Enap

Análise de Dados em Linguagem R

Módulo 3

Fundamentos de *Machine Learning*

Fundação Escola Nacional de Administração Pública

Presidente

Diogo Godinho Ramos Costa

Diretor de Desenvolvimento Profissional

Paulo Marques

Coordenador-Geral de Educação a Distância

Carlos Eduardo dos Santos

Equipe responsável

Ana Carla Gualberto Cardoso (Diagramação, 2020).

Ana Paula Medeiros Araújo (Direção e Produção Gráfica, 2020).

Douglas Gomes Ferreira (Conteudista, 2020).

Guilherme Teles da Mota (Implementação Rise, 2020).

Iara da Paixão Corrêa Teixeira (Designer Instrucional, 2020).

Juliana Bermudez Souto de Oliveira (Revisão Textual, 2020).

Larisse Padua da Silva (Produção Audiovisual, 2020).

Michelli Batista Lopes (Produção Audiovisual e Implementação, 2020).

Patrick Coelho (Implementação Moodle, 2020).

Sheila Rodrigues de Freitas (Coordenação Web, 2020).

Desenvolvimento do curso realizado no âmbito do acordo de Cooperação Técnica FUB / CDT / Laboratório Latitude e Enap.

Curso produzido em Brasília, 2020.

Enap Escola Nacional de Administração Pública

Enap, 2020

Enap Escola Nacional de Administração Pública

Diretoria de Educação Continuada

SAIS - Área 2-A - 70610-900 — Brasília, DF

Sumário

Módulo 3 Fundamentos de *Machine Learning*

DESTAQUE

Ao final deste módulo, você deverá ser capaz de classificar alguns algoritmos de regressão, classificação e clusterização relacionados à *machine learning*. Além disso, você também desenvolverá aptidão para classificar as técnicas e etapas de construção do modelo preditivo de *machine learning* e seus principais conceitos.

Machine learning ou aprendizagem de máquina é uma representação que tem como objetivo criar um modelo a partir de dados históricos para generalizar decisões.

PERGUNTA

Mas o que vem a ser um modelo nesse contexto?

RESPOSTA

Modelo é uma representação dos relacionamentos existentes nos dados por meio de uma fórmula matemática.

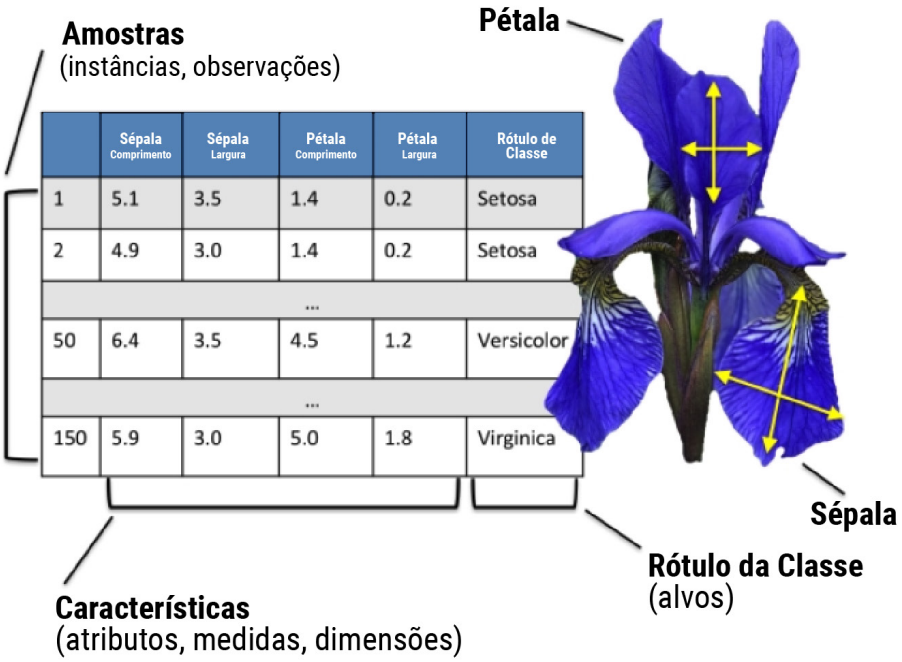
Antes de estudarmos os algoritmos, vamos aprender alguns termos utilizados para se referir a partes específicas de um conjunto de dados.

Instâncias ou observações
São as linhas do *dataset*.

Variável resposta/dependente, classe, *label* ou *target*
É a variável/coluna que se quer prever.

***Features*, atributos, dimensões ou variáveis independentes/explicativas**
São colunas do *dataset* que podem ser utilizadas para prever a variável *target*.

A imagem a seguir ilustra alguns conceitos fundamentais para darmos continuidade ao nosso estudo.



	Sépala Comprimento	Sépala Largura	Pétala Comprimento	Pétala Largura	Rótulo de Classe
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Fonte: Raschka (2015).

Após aprendermos alguns conceitos importantes relacionados aos algoritmos de *machine learning*, vamos estudá-los mais detalhadamente.

Unidade 1 - Algoritmos de *Machine Learning*

Já entendemos que um modelo é uma representação dos dados. Uma vez criado, podemos utilizá-lo em novos conjuntos de dados para realizar predições. Agora, você vai conhecer a relação do modelo com os algoritmos de *machine learning*.

Os algoritmos de *machine learning* são aplicados a um conjunto de dados com objetivo de identificar os relacionamentos existentes e gerar um modelo a partir desses dados.

Existem diversos algoritmos que podem ser utilizados em *machine learning*. Normalmente, são agrupados em duas categorias:

1. **Tipo de aprendizagem**: supervisionada, não supervisionada e outros.
2. **Categorias de problemas**: classificação, regressão, agrupamento, entre outros.

Primeiramente, vamos conhecer os algoritmos de aprendizagem supervisionada, a saber: regressão linear, KNN e árvores de decisão. Posteriormente, vamos verificar o objetivo do *K-means*, algoritmo de aprendizagem não supervisionada e, por último, apresentaremos uma lista com vários outros algoritmos de *machine learning* com a identificação do problema e a classificação quanto ao tipo de aprendizagem que o utiliza.

1.1 Algoritmos de aprendizagem supervisionada

Na aprendizagem supervisionada, a predição é estimada com base na relação entre os dados de entrada (*features*) e os dados de saída (variável resposta). Para cada entrada, é apresentado o resultado esperado.

O algoritmo é responsável por mapear uma função que descreve os padrões ocultos nos dados. Para esse tipo de aprendizagem, **são necessários dados rotulados, que são os dados de entrada associados com o resultado esperado. A função mapeada pelo algoritmo é utilizada para prever novos valores quando apresentada a novos conjuntos de dados.**

A aprendizagem supervisionada pode ser utilizada para resolver problemas de classificação e regressão. A classificação tem como resultado uma saída categórica/discreta. Já a regressão tem como resultado uma saída numérica.

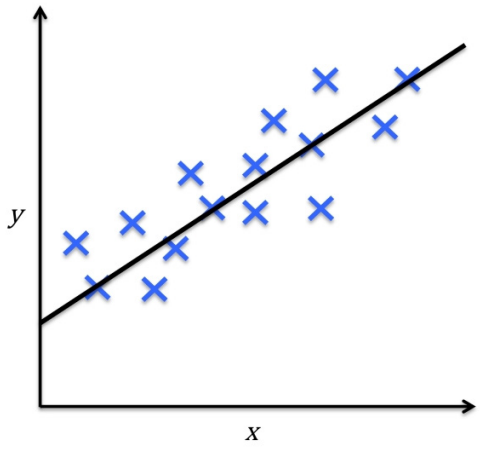
Vamos agora conhecer alguns algoritmos de aprendizagem supervisionada:

- **Regressão Linear**

A análise de regressão simples estuda o relacionamento entre a variável dependente y (variável resposta) e a variável independente x (variável explicativa).

O objetivo é prever o valor de uma variável contínua. A regressão linear assume que existe uma relação linear entre a variável resposta e a variável explicativa.

Acompanhe no gráfico a seguir a representação de uma regressão linear:



Fonte: Raschka (2015).

De acordo com a imagem, o valor de y **é calculado** com a seguinte f**ó**rmula:

$$y = W*x+B$$

W e B são os parâmetros do modelo, onde W = *weight* e B = *bias*.

Dessa forma, o treinamento de um modelo basicamente consiste em estimar os valores de W e B. Após o modelo encontrar os melhores valores com base em uma métrica de avaliação, é possível realizar as previsões.

Considere uma base composta por dados de estudantes, tais como: horas de estudo por dia, quantidade de faltas e nota final. A partir de uma grande quantidade de dados históricos, **é possível** apresentar esses dados a um algoritmo, criar um modelo e utilizá-lo para prever a nota final de outros alunos, conforme a seguinte tabela:

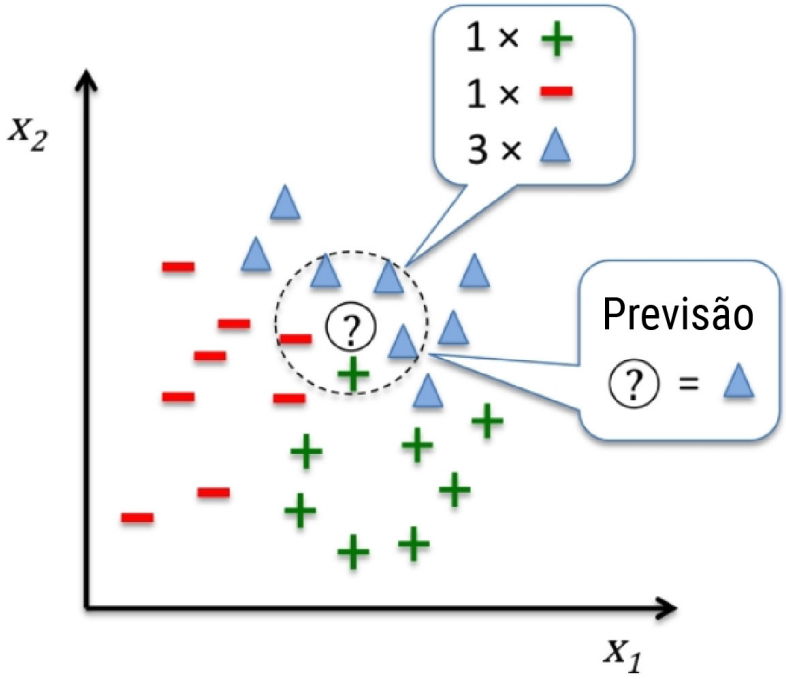
Tabela: Previsão de nota dos alunos

Aluno	Horas de Estudo	Faltas	Nota Final
A	4	2	7,0
B	2	5	5,0
C	6	0	9,5

Fonte: Elaborado pelo autor.

➢ **KNN**

O *K-Nearest Neighbors* (KNN) é um algoritmo de classificação que se baseia nos vizinhos mais próximos. Quando um novo dado é apresentado ao algoritmo, ele irá classificá-lo com base nos exemplos mais próximos apresentados na fase de treinamento.



Fonte: Raschka (2015).

O parâmetro k representa a quantidade de vizinhos mais próximos que deve ser considerada pelo algoritmo. Analisando o gráfico apresentado e considerando o valor de k = 3, temos que o novo

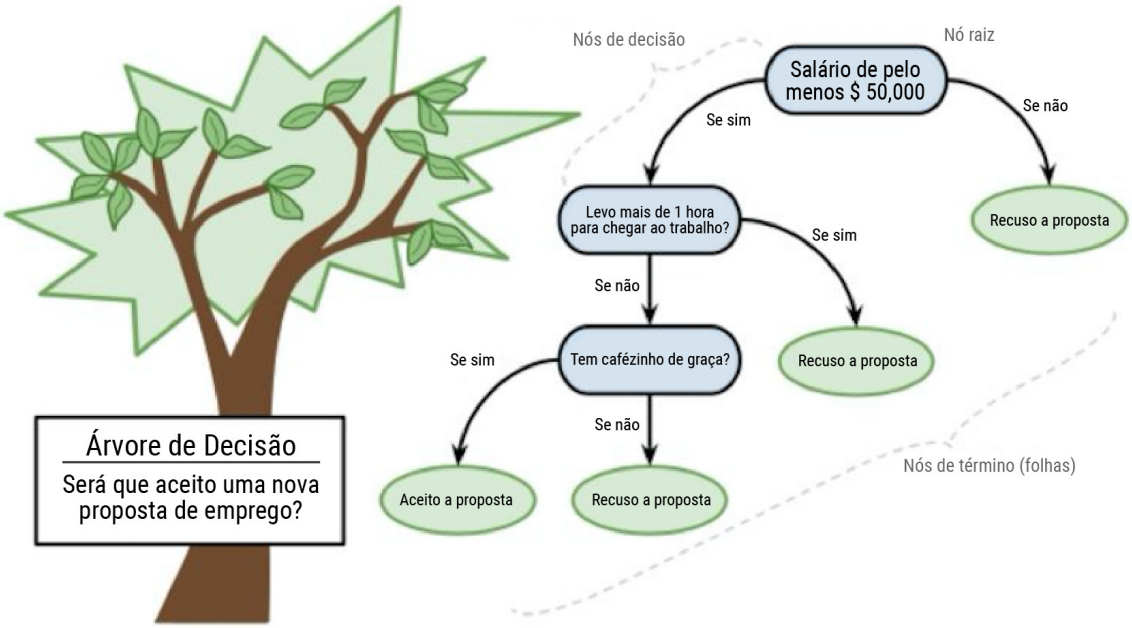
elemento (?) é classificado como triângulo, pois é a quantidade de elementos mais próximos da nova observação.

➢ Árvore de Decisão

Será que você já utilizou alguma árvore de decisão em sua vida pessoal ou profissional? Você já tomou alguma decisão se baseando em respostas para os diversos cenários inerentes àquele contexto? Você já utilizou um fluxograma em seu trabalho? Se a sua resposta foi afirmativa para qualquer um desses questionamentos, então você já fez uso dessa ferramenta que, simplificadamente, é a representação de uma tabela de decisão sob a forma de uma árvore.

Árvore de decisão é uma estrutura que armazena regras de decisão e possui nós, ramos e folhas. Os nós representam as variáveis, os ramos representam os valores possíveis de cada nó e as folhas representam o valor final de um nó.

Na imagem a seguir, temos a representação de uma árvore de decisão:



Fonte: Lantz (2015).

Com isso, o modelo aprendeu as regras de quando aceitar e quando não aceitar uma proposta de emprego. Dessa forma, ao ser apresentado um novo dado, uma nova proposta de emprego, o modelo é capaz de tomar essa decisão.

1.2 Algoritmos de aprendizagem não supervisionada

Esse tipo de aprendizagem é usado quando não temos os dados rotulados, ou seja, quando não temos a saída esperada para uma determinada entrada. Assim, o algoritmo aprende sem informações adicionais para produzir uma saída.

Os algoritmos de aprendizagem não supervisionada são baseados em medidas de similaridade e padrões ocultos nos dados. Normalmente, é utilizado para resolver problemas de agrupamento (*clustering*), associação e detecção de anomalias.

Clustering é uma atividade frequentemente utilizada para agrupar os dados que possuem características similares.

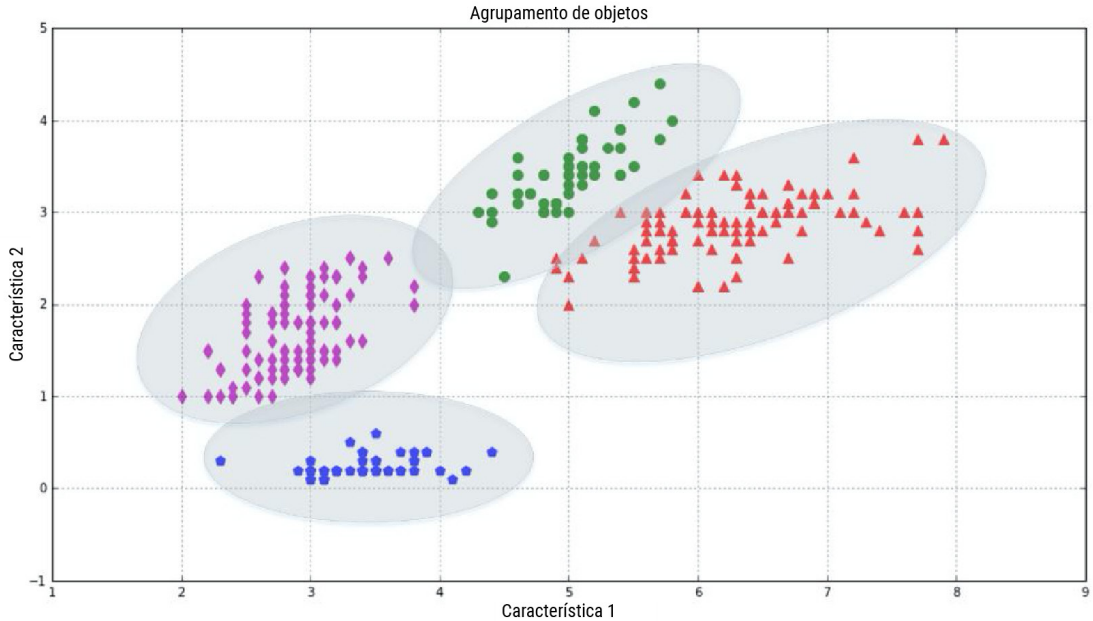
Um algoritmo muito conhecido para essa tarefa é o *K-means*. Acompanhe:

➢ ***K-means***

No artigo *Introdução* ***Básica à*** *Clusterização*, do Laboratório de Aprendizado de Máquina em Finanças e Organizações (LAMFO/UNB), Honda (2017) diz que "o algoritmo se chama assim pois **encontra k *clusters* diferentes** no conjunto de dados. O **centro de cada *cluster* será chamado centroide** e terá a média dos valores neste cluster." Ainda de acordo com o autor, o *K-means* é um dos algoritmos mais básicos para a clusterização.

O objetivo desse algoritmo é dividir os dados em grupos com base na similaridade dos dados (*clusters*), ou seja, temos dados que são similares dentro de um grupo, porém diferentes quando comparados com os dados de outros grupos.

A sua representação gráfica é apresentada na figura a seguir.



Fonte: Bonaccorso (2017).

Nesse exemplo, pode-se perceber que cada cluster possui dados diferentes, porém são similares quando comparados com os dados do mesmo cluster.

1.3 Outros algoritmos de *machine learning*

Atualmente, temos inúmeros algoritmos de *machine learning*. Na tabela a seguir, apresentamos algoritmos frequentemente utilizados em aprendizagem supervisionada e não supervisionada:

Algoritmos de *Machine Learning*		
Algoritmo	**Problema**	**Tipo de Aprendizagem**
Logistic Regression	Classificação	Supervisionada
K-Nearest Neighbor	Classificação	Supervisiona
Naive Bayes	Classificação	Supervisionada
Decision Trees	Classificação	Supervisionada
Regression Trees	Regressão	Supervisionada
Linear Regression	Regressão	Supervisionada
Neural Networks	Classificação/Regressão	Supervisionada
Support Vector Machines	Classificação/Regressão	Supervisionada
Random Forest	Classificação/Regressão	Supervisionada
PCA	Redução de dimensionalidade	Não supervisionada
Association Rules	Detecção de padrões	Não supervisionada
K-means Clustering	Agrupamento	Não supervisionada
DBSCAN	Agrupamento	Não supervisionada

Fonte: Elaborada pelo autor.

Depois de conhecer os diversos algoritmos utilizados na aprendizagem de máquina, é importante escolher o mais adequado ao problema proposto.

Unidade 2 - Construção do Modelo Preditivo

DESTAQUE

Agora, vamos aprender algumas etapas envolvidas na construção do modelo de *machine learning*, considerando a aprendizagem supervisionada.

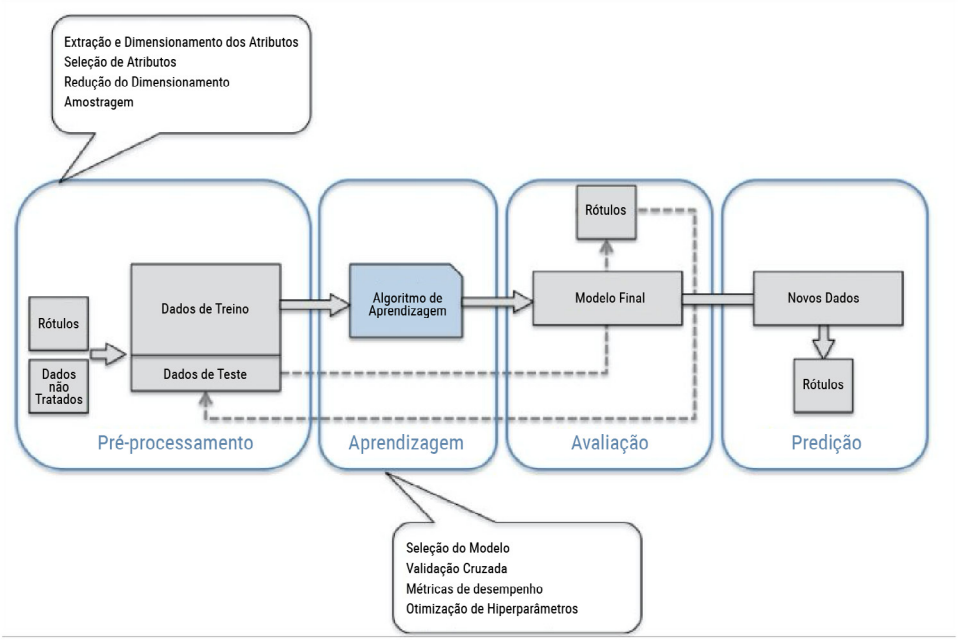
2.1 Técnicas e etapas de construção do modelo de *machine learning*

Ao criar um modelo de *machine learning*, dificilmente teremos os dados prontos. Por esse motivo, faz-se necessário realizar algumas transformações nos dados antes de apresentá-los ao algoritmo.

Na fase de pré-processamento, os dados são divididos em dados de treino e dados de teste. Os dados de treino são apresentados ao algoritmo para que ele aprenda o relacionamento entre as variáveis e crie o modelo. Os dados de teste, por sua vez, são utilizados para avaliar o quanto o algoritmo aprendeu.

Ao apresentar os dados de teste ao modelo, as previsões são realizadas com base no que foi aprendido na fase de treinamento. Essas previsões são comparadas com as respostas esperadas para calcular o desempenho do modelo. Uma vez criado e validado, o modelo pode ser utilizado para realizar novas previsões quando for apresentado a novos dados.

A imagem a seguir apresenta uma esquematização das atividades envolvidas na construção de um modelo preditivo:



Fonte: Raschka (2015).

Após analisar a imagem, vamos detalhar cada uma dessas etapas de construção do modelo de *machine learning*:

1. Pré-processamento dos Dados

Essa fase tem como objetivo melhorar a qualidade dos dados que serão apresentados ao algoritmo.

Algumas técnicas utilizadas são:

Feature selection
É utilizada para selecionar os atributos mais relevantes que serão utilizados para treinar o modelo.

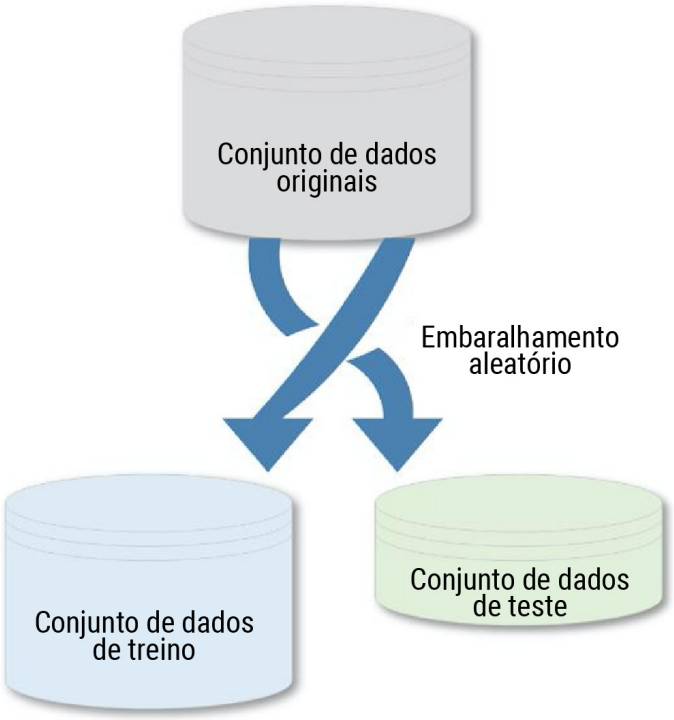
Feature engineering
É a arte de criar variáveis a partir de um conjunto de dados para melhorar a performance do modelo.

Normalização
Um *dataset* pode conter variáveis em diferentes escalas e, assim sendo, recomenda-se padronizar esses dados para uma mesma escala. Para isso, usamos a técnica da normalização.

Redução de dimensionalidade
É utilizada quando se tem conjuntos de dados com muitas dimensões (colunas), o que acaba prejudicando a capacidade de generalização de alguns algoritmos.

Divisão dos dados em treino e teste
Em aprendizagem supervisionada, deve-se dividir os dados em dois *datasets*: treino e teste. Os dados de treino são usados para criar o modelo e os dados de teste são usados para verificar a performance do modelo. Além disso, essa divisão deve ser feita de forma aleatória.

A figura a seguir ilustra a técnica da divisão dos dados em treino e teste:



Fonte: Boschetti, Massaron (2016).

2. Aprendizagem – Construção do Modelo

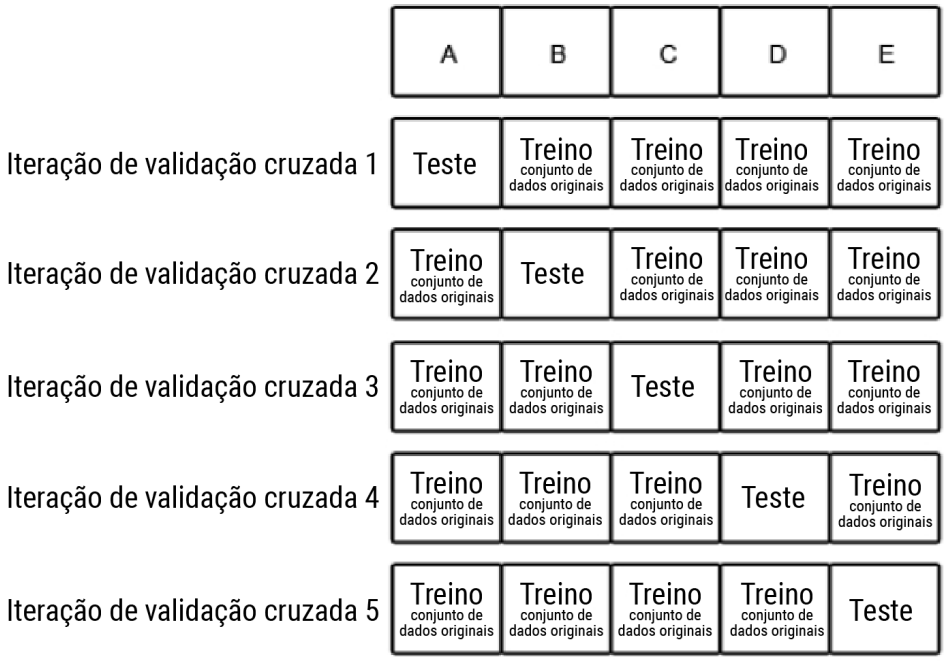
Nessa fase, o modelo é construído a partir dos dados que são apresentados ao algoritmo.

Algumas técnicas utilizadas são:

- ***Cross-validation***

É utilizada para treinar e validar um modelo com o mesmo conjunto de dados, dividindo-os em partições.

Analisando a imagem a seguir, temos que, a cada iteração, o algoritmo troca os dados de treino e teste com o objetivo de obter um melhor desempenho.



Fonte: Hackeling (2014).

- **Métricas de desempenho**

Existem diversas métricas para medir o desempenho de um modelo. Apenas para exemplificar, podemos medir a acurácia, que é o percentual de previsões corretas em problemas de classificação.

- **Otimização de hiperparâmetros**

Cada algoritmo possui um conjunto de hiperparâmetros que podem ser alterados. Assim, essa técnica busca encontrar a combinação certa de valores com o objetivo de melhorar a performance do modelo.

3. Avaliação do Modelo

Nesta fase, os dados de teste são apresentados ao modelo e, com isso, são geradas previsões. Essas previsões são comparadas com os resultados desejados para avaliar o desempenho do modelo.

4. Predição

Se o modelo avaliado apresentar um bom resultado, poderá ser utilizado para receber novos dados e realizar previsões.

Referências

Unidade 1 - Algoritmos de *Machine Learning*

BONACCORSO, G. **Machine Learning Algorithms**. Birmingham: Packt Publishing, 2017.

CIABURRO, G. **Regression Analysis with R**. Birmingham: Packt Publishing, 2018.

DASGUPTA, N. **Practical Big Data Analytics**. Birmingham: Packt Publishing, 2018.

GOLLAPUDI, S. **Practical Machine Learning**. Birmingham: Packt Publishing, 2016.

HACKELING, G. **Mastering Machine Learning with scikit-learn**. Birmingham: Packt Publishing, 2014.

HONDA, H. Introdução básica à Clusterização. **Laboratório de Aprendizado de Máquina em Finanças e Organizações**, Brasília, 5 out. 2017. Disponível em: https://lamfo-unb.github.io/2017/10/05/Introducao_basica_a_clusterizacao/. Acesso em: 26 maio 2020.

KUMAR, A. **Learning Predictive Analytics with Python**. Birmingham: Packt Publishing, 2016.

LANTZ, B. **Machine Learning with R**. 2. ed. Birmingham: Packt Publishing, 2015.

OZDEMIR, S. **Principles of Data Science**. Birmingham: Packt Publishing, 2016.

RASCHKA, S. **Python Machine Learning**. Birmingham: Packt Publishing, 2015.

BOSCHETTI, A.; MASSARON, L. **Python Data Science Essentials**. 2. ed. Birmingham: Packt Publishing, 2016.

Unidade 2 - Construção do Modelo Preditivo

BONACCORSO, G. **Machine Learning Algorithms**. Birmingham: Packt Publishing, 2017.

BOSCHETTI, A.; MASSARON, L. **Python Data Science Essentials**. 2. ed. Birmingham: Packt Publishing, 2016.

CHINNAMGARI, S. K. **R Machine Learning Projects**: Implement supervised, unsupervised, and reinforcement learning techniques using R 3.5. Birmingham: Packt Publishing, 2019.

HACKELING, G. **Mastering Machine Learning with scikit-learn**. Birmingham: Packt Publishing, 2014.

OZDEMIR, S. **Principles of Data Science**. Birmingham: Packt Publishing, 2016.

RASCHKA, S. **Python Machine Learning**. Birmingham: Packt Publishing, 2015.