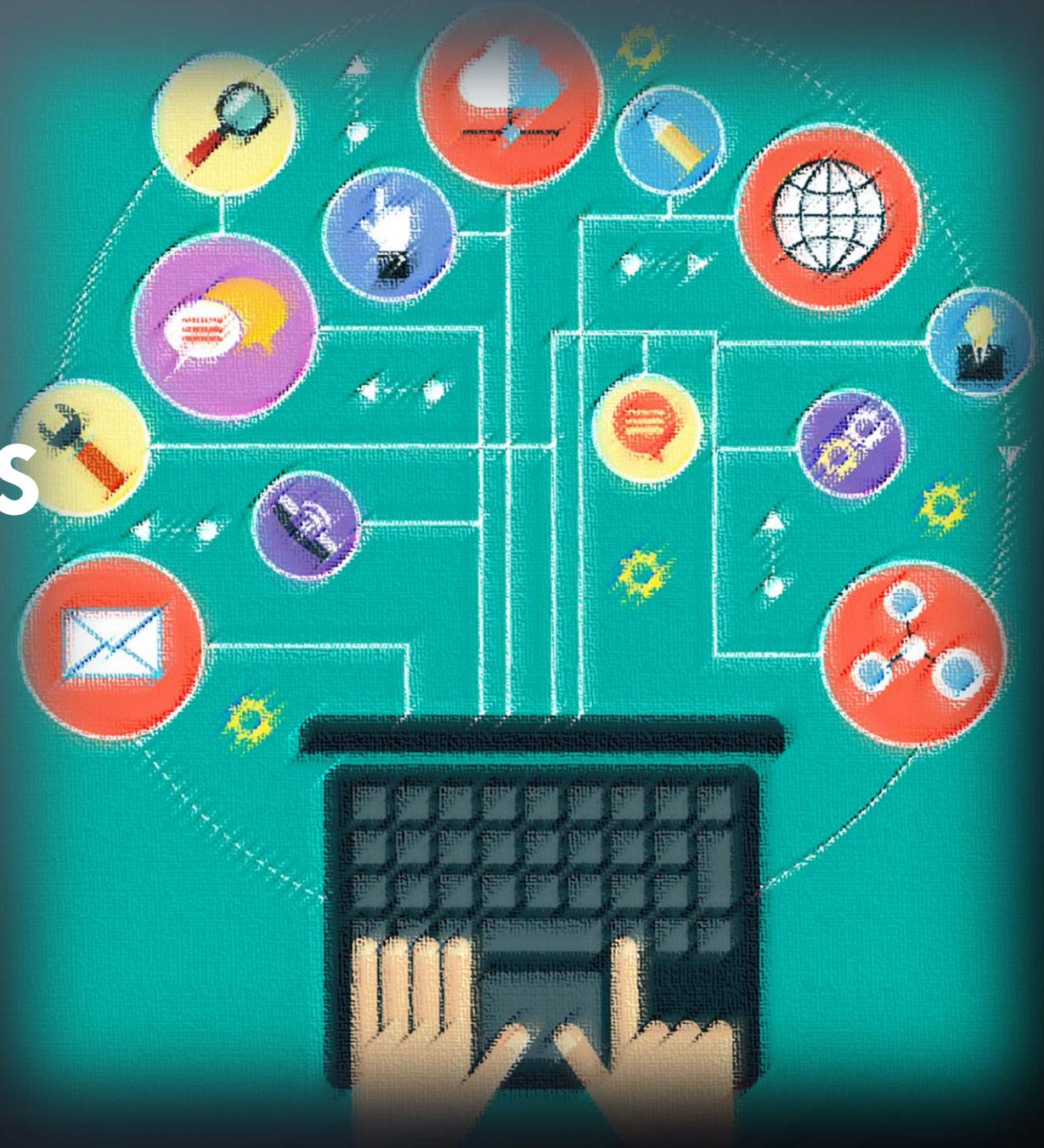


# CIÊNCIA DE DADOS EDUCACIONAIS

PRINCIPAIS CONCEITOS



# CIÊNCIA DE DADOS EDUCACIONAIS

TEMA: PRÉ-PROCESSAMENTO,  
MINERAÇÃO DE DADOS E  
VISUALIZAÇÃO DE DADOS





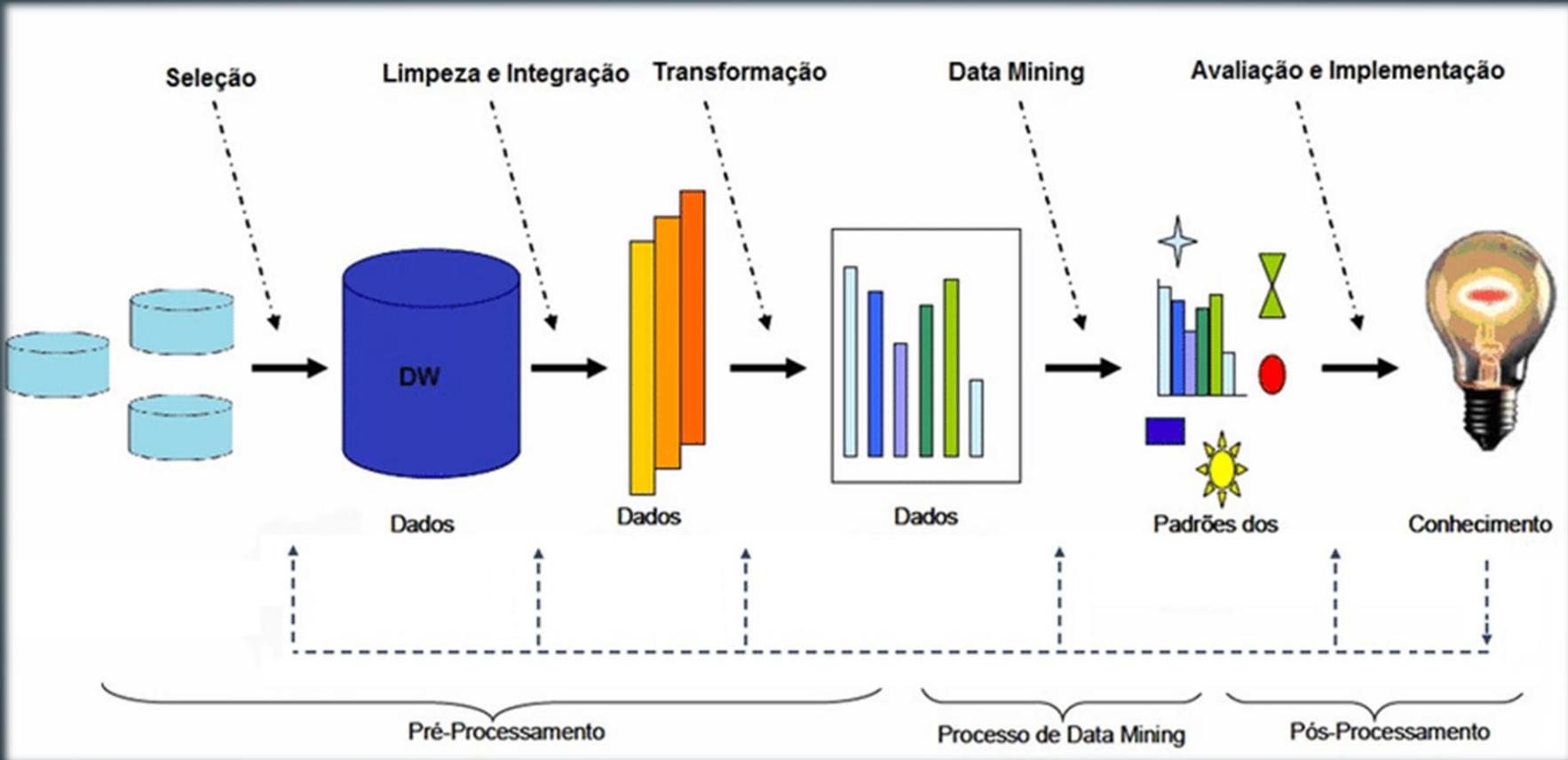
## DESCOBERTA DO CONHECIMENTO

Conhecido pela sigla (KDD) (Knowledge Discovery in Databases) constitui no processo que envolve as etapas de **pré-processamento, mineração e pós-processamento** dos dados no sentido de extrair conhecimento útil de grandes concentrações de dados.

# DESCOBERTA DO CONHECIMENTO

Trata-se de um processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis implícitos nos dados.





O processo CRISP-DM (Cross-Industry Standard Process for Data Mining) é composto por seis fases.

# PROCESSO CRISP-DM

5. AVALIAÇÃO

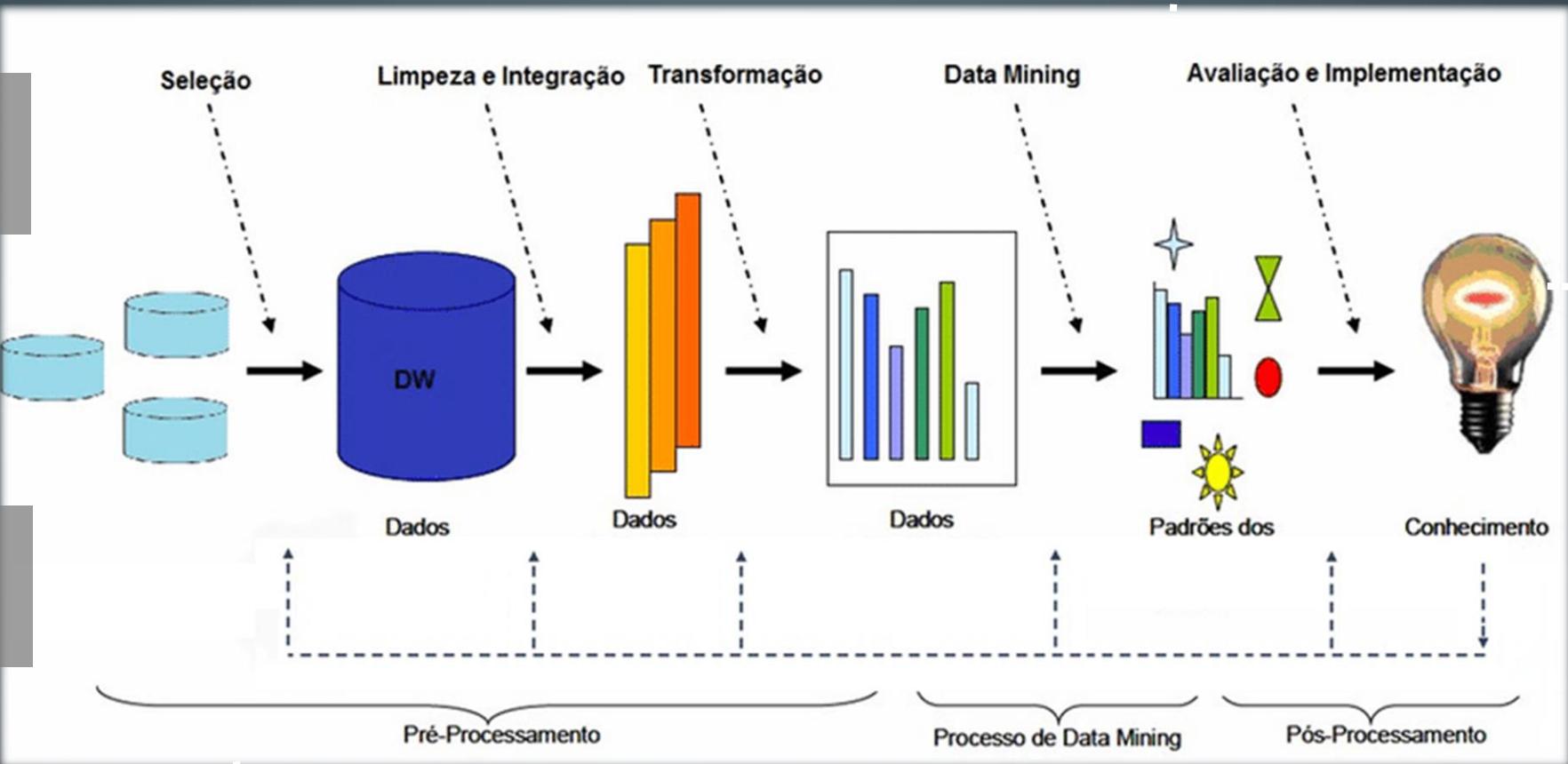
1.  
ENTENDIMENTO  
DO NEGÓCIO

2.  
ENTENDIMENTO  
DOS DADOS

3. PREPARAÇÃO DOS DADOS

4. MODELAGEM

6.  
DIPONIBILIZAÇÃO



# ENTENDIMENTO DO NÉGOCIO

Levantamento das problemáticas  
educacionais da instituição

Entrevistas com especialistas,  
aplicações de questionários

Avaliação das situações e  
recursos, determinação dos  
objetivos do MD em relação  
ao domínio e criação de um  
plano de projeto.

# ENTENDIMENTO DOS DADOS



- ✓ Coleta dos dados iniciais
- ✓ Realização da descrição dos dados
- ✓ Exploração e análise de qualidade

# PREPARAÇÃO DOS DADOS

Seleção de dados

Limpeza

Etapas de pré-processamento

Construção

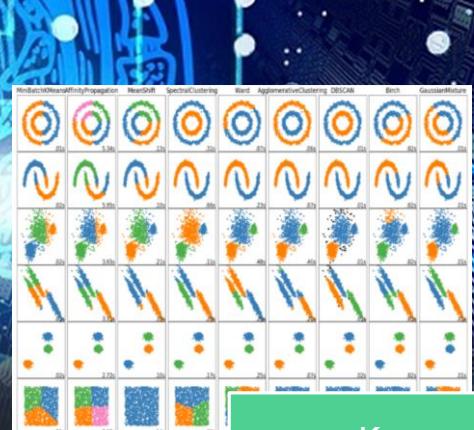
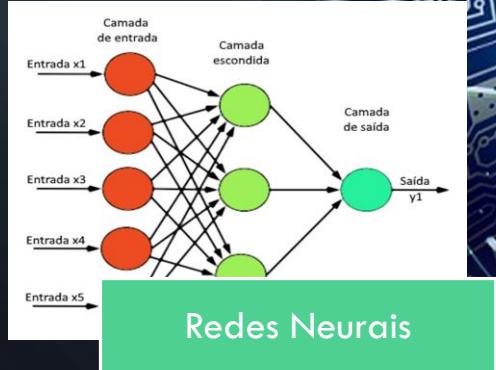
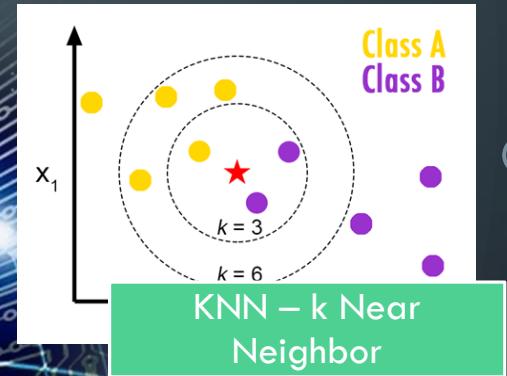
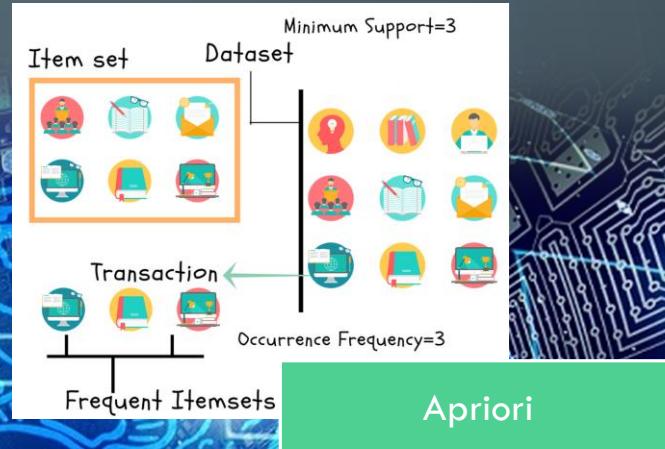
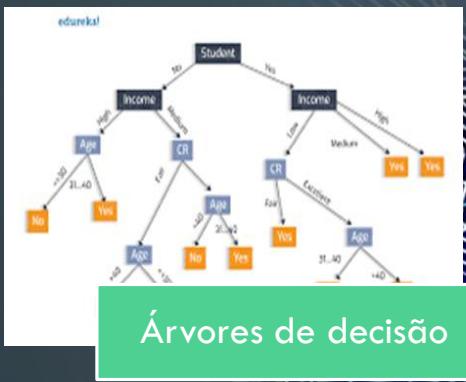
Integração

# DATA CLEAN

Formatação dos  
dados



# Seleção de técnicas de modelagem



K-means



Geração de testes para construção de um modelo

# Avaliação

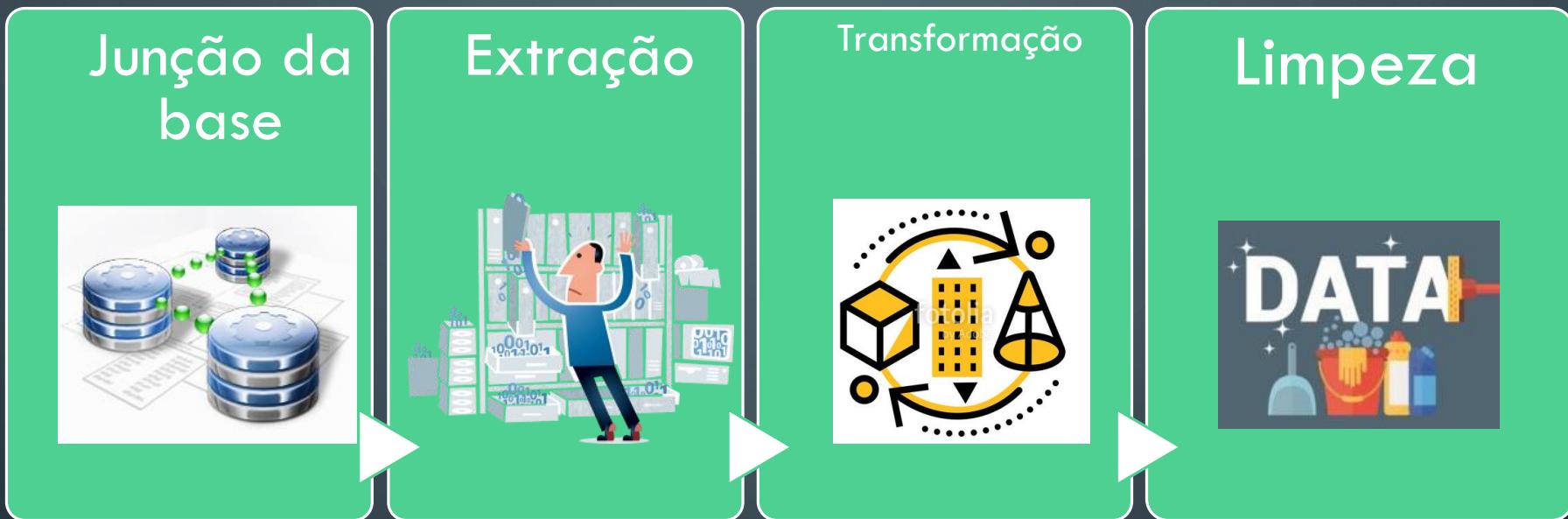
Avaliação dos resultados e  
revisão dos procedimentos

# Disponibilização dos dados



# PRÉ- PROCESSAMENTO

É feita a junção das bases de dados, extração, transformação e limpeza dos dados para a construção da tabela de análise final



Melhorar a qualidade dos dados por meio da eliminação ou minimização dos seguintes problemas: Ruído, imperfeições com valores incorretos, inconsistentes, duplicados ou ausentes.



Levar a construção de um modelo de dados mais fiéis à distribuição real dos dados, reduzindo sua complexidade computacional.

# PRÉ- PROCESSAMENTO

São utilizadas técnicas de amostragem de dados, tratamento dos dados desbalanceados, modificações para adequação dos tipos de atributos, limpeza, integração, transformação dos dados e redução de dimensionalidade.

# AMOSTRAGEM DE DADOS

Busca-se um conjunto representativo do dados originais



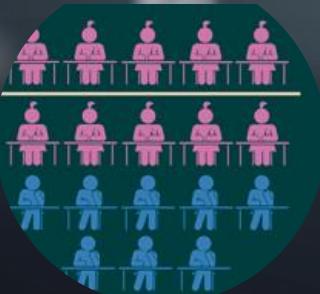
Busca-se uma eficiência computacional e de uma boa taxa de acurácia

Existem três técnicas de amostragem: A aleatória simples, a estratificada e a progressiva



Com ou sem  
repetição

Manter a proporção  
das classes



Começa pequeno e vai  
aumentando progressivamente



# DADOS DESBALANCEADOS

Farovericimento da classe majoritária

As técnicas são: Redefinir o tamanho dos conjuntos dos dados, utilizar diferentes custos de classificação para as classes diferentes e induzir um modelo para uma classe

✗ Overfitting

✗ Underfitting



Essa técnica tem um baixo desempenho quando boa parte dos objetos da classe majoritária são semelhantes



É induzir o modelo para um classe minoritária ou majoritária

# LIMPEZA DOS DADOS



Dados ruidosos



Dados inconsistentes



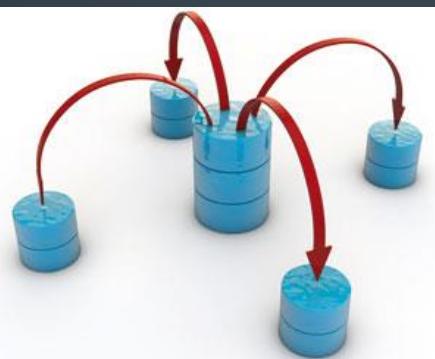
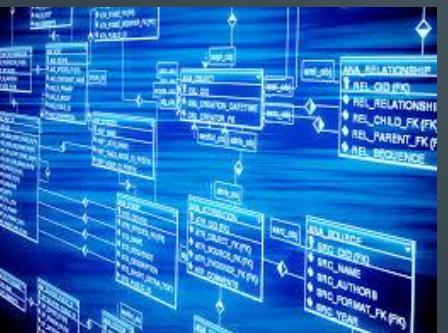
Dados redundantes



Dados incompletos

# INTEGRAÇÃO DE BASES DE DADOS

Os dados dos arquivos de registro da Web podem ser integrados a outros dados para permitir uma maior abrangência e / ou análise significativa.



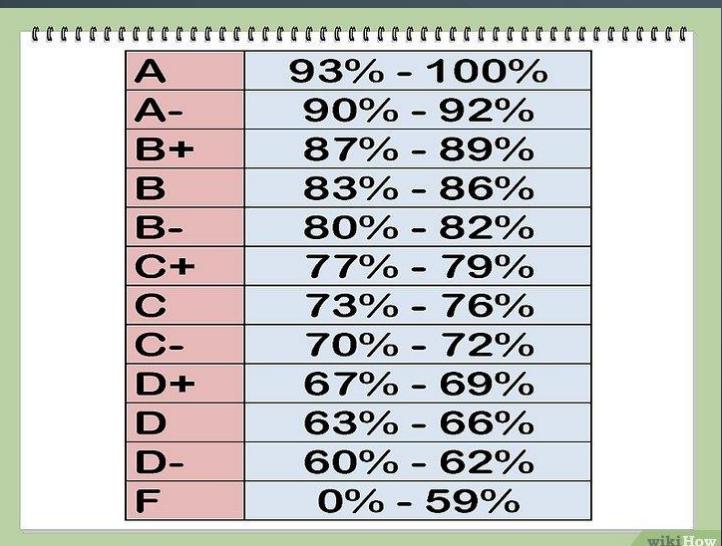
Informações sobre os alunos: Idade, sexo, Experiência na Web, Desempenho do curso e Estilo aprendizagem



# TRANSFORMAÇÃO DOS DADOS

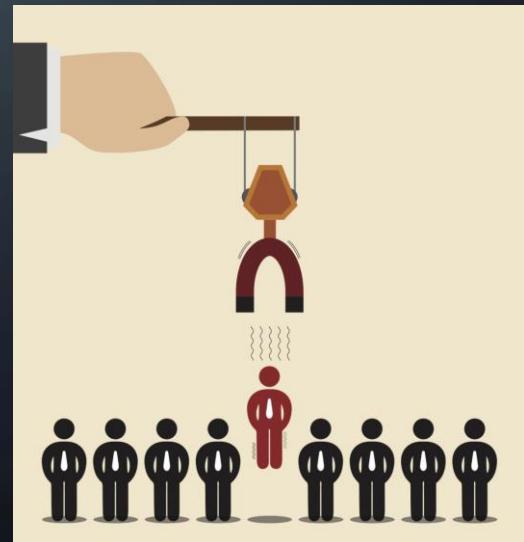
Dois processos que podem ser realizados como parte da transformação de dados.

## Discretização de dados



<b>A</b>	<b>93% - 100%</b>
<b>A-</b>	<b>90% - 92%</b>
<b>B+</b>	<b>87% - 89%</b>
<b>B</b>	<b>83% - 86%</b>
<b>B-</b>	<b>80% - 82%</b>
<b>C+</b>	<b>77% - 79%</b>
<b>C</b>	<b>73% - 76%</b>
<b>C-</b>	<b>70% - 72%</b>
<b>D+</b>	<b>67% - 69%</b>
<b>D</b>	<b>63% - 66%</b>
<b>D-</b>	<b>60% - 62%</b>
<b>F</b>	<b>0% - 59%</b>

## Seleção de recursos



# REDUÇÃO DE DIMENSIONALIDADE

O objetivo é melhorar o desempenho do modelo induzido, reduzindo custo computacional e tornar os resultados mais compreensíveis



Agregação

Análise de Componentes principais

Correlaciona estatisticamente os exemplos

Perda de informação.  
Dificulta a compreensão dos resultados

Seleção de atributos



$$A = \{ \text{black, orange, yellow, green, blue, pink, red} \}$$

Procura um sub conjunto ótimo de atributos

Técnicas: Embutidas,  
Baseada em filtros e  
Baseadas em wrapper

## MINERAÇÃO DE DADOS

Focado nos métodos de análises, tais como: classificação, regressão, agrupamento e análise de associação.

A tarefa de mineração de dados é extrair  
algum tipo de conhecimento da base de dados



As tarefas podem ser definidas como : Modelo Preditivo, Análise  
de Agrupamento e Regras de Associação.

# MODELOS PREDITIVOS

- É a tentativa de prever o que acontecerá no futuro.
- Essa tentativa se dá através da construção de um modelo.
- Esse modelo é construído utilizando a base de dados.
- Esse processo de construção recebe o nome de aprendizado.

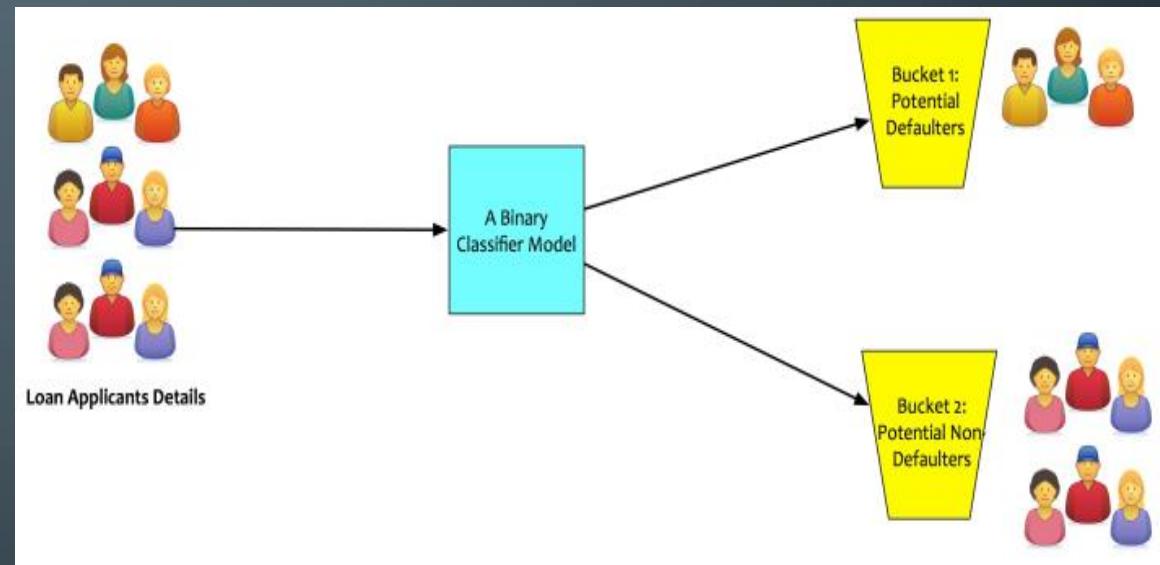


- Este aprendizado é chamado de supervisionado.
- É uma função que permite mapear um conjunto de atributos em um dos valores de atributo especial.
- O modelo é capaz de inferir novas situações que não estavam armazenadas na base de dados.

Dependendo do tipo do atributo especial, chamamos o modelo de preditivo de classificação ou previsão.

# PREDITIVO - CLASSIFICAÇÃO

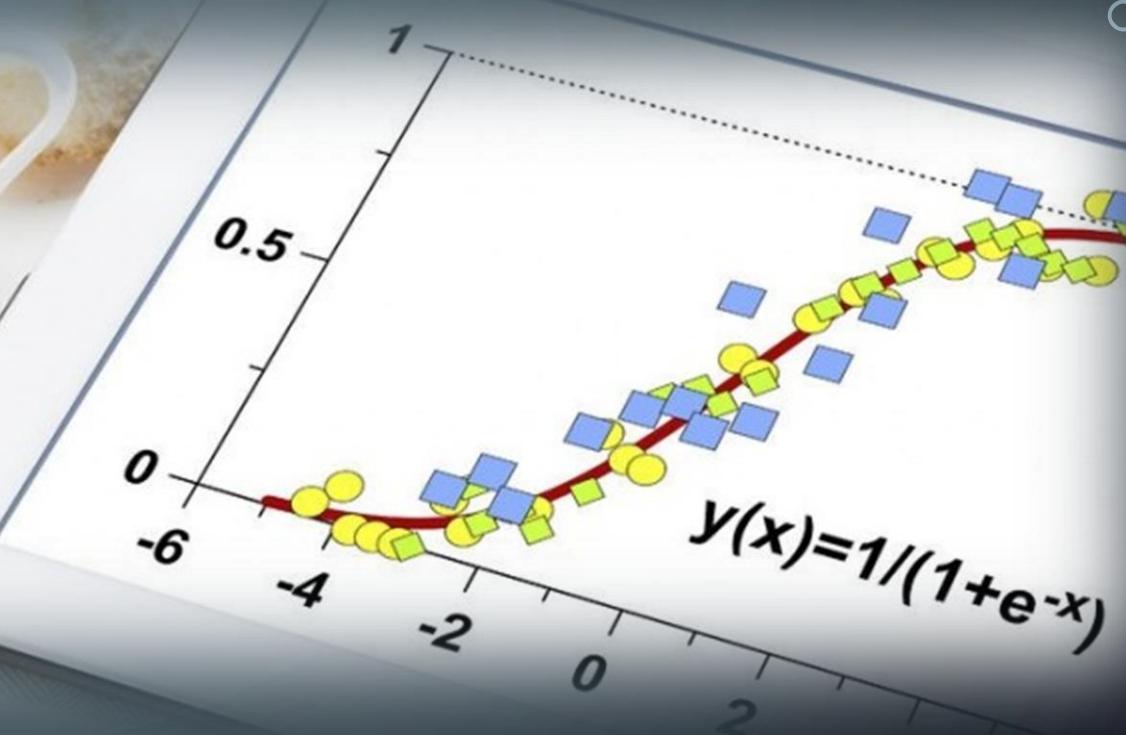
- É quando o atributo especial (Rótulo) da base de dados de treinamento é formado por uma categoria.
- As classificações podem ser binárias ou multiclasse.



No contexto da educação, um processo de classificação poderia ser feito para prever se o aluno merece ou não uma bolsa de estudos, se ele está apto para participar de um intercâmbio, se ele se formará no prazo ou não.

# PREDITIVO - REGRESSÃO

- O atributo especial (rótulo) é contínuo.
- Temos um única entrada com valor real e uma única saída com valor real.
- O objetivo é encontrar uma função matemática que modele os dados

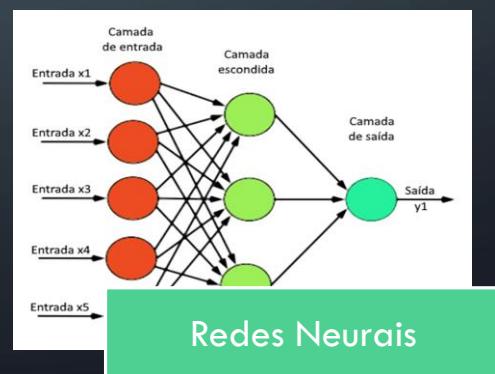
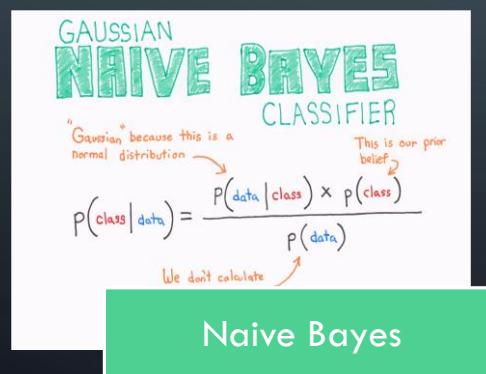
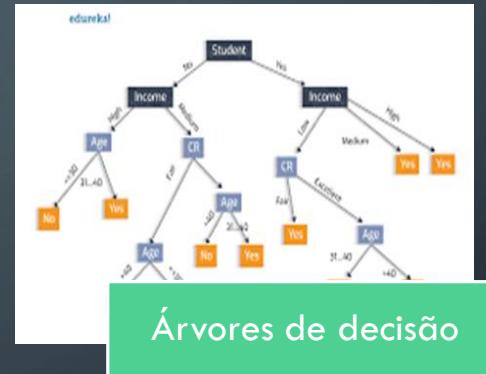
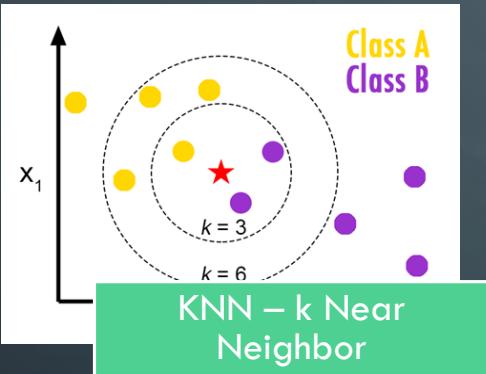


- Tem um hiperplano como superfície de decisão.
- Os modelos podem ser lineares ou não.

No contexto da educação, a previsão poderia ser feita para inferir a nota do aluno em uma disciplina, a nota média do aluno no final do curso e a quantidade de faltas no final do semestre.

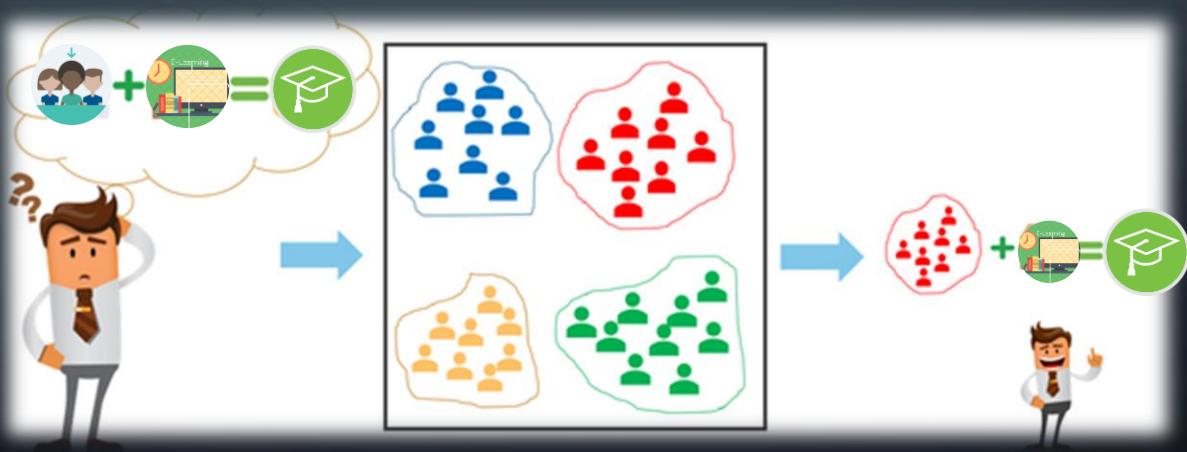
# PREDITIVO – CLASSIFICAÇÃO OU REGRESSÃO

Os algoritmos de classificação e previsão são divididos em aprendizado de máquina e inteligência computacional.

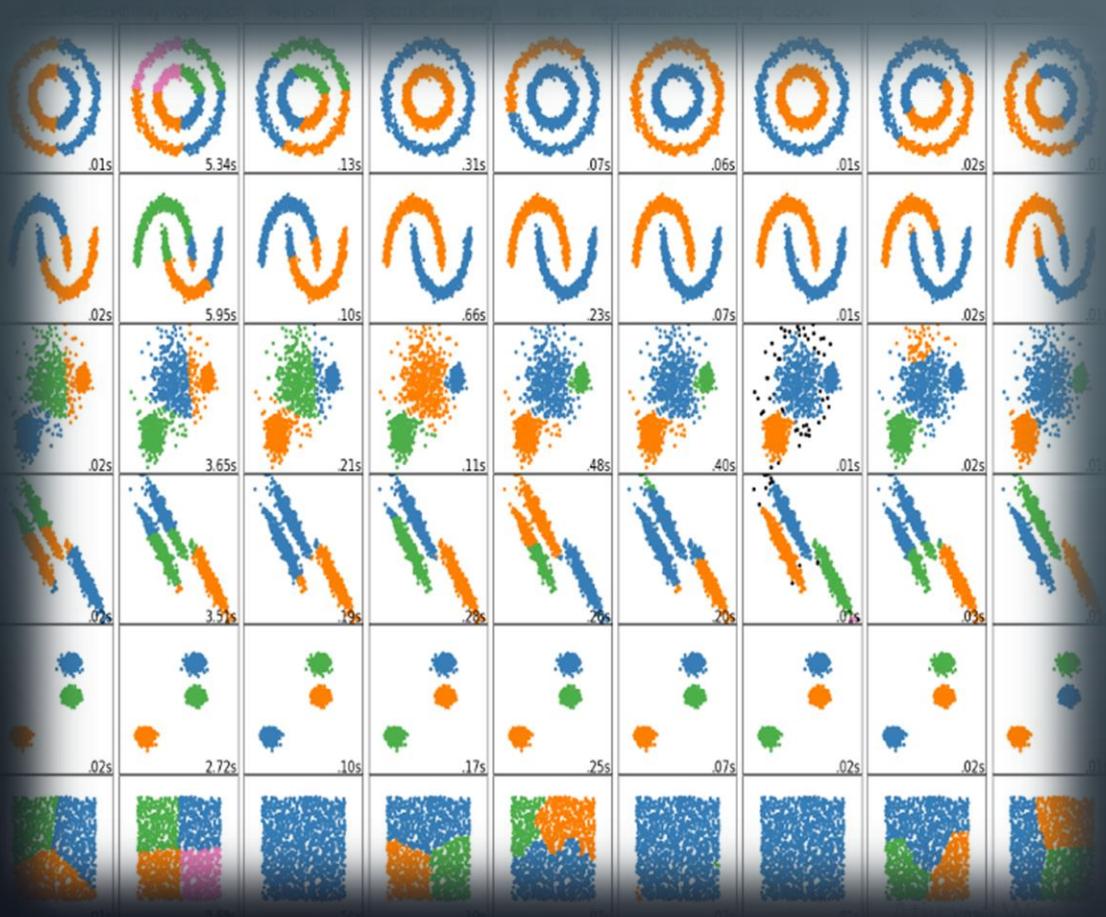


# ANÁLISE DE AGRUPAMENTO

- É a descoberta de grupos ou clusters.
- Encontrar padrões com atributos semelhantes na base de dados
- A segmentação é feita por medidas de similaridade
- É um aprendizado não supervisionado
- Na educação podemos descobrir estilos de aprendizado dos alunos e disciplinas de interesses em comum.
- Não se encontra disponível o atributo especial (rótulo)



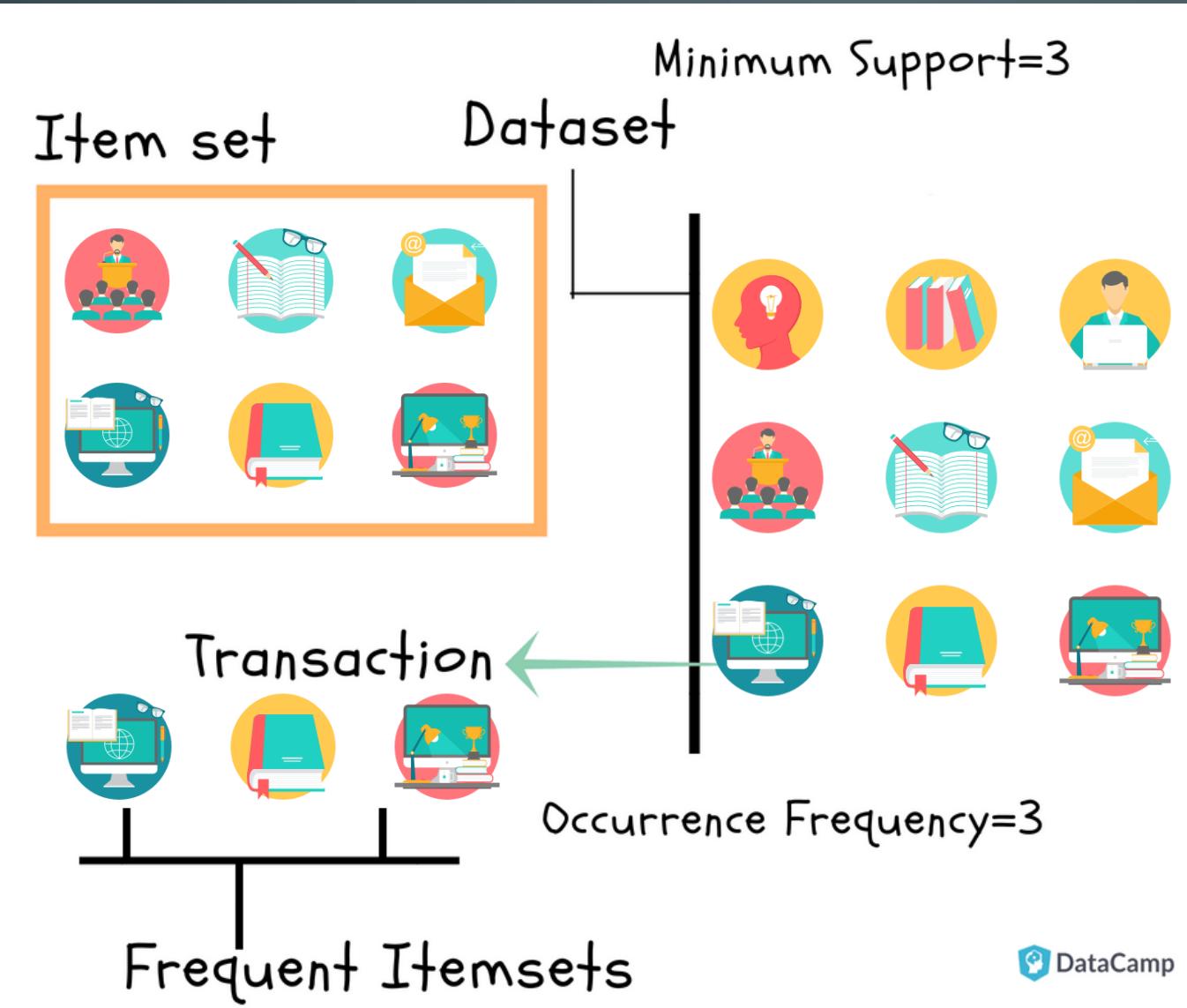
# ANÁLISE DE AGRUPAMENTO



Os algoritmos são:

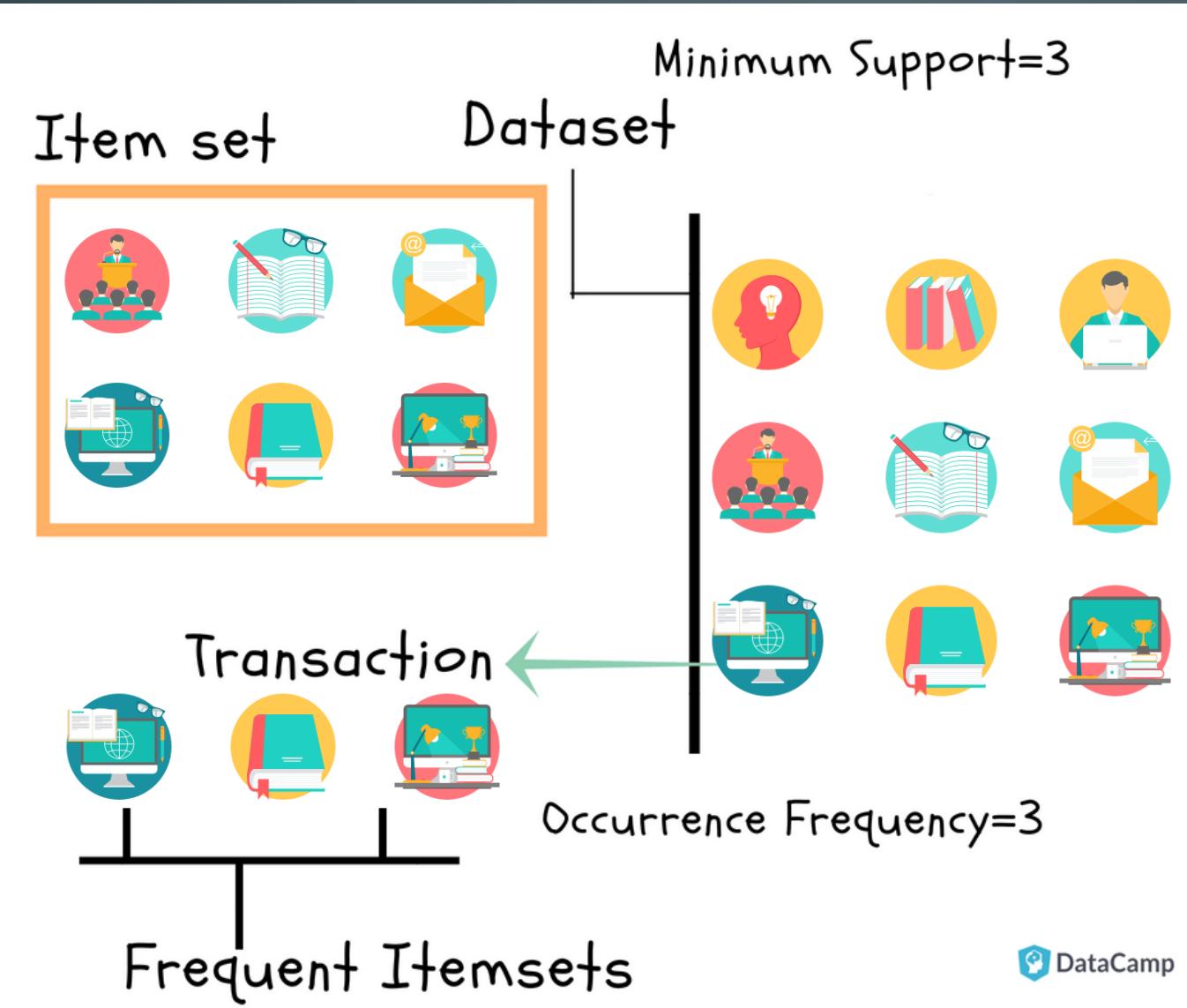
1. Agrupamento hierárquico,
2. k-médias (k-means)
3. Agrupamento espacial baseado em densidade
4. Redes Neurais com Mapas Auto-Organizáveis ou SOM

# REGRAS DE ASSOCIAÇÃO



- É derivar regras de conhecimento, referindo-se a relacionamentos entre objetos de um conjunto de dados, visando características e tendências.
- Os algoritmos não fazem parte de nenhum tipo de aprendizado, apenas são metodologias para geração de regras

# REGRAS DE ASSOCIAÇÃO



- ✓ Formalmente pode definir como uma associação entre itens do tipo “uma transação que contém os itens X também possui o conjunto de itens Y” ( $X \Rightarrow Y$ ) , onde  $X \subseteq I$  e  $Y \subseteq I$  e  $X \cap Y = \emptyset$ ,  $I = \{1,2,...,m\}$  um conjunto de literais, chamados de itens.
- ✓ Assim, a regra tem a forma ‘Se X, então Y, onde X é denominado de corpo da regra e Y de cabeça da regra.
- ✓ A cada regra derivada pelo algoritmo , verifica-se a sua validade e importância. Para isso, faz-se uso de duas medidas básicas: o suporte e a confiança, comparando-se respectivos limiares estabelecidos (suporte mínimo e confiança mínima)

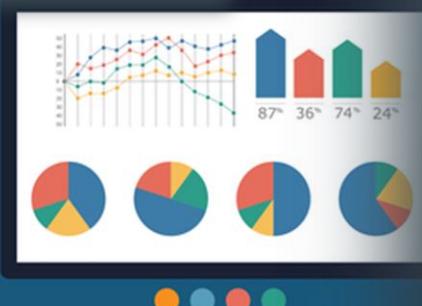
# REGRAS DE ASSOCIAÇÃO

Em mineração de dados educacionais é a mineração de regras em um banco de notas de alunos em disciplinas. Neste contexto seria possível derivar regras **como 90% dos alunos que têm bom desempenho nas disciplinas de Matemática e Lógica são bem sucedidos também em Programação”.**

Na literatura os dois algoritmos mais utilizados são o Apriori e o FP-Growth. A maioria desses algoritmos exige que o usuário defina dois limites , o suporte mínimo e a confiança mínima e encontre todas as regras que excedem os limites especificados pelo usuário. Esses algoritmos derivam regras apenas conjuntivas, limitando-se a utilização do operador lógico AND.



## VISUALIZAÇÃO DE DADOS



É o uso de representações visuais, interativas e sustentadas por computador, de dados abstratos para amplificar a cognição”

**É o uso de representações visuais, interativas e sustentadas por computador, de dados abstratos para amplificar a cognição”**

Por essa definição, quatro termos são a chave para entender este domínio: representação visual, interação, abstração dos dados e amplificação cognitiva.

#### Representação visual

Visualização é uma atividade cognitiva, facilitada por representações externas gráficas de que as pessoas constroem a representação mental interna do mundo

#### Interação

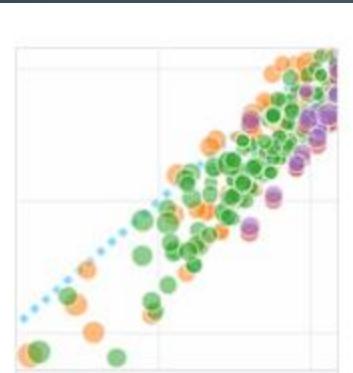
Quando a exibição visual muda e permite que eles manipulem a visualização ou os dados subjacentes para explorar essas mudanças

#### Abstração dos dados

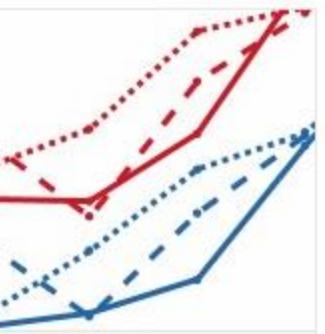
Está intimamente relacionada a estruturas e modelos matemáticos

#### Amplificação cognitiva

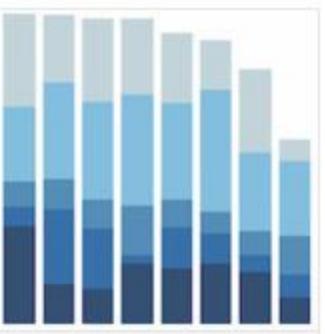
Usam as representações visuais que ajudam a codificar informações em um meio manipulável



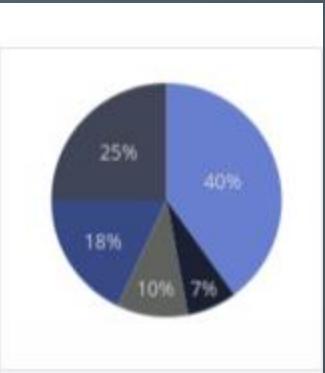
Scatter Plots



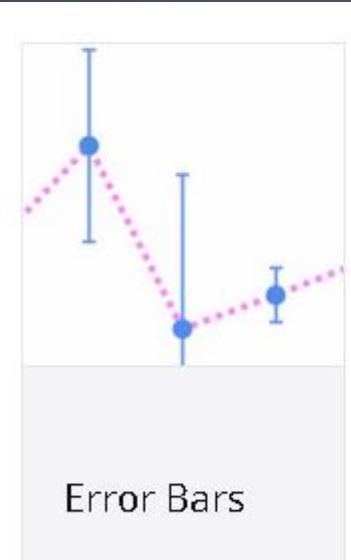
Line Charts



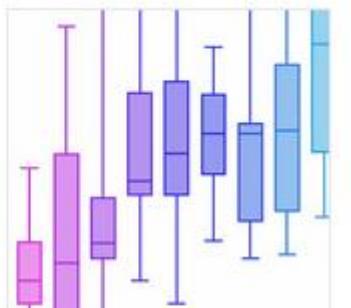
Bar Charts



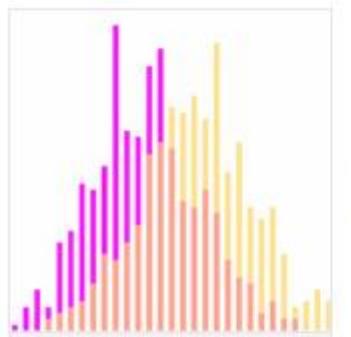
Pie Charts



Error Bars



Box Plots



Histograms



Distplots

# 10 RAZÕES

pelas quais deve investir na

## VISUALIZAÇÃO DE DADOS

Forma simples e rápida de transmitir e interpretar informações de negócio

1

Melhor entendimento do desempenho operacional e das atividades de negócio

2

Rápida identificação de tendências e áreas que necessitam de atenção

3

Análise de padrões e compreender o impacto das estratégias implementadas

4

Interação direta e costumizada com os dados e previsão de cenários

5

Interpretação significativa de um grande volume de dados

6

Otimização de processos e tomada de decisões baseadas em factos

7

Aumenta a produtividade e ROI

8

Agiliza processos e garante a assertividade da estratégia

9

Redução de risco e otimização de tempo e recursos

10

