

CURSO SUPERIOR DE ADS

Descoberta de Conhecimento em Bases de Dados (KDD)



Prof. Fernando Marlon Soares Figueiredo

Disciplina: Ciência de Dados e Bigdata



AULA DE HOJE – DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

Introdução à KDD (Knowledge Discovery in Databases)

Etapas do Processo KDD

Exercícios práticos



Antes...

O que é Big Data?

O termo Big Data refere-se a conjuntos de dados extremamente grandes e complexos que não podem ser processados usando métodos tradicionais.

Esses dados podem incluir informações de redes sociais, transações financeiras, dados meteorológicos e muito mais.

O Big Data é caracterizado por quatro Vs: volume, velocidade, variedade e valor.

4 V's

- **Volume:** O Big Data é caracterizado por grandes volumes de dados. Esses dados podem ser tão grandes que não podem ser armazenados em um único servidor.
- **Velocidade:** O Big Data é caracterizado por sua velocidade. Os dados são gerados em tempo real, o que significa que é necessário processá-los rapidamente para obter insights.
- **Variedade:** O Big Data é caracterizado por sua variedade. Os dados podem vir de várias fontes diferentes, incluindo mídias sociais, sensores, transações financeiras e muito mais.
- **Valor:** O Big Data é caracterizado pelo valor que pode ser extraído dele. Os dados podem ser usados para obter insights valiosos sobre os clientes, processos de negócios e muito mais.

Em síntese

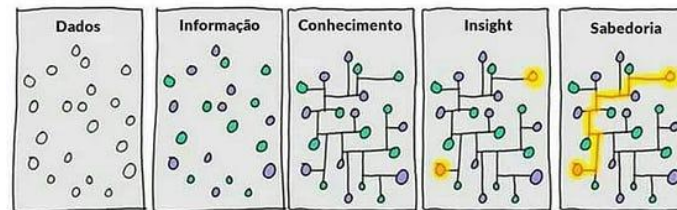


Muitos dados



Dados sozinhos *vs juntos*

- Dados sozinhos não podem dizer muita coisa. Por exemplo, eu posso te dizer que meu cabelo é preto. Isso não te leva a tirar muitas conclusões sobre minha personalidade. Porém, se eu te der uma pequena tabela em excel com alguns dos meus filmes favoritos, provavelmente você terá conclusões mais significativas sobre mim.
- Objetivo do Big Data: Obter insights valiosos a partir de grandes volumes de dados.



Desafios

Nem tudo são flores

- Questões de Privacidade: A era do Big Data traz preocupações sobre privacidade, especialmente com o aumento de dispositivos conectados (IoT).
- A coleta massiva de dados por empresas e governos levanta questões sobre quem mais tem acesso a informações pessoais?

IoT – Internet das Coisas

- Denominado por muitos como 4ª Revolução Industrial
- Junto com tecnologias como computação em nuvem, robótica, inteligência artificial, impressão 3D e realidade virtual e aumentada vem mudando radicalmente o mundo em que vivemos.
- Trata-se basicamente da conexão de bilhões de dispositivos à internet.



E o volume cada vez maior de dados

- Também é válido citar o exemplo das redes sociais. Hoje a Meta possui mais de 1 bilhão de usuários, ou seja, mais de 3,14 bilhões de pessoas conectadas, gerando dados através de curtidas, posts e compartilhamentos.
- Vantagem: úteis a criação de serviços inovadores, como os que já vieram: Netflix, Amazon, Spotify, etc.
- Mas e quem mais possui acesso a nossos dados?
- E no contexto de IA generativa, onde vão nossos dados que ficam no CHAT GPT?

Caso Snowden

Vazamento de Dados (2013): Edward Snowden, ex-funcionário da NSA, vaza documentos secretos expondo programas de vigilância em massa dos EUA.

Fuga e Asilo: Após as revelações, Snowden foge para Hong Kong e depois para a Rússia, onde recebe asilo temporário e, mais tarde, cidadania russa.

Debate Global: As revelações de Snowden geram um debate mundial sobre privacidade, segurança e vigilância governamental.

Impacto Legal e Legislativo: As ações de Snowden levam à revisão de políticas de vigilância nos EUA, incluindo a aprovação do USA Freedom Act em 2015.

Vida em Exílio: Snowden continua vivendo na Rússia, sendo um ativista vocal pela privacidade e enfrentando acusações de espionagem nos EUA.



Apesar disso,
segue sendo
um Oceano de
possibilidades

Sugestão

Estatística Prática

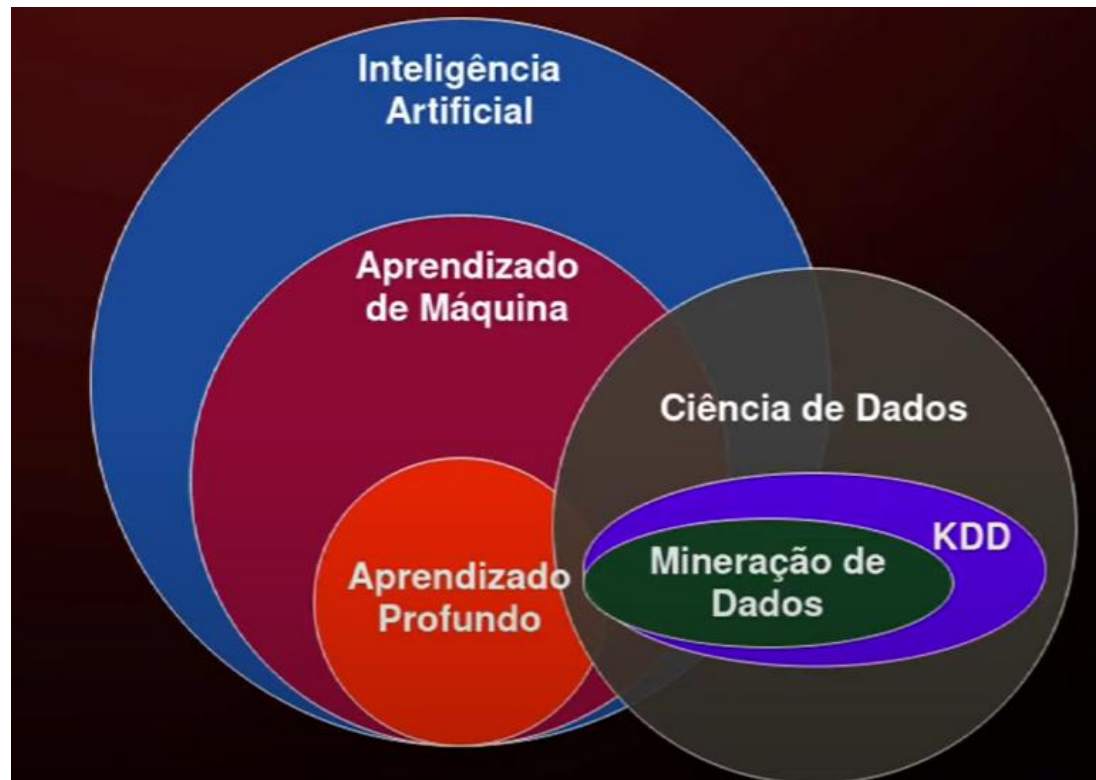
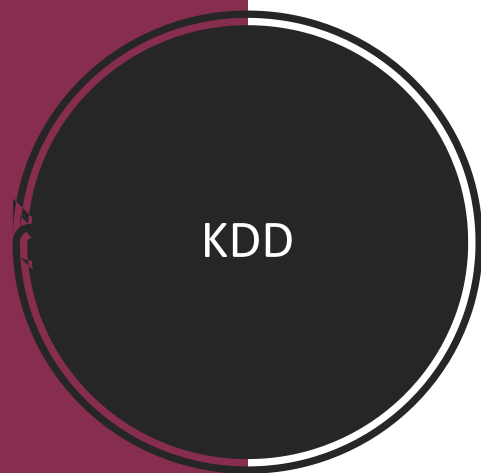
para Cientistas de Dados

50 CONCEITOS ESSENCIAIS



Knowledge Discovery in Databases (KDD)





Em síntese

- KDD:
 - Processo de geração de conhecimento
- Mineração de Dados
 - Técnicas para identificar relações ou padrões nos dados
 - Parte do KDD

Definição

- KDD é um processo abrangente que visa descobrir conhecimentos úteis e significativos a partir de grandes volumes de dados
- KDD ou Descoberta de Conhecimento em Bancos de Dados, é um processo interdisciplinar que combina técnicas de bancos de dados, estatística, aprendizado de máquina e visualização de dados para extrair conhecimento útil a partir de dados brutos.

O objetivo principal do KDD é transformar dados em informações, informações em conhecimento e conhecimento em ação.

Etapas do processo de KDD



1. Seleção de Dados A primeira etapa do processo de KDD é a seleção de dados. Nessa fase, são selecionados os dados relevantes para a análise. Isso pode incluir a identificação das **fontes de dados apropriadas**, a definição dos critérios de inclusão/exclusão e a coleta dos dados necessários. Por exemplo, suponhamos que desejamos realizar uma análise de vendas de uma empresa de varejo. Selecionamos os dados de vendas, incluindo informações como data, valor da venda, produtos vendidos e informações dos clientes.

2. Pré-processamento de Dados

Após a seleção dos dados, é necessário realizar o pré-processamento dos mesmos. Nessa etapa, os dados brutos são preparados para análise. Isso envolve a **limpeza dos dados**, tratamento de **valores ausentes**, lidar com ruídos e anomalias, e realizar a normalização dos dados, se necessário. Por exemplo, em nosso conjunto de dados de vendas, podemos **remover entradas duplicadas**, **preencher valores ausentes** e ajustar **formatos de data**.

Etapas do processo de KDD



3. Transformação de Dados

Uma vez que os dados tenham passado pelo pré-processamento, é hora de transformá-los em uma forma adequada para a análise. Isso pode envolver técnicas como **agregação, redução de dimensionalidade e codificação de dados**. Por exemplo, podemos agrupar as vendas por mês para obter uma visão mais ampla dos padrões sazonais de vendas.

4. Mineração de Dados

A etapa de mineração de dados é o coração do processo de KDD. Aqui, aplicamos algoritmos de mineração para descobrir padrões, relações e insights nos dados. Existem diferentes técnicas de mineração de dados disponíveis, como **associação, classificação, clusterização e detecção de anomalias**. Voltando ao nosso exemplo de vendas, poderíamos usar técnicas de mineração para identificar quais produtos são frequentemente comprados juntos por clientes.

Etapas do processo de KDD



5. Avaliação e Interpretação dos Resultados

Uma vez que os padrões e *insights* tenham sido descobertos, é importante **avaliar** e **interpretar** os **resultados**. Isso envolve a análise dos padrões encontrados, a validação de sua significância estatística e a interpretação do seu impacto no contexto do problema em estudo. Voltando ao exemplo de vendas, podemos avaliar a força das associações encontradas entre os produtos e identificar oportunidades de *cross-selling*.

6. Utilização do Conhecimento Descoberto

A última etapa do processo de KDD é a utilização do conhecimento descoberto para tomar decisões informadas. Os *insights* obtidos podem ser usados para melhorar processos, otimizar estratégias, fazer previsões e tomar medidas corretivas, dependendo do contexto do problema. No caso das vendas, o conhecimento descoberto pode direcionar campanhas de marketing segmentadas com base nas associações encontradas entre os produtos.

KDD (Knowledge Discovery in Databases)

- KDD ou "Processo de Descoberta de Conhecimento", segundo Fayyad, Piatetsky-Shapiro e Smyth, é um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de conjuntos de dados.
- A característica "não trivial" diz respeito à complexidade existente na execução e manutenção dos processos de KDD;

"Interativo" representa a relevância de possuir um elemento que controle o processo;

"Iterativo" indica a possibilidade de repetições em qualquer uma das etapas do processo;

"Conhecimento útil" aponta para a indicação de que o objetivo foi alcançado.



Fonte: tradução adaptada de (FAYYAD *et al.* 1996a)

Aplicações

Negócios: Análise de comportamento do consumidor, detecção de fraudes, e otimização de marketing. *Exemplo:* Analisar padrões de compra para personalizar campanhas de marketing.

Saúde: Análise de dados de pacientes para prever surtos de doenças e personalizar tratamentos. *Exemplo:* Utilizar dados de prontuários médicos para prever o risco de doenças crônicas.

Finanças: Análise de risco de crédito, detecção de fraudes e previsão de mercados financeiros. *Exemplo:* Modelar o risco de crédito de clientes para decidir a aprovação de empréstimos.

Desafios e considerações éticas

Qualidade dos Dados: Importância de dados limpos e precisos.

- *Exemplo:* Dados incompletos ou errados podem levar a conclusões equivocadas.

Privacidade e Segurança: Proteção de dados sensíveis e conformidade com regulamentações.

- *Exemplo:* Garantir a anonimização de dados pessoais para proteger a privacidade dos indivíduos.

Bias e Equidade: Evitar vieses nos algoritmos e garantir a justiça nas análises.

- *Exemplo:* Garantir que algoritmos de crédito não discriminem grupos específicos de pessoas.

Ferramentas e Tecnologias

- **Ferramentas de Software:** Introdução a ferramentas populares de mineração de dados, como R, Python (pandas, scikit-learn), e Weka. *Exemplo:* Usar Python e a biblioteca scikit-learn para construir e avaliar modelos de machine learning.
- **Plataformas de Big Data:** Uso de plataformas como Hadoop e Spark para lidar com grandes volumes de dados.
- *Exemplo:* Processar grandes conjuntos de dados em um cluster Hadoop para análises em larga escala.

Diferença entre Mineração de Dados e KDD (Descoberta de Conhecimento em Banco de Dados)

A mineração de dados é, na verdade, uma etapa dentro do processo mais amplo de KDD (Knowledge Discovery in Databases - Descoberta de Conhecimento em Banco de Dados).

Enquanto a mineração de dados é uma etapa técnica que aplica algoritmos para encontrar padrões em dados, KDD é um processo mais abrangente que inclui todas as etapas necessárias para transformar dados em bruto em conhecimento útil.

Técnicas de Mineração de Dados

Classificação: Atribuição de itens a categorias predefinidas.

Exemplo: Classificar e-mails como spam ou não spam.

Regressão: Modelagem da relação entre variáveis.

Exemplo: Prever vendas futuras com base em dados históricos.

Clusterização: Agrupamento de itens em clusters com características semelhantes.

Exemplo: Segmentar clientes com base em padrões de compra.

Regras de Associação: Descoberta de relações entre variáveis.

Exemplo: Encontrar produtos frequentemente comprados juntos (análise de cesta de compras).

REFERÊNCIAS

- **Referências**
- GUILHON, André et al. (org.). Jornada Python: uma jornada imersiva na aplicabilidade de uma das mais poderosas linguagens de programação do mundo. Rio de Janeiro, RJ: Brasport, 2022. E-book. Disponível em: <https://plataforma.bvirtual.com.br>. Acesso em: 30 jun. 2024.
- Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37–54.
- Witten, I. H., Frank, E., & Hall, M. A. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
- Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.
- Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons.
- <https://medium.com/blog-do-zouza/knowledge-discovery-in-databases-kdd-462ea2775715>