# Umělá inteligence 1 – cvičení pro dolování znalosti z dat pomocí systému WEKA

Vytvořte 5 (pět) textových souborů umělých dvourozměrných dat podle příkladů na obrázcích uvedených níže. Data představují dvě klasifikační třídy (černé a bílé body) oddělené hledanou neznámou hranicí, znázorněnou na obrázcích čárkovanou čarou. Každý z pěti čtverců představuje oblast o rozměru 100 x 100 (tj. například osa *x* je od 0.0 do 100.0, osa *y* také). Do každého čtverce vygenerujte libovolným způsobem 500 černých a 500 bílých bodů tak, aby byly ***náhodně*** rozmístěny vždy na své příslušné straně hranice. Celkem tedy je pro obě třídy 1000 trénovacích příkladů (500+500). Data (body) jsou popsána svými číselnými souřadnicemi *x* a *y* včetně příslušnosti do jedné ze dvou tříd. Poznámka: přiložené obrázky jsou pouze ilustrativním příkladem, jak by mohla data vypadat – vytvořte svá vlastní data (např. pomocí MS Excel nebo generátoru náhodných čísel v libovolném programovacím jazyce).

Pro jednoduchost vytvořte formát dat jako prostý textový `*.csv` soubor (*comma separated values* = čárkami oddělené hodnoty), kde v prvním řádku jsou ***čárkami*** oddělené libovolně zvolené názvy proměnných (např. *x, y, class*). V dalších řádcích jsou (v tomto pořadí) vždy č*íselné souřadnice bodů* (reálná čísla) a jejich *příslušnost do třídy* (nominální hodnota) oddělené ***čárkami*** (nikoliv středníky! – česká verze MS Excel používá středník, protože v des. číslech je čárka; nutno použít desetinnou tečku a jako oddělovač čárku). Příklad:

```
x, y, class
5.3, 7.8, black
64.0, 92.1, white
19.8, 49.5, black
.
.
.
```

Celkem 1000 řádků dat (plus první řádek s názvy proměnných), 3 sloupce. Datové soubory uložte jako prostý ASCII text s příponou .csv, např. *data1.csv, data2.csv,* apod. Po spuštění systému WEKA postupně pro každý z pěti souborů vyzkoušejte účinnost těchto vybraných klasifikátorů:

- J48 (rozhodovací strom na bázi minimalizace entropie),
- Naivní Bayes (pravděpodobnostní systém se zjednodušením),
- Naive Bayes Multinominal
- IB1 (1-NN), IBk (k-NN pro $k = 2, 3, 4, 5$),
- Multilayer Perceptron (pro několik hodnot počtu trénovacích epoch, učící konstanty a momenta),
- PART (generátor pravidel – rules),
- v případě zájmu libovolný další (6 uvedených jsou povinné).

Výpočty probíhají velice rychle. Pro testovaní použijte metodu krosvalidace s dělením na 10 částí a dále 1-z-N (kde N je celkový počet trénovacích příkladů, tj. 100). Pro vyhodnocení chyby klasifikátoru použijte ***accuracy*** (počet černých zařazených chybně k bílým a naopak). Pro J48 zaznamenejte pro každý z pěti případů rozměr stromu (počet uzlů a počet listů – poskytuje přímo WEKA). Výsledky trénování a testování lze pomocí *copy+paste* převzít přímo z výsledkového okna WEKA (do zprávy o výsledcích uveďte kopie jen pro nejlepší výsledky, nikoliv všechny). Výsledky testování uvedených klasifikátorů stručně slovně vyhodnoťte, uveďte, který klasifikátor s kterými parametry na kterých datech poskytoval nejlepší a nejhorší *accuracy* a pokuste se vysvětlit, proč. Výsledky a jejich vyhodnocení zapište volnou stručnou formou do textového souboru (MS Word, LaTeX, apod.), převeďte do pdf souboru, přiložte k němu své datové soubory, zkomprimujte do jednoho *.zip nebo *.rar souboru a odevzdejte do odevzdávárny UI-1 v UIS. Odevzdaná úloha je jedním z předpokladů k řádnému ukončení předmětu. Na projekt jsou vyhrazena dvě cvičení. Věnujte pozornost lhůtě odevzdání uvedené v odevzdávárně.