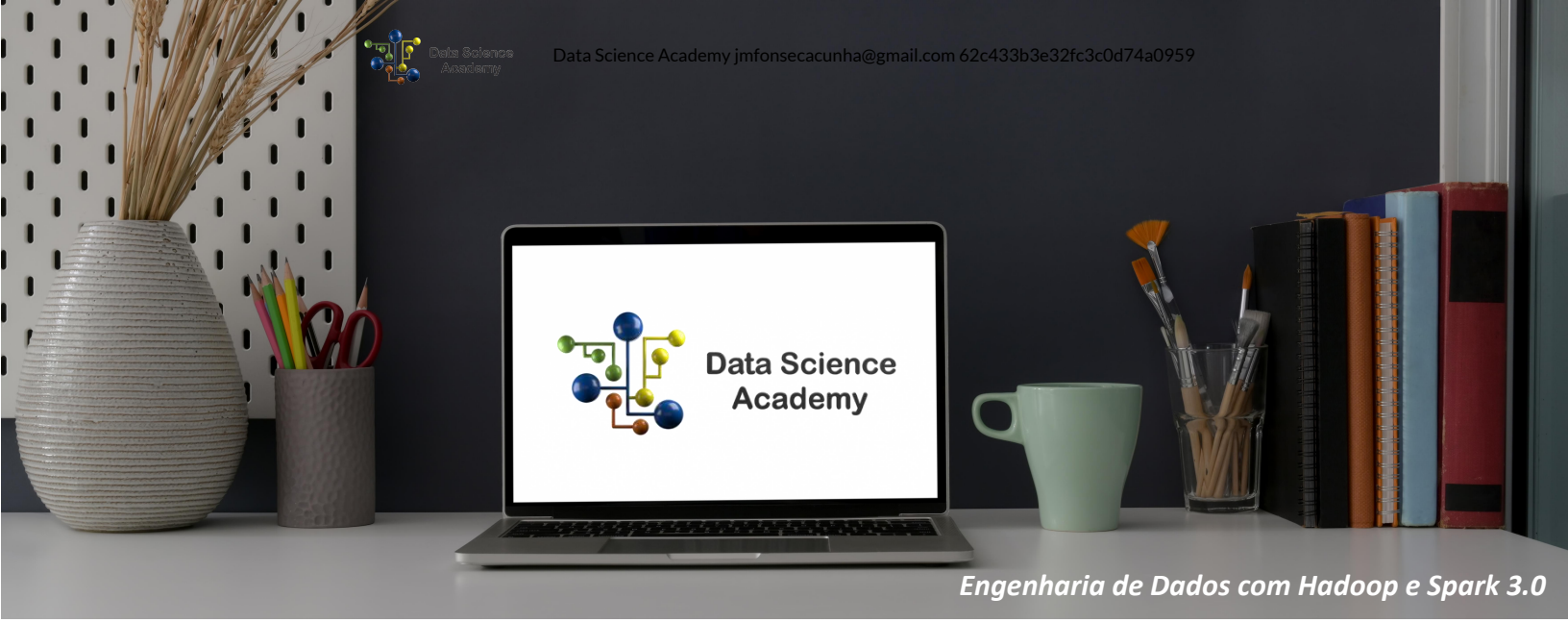



# Engenharia de Dados com Hadoop e Spark Versão 3.0

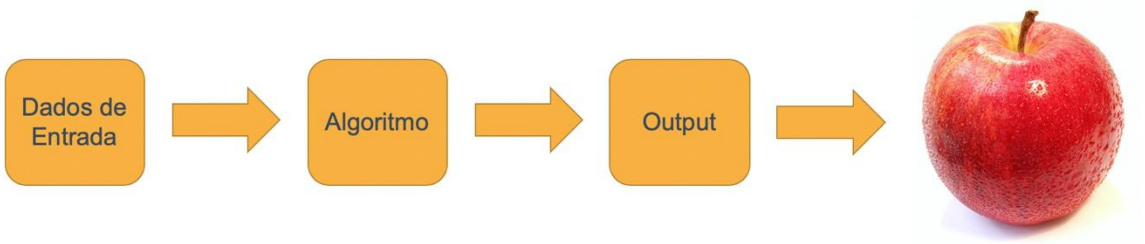
## Machine Learning Algoritmos de Classificação

Vamos nos concentrar na Classificação, uma técnica de aprendizagem supervisionada e que compõe a maioria dos algoritmos suportados pelo Apache Mahout.

Classificação é a categorização de potenciais respostas e em ML queremos automatizar este processo. Temos classificação multiclasse, quando várias classes são possíveis no processo de classificação e temos a classificação binária, quando apenas duas classes são possíveis. Em Machine Learning, usamos um algoritmo que prevê, baseado em uma série de regras, a que classe um objeto ou indivíduo vai pertencer. Normalmente as classes são mutuamente exclusivas, ou seja, um indivíduo não pode pertencer a mais de uma classe ao mesmo tempo. O algoritmo de Machine Learning calcula a probabilidade de um indivíduo ou objeto pertencer a uma determinada classe.

Classificação é uma técnica de aprendizagem supervisionada. Nesta técnica, baseado em fatos históricos, o algoritmo é capaz de prever um valor desconhecido. Em aprendizagem supervisionada, nós já sabemos o possível output, ou seja, um objeto pode pertencer a uma ou outra categoria obrigatoriamente. Mas existem outras técnicas como a aprendizagem não supervisionada, em que não sabemos os possíveis outputs e nesse caso o algoritmo precisa realizar esta previsão.

Vejamos um simples exemplo: suponha que você tenha uma cesta de frutas e você quer classificá-las. Na sequência, faríamos as medidas de todas as frutas e colocaríamos em um banco de dados. Ao usar aprendizagem supervisionada, nós precisaríamos conhecer os possíveis outputs, por exemplo: uma maçã é vermelha e tem 4 centímetros de diâmetro.




De acordo com os dados de entrada, o algoritmo pode prever se a fruta é uma maça ou não

Apresentamos todos os dados ao algoritmo e o treinamos. Ele então aprende sobre o que é uma maça e cada vez que encontrar algo parecido, ele classifica como maça.

Mas e se não tivéssemos os possíveis outuputs, ou seja, não sabemos o que é uma maça? Nesse caso, poderíamos usar aprendizagem não supervisionada e nosso algoritmo se encarrega de aprender sobre todas as características e fazer a classificação das frutas.