não continuaram engajados no curso. Os que realizaram apenas 01 ou 02 atividades são 35 alunos e representam 44% dos desistentes. Esse grupo significativo de desistentes precisariam com mais urgência de ações por parte da gestão do curso para diminuir sua evasão.

Do lado esquerdo da árvore, com alunos classificados como “Bom”, também são desistentes, mesmo tendo enviado 3 atividades, cabe então a análise do quantitativo de notas desses alunos para verificação se há alguma influência em sua desistência, desta forma após o atributo de frequência de entregas (Freq_Ativ), o mais importante foi a atividade AS_III. Nesse caminho, há um grupo de 50 alunos que mesmo fazendo 3 atividades e AS_III não chegaram até o final do curso, decidiram não pegar sua certificação. Mas com as análises obtidas parece que as notas não afetam a desistência dos alunos, pois mesmo indo bem, estes acabaram abandonando o curso.

Figura 3: Árvore de Decisão



Fonte: Autor

Os outros caminhos não foram significativos em termos do número de alunos desistentes. A árvore de decisão poderia ter gerado resultados mais detalhados, caso tivesse sido agregado para a análise, outros atributos, principalmente de caráter temporal,

como a data do último acesso do aluno ao ambiente e a datas de entrega das atividades, que se pretende incluir em trabalhos futuros.

5.2 Experimento B

– Agrupamento (Clustering): Para o experimento com a técnica de clusterização foi utilizado o mesmo conjunto de dados do experimento de árvore de decisão. Na sequência, foi utilizado o operador Select Attributes para determinar os atributos utilizados no processo, e os selecionados foram apenas atributos numéricos, como AS_I, AS_II, AS_III, além de Num_Ativ e Resultado. No momento seguinte, foi utilizado o operador Clustering com o algoritmo K-means, com parâmetro de k = 4. Após a execução do algoritmo k-means, o sistema gerou 4 grupos ou clusters com a seguinte distribuição de alunos: Cluster 0 com 138 alunos; Cluster 1 com 24 alunos; Cluster 2 com 61 alunos; Cluster 3 com 204 alunos.

Em relação aos alunos concluintes, no cluster 0 e no cluster 3 estão agrupados todos esses alunos, com 138 alunos e 203 alunos, respectivamente. O cluster 3 é o que reúne maior número de alunos aprovados e com melhor desempenho geral, pois todos fizeram as 04 atividades. Os alunos do cluster 0 também fizeram todas as atividades, mas tiveram desempenho inferior aos alunos do cluster 3 em todas elas. Os alunos do cluster 3 mantiveram um aproveitamento em relação à nota máxima de cada atividade de 98,44% em média. Já os alunos do cluster 0 tiveram aproveitamento de 84,68%. Os alunos do cluster 3 mantiveram, desse modo, um padrão de comportamento mais homogêneo nos resultados das 04 atividades. Na AS_II, o aproveitamento desse grupo foi de 97,48% e 98,52% na atividade AS_III. Para os alunos do cluster 0, o aproveitamento caiu de 87,12% da AS_II para 77,20% na AS_III. Portanto, os alunos do cluster 3 permaneceram mais engajados até o final.

O cluster 2 com 61 alunos apresenta como característica principal, agrupar alunos desistentes, no caso desse agrupamento há alunos com notas variando de 4,375 até 7,50 e número de 2 ou 3 atividades entregues no decorrer do curso. No cluster 1 somente também com apenas alunos desistentes, há um conjunto de 24 alunos os quais obtiveram nota final “média” de 2,30, valor muito inferior ao mínimo exigido.

Os alunos do cluster 1 tiveram maior aproveitamento na AS_I com 6,69. Nas atividades seguintes, os alunos continuaram a apresentar um comportamento similar em termos de desempenho, mas bastante inferior aos outros grupos, com queda contínua nas notas das atividades AS_II, AS_III, com média igual a 2,4; 2,1, respectivamente. Sendo assim, após a entrega e resultado ruim na primeira atividade, os alunos foram perdendo o interesse e abandonando o curso. Os clusters 1 e 2 despertaram atenção para entender melhor o padrão de comportamento desse grupo de alunos.

Para refinar um pouco mais a análise foi gerado um novo agrupamento somente com os 85 alunos reprovados. Nesse caso foram gerados dois clusters e os alunos foram distribuídos da seguinte maneira. Cluster 0 com 35 alunos e Cluster 1 com 50 alunos

Por esse agrupamento, o cluster 0 representa 41% dos reprovados e o cluster 1 representa a maioria da base total de alunos, com 59%. Uma análise possível aponta que para permanecer na média de aprovação, considerando-se a nota final maior ou igual a 7,0 cada aluno precisa atingir essa média nas atividades. No cluster 0, que representa o menor grupo de desistentes, os alunos superaram esse valor nas atividades AS_I e AS_II, com notas médias de 8,5 e 7,7 respectivamente, das atividades que realizaram.

ISSN 2318-4051

A partir da terceira atividade, esse grupo começou a apresentar desempenho bastante inferior a nota mínima, com 2,7 na AS_III. Outro aspecto relevante é que mesmo entregando em média 03 atividades os alunos desse grupo desistiram. No cluster 1 que representa a maioria dos desistentes, os alunos tiveram nota média na AS_I de 6,6. A partir da atividade AS_II, os alunos desse grupo praticamente não tiveram aproveitamento nas atividades, representando um abandono do curso, e quase todos entregaram apenas a primeira atividade.

6. Considerações finais

A pesquisa teve como objetivo analisar as contribuições e restrições da aplicação de métodos de mineração de dados educacionais em um conjunto de dados de um MOOC. Para atingir tal objetivo, foi analisado o Curso Esportes e Atividades ao ar Livre, da plataforma Lúmina da UFRGS, este obteve 702 matrículas.

Foram considerados na etapa de mineração de dados algoritmos, bastante utilizados em contextos educacionais, árvore de decisão e agrupamento. Após os experimentos, os resultados trouxeram uma clareza maior a respeito do assunto, pois foram descobertos conhecimentos novos e que podem ser úteis para os professores e gestores do curso.

Foram considerados para análise, 427 alunos que realizaram pelo menos uma atividade durante o curso. As duas primeiras fases do processo de KDD, pré-processamento e transformação, foram muito trabalhosas, pois mesmo com os recursos de filtros e fórmulas nativos da planilha Excel, as tarefas, envolvendo uma base dados que não tem uma boa qualidade como a gerada pela plataforma é uma tarefa bastante complexa, levaram aproximadamente 60% do tempo total do processo de KDD.

No experimento com a árvore de decisão foi possível verificar alguns padrões de comportamento dos alunos. Por meio desse algoritmo foram destacados 02 grupos de alunos reprovados que necessitam de maior nível de atenção. Provavelmente são os 18 alunos que que fizeram apenas uma atividade e desistiram do curso, assim como outro conjunto de 23 alunos, que fizeram apenas duas atividades e interromperam o curso. Tais grupos demonstraram baixo nível de engajamento e seria oportuno para as próximas edições, o desenvolvimento de um modelo de predição, que baseado nessas regras, pudesse prever o comportamento dos novos alunos. Aqueles com comportamento semelhante aos indicados anteriormente, deveriam receber atenção maior por parte dos professores e gestores do curso, por exemplo, recebendo mensagens específicas e atividades adicionais.

O algoritmo de agrupamento trouxe contribuições mais significativas em relação ao de árvore de decisão. Em um primeiro momento, toda a base de dados foi utilizada, sendo empregado o algoritmo k-means com 04 clusters. Dos grupos gerados, foi possível verificar que foram 02 clusters de concluintes e 02 de desistentes, mas com rendimentos diferenciados para cada um desses agrupamentos. Em relação aos clusters de concluintes, embora todos tenham entregado as 04 atividades, os alunos do cluster 3 mantiveram um padrão de comportamento mais homogêneo e engajado, com ótimo aproveitamento até a última atividade.

Contudo, é o grupo de desistentes que merece mais atenção. No cluster 01, que reuniu os alunos com menos rendimento, os alunos tiveram aproveitamento aceitável somente na AS_I. A partir dela, os alunos foram diminuindo o aproveitamento e aban-

ISSN 2318-4051

donando o curso. A partir da constatação que o cluster 2 também apresentava alunos com baixo rendimento, foi feito um novo agrupamento (k=2) com uma nova base somente de desistentes (85) para entender melhor esse grupo.

Nessa nova análise, o cluster 0 (35 alunos) tem alunos que tiveram nota superior a 7,0 apenas nas AS_I e AS_II e no cluster 1 (50 alunos), o desempenho foi pior, pois a maioria teve aproveitamento satisfatório somente na AS_I, e a partir dela os alunos praticamente não tiveram aproveitamento nas tarefas. Esse padrão de comportamento é semelhante ao da árvore de decisão. Nesse caso, conhecer o comportamento de cada grupo pode apoiar o gestor ou professor das próximas turmas. Seria importante analisar semanalmente o comportamento dos alunos a fim de verificar se o comportamento da turma anterior se repete. Por exemplo, analisar aqueles que não entregaram atividades 1 e 2 até determinada data. Tal comportamento poderia indicar um aluno com alto potencial de evasão ou reprovação.

Esse conhecimento gerado após a utilização de algoritmos de MDE pode ser útil em cursos a distância e, especialmente, em MOOCs, principalmente para compreender o ponto de vista dos alunos. Em um curso a distância, a tutoria tem papel preponderante no contato com os alunos, orientações, solução de dúvidas, dentre outros. No caso dos MOOCs, essa questão da tutoria torna-se inviável para a gestão do curso, em razão do número de tutores necessários para atender, uma quantidade tão elevada de alunos.

Desse modo, os recursos de tutoria deveriam ser investidos quando são mais necessários. Conhecendo o comportamento de determinados grupos, os professores e gestores podem enviar mensagens ou propor atividades específicas para esse grupo de alunos, por exemplo, com risco de abandonar o curso. A oferta de um curso massivo representa um considerável desafio em termos de gestão, pois uma grande quantidade de alunos gera além da grande quantidade de dados, aspectos envolvendo a parte operacional do curso, como responder as centenas de mensagens dos alunos sobre diversos assuntos e verificar os temas mais citados nos fóruns de discussão.

O desafio tecnológico também está presente, pois é preciso as limitações da plataforma Lúmina, que proporciona uma integração entre os participantes, mas até certo nível, não configurando um sistema totalmente interativo. Uma análise superficial dos fóruns de discussão mostrou que a participação do professor nessa atividade ocasionou uma motivação no participantes, um estimulo a mais para a permanência dos alunos até o final.

Uma contribuição importante desse trabalho é mostrar a possibilidade da criação de um sistema de alertas para professores e gestores que, a partir das regras geradas pelos algoritmos de MDE, como árvore de decisão, identifique alunos com risco de evasão e possibilite ao professor ou gestor atuar de maneira antecipada, enviando mensagens de acordo com os alertas recebidos pelo sistema.

Em termos de trabalhos futuros para análises quantitativas sugerem-se novos estudos a respeito da aplicação de outros algoritmos em contextos educacionais, como redes neurais, regressão linear e regras de classificação. Ainda em termos de trabalhos futuros, mas pensando em análises mais qualitativas cabe citar novamente que o curso apresentou 342 alunos concluintes (49%), esse alto índice de concluintes não é comum em cursos MOOC.

Por isso, coube uma primeira análise de forma mais quantitativa de quais poderiam ser os fatores que levaram tantos alunos a conclusão, contudo apesar de se ter identificado alguns fatores relevantes para a conclusão do curso, como a frequência de ativi-

dades realizadas, e também, por meio dos algoritmos aplicados, poder identificar alguns fatores que determinam o abandono do curso, como baixo desempenho, citado acima, não foi possível definir com clareza quais atributos tornam esse curso em especifico bem sucedido.

Pressupõe-se que a participação do professor que ofertou o curso nos Fóruns de discussão possa ser um fator preponderante para os bons resultados, contudo uma verificação de forma qualitativa nas postagens dos alunos, assim como uma investigação mais aprofundada na avaliação final do curso, pelos alunos, poderia elucidar melhor quais os motivos para as altas taxas de conclusão.

Referências

ALLEN, I., & SEAMAN, J. (2015). Online Learning Consortium. Acesso em 10 de 03 de 2016, disponível em Online Report Card – Tracking Online Education in the United States, 2015: http://onlinelearningconsortium.org/read/online-report-card-tracking-onlineeducation-united-states-2015.

ALRAIMI, K., ZO, H., & CIGANEK, A. (2015). Understanding the MOOCs continuance: The role of openness and. Computers & Education, pp. 28-38.

ASIF, R., MERCERON, A., & PATHAN, M. (2014). Predicting student academic performance at degree level: a case study. International Journal of Intelligent Systems and Applications, 7(1), 49-61.

BAKER, R. (2010). Data mining for education. International encyclopedia of education, 7, 112- 118.

BAKER, S. (2014). Educational data mining: An advance for intelligent systems in education. IEEE Intelligent systems, 29(3), pp. 78-82.

BALA, M., & OJHA, D. (2012). Study of applications of data mining techniques in education. International Journal of Research in Science and Technology, 1(4), 1-10.

CALDERS, T., & PECHENIZKIY, M. (2012). Introduction to The Special Section onEducational Data Mining. ACM SIGKDD Explorations Newsletter, 13(2), 3-6.

CAMPAGNI, R., MERLINI, D., SPRUGNOLI, R., & VERRI, M. (2015). Data mining models for student careers. Expert Systems with Applications, 42(13), 5508-5521.

CHATTI, M., DYCKHOFF, A., SCHROEDER, U., & THÜS, H. (2012). A reference model for learning analytics. International Journal of Technology Enhanced Learning, 4(5-6), pp. 318- 331.

CLOW, D. (2013). MOOCs and the Funnel of Participation. Proceedings LAK '13, (pp. 186-189). Leuven, Bélgica.

COFFRIN, C., BARBA, P., CORRIN, L., & KENNEDY, G. (2014). Visuzalizing patterns of student engagement and performance in MOOCs. Proceedings - LAK2014 - Learning Analytics and Knowledge. Indianapolis, USA.

COOPER, S., & SAHAMI, M. (2013). Reflections on Stanford's MOOCs. New possibilities in online education create new challenges. Communications of the acm, 56(2), 28-30.

CROSSLEY, S., PAQUETTE, L., DASCALU, M., MCNAMARA, D., & BAKER, R. (2016). Combining ClickStream Data with NLP Tools to Better. Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. ACM (pp. 6-14). Edinburgh, U.K.: ACM - Association for Computing Machinery.

DUTT, A., AGHABOZRGI, S., ISMAIL, M., & MAHROEIAN, H. (2015). Clustering Algorithms Applied in Educational Datamining. International Journal of Information and Electronics Engineering, 5(2), 112-116.

ELMASRI, R., & NAVATHE, S. (2011). Sistemas de Banco de Dados (6a. ed.). São Paulo: Pearson Addison Wesley.

FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), pp. 27-34.

GUO, P., KIM, J., & RUBIN, R. (2014). How video production affects student engagement: An empirical study of mooc videos. Proceedings of the first ACM conference on Learning@ scale conference (pp. 41-50). Atlanta, Georgia, USA: ACM - Association for Computing Machiinery.

HAN, J., PEI, J., & KAMBER, M. (2011). Data mining: concepts and techniques (3. ed.). Waltham, MA: Elsevier. Hew, K., & Cheung, W. (2014). Students and Instructors use of massive open online courses (MOOCs): motivations and challenges. Educacional Research Review, pp. 45-58.

HU, Y., LO, C., & SHIH, S. (2014). Developing early warning systems to predict students' online learning. Computers in Human Behavior, 36, pp. 469-478. Hyman, P. (2012). In the Year of Disruptive Education. Communications of the acm, 55(12), 20-22.

JEEVALATHA, T., ANANTHI, N., & KUMAR, D. (2014). Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms. International Journal of Computer Applications, 108(15), 27-31.

JORDAN, K. (2015). Massive Open Online Course Completion Rates Revisited: Assessment, Length and Attrition. The International Review of Research in Open and Distributed Learning, 16(3).

KALTURA. (2016). The State of Video in Education 2016: A Kaltura Report. Acesso em 20 de julho de 2019, disponível em Kaltura: https://corp.kaltura.com.

KHALIL, M., & EBNER, M. (2017). Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories. Journal of Computing in Higher Education, 29(1), 1-19. Khan, S. (2012). The one world schoolhouse: Education reimagined. New Yourk: Twelve.

MUÑOZ-MERINO, P., RUIPÉREZ-VALIENTE, J., ALARIO-HOYOS, C., PEREZ-SANAGUSTIN, M., & KLOOS, C. (2014). Precise Effectiveness Strategy for Analyzing the Effectiveness of Students. Computer in Human Behavior, pp. 1-11.

NANFITO, M. (2014). MOOCs: Opportunities, impacts, and challenges: massive open online courses in colleges and universities. Createspace - Amazon. Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related. Expert Systems with Applications, 41(14), 6400-6407.

PARDO, A., & KLOOS, C. (2011). Stepping out of the box: towards analytics outside the learning management system. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 163-167). Banff, Canada: ACM.

RAMAMOHAN, Y., VASANTHARAO, K., CHAKRAVARTI, C., & RATNAM, A. (2012). A study of data mining tools in knowledge discovery process. International Journal of Soft Computing and Engineering (IJSCE), 2(3), 2231-2307. 130

RIGO, S., CAMBRUZZI, W., BARBOSA, J., & CAZELLA, S. (2014). Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. Revista Brasileira de Informática na Educação, 22(1), 132- 146.

ROMERO, C., & VENTURA, S. (2010). Educational Data Mining: A Review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, 40(6), pp. 601-618.

ROMERO, C., & VENTURA, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

ROMERO, C., ZAFRA, A., LUNA, J., & VENTURA, S. (2013). Association rule mining using genetic programming using genetic programming to provide feedback to instructors from multiple-choice quiz data. Expert Systems, 30(2), 162-172. Sandeen, C. (2013). Integrating MOOCs into Traditional Higher Education: The emerging "MOOC 3.0" Era. The Magazine of Higher Learning, pp. 34-39.

SELVAN, A., BELEYA, P., MUNIANDY, M., HENG, L., & REMENDRAN, C. (2015). Minimizing Student Attrition in Higher Learning Institutions in Malaysia Using Support Vector Machine. Journal of Theoretical and Applied Information Technology, 71(3), 377-385.

SHAHIRI, A., HUSAIN, W., & RASHID, N. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, pp. 414-422.

SHALEENA, K., & SHAIJU, P. (2015). Data Mining Techniques for Predicting Student Performance. Engineering and Technology (ICETECH) (pp. 1-3). Coimbatore, TN, India: IEEE. Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. Educase Review, 46(5), pp. 30-40.

WILKOWSKI, J., DEUTSCH, A., & RUSSELL, D. (2014). Student Skill and Goal Achievement in the Mapping with Google MOOC. L@S 2014 - Student Skills and Behavior (pp. 3-10). Atlanta, Georgia, USA.: ACM.

YADAV, S., BHARADWAJ, B., & PAL, S. (2012). Data Mining Applications: A comparative for predicting student's performance. International Journal of Innovative Technology & Creative Engineering, 1(12), pp. 13-19. You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. The Internet and Higher Education, 29, pp. 23-30.