

# Projekt z predmetu Strojové učenie

**2018/2019**

Cieľom projektu je riešenie praktického problému pomocou algoritmov strojového učenia pomocou balíka **WEKA** (alebo iného nástroja na dolovanie údajov). Cieľom projektu je získať skúsenosti s aplikáciou týchto algoritmov.

Riešený problém: **konštrukcia klasifikátora** (vo forme rozhodovacieho stromu, booleovskej formuly, rozhodovacieho zoznamu, neurónovej siete a pod.) na základe tréningovej množiny. Riešený problém môže byť stanovený aj inak po dohode s vyučujúcim.

Vašou úlohou je vybrať si vhodnú údajovú sadu (dataset) rozdelený na tréningovú a testovaciu množinu. Zvoľte si aspoň dva vhodné algoritmy implementované v balíku WEKA, použite vybrané algoritmy na konštrukciu klasifikátora na základe tréningovej množiny, spracujte výsledky a skúsenosti v tvare textu v rozsahu  $\geq 5$  (A4) strán, a odovzdajte v elektronickej podobe (PDF, Word).

Popis riešenia problému má pozostávať z nasledujúcich častí:

- **popis dát:** aké dáta využívate, čo reprezentujú, z akého zdroja pochádzajú, aké majú vlastnosti ( $\leq 3$  body) ,
- **popis použitých algoritmov:** ako reprezentujú svoje hypotézy, ako fungujú ( $\leq 4$  body),
- **popis evaluačných kritérií:** ako a čím ste merali kvalitu získaných výsledkov - na osobitnej testovacej množine, pomocou cross-validation, atď. ( $\leq 3$  body),
- **popis výsledkov:** nájdené hypotézy a ich presnosť, porovnanie úspešnosti použitých algoritmov, prezentácia výsledkov pomocou tabuliek/grafov ( $\leq 5$  bodov),
- **interpretácia výsledkov** vlastnými slovami, porovnanie výsledkov s očakávaniami: prečo bol ten-ktorý algoritmus lepší/horší, diskusia (napr. porovnať zrozumiteľnosť nájdených riešení, výpočtovú rýchlosť jednotlivých algoritmov, návrhy na možné vylepšenia) ( $\leq 5$  bodov).

Vyššie uvedené body budú udelené na základe kvality jednotlivých častí referátu. Za projekt možno získať 20 bodov.

Vhodné dáta (trénovacie a testovacie množiny) je možné nájsť napríklad na stránke:

**<https://www.kaggle.com/>**

Je možné použiť aj iné zdroje. Projekty budú prezentované na cvičeniach na konci semestra. Predpokladáme, že v skupine každý študent bude pracovať s inými dátami. Projekt je potrebné odovzdať do **28. apríla 2019**.