**TECHNICKÁ UNIVERZITA V KOŠICIACH**
**FAKULTA ELEKTROTECHNIKY A INFORMATIKY**

# Dokumentácia
# Zadanie 1 – Regresná úloha

Inteligentné systémy v kyberbezpečnosti

**Košice, 2024** **Vladyslav Boiko**

# Teoretická časť

Použitý model pre regresiu je **lineárna regresia**. Tento model patrí medzi základné metódy strojového učenia a predpokladá lineárny vzťah medzi nezávislými premennými (vstupné údaje) a závislou premennou (cieľová hodnota). Výhoda lineárnej regresie spočíva v jednoduchosti, rýchlej implementácii a interpretovateľnosti výsledkov.

# Návrh riešenia

Metodológia:

1. Načítanie údajov:

   - Údaje boli načítané z *.npy* súborov pomocou knižnice **numpy**.

   - Dátová množina bola rozdelená na kategóriové(*X_data_text_columns, X_eval_text_columns*) a numerické stĺpce(*X_data_numeric_columns, X_eval_text_columns*)

2. Predspracovanie údajov:

   - Kategóriové stĺpce boli kódované pomocou **OneHotEncoder**. Táto metóda zabezpečila správnu reprezentáciu textových údajov.
   - Numerické stĺpce boli spracované **SimpleImputer**, čo zamedzilo problémom s chýbajúcimi hodnotami. Všetky chýbajúce položky sa nahradia priemernou hodnotou stĺpca

3. Spájanie dát:

   - Po transformácii boli kategóriové a numerické stĺpce skombinované do jednej matice pomocou *numpy.hstack*.

4. Rozdelenie údajov:

   - Dátová množina bola rozdelená na tréningovú a testovaciu množinu pomocou *train_test_split* (pomer 75 train : 25 test).

5. Tréning modelu
6. Predikcia a hodnotenie:
   - Model bol vyhodnotený na testovacej množine pomocou metriky $R^2$ skóre, ktorá meria kvalitu predikcie.

- Finálna predikcia pre evaluačné údaje bola uložená do súboru *y_predikcia.npy*.

# Diskusia a výsledky

Počas tejto úlohy som porovnával rôzne modely, ako napríklad **Ridge**, **Lasso** a **LinearRegression**. Spomedzi nich dosiahla lineárna regresia najlepší výsledok r2_score 0,981. V procese implementácie sa používal aj GridSearchCV, ktorý určuje najlepšie parametre pre rôzne modely, ale vo finálnej verzii bol odstránený, pretože vždy vyberál rovnaké predvolené parametre a osobne si myslím, že je lepšie kontrolovať parametre samostatne a vedieť, ktoré parametre vedú k určitému výsledku.