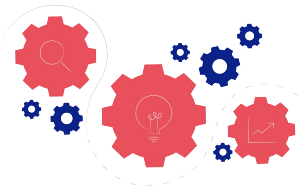




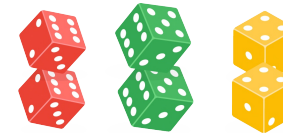
By @kakashi_copiador



Conjunto de **processos, métodos, teorias, ferramentas** e **tecnologias open-end**...



...Utilizadas p/ **explorar, organizar e analisar**, de forma semiautomática, uma grande quantidade de dados brutos...



... Com o intuito de **identificar, descobrir, extrair, classificar e agrupar** informações implícitas desconhecidas...



... Além de **avaliar correlações, tendências e padrões consistentes de comportamento** potencialmente úteis.

Como regras de **associação** ou **seqüências temporais**...

...De forma **não trivial**...

Mineração de Dados

Necessita, por vezes, utilizar **processamento paralelo**.

Utilizado em áreas como Marketing, Finanças, Manufatura, Saúde, etc.

Também chamada de **Data Mining** ou **Prospecção de Dados**.



Apesar de geralmente ser utilizada em **conjunto com Data Warehouses**, não é obrigatório que o seja.

... Por meio de **técnicas estatísticas e matemáticas**.

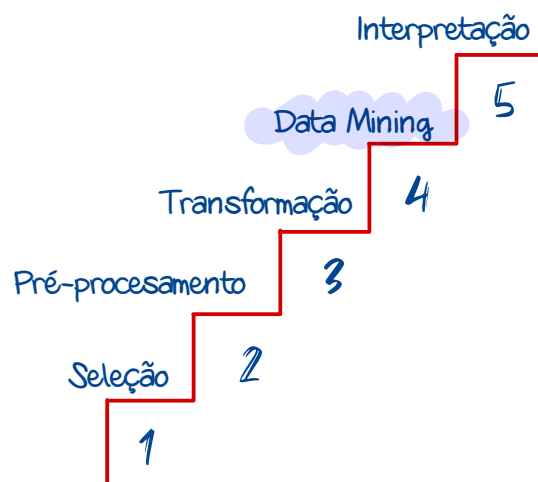
Ex:

- redes neurais;
- algoritmos genéticos;
- inteligência artificial;
- lógica nebulosa;
- análise de conglomerados (clusters);
- etc.



Processo de Descoberta de Conhecimento

- Knowledge Discovery in Databases (KDD) - Descoberta de Conhecimento em Banco de Dados.
- A mineração de dados é uma das fases do KDD.



- **Seleção**: selecionar um conjunto de dados ou se concentrar em um subconjunto de variáveis ou amostras.
- **Limpeza e Pré-processamento**: remoção de erros, coleta de informações, etc.
- **Transformação**: os dados são transformados e consolidados em formas apropriadas à mineração (sumarizando-os ou agregando-os).
- **Mineração de dados**: algoritmos e técnicas p/ extrair possíveis padrões úteis de dados.
- **Interpretação**: os padrões encontrados são avaliados e interpretados.



Mineração de Dados

Objetivos



PREVISÃO

- Prever comportamentos futuros com base em comportamentos passados.

IDENTIFICAÇÃO

- Identificar, através de padrões de dados, a existência de um item, um evento ou uma atividade.

CLASSIFICAÇÃO

- Particionar os dados p/ que diferentes categorias possam ser identificadas com base em combinações de parâmetros.

OTIMIZAÇÃO

- Otimizar o uso de recursos limitados (como tempo, espaço, dinheiro ou materiais) e maximizar variáveis de saída (como vendas e lucros), sob determinado conjunto de restrições.



Predição

Ou Previsão

Busca descrever a natureza de **ocorrências futuras** de certos eventos com base nos **acontecimentos passados**.



Classificação

Regressão

Objetiva prever um **número/valor**.

- Compreende a busca por uma **função** que mapeie os registros de um BD em um intervalo de **valores reais**.

Categorização de dados p/ a identificação de uma **classe** por meio de **múltiplos atributos**.

- As **classes** são **predefinidas** antes da análise dos resultados (**aprendizado supervisionado**).

PRINCIPAIS FERRAMENTAS DE CLASSIFICAÇÃO:

Árvores de Decisão

- Representação gráfica das **regras de classificação**, demonstrando visualmente as **condições e probabilidades** p/ se chegar a algum resultado.

Redes Neurais (Artificiais)

- Algoritmos inspirados no cérebro de animais, com capacidade de **aprender com experiências passadas**, **prever um comportamento** ou **reconhecer um padrão**, desde que apresentados e um conjunto de dados estruturados.

Mineração de Dados Técnicas e Tarefas



Clusterização

Ou Agrupamento

Particiona dados em **segmentos** previamente desconhecidos com **características semelhantes**.



Associação



Busca **descobrir relacionamentos** entre variáveis correlacionando a presença de um item com uma faixa de valores p/ outro conjunto de variáveis.

REGRAS DE ASSOCIAÇÃO: $X \rightarrow Y$

- Visam descobrir o **relacionamento entre variáveis** de um banco de dados.

PADRÕES SEQUENCIAIS

- Buscam descobrir padrões **sequências de ações ou eventos** de forma equivalente a certos relacionamentos temporais.

PADRÕES TEMPORAIS

- Semelhante à técnica dos Padrões Sequenciais, mas sempre **envolve um fator temporal** que permite diferenciá-los.

um intervalo regular de tempo

Uma regra de associação deve satisfazer alguma **medida de interesse** do analista de dados:

- **SUORTE/PREVALÊNCIA**: **frequência** com que um conjunto de itens **ocorre** no BD.
- **CONFIANÇA/FORÇA**: **probabilidade** de que exista uma **relação** entre itens.



@mapasdathai