Tópicos em Inteligência Artificial: eXplainable AI

T-P-I: 4-0-4
Carga horária: 144 horas

Objetivos: Algoritmos de Inteligência Artificial faz cada vez mais parte de nosso dia a dia, seja em simples funções para eletrônicos ou na ajuda de tomada de decisões. Em situações de alto-risco, como na área da saúde e questões que afetam diretamente a sociedade, é importante entender o processo de tomada de decisão para validar ou invalidar a resposta do modelo, compreender as características do fenômeno estudado, garantir que o modelo possui determinadas propriedades, debugar o modelo, dentre outras possibilidades. Esse curso tem o objetivo de estudar e revisar artigos importantes da área de interpretação e explicação de modelos de aprendizado de máquina, geração de modelos com restrição de propriedades, e estudo de modelos "caixa-branca" que permite uma interpretação sem necessidade de uso de ferramentas externas.

Conteúdo programático: Interpretabilidade x Explanabilidade, Tipos de interpretação de modelos, Importância de atributos, Explicação baseada em protótipos, Algoritmos geradores de regras de decisão, Explicadores agnósticos de modelo, Visualização do comportamento, Equidade de tratamento, Modelos caixa-branca.