



# METODOLOGIA CRISP-DM: POTENCIALIDADES NA DESCOBERTA DO CONHECIMENTO EM DADOS EDUCACIONAIS

*Methodology CRISP-DM: Potentialities in the Discovery of  
Knowledge in Education Data*

ANTONIO FRANCISCO LIMA DE OLIVEIRA PÁDUA

FABIANA ARAUJO SOUSA

## Resumo

CRISP-DM é uma metodologia de mineração de dados em formato cíclico, composto por 6 fases, que direcionam a descoberta do conhecimento para tomada de decisão sobre dados em grande volume. A versatilidade da utilização desta metodologia em diferentes contextos vem despertando o interesse de pesquisadores que estudam temáticas de cenários educacionais. Assim, o objetivo deste artigo é mostrar as potencialidades da aplicabilidade da metodologia CRISP-DM registradas em trabalhos científicos que utilizaram sua estrutura como bússola norteadora na extração do conhecimento em dados provenientes de fontes escolares possibilitando criar estratégias na resolução de problemas em âmbito educacional. Metodologicamente, utilizamos o levantamento bibliográfico em artigos e dissertações publicados em portais científicos digitais cuja temática envolvesse mineração de dados educacionais e a metodologia CRISP-DM. Esta investigação comprovou que a descoberta do conhecimento sobre dados educacionais utilizando técnicas de mineração de dados sob as fases da metodologia CRISP-DM é útil para desvelar o conhecimento escondido em dados escolares oportunizando solucionar problemas ou fenômenos presentes no cenário escolar.

**Palavras-chave:** mineração de dados educacionais, CRISP-DM, descoberta do conhecimento

## Abstract

*CRISP-DM is a cyclic data mining methodology, composed of 6 phases, that guide the discovery of knowledge for decision making on large-volume data. The versatility of the use of this methodology in different contexts has aroused the interest of researchers who study educational scenarios. Thus, the objective of this article is to show the applicability of the CRISP-DM methodology registered in scientific studies that used its structure as a guiding compass in the extraction of knowledge in data from school sources, making it possible to create strategies in problem solving in education. Methodologically, we used the literature review in articles and dissertations published in digital scientific portals whose subject matter involved data mining and the CRISP-DM methodology. This research proved that the discovery*

*of knowledge about educational data using data mining techniques under the CRISP-DM methodology is useful to uncover hidden knowledge in school data, making it opportunize to solve problems or phenomena present in the school scenario.*

**Key words:** *educational data mining, CRISP-DM, knowledge discovery*

## **Introdução**

A utilização de ferramentas tecnológicas permite a organização, análise e a extração de conhecimento sobre grande volume de dados. Assim, considerando uma das tecnologias mais propícias quando o contexto é buscar conhecimento em dados volumosos, podemos citar a Mineração de Dados (MD) ou *Data Mining* (DM), criada no final da década de 80, por profissionais de organizações, que dedicaram sua atenção a grandes volumes de dados armazenados, subutilizados ou ignorados pelos seus possuidores.

A Mineração de Dados (MD) objetiva descobrir o conhecimento por meio da realização de fases e tarefas dentro de um contexto que requer tomada de decisão diante de um problema. Dentre as diversas metodologias de mineração de dados, existe a CRISP-DM, objeto deste estudo, que visa mostrar a sua aplicabilidade na criação de estratégias para a resolução de problemas em âmbito educacional.

Outrossim, para alcançarmos o objetivo da nossa pesquisa realizamos um levantamento bibliográfico de artigos e dissertações de pesquisadores que utilizaram a metodologia CRISP-DM em dados educacionais, que serão expostos nos tópicos a seguir.

## **Referencial Teórico**

A Mineração de Dados (MD) é a etapa mais importante de um processo mais amplo conhecido como Descoberta de Conhecimento em Bases de Dados (DCBD) ou *Knowledge Discovery in Databases* (KDD).

De acordo com Fayyad, Piatetsky e Smyth (1996), a DCBD consiste em um processo não trivial da extração de conhecimento dos dados em um formato mais amplo, enquanto a MD representa apenas uma etapa específica do DCBD, na qual a identificação de padrões é executada com o auxílio de algoritmos específicos.

Para Camilo e Silva (2009), ainda não há uma unanimidade quanto à definição de DCBD e MD, pois alguns pesquisadores entendem como expressões semelhantes, no entanto todos concordam que o processo de mineração é interativo, por possuir fases que contêm tarefas e decisões a serem tomadas por um humano conhecedor do problema envolvido, e é iterativo, pois todas as fases são sistematicamente interligadas.

Nesse sentido, a Figura 1 representa o processo da DCBD, segundo Fayyad, Piatetsky e Smyth (1996).



Figura 1 - Descoberta de Conhecimento em Bases de Dados (DCBD)  
Fonte: Fayyad, Piatetsky e Smyth (1996).

As etapas da DCBD podem ser entendidas como:

- Seleção:** considerada como a primeira etapa da DCBD, nesta instância é criado um conjunto ou subconjunto de dados que será o foco da descoberta de novos conhecimentos. Ele deve conter as informações necessárias para que os algoritmos de mineração possam alcançar o objetivo do pesquisador.
- Pré-processamento:** momento em que os dados passam por uma limpeza ou eliminação de ruídos, e que inclui operações básicas para remoção de inconsistências.
- Transformação:** etapa da formatação necessária para agregar valor semântico às informações ou características úteis para representar os dados da base.
- Mineração de dados:** aplicação das técnicas de MD usando algoritmos para alcançar os objetos definidos na etapa da Seleção.
- Interpretação e Avaliação:** compreensão dos padrões obtidos incluindo a visualização dos modelos que resumem a estrutura e as informações presentes nos dados juntamente com as medidas técnicas que avaliam.

Nos últimos anos, dentre as etapas da DCBD destacamos a MD, por está suscitando o interesse de pesquisadores quanto a sua utilidade sobre dados educacionais, no qual denominam-a Mineração de Dados Educacionais (MDE).

De acordo com Baker, Isotani e Carvalho (2011), a MDE ou *Educational Data Mining* (EDM) é uma área recente de pesquisa cujo objetivo é desenvolver métodos para analisar dados volumosos provenientes de fontes relacionadas a um cenário escolar, úteis na descoberta do conhecimento e no direcionamento da tomada de decisão diante de um problema presente no âmbito educacional.

Historicamente, Marques (2014), afirma que a MDE está bem consolidada internacionalmente desde o primeiro *Workshop, Educational Data Mining*, durante o *20th National Conference on Artificial Intelligence*, em *Pittsburg-EUA*, no ano de 2005, e que no Brasil está conquistando seu espaço de forma gradativa.

Além disso, Silva *et al.* (2015) entende que a MDE é o processo da descoberta de conhecimento sobre dados brutos armazenados por sistemas escolares capazes de nortear desenvolvedores de softwares e pesquisadores que buscam soluções para tomada de decisão no ambiente educacional, no que remete a identificação precoce de comportamentos de aprovação, reprovação ou evasão escolar.

Logo, podemos dizer que a MDE é uma metodologia que possibilita a exploração do conhecimento sobre dados brutos oriundos do contexto educacional capaz de direcionar pesquisadores e gestores educacionais que buscam identificar dados e variáveis que evidenciam, sobretudo na identificação de discentes ou grupo de discentes tendentes ao sucesso ou ao insucesso escolar.

Dentre as várias metodologias de MD, destacamos a *Cross-Industry Standard Process of Data Mining* (CRISP-DM). Mas o que podemos entender sobre metodologia CRISP-DM?

A metodologia CRISP-DM ou *Cross-Industry Standard Process of Data Mining* (Processo Padrão Inter-Indústrias para Mineração de Dados) foi criada durante a década de 90 e sua origem se deve principalmente à necessidade da elaboração de modelos com foco na qualidade através da padronização de conceitos e técnicas, busca de informações e tomada de decisões. Ela recomenda um modelo de MD em formato compreensivo.

Sua estrutura propõe auxiliar os pesquisadores desde o planejamento até a execução da MD, passando pela especificação do processo da descoberta do conhecimento até a apresentação dos resultados alcançados.

Segundo Chapman (2000), a metodologia CRISP-DM é composta por 6 fases, organizadas de maneira cíclica, cujo fluxo é não unidirecional, possibilitando ir e voltar entre as suas fases e tarefas. As fases da metodologia CRISP-DM são:

- a) **Business Understanding** (Entendimento do Negócio)
- b) **Data Understanding** (Entendimento dos Dados)
- c) **Data Preparation** (Preparação dos Dados)
- d) **Modeling** (Modelagem)
- e) **Evaluation** (Avaliação)
- f) **Deployment** (Implementação do Modelo).

Com isso, a Figura 2 representa a sequência de cada uma das fases que compõem o objeto do nosso estudo.

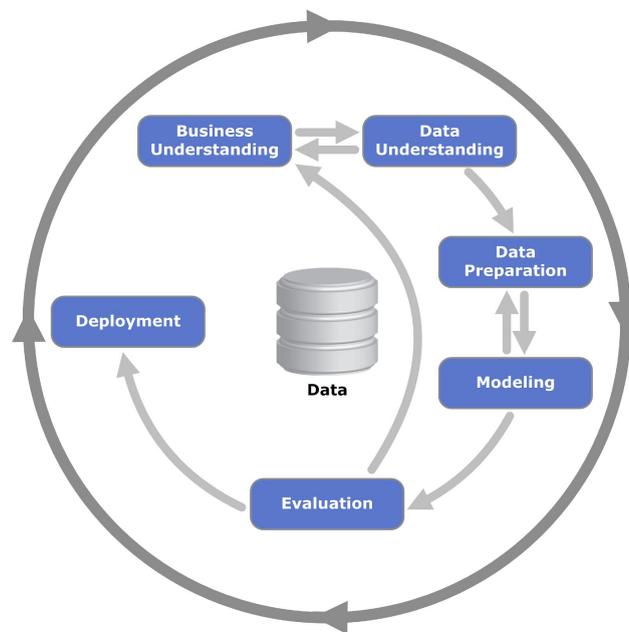


Figura 2 – Metodologia CRISP-DM

Fonte: <http://www.bigdatabusiness.com.br>

De acordo a Figura 2, percebemos que a metodologia CRISP-DM define um processo de MD em formato cíclico e que, implicitamente, cada fase é composta de tarefas que deverão ser realizadas para que seja determinada qual fase ou tarefa será a próxima a suceder.

Vale ressaltar que embora a referida metodologia tenha um formato padrão de fases do processo de MD, a mesma não impede a possibilidade da inserção de novas amostras,

variáveis ou atributos a serem minerados. Mas no que consiste cada uma dessas fases? Quais as contribuições metodologia CRISP-DM nas análises de dados educacionais?

De acordo a Figura 3, Chapman (2000) define as fases do nosso objeto de estudo como:



Figura 3 – Fases da metodologia CRISP-DM

Fonte: Chapman (2000).

Ao analisarmos a metodologia CRISP-DM sob a perspectiva da sua potencialidade na descoberta do conhecimento em contexto educacional, encontramos trabalhos cuja temática abordava mineração de dados educacionais utilizando a metodologia supracitada e percebemos que seu maior talento está na capacidade de se ajustar a diferentes situações da realidade escolar.

Em 2014, Santana, Maciel e Rodrigues, publicaram nos Anais do Simpósio Brasileiro de Informática na Educação, os resultados de um estudo cuja proposta foi aplicar técnicas de classificação sobre dados educacionais de curso semipresencial, para obter resultados que direcionassem a tomada de decisão por parte de professores e gestores. Os dados para esta pesquisa foram extraídos do perfil de uso do Ambiente Virtual de Aprendizagem (AVA), considerando o desempenho do aluno como variável alvo.

Os experimentos realizados foram inspirados na metodologia CRISP-DM e os resultados apontaram que as variáveis da dimensão perfil do aluno ao usar o AVA pode prever o desempenho do estudante.

Também em 2014, Adeodato, Santos Filho e Rodrigues demonstraram que além do conhecimento técnico, o ENEM coleta dados sócio-econômico e culturais dos alunos. No entanto, os autores afirmam que pouco conhecimento é extraído das bases que armazenam estes dados.

Com isso, utilizaram a metodologia CRISP-DM sobre os microdados do ENEM de 2011 para avaliar a qualidade do ensino médio brasileiro e caracterizar a “boa” escola como um dos objetivos da pesquisa.

Sob o mesmo entendimento da importância da utilização da metodologia CRISP-DM como modelo a ser utilizado para a aplicação de técnicas de mineração de dados do ambiente de gestão educacional, Pasta (2011) realizou um estudo de caso em uma Instituição de Ensino Superior de Blumenau-SC e concluiu que a aplicação de ferramentas de mineração de dados pode ser um recurso com potencialidades para a eficiência da gestão do conhecimento sobre dados escolares nas IES.

Ao enfatizar a importância do combate à evasão escolar, Veloso (2015) realizou um estudo sobre a predição da evasão em cursos técnicos de nível médio sob a perspectiva da mineração de dados por entender que o perfil de alunos já evadidos, podem contribuir na predição de alunos propensos a evasão.

Veloso (2015) extraiu o conhecimento sobre os dados coletados aplicando a metodologia CRISP-DM e concluiu que a mineração de dados educacionais pode auxiliar educadores em estudos sobre outras áreas.

Em outro trabalho, Barbosa *et al.* (2014) enfatizaram a aplicabilidade da descoberta de conhecimentos em banco de dados educacionais utilizando a metodologia CRISP-DM na identificação das causas da evasão na EAD de cursos técnicos, graduação e de pós-graduação.

A metodologia CRISP-DM também pode ser utilizada na mineração de textos.

Em 2016, Borges *et al.* mostraram que a descoberta de conhecimento em texto é o processo que combina técnicas de extração e de recuperação da informação, processamento da linguagem natural e organização de documentos com métodos de mineração de dados.

Para estes pesquisadores, lidar com dados não-estruturados, torna a descoberta de conhecimento em texto uma atividade considerada mais complexa que a descoberta de conhecimento em base de dados.

Através de artigos relacionados ao campo da biologia, aplicaram as técnicas de mineração de texto, combinadas com a metodologia CRISP-DM, com o objetivo de encontrar as palavras específicas desta área de conhecimento, buscando facilitando a tomada de decisão das palavras a serem escolhidas por discentes com deficiência motora durante o processo de escrita.

Com isso, podemos afirmar que além da metodologia CRISP-DM nos fornecer um panorama compreensivo do processo de MD, semelhante a outras estruturas já existentes, ela também pode nortear pesquisadores educacionais a traçar estratégias de como será o caminho percorrido para se extrair o conhecimento sobre dados escolares em grande volume em diferentes áreas do conhecimento.

Neste trabalho, escolhemos o CRISP-DM como metodologia por ser considerado atualmente o padrão de maior aceitação entre especialistas, de acordo com o *site* de DCBD chamado Kdnuggets.

O referido *site* é líder em *Big Data* e *Data Mining* e coordenado por Gregory Piatetsky-Shapiro, um dos principais pesquisadores na área. Ainda segundo o *site*, a metodologia CRISP-DM é a mais utilizada para alcançar objetivos na resolução de problemas envolvendo *data mining*, sendo a preferida por 43% dos profissionais que trabalham com MD.

## Metodologia

Para realizarmos uma pesquisa bibliográfica é preciso fazer uma análise minuciosa de todas as fontes documentais – livros, jornais, boletins, revistas, pesquisas, monografias, teses e outras em relação ao tema da pesquisa - que servem de suporte na investigação do objeto de estudo.

Segundo Prodanov e Freitas (2013, p. 80).

O levantamento bibliográfico é um apanhado geral sobre os principais documentos e trabalhos realizados a respeito do tema escolhido, abordados anteriormente por outros pesquisadores para a obtenção de dados para a pesquisa.

Portanto, o levantamento bibliográfico é um procedimento metodológico que fornece informações e contribuições para o pesquisador buscar possíveis soluções para seu problema de pesquisa. Sendo assim, não poderá ser aleatório.

De acordo com Fonseca (2002), a pesquisa bibliográfica é aquela realizada a partir da investigação de referências teóricas já apreciadas, e publicadas por meios impressos e

eletrônicos. Todo e qualquer trabalho científico é iniciado com uma pesquisa bibliográfica, que proporciona ao pesquisador conhecer o que já foi produzido sobre o tema em questão.

Existem pesquisas científicas que se baseiam unicamente na pesquisa bibliográfica, cujo objetivo é procurar e identificar referências teóricas publicadas com o objetivo de compilar informações ou conhecimentos preexistentes sobre o qual se procura a resposta.

Sendo assim, para esta investigação fizemos um apanhado geral dos principais trabalhos já realizados, atuais e relevantes sobre a utilização da metodologia CRISP-DM aplicados em dados educacionais.

Esse levantamento foi realizado em revistas científicas, artigos e dissertações de autores que estudaram ou estudam sobre a temática mineração de dados educacionais. Logo, possibilitou um amplo alcance de informações sobre a temática da pesquisa, auxiliando para uma melhor definição do quadro conceitual que envolve o objeto de estudo, bem como “coloca o pesquisador em contato direto com tudo que foi escrito sobre o tema” (MARCONI e LAKATOS, 2008).

### Considerações Finais

Diante deste trabalho podemos concluir que a MD é útil em áreas que possuem dados que nunca ou pouco foram explorados para o planejamento do futuro.

Além disso, entendemos que a MDE utilizando a metodologia CRISP-DM é digna de ser utilizada em diferentes contextos da educação, da mesma forma que a utilizamos para alcançar nosso objetivo geral.

Justificamos esta afirmação por percebermos que sua estrutura não obriga ao pesquisador executar todas as etapas, não impõe a técnica a ser adotada, nem tão pouco estabelece qual o algoritmo e software serão utilizados para que o conhecimento seja extraído. Ela apenas direciona e permite a tomada de decisão a cada tarefa realizada pelo pesquisador, no sentido de dar prosseguimento ou retornando a etapa anterior.

A descoberta do conhecimento sobre banco de dados enfatizando a MD, seguindo os moldes da metodologia CRISP-DM, pode motivar pesquisadores e instituições educacionais a construir e implementarem sistemas mais fáceis de serem utilizados por outros atores do cenário escolar, pois os estudos e resultados publicados ainda estão limitados aos profissionais da informática e da estatística.

Ao longo deste trabalho, buscamos investigar e mostrar que a metodologia CRISP-DM é de fato, aplicável em dados de contexto educacional.

A versatilidade da utilização da metodologia CRISP-DM sobre dados volumosos oriundos de diferentes fontes de dados e em especialmente de contextos educacionais citadas nas diferentes publicações analisadas, comprova que esta metodologia é uma ferramenta muito útil para descobrir o conhecimento no que tange a identificação de variáveis preditoras responsáveis por circunstâncias que podem explicar a evasão escolar e/ou sucesso escolar.

Nesse sentido, entendemos que pesquisas que visam analisar dados educacionais utilizando a metodologia CRISP-DM podem ampliar o horizonte no que diz respeito à possibilidade da identificação de variáveis nunca ou pouco discutidas na literatura.

### Referências

ADEODATO, P. J. L.; SANTOS FILHO, M. M.; RODRIGUES, R. L. Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar. **Anais do Simpósio Brasileiro de Informática na Educação**. v. 25. n. 1, 2014.

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

BARBOSA, W.; MÁXIMO, D.; JATOBÁ, A.; LEITE, A.; SOARES, E. . Uma Proposta para identificação de causas da evasão na educação a distância através de mineração de dados.

2014. Disponível em:<erbase2014.uefs.br/artigos/125801.pdf>. Acesso em: 18 mar. 2018.

BORGES, T; SILVA, K. A.; SANTOS, M. R. P.; SILVA, M.; GIACON, J. F. **A utilização do método Cross-Industry Standard Process for Data Mining no processo de mineração de textos**: extração de termos para criação de uma tecnologia assistiva para o auxílio à alunos com deficiência motora. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erbd/2016/001.pdf>. Acesso em: 3 de jul. de 2018.

CAMILO, C. O.; SILVA, J. C. da. **Mineração de Dados**: conceitos, tarefas, métodos e ferramentas. Instituto de Informática da Universidade Federal de Goiás, 2009.

CHAPMAN, P. **CRISP-DM 1.0**: Step-By-Step Data Mining Guide. [S.l.]: 2000. Disponível em: <<http://www.crisp-dm.org/download.htm>>. Acesso em: 28 jan. 2018.

COSTA, E.; BAKER, R. S. J. D.; AMORIM, L.; MARINHO, J. M. T. **Mineração de Dados Educacionais**: Conceitos, Técnicas, Ferramentas e Aplicações. Jornada de Atualização em Informática na Educação - JAIE 2012.

FAYYAD, U.; SHAPIRO, G. P.; SMYTH, P. **From Data Mining to Knowledge Discovery**: An Overview. Menlo Park, CA: AAAI Press/The MIT Press, 1996.

FONSECA, J. J. S. **Metodologia da pesquisa científica**. Fortaleza: UEC, 2002. Apostila

KDNUGGETS. Disponível em: <<http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>>. Acesso em: 3 de jul. de 2018.

MACHADO, R. D.; NARA, E. O. B.; SCHREIBER, J. N. C.; SCHWINGEL, G. A. Estudo bibliométrico em mineração de dados e evasão escolar. In: CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO, 11., 2015, Rio de Janeiro. **Anais...**, Rio de Janeiro: INOVARSE, 2015.

MARCONI, M. A; LAKATOS, E. M. **Fundamentos de metodologia científica**. São Paulo: Atlas, 2008

MARQUES, J. L. Q. **Mineração de Dados Educacionais**: um estudo de caso utilizando o Ambiente Virtual do SENAI. 2014. 72 f. Monografia (Licenciatura Plena em Computação) - Universidade Estadual da Paraíba, Campina Grande, 2014.

PASTA, A. **Aplicação da técnica de data mining na base de dados do ambiente de gestão educacional**: um estudo de caso de uma instituição de ensino superior de Blumenau-sc.

Disponível em:<<http://www.uniedu.sed.sc.gov.br/wp-content/uploads/2013/10/Arquelau-Pasta.pdf>>. Acesso em: 3 de jul. de 2018.

PRODANOV, Cleber Cristiano e FREITAS, Ernani Cesar de. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013.

SILVA, R. E. D.; RAMOS, J. L. C.; RODRIGUES, R. L.; GOMES, A. S.; FONSÊCA, J. A. V. Mineração de dados educacionais na análise das interações dos alunos em um Ambiente

Virtual de Aprendizagem. **Anais do Simpósio Brasileiro de Informática na Educação**. v. 26. n. 1., 2015.

VELOSO, L. A. **A predição da evasão escolar dos cursos técnicos de nível médio: um estudo de caso no SENAI**. 2015.94f. Dissertação (Mestrado em Gestão do Conhecimento e da Tecnologia da Informação) – Universidade Católica de Brasília, Brasília, 2015.