Próximo
Desafio

Apresentação

Em *Ciência de dados*, existem cenários onde o pesquisador necessita de análises de dependência ou associação entre dados bivariáveis (ou seja, entre duas variáveis). A partir disso, as correlações lineares necessárias para esse tipo de estudo são medidas. Uma das mais conhecidas é o coeficiente de Pearson, que mede a relação linear entre variáveis contínuas. Para calcular as correlações lineares baseadas no coeficiente de Pearson, são utilizadas como ferramentas as bibliotecas Pandas e Numpy.

Além de variáveis contínuas, bases de dados também apresentam variáveis categóricas. Tipicamente, as correlações usando tabelas cruzadas (*cross-tabs*) servem para determinar a correlação entre variáveis categóricas em bases de dados. É necessário verificar se múltiplas variáveis são altamente correlacionadas. Assim, a multicolinearidade indica se há existência forte de correlação entre duas (ou mais) variáveis independentes. A multicolinearidade está presente em bases de dados, portanto se torna necessário o entendimento sobre ela. Consequentemente, é importante lidar com essa característica na base de dados e como tratá-la, caso necessário.

Nesta Unidade de Aprendizagem, você aprenderá os cálculos de correlações lineares baseadas no coeficiente de Pearson na forma matemática e utilizando as bibliotecas Pandas e Numpy, identificará as principais técnicas de correlação presentes em bases de dados com variáveis categóricas a partir de tabelas cruzadas e reconhecerá a multicolinearidade a partir de um conjunto de dados, com definições matemáticas e códigos em Python. Além disso, as técnicas de normalização para análise de correlações de variáveis categóricas e contínuas serão apresentadas ao longo desta Unidade de Aprendizagem.

Bons estudos.