INTERFACE TECNOLÓGICA

DATA MINING: conceitos e consequências

DATA MINING: concepts and consequences

Luiz Amélio Sodaite Rossini – amelio.rossini@gmail.com
Renan Ricardo de Polli Silva – renan.polli0607@gmail.com
Eder Carlos Salazar Sotto – eder.sotto@fatec.sp.gov.br
Liriane Soares de Araújo – lirianearaujo@hotmail.com
Faculdade de Tecnologia de Catanduva (FATEC) – SP – Brasil
DOI: 10.31510/infa.v15i2.486

RESUMO

A Mineração de Dados (*Data Mining*) deve ser entendida como um conjunto de esforços empregados para a descoberta de padrões de acordo com bases de dados. Dessa maneira, há condições de gerar conhecimento útil para a tomada de decisões, através de algoritmos computacionais que recebem fatos do mundo real (entrada) e devolvem um padrão de comportamento (saída), expresso como modelagem de um perfil. Sendo assim, o objetivo deste artigo é definir a Mineração de Dados e os conceitos inerentes a ela, bem como elencar algumas ferramentas utilizadas para extração de conhecimento a partir dos dados. A metodologia do trabalho consiste em levantamento bibliográfico. Espera-se como resultado demonstrar a imprescindibilidade da Mineração de Dados no apoio à decisão nas organizações e contribuir para a produção científica e acadêmica.

Palavras-chave: *Data Mining*. Dado. Informação. Conhecimento.

ABSTRACT

Data Mining should be understood as a set of efforts for discovery of patterns according to databases. In this way, there are conditions to generate useful knowledge for decision-making, through computational algorithms that receive facts from the real-world (input) and they return a behavior pattern (output), expressed as modeling of a profile. Therefore, the objective of this article is to define Data Mining and the concepts inherent to it, as well as to list some tools used to extract knowledge from the data. The methodology of this article consists of a bibliographical survey. It is expected as a result to demonstrate the indispensability of the Data Mining in companies for decision support and to contribute to the scientific and academic production.

Keywords: Data Mining. Data. Information. Knowledge.

1 INTRODUÇÃO

No mundo interconectado do século XXI, em que novidades tecnológicas despontam quase que diariamente, as pessoas são alvo de todo tipo de informação, ainda que não a busquem. Há um grande entorpecimento de informações, muitas vezes supérfluas, mas que casualmente determinam uma nova maneira de pensar ou de agir, sobretudo quando se deixam levar por elas sem a realização de uma análise crítica.

As organizações não estão imunes à realidade de geração e interpretação de informações. Nesse contexto, a Mineração de Dados (*Data Mining*) surge como ferramenta indispensável para a aplicação de algoritmos na busca por conhecimento implícito e útil. Ela é considerada uma das etapas do processo de Descoberta de Conhecimento em Bases de Dados (*KDD*) ou até mesmo equivocadamente vista como sinônimo deste termo (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Além disso, a Mineração de Dados (*Data Mining*) pode ser compreendida como um processo automático ou semiautomático capaz de efetuar uma exploração analítica de grandes bases de dados, com o intuito de descobrir padrões essenciais na assimilação de informações importantes e oferecer suporte à geração de conhecimento (SILVA; PERES; BOSCARIOLI, 2016).

O objetivo deste artigo é definir a Mineração de Dados (*Data Mining*) e os conceitos relacionados a ela, assim como citar algumas ferramentas utilizadas para extração de conhecimento a partir dos dados. A metodologia do trabalho consiste em levantamento bibliográfico.

Estudar a Mineração de Dados (*Data Mining*) implica uma visão mais clara acerca do funcionamento da Inteligência Artificial e do Aprendizado de Máquina, que vêm em decorrência de uma mineração estruturada, além de permitir um maior discernimento na hora de tomar decisões, principalmente nas empresas de médio e grande porte. Esta pesquisa também contribui para a produção científica e acadêmica, por meio da qual estudantes adquirem conhecimento tácito e permanecem engajados no ambiente mercadológico.

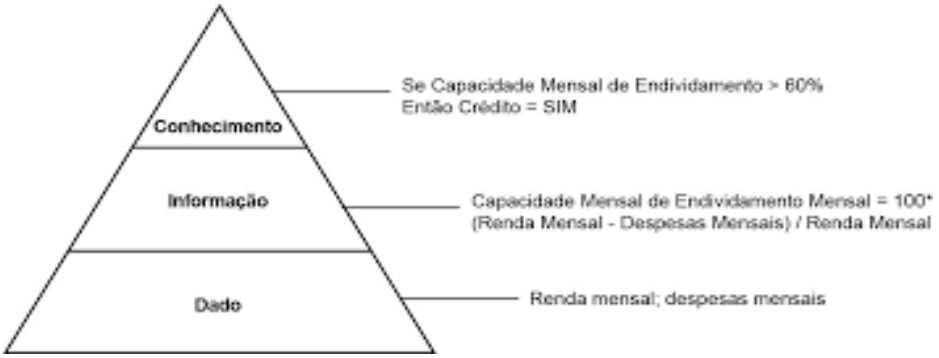
O artigo está estruturado com uma introdução, que compõe a seção 1; um referencial teórico sobre Mineração de Dados (*Data Mining*) e seus principais conceitos na seção 2; procedimentos metodológicos de pesquisa na seção 3; levantamento de resultados e discussão na seção 4; considerações finais na seção 5 e, por fim, as referências utilizadas.

2 MINERAÇÃO DE DADOS

Antes de tudo, é imprescindível a compreensão de três termos, uma vez que o assunto tratado é inerente a banco de dados e suas aplicações: dado, informação e conhecimento. Enquanto dado é apenas um valor coletado e armazenado, a informação, por sua vez, é o dado trabalhado e interpretado que gera certo significado. Já o conhecimento é a informação analisada criteriosamente e aplicada a uma finalidade específica (AMARAL, 2016b).

Para exemplificar as diferenças e a hierarquia entre dado, informação e conhecimento, evidencia-se na Figura 1 a base de dados de uma organização financeira hipotética que armazena as rendas e as despesas mensais de seus clientes.

Figura 1 - Hierarquia entre Dado, Informação e Conhecimento



Fonte: Goldschmidt, Passos e Bezerra, 2015, p. 2 (adaptado).

Na base da pirâmide, os dados podem ser interpretados como itens essenciais que são captados e armazenados por ferramentas da Tecnologia da Informação, com a finalidade de expressar fatos do mundo real de forma a serem aplicados no meio computacional. As informações representam os dados processados e, para isso, diversas ferramentas da Tecnologia da Informação também são utilizadas. O conhecimento corresponde a um padrão ou conjunto de padrões, através do qual dados e informações podem ser relacionados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

De acordo com Amaral (2016b), o dado pode estar em formato eletrônico analógico ou digital. Ele ainda pode existir em formato não eletrônico, normalmente impresso em papel e até mesmo nas pedras esculpidas pelo homem. A informação não eletrônica, impressa em papel, é de grande abundância no mundo – só na Biblioteca do Congresso Americano existem mais de 150 milhões de livros armazenados. O dado analógico é transmitido por ondas e pode

sofrer interferência, ao passo que o dado digital é transferido em pacotes de bits, mais eficientes e sofrendo menos interferência.

Segundo Porto e Ziviani (2014), o sucesso do uso dos Sistemas de Bancos de Dados Relacionais reduziu a representação dos dados a tabelas bidimensionais. Ao se tratar de grandes volumes de dados, a propriedade de sua representação se reflete positivamente no desempenho do acesso aos dados, o que, por sua vez, permite a criação de aplicações mais complexas do que aquelas apoiadas por bancos de dados relacionais. Assim, domínios que tem sido em grande parte negligenciados pelo suporte de banco de dados, tais como simulações numéricas, análises sísmicas e redes de interação gênica, que são apenas alguns dos muitos domínios nesta área, demandam representações de dados em sintonia com representações complexas, tais como: espaço, tempo, grafos e sequências.

Data Mining (ou Mineração de Dados) deve ser entendida como um conjunto de esforços empregados para a descoberta de padrões de acordo com bases de dados. A partir daí, há condições de gerar conhecimento útil para a tomada de decisões, através de algoritmos computacionais que recebem fatos do mundo real (entrada) e devolvem um padrão de comportamento (saída), expresso como modelagem de um perfil. *Data Mining* e o processo de descoberta de padrões de comportamento por meio de bases de dados estão relacionados com a Inteligência de Negócios (*Business Intelligence* ou BI, em inglês), que é um conjunto de aplicações, infraestrutura, ferramentas e melhores práticas que permitem o acesso e a análise da informação (SILVA; PERES; BOSCARIOLI, 2016).

2.1 Ciência de Dados

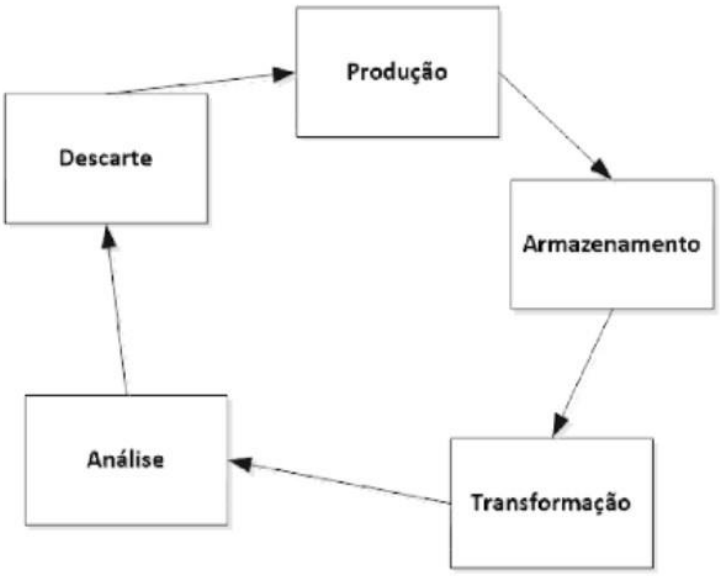
Toda ciência almeja a obtenção de conhecimento e informação de modo sistemático. Com a Ciência de Dados não é diferente. Embora o termo tenha surgido por volta dos anos 1960, ela é considerada uma ciência nova, que estuda todo o ciclo de vida do dado, desde a sua produção até o seu descarte. Por vezes, esta ciência é erroneamente considerada uma estatística com nome mais sofisticado, que está orientada à pura análise de dados. Na verdade, a Ciência de Dados é mais abrangente do que a estatística (descritiva ou inferencial) e é composta por outras ciências, modelos, tecnologias, processos e procedimentos relativos ao dado (AMARAL, 2016b).

Date (2003 apud GOLDSCHMIDT; PASSOS; BEZERRA, 2015, p. 12) afirma que “um banco de dados é uma coleção integrada de dados, organizada de tal forma a facilitar o

armazenamento eficiente, assim como sua modificação e recuperação". Pode-se considerar também como banco de dados, de forma ainda incipiente, uma simples agenda telefônica ou o arquivo físico de uma empresa, desde que estejam organizados em ordem alfabética, por exemplo, ou levem em consideração algum outro requisito de armazenagem.

Considerando o ciclo de vida do dado, só existe sentido em sua produção desde que o mesmo seja armazenado em algum dispositivo eletrônico para utilização futura. Depois de ser produzido e mantido em determinada mídia, o dado passa por uma transformação. Um exemplo simples é a formatação de um arquivo XML de uma nota fiscal eletrônica para exibi-la em um formulário. A Figura 2 ilustra o ciclo de vida do dado.

Figura 2 - Ciclo de vida do dado



Fonte: Amaral, 2016b, p. 6 (adaptado).

Uma vez transformado, o dado passa por operações de análise, que consistem na extração de informação e conhecimento, como ocorre numa consulta SQL para visualizar as vendas diárias. Por fim, o dado será, cedo ou tarde, vítima de um processo de descarte (AMARAL, 2016b).

Segundo Amaral (2016b, p. 6), Ciência de Dados, portanto, pode ser definida "[...] como os processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida: da produção ao descarte [...]". Já o profissional desta área, conforme o autor, denominado cientista de dados, deve possuir um perfil delineado à altura de quem procura se destacar no mercado. O cientista de dados deve ter um conhecimento abrangente e horizontal,

implementar desafios dispondo das melhores práticas de gerência de projetos, além de possuir aspectos de liderança para comandar uma equipe de especialistas.

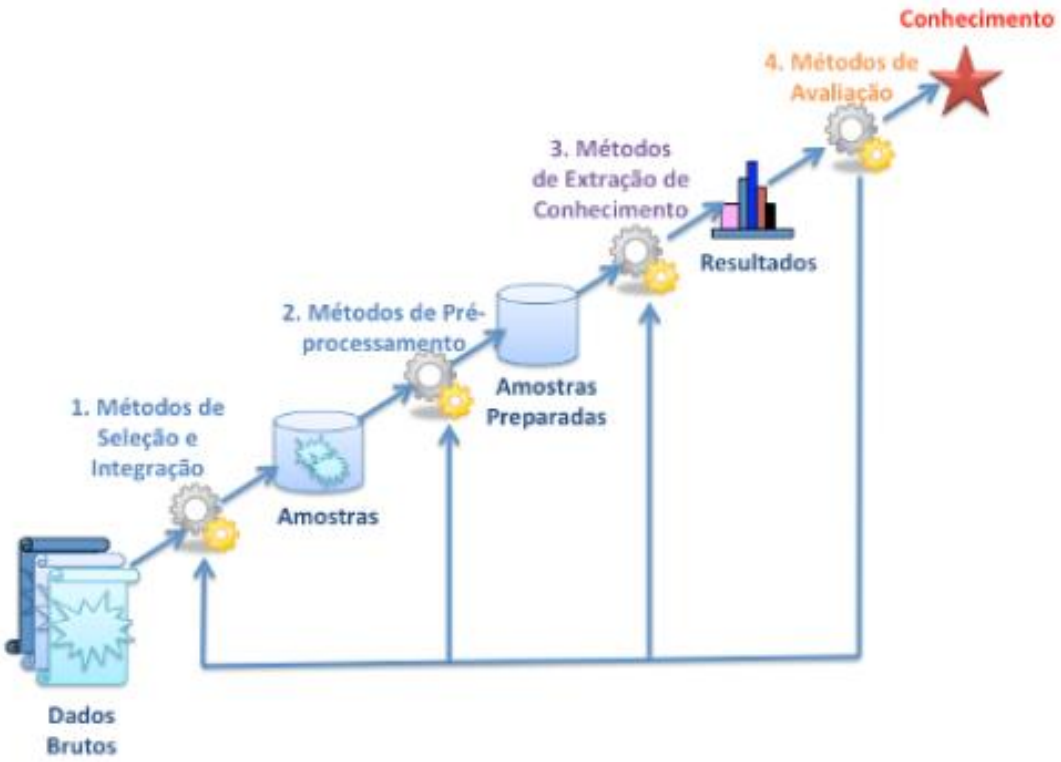
2.2 Descoberta de Conhecimento em Bases de Dados (*KDD*)

Sem o auxílio de ferramentas computacionais apropriadas, torna-se inviável para o ser humano analisar grandes quantidades de dados. Sendo assim, não se deve abrir mão de ferramentas que possibilitam a análise, a interpretação e a relação entre dados, a fim de que as melhores estratégias de ação sejam elaboradas e selecionadas. Para isso, surge uma área denominada Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases – KDD*), que tem a Mineração de Dados (*Data Mining*) como uma de suas etapas (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

O processo de *KDD* é constituído pelas seguintes fases: Obtenção de Dados, Pré-Processamento, Mineração de Dados e Pós-Processamento. Na primeira fase, os dados de uma determinada área de interesse são organizados, com a finalidade de descobrir algum conhecimento útil. Na segunda fase, os dados são organizados em um repositório único, como um *Data Warehouse*, que é um depósito de dados corporativos voltados ao apoio à decisão (AMARAL, 2016b). Na fase da Mineração de Dados, há a resolução de tarefas como predição, agrupamento ou associação. Por último, os resultados alcançados passam por validação, avaliação e formatação, o que permite sua visualização em gráficos, tabelas e relatórios. Todas as etapas do processo *KDD* podem ser executadas mais de uma vez, na sequência habitual ou fora dela (SILVA; PERES; BOSCARIOLI, 2016).

Segundo Han, Kamber e Pei (2011), o *KDD* só pode ser obtido após a execução dos métodos de Pós-Processamento, também chamados de métodos de avaliação, conforme pode ser verificado na Figura 3.

Figura 3 - Processo KDD



Fonte: Han, Kamber e Pei, 2011, p. 7 (adaptado).

A Mineração de Dados é, pois, a etapa responsável por gerar um modelo de conhecimento que é visualizado, analisado e interpretado no Pós-Processamento, no qual também os resultados obtidos são avaliados e novas alternativas para a investigação dos dados são definidas de acordo com especialistas (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

3 PROCEDIMENTOS METODOLÓGICOS

O presente artigo tem como procedimento metodológico principal levantamento bibliográfico. Para Severino (2007, p. 122), uma pesquisa bibliográfica "(...) é aquela que se realiza a partir do registro disponível, decorrente de pesquisas anteriores, em documentos impressos, como livros, artigos, teses etc.". Pesquisas que adotam essa abordagem tendem ao enriquecimento pessoal, intelectual e profissional de quem as elabora. Elas contribuem, ao mesmo tempo, para a produção científica e acadêmica e transformam o processo de aprendizagem de puramente teórico para oportunamente prático.

4 RESULTADOS E DISCUSSÃO

As tecnologias emergentes são inovações ou aperfeiçoamento de uma tecnologia já existente das diversas áreas da tecnologia que estão em tendência. Essas evoluções trazem melhorias nos processos, ajudam na tomada de decisão e aumento da lucratividade para todos os que usufruem delas – aplicadas em uma organização ou para uso pessoal. Dentre tantas tecnologias, é possível apontar algumas que utilizam a ciência de dados e que estão em crescente uso, tais como: *Data Mining* (Mineração de Dados), *Artificial Intelligence* (Inteligência Artificial) e *Machine Learning* (Aprendizado de Máquina).

4.1 *Data Mining*

A *Data Mining* ou Mineração de Dados utiliza ferramentas que podem analisar os dados, descobrir problemas ou oportunidades escondidas nos relacionamentos dos dados que podem auxiliar na tomada de decisão. As principais ferramentas para Mineração de Dados são *RapidMiner* e *Weka*.

A Mineração de Dados pode ser usada em diversas áreas e aplicações, como na educação, para identificar possíveis melhorias no ensino; em indústrias, para identificar anomalias no processo produtivo; em empresas e comércio, para identificar possíveis consumidores de determinados produtos e serviços; além da medicina, finanças, robótica, dentre outras áreas (AMARAL, 2016a).

4.2 *Artificial Intelligence*

A *Artificial Intelligence* ou Inteligência Artificial (mencionada pela sigla em português IA) são algoritmos de aprendizagem que são capazes de simular a capacidade humana de raciocínio, assim produzindo máquinas, equipamentos e/ou softwares inteligentes capazes de resolver problemas. A IA possui características como capacidade de raciocínio aplicando regras lógicas; aprendizagem com erros e acertos para a tomada de decisão mais eficaz; reconhecimento de padrões visuais, sensoriais e comportamentais. Para desenvolver a IA, as principais linguagens de programação em ascendência são Java, Python e Linguagem R.

A Inteligência Computacional, por sua vez, é um ramo da IA e se relaciona também com o processo de *KDD* (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). A IA pode ser

aplicada em diversas áreas, como no *Smartphone* para reconhecimento de voz; em *e-mail* para identificar um *spam*; em empresas para a produção automática de um determinado produto; em veículo autônomo que é capaz de identificar obstáculos e chegar ao seu destino; além de sistemas de *Streaming* para recomendação de vídeos, filmes e séries.

4.3 *Machine Learning*

A *Machine Learning* ou Aprendizado de Máquina é uma divisão da IA e tem como característica ensinar a máquina (neste caso, a máquina refere-se ao software que será programado). Este ensinamento consiste em identificar padrões e analisar outros dados que possam chegar a um determinado resultado, como letras, números e imagens (AMARAL, 2016b). No caso de imagens, a máquina as divide em camadas e identifica as matrizes, os elementos das matrizes e padrões entre as imagens, ou seja, as camadas das imagens são quebradas em *pixels* e os *pixels* das outras imagens são comparados para que se possa criar padrões. Como o Aprendizado de Máquina é uma divisão da IA, suas ferramentas para desenvolvimento são as mesmas.

Após criada a máquina, seus dados são alimentados pelos próprios usuários, como a publicação de uma foto em uma rede social. Dessa forma, a máquina vai identificar automaticamente seus amigos na foto e o usuário, por sua vez, irá identificar os que a máquina não foi capaz de reconhecer. Com essa ação, o usuário alimenta o banco de dados da máquina, que aprende a identificar as pessoas para as próximas publicações. Já em um assistente virtual, a máquina identifica a voz ou o texto escrito pelo usuário. Depois disso, é feita uma varredura no banco de dados e uma resposta é devolvida. Em suma, a voz ou o texto escrito alimenta o banco de dados.

5 CONSIDERAÇÕES FINAIS

O cientista de dados utiliza o banco dados para descobrir o real valor contido nos dados, prevendo, assim, ações futuras. Ele é responsável por aplicar o algoritmo capaz de minerar os dados, aplicando a Inteligência Artificial em regras lógicas, de modo que esse algoritmo aprende a identificar padrões que resultam em dados analíticos. O cientista de dados, por sua vez, deve manter os padrões dos dados, pois os dados não padronizados poderão gerar modelos de máquinas inapropriados. No futuro, a quantidade de profissionais

de programação poderá diminuir significativamente, uma vez que a máquina é capaz de se autoprogramar e programar outras máquinas.

A partir dos estudos realizados, é possível concluir que a Mineração de Dados (*Data Mining*) é realmente imprescindível no apoio à decisão nas organizações, uma vez que estas geram diariamente um aglomerado de informações e precisam lidar com elas de maneira eficiente, para que possam satisfazer as necessidades e os desejos de seus clientes e oferecer vantagem competitiva diante de seus concorrentes.

REFERÊNCIAS

AMARAL, F. **Aprenda mineração de dados**: teoria e prática. Rio de Janeiro: Alta Books, 2016a.

___. **Introdução à ciência de dados**: mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016b.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data mining**: conceitos, técnicas, algoritmos, orientações e aplicações. Rio de Janeiro: Elsevier, 2015.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** concepts and techniques. 3. ed. Waltham: Morgan Kaufmann, 2011.

PORTO, F.; ZIVIANI, A. Ciência de dados. In: III SEMINÁRIO DE GRANDES DESAFIOS DA COMPUTAÇÃO NO BRASIL, 2014, Rio de Janeiro.

SEVERINO, A. J. **Metodologia do trabalho científico**. 23. ed. São Paulo: Cortez, 2007.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicações em R. Rio de Janeiro: Elsevier, 2016.