



Análise de Dados com Linguagem Python

Análise de Dados com Linguagem Python

Técnicas de Imputação de Variáveis Quantitativas

A imputação para variáveis quantitativas visa substituir valores ausentes por valores numéricos que representem a informação contida na variável.

A principal técnica de imputação nesse caso é usar uma medida de tendência central da variável e então imputar os valores ausentes com essa medida. As medidas de tendência central são: média, mediana e moda.

A média é a média aritmética dos valores da variável. Normalmente usamos a média quando a variável segue uma distribuição normal e não tem outliers (valores além de 1.5 a 3 acima ou abaixo do intervalo interquartil).

A mediana é o valor do meio da distribuição quando os dados estão ordenados. Normalmente usamos a mediana quando a média não pode ser usada.

A moda é o valor que parece com mais frequência na variável e pode ser uma opção quando a média não puder ser usada.

Outra técnica para imputação de variáveis quantitativas é usar o forward ou backward fill preenchendo o valor ausente com o último valor válido, fazendo isso para frente (forward) ou para trás (backward).

Podemos ainda usar Machine Learning para prever o valor ausente e então realizar a imputação.