

# Testování a validace modelů

Michael Anděl



Evropský sociální fond  
Praha & EU: Investujeme do vaší budoucnosti

*Fakulta elektrotechnická, ČVUT*

- ▶ V RapidMineru vytvořte co nejpřesnější prediktivní model pro data *horse colic*, který bude predikovat `surgical lesion`!
- ▶ Bodována bude přesnost modelu na neznámých datech.

## Jak na to:

1. Založte proces v RM. Importujte data *horse colic*. Data obsahují různé typy příznaků (spojité, diskrétní/ordinální, nominální). Věnujte tedy pozornost správnému otagování atributů během importu.
2. Implementujte bloky vám známých klasifikátorů (z přednášek, cvičení).
3. Pomocí operáru `X-Validation` odhadněte přesnost vašich klasifikátorů a vyberte „nejlepší“, resp. ten o kterém předpokládáte že bude nejlepší. Věnujte pozornost volbě evaluační metriky.
4. Připravte testovací podproces, který:
  - 4.1 stejným způsobem jako v 1) načte „budoucí“ testovací data a otaguje atributy
  - 4.2 použije vámi vybraný naučený model z 2) a použije je na načtená data a vyčíslí klasifikační přesnost, tj. `accuracy`

- ▶ Za odevzdaný protokol můžete získat až 2 body. V protokolu bude:
  - ▶ Krátký popis klasifikační úlohy, tj. co klasifikujeme, jaké typy atributů máte, resp. jste otagovali.
  - ▶ Popis validačního protokolu
  - ▶ Nejlepší model, jeho interpretace, popř. vizualizace, a odhad jeho přesnosti.
- ▶ Odevzdejte váš RM proces
- ▶ Klasifikátor vámi zvolený jako nejlepší bude ohodnocen na (vám dosud neznámých) testovacích datech.
- ▶ Kromě bodů za protokol tedy můžete dále získat body odpovídající procentům přesnosti vašeho klasifikátoru:
  - ▶  $\langle 85, 99 \rangle$  ... 6 b.
  - ▶  $\langle 80, 85 \rangle$  ... 5 b.
  - ▶  $\langle 75, 80 \rangle$  ... 4 b.
  - ▶  $\langle 70, 75 \rangle$  ... 3 b.
  - ▶  $\langle 60, 70 \rangle$  ... 2 b.
  - ▶  $\langle 55, 60 \rangle$  ... 1 b.
- ▶ **Deadline je až za 2 týdny od zadání!!!**