

Umělá inteligence - semestrální projekt

Analýza dat a aplikace vybrané metody strojového učení

Jména studentů (práce ve dvojici), e-maily

21. listopadu 2025

Pokyny k práci

Tento dokument slouží jako závazná šablona reportu. Vyplňte jednotlivé sekce v uvedeném rozsahu.

Práce se řeší *ve dvojcích*.

Datové sady musí pocházet z *reálných nebo otevřených zdrojů* (ne knihovní benchmarky jako Iris, MNIST, Breast Cancer, Wine, apod.).

Je povoleno využívat nástroje generativní AI (např. ChatGPT, Copilot). V závěru práce musí být uvedeno samostatné prohlášení popisující, k čemu byla AI využita (pouze pokud byla skutečně využita) (např. konzultace postupu, generování části kódu, kontrola textu). Využití AI však nesmí nahrazovat vlastní porozumění — každý student musí být schopen při ústním zkoušení detailně vysvětlit použitou metodu, její parametry, datovou přípravu, kód i interpretaci výsledků. Nesplnění této podmínky může ovlivnit hodnocení práce i ústní část zkoušky.

Celkový doporučený rozsah textu je *8–12 stran bez příloh*.

1 Možné otevřené zdroje dat

- Kaggle Datasets: <https://www.kaggle.com/datasets>

- **Google Dataset Search:** <https://datasetsearch.research.google.com>
- **UCI Machine Learning Repository** (vyjma známých benchmarků):
<https://archive.ics.uci.edu>
- **OpenML** (vyhýbat se klasickým učebnicovým datasetům):
<https://www.openml.org>
- **Data.gov (USA):** <https://www.data.gov>
- **EU Open Data Portal:** <https://data.europa.eu>
- **České otevřené datové portály:**
 - <https://data.gov.cz>
 - <https://opendata.praha.eu>
 - <https://opendata.brno.cz>
- **OpenWeather, NOAA a další meteorologická data**
- **GitHub repozitáře s raw daty** (např. data ze senzorů, logů apod.)
- **Data z vlastního okolí** (pokud nejsou osobní – např. logy z chytré domácnosti, sportovní senzorika apod.)

2 Úvod (0.5–1 strana)

- Stručné představení vybrané metody a problému.
- Odůvodnění výběru dat a jejich relevance.
- Struktura práce.

3 Popis datové sady (1–2 strany)

- Původ dat (zdroj + odkaz).
- Typy atributů, velikost, časové pokrytí (pokud relevantní).

- Exploratorní analýza: statistiky, distribuce, základní grafy.
- Identifikace problémů datasetu (nevyváženosť, šum, korelace, chybějící hodnoty, outliersy...)
- Případně profil vytvořený pomocí `ydata-profiling` (přiložte do přílohy).

4 Formulace cíle (0.5 strany)

- Přesná formulace predikčního/klasifikačního cíle.
- Typ úlohy a zdůvodnění výběru metody.

5 Předzpracování dat (1–2 strany)

- Vyřešené datové problémy a popis postupů.
- Kroky předzpracování (čištění, imputace, škálování, výběr atributů, transformace...).
- Zdůvodnění vzhledem k vybrané metodě.

6 Metoda strojového učení (1–2 strany)

- Princip a stručný matematický popis.
- Popis parametrů a jejich nastavení.
- U složitějších metod popis architektury nebo strategie ladění.

7 Experimenty a výsledky (2–3 strany)

- Oddělení trénovacích, validačních a testovacích dat.
- Baseline, hyperparametrické ladění, cross-validation (pokud vhodné).
- Vybrané metriky:

- regrese: MSE, RMSE, MAE, R^2 , Adjusted R^2 ,
 - klasifikace: accuracy, precision, recall, F1, AUC-ROC, confusion matrix.
- Grafy: průběhy učení, ROC, predikce vs. realita.
 - Interpretace výsledků – co znamenají a proč jsou takové.

8 Diskuse (1 strana)

- Zhodnocení silných a slabých stránek výsledku.
- Co fungovalo a co ne, a proč.
- Možné alternativy a budoucí vylepšení.

9 Závěr (0.5–1 strana)

- Shrnutí postupu a klíčových výsledků.
- Vyjádření, zda byl cíl splněn a proč.
- Krátký přehled naučených poznatků.

Prohlášení o využití generativní AI (pokud relevantní)

Tato sekce je povinná pouze v případě, že byla při zpracování projektu využita generativní umělá inteligence. **Vzorové prohlášení:**

V rámci vypracování této semestrální práce jsme využili nástroje generativní umělé inteligence (např. ChatGPT, Gemini, Mistral, GitHub Copilot) k následujícím účelům:

- konzultace teoretického postupu a ověření správnosti vysvětlení,
- návrh části kódu / kontrola kódu,
- jazyková korektura textu.

Veškerým použitým postupům, kódu i interpretaci výsledků rozumíme a jsme schopni je samostatně vysvětlit v rámci ústního zkoušení. Současně bereme na vědomí, že plně odpovídáme za správnost obsahu, výpočtů, kódu i závěrů uvedených v této práci.

Přílohy

- Výstupy profilování dat.
- Vybrané části kódu.
- Dodatečné grafy a tabulky.