# Průběžná zpráva k projektu dne 31. 12. 2024

V projektu bylo přislíbeno otestovat metaheuristické knihovny na modelech strojového učení. Z předběžného seznamu byly postupně vyřazeny všechny knihovny (PyGMO, MEALPY, DEAPa Opytimizer) a byly nahrazeny knihovnami **Hyperopt**, **Optuna** a **Skopt**. Důvodem byly problémy s implementací jednotlivých knihoven a špatných výsledků. Nakonec jsme se přiklonili ke knihovnám přizpůsobeným k optimalizaci parametrů strojového učení.

Script s porovnáním výsledků se nachází v *01\_experiment\_ML\_MH.py*. Výsledky pak jsou v souborech *experiment\_ML\_MH\_testing.csv* a *experiment\_ML\_MH.csv*. První zmíněný je testovací mód a v druhém už jsou výsledky experimentu.

Ve výpočetním experimentu byly použity tři volně dostupné knihovny (**California Housing**, **Diabetes**, **Wine**) určené k predikci cílové proměnné pomocí vstupních proměnných. Jako vstupní proměnné bylo použito všech dostupných proměnných, aby se jednotlivé algoritmy otestovaly pro extrémní podmínky, do kterých budou nasazeny. Následně je na jednotlivých algoritmech a optimalizaci parametrů, aby se našly ideální proporce pro vytvoření ideálního modelu.

Z řady algoritmů strojového učení byly použity modely RandomForestRegressor (**RFR**), GradientBoostingRegressor (**GBR**) a XGBRegressor (**XGBoost**).

V souboru *Experimental\_results.pdf* nalezneme experimentální výsledky již přehledně zapsané. Jedná se o tabulku naformátovanou v prostředí LaTeX pomocí scriptu *df\_to\_LaTeX.py*. Hodnotu MAE (Mean Absolute Error) chceme minimalizovat a hodnotu R2 naopak maximalizovat. Co se týče výpočetního času pro jednotlivé algoritmy, tak zde opět pro vyhodnocení projektu budeme hledat co nejnižší hodnoty.

Závěry výpočetních experimentů se nachází na další stránce.

Dále byly dopředu vytvořeny scripty pro zisk dat pro další část projektu. Jedná se o složky processing, kde se využívá scriptů:

* *Get\_coordinates\_data.py* -> používáno pro získání GEO koordinárů (středy okresů České republiky)
* *Get\_OpenMeteo.py* -> používáno pro zisk historických dat o počasí ze stránky *OpenMeteo.com*. Zároveň lze z této stránky získat i predikční data a testovat model pro případné nasazení v reálné situaci.

Stažená data pak mohou být nalezena ve složce *output\_data*.

Data týkající se plynárenského stavu v České republice pak zpracovává script *00\_ote\_year\_report.py*, který získává data od OTE (Operátor trhu s elektřinou). Zároveň lze script i jednoduše upravit pro stahování dat o elektřině.

## Závěr experimentu:

Dataset **California Housing**:

* Knihovny Hyperopt a Skopt dosahují velmi dobrých výsledků. Nicméně knihovna **Hyperopt** dosáhla na lepší výsledky.
* Vzhledem k délce výpočtu a dosaženého výsledku pak jednoznačně nejlépe vychází kombinace **XGBoost** a **Hyperopt**.

Dataset **Diabetes**:

* Jednoznačně nejlepších výsledků dosahuje knihovna **Hyperopt.** Jak vzhledem k výsledkům, tak i vzhledem k časové délce výpočtu.

Dataset **Wine**:

* Opět nejlépe vychází výsledky knihovny **Hyperopt**. Knihovna **Skopt** dosahuje velmi podobných výsledků, ovšem s mnohem horším výpočetním časem.

**Obecné shrnutí**:

* Knihovna **Optuna** bohužel při výpočtu dvou algoritmů ze tří selhává a nedosahuje ani na rozumná řešení. K její aplikaci se budeme muset vrátit a nalézt důvod proč je knihovna při nalézání řešení tak neefektivní. Nicméně pro model **RFR** funguje dobře.
* Knihovna **Skopt** má výpočetně velmi dlouhé časy i pro jednodušší datasety. Dosahuje poměrně dobrých výsledků, ale výsledky nejsou tak dobré jako pro nejlepší knihovnu.
* Nejlepších výsledků dosahuje knihovna **Hyperopt**. Vzhledem k náročnosti jednotlivých modelů vidíme, že pro výpočetně nejnáročnější knihovnu **California Housing** dosahuje algoritmus **XGBoost**. Pro menší knihovny je volba modelu strojového učení redundantní vhledem k časové délce výpočtu.

Závěrem doporučujeme pro navazující studii algoritmus strojového učení **XGBoost** v kombinaci s optimalizační knihovnou **Hyperopt**.