# IB031 Úvod do strojového učení
# Pokyny k projektům

20. února 2023

## 1 Zadání projektu

Projekty se vypracovávají ve skupinách po max. 3 studentech a všichni studenti musí být ze stejné seminární skupiny. Každý projekt vymezuje dataset a případné další podmínky na vypracování, ve zbytku rozhodnutí mate úplnou volnost. Typicky každý student ve skupině provede následující.

1. zapojí se do vypracování explorační analýzy dat
2. zapojí se do předzpracování dat
3. vybere si jeden konkrétní model pro strojové učení
4. sepíše krátké obecné vysvětlení fungování svého modelu
5. natrénuje svůj vybraný model na předzpracovaném datasetu
6. interpretuje svůj natrénovaný model
7. vyhodnotí svůj natrénovaný model na datasetu
8. zapojí se do sepisování krátkého shrnutí výsledků z porovnání

## 2 Popis jednotlivých úkolů

### 2.1 Explorační analýza

Prozkoumejte dataset, tj. podívejte se, kolik je v datasetu dat a jaká jsou, kolik a jakého typu jsou hodnoty jednotlivých sloupců, jak spolu jednotlivé položky korelují. Ke každému datsetu je vždy i úvodní komentář, který shrnuje, co je v datech a případně i jak byla dat sbírána. Tento komentář i případné odkazy v něm uvedené je dobré přečíst před započetím analýzy. Výstup této analýzy budou typicky tabulky a grafy. Svá pozorování okomentujte pár větami.

### 2.2 Předzpracování

Připravte dataset tak, aby se na něm mohly učit jednotlivé modely. Do tohoto kroku patří veškerá manipulace s daty. Může se jednat např. o převody datových typů (např. na datetime), práce s chybějícími hodnotami, škálování a normalizace, feature selection, feature extraction, rozdělení na trénovací a testovací množinu, resampling, přidání dalších externích dat a další. Ne všechny vyjmenované věci je potřeba udělat, záleží na datech a modelech. Když už se k nějakému předzpracování rozhodnete, stručně okomentuje co a proč děláte.

### 2.3 Výběr modelu

Výběr model je omezený pouze typem úlohy (regrese, klasifikace, shlukování, detekce anomálií) vašeho konkrétního zadání, jinak máte volný výběr. Můžete zvolit model z knihovny `scikit-learn`, ale klidně můžete použít jinou knihovnu nebo si dokonce implementovat vlastní model.

### 2.4 Vysvětlení modelu

Vysvětlení modelu bude v rozsahu cca dvou odstavců. Cílem je představit techniku a stručně a výstižně popsat její fungování. Můžete předpokládat, že čtenář má základní povědomí o strojovém učení, ale nezná váš konkrétní model. Ve vysvětlení se zaměřte na popis učení modelu (co se počítá během učení, co je výsledkem učení, ...).

### 2.5 Natrénování modelu

Natrénujte svůj vybraný model na předzpracovaném datasetu. Přesný způsob trénování je na vás, ale měl by být porovnatelný napříč modely. Tedy není dobré, aby se každý model trénoval na jiné podmnožině dat. Model se nemusí učit na všech sloupcích, může si vybrat jen jejich podmnožinu. Součástí trénování bude i ladění hyperparametrů.

### 2.6 Interpretace modelu

Interpretujte svůj vybraný a natrénovaný model. Konkrétní interpretace bude záviset od zvoleného modelu, ale cílem je vysvětlit, co se model naučil a proč se naučil zrovna toto. Interpretace může zahrnovat analýzu naučených parametrů, význam a vliv jednotlivých sloupců na predikce modelu a analýzu citlivosti na hodnoty v jednotlivých sloupcích.

### 2.7 Vyhodnocení modelu

Na svém datasetu proveďte vyhodnocení modelu pomocí vhodně zvolené míry v závislosti na řešené úloze (regrese, klasifikace, shlukování, detekce anomálií). Volbu míry stručně zdůvodněte a okomentujte, co měří. Vyhodnocení bude obsahovat také srovnání s „baseline“ modelem, který bude problém řešit na naivním způsobem (např. konstantní hodnota, jednoduchá statistika, náhoda, ...)

### 2.8 Shrnutí výsledků

V pár větách srovnejte výsledky jednotlivých modelů a shrňte výsledky z vyhodnocení. Zejména zajímavá jsou zjištění, který model funguje nejlépe a zdůvodnění proč. Stejně tak zda je některý z modelů robustnější při volbě různých rozdělení dat na trénovací a testovací množinu.

## 3 Odevzdání

Své výsledky budete dvakrát prezentovat na cvičení a zároveň odevzdáte finální verzi do odevzdávárny v ISu.

První prezentace proběhne na 8. cvičení (týden 3. 4. až 6. 4.). Během této prezentace krátce představíte dataset, výsledky explorační analýzy a kam jste se s vypracováním projektu dostali. Zároveň byste měli mít rozmyšlené, jak chcete dataset předzpracovat a jaké modely budete učit. Prezentace bude opravdu krátká, pouze několik minut. Přesné požadavky na formu prezentace si určí každý cvičící.

Druhá prezentace proběhne na posledním 13. cvičení (týden 8. 5. až 13. 5., 15. 5. náhrada pondělní skupiny). Na této prezentaci už představíte kompletní řešení svého projektu, vaše výsledky a pozorování. Zároveň každý z členů týmu řekne jednu věc, kterou si z projektu odnesl. Prezentace bude maximálně na 5 minut.

Hotový projekt odevzdáte jako jediný zip nebo tar.gz archiv do odevzdávárny v ISu. Archiv bude obsahovat:

- jediný IPython notebook, kde bude veškerý kód proložený komentáři a popisnými texty,
- PDF dokument vygenerovaný z vyhodnoceného notebooku (tzn. včetně výstupů),
- případné další zdroje dat, které jsou používali.

Deadline pro odevzdání projektu je **konec 13. cvičení**.

# 4 Hodnocení

Níže je rubrika shrnující co a jak budu na projektech hodnotit.

| popis požadavku | body |
|---|---|
| explorační analýzu datasetu | 2 |
| vhodné předzpracování dat podle typu řešené úlohy a vybraných modelů | 2 |
| pokročilé předzpracování jako extrakce rysů nebo externí data | 2 |
| odůvodnění a okomentování použitých technik předzpracování dat | 2 |
| popis fungování pro každý vybraný model | 2 |
| všechny modely natrénované na datasetu | 1 |
| vhodná volba parametrů modelů a jejich ladění | 3 |
| interpretace naučených modelů | 3 |
| vyhodnocení modelů pomocí několika vhodně zvolených mír | 2 |
| porovnání modelů s naivním ,,baseline“ modelem/přístupem | 2 |
| krátké shrnutí výsledků a pozorování | 2 |
| vysvětlující komentáře dokumentující jednotlivá rozhodnutí v projektu | 2 |
| správná metodologie učení a vyhodnocování modelů | 5 |