

Descreva os passos envolvidos na mineração de dados, quando vista como um processo de descoberta de conhecimento.

Mineração de dados ou *Data Mining* consiste em um processo analítico projetado para explorar grandes quantidades de dados (tipicamente relacionados a negócios, mercado ou pesquisas científicas), na busca de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis, anomalias, correlações e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados de forma automática. Para tal processo, a partir do conjunto de dados “bruto”, o processo de descoberta de conhecimento pode ser dividido em cinco etapas: seleção de dados, pré-processamento dos dados, transformação dos dados, aplicação de regras e padrões e interpretação/validação.

- **Seleção de dados:**

Esta etapa consiste na seleção dos dados que serão mais relevantes para análise. Ao mesmo tempo, esta etapa é de grande importância, uma vez que é preciso definir os objetivos de negócios onde se quer chegar (no caso de contexto empresarial), formular as perguntas que deseja responder e quais parâmetros são importantes para a pesquisa.

- **Pré-processamento:**

Esta etapa consiste no processamento dos dados com o objetivo de tratar inconsistência nos dados, ruídos como duplicatas, *outliers*, valores perdidos. Dependendo do tipo do conjunto de dados, nesta etapa realiza-se também uma redução no número de dimensões no vetor de características com o objetivo de melhorar a eficiência computacional.

- **Transformação:**

A maioria dos dados reais não apresentam uniformidade entre si, muitas das vezes com escala e tipos diferentes. Esta etapa consiste em deixar os dados de forma uniforme para que aplicação das regras posteriormente não cause resultados tendenciosos. O tipo de transformação a ser aplicada irá depender do tipo de dados disponíveis e o problema que deseja responder.

- **Aplicação de regras e padrões:**

Consiste basicamente no processo de extração das novas informações. Qual tipo de técnica/tarefa a ser usada irá depender do tipo dos dados disponíveis e qual tipo de informação que se deseja extrair. Há vários tipos de algoritmos e técnicas, como por exemplo, regra de associação, redes neurais, árvores de decisão, clusterização, classificação, regressão, sumarização, padrão de sequência.

- **Interpretação/validação:**

Depois da aplicação das regras através das técnicas de mineração de dados, esta é a etapa da interpretação das informações geradas. Cabe ao analista identificar os padrões identificados e verificar se os mesmos respondem às perguntas geradas no início da análise. Em caso negativo, deve-se ajustar os parâmetros dos algoritmos, realizar validação dos dados e realizar as etapas anteriores, caso necessário.