



# Engenharia de Dados com Hadoop e Spark

---



# Bem-vindo(a)





# Hadoop Machine Learning com Apache Mahout



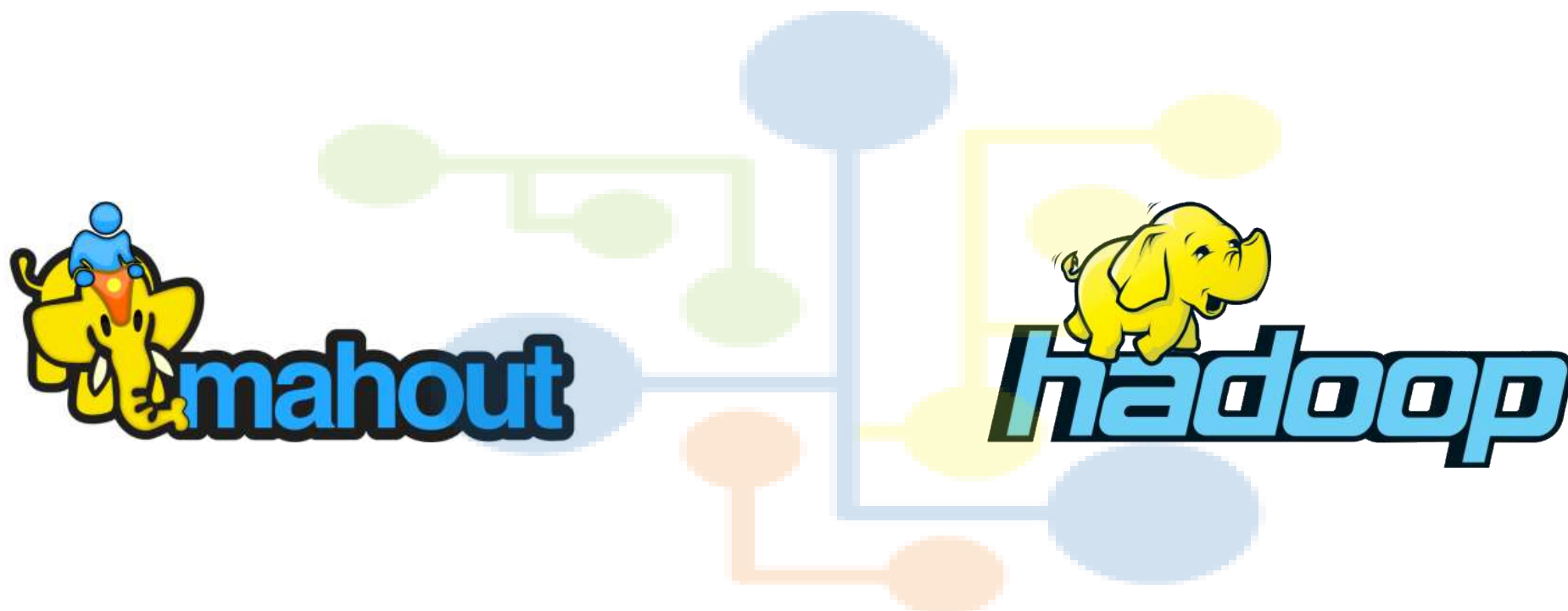
# Conhecendo o Apache Mahout

# Apache Mahout



Data Science  
Academy

Data Science Academy [rodrigo.c.abreu@hotmail.com](mailto:rodrigo.c.abreu@hotmail.com) 5e207d48e32fc335fa60447d



## Principais Algoritmos de Machine Learning disponíveis no Apache Mahout:

- Algoritmos de Classificação
- Sistemas de Recomendação
- Clustering
- Redução de Dimensionalidade



## Tipos de Algoritmos Suportados:

- Algoritmos Sequenciais
  - Regressão Logística
  - Modelos Ocultos de Markov
  - Perceptrons de Multi-camadas
- Algoritmos Paralelos
  - Random Forest
  - Naive Bayes
  - K-Means





# Apache Mahout x Outros Frameworks de Machine Learning

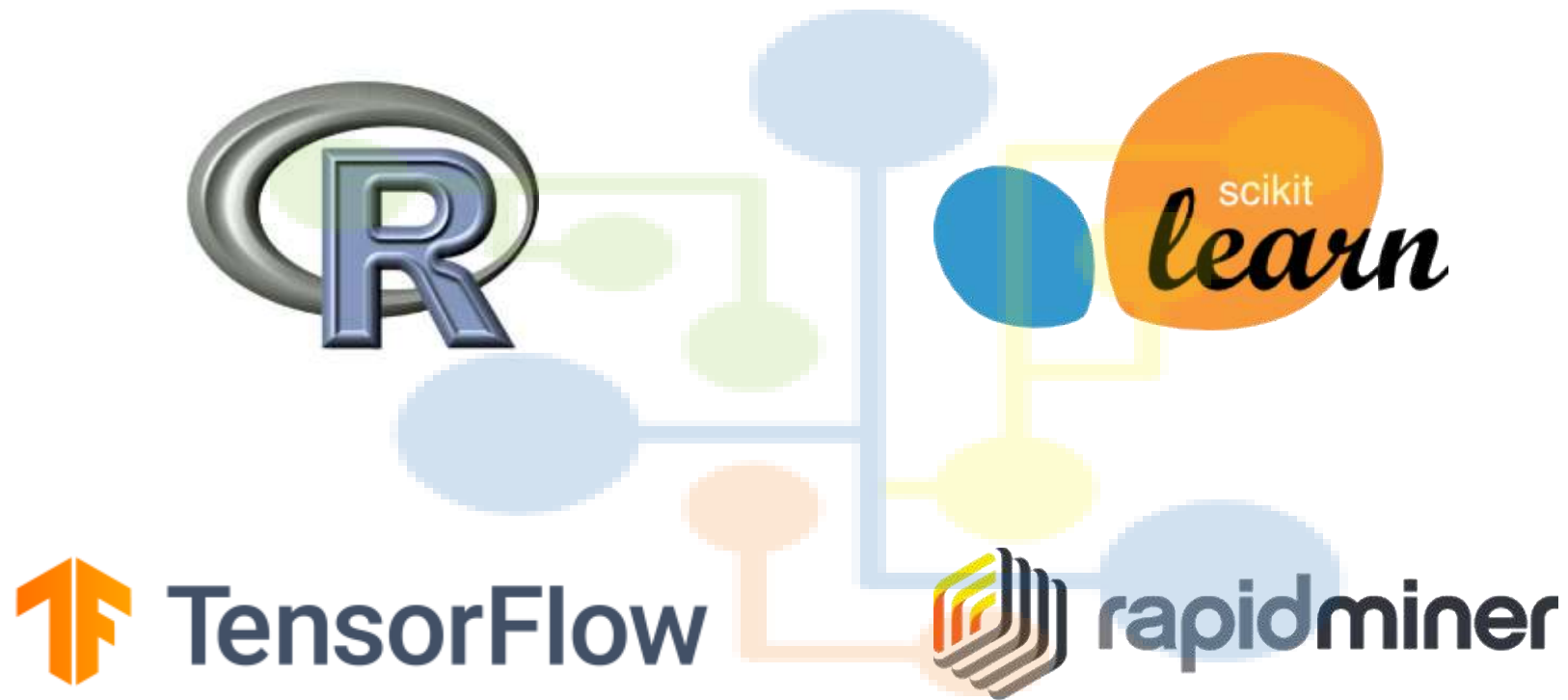




Data Science  
Academy

Data Science Academy [rodrigo.c.abreu@hotmail.com](mailto:rodrigo.c.abreu@hotmail.com) 5e207d48e32fc335fa60447d

# Apache Mahout





# Apache Mahout

Então por que vamos estudar o Apache Mahout, se existem frameworks de Machine Learning mais fáceis e mais completos?

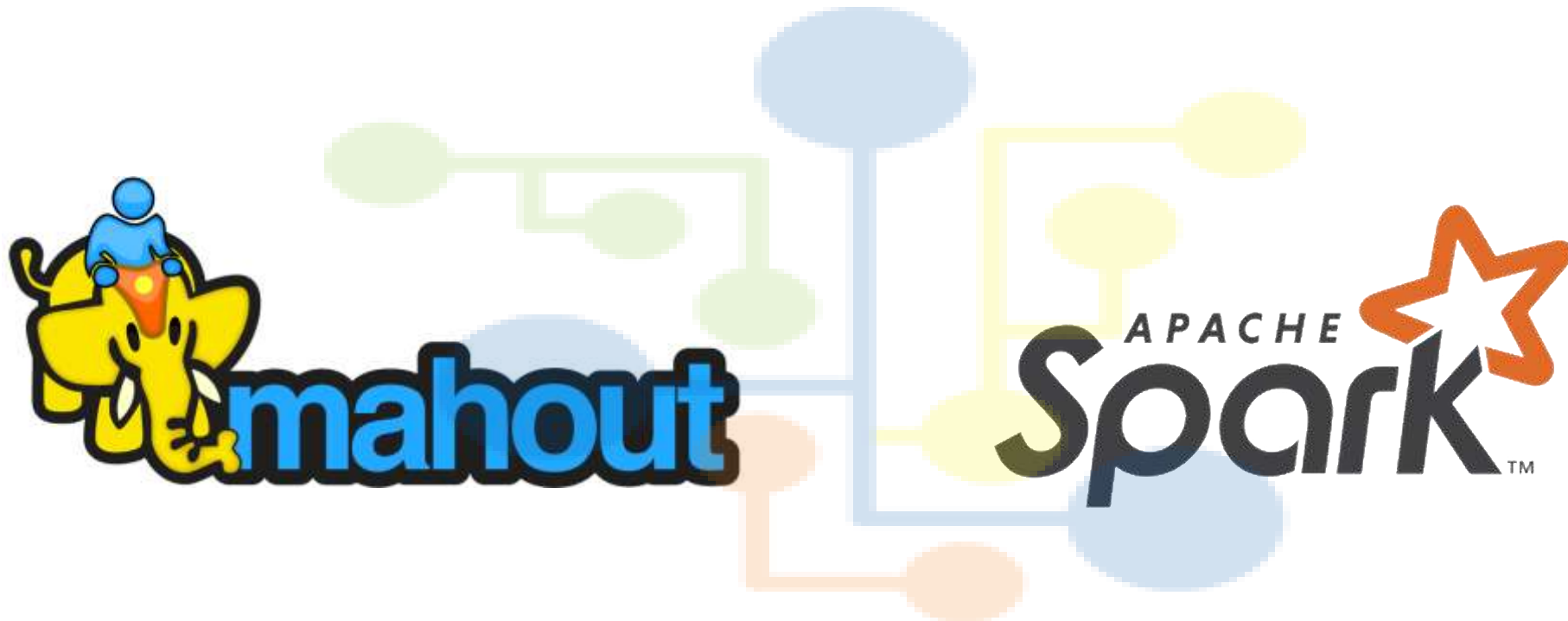


# Apache Mahout





# Apache Mahout



## Características do Apache Mahout



Os algoritmos do Mahout são escritos para funcionar sobre o Hadoop e dessa forma, eles funcionam em ambiente distribuído.



## Características do Apache Mahout

O Framework Mahout está pronto para uso e permite realizar mineração de dados em grandes conjuntos de dados.



# Apache Mahout

## Características do Apache Mahout

O Mahout é eficiente na análise de grandes conjuntos de dados.



## Características do Apache Mahout

Possui diversas implementações de Clustering, como:  
K-means, Fuzzy K-Means, Canopy e Mean-Shift.





## Características do Apache Mahout

Suporta a execução do algoritmo de classificação Naive Bayes de forma distribuída.



# Apache Mahout

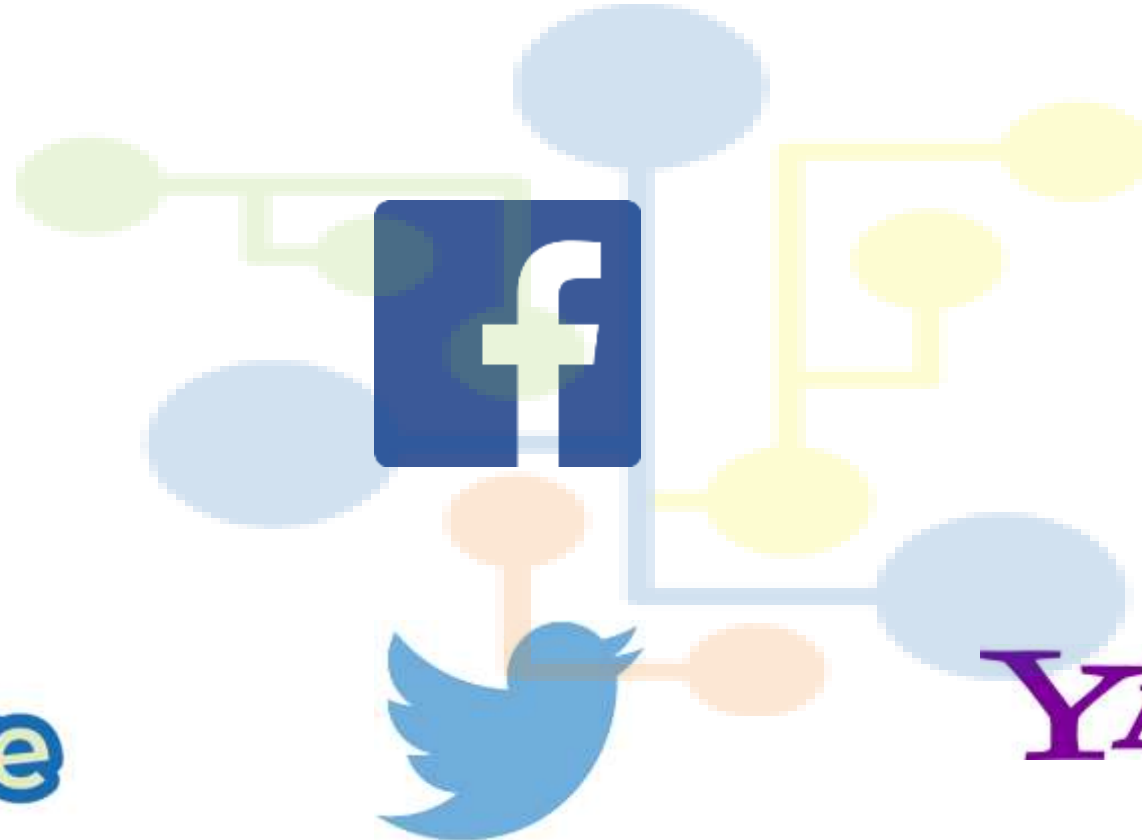
## Características do Apache Mahout

Inclui bibliotecas para manipulação de vetores e matrizes.



# Apache Mahout

Quem utiliza o Mahout

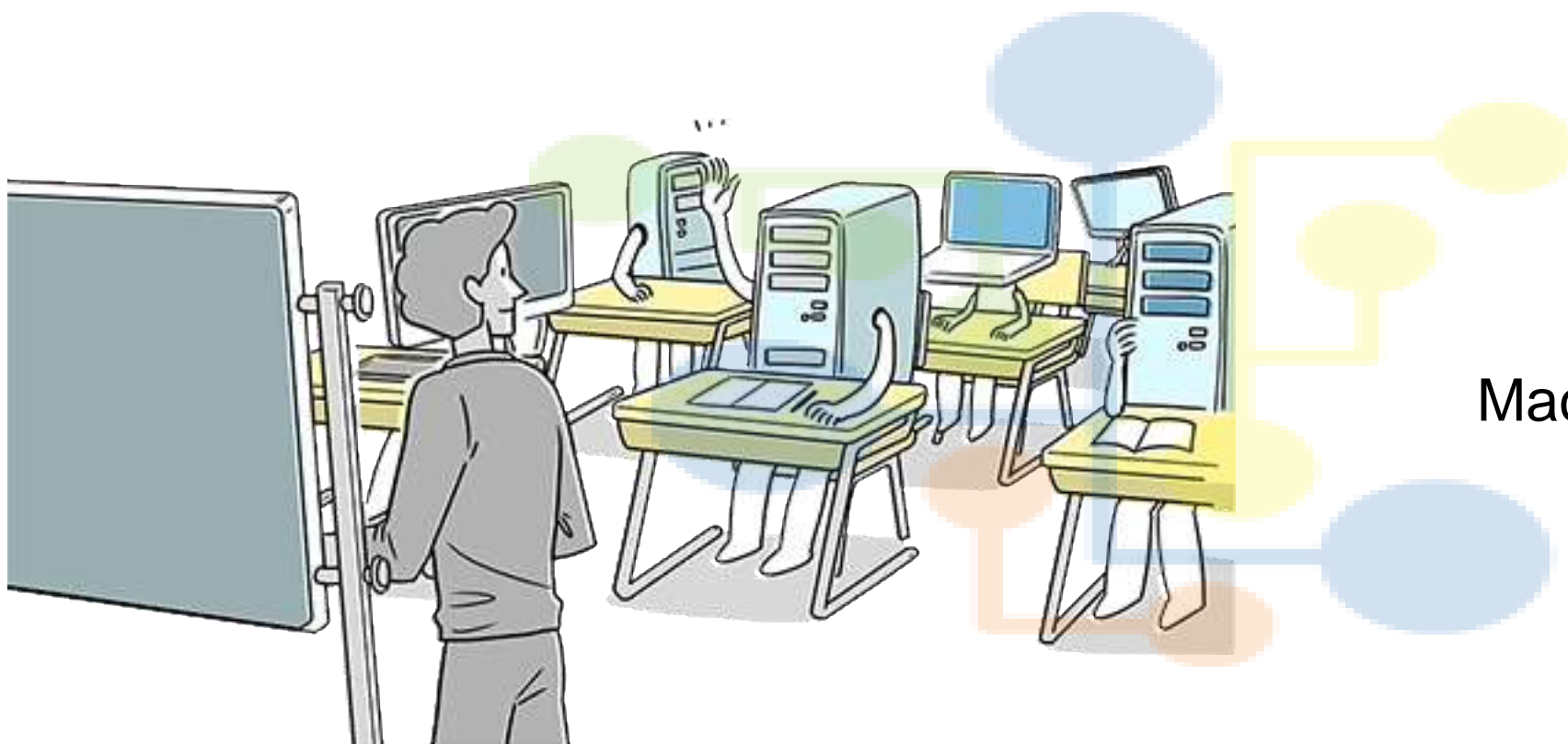




# Machine Learning - Algoritmos de Classificação



# Algoritmos de Classificação



Machine Learning



# Algoritmos de Classificação



Features  
(Atributos)



# Algoritmos de Classificação

A diagram showing two main categories of machine learning: Supervised Learning (Aprendizagem Supervisionada) in a green box on the left and Unsupervised Learning (Aprendizagem Não-Supervisionada) in a dark blue box on the right. They are connected by a central vertical line with several colored circles (blue, green, yellow, orange) and horizontal lines branching out to the left and right, suggesting a spectrum or relationship between the two types.

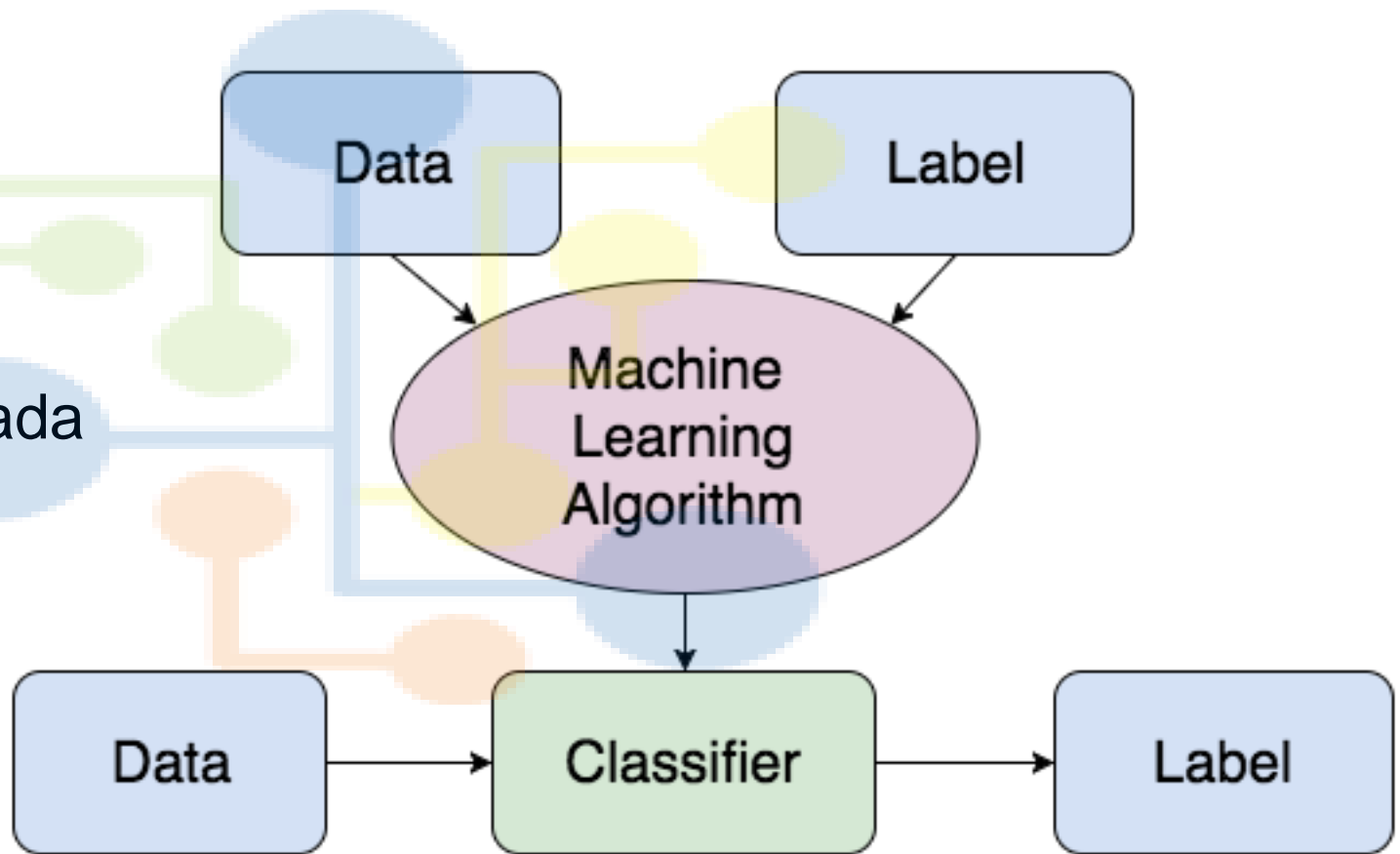
Aprendizagem  
Supervisionada

Aprendizagem  
Não-Supervisionada



# Algoritmos de Classificação

Classificação  
Aprendizagem Supervisionada

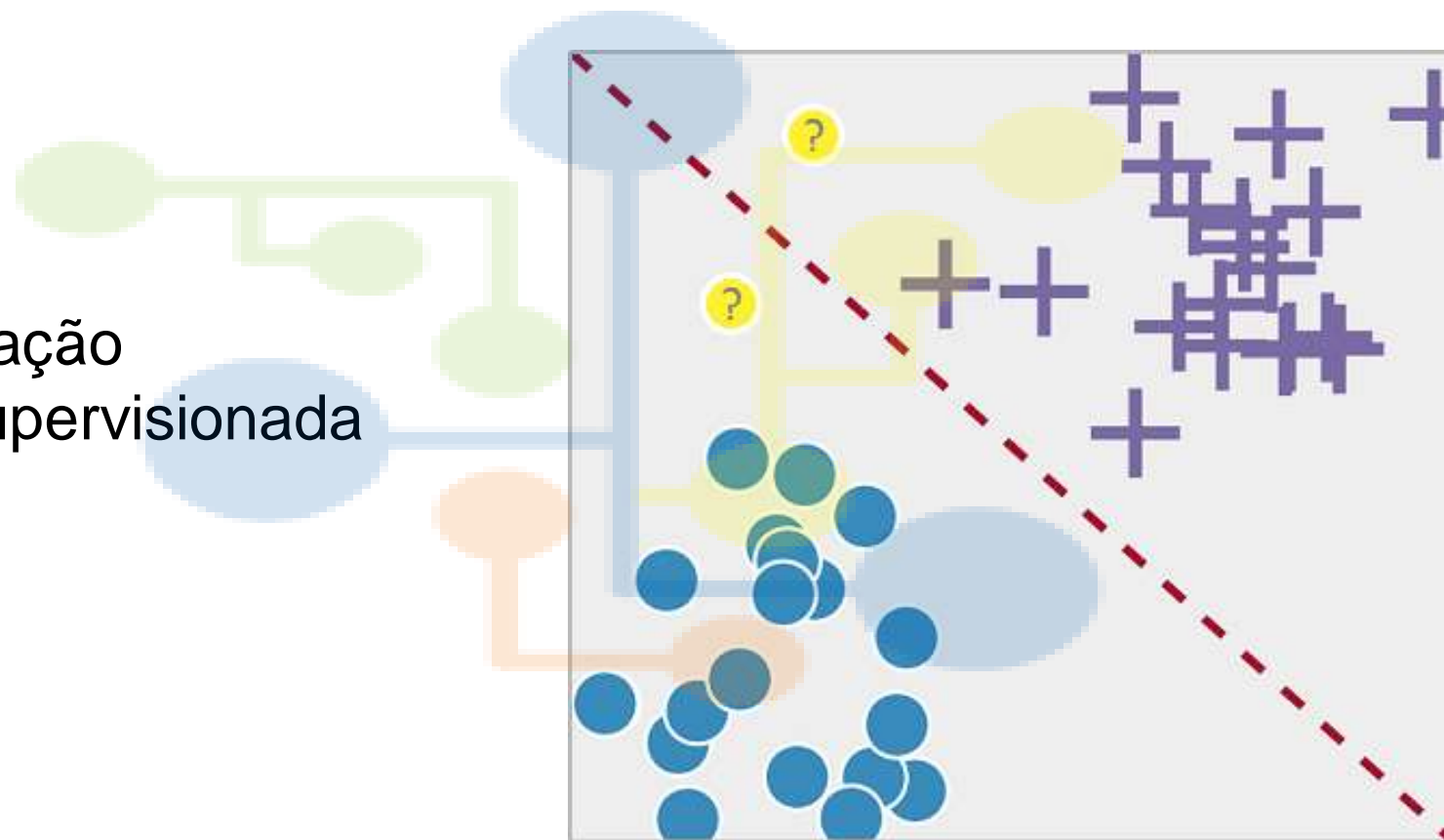






# Algoritmos de Classificação

Classificação  
Aprendizagem Supervisionada





Data Science  
Academy

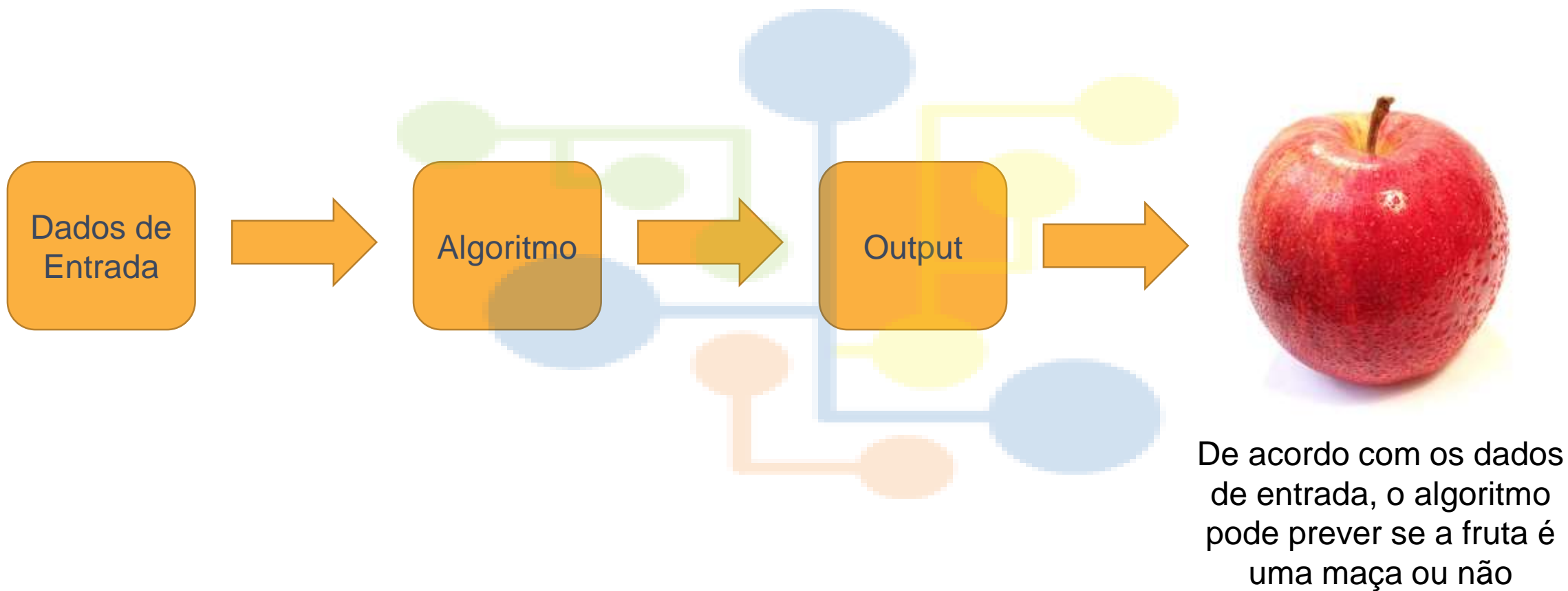
Data Science Academy [rodrigo.c.abreu@hotmail.com](mailto:rodrigo.c.abreu@hotmail.com) 5e207d48e32fc335fa60447d

# Algoritmos de Classificação



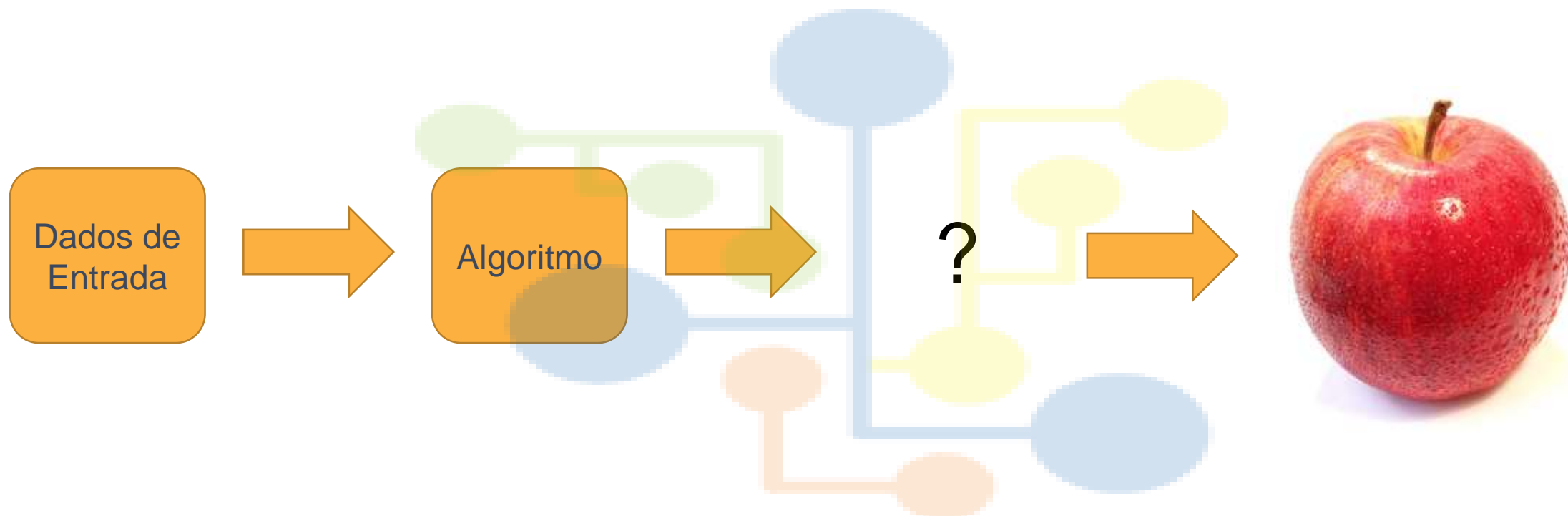


# Algoritmos de Classificação





# Algoritmos de Classificação





# Algoritmos de Classificação

Classificação

Categorização de E-mails

Filtros de Spam

Detecção de Fraudes

Eligibilidade de Clientes



# Algoritmos de Classificação

Variáveis Explonatórias

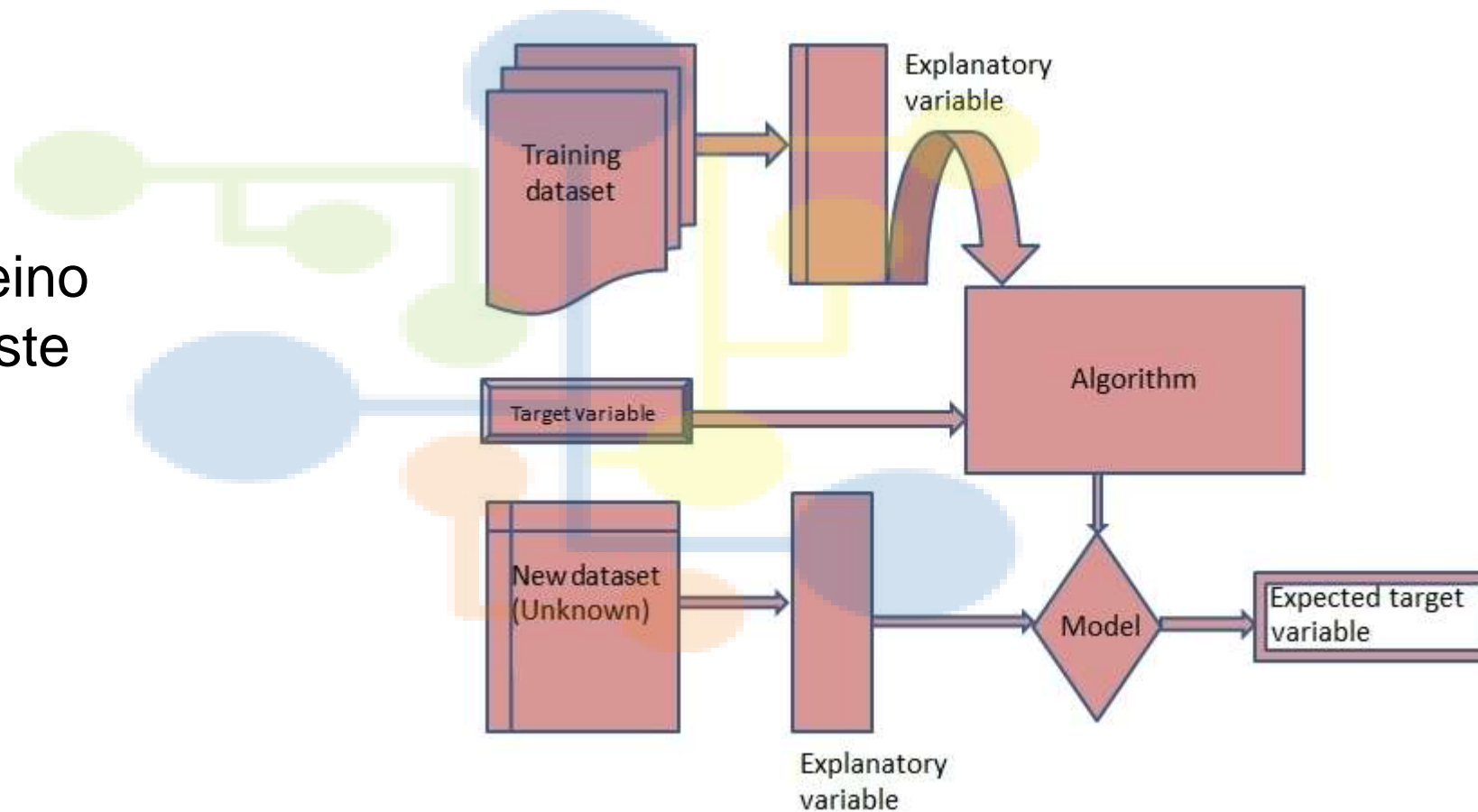
Variável Tatget  
(Class label)

Idade do Cliente	Renda do Cliente	Saldo Atual	Conceder Crédto
26	R\$ 5.000	R\$ 30.000	Sim
34	R\$ 9.500	R\$ 1.200	Não



# Algoritmos de Classificação

- Dataset de treino
- Dataset de teste
- Modelo





# Algoritmos de Classificação

## Processo de criação do modelo preditivo

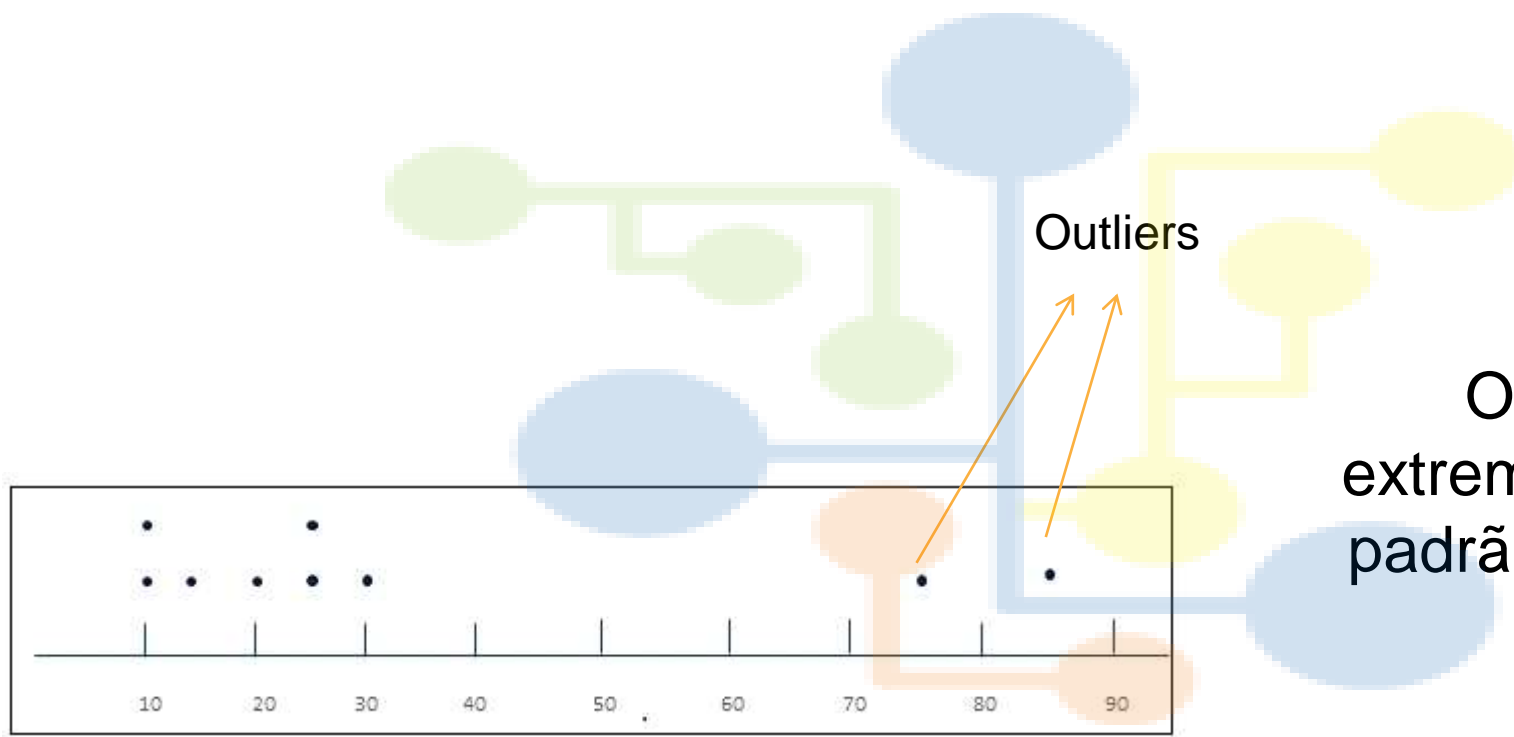
- Limpeza dos Dados
- Definição da variável target
- Definição das variáveis explanatórias
- Seleção do algoritmo
- Treinamento do modelo
- Teste do modelo
- Avaliação do modelo
- Otimização e ajuste do modelo
- Deploy







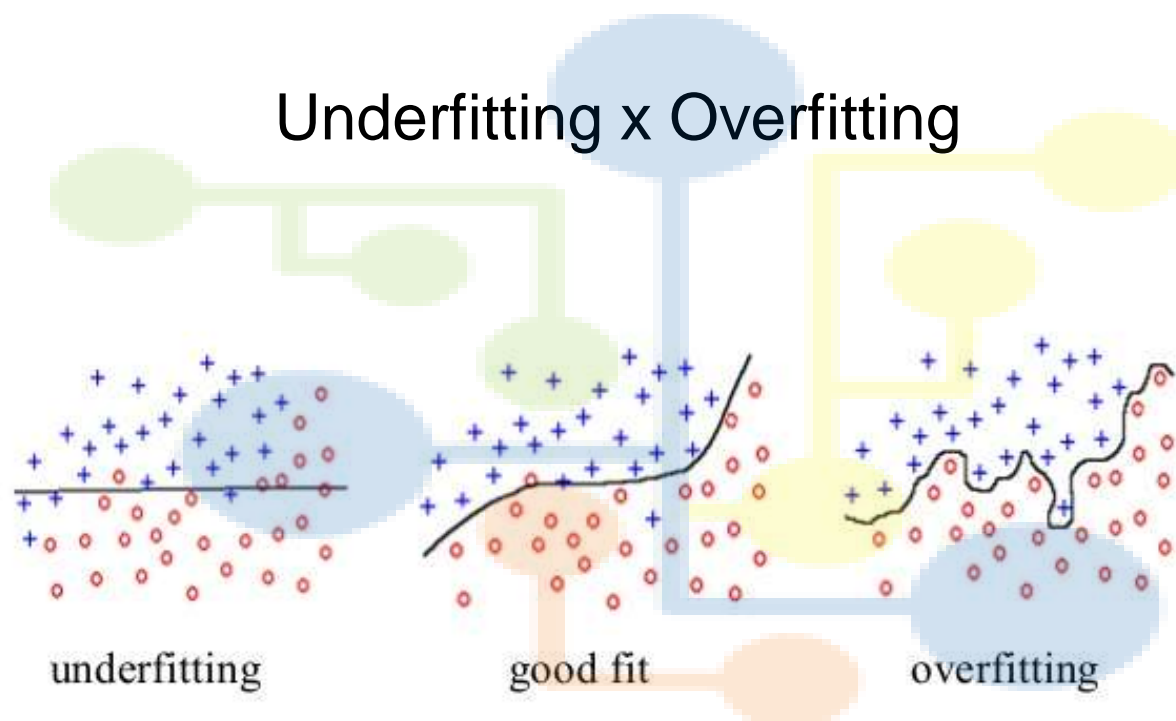
# Algoritmos de Classificação



Outliers são valores extremos que não seguem o padrão esperado nos dados



# Algoritmos de Classificação

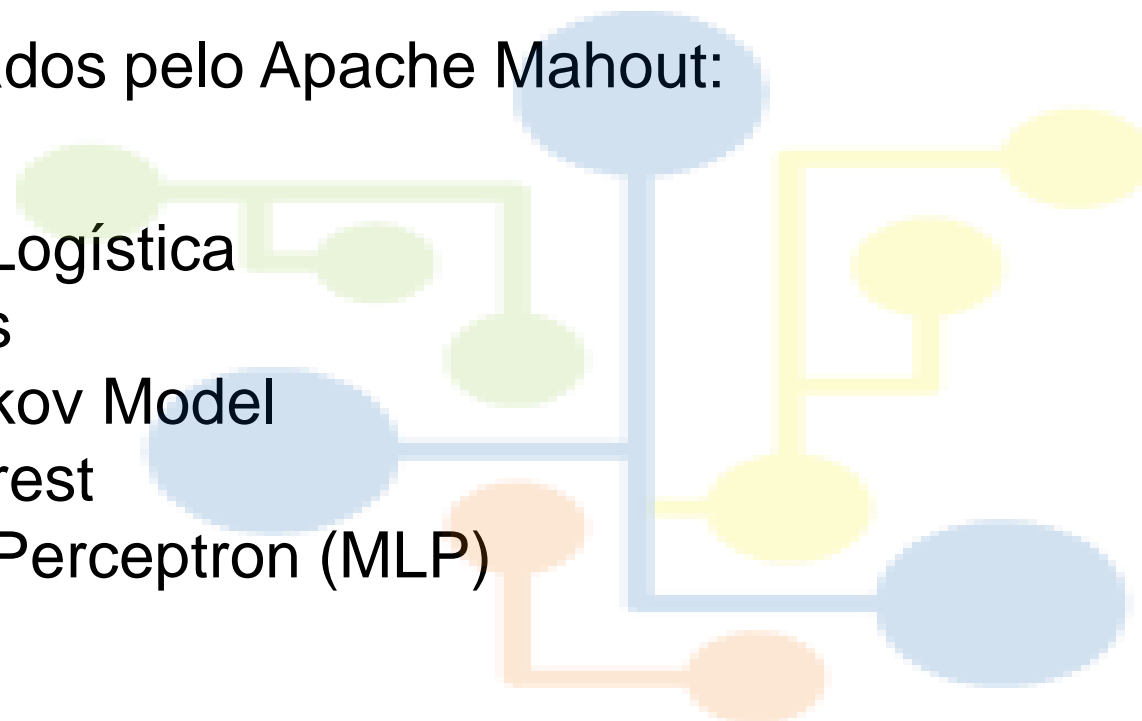




# Algoritmos de Classificação

Algoritmos suportados pelo Apache Mahout:

- Regressão Logística
- Naive Bayes
- Hidden Markov Model
- Random Forest
- Multi-Layer Perceptron (MLP)





# Algoritmos de Classificação

Métricas de Ferramentas para Avaliar o Modelo de Classificação:

- Confusion Matrix
- Gráfico ROC
- AUC
- Entropy Matrix





# Algoritmos de Classificação

Se o volume de dados for entre 1 e 10 milhões de registros, o Apache Mahout pode ser uma excelente opção



# Obrigado

---

