Umělá inteligence - semestrální projekt

Analýza dat a aplikace vybrané metody strojového učení

Ondřej Chalupa
chaluon3@student.cvut.cz

Matěj Hrdlička
hrdlima8@student.cvut.cz

1. ledna 2026

Pokyny k práci

Tento dokument slouží jako závazná šablona reportu. Vyplňte jednotlivé sekce v uvedeném rozsahu.
Práce se řeší *ve dvojicích*.

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

- **Stručné představení vybrané metody a problému**

  Tato práce se zabývá aplikací strojového učení, konkrétně metody vícevrstvého perceptronu, na oblast analýzy dopravních přestupků v hlavním městě Praze. Data o dopravních přestupcích nám přišla zajímavá. Zjistili jsme, že záznamy o dopravních přestupcích v Praze přidává buď Policie České republiky, nebo Městská policie Praha. Chtěli jsme zjistit, jestli je možné na základě toho, že víme, kdy, kde a jaký přestupek byl spáchán, určit, zda přestupek řešila Policie České republiky, nebo Městská policie Praha.

- **Odůvodnění výběru dat a jejich relevance**

  Data jsme našli na portálu Open Data hlavního města Prahy. Na základě toho, že nám data o dopravních přestupcích přišla zajímavá, jsme si vymysleli problém, který s nimi můžeme vyřešit.

- **Struktura práce**

  Práce se skládá z několika souborů. První soubor `data_process.py` obsahuje kód, který připraví data pro trénování nebo pro použití modelu na predikci. Soubor `train.py` obsahuje kód pro vytrénování modelu na datech a `predict.py` použije model pro predikci a určí jeho přesnost.

3 Popis datové sady

- **Původ dat (zdroj a odkazy)**

  Data pocházejí z portálu Open Data hlavního města Prahy a jsou vedena pod názvem *Dopravní přestupky MHMP*. Pro trénování modelu byla zvolena data z roku 2023, zatímco pro validaci a testování byly využity sady z let 2022 a 2024.

  Odkazy na zdroje:

  - Rok 2022: Dopravní přestupky MHMP 2022
  - Rok 2023: Dopravní přestupky MHMP 2023
  - Rok 2024: Dopravní přestupky MHMP 2024

- **Typy atributů, velikost a časové pokrytí**

  Datasety pokrývají období od ledna 2022 do června 2024. Sady z let 2022 a 2023 obsahují záznamy za celých 12 měsíců, dataset z roku 2024 obsahuje data pouze za první polovinu roku. Celkový přehled uvádí Tabulka 1.

Tabulka 1: Přehled počtu záznamů v jednotlivých letech

Dataset	**Časové pokrytí**	**Počet záznamů**
MHMP 2022	Leden – Prosinec	846 832
MHMP 2023	Leden – Prosinec	946 701
MHMP 2024	Leden – Červen	372 084
Celkem		**2 165 617**

Každý záznam reprezentuje jeden evidovaný přestupek. Mezi klíčové atributy patří:

Časové údaje: Datum (*2022-04-15*) a Čas (*09:14*).

Lokalizace: Místo (*Lohniského 4*) a Čtvrť (*Praha 5*).

Subjekty: Oznamovatel (*MPP/PČR*), Státní příslušnost (*CZ*), Značka vozidla.

Právní kvalifikace: Porušený zákon (např. *§ 125d zák. č. 361/2000 Sb.*).

- **Exploratorní analýza: statistiky, distribuce a grafy**

  Analýza provedená na datech z roku 2023 odhalila následující charakteristiky:

  - **Distribuční poměr oznamovatelů:** Data vykazují silnou převahu záznamů od Městské policie (MPP), která tvoří přibližně 91,6 % datasetu (866 709 záznamů). Policie ČR (PČR) se podílí zbylými 8,4 %.
  - **Geografické rozložení:** Přestupky jsou nejčastěji evidovány v obvodech **Praha 4** (207 400 záznamů) a **Praha 6** (167 246 záznamů). Centrum města (Praha 1) je v četnosti až čtvrté.
  - **Typy přestupků:** Nejčastějším porušením je § 125c odst. 1 písm. k) (nedovolené zastavení/stání). Významnou část tvoří i překročení rychlosti do 20 km/h v obci.
  - **Struktura chybějících dat:** Atributy jako Datum či Oznamovatel jsou vyplněny vždy. Nejvíce chybějících hodnot vykazuje:
    - tovární značka vozidla - 12.4%
    - čas - 4.6%
    - státní příslušnost - 2.4%

- **Identifikace problémů datasetu (nevyváženost, šum, chybějící hodnoty)**

  Během přípravy dat byly identifikovány tyto problémy:

  - **Vysoká nevyváženost tříd:** Poměr MPP vůči PČR je cca 11:1.
  - **Nekonzistentní označování:** Atribut právní kvalifikace obsahuje duplicitní významy (např. dlouhý zápis `§ 125c odst. 1...` vs. zkrácený `§ 125c/1k...`).
  - **Chybějící hodnoty:** Atribut *TOVZN* (Tovární značka) chybí u 12,4 % záznamů, atribut *CASSK* (Čas) u 4,5 % záznamů.
  - **Nestrukturovaný text:** Atribut *MISTOSK* obsahuje volný text s překlepy a nejednotným formátováním (např.: *Maltézské náměstí 16, Praha 1, u domu*)

- **Profil dat (ydata-profiling)**

  Detailní automaticky generovaný profil datové sady vytvořený nástrojem `ydata-profiling` je přiložen v Příloze A.

4 Formulace cíle (0.5 strany)

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

5 Předzpracování dat

Tato kapitola popisuje technické kroky vedoucí k transformaci surových dat do podoby vhodné pro trénování klasifikačního modelu. Proces zahrnoval čištění dat, extrakci nových příznaků (feature engineering) a finální selekci atributů na základě jejich statistické významnosti.

- **Vyřešené datové problémy a popis postupů**

  Na základě závěrů exploratorní analýzy byly adresovány následující nedostatky v kvalitě a struktuře dat:

  - **Nestrukturovaný text u lokace:** Atribut *MISTOSK* původně obsahoval volný text kombinující názvy ulic, čísla popisná a další upřesnění (např. „před domem č.p. 5“). Pro potřeby modelování bylo nutné tento text kategorizovat. Pomocí regulárních výrazů a vyhledávání klíčových slov (např. „tunel“, „náměstí“, „Evropská“, „spojka“) byly záznamy rozděleny do čtyř obecných tříd: *náměstí, tunel, hlavní ulice* a *ostatní*.
  - **Nekonzistentní zápis zákonů:** Atribut *ZAKON* vykazoval vysokou variabilitu v zápisu totožných přestupků (rozdílné mezery, zkratky odstavců). Byla provedena normalizace textu, kdy byly pomocí vzorů (regex) sjednoceny varianty zápisu na základní tvar paragrafu a odstavce (např. unifikace na `§125c/1k`).
  - **Chybějící a odlehlé hodnoty:** U atributu tovární značky (*TOVZN*) byla chybějící data (cca 12 %) imputována zástupnou hodnotou `UNSPECIFIED`. V případě chybějící státní příslušnosti byla použita

hodnota `UNKNOWN`. Pokud u záznamu chyběl čas nebo byl v chybném formátu, byla hodina přestupku nastavena na indikátor `-1`.

- **Kroky předzpracování (transformace a tvorba příznaků)**

  Cílem transformací bylo snížit dimenzionalitu dat a zachytit obecnější trendy namísto specifických detailů, které by mohly vést k přeučení modelu.

  Časové atributy: Ze spojitého času byla vytvořena sada kategorických příznaků s cílem zachytit cyklické chování a trendy. Konkrétně byly derivovány:
  - **Měsíc** (1–12).
  - **Hodina** (0–11), pro chybějící čas nahrazeno hodnotou `-1`.
  - **Typ dne** (binární rozdělení na pracovní den a víkend).
  - **Roční období** (jaro, léto, podzim, zima).
  - **Denní doba** (ráno, odpoledne, večer, noc); v případě chybějícího času nastaveno jako `none`.

  Typ vozidla (`CAR_TYPE`): Atribut značky vozidla byl diskretizován dle předem vytvořeného seznamu nejfrekventovanějších značek (např. *Škoda, Volvo, Hyundai*). Značky s nižším výskytem, které nebyly součástí seznamu, byly sloučeny do kategorie `OTHER`. Záznamy bez uvedené značky byly označeny jako `UNSPECIFIED`.

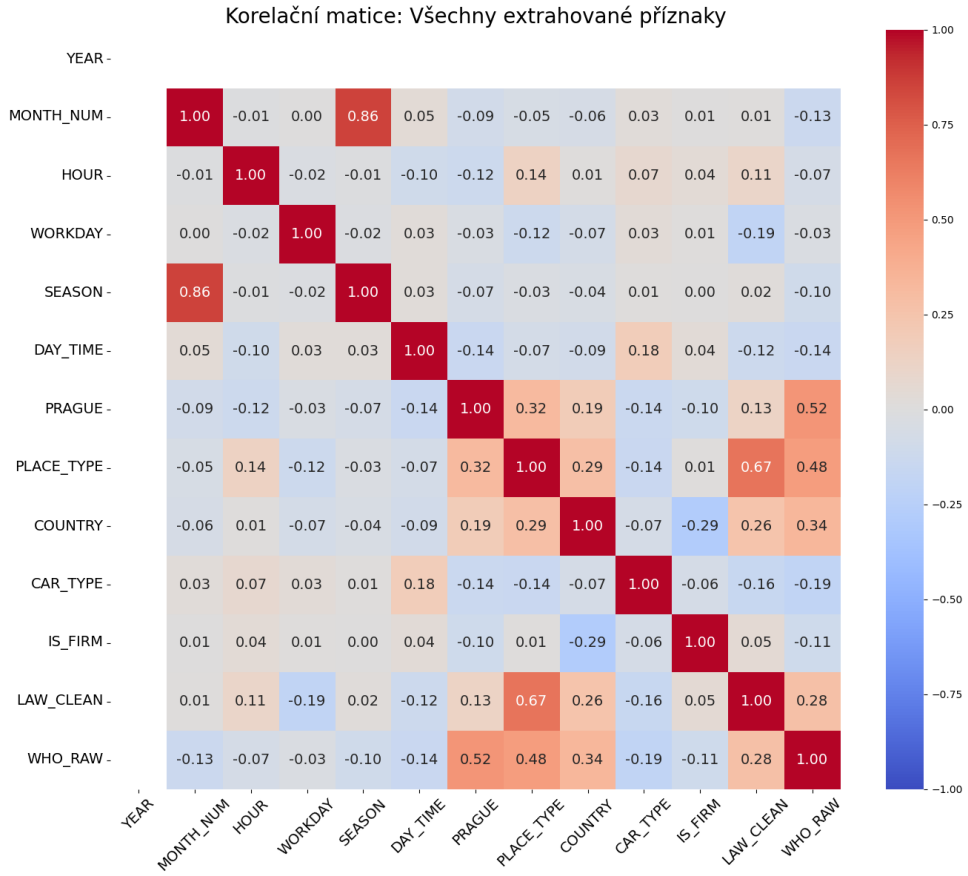
- **Zdůvodnění výběru atributů**

  Finální sada atributů pro trénování byla zvolena na základě analýzy korelací a iterativního testování výkonu modelu. Byla vypočtena korelační matice (viz Obrázek 1), která vizualizuje vztahy mezi nově vytvořenými příznaky a cílovou proměnnou *OZNAM*.

  Na základě této analýzy byly pro finální trénink modelu vybrány atributy: `SEASON`, `DAY_TIME`, `PRAGUE`, `PLACE_TYPE`, `CAR_TYPE`, `LAW_CLEAN` a `IS_FIRM`.

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.



Obrázek 1: Korelační matice finálních atributů (vizualizace závislosti mezi extrahovanými příznaky)

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.