Seminário

EXPLICAÇÕES HUMANO-INTERPRETÁVEIS PARA MODELOS BLACK-BOX DE MACHINE LEARNING: UMA APLICAÇÃO À DETEÇÃO DA FRAUDE

27-10-2021 / 18h00 / Edifício B – Auditório

EI / EAU

ORADOR

Vladimir Balayan

RESUMO

Neste seminário, explicações humano-interpretáveis para modelos black-box de Aprendizagem Automática/Machine Learning (ML) são analisadas e aplicadas a um problema de deteção de fraude proposto pela empresa Feedzai, que utiliza ML para prevenir o crime financeiro. Um dos principais produtos Feedzai é uma aplicação de gestão de casos utilizada pelos analistas de fraude para analisar transações financeiras suspeitas, sinalizadas pelos modelos ML. Os analistas de fraude são peritos que não possuem conhecimento profundo de ML; consequentemente, os atuais métodos explicativos de inteligência artificial não se adequam às suas necessidades de informação. O objetivo deste trabalho consistiu no desenvolvimento de uma estrutura baseada em redes neurais que agrega a tarefa de tomada de decisão com as explicações associadas ao domínio de conhecimento, fornecendo uma perceção clara sobre as previsões do modelo.

REFERÊNCIA BIOGRÁFICA

Vladimir Balayan é investigador em Ciência de Dados na Feedzai, no grupo FATE (Responsável IA) e Mestre em Análise e Engenharia de Big Data pela Universidade Nova. O foco da sua investigação é a interpretação e clarificação de modelos de ML e a sua tese, orientada por Pedro Saleiro e Ludwig Krippahl, foca a criação de explicações humano-interpretáveis para a tomada de decisão no domínio da deteção de fraude. Os seus trabalhos (como autor e co-autor) foram publicados nos workshops de NeurIPS'2020 e ICLR'2021.

POLITÉCNICO DE LEIRIA
ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO
WWW.ESTG.IPLEIRIA.PT | FACEBOOK ESTGLeiria