

CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais

Jorge Luis C. Ramos¹, João C. Sedraz Silva¹, Rodrigo L. Rodrigues², Pamella Letícia Silva de Oliveira²

¹Universidade Federal do Vale do São Francisco (UNIVASF) - Juazeiro - BA – Brasil

²Universidade Federal Rural de Pernambuco (UFRPE) - Recife-PE, Brasil.

Abstract. *This theoretical article presents a proposal to adapt the consolidated CRISP-DM data mining model to educational data scenarios, whether they come from online teaching platforms or from sources such as national course exams or even academic management systems. Despite fully following the six steps of the original model, the proposed model, here called CRISP-EDM, presents some particularities due to the types of data and the educational domain. The model has already been used in two approved doctoral theses and is now presented in more detail in this work. This is expected to contribute to research in the field of educational data mining.*

Resumo. *Este artigo teórico apresenta uma proposta de adaptação do consolidado modelo de mineração de dados CRISP-DM, para cenários de dados educacionais, sejam eles provenientes de plataformas de ensino online ou de fontes como exames nacionais de curso ou mesmo de sistemas de gestão acadêmica. Apesar de seguir integralmente as seis etapas do modelo original, o modelo proposto, aqui denominado de CRISP-EDM, apresenta algumas particularidades em razão dos tipos de dados e do domínio educacional. O modelo já foi usado em duas teses de doutorado aprovadas e agora é apresentado com mais detalhes neste trabalho. Espera-se com isso, dar uma contribuição às pesquisas na área de mineração de dados educacionais.*

1. Introdução

Com a evolução contínua da *Data Mining* (DM), alguns esforços foram feitos, buscando estabelecer normas e padrões na área, tanto por iniciativas acadêmicas quanto pela indústria. A maior parte desses esforços foram centrados na definição de uma linguagem para DM que possa ser aceita como padrão da mesma forma que SQL foi aceita como um padrão para bancos de dados relacionais [Azevedo e Santos, 2008]. Nesse sentido, metodologias como o PMI¹, Agile², RUP³ e outras similares são adequadas para processos de desenvolvimento de software e não para projetos que envolvam extração e análise de dados. Os esforços da indústria foram principalmente na definição dos processos ou metodologias que podem guiar a implementação de aplicações de DM em diversas áreas.

Diante dessa necessidade, um consórcio liderado por várias empresas consumidoras e fornecedoras em potencial de serviços de *data mining*, com a participação de mais de 200 usuários da área, provedores de serviço e de ferramentas de mineração de dados, começou, em na segunda metade da década de 1990, a desenvolver o CRISP-DM (Acrônimo de *CRoss-Industry Standard Process for Data Mining*)⁴, um modelo de mineração de dados não proprietário, neutro, documentado e disponível livremente. É um

¹ <https://www.pmi.org>

² <http://agilemethodology.org/>

³ <http://www.wthree.com/rup/>

⁴ <http://www.crisp-dm.org/>

modelo específico de processo que descreve abordagens comumente usadas em mineração de dados para resolver problemas desse domínio [Shearer, 2000].

Segundo MARTÍNEZ-PLUMED *et al.* (2019), em muitas pesquisas feitas com empresas e usuários, o CRISP-DM ainda é o padrão para o desenvolvimento de projetos de mineração de dados e descoberta de conhecimento. O campo avançou bastante em vinte anos com a ciência de dados, sendo agora o método preferido na mineração de dados.

Com a consolidação do método para processos de DM, é uma consequência natural que os processos de mineração em contextos educacionais, também possam se valer deste percurso metodológico para uma melhor efetivação de projetos de EDM.

Este trabalho sugere e descreve o processo de adoção do CRISP-DM em cenários de mineração de dados educacionais, deixando ainda mais evidente a sua utilização nesses contextos e como uma proposta de adaptação na sua denominação para CRISP-EDM.

2. Mineração de Dados Educacionais

A Mineração de Dados Educacionais (EDM – do inglês *Educational Data Mining*) representa um processo de aplicação de técnicas de mineração em dados relacionados aos aspectos educacionais para abordar as questões de pesquisa no domínio da educação [García *et al.*, 2011]. Por um lado, o aumento tanto do software educativo instrumental, bem como de bases de dados de informação dos estudantes criaram grandes repositórios de dados que refletem como os alunos aprendem. Por outro lado, o *e-learning* tem gerado grandes quantidades de dados que, devidamente, explorados e classificados, podem fornecer importantes informações sobre os alunos e os cursos [Koedinger *et al.*, 2009].

A mineração de dados educacionais ajuda a identificar variáveis de aprendizagem e também é usada para examinar essas variáveis, a fim de entender o processo de aprendizado apoiado em tecnologias de comunicação e informação.

Assim, é possível compreender de forma mais eficaz e adequada, os alunos, como eles aprendem, o papel do contexto no qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem. Por exemplo, é possível identificar em que situação um tipo de abordagem instrucional (isto é, aprendizagem individual ou colaborativa) proporciona melhores benefícios educacionais ao aluno. Também é possível verificar se o aluno está desmotivado ou confuso e, assim, personalizar o ambiente e os métodos de ensino para oferecer melhores condições de aprendizagem [Baker *et al.*, 2011].

O processo de EDM converte os dados brutos de sistemas educacionais em conhecimento que pode ser usada por desenvolvedores de *software* educacional, professores, pesquisadores educacionais, entre outros. Esse processo não difere muito de outras áreas de aplicação de mineração de dados, porque ele se baseia nos mesmos passos do processo de mineração de dados em geral, conforme mostra a Figura 1 [García *et al.*, 2011].



Figura 1 - Etapas da Mineração de Dados Educacionais. Fonte: [García *et al.*, 2011].

Para que esse ciclo seja eficaz e produza o conhecimento relevante, a qual possa não somente ser utilizado pelos atores da *e-learning* mas também retroalimentar o sistema de ensino e ajudar a promover melhoras no processo, as etapas da EDM precisam acontecer de maneira efetiva e consistente, possibilitando que resultados intermediários e finais sejam analisados e validados, para uma adequada tomada de decisão, seja ela pedagógica ou gerencial.

3. CRISP-DM

De modo simplificado, o CRISP-DM é uma metodologia abrangente de mineração de dados e um modelo de processo que fornece para qualquer usuário de DM, seja ele iniciante ou especialista, um modelo completo para realização de um projeto de DM. O projeto é dividido em seis fases: entendimento do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação [Shearer, 2000]. O ciclo do CRISP-DM é exibido na Figura 2.

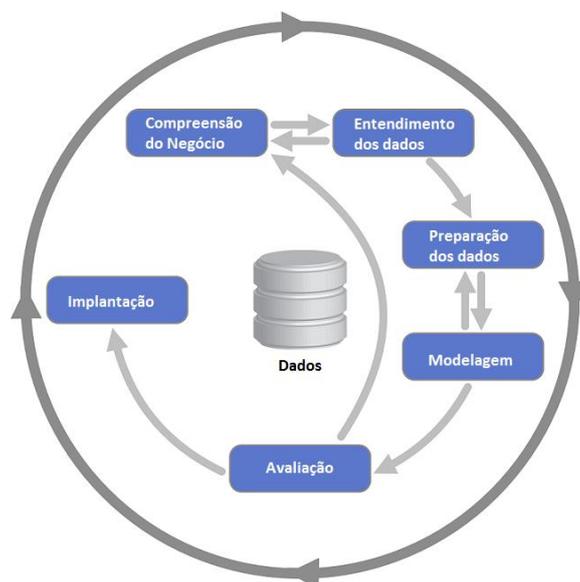


Figura 2 - Fases do CRISP-DM. Fonte: [Shearer, 2000]

A sequência de fases não é rigorosa. Na verdade, a maioria dos projetos avança e retorna entre as etapas se necessário. Como metodologia, inclui descrições das fases normais de um projeto, as tarefas requeridas em cada fase e uma explicação sobre as relações entre as tarefas. Como modelo de processo, o CRISP-DM fornece uma visão geral do ciclo de vida da mineração de dados [Chapman, 2000].

A primeira etapa consiste no conhecimento do domínio do negócio, ou seja, conhecer e compreender os objetivos do projeto de mineração a partir da perspectiva do negócio, a partir do qual esse conhecimento se transformará em um problema de mineração de dados. Essa etapa é considerada como uma das mais importantes do processo, a partir da qual é desenvolvido um plano preliminar do projeto de mineração para busca dos objetivos [Shearer, 2000].

A segunda etapa, na qual acontece a compreensão dos dados, começa com uma coleta de dados inicial, para estabelecer uma familiaridade com os dados, identificar possíveis problemas de qualidade dos dados, descobrir ideias iniciais ou para detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas. Esse passo

é essencial para evitar problemas inesperados durante a etapa seguinte (preparação de dados) que é normalmente a fase mais longa de um projeto [Shearer, 2000].

Na etapa seguinte, acontece a preparação dos dados que abrange todas as atividades necessárias para construir o conjunto de dados final, a partir dos dados brutos iniciais. Esses dados preparados alimentarão a ferramenta de modelagem na etapa seguinte. Tarefas de preparação de dados são susceptíveis de serem executadas várias vezes e não em qualquer ordem preestabelecida. Essas tarefas incluem a seleção de tabelas, registros e atributos, assim como a transformação e a limpeza dos dados [Chapman *et al.*, 2000].

A preparação de dados é um dos aspectos mais importantes e, muitas vezes, é o que exige mais tempo em mineração de dados, pois estima-se que essa etapa, geralmente, leva entre 50-70% do tempo e esforço de um projeto. Dedicar recursos adequados para os estágios iniciais de entendimento do negócio e esforços de dados pode minimizar a sobrecarga relacionada, mas, ainda, será necessário dedicar muito esforço para preparar e empacotar os dados para a mineração [IBM, 2016].

A modelagem acontece na quarta etapa. De acordo com o problema de mineração, várias técnicas podem ser usadas. A modelagem é geralmente executada em várias iterações, nas quais os analistas de dados executam vários modelos, usando as configurações padrão e vão ajustando os parâmetros para valores otimizados. É comum também retornar para a fase de preparação de dados para manipulações exigidas pelo modelo [Shearer, 2010; IBM, 2016].

Antes de proceder à implantação definitiva do modelo construído, é importante avaliá-lo mais profundamente e rever a sua construção para ter certeza que atinge, adequadamente, os objetivos planejados. A quinta etapa, avaliação, é a etapa chave para garantir que a organização pode utilizar os resultados obtidos e os conhecimentos descobertos. Um objetivo fundamental é determinar se há algum problema importante do negócio que não foi suficientemente considerado. A seta direcionada à etapa inicial indica a possibilidade de retorno em função da avaliação dos resultados não se mostra satisfatória para os objetivos do projeto [Chapman *et al.*, 2000; IBM, 2016].

A sexta e última etapa é a implantação, na qual os novos conhecimentos descobertos são usados para proporcionar melhorias na organização. Nessa etapa, todo o conhecimento adquirido deve ser organizado e apresentado de uma forma que o cliente possa usá-lo efetivamente dentro dos processos de tomada de decisão. Dependendo dos requisitos, a fase de implantação pode ser tão simples como gerar um relatório ou tão complexo como a implementação de um processo de mineração de dados aplicável em toda a empresa [Shearer, 2000; Chapman *et al.*, 2000].

3. Amostra de trabalhos que descrevem o uso do CRISP-DM em EDM.

Como a EDM pode ser considerada uma extensão da mineração de dados, é natural que trabalhos nessa área também se apropriem da metodologia CRISP-DM em seus estudos. Em razão da sua generalização, mesmo sendo um método criado pela indústria, a adoção da metodologia é sugerida como boa prática em EDM no trabalho de Sheth e Patel (2010). Conforme esses autores, o uso da metodologia pode ajudar a entender os requisitos do projeto e dados claramente, além de identificar os recursos necessários para a mineração com antecedência. Os trabalhos resumidos a seguir, apontam uma tendência no uso do CRISP-DM também em cenários de dados educacionais.

Sahay e Mehta (2010) apresentaram uma pesquisa para desenvolvimento de *software* via web para auxiliar o ensino superior em instituições americanas, na avaliação e previsão de questões-chave relacionadas ao sucesso do aluno. Pelo uso do KDD apoiado pelo CRISP-DM, a proposta do sistema incluiu o uso de vários algoritmos de mineração de dados e ferramenta de qualidade, tais como a implantação da função de qualidade para estudar e prever problemas relacionados à gestão de matrícula, taxa de abandono, tempo para conclusão de curso e também pode sugerir formas de melhorar os cursos e programas.

Kovacic (2010) apresentou, baseado na metodologia CRISP-DM, um estudo que explorou variáveis sócio-demográficas (idade, gênero, etnia, educação, status de trabalho e deficiência) e ambientes de estudo (período no qual o aluno está matriculado: 1º, 2º ou 3º e tipo de curso: bacharelado em ciências aplicadas ou bacharelado em negócios), que podem influenciar na permanência ou, então, na evasão de estudantes em uma universidade na Nova Zelândia. Foram examinados em que medida esses fatores podem ajudar na identificação prévia de alunos bem ou malsucedidos no curso.

Já o trabalho de Kabakchieva *et al.* (2011) apresentou um projeto de pesquisa de mineração de dados baseado no CRISP-DM em uma universidade búlgara, com o principal objetivo de revelar o grande potencial de aplicações de EDM para gestão universitária. O objetivo do trabalho foi descobrir padrões interessantes nos dados disponíveis, os quais poderiam contribuir para prever o desempenho do estudante na universidade com base em suas características pessoais e pré-universitárias.

O trabalho de Vialardi *et al.* (2011) apresentou a lógica por trás do projeto de um sistema de recomendação para apoiar o processo de matrícula, usando registro de desempenho acadêmico dos alunos. Para construir o sistema, a metodologia CRISP-DM foi aplicada a uma base de dados de alunos do Departamento de Ciência da Computação na Universidade de Lima no Peru.

A partir do uso do CRISP-DM, o estudo feito por Fernández e Mora (2016), realizou um processo alternativo de análise de dados de estudantes de engenharia de computação e sistemas de informação, em um ambiente de aprendizado que combinava educação on-line e presencial, com uma aplicação sequencial várias técnicas de mineração de dados (agrupamento, classificação e regras de associação) nos registros de um LMS, para reforçar a medição do desempenho acadêmico do aluno.

No artigo de Oreski *et al.* (2017), foi aplicada uma metodologia abrangente, baseada no processo CRISP-DM para mineração de dados, a fim de obter *insights* sobre a previsão do desempenho acadêmico. O artigo explorou em que medida as variáveis sócio-demográficas, a educação anterior, a motivação e o estilo de aprendizagem podem ajudar na identificação de alunos bem-sucedidos. Com base no método da árvore de decisão, o perfil do estudante bem-sucedido típico de tecnologia da informação foi identificado.

O estudo de Espitia *et al.* (2018), também focou na questão da evasão dos alunos, usando técnicas do KDD com o CRISP-DM para identificar padrões de comportamento dos alunos dos cursos de Engenharias Civil, Elétrica, Eletrônica, Computação e Sistemas, para reduzir possíveis desmotivações e desgaste dos alunos durante o seu curso.

O artigo de Silveira *et al.* (2019), teve como objetivo responder à questão sobre quais eram os principais fatores que motivam os estudantes de engenharia a abandonar os

curso de engenharia em uma universidade brasileira. A partir de dados das redes sociais e do desempenho de estudantes no período de 2009 a 2019, e ancorado na metodologia CRISP-DM, foi usado processo de classificação supervisionada, usando três algoritmos classificadores.

A pesquisa realizada por Ribeiro e Canedo (2020) identificou os fatores que afetam os alunos que abandonaram seu curso de graduação, usando aprendizagem de máquina, com a adoção das etapas do CRISP-DM para descobrir um modelo para prever quais alunos com risco de evasão e também quais as variáveis que mais contribuíam para esse abandono, ajudando a promover ações para reduzir a taxa de desistência dos cursos.

4. Adaptação do CRISP DM para o CRISP-EDM

Para a melhor adequação ao contexto da mineração de dados educacionais, o processo CRISP-DM foi adaptado para o contexto de dados educacionais e intitulado de CRISP-EDM. A Figura 3 mostra o ciclo de vida de um projeto de mineração de dados educacionais, o qual também consiste nas 6 (seis) fases de acordo com o processo CRISP-DM. A sequência de fases não é obrigatória, podendo ocorrer a transição para diferentes fases, dependendo do resultado de cada uma delas. As setas indicam as mais importantes e mais frequentes dependências entre as fases.

Cada etapa do CRISP-EDM contempla técnicas e abordagens devidamente adequadas ao domínio educacional sob análise. Algumas das etapas podem ser divididas em etapas menores para permitir um melhor desenvolvimento das mesmas.

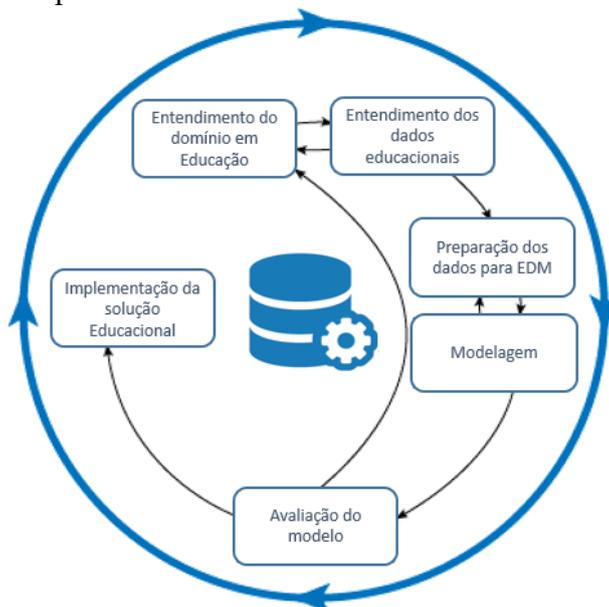


Figura 3 – Modelo do CRISP-EDM. Fonte: Autores, adaptado de Shearer, (2000)

O ciclo externo na figura simboliza o ciclo natural da mineração de dados. Um processo de mineração de dados continua após a solução ter sido desenvolvida, como por exemplo com a validação e atualização dos modelos desenvolvidos ou atualizações de modelos ou soluções implementadas.

A Etapa 01, no caso de um projeto educacional, o negócio ou domínio, é uma das fases mais importantes do processo. É nesta fase que é feito o levantamento das problemáticas educacionais da instituição para as quais se deseja buscar soluções por meio da mineração. Pode ser um ambiente de *e-learning* de uma universidade, dados do

exame nacional de estudantes de ensino médio, dados de sistemas de gestão acadêmica entre outros. Conhecer bem o domínio pode ajudar na definição dos objetivos da mineração e na escolha apropriada de variáveis que serão coletadas para o processo.

Ela pode também ser realizada durante a fase de justificativa e na análise da literatura, com o entendimento do domínio a ser explorado e do principal problema a ser tratado com o uso da EDM na pesquisa.

Em um projeto de EDM, a Etapa 02 pode acontecer, por exemplo, na análise inicial dos dados de um ambiente virtual de aprendizagem ou na compreensão dos dados produzidos por um censo escolar. Importante e recomendável também que, sempre que possível, a escolha dos dados seja a partir de alguma teoria educacional já consolidada, para respaldar tal escolha das variáveis. Como exemplo, variáveis comportamentais baseadas na Teoria da Distância Transacional [Moore, 1993] ou da Autorregulação da Aprendizagem [Zimmerman, Pons, 1986].

Em um contexto educacional, a preparação dos dados (Etapa 03) pode envolver, normalmente, a junção das bases de dados, extração, transformação e limpeza das variáveis para a construção da tabela de análise. Com exemplo, pode ser necessária uma limpeza dos dados nos *logs* de acessos dos alunos ao ambiente virtual e transformações de notas dos estudantes (variáveis contínuas) em conceitos (discretas). Importante verificar que, em bases oriundas de ambientes virtuais de aprendizagem, muitas vezes, há a ocorrência significativa de *outliers*, decorrentes de alunos muito interativos e *missing values*, de alunos com muito baixa ou nenhuma interação.

Problemas como diferentes bases ou versões de ambientes e gerenciadores de bancos de dados podem ocorrer e exigem cuidado no tratamento e unificação desses dados, para não comprometer a integridade dos mesmos.

Na EDM, a Etapa 04 (Modelagem) são definidas as técnicas de modelagem de dados, especificamente um conjunto de algoritmos de aprendizagem de máquina, bem como seus ajustes de parâmetros. Por exemplo, pode-se definir um modelo de previsão de tendência à evasão de alunos em educação a distância (EAD) ou, então, uma definição de agrupamento de alunos de acordo com suas características de interação em um ambiente virtual. Geralmente, opta-se inicialmente por parâmetros *default* nos algoritmos de mineração e, à medida em que o processo avança e os resultados preliminares forem obtidos, pode-se promover ajustes nesses parâmetros, no sentido de se buscar um refinamento nos resultados finais ou mesmo estabelecer comparativos entre os algoritmos disponíveis para a tarefa.

A Etapa 05, que prevê a avaliação dos modelos, deve-se verificar como os modelos desenvolvidos com as técnicas de mineração são executados nos dados reais. É importante validar os modelos de mineração, entendendo suas qualidades e características antes de implantá-los em um ambiente de produção. Um exemplo dessa etapa em um contexto educacional, seria avaliar se os modelos preditivos de desempenho de alunos atendem às taxas de acurácia, precisão e *recall* compatíveis com outros trabalhos ou com resultados esperados, ou, então, se as regras de associação geradas de acordo com o perfil dos alunos em *e-learning* possuem índices de suporte e confiança satisfatórios.

Na Etapa 06, a implantação do modelo pode ser feita, por exemplo, a partir do uso de *plug-ins* ou módulos de sistemas que realizam as tarefas de mineração na plataforma de dados educacionais e apresente, de maneira simplificada e geralmente por meio de

relatórios ou visualizações (*dashboards*), os resultados da aplicação dos modelos nos dados, auxiliando então na tomada de decisão que resulte em intervenções para melhoria do processo educacional. Diferente do modelo clássico do CRISP-DM, onde o conhecimento gerado pode ser destinado somente a especialistas no domínio estudado, ou mesmo somente para os analistas de dados, no processo de EDM é importante que o conhecimento obtido possa ser disponibilizado e validado por diversos atores do processo educacional, implicando em uma validação da implantação nessa última etapa, sendo essa a principal demanda para adaptação do CRISP para a EDM. Então, a implantação pode ser finalizada com a validação desses instrumentos.

Conforme preconiza o modelo CRISP-DM, em algumas etapas, pode haver um retorno a etapas anteriores para ajustes e verificações nos procedimentos adotados como, por exemplo, os modelos inicialmente obtidos não apresentavam resultados relevantes em suas métricas de avaliação ou alguma variável mostrou-se pouco significativa no modelo.

Como forma de validar a adaptação do CRISP-DM para o CRISP-EDM, tal proposta de adaptação foi apresentada e aprovada nas teses de doutorado de Rodrigues (2016) e Ramos (2016), sendo também adotada em trabalhos posteriores sem que, entretanto, tal método fosse detalhado como neste trabalho.

5. Considerações Finais

Modelos de mineração de dados, como o CRISP-DM, são muito utilizados como balizadores metodológicos para processos de preparação, modelagem e análise em grandes volumes de dados. Por isso, muitos estudos sobre mineração e descoberta de conhecimentos em bases de dados têm sido desenvolvidos utilizando essa metodologia.

A geração de dados em ambientes educacionais, sejam de ensino ou gerenciais tem contribuído de maneira significativa para obtenção de conhecimentos importantes nesses contextos. E agora, em um cenário de avanço exponencial do ensino e da gestão acadêmica remota, em razão das limitações e isolamento social impostos pela pandemia da COVID-19, essa quantidade de dados pode representar um importante meio de apoio e consolidação dessas atividades, desde que sejam devidamente analisados e deles sejam extraídas informações úteis, que possam auxiliar na tomada de decisões, sejam elas pedagógicas ou gerenciais.

As diversas áreas de utilização da ciência de dados na educação, tais como modelagem do estudante, detecção de comportamentos, previsão de desempenho, detecção precoce de evasão, *feedback* e apoio ao estudante e professores, dentre outros, estão sendo beneficiadas diretamente com os avanços nos estudos na área de EDM.

É evidente que o uso de técnicas de mineração de dados em contextos educacionais oferece oportunidades para educadores e pesquisadores obterem conhecimentos úteis e relacionamentos mais interessantes entre as variáveis em grandes conjuntos de dados. O modelo CRISP-EDM, pelas suas características, oferece possibilidades de integração de equipes multidisciplinares, além de oferecer um padrão já bem estabelecido na área.

Vale ressaltar que as técnicas de EDM devem vir embutidas em aplicações computacionais de fácil uso, com interfaces amigáveis e com ferramentas de visualização, de modo a proporcionar aos usuários sem os conhecimentos específicos, maneiras

transparentes de se beneficiarem das capacidades descritivas e preditivas da mineração de dados. Os resultados da mineração têm que ser transparentes e legíveis para qualquer ator do processo educacional. A última etapa do CRISP-EDM deve contemplar essa implementação e apresentação dos resultados.

O CRISP-EDM é uma proposta complementar ao processo original, direcionando esse consolidado percurso metodológico para ambientes de dados educacionais. A simples adição da letra “E” na sigla, pode representar não somente o “*educational*”, mas também termos relevantes como uma “*extension*” do método original ou a “*evaluation*” da solução implementada no contexto educacional.

Espera-se, portanto, que este trabalho possa contribuir com as pesquisas na área e que seja cada vez mais destacada a importância da ciência de dados educacionais.

Referências

AZEVEDO, A. I. R. L.; SANTOS, M. F. KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conference on Data Mining, 2008, IADIS. p.182–185.

BAKER, R. S. J. D.; ISOTANI, S.; CARVALHO, A. M. J. B. D. Mineração de dados educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, v. 19, n. 2, 2011.

CHAPMAN, P. *et al.* CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, v. 9, p. 13, 2000.

ESPITIA, E. *et al.* Applying CRISP-DM in a KDD process for the analysis of student attrition. In: Colombian Conference on Computing. Springer, Cham, 2018. p. 386-401.

FERNÁNDEZ, D. B.; MORA, S. L.. Uso de la metodología CRISP-DM para guiar el proceso de minería de datos en LMS. In: Tecnología, innovación e investigación en los procesos de enseñanza-aprendizaje. Octaedro, 2016. p. 2385-2393.

GARCÍA, E. *et al.* A collaborative educational association rule mining tool. The Internet and Higher Education, v. 14, n. 2, p. 77-88, 2011. ISSN 1096-7516.

KABAKCHIEVA, D.; STEFANOVA, K.; KISIMOV, V. Analyzing university data for determining student profiles and predicting performance. Educational Data Mining 2010, 2010.

KOEDINGER, K. R. *et al.* Educational software features that encourage and discourage “gaming the system”. Proceedings of the 14th International Conference on Artificial Intelligence in Education, 2009. p.475-482.

KOVACIC, Z. Early prediction of student success: Mining students' enrolment data. Informing Science & IT Education Conference (InSITE), 2010.

MARTÍNEZ-PLUMED, F. *et al.* CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 2019. doi: 10.1109/TKDE.2019.2962680.

MODELER, IBM SPSS. IBM SPSS modeler text analytics 16 user guide. 2016.

MOORE, M. G. Theory of transactional distance. In: (Ed.). Theoretical Principles of Distance Education. New York: Routledge, 1993. p.22-29.

ORESKI, D.; PIHIR, I.; KONECKI, M.. Crisp-DM process model in educational setting. Economic and Social Development: Book of Proceedings, p. 19-28, 2017.

RAMOS, J. L. C. Uma abordagem preditiva da evasão na educação a distância a partir dos construtos da distância transacional. Tese de Doutorado (Centro de Informática – UFPE), 2016. Disponível em <http://bit.ly/TeseJorge>. Acesso em 02 jul 20.

RIBEIRO, R. C.; CANEDO, E. D. Using Data Mining Techniques to Perform School Dropout Prediction: A Case Study. In: 17th International Conference on Information Technology–New Generations (ITNG 2020). Springer, Cham, 2020. p. 211-217.

RODRIGUES, R. L. Uma abordagem de mineração de dados educacionais para previsão de desempenho a partir de padrões comportamentais de autorregulação da aprendizagem. Tese de Doutorado (Centro de Informática – UFPE), 2016. Disponível em <http://bit.ly/TeseRodrigo>. Acesso em 10 jul 20.

SAHAY, A.; MEHTA, K. Assisting higher education in assessing, predicting, and managing issues related to student success: A web-based software using data mining and quality function deployment. Academic and Business Research Institute Conference, 2010.

SHEARER, C. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, v. 5, n. 4, p. 13-22, 2000.

SHETH, J.; PATEL, B. Best practices for adaptation of Data mining techniques in Education Sector. National Journal of System and Information Technology, v. 3, n. 2, p. 186, 2010. ISSN 0974-3308.

SILVEIRA, R. F. *et al.* Educational Data Mining: Analysis of Drop out of Engineering Majors at the UnB-Brazil. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE, 2019. p. 259-262.

VIALARDI, C. *et al.* A data mining approach to guide students through the enrollment process based on academic performance. User modeling and user-adapted interaction, v. 21, n. 1-2, p. 217-248, 2011. ISSN 0924-1868.

ZIMMERMAN, B. J.; PONS, M. M. Development of a structured interview for assessing student use of self-regulated learning strategies. American educational research journal, v. 23(4), p. 614-628, 1986.