

Knowledge Discovery from Databases (KDD)

Descoberta de Conhecimento de Bases de Dados

Danilo Sipoli Sanches

Departamento Acadêmico de Computação
Universidade Tecnológica Federal do Paraná
Cornélio Procópio



Descoberta de conhecimento

A descoberta de conhecimento em bases de dados tem como objetivo encontrar padrões intrínsecos aos dados nela contidos, apresentando-os de forma a facilitar sua assimilação como conhecimento.

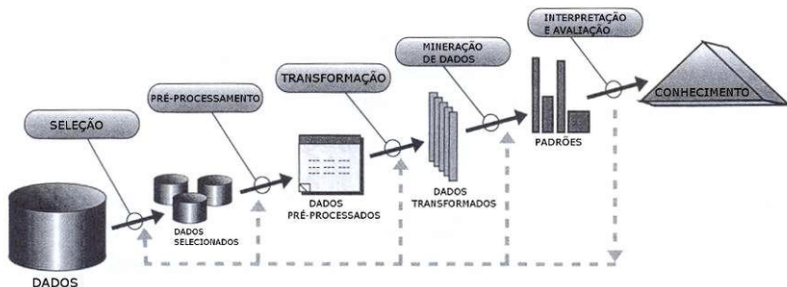


Figura: Etapas do processo de descoberta de conhecimento (KDD¹).

¹Fayyad U. et al. Knowledge Discovery and Data Mining: Towards a Unifying Framework, KDD 96, 1996

- ① Definição dos objetivos a serem atingidos;
- ② Criar uma nova base de dados e selecionar um conjunto de dados;
- ③ Limpeza e preparação:
 - Eliminar ruídos
 - Eliminar registros duplicados
 - Normalização de Valores
 - Transformação de campos
- ④ Classificação/Regressão (Algoritmos de AM);
- ⑤ Avaliação do Modelo.

Uma quantidade enorme de dados é coletada e armazenada.

- Dados da internet (Google);
- Facebook possui bilhões de usuários ativos;
- Amazon processa milhões de visitas por dia;
- Transações bancárias e de cartões de crédito.
- Os computadores tornaram-se mais baratos e mais eficientes para analisar o grande volume de dados.

Principais empresas de Big Data².



²<https://www.analyticsteps.com/blogs/companies-uses-big-data>

O que é mineração de dados

Não é:

- Fazer uma consulta a um banco de dados pelos dados de um cliente;
- Buscar documentos baseado na palavra "Paraná".

Mas é:

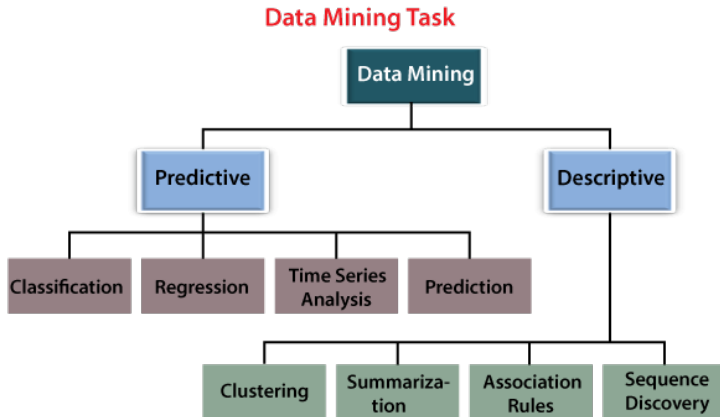
- Construir um modelo dos interesses do cliente para classificar itens como de seu interesse;
- Agrupar documentos com base em suas similaridades (praias do Paraná, futebol do Paraná, café do Paraná, etc).

Tarefas da Mineração de Dados

Tarefas Típicas:

- Agrupamento: ex. grupos de clientes com padrões de compra comum;
- Classificação/Categorização: ex. classificação biométrica de usuário;
- Regressão: ex. predição de demanda de um produto baseado em gasto com propaganda;
- Regras de Associação/Dependência: ex. quem compra A compra também B.

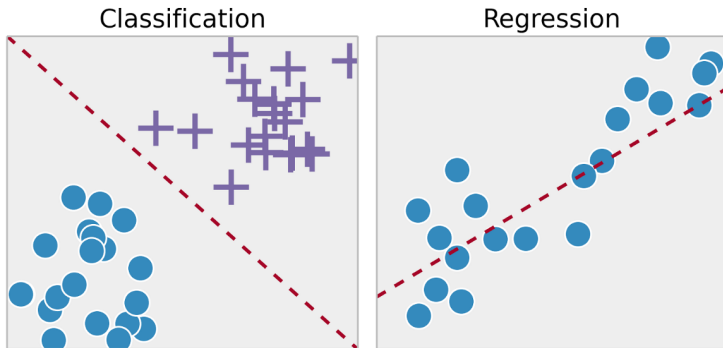
<https://www.tutorialandexample.com/data-mining-tasks>



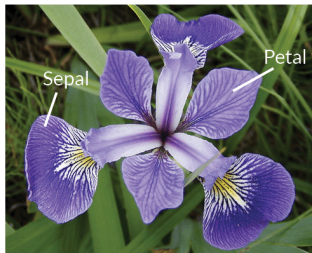
- Esse é o modelo mais conhecido, pois ajuda a prever cenários futuros com base na análise de padrões da base de dados. Assim, é possível tomar decisões mais precisas;
- Os métodos usados pela análise preditiva são dados estatísticos e históricos;
- Ela é indicada para projetar comportamentos futuros do público e do mercado, além de avaliar flutuações da economia e tendências de consumo.

- Nesse tipo de análise, os dados são todos resumidos, organizados e descritos por meio de métricas de estatísticas;
- Depois de definidos os aspectos mais importantes de um grupo de características, os dados são então relacionados entre dois ou mais conjuntos;
- Em geral, as ferramentas utilizadas em um estudo descritivo são as tabelas e os gráficos, mas existem outras formas de serem captadas: por meio de porcentagens, médias e índices;

Modelos Preditivos: Classificação e Regressão



Exemplo de Classificação - Base de dados Iris



Iris Versicolor



Iris Setosa



Iris Virginica

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

Modelos Descritivos: Agrupamento

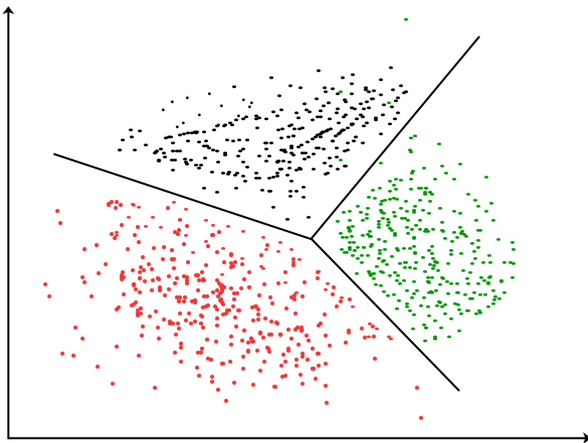
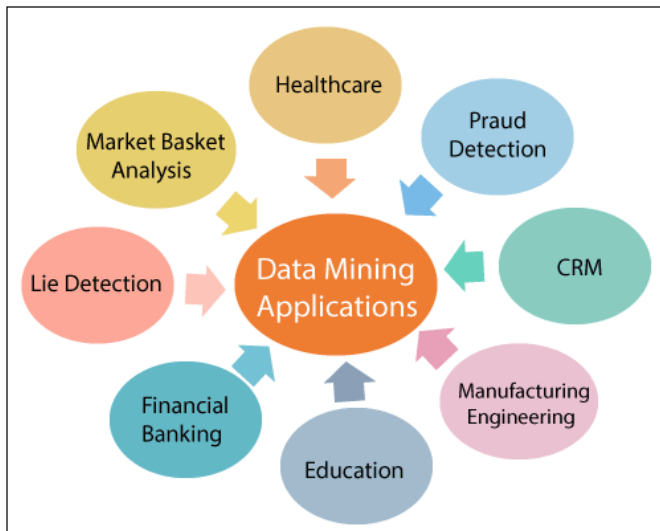


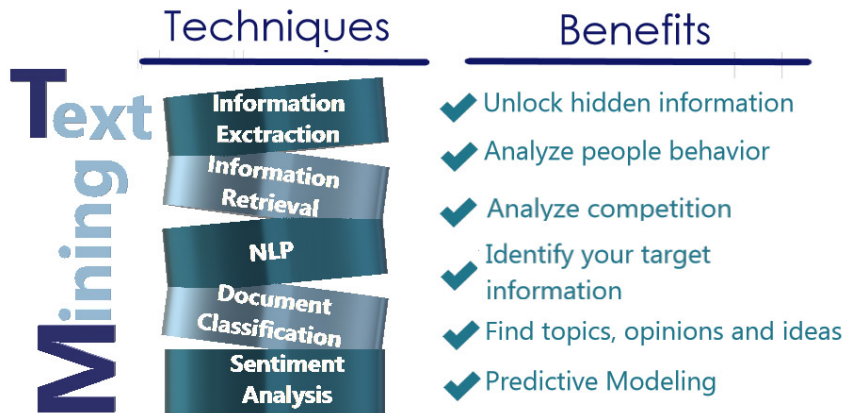
Figura: Exemplo de agrupamento de dados.

Exemplo de Aplicações





<https://www.inesisgroup.com/methods/text-mining/>




designed by INESIS GROUP - www.inesisgroup.com



<https://blog.superannotate.com/where-to-get-public-datasets/>

kaggle

Google
Dataset Search Beta

 VisualData



UC Irvine
Machine Learning
Repository



 OpenML

DATA
HUB



HealthData.gov

The NLP Index

- [1] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10 (7):1895–1923, 1998.
- [2] J. Gama, K. Faceli, A. Lorena, and A. De Carvalho. *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen - LTC, 2011. ISBN 9788521618805. URL <https://books.google.com.br/books?id=4DwelAEACAAJ>.