06

Trabalhando com dados faltantes

Durante a aula discutimos sobre valores faltantes em uma base de dados, os erros que podem ocasionar na criação de um modelo de machine learning e como tratar essa informação. Então, vamos praticar!

Considere que temos a seguinte base de dados:

Diagnóstico	Tumor maior que 5 cm?	Região do tumor dolorida?	Histórico familiar?	Detectado a mais de 6 meses?
Benigno	0	1	1	
Maligno	1	0	1	
Maligno	1	0	1	
Benigno	0	1	1	0
Benigno	0	1	1	
Maligno	1	0	1	1
Maligno	1	0		
Benigno	0	1	1	

Sabendo que o classificador utilizado na criação do modelo de machine learning não suporta entrada vazia, qual tratamento na base de dados é preciso para classificar a coluna de Diagnóstico?

Selecione uma alternativa

A Temos duas colunas com dados incompletos, de modo que a coluna “Detectado a mais de 6 meses?” apresenta apenas 2 valores preenchidos, ou seja, não é possível prever algo sobre esses dados faltantes, logo podemos excluí-la. A coluna “Histórico familiar?” apresenta apenas 1 valor faltante, ou seja, não há como realizar nenhuma previsão sobre esses dados faltantes, logo podemos excluí-la.

B Temos duas colunas com dados incompletos. A coluna “Detectado a mais de 6 meses?” apresenta apenas 2 valores preenchidos, ou seja, não é possível prever os dados faltantes, logo podemos excluí-la. A coluna “Histórico familiar?” apresenta apenas 1 valor faltante e, como podemos verificar, há uma forte tendência de que esse valor não preenchido seja 1, logo, podemos completar esse dado faltante com 1.

C Temos duas colunas com dados incompletos, de modo que a coluna “Detectado a mais de 6 meses?” apresenta apenas 2 valores preenchidos, então é possível prever esses dados faltantes. Dessa forma, podemos preencher benigno com 0 e maligno com 1. A coluna “Histórico familiar?” apresenta apenas 1 valor faltante, e como podemos verificar há uma forte tendência de que esse valor não preenchido seja 1, logo podemos completar o dado faltante valor 1.