STUDENTŲ ĮTRAUKIMO Į MOKSLINĘ VEIKLĄ KONKURSO TEMA

Temos pavadinimas:
Duomenų dimensijos mažinimo įtakos mašininio mokymo metodų tikslumui tyrimas

Tikslas:
Nustatyti, kaip kinta mašininio mokymo metodų rezultatų tikslumas, keičiant duomenų rinkinio dimensiją

Trumpas temos vykdymo aprašymas (ne daugiau kaip 2000 ženklų):

Taikant mašininio mokymo metodus duomenų analizei naudojami įvairūs duomenų rinkiniai. Pvz. www.kaggle.com Paprastai laikoma, kad kuo didesnis duomenų atributų (dimensijų) skaičius, tuo galima gauti didesnį mašininio mokymo metodo tikslumą. Tačiau taip nėra, nes dalis duomenų rinkinyje gali būti nepilni, priklausomi vienas nuo kito, todėl nekontroliuojamas žemos kokybės duomenų naudojimas gali sukelti per daug triukšmo, žymiai sulėtinti mašininio mokymo algoritmo darbą, didinti skaičiavimų laiką ir mažinti tikslumą.

Darbe reikia panagrinėti duomenų rinkinio dimensijų mažinimo metodus ir juos pritaikyti pasirinktam pasirinktam duomenų rinkiniui pvz. „KDD Cup“. Tada palyginti mašininio mokymo metodų (pvz. Naivus Bajesas, SVM, sprendimų medis) tikslumą, gautą naudojant sumažintų dimensijų rinkinius. Rezultatai pateikiami grafikų pavidalu, atliekama gautų rezultatų analizė.

Darbas atliekamas naudojant vieną iš įrankių: KMINE, Weka, Python Scikit-Learn. Tai priklauso nuo studento galimybių ir pasiruošimo.

Temą siūlantis mokslininkas/dėstytojas:
prof.dr. Dalius Mažeika