Umělá inteligence - semestrální projekt

Analýza dat a aplikace vybrané metody strojového učení

Ondřej Chalupa
chaluon3@student.cvut.cz

Matěj Hrdlička
hrdlima8@student.cvut.cz

12. 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. dodatečně jsme pak použili data z roků 2021 a 2020 pro další testování.

Odkazy na zdroje:

- Rok 2024: Dopravní přestupky MHMP 2024
- Rok 2023: Dopravní přestupky MHMP 2023
- Rok 2022: Dopravní přestupky MHMP 2022
- Rok 2022: Dopravní přestupky MHMP 2021
- Rok 2022: Dopravní přestupky MHMP 2020

- **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 2020	Leden – Prosinec	562 348
MHMP 2021	Leden – Prosinec	665 824
MHMP 2022	Leden – Prosinec	846 832
MHMP 2023	Leden – Prosinec	946 701
MHMP 2024	Leden – Červen	372 084
Celkem		**3 393 789**

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 generovaný profil datové sady je přiložen v příloze.

4 Formulace cíle

- **Přesná formulace predikčního/klasifikačního cíle.**

  Hlavním cílem této semestrální práce je navrhnout, natrénovat a vyhodnotit model strojového učení, který bude schopen rozlišit, kdo zaznamenal dopravní přestupek na území hlavního města Prahy. Konkrétně se jedná o rozlišení mezi Městská policie Praha a Policie České republiky.

- **Typ úlohy a zdůvodnění výběru metody.**

  Z hlediska strojového učení se jedná o úlohu učení s učitelem, konkrétně o binární klasifikaci. Pro řešení úlohy byla zvolena metoda Vícevrstvého perceptronu. Předpokládáme že vztahy mezi vstupními daty a cílovou třídou jsou komplikované ale je v nich možné najít nějaký vzor, na základě kterého se policie pohybuje. Například ulice "Evropská"může implikovat PČR v noci (měření rychlosti), ale MPP ve dne (parkování). proto jsme zvolili metodu MLP která by tyto vztahy mohla najít.

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:

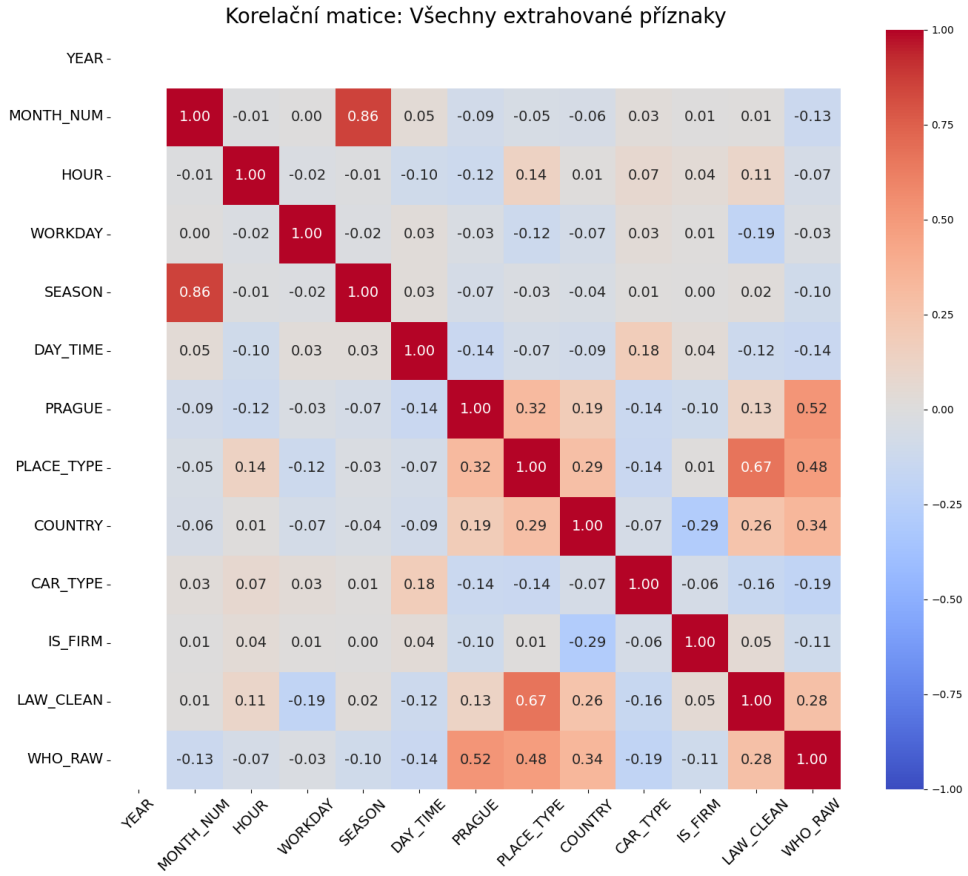
  - **Měsíc** (1–12).
  - **Hodina** (0–11), pro chybějící čas nahrazeno hodnotou `-1`.
  - **Typ dne** (binární rozdělení na pracovní den a víkend).
  - **Roční období** (jaro, léto, podzim, zima).
  - **Denní doba** (ráno, odpoledne, večer, noc); v případě chybějícího času nastaveno jako `none`.

  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`.



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

6 Metoda strojového učení

Pro řešení úlohy byla zvolena metoda Vícevrstvého perceptronu (MLP). Jedná se o dopřednou umělou neuronovou síť (Feedforward Neural Network), která se skládá z jedné vstupní vrstvy, jedné či více skrytých vrstev a jedné výstupní vrstvy.

Základním stavebním kamenem je umělý neuron, který provádí vážený součet svých vstupů, přičítá k němu prahovou hodnotu (bias) a výsledek transformuje pomocí nelineární aktivační funkce. Díky vrstvení těchto neuronů a použití nelinearit je MLP schopen aproximovat i velmi složité rozhodovací hranice v mnohadimenzionálním prostoru příznaků, což je pro naše kategorická data klíčové.

My jsme pro vytvoření modelu jsme použili knihovnu `scikit-learn` a třídu `MLPClassifier`. Na základě experimentů s počtem neuronů a skrytých vrstev byla zvolena architektura s dvěmi skrytými vrstvami. První skrytá vrstva s 50 neurony a druhá s 25 neurony. Aktivační funkce v obou vrstvách je ReLu.

7 Experimenty a výsledky

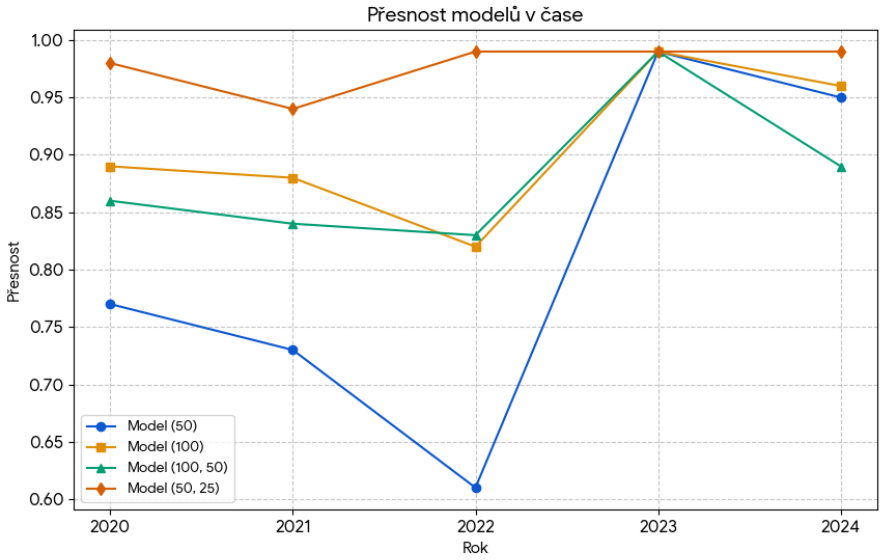
Při tréninku jsme použili data z roku 2023. Ty byla rozdělené na trénovací (80 % dat) a testovací (20 % dat). Následně jsme pak ještě natrénovaný model testovali na datech z jiných let. Jako první jsme vytvořili model se dvěma skrytými vrstvami se 100 neuronů v první a 50 neurony v druhé vrstvě. Tento model jsme používali při úpravě dat.

Při analýze dat jsme si například všimli, že v datech často chybí značka vozidla a čas. Když jsme se podívali na to, kdo udělal záznam v řádcích s chybějícími hodnotami, tak jsme zjistili, že se z většiny jednalo o PČR. Například když chyběla značka auta, tak se z 59,4 % jednalo o PČR, Když si vezmeme že rozložení záznamů PČR a MPP je v datech 1:11, tak se zdá že absence údajů je vlastně dobrý ukazatel toho, zda se jedná o PČR či nikoliv. Přidali jsme sloupce `MISSING_TIME` a `MISSING_BRAND`. Výsledky ale byly horší než když jsme jen do sloupců `HOUR` dali hodnotu -1 a `CAR_TYPE` dali hodnotu `UNSPECIFIED`. Řádky s chybějícím časem a značkou auta jsme i zkusili odstranit a natrénovat model na datech bez nich. Model pak na datech pro rok 2023 a 2024 měl přesnost 99.9 %. Pro rok 2022 ale jen 92 %.

Následně jsme zkusili měnit počet neuronů a vrstev a rozšířili jsme testovací data ještě o roky 2021 a 2020. Zkusili jsme natrénovat model s jednou skrytou vrstvou o 50 a poté o 100 neuronech. Následně modely se dvěma vrstvami, jeden model o 50 neuronech v první skryté vrstvě a 25 v druhé a jeden s původními 100 neurony v první vrstvě a 50 v druhé.

Všechny modely měly přesnost 99 % při určování záznamů od MPP ale při určování PČR byla přesnost u některých modelů výrazně nižší (nejhorší 61 % měl model s jednou skrytou vrstvou na datech z roku 2022). Graf Přesnosti na třídě PČR vidíme na obrázku 2 níže.

nejlepších výsledků dosáhl model s 50 neurony v první a 25 neurony v druhé vrstvě.



Obrázek 2: Graf přesnosti na třídě PČR

Tabulka 2: Přesnost modelů

Rok	**Model (50)**	**Model (100)**	**Model (100, 50)**	**Model (50, 25)**
2024	0.95	0.96	0.89	**0.99**
2023 - trénink	0.99	0.99	0.99	**0.99**
2022	0.61	0.82	0.83	**0.99**
2021	0.73	0.88	0.84	**0.94**
2020	0.77	0.89	0.86	**0.98**

Tabulka 3: výsledky modelu(50,25)

Rok	**Precision (PČR)**	**Recall (PČR)**	**F1-Score (PČR)**	**přesnost**
2024	0.9707	0.9460	0.9582	0.9950
2023 (Trénink)	0.9917	0.9593	0.9753	0.9959
2022	0.9329	0.9306	0.9317	0.9899
2021	0.9096	0.6135	0.7372	0.9443
2020	0.9443	0.9019	0.9226	0.9846

8 Diskuse

V této části rozebíráme, proč náš model funguje, jaké má výsledky a na co jsme během práce přišli.

- **Zhodnocení silných a slabých stránek výsledku.**

  Největším problémem datasetu byla nevyváženost počtu záznamů PČR a MPP (11:1), což by mohlo vést k tomu, že by se model nenaučil specifické odlišnosti záznamů PČR, ale naučil by se bezhlavě hádat MPP s vizí úspěšnosti nad 90%. S tímto problémem si ale model se dvěma vrstvami (50 a 20 neuronů) poradil, což lze pozorovat na základě těchto metrik pro rok 2024:

  - **Precision (0,97):** Tato metrika říká, že v případech kde model označil záznam jako PČR, tak se v 97% trefil. Pokud by se model snažil násilně vyhledávat PČR, mohlo by vzniknout více případů FP (false positive), k tomu však nedošlo.
  - **Recall (0,95):** Recall 95% pro třídu PČR znamená, že z celkového počtu záznamů PČR v datasetu jich model objevil 95%. Pokud by model raději označoval MPP kvůli jejich velkému zastoupení, bylo by číslo nižší.
  - **F1-Score (0,96):** Parametr F1-Score je harmonickým průměrem parametrů Precision a Recall. Nízká hodnota by mohla naznačovat náchylnost vůči datové nevyváženosti. Tento výsledek ale ukazuje, že náš model je proti ní velmi naopak velmi odolný.

  Tyto hodnoty ukazují, že se nám podařilo vyřešit problém s nevyvážeností dat. Model nehádá jen naslepo většinovou třídu, ale skutečně se naučil poznat specifika státní policie.

  Další silnou stránkou modelu je, že je použitelný na záznamech ve velkém časovém rozpětí (rok 2020 až polovina roku 2024). To ukazuje že model zachytil hlubší závislosti než jen nějaké krátkodobé trendy jednoho roku. Naopak problémovou stránkou je, že model pravděpodobně velmi spoléhá pro rozpoznání PČR na neúplnost dat (PČR často nechává některá pole prázdná). Pokud by PČR změnila své návyky a začala lépe udržovat záznamy, model už by možná nebyl použitelný.

- **Co fungovalo a co ne, a proč.**

  Klíčem k úspěchu bylo nemazat neúplné řádky. Původně jsme například chtěli data s chybějícím časem vyhodit, ale ukázalo se, že právě prázdná políčka nejlépe prozrazují PČR. Také se ukázalo, že jednoduchá síť na tento úkol nestačí. Model s jednou vrstvou dělal chyby. Až architektura se dvěma vrstvami (50 a 25 neuronů) dokázala zachytit složitější vztahy v datech.

- **Možné alternativy a budoucí vylepšení.**

  Nevýhodou neuronové sítě je že nevidíme přesně, proč se rozhodla tak jak se rozhodla. Není proto možné z výsledků s jistotou vyčíst nějaký patern pro rozdělení. Kdybychom chtěli větší transparentnost, mohli bychom příště zkusit **Rozhodovací stromy (Decision Trees)**. Ty by nám formou diagramu přesně ukázaly pravidla (např. „pokud chybí čas → je to PČR“), což by bylo pro vysvětlování chování policie srozumitelnější.

  Dalším vylepšením by mohlo být důkladnější zpracování dat. Například místo je určeno jednoduchým vyhledáváním klíčových slov jako je "tunel", "spojka"atd. Důmyslnějším rozdělením by třeba bylo možné lokaci rozdělit lépe a vyčíst nějaké další příznaky.

9 Závěr

- **Shrnutí výsledků a výkonnosti modelu.**

  Úkolem práce bylo zpracovat data a natrénovat model schopný automaticky klasifikovat původce přestupku a rozlišit mezi Městskou policií a Policií ČR. K tomuto účelu jsme využili vícevrstvou neuronovou síť (MLP). Jako nejefektivnější se po sérii experimentů ukázal model se dvěma skrytými vrstvami o velikosti 50 a 25 neuronů.

  Tento model konzistentně dosahuje na datech od roku 2020 po rok 2024 přesnost nad 98%. Vysoká hodnota F1-skóre (0,95) potvrzuje, že model je spolehlivý i při detekci menšinové třídy PČR a netrpí tendencí ignorovat méně časté záznamy.

- **Splnění cílů práce.**

  Cíl práce lze považovat za úspěšně splněný. Podařilo se vytvořit plně funkční klasifikátor, který s vysokou jistotou identifikuje, zda záznam pochází od městské či státní policie.

Kritickým bodem celého řešení bylo vypořádání se s výraznou nevyvážeností datasetu, kde poměr záznamů MPP ku PČR činil 11:1. Výsledné metriky potvrzují, že zvolená architektura a metoda zpracování dat tento nepoměr úspěšně kompenzovaly, aniž by došlo ke zkreslení predikcí ve prospěch dominantní třídy.

- **Získané poznatky a přínos.**

  Během vývoje jsme ověřili, že při práci s reálnými daty může být i absence informace klíčovým příznakem – v našem případě se ukázalo, že chybějící údaje v záznamech jsou silným indikátorem pro PČR.

  Zároveň se v praxi potvrdilo, že u silně nevyvážených dat nelze spoléhat pouze na celkovou přesnost modelu. Pro objektivní vyhodnocení kvality sítě bylo nezbytné sledovat metriky Precision a Recall, které nám poskytly jistotu, že model skutečně chápe strukturu dat a neuchyluje se k prostému statistickému hádání.

Prohlášení o využití generativní AI (pokud relevantní)

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

- **Výstup profilování dat** (ydata-profiling): `vystup_report.html`
- **Vybrané části kódu**: `data_process.py`, `train.py`, `predict.py`