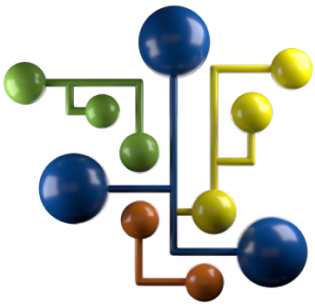

Data Science Academy

www.datascienceacademy.com.br

Microsoft Power BI Para Data Science, Versão 2.0

Por Que Dividimos os Dados em Treino e Teste?

Ao trabalhar com Ciência de Dados você receberá um conjunto de dados contendo linhas e colunas. Cada coluna representa um atributo (uma variável) e cada linha representa uma observação, um registro do evento.

Em Machine Learning teremos que treinar o modelo e depois de treinado teremos que testá-lo. Podemos treinar e testar com os mesmos dados? Não. Ao testar o modelo devemos apresentar dados que o modelo não recebeu durante o treinamento, exatamente para avaliar sua performance. Não treinamos o modelo para funcionar apenas com os dados de treino, mas sim com novos dados que serão apresentados ao modelo para resolver o problema de negócio.

Por essa razão fazemos a divisão dos dados em treino e teste. Por exemplo: se tivermos um dataset com 1000 linhas, podemos reservar 800 linhas para treinar o modelo e 200 linhas para testar o modelo, fazendo a divisão com uma proporção 80/20. Não existe proporção ideal, sendo essa mais uma decisão do Cientista de Dados.

Nos cursos da Formação Cientista de Dados ensinamos diversas técnicas de divisão de dados em treino e teste.