Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-182905-105345

Bc. Monika Zjavková

Predikčný model na stanovenie dermatotoxicity

Prílohy

Vedúci práce: Ing. Marta Šoltésová Prnová PhD.

Máj 2025

Dodatok A

Inštalačná príručka

Aplikáciu je možné spustiť prostredníctvom virtuálneho prostredia, ktoré zabezpečuje izolované prostredie s vlastnými knižnicami a závislosťami. Postup inštalácie je nasledovný:

1. **Stiahnutie zdrojového kódu**
   Najprv si stiahnite zdrojový kód aplikácie a prejdite do príslušného adresára v termináli.

2. **Vytvorenie virtuálneho prostredia**
   V koreňovom adresári aplikácie vytvorte nové virtuálne prostredie:

   ```
   python -m venv venv
   ```

3. **Aktivácia virtuálneho prostredia**
   Aktivujte prostredie podľa vášho operačného systému:

   Linux/macOS:

   ```
   source venv/bin/activate
   ```

Windows:

```
venv\Scripts\activate
```

4. **Inštalácia závislostí**

   Po aktivácii virtuálneho prostredia nainštalujte požadované knižnice pomocou súboru `requirements.txt`:

   ```
   pip install -r requirements.txt
   ```

5. **Spustenie aplikácie**

   Aplikáciu spustíte nasledovným príkazom:

   ```
   python manage.py runserver
   ```

   Aplikácia bude následne dostupná na adrese: `http://localhost:8000/`

Dodatok B

Používateľská príručka

B.1 Predikcia podráždenia pokožky

Aplikácia umožňuje predikovať potenciál chemickej látky spôsobiť podráždenie pokožky. Užívateľské rozhranie je navrhnuté jednoducho a intuitívne. Postupujte podľa nasledujúcich krokov:

1. Spustite aplikáciu vo webovom prehliadači. Aplikácia je dostupná na adrese `http://localhost:8000`.
2. Do vstupného poľa zadajte chemickú štruktúru vo formáte SMILES alebo zadajte CAS číslo danej zlúčeniny.
3. Stlačte tlačidlo **Predict** pre spustenie predikcie.
4. Po predikcii sa zobrazia nasledujúce informácie:
   - **Výsledok predikcie**: Označenie látky ako `Irritant` alebo `Non-Irritant`.
   - **Dôveryhodnosť modelu (Trust Score)** v percentách.
   - **Vizualizácia molekuly** na základe SMILES.

- **Chemické informácie**: názov zlúčeniny, molekulový vzorec, SMILES reťazec a CAS číslo.
- **Informácie o datasete**: zdroj údajov, veľkosť datasetu, formát dát a validácia.
- **Graf ROC krivky** modelu.
- **Hodnoty SHAP**: vplyv jednotlivých modelov (KNN, Random Forest, XGBoost, SVM, Decision Tree) na výstupnú predikciu.

Dodatok C

Technická dokumentácia

Táto kapitola popisuje technické aspekty riešenia diplomovej práce, vrátane architektúry systému, štruktúry súborov, použitých technológií a implementačných rozhodnutí. Výsledný softvér pozostáva z dvoch samostatných častí:

1. fáza vývoja modelov, ktorá zahŕňa spracovanie dát, výber príznakov, trénovanie a optimalizáciu modelov pomocou Jupyter notebookov,

2. produkčná fáza – webová aplikácia postavená na Django frameworku, ktorá poskytuje jednoduché a prístupné rozhranie na predikciu dermatotoxicity chemických látok pomocou uloženého modelu.

C.1 Fáza vývoja modelov (offline spracovanie)

Tréningová časť bola realizovaná výhradne pomocou Jupyter notebookov. Každý krok spracovania a vývoja modelu je oddelený do samostatného notebooku pre lepšiu prehľadnosť a flexibilitu.

Použité notebooky

- `data_processing.ipynb` – načítanie datasetu, čistenie údajov, spracovanie chemických štruktúr do deskriptorov,
- `feature_selection.ipynb` – zníženie počtu príznakov pomocou korelačnej analýzy, rozptylu, LASSO a SHAP hodnôt,
- `hyperparameter_optimization_X.ipynb` – samostatné notebooky pre optimalizáciu každého modelu.
- `ensemble_model.ipynb` – porovnanie modelov, prípadne kombinácia a výber finálneho modelu na nasadenie.

Uloženie modelu

Najlepší model bol uložený pomocou knižnice `joblib` a exportovaný do súboru, ktorý je následne použitý v produkčnej časti aplikácie.

C.2 Webová aplikácia (Django)

Django aplikácia slúžila na nasadenie uloženého modelu. Obsahuje všetky potrebné komponenty na vizualizáciu a interpretáciu výsledkov:

- `views.py` – načítanie modelov, spracovanie vstupu, volanie predikcie, generovanie výstupov,
- `urls.py` – routovanie požiadaviek,
- `templates/` – HTML šablóny s Bootstrapom,
- `static/` – CSS, JS a statické zdroje (napr. vykreslené grafy),

- `utils/` – pomocné skripty pre spracovanie SMILES, vykresľovanie molekúl, načítanie SHAP hodnôt.

Po zadaní chemickej látky používateľom aplikácia:

1. transformuje SMILES do deskriptorov,
2. aplikuje rovnaké predspracovanie ako počas tréningu (napr. škálovanie),
3. použije uložený model na vykonanie predikcie,
4. zobrazí výsledky vrátane interpretácie (napr. SHAP grafy, vizualizácia molekuly),

C.3 Použité knižnice

V rámci vývoja riešenia boli využité viaceré knižnice na oblasti spracovania dát, strojového učenia, vizualizácie, štatistiky a webový vývoj. V nasledujúcom prehľade sú uvedené najdôležitejšie z nich:

- **RDKit** – nástroj na výpočty a vizualizácie chemických štruktúr.
- **pandas** – knižnica na manipuláciu s dátami vo forme tabuliek (DataFrame).
- **NumPy** – základná knižnica pre numerické výpočty. Využíva sa najmä na prácu s poľami, lineárnu algebru a matematické transformácie.
- **scikit-learn** – hlavný framework pre implementáciu modelov strojového učenia. Bol použitý na trénovanie aj na metriky vyhodnocovania.
- **XGBoost** – implementácia gradient boosting algoritmu. V projekte použitá ako jeden z klasifikačných modelov.
- **SHAP** – nástroj na vysvetľovanie rozhodnutí modelov strojového učenia.

Pomocou SHAP hodnôt bola vyhodnotená dôležitosť jednotlivých deskriptorov v predikcii.

- **Optuna** – framework pre automatizovanú optimalizáciu hyperparametrov.
- **matplotlib** a **seaborn** – knižnice na vizualizáciu dát a výsledkov. Použité na generovanie grafov (ROC krivky, SHAP diagramy.)
- **joblib** – slúži na ukladanie modelov na disk, aby mohli byť neskôr načítané v produkčnej aplikácii.
- **SciPy** – doplnková knižnica pre štatistické výpočty (napr. distribúcie, testy normality, rozptylové analýzy).
- **Django** – webový framework v jazyku Python, použitý na vytvorenie používateľského rozhrania, prípravu logiky pre načítanie modelu, predikciu a zobrazenie výstupu.

Dodatok D

Časový plán práce

Táto kapitola poskytuje prehľad časového harmonogramu a hodnotí jeho realizáciu. Nasleduje rozdelenie podľa semestrov s podrobnosťami jednotlivých krokov.

D.1 Letný semester 2023/2024

V letnom semestri som sa zameriavala na analýzu problému, návrh aplikácie a predspracovanie dát. Pracovný plán prebiehal nasledovne:

- **Týždeň 1-2:** Preštudovanie literatúry o QSAR modelovaní a predikcii dermatotoxicity.
- **Týždeň 3-4:** Písanie kapitoly o analýze problému a výbere modelov strojového učenia.
- **Týždeň 5-6:** Zoznamovanie sa s dostupnými datasetmi a ich štruktúrou (SMILES kódy, chemické deskriptory).
- **Týždeň 7-8:** Návrh architektúry aplikácie a vytvorenie úvodných diagramov

pracovných procesov.

- **Týždeň 9-10:** Predspracovanie a čistenie dát, výpočet deskriptorov pomocou RDKit.
- **Týždeň 11-12:** Písanie kapitoly o návrhu riešenia.

D.2 Zimný semester 2024/2025

Zimný semester bol venovaný implementácii algoritmov, výberu vlastností (feature selection), optimalizácii a validácii modelov.

- **Týždeň 1-3:** Implementácia algoritmov strojového učenia (Random Forest, SVM, XGBoost).
- **Týždeň 4:** Validácia modelov na tréningových a testovacích dátach.
- **Týždeň 5-6:** Implementácia metód výberu vlastností (SHAP, filter, wrapper metódy).
- **Týždeň 7-8:** Optimalizácia hyperparametrov pomocou grid search a ďalších metód.
- **Týždeň 9:** Implementácia vizualizácie pomocou SHAP a interpretácie výsledkov.
- **Týždeň 10:** Testovanie aplikácie a identifikácia chýb v predikciách.
- **Týždeň 11:** Oprava nedostatkov a vylepšenia aplikácie.
- **Týždeň 12:** Písanie kapitoly o implementácii riešenia.

D.3 Letný semester 2024/2025

V letnom semestri 2025 sme sa zamerali na finálne fázy diplomovej práce, ktoré zahŕňali vývoj ensemble modelov, vytvorenie webovej aplikácie, testovanie a písanie záverečných častí dokumentácie. Pracovný plán bol rozvrhnutý nasledovne:

- **Týždeň 1–2:** Návrh a implementácia ensemble prístupu — testovanie metamodelu a hlasovania na kombinovanie viacerých predikčných algoritmov.
- **Týždeň 3–5:** Vývoj backendovej časti webovej aplikácie vrátane integrácie predikčného modelu, predspracovania vstupov a výpočtu výsledkov.
- **Týždeň 6–8:** Implementácia frontendovej časti aplikácie so zameraním na používateľsky prívetivé rozhranie, vizualizáciu výstupov a zobrazovanie SHAP interpretácií.
- **Týždeň 9:** Riešenie technických a funkčných nedostatkov zistených počas vývoja a interného testovania aplikácie.
- **Týždeň 10–11:** Písanie kapitoly o webovej aplikácii a o výsledkoch ensemble prístupu vrátane grafických výstupov a diskusie.
- **Týždeň 12:** Finálne testovanie funkčnosti aplikácie, kontrola výstupov a validácia modelov v produkčnom prostredí.

D.4 Zhodnotenie časového harmonogramu

Časový harmonogram bol navrhnutý tak, aby pokrýval všetky hlavné fázy práce od úvodného štúdia problematiky, cez analýzu a návrh riešenia až po implementáciu a validáciu modelov. Celkový plán sa ukázal byť realistický, avšak počas realizácie bolo nutné urobiť niekoľko úprav.

V letnom semestri sme sa sústredili na preštudovanie literatúry, pochopenie dostupných datasetov a návrh aplikácie. Táto fáza prebehla prevažne podľa plánu, pričom najväčšou výzvou bolo spracovanie dát a zabezpečenie ich dostatočnej kvality pre ďalšiu analýzu. Písanie kapitol o analýze problému a návrhu riešenia prebehlo bez výrazných oneskorení.

V zimnom semestri sme sa zamerali na implementáciu algoritmov strojového učenia a optimalizáciu výsledkov. Hoci bol plán všeobecne dodržaný, implementácia metód výberu vlastností (feature selection) si vyžiadala viac času, ako sa pôvodne očakávalo. Dôvodom bola potreba experimentovania s viacerými metódami (SHAP, filter a wrapper) na dosiahnutie optimálnych výsledkov. Napriek tomu sa podarilo zachovať harmonogram implementácie.

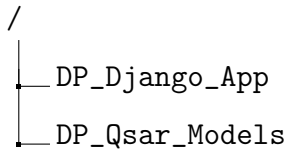
Letný semester 2025 bol zameraný na rozšírenie modelu pomocou ensemble techník a jeho integráciu do plne funkčnej webovej aplikácie. Táto fáza si vyžiadala systematické plánovanie — backend bol navrhnutý tak, aby zvládol dynamické vstupy, zatiaľ čo frontend sa sústredil na zrozumiteľnú vizualizáciu výstupov a interpretácií modelu. Rezerva v závere semestra umožnila doladiť nedostatky a sústrediť sa na kvalitu používateľskej skúsenosti.

Dodatok E

Opis elektronického média

Evidenčné číslo práce: FIIT-182905-105345

Digitálna zložka diplomovej práce je rozdelená následovne:



DP_Django_App obsahuje implementáciu webovej aplikácie. DP_Qsar_Models obsahuje Jupyter notebooky, kde sa nachádza spracovanie dát, výber deskriptorov a optimalizácia jednotlivých modelov. Zloženie a obsah oboch priečinkov sú bližšie popísané v prílohe C.