Magyarázható modellek prototípusok segítségével

Csató Lehel

Babeş-Bolyai Tudományegyetem, Kolozsvár, Matematika és Informatika Kar

lehel.csato@ubbcluj.ro

Napjainkban a mesterséges-intelligencia modellek vagy a kapcsolódó technológiák majd minden területet átitatnak: le tudunk olvasni autó-rendszámtáblákat, tudunk szelektálni zöldségeket egy futószalagon, illetve a mobiltelefonunk nagyon sok alkalmazást futtathat, melyek mindegyike a mesterséges intelligencia – mint eszköztár – alkalmazásával old meg egy feladatot. Az általános elterjedtség ellenére ezen modellek fekete dobozként működnek: például ha egy osztályozási feladatra javasolnak megoldást, nem tudjuk meghatározni a választott kimenet okát. Az átláthatatlanság sok alkalmazási területen hátrány, például orvosi diagnosztikában vagy nagybiztonságú rendszerek fejlesztése esetén. Az általunk bemutatott prototípus-alapú döntési rendszerek a modellek magyarázhatóságát igyekeznek növelni úgy, hogy a modell felépítése során strukturális elemekkel tesszük azt magyarázhatóbbá. Bemutatunk különböző architektúrákat és futási eredményeket, melyeket ezen modellek implementálása során kaptunk.