# Ciência de Dados

## Descoberta de Conhecimento em Bases de Dados

# Descoberta de Conhecimento em Bases de Dados

A Descoberta de Conhecimento em Bases de Dados (KDD, do inglês *Knowledge Discovery on Databases*) (FAYYAD et al 1996) é uma metodologia para extração de informações a partir de bases de dados.

O KDD é um processo interdisciplinar que envolve conhecimentos e técnicas de diversas áreas. Entre elas estão a Computação, a Matemática e a Estatística. Os conhecimentos aplicados são variados, como Bancos de Dados, Aprendizado de Máquina, Técnicas de visualização de dados, entre outros.

# Fases do processo

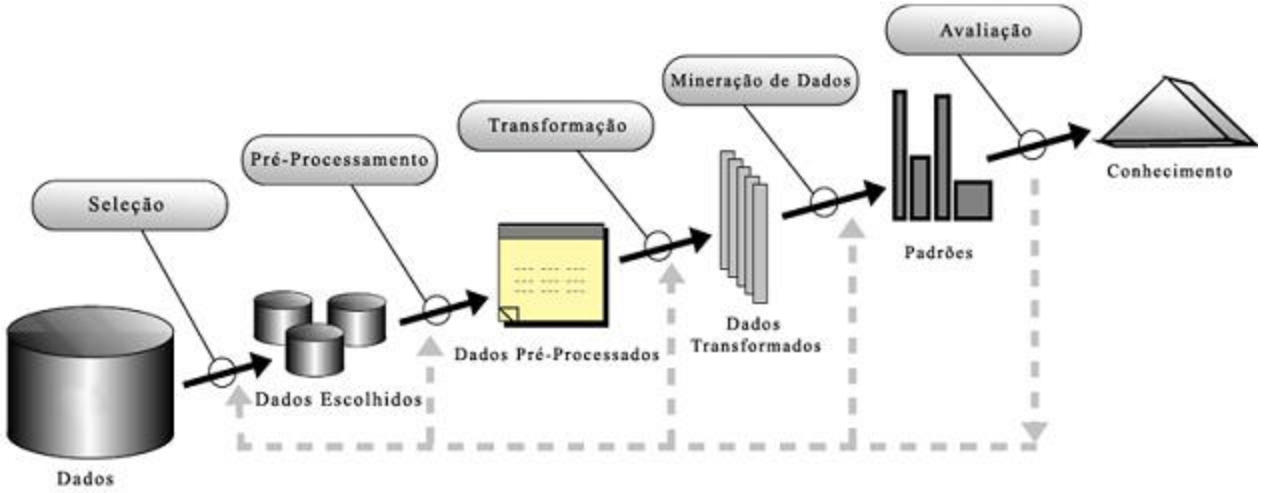


Imagem: Fayyad et Al, 1996 via Lira et al, 2016

## Seleção

Nesta primeira fase do processo são selecionados os dados relevantes para a análise. Aqui podemos incluir os critérios de escolha dos dados, os meios de coleta e as fontes de dados utilizadas.

Aqui são definidos os conjuntos de dados contendo as possíveis variáveis (atributos) e registros (instâncias) que serão objeto de análise. Muitas vezes, este processo é realizado por um especialista da área de domínio dos dados (FERREIRA et al, 2018). Podem ser envolvidas consultas em bancos de dados, filtragem de dados com base em certos atributos ou características, ou outra técnica de seleção apropriada

## Pré-processamento

Esta fase consiste em transformar os dados de maneira adequada para serem analisados, de forma a garantir que estejam em em formato adequado para as fases seguintes do processo.

Algumas das atividades que podem ocorrer nesta fase:

## Limpeza de Dados

Esta etapa envolve a identificação e correção de erros nos dados, como valores ausentes, *outliers*, ruídos e inconsistências. Isso pode incluir técnicas como preenchimento de valores ausentes, remoção de *outliers* e correção de erros de digitação.

## Integração de Dados

Caso os dados sejam provenientes de várias fontes, pode ser necessário integrá-los em um único conjunto de dados coeso. Isso envolve mapear esquemas de dados, resolver inconsistências nos formatos dos dados e combinar diferentes conjuntos de dados em um único conjunto.

## Tratamento de valores ausentes

Valores ausentes são comuns em conjuntos de dados do mundo real devido a vários motivos, como erros de medição, falhas na coleta de dados, ou simplesmente porque as informações não estão disponíveis.

É importante lidar com valores ausentes antes de realizar qualquer análise ou modelagem de dados, pois a presença de valores faltantes pode levar a resultados incorretos ou enviesados. A escolha da técnica de tratamento de valores ausentes depende da natureza dos dados, do problema abordado e do algoritmo de mineração de dados. Algumas técnicas para tratamento de valores ausentes são:

## Remoção de Dados

Uma abordagem simples é remover as observações (linhas) ou os atributos (colunas) que contêm valores ausentes. Isso é viável quando os valores ausentes são apenas uma pequena parte do conjunto de dados e sua remoção não causa perda significativa de informações.

## Preenchimento de Valores Ausentes

Esta abordagem envolve substituir os valores ausentes por valores estimados. Isso pode ser feito de várias maneiras, incluindo:

**Preenchimento com a Média/Mediana/Moda**

Os valores ausentes são substituídos pela média, mediana ou moda dos valores existentes na mesma variável.

**Preenchimento por Imputação**

Os valores ausentes são estimados com base em outras variáveis correlacionadas. Isso pode ser feito usando técnicas como regressão, k-vizinhos mais próximos (KNN) ou algoritmos de aprendizado de máquina.

**Preenchimento por Valor Fixo**

Os valores ausentes são substituídos por um valor fixo escolhido manualmente, como 0 ou -1. Essa abordagem é adequada quando a ausência dos valores é informativa.

**Imputação Multivariada**

Nesta abordagem, os valores ausentes são estimados com base em todas as outras variáveis do conjunto de dados. Isso leva em consideração a relação entre as variáveis e pode fornecer estimativas mais precisas dos valores ausentes.

## Modelagem de Aprendizado de Máquina

Em algumas situações, é possível utilizar modelos de aprendizado de máquina para prever os valores ausentes com base nos padrões existentes nos dados. Isso pode ser feito treinando um modelo nos dados completos e usando-o para prever os valores ausentes.

## Transformação

Esta fase consiste em preparar os dados para serem analisados, de forma a auxiliar os algoritmos na fase posterior, a Mineração de Dados.

## Normalização

Normalizar os dados é uma etapa comum na transformação de dados. Isso envolve a escala dos valores dos atributos para que fiquem em uma faixa específica, como 0 a 1 ou -1 a 1. A normalização é útil quando os atributos possuem escalas diferentes e quando algoritmos de aprendizado de máquina podem ser sensíveis a essas diferenças.

## Padronização

A padronização é semelhante à normalização, mas em vez de escalar os valores para uma faixa específica, os dados são transformados para que tenham uma média zero e um desvio padrão de um. Isso é útil para algoritmos que assumem que os dados estão distribuídos de forma normal.

## Discretização de Dados

Em alguns casos, pode ser útil transformar dados contínuos em dados discretos, especialmente para algoritmos que requerem variáveis categóricas. Isso pode ser feito por meio de técnicas de discretização, como a divisão em intervalos fixos ou com base em critérios estatísticos.

## Redução de Dimensionalidade

Em alguns casos, pode ser útil reduzir a dimensionalidade dos dados, especialmente se houver muitos atributos ou se alguns atributos forem redundantes. Técnicas como Análise de Componentes Principais (PCA) ou Seleção de Características podem ser utilizadas para este fim.

## Codificação de variáveis categóricas

É um processo no qual as variáveis categóricas são convertidas em uma forma numérica que pode ser utilizada por algoritmos de aprendizado de máquina. As variáveis categóricas são aquelas que representam diferentes categorias ou grupos e não têm uma ordem natural entre elas. Por exemplo, cores (vermelho, azul, verde), tipos de produtos (eletrônicos, vestuário, alimentos) ou estados civis (solteiro, casado, divorciado).

Alguns algoritmos requerem que os dados de entrada sejam numéricos, o que significa que as variáveis categóricas precisam ser convertidas em números antes de serem utilizadas. A escolha da técnica de codificação depende do tipo de variável categórica e do algoritmo que está sendo utilizado. Algumas técnicas para realizar a codificação de variáveis categóricas são:

**Codificação One-Hot (One-Hot Encoding)**

Nesta abordagem, cada categoria única é representada por uma variável binária (0 ou 1). Para uma variável com $n$ n categorias únicas, são criadas $n$ novas variáveis, onde cada variável indica se a observação pertence à categoria correspondente. Por exemplo, se tivermos a variável categórica "cor" com três categorias únicas (vermelho, azul, verde), ela seria codificada como três variáveis binárias: cor_vermelho, cor_azul e cor_verde.

## Codificação de Números Inteiros (Integer Encoding)

Nesta abordagem, cada categoria única é mapeada para um número inteiro. Por exemplo, se tivermos a variável categórica "estado civil" com três categorias únicas (solteiro, casado, divorciado), elas podem ser mapeadas para os números 1, 2 e 3, respectivamente.

## Codificação de Frequência (Frequency Encoding)

Nesta técnica, cada categoria é substituída pela frequência com que aparece no conjunto de dados. Isso pode ser útil quando a frequência das categorias é informativa para o modelo.

## Codificação Ordinal

Esta técnica é usada quando as categorias têm uma ordem natural entre elas. Neste caso, as categorias são codificadas como números inteiros de acordo com a ordem natural. Por exemplo, para uma variável categórica "tamanho de roupa" com categorias pequeno, médio e grande, elas podem ser codificadas como 1, 2 e 3, respectivamente.

## Mineração de Dados

Esta etapa consiste em aplicar algoritmos de mineração de dados para descobrir padrões e relações presentes nos dados. Os algoritmos são executados com dados preparados nas fases anteriores para extrair padrões, relações e informações úteis. Essa fase é central no processo de descoberta de conhecimento, visto que novos dados e informações são obtidos.

## Tarefas na Mineração de Dados

Nesta fase, diferentes algoritmos podem ser aplicados, de acordo com os dados disponíveis e o objeto de estudo. A organização destes algoritmos podem ser abordados de maneira preditiva ou descritiva.

**Tarefas preditivas**

As tarefas preditivas se concentram na previsão de resultados futuros com base em dados históricos, as tarefas descritivas se concentram em entender e descrever os padrões presentes nos dados.

As tarefas preditivas visam prever um resultado futuro com base em dados históricos e características conhecidas.

Exemplos de tarefas preditivas:

- Classificação: envolve atribuir categorias ou rótulos a instâncias de dados com base em suas características. Por exemplo, classificar e-mails como spam ou não spam, prever se um paciente tem uma determinada doença com base em seus sintomas, ou classificar transações financeiras como fraudulentas ou legítimas.
- Regressão: o objetivo é prever um valor contínuo com base em variáveis de entrada. Por exemplo, prever o preço de uma casa com base em características como tamanho, localização e número de quartos, ou prever a demanda por um produto com base em fatores como preço, promoções e época do ano.

- Análise de Séries Temporais: Prever valores futuros com base em padrões observados em dados temporais passados. Por exemplo, prever a demanda por um produto ao longo do tempo.

## Tarefas descritivas

As tarefas descritivas visam descrever os padrões e relações presentes nos dados, geralmente sem a preocupação de prever resultados futuros.

Exemplos de Tarefas Descritivas:

- Agrupamento: envolve identificar grupos ou clusters de instâncias de dados que compartilham características semelhantes. Isso é útil para descobrir estruturas subjacentes nos dados e identificar padrões naturais de agrupamento. Por exemplo, segmentar clientes em grupos com base em seus hábitos de compra, agrupar documentos semelhantes em coleções de texto, ou identificar grupos de genes com expressão semelhante em dados genômicos.

- Associação: A associação envolve descobrir padrões de relação entre diferentes itens em conjuntos de dados transacionais. Isso é comumente usado em análise de cestas de compras para identificar produtos que são frequentemente comprados juntos. Por exemplo, identificar que clientes que compram pão também tendem a comprar manteiga..

## Avaliação e interpretação de resultados

Após os resultados obtidos na fase de mineração de dados, é necessário analisar tais resultados e apresentá-los de maneira que compreensível.

Algumas das atividades que podem ocorrer durante esta fase são:

## Avaliação de Modelos

Os modelos gerados durante a fase de mineração de dados são avaliados quanto à sua qualidade e desempenho. Isso pode ser feito usando métricas apropriadas para o tipo de análise realizada, como acurácia, precisão, recall, F1-score, erro médio quadrático, entre outras.

## Validação Cruzada

A validação cruzada é uma técnica comum usada para avaliar a capacidade de generalização dos modelos. Os dados são divididos em conjuntos de treinamento e teste múltiplos, e o modelo é treinado e testado várias vezes, com diferentes divisões dos dados.

## Interpretação de Resultados

Os resultados dos modelos são interpretados para extrair insights e conhecimentos úteis. Isso pode envolver identificar padrões interessantes nos dados, entender as relações entre as variáveis, descobrir insights sobre o comportamento dos clientes ou usuários, entre outros.

## Visualização de Dados

A visualização de dados é uma ferramenta poderosa para comunicar os resultados da análise de forma compreensível e intuitiva. Gráficos, tabelas, mapas e outras técnicas de visualização são utilizados para apresentar os resultados de maneira eficaz.

## Comparação de Modelos

Se foram testados vários modelos durante a fase de mineração de dados, é importante comparar seus desempenhos e identificar o modelo mais adequado para os objetivos da análise.

## Documentação dos Resultados

Todos os resultados obtidos durante a fase de avaliação e interpretação são documentados de maneira clara e completa. Isso inclui os métodos utilizados, as métricas de avaliação, as conclusões tiradas e qualquer insight ou descoberta relevante.

## Comunicação dos Resultados

Os resultados da análise são comunicados às partes interessadas de maneira clara e acessível. Isso pode envolver a preparação de relatórios, apresentações ou outros materiais de comunicação que ajudem a transmitir os insights obtidos de forma eficaz.

## Conhecimento descoberto

Esta é a fase final, em que trata utilização dos resultados obtidos. Nesta, o conhecimento descoberto é interpretado, validado e utilizado para tomar decisões informadas e gerar valor em um contexto específico.

# Referências

FACELI, Katti et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. 2. ed.
São Paulo: LTC, 2021. 304 p. ISBN 978-8521637349.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH P. From Data Mining to Knowledge Discovery in Databases. AI Magazine. Vol. 17, n. 3. 1996

FERREIRA, J.C.; ROSA, C.R.M.; STEINER, M.T.A. Knowledge Discovery in Database e Data Mining: Uma contribuição bibliomética. In: Anais do XXXVIII Encontro Nacional de Engenharia de Produção. Maceió, Alagoas, Brasil. 2018.

FRAWLEY, W.J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C.J. Knowledge Discovery in Databases: An Overview. AI Magazine. Vol. 13, n. 3. 1992.

LIRA, K.C.; de OLIVEIRA, M.A.; MAGALHÃES, R.P.; GONÇALVES, E.J.T. Utilizando Mineração de Dados e Sistemas Multiagentes na Análise da Evasão em Educação a Distância por Meio do Perfil dos Alunos. In: Anais do Encontro Nacional de Inteligência Artificial e Computacional. Recife, 2016.