

**SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE**

**FAKULTA CHEMICKEJ A POTRAVINÁRSKEJ  
TECHNOLÓGIE**

Evidenčné číslo: FCHPT-FCHPT-184466-117370

# **Softvérové senzory pre priemysel**

**BAKALÁRSKA PRÁCA**



**SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE**

**FAKULTA CHEMICKÉJ A POTRAVINÁRSKEJ  
TECHNOLÓGIE**

Evidenčné číslo: FCHPT-FCHPT-184466-117370

# **Softvérové senzory pre priemysel**

**BAKALÁRSKA PRÁCA**

|                      |                                                           |
|----------------------|-----------------------------------------------------------|
| Študijný program:    | Riadenie procesov                                         |
| Študijný odbor:      | Kybernetika                                               |
| Školiace pracovisko: | Oddelenie informatizácie a riadenia procesov - ÚIAM FCHPT |
| Vedúci práce:        | doc. Ing. Radoslav Paulen, PhD.                           |

**2026**

**Semen Suslov**





## ZADANIE BAKALÁRSKEJ PRÁCE

Študent: **Semen Suslov**  
ID študenta: 117370  
Študijný program: riadenie procesov  
Študijný odbor: kybernetika  
Vedúci práce: doc. Ing. Radoslav Paulen, PhD.  
Vedúci pracoviska: prof. Ing. Miroslav Fikar, DrSc.  
Miesto vypracovania: Oddelenie informatizácie a riadenia procesov FCHPT STU

Názov práce: **Softvérové senzory pre priemysel**

Jazyk, v ktorom sa práca vypracuje: slovenský jazyk

Špecifikácia zadania:

Hlavným cieľom projektu je návrh (inferenčných) softvérových senzorov, ktoré predstavujú v priemyselných procesoch lacnú alternatívu snímania problematických, ťažko merateľných veličín (napr. zloženie produktov alebo degradácia zariadení). Ako prípadové štúdie budú použité dáta z priemyselných destilačných kolón prevádzkovaných v závode Slovnaft, a.s. Pri návrhu inferenčných senzorov bude porovnaná presnosť viacerých prístupov (pomocou štatistických metód) a určený najvhodnejší softvérový senzor.

Úlohy:

- Spracovanie dát (odstránenie systematických chýb, normalizácia)
- Návrh štruktúr matematických modelov senzorov
- Výpočet parametrov senzorov
- Vyhodnotenie kvality navrhnutých softvérových senzorov

Zoznam odbornej literatúry:

1. MOJTO, Martin a PAULEN, Radoslav. *Data-driven Design of Linear Soft Sensors*. Dizertačná práca. 2023.
2. KRIŽAN, Darko a PAULEN, Radoslav. *Software Sensors for Industry*. Diplomová práca. 2021.

Termín odovzdania bakalárskej práce: 08. 05. 2026  
Dátum schválenia zadania bakalárskej práce: 23. 02. 2026  
Zadanie bakalárskej práce schválil: doc. Ing. Juraj Oravec, PhD. – garant študijného programu



# Podakovanie

Touto cestou by som chcel úprimne poďakovať môjmu školiteľovi, doc. Ing. Radoslavovi Paulenovi, PhD., za jeho odborné vedenie, cenné rady a neustálu podporu počas vypracovania tejto práce. Jeho prístup bol nielen pedagogický, ale aj inšpirujúci. Motivoval ma prekračovať vlastné limity, hľadať hlbší zmysel v práci a neustále sa rozvíjať. Vďaka jeho vedeniu som mal možnosť zúčastniť sa konferencie, získať nové skúsenosti a uvedomiť si svoj vlastný potenciál.

Zároveň by som chcel vyjadriť poďakovanie celému kolektívu Ústavu informatiky, automatizácie a matematiky za príjemné a podnetné prostredie, v ktorom je ochota pomôcť a spolupracovať samozrejmosťou.

Osobitné poďakovanie patrí všetkým členom UIAM a najmä doc. Paulenovi za významný prínos k môjmu odbornému aj osobnému rastu.

## Čestné vyhlásenie

Vyhlasujem, že predloženú záverečnú prácu som vypracoval samostatne pod vedením vedúceho záverečnej práce, s použitím odbornej literatúry a ďalších informačných zdrojov, ktoré sú citované v práci a uvedené v zozname použitej literatúry. Ako autor záverečnej práce ďalej prehlasujem, že som v súvislosti s jej vytvorením neporušil autorské práva tretích osôb.

Podpis: \_\_\_\_\_

## Vyhlásenie o použití umelej inteligencie

Pri vypracovaní tejto práce nástroje umelej inteligencie (napr. ChatGPT) boli použité výlučne na jazykovú korektúru a pomocné formulácie textu.

Obsah práce, analýza dát, návrh modelov a interpretácia výsledkov boli vypracované autorom samostatne. Autor zodpovedá za správnosť a originalitu všetkých uvedených informácií.

Podpis: \_\_\_\_\_

# Abstrakt

Softvérové senzory sú matematické modely, ktoré predikujú ťažko merateľné veličiny na základe ľahko dostupných procesných meraní. Práca sa zaoberá návrhom a porovnaním softvérových senzorov určených na predikciu molárneho zlomku produktu na dne destilačnej kolóny. V každej z piatich etáp experimentu líšiacich sa prístupom k výberu príznakov boli porovnávané dva modely: lineárna regresia a neurónová sieť, hodnotené pomocou opakovanej krížovej validácie. Kľúčovým zistením je, že výber príznakov mal väčší vplyv na presnosť predikcie než voľba architektúry modelu. Najlepší výsledok dosiahla lineárna regresia, ktorá predstavuje kompromis medzi presnosťou, interpretovateľnosťou a nákladmi na implementáciu.

**Kľúčové slová:** softvérové senzory, inferenčné senzory, lineárna regresia, neurónová sieť, inžinierstvo príznakov, destilačná kolóna, krížová validácia



# Abstract

Soft sensors are mathematical models that predict difficult-to-measure quantities based on easily available process measurements. This thesis focuses on the design and comparison of soft sensors for predicting the molar fraction of the product at the bottom of a distillation column. In each of the five experimental stages, differing in their approach to feature selection, two models were compared: linear regression and a neural network, evaluated using repeated cross-validation. The key finding is that feature selection had a greater impact on prediction accuracy than the choice of model architecture. The best result was achieved by linear regression, which represents a trade-off between accuracy, interpretability, and implementation cost.

**Keywords:** soft sensors, inferential sensors, linear regression, neural network, feature engineering, distillation column, cross-validation



# Obsah

|                                               |            |
|-----------------------------------------------|------------|
| <b>Podakovanie</b>                            | <b>iii</b> |
| <b>Abstrakt</b>                               | <b>v</b>   |
| <b>Abstract</b>                               | <b>vii</b> |
| <b>1 Úvod</b>                                 | <b>1</b>   |
| <b>2 Teoretická časť</b>                      | <b>3</b>   |
| 2.1 Lineárna a nelineárna regresia . . . . .  | 3          |
| 2.1.1 Lineárna regresia . . . . .             | 3          |
| 2.1.2 Metóda najmenších štvorcov . . . . .    | 4          |
| 2.1.3 Neurónová sieť . . . . .                | 5          |
| 2.1.4 Architektúra neurónovej siete . . . . . | 5          |
| 2.1.5 Stratová funkcia . . . . .              | 6          |
| 2.1.6 Aktivačné funkcie . . . . .             | 6          |
| 2.1.7 Optimalizačný algoritmus Adam . . . . . | 7          |
| 2.1.8 Dropout . . . . .                       | 7          |
| 2.1.9 Krivky učenia . . . . .                 | 7          |
| 2.2 Ukazovatele presnosti modelu . . . . .    | 8          |

|          |                                                                         |           |
|----------|-------------------------------------------------------------------------|-----------|
| 2.2.1    | Odmocnina strednej kvadratickej chyby (RMSE) . . . . .                  | 8         |
| 2.2.2    | $R^2$ (koeficient determinácie) . . . . .                               | 9         |
| 2.2.3    | Štandardná odchýlka . . . . .                                           | 9         |
| 2.3      | Metodológia . . . . .                                                   | 10        |
| 2.3.1    | Rozdelenie datasetu . . . . .                                           | 10        |
| 2.3.2    | Normalizácia dát . . . . .                                              | 10        |
| 2.3.3    | Vplyv úniku informácií pri normalizácii . . . . .                       | 11        |
| 2.3.4    | Krížová validácia . . . . .                                             | 11        |
| 2.3.5    | Opakovaná krížová validácia . . . . .                                   | 11        |
| 2.3.6    | Analýza korelácie . . . . .                                             | 12        |
| 2.3.7    | Multikolinearita . . . . .                                              | 12        |
| 2.3.8    | Odporúčaný pomer parametrov a počtu pozorovaní . . . . .                | 14        |
| 2.4      | Aplikácia Phiture ako nástroj pre vývoj softvérových senzorov . . . . . | 15        |
| 2.4.1    | Architektúra a pracovný postup . . . . .                                | 15        |
| 2.4.2    | Využitie pri návrhu softvérových senzorov . . . . .                     | 20        |
| <b>3</b> | <b>Experimentálna časť</b>                                              | <b>21</b> |
| 3.1      | Opis datasetu . . . . .                                                 | 21        |
| 3.1.1    | Destilačná kolóna . . . . .                                             | 21        |
| 3.1.2    | Opis premenných . . . . .                                               | 22        |
| 3.2      | Implementácia práce . . . . .                                           | 23        |
| 3.2.1    | Použité nástroje a knižnice . . . . .                                   | 23        |
| 3.2.2    | Rozdelenie datasetu . . . . .                                           | 23        |
| 3.2.3    | Normalizácia dát . . . . .                                              | 24        |
| 3.2.4    | Implementácia krížovej validácie . . . . .                              | 24        |

---

|       |                                                                     |    |
|-------|---------------------------------------------------------------------|----|
| 3.2.5 | Výber hlavnej metriky hodnotenia . . . . .                          | 25 |
| 3.2.6 | Implementácia modelov . . . . .                                     | 26 |
| 3.2.7 | Algoritmus výberu príznakov . . . . .                               | 26 |
| 3.2.8 | Vyčerpávajúce prehľadávanie priestoru príznakov . . . . .           | 26 |
| 3.3   | Výber architektúry neurónovej siete . . . . .                       | 27 |
| 3.3.1 | Obmedzenie počtu parametrov . . . . .                               | 27 |
| 3.3.2 | Porovnanie architektúr . . . . .                                    | 27 |
| 3.3.3 | Krivky učenia . . . . .                                             | 28 |
| 3.3.4 | Absencia výrazného pretrénovania . . . . .                          | 30 |
| 3.3.5 | Voľba architektúry . . . . .                                        | 30 |
| 3.4   | Etapa 1 — Základný výber príznakov . . . . .                        | 30 |
| 3.4.1 | Postup . . . . .                                                    | 30 |
| 3.4.2 | Výsledky . . . . .                                                  | 32 |
| 3.5   | Etapa 2 — Nelineárne transformácie príznakov . . . . .              | 34 |
| 3.5.1 | Postup . . . . .                                                    | 34 |
| 3.5.2 | Výsledky . . . . .                                                  | 38 |
| 3.6   | Etapa 3 — Fyzikálne odôvodnené príznaky . . . . .                   | 39 |
| 3.6.1 | Postup . . . . .                                                    | 39 |
| 3.6.2 | Výsledky . . . . .                                                  | 43 |
| 3.7   | Etapa 4 — Výber príznakov z predchádzajúcich etáp . . . . .         | 43 |
| 3.7.1 | Postup . . . . .                                                    | 44 |
| 3.7.2 | Výsledky . . . . .                                                  | 47 |
| 3.8   | Etapa 5 — Vyčerpávajúce prehľadávanie priestoru príznakov . . . . . | 48 |
| 3.8.1 | Postup . . . . .                                                    | 48 |

---

|          |                                                                    |           |
|----------|--------------------------------------------------------------------|-----------|
| 3.8.2    | Výsledky . . . . .                                                 | 48        |
| 3.9      | Celkové porovnanie etáp 1–5 . . . . .                              | 49        |
| <b>4</b> | <b>Diskusia</b>                                                    | <b>51</b> |
| 4.1      | Porovnanie etáp a výberu príznakov . . . . .                       | 51        |
| 4.2      | Porovnanie modelov: lineárna regresia vs. neurónová sieť . . . . . | 52        |
| 4.3      | Odporúčanie finálnej konfigurácie softvérového senzora . . . . .   | 53        |
| 4.3.1    | Fyzické senzory potrebné pre rôzne konfigurácie . . . . .          | 53        |
| 4.3.2    | Odporúčaný model a príznaky . . . . .                              | 54        |
| 4.4      | Obmedzenia a zdroje neistoty . . . . .                             | 55        |
| <b>5</b> | <b>Záver</b>                                                       | <b>57</b> |
|          | <b>Literatúra</b>                                                  | <b>59</b> |

# Zoznam obrázkov

|     |                                                                                                                                                    |    |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | Schéma neurónu . . . . .                                                                                                                           | 5  |
| 2.2 | Schéma rozdelenia datasetu na trénovaciu, testovaciu a validačnú množinu                                                                           | 10 |
| 2.3 | Modul vyhľadávania vedeckých publikácií. . . . .                                                                                                   | 16 |
| 2.4 | Stránka po dokončení sťahovania PDF dokumentov. Zobrazuje počet úspešne stiahnutých súborov a umožňuje ich správu pred ďalším spracovaním. . . . . | 17 |
| 2.5 | Modul extrakcie matematických výrazov z PDF dokumentov. Používateľ môže označiť relevantné výrazy určené na ďalšie spracovanie. . . . .            | 18 |
| 2.6 | Modul konverzie LaTeX zápisu do Python kódu pomocou knižnice <code>sympy</code> . . . . .                                                          | 19 |
| 3.1 | Krivky učenia pre architektúru s 8 neurónmi (hodnota generátora rozdelení 34). . . . .                                                             | 29 |
| 3.2 | Krivky učenia pre architektúru so 6 neurónmi (hodnota generátora rozdelení 34). . . . .                                                            | 29 |
| 3.3 | Architektúra navrhutej neurónovej siete so šiestimi neurónmi v skrytej vrstve . . . . .                                                            | 31 |
| 3.4 | Matica vzájomnej korelácie vstupných príznakov v etape 1. . . . .                                                                                  | 33 |

---

|      |                                                                                                                            |    |
|------|----------------------------------------------------------------------------------------------------------------------------|----|
| 3.5  | Korelácia transformovaných príznakov s cieľovou premennou $x_B$ v etape 2 (zobrazené len príznaky s $ r  > 0,4$ ). . . . . | 35 |
| 3.6  | Matica vzájomnej korelácie vybraných príznakov etapy 2. . . . .                                                            | 37 |
| 3.7  | Korelácia fyzikálne odôvodnených príznakov s cieľovou premennou $x_B$ v etape 3. . . . .                                   | 40 |
| 3.8  | Matica vzájomnej korelácie vybraných príznakov etapy 3. . . . .                                                            | 42 |
| 3.9  | Korelácia vybraných príznakov s cieľovou premennou $x_B$ v etape 4. . .                                                    | 45 |
| 3.10 | Matica vzájomnej korelácie vybraných príznakov etapy 4. . . . .                                                            | 46 |

# Zoznam tabuliek

|      |                                                                         |    |
|------|-------------------------------------------------------------------------|----|
| 3.1  | Porovnanie výsledkov krížovej validácie pre rôzne hodnoty $k$ . . . . . | 25 |
| 3.2  | Porovnanie architektúr neurónových sietí . . . . .                      | 27 |
| 3.3  | Vybrané príznaky etapy 1 . . . . .                                      | 31 |
| 3.4  | Najlepšie výsledky modelov v etape 1 . . . . .                          | 32 |
| 3.5  | Matica vzájomnej korelácie vybraných príznakov etapy 1 . . . . .        | 32 |
| 3.6  | Test RMSE lineárnej regresie ( $p = 5$ ) v etape 1. . . . .             | 34 |
| 3.7  | Vybrané príznaky etapy 2 . . . . .                                      | 35 |
| 3.8  | Najlepšie výsledky modelov v etape 2 . . . . .                          | 36 |
| 3.9  | Matica vzájomnej korelácie vybraných príznakov etapy 2 . . . . .        | 38 |
| 3.10 | Test RMSE neurónovej siete ( $p = 5$ ) v etape 2. . . . .               | 38 |
| 3.11 | Vybrané príznaky etapy 3 . . . . .                                      | 40 |
| 3.12 | Najlepšie výsledky modelov v etape 3 . . . . .                          | 41 |
| 3.13 | Matica vzájomnej korelácie vybraných príznakov etapy 3 . . . . .        | 41 |
| 3.14 | Test RMSE lineárnej regresie ( $p = 5$ ) v etape 3 . . . . .            | 43 |

---

|      |                                                                      |    |
|------|----------------------------------------------------------------------|----|
| 3.15 | Porovnanie najlepších výsledkov CV RMSE v etapách 1–3 . . . . .      | 43 |
| 3.16 | Súbor 15 kandidátnych príznakov pre etapu 4 . . . . .                | 44 |
| 3.17 | Matica vzájomnej korelácie vybraných príznakov etapy 4 . . . . .     | 45 |
| 3.18 | Vybrané príznaky etapy 4 . . . . .                                   | 46 |
| 3.19 | Najlepšie výsledky v etape 4 . . . . .                               | 47 |
| 3.20 | Test RMSE lineárnej regresie ( $p = 5$ ) v etape 4. . . . .          | 47 |
| 3.21 | Najlepšie výsledky vyčerpávajúceho prehľadávania (etapa 5) . . . . . | 48 |
| 3.22 | Test RMSE lineárnej regresie ( $p = 5$ ) v etape 5. . . . .          | 49 |
| 3.23 | Celkové porovnanie najlepších výsledkov CV RMSE naprieč etapami .    | 49 |
| 4.1  | Počet fyzických meraní pre vybrané konfigurácie softvérového senzora | 54 |

V moderných priemyselných procesoch má monitorovanie technologických veličín kľúčovú úlohu pri zabezpečení kvality výroby, bezpečnosti prevádzky a efektívneho riadenia procesov. V mnohých prípadoch však nie je možné niektoré dôležité veličiny merať priamo pomocou fyzických senzorov. Dôvodom môžu byť vysoké náklady na meracie zariadenia, technické obmedzenia, oneskorené merania alebo nevyhovujúce podmienky prostredia, ako sú vysoké teploty alebo tlaky (Bakhtadze a kol., 2008).

Jedným z prístupov k riešeniu tohto problému je využitie softvérových senzorov (angl. *soft sensors*). Softvérové senzory umožňujú odhadovať ťažko merateľné alebo nemerateľné veličiny na základe iných, ľahko dostupných procesných meraní. Namiesto priameho merania sa využívajú matematické modely, ktoré popisujú vzťah medzi vstupnými procesnými veličinami a požadovanou výstupnou hodnotou (Kadlec a kol., 2009).

Softvérové senzory majú široké uplatnenie v chemickom priemysle, kde sú sledovanými veličinami napríklad zloženie zmesi, viskozita, stupeň polymerizácie alebo rozdelenie veľkosti častíc. Ich hlavnou výhodou je možnosť poskytovať odhady v reálnom čase, znižovať potrebu drahých analyzátorov a zvyšovať flexibilitu riadiacich systémov.

Z hľadiska štruktúry môžu mať softvérové senzory rôznu podobu. Najjednoduchšie prístupy sú založené na lineárnych regresných modeloch, ktoré ponúkajú vysokú interpretovateľnosť a nízku výpočtovú náročnosť (James a kol., 2013). Pokročilejšie riešenia využívajú metódy strojového učenia, ako sú neurónové siete, ktoré dokážu zachytiť aj nelineárne vzťahy medzi veličinami (Goodfellow a kol., 2016). Výber konkrétneho modelu závisí od charakteru procesu, dostupnosti dát a požadovanej presnosti.

Pre úspešný návrh softvérového senzora sú nevyhnutné kvalitné dáta, ktoré obsahujú merania vstupných veličín a zodpovedajúce hodnoty cieľovej premennej. Dôležitým

krokom je tiež výber vhodných vstupných príznakov, analýza ich vzájomných vzťahov a eliminácia redundantných veličín. Samotný model je následne potrebné natrénovať, validovať a otestovať, aby bola zabezpečená jeho spoľahlivosť pri reálnom nasadení (James a kol., 2013).

V rámci tejto práce bol použitý dataset obsahujúci merania koncentrácie a hodnoty rôznych procesných snímačov z destilačnej kolóny v spoločnosti Slovanft, a.s. Cieľom je návrh a porovnanie modelov softvérového senzora určeného na odhad koncentrácie v konkrétnej destilačnej kolóne na základe procesných meraní a hodnotenie presnosti navrhnutých modelov pomocou vhodných metrík.

## Teoretická časť

---

Táto kapitola poskytuje teoretický základ pre návrh softvérových senzorov. Sú tu predstavené metódy lineárnej a nelineárnej regresie, princípy trénovania modelov a prístupy k výberu a hodnoteniu príznakov.

### 2.1 Lineárna a nelineárna regresia

V tejto kapitole sú popísané dva prístupy k modelovaniu: lineárna regresia a neurónová sieť, ako nelineárna alternatíva.

#### 2.1.1 Lineárna regresia

Lineárna regresia patrí medzi základné modely strojového učenia a používa sa na opis lineárnej závislosti medzi množinou vstupných príznakov a cieľovou premennou. Predpokladá sa, že vzťah medzi príznakmi a výstupnou hodnotou je možné aproximovať lineárnou funkciou.

Pre jeden príznak má model tvar:

$$\hat{y} = c_0 + c_1x. \quad (2.1)$$

Pre prípad  $p$  vstupných príznakov sa model rozšíri na:

$$\hat{y} = c_0 + c_1x_1 + c_2x_2 + \dots + c_px_p, \quad (2.2)$$

kde  $\hat{y}$  je predikovaná hodnota,  $x_1, x_2, \dots, x_p$  sú vstupné príznaky,  $c_0$  je absolútny člen (posun),  $c_1, \dots, c_p$  sú koeficienty modelu.

### 2.1.2 Metóda najmenších štvorcov

Na určenie koeficientov lineárnej regresie sa používa metóda najmenších štvorcov. Cieľom metódy najmenších štvorcov je nájsť také hodnoty vektora  $\mathbf{c}$ , ktoré minimalizujú súčet štvorcov odchýlok medzi nameranými a predikovanými hodnotami. Nech  $\mathbf{c} = (c_0, c_1, \dots, c_p)^\top$  je vektor regresných koeficientov. Potom úloha metódy najmenších štvorcov má tvar:

$$\min_{\mathbf{c}} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (2.3)$$

kde  $n$  je počet meraní,  $i = 1, \dots, n$  označuje jednotlivé pozorovania,  $y_i$  je nameraná hodnota výstupu a  $\hat{y}_i$  je predikcia modelu.

V prípade viacerých príznakov je výhodné zapísať lineárnu regresiu v maticovej forme. Nech:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \quad (2.4)$$

je matica príznakov rozmeru  $n \times (p+1)$ , kde prvý stĺpec zodpovedá absolútnemu členu,  $n$  je počet meraní a  $p$  je počet príznakov. Vektor predikcií modelu je potom zapísaný ako:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{c}. \quad (2.5)$$

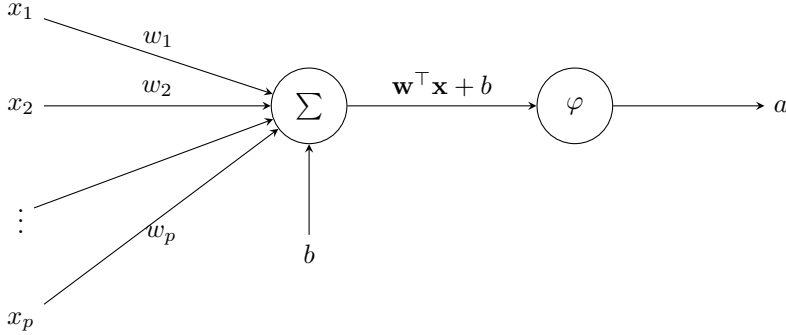
$\mathbf{y}$  je vektor cieľových hodnôt:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad (2.6)$$

V prípade, že  $n > p$ , má matica  $\mathbf{X}$  obdĺžnikový tvar, a teda jej inverzia neexistuje. Riešenie úlohy metódy najmenších štvorcov je preto vyjadrené pomocou pseudoinverznej Mooreovej-Penroseovej matice  $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  a má analytický tvar:

$$\mathbf{c} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (2.7)$$

Výhodou lineárneho regresného modelu je jeho vysoká interpretovateľnosť. Každý vstupný príznak je reprezentovaný samostatným koeficientom, ktorého znamienko určuje smer vplyvu príznaku na cieľovú premennú a jeho hodnota vyjadruje relatívnu veľkosť tohto vplyvu. Lineárna regresia určuje optimálnu lineárnu kombináciu vstupných príznakov, ktorá najlepšie aproximuje dáta. Zároveň je výpočtovo nenáročná a stabilná aj pri obmedzenom množstve dát.



Obr. 2.1: Schéma neurónu

### 2.1.3 Neurónová sieť

Neurónové siete je možné použiť na riešenie regresných úloh, v ktorých je potrebné predikovať cieľovú premennú na základe množiny vstupných príznakov. Na rozdiel od lineárnej regresie sú neurónové siete schopné aproximovať nelineárne závislosti vďaka použitiu nelineárnych aktivačných funkcií. Vo všeobecnom tvare neurónová sieť realizuje zobrazenie:

$$\hat{y} = f(\mathbf{x}; \mathbf{c}) \quad (2.8)$$

kde  $\mathbf{x} \in \mathbb{R}^p$  je vektor vstupných príznakov,  $\mathbf{c}$  predstavuje parametre siete (váhy a posuny) a  $f(\cdot)$  označuje funkciu určenú architektúrou neurónovej siete.

### 2.1.4 Architektúra neurónovej siete

Základnou jednotkou neurónovej siete je umelý neurón. Ten realizuje lineárnu kombináciu vstupov a následnú aplikáciu aktivačnej funkcie:

$$a = \varphi(\mathbf{w}^\top \mathbf{x} + b), \quad (2.9)$$

kde  $\mathbf{w} \in \mathbb{R}^p$  je vektor váh,  $b \in \mathbb{R}$  je posun a  $\varphi(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$  je aktivačná funkcia. Premenná  $a$  predstavuje výstup neurónu. Spojením viacerých neurónov pracujúcich paralelne, z ktorých každý má vlastný vektor váh a posun, vzniká lineárna vrstva. Výstup takejto vrstvy možno zapísať v maticovom tvare ako:

$$\mathbf{z} = \mathbf{C}\mathbf{x} + \mathbf{b}, \quad (2.10)$$

kde  $\mathbf{C}$  je matica váh,  $\mathbf{b}$  je vektor posunov a  $\mathbf{z}$  predstavuje lineárny výstup vrstvy. Každý riadok matice  $\mathbf{C}$  zodpovedá váham jedného neurónu.

### 2.1.5 Stratová funkcia

Stratová funkcia  $\mathcal{L}$  (angl. *loss function*) kvantifikuje rozdiel medzi predikovanými hodnotami modelu a skutočnými hodnotami. Cieľom trénovania neurónovej siete je nájsť také hodnoty váh a posunov, ktoré minimalizujú  $\mathcal{L}$ . Pre regresné úlohy sa bežne používa stredná kvadratická chyba (MSE) alebo jej odmocnina (RMSE).

### 2.1.6 Aktivačné funkcie

Po lineárnej transformácii vstupu je na výstup neurónu aplikovaná aktivačná funkcia  $\varphi(\cdot)$ . Jej úlohou je zaviesť do modelu nelinearitu. Bez aktivačných funkcií by zloženie dvoch neurónov bolo ekvivalentné jednému neurónu s novým parametrom. Ako príklad uvažujme dva neuróny bez aktivačnej funkcie, kde prvý neurón má výstup  $a_1 = w_1 \cdot x$  a druhý neurón následne  $a_2 = w_2 \cdot a_1$ . Celkové zobrazenie možno zapísať ako:

$$a_2 = w_2 \cdot w_1 \cdot x = w_{\text{nové}} \cdot x. \quad (2.11)$$

Pôvodné váhy  $w_1$  a  $w_2$  sa skladajú do jediného parametra  $w_{\text{nové}} = w_2 \cdot w_1$ . Medzi často používané aktivačné funkcie patria sigmoidálna funkcia, hyperbolický tangens a funkcia ReLU (Rectified Linear Unit). Aktivačná funkcia ReLU je definovaná vzťahom:

$$\text{ReLU}(x) = \max(0, x). \quad (2.12)$$

To znamená, že funkcia nadobúda hodnotu 0 pre záporné vstupy a lineárne rastie pre kladné hodnoty. Jej derivácia má tvar:

$$\text{ReLU}'(x) = \begin{cases} 0 & x < 0, \\ 1 & x > 0. \end{cases} \quad (2.13)$$

Na rozdiel od sigmoidálnej funkcie a hyperbolického tangensu, ktoré sú tzv. saturované funkcie (ich derivácia sa pre veľké hodnoty vstupu blíži k nule), ReLU pre kladné vstupy má deriváciu rovnú 1. Pri aktualizácii parametrov sa gradient funkcie straty počíta pomocou pravidla reťazenia a obsahuje násobenie deriváciou aktivačnej funkcie:

$$\frac{\partial \mathcal{L}}{\partial w_i} = \frac{\partial \mathcal{L}}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_i} = \frac{\partial \mathcal{L}}{\partial a} \cdot \varphi'(z) \cdot x_i, \quad (2.14)$$

kde  $\mathcal{L}$  je funkcia straty,  $w_i$  je  $i$ -ta váha neurónu,  $a = \varphi(z)$  je výstup neurónu po aplikácii aktivačnej funkcie,  $z = \mathbf{w}^\top \mathbf{x} + b$  je lineárna kombinácia vstupov,  $\varphi'(z)$  je derivácia aktivačnej funkcie a  $x_i$  je  $i$ -ta vstupná premenná. Ak je derivácia rovná 1, gradient sa pri spätnom šírení chyby nezmenšuje, čo prispieva k stabilnejšiemu a rýchlejšiemu učeniu modelu. Ďalšou výhodou ReLU je jej výpočtová jednoduchosť, keďže nevyžaduje exponenciálne operácie, ale iba porovnanie s nulou. Zároveň funkcia

ReLU zavádza určitú mieru riedkosti aktivácií, keďže pre záporné vstupy vracia nulovú hodnotu, čo môže viesť k jednoduchšej reprezentácii dát (neuróny s nulovou aktiváciou neprispievajú k výslednému výstupu). Na druhej strane táto vlastnosť môže predstavovať nevýhodu. Ak neurón produkuje záporný výstup, jeho derivácia je nulová a váhy sa ďalej neaktualizujú, čo vedie k tzv. problému "mŕtveho neurónu" (Goodfellow a kol., 2016).

### 2.1.7 Optimalizačný algoritmus Adam

Adam (Adaptive Moment Estimation) je optimalizačný algoritmus zo skupiny metód stochastického gradientného zostupu s adaptívnou rýchlosťou učenia (Kingma and Ba, 2014). Na rozdiel od základného gradientného zostupu Adam prispôbuje rýchlosť učenia pre každý parameter individuálne na základe histórie gradientov. Cieľom tréningu je minimalizovať stratovú funkciu  $\mathcal{L}(\mathbf{c})$ , kde  $\mathbf{c}$  je vektor všetkých parametrov modelu. V každom kroku  $t$  sa vypočíta gradient  $\mathbf{g}_t = \nabla_{\mathbf{c}} \mathcal{L}(\mathbf{c}_t)$ , ktorý udáva smer najstrmšieho rastu funkcie straty. Adam na jeho základe aktualizuje odhady prvého momentu  $\hat{s}_t$  a druhého momentu  $\hat{r}_t$ . Parametre sa potom aktualizujú ako:

$$\mathbf{c}_{t+1} = \mathbf{c}_t - \frac{\alpha}{\sqrt{\hat{r}_t} + \epsilon} \hat{s}_t, \quad (2.15)$$

kde  $\alpha$  je učiaca rýchlosť a  $\epsilon$  je malá konštanta zabraňujúca deleniu nulou.

### 2.1.8 Dropout

Dropout je metóda regularizácie neurónových sietí, ktorá pomáha predchádzať pretrénovaniu modelu (angl. *overfitting*) a zlepšuje jeho schopnosť generalizácie (Srivastava a kol., 2014). Jej princíp spočíva v náhodnej deaktivácii časti neurónov počas každého tréningového kroku s určitou pravdepodobnosťou. Deaktivované neuróny neprispievajú k výpočtu ani k aktualizácii váh v danom kroku, čo núti sieť sa učiť robustnejšie a znižuje závislosť medzi jednotlivými neurónmi. Výsledkom je model, ktorý sa lepšie generalizuje na nové dáta. Počas testovania sú všetky neuróny aktívne.

### 2.1.9 Krivky učenia

Krivky učenia predstavujú dôležitý nástroj na analýzu procesu tréningu modelov strojového učenia. Ide o grafické znázornenie hodnoty stratovej funkcie (*loss*) v závislosti od počtu epoch počas tréningu. Najčastejšie sa zobrazujú dve krivky: tréningová krivka (*training loss*) a validačná krivka (*validation loss*) (Goodfellow a kol., 2016).

Krivky učenia umožňujú identifikovať, či sa model učí správne, alebo dochádza k

problémom, ako je podučenie (angl. *underfitting*) alebo pretrénovanie (angl. *overfitting*).

Pri správnom učení by mali obe krivky postupne klesať a stabilizovať sa na podobných hodnotách. Pretrénovanie sa typicky prejavuje tak, že trénovacia chyba naďalej klesá, zatiaľ čo validačná chyba sa prestane zlepšovať, začne stagnovať alebo dokonca rásť. Charakteristickým znakom je aj výrazné rozchádzanie sa trénovacej a validačnej krivky. V takom prípade je vhodné trénovanie zastaviť pomocou predčasného zastavenia učenia (angl. *early stopping*), aby sa predišlo zhoršeniu schopnosti modelu generalizovať na nové dáta.

Naopak, ak obe krivky zostávajú na vysokej hodnote bez výrazného poklesu, ide pravdepodobne o podučenie, čo znamená, že model nie je dostatočne komplexný alebo nebol trénovaný dostatočne dlho.

## 2.2 Ukazovatele presnosti modelu

Na hodnotenie kvality regresného modelu je potrebné použiť kvantitatívne ukazovatele, ktoré merajú rozdiel medzi predikovanými a skutočnými hodnotami cieľovej premennej. Pri regresných úlohách sa často používajú metriky ako Mean Squared Error (MSE), Root Mean Squared Error (RMSE),  $R^2$  a štandardná odchýlka (STD). Tieto ukazovatele vyjadrujú mieru odchýlky predikcií modelu od skutočných hodnôt.

### 2.2.1 Odmocnina strednej kvadratickej chyby (RMSE)

RMSE je jedna z najčastejšie používaných metrík kvality pre regresné úlohy. Udáva priemernú veľkosť odchýlky predikovaných hodnôt od nameraných hodnôt. RMSE je definovaná ako:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (2.16)$$

kde  $y_i$  je nameraná hodnota,  $\hat{y}_i$  je predikcia modelu,  $n$  je počet pozorovaní, a  $i$  označuje index pozorovania. Čím je hodnota RMSE menšia, tým je presnosť modelu vyššia. Metodika je citlivá na veľké chyby, čo ju robí vhodnou v úlohách, kde sú veľké odchýlky nežiaduce.

### 2.2.2 $R^2$ (koeficient determinácie)

Koeficient determinácie  $R^2$  vyjadruje, ako dobre model predikuje cieľovú premennú v porovnaní s referenčným modelom, ktorý vždy predikuje aritmetický priemer hodnoty z celého datasetu.

Koeficient  $R^2$  sa vypočíta podľa:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (2.17)$$

kde  $\bar{y}$  je aritmetický priemer cieľovej premennej a ostatné symboly sú definované rovnako ako v rovnici pre RMSE. Hodnota  $R^2$  blízka 1 indikuje, že model vysvetľuje väčšinu variability dát. Hodnota blízka 0 znamená, že model nie je lepší než model založený na výpočte priemernej hodnoty. Záporné hodnoty naznačujú, že model dosahuje horšie výsledky než tento jednoduchý model.

### 2.2.3 Štandardná odchýlka

Štandardná odchýlka (STD) je štatistická miera variability, ktorá vyjadruje, ako veľmi sa jednotlivé hodnoty odlišujú od ich priemeru. Používa sa na kvantifikáciu rozptylu dát a poskytuje informáciu o stabilite sledovanej veličiny.

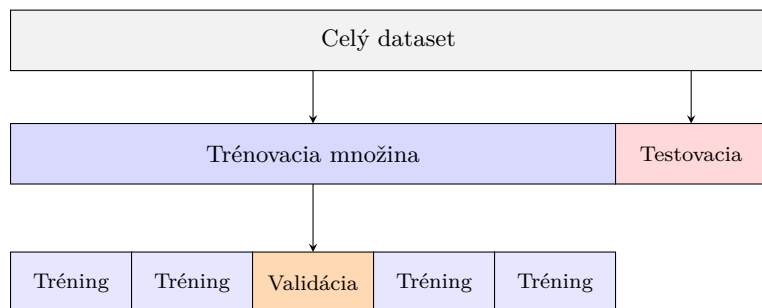
Pre množinu hodnôt  $\{x_1, x_2, \dots, x_n\}$  je štandardná odchýlka definovaná ako:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}, \quad (2.18)$$

kde  $\mu$  je aritmetický priemer daných hodnôt:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i. \quad (2.19)$$

Štandardná odchýlka nadobúda nezáporné hodnoty, pričom vyššia hodnota indikuje väčšiu variabilitu dát, zatiaľ čo nižšia hodnota znamená, že hodnoty sú bližšie k priemeru.



Príklad jedného rozdelenia  $k$ -násobnej krížovej validácie

**Obr. 2.2:** Schéma rozdelenia datasetu na trénovaciu, testovaciu a validačnú množinu

## 2.3 Metodológia

### 2.3.1 Rozdelenie datasetu

Pred trénovaním modelu sa dataset rozdeľuje na trénovaciu a testovaciu množinu. Trénovacia množina slúži na odhadovanie parametrov modelu. Testovacia množina slúži na záverečné hodnotenie schopnosti modelu generalizovať na nové dáta a poskytuje odhad jeho výkonnosti pri reálnom nasadení. Pri použití krížovej validácie sa trénovacia množina rozdeľuje na  $k$  častí, pričom jedna z nich slúži ako validačná množina a zvyšné tvoria trénovaciu množinu pre daný krok (viď podkapitola 2.3.4). Rozdelenie datasetu sa riadi hodnotou generátora náhodných čísel (angl. *seed*, ďalej len *hodnota generátora rozdelení*), ktorá zabezpečuje deterministické a reprodukovateľné rozdelenie dát.

### 2.3.2 Normalizácia dát

Dôvodom normalizácie je skutočnosť, že jednotlivé vstupné príznaky majú rozdielne fyzikálne jednotky a rozsahy hodnôt, čo môže negatívne ovplyvniť proces učenia neurónovej siete. Príznaky s veľkými absolútnymi hodnotami môžu dominovať pri výpočte gradientov. Napríklad, ak jeden príznak nadobúda hodnoty rádovo okolo 1 a iný rádovo okolo 1000, optimalizačný algoritmus môže priradovať výrazne odlišné váhy týmto príznakom, čo môže viesť k neúmernému zväčšovaniu váh (Goodfellow a kol., 2016). Normalizácia zabezpečuje, že všetky vstupné príznaky sú v rovnakom číselnom rozsahu, čím sa zlepšuje rýchlosť konvergencie optimalizačných algoritmov.

V tejto práci bola použitá min-max normalizácia, pri ktorej sú hodnoty jednotlivých

príznakov transformované do intervalu  $(0,1)$  podľa vzťahu:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \quad (2.20)$$

kde  $x$  predstavuje pôvodnú hodnotu príznaku,  $x_{\min}$  a  $x_{\max}$  sú minimálna a maximálna hodnota daného príznaku a  $x_{\text{norm}}$  je normalizovaná hodnota.

### 2.3.3 Vplyv úniku informácií pri normalizácii

Parametre normalizácie  $x_{\min}$  a  $x_{\max}$  musia byť odhadnuté výlučne z trénovacej množiny a následne aplikované aj na validačnú a testovaciu množinu (viď podkapitola 3.2.2). Použitie parametrov odhadnutých z celého datasetu vrátane testovacích dát by predstavovalo únik informácií (*data leakage*) — model by nepriamo získal informáciu o rozsahu testovacích dát ešte pred hodnotením, čo by viedlo k príliš optimistickému odhadu jeho skutočnej výkonnosti. Miera vplyvu prípadného úniku informácií závisí od rozdielu medzi parametrami normalizácie odhadnutými na celom datasete a parametrami odhadnutými výlučne na trénovacej množine. V prípadoch, keď trénovacia množina pokrýva väčšinu rozsahu hodnôt celého datasetu, sú tieto rozdiely minimálne a vplyv na výsledné metriky je zanedbateľný. Naopak, výrazný vplyv možno očakávať vtedy, keď sa v testovacej množine nachádzajú hodnoty výrazne vybočujúce z rozsahu trénovacej množiny.

### 2.3.4 Krížová validácia

Krížová validácia je metóda štatistického hodnotenia modelu, pri ktorej sa dáta opakovane rozdeľujú na trénovacie a validačné podmnožiny (viď podkapitola 2.3.1). Najčastejšie používaným variantom je  $k$ -násobná krížová validácia, pri ktorej sa dataset rozdelí na  $k$  rovnakých častí (angl. *fold*). Model sa postupne trénuje na  $k - 1$  častiach a validuje sa na zostávajúcej časti, pričom tento proces sa opakuje  $k$ -krát tak, aby každá časť slúžila práve raz ako validačná množina.

Výsledná kvalita modelu sa určí ako priemer hodnotených metrík zo všetkých  $k$  behov. Týmto spôsobom sa znižuje vplyv náhodného rozdelenia dát, získava sa stabilnejší odhad kvality modelu a umožňuje sa objektívne porovnanie viacerých modelov.

### 2.3.5 Opakovaná krížová validácia

Opakovaná krížová validácia je rozšírením štandardnej  $k$ -násobnej krížovej validácie. Celý postup krížovej validácie sa opakuje viackrát, pričom pred každým opakovaním

sú dáta premiešané s iným poradím. Každé opakovanie tak produkuje iné rozdelenie dát na časti, čo ďalej znižuje závislosť výsledku od konkrétneho rozdelenia. Výsledná metrika je priemer hodnotení zo všetkých opakovaní a všetkých častí, čo poskytuje stabilnejší a spoľahlivejší odhad kvality modelu než jednorazová krížová validácia.

### 2.3.6 Analýza korelácie

Vzťahy medzi premennými možno analyzovať pomocou Pearsonovho korelačného koeficientu, ktorý vyjadruje mieru lineárnej závislosti a je definovaný vzťahom:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (2.21)$$

kde  $x_i$  a  $y_i$  sú hodnoty dvoch skúmaných premenných,  $\bar{x}$  a  $\bar{y}$  sú ich priemerné hodnoty,  $n$  je počet pozorovaní a  $i$  označuje index pozorovania. Koeficient  $r_{xy}$  nadobúda hodnoty v intervale  $\langle -1, 1 \rangle$ , kde 0 znamená žiadnu lineárnu závislosť a čím viac  $|r_{xy}| \rightarrow 1$ , tým silnejšia je lineárna závislosť.

Cieľom analýzy je identifikovať príznaky, ktoré majú silnú závislosť s cieľovou premennou, a zároveň odhaliť vzájomne silne korelované príznaky, ktoré môžu spôsobovať multikolinearitu (viď podkapitola 2.3.7).

Pri detekcii multikolinearity pomocou korelačného koeficientu je potrebné určiť hraničnú hodnotu  $|r|$ , nad ktorou sa dvojica príznakov považuje za nadmerne závislú. Táto voľba nemá jednoznačné teoretické zdôvodnenie. Podľa Gujarati (2009) je multikolinearita otázkou miery (*question of degree*), nie kvalitatívnou vlastnosťou. Autori uvádzajú empirické pravidlo, podľa ktorého korelácia presahujúca  $|r| > 0,8$  označuje problematickú multikolinearitu, ktorá vedie k nestabilite odhadov parametrov lineárnych modelov.

### 2.3.7 Multikolinearita

Multikolinearita je jav, pri ktorom sú dva alebo viac vstupných príznakov v silnej lineárnej závislosti. V takom prípade pre každý príznak  $x_i$  existuje aproximácia vo forme lineárnej kombinácie ostatných príznakov:

$$x_i \approx \sum_{\substack{j=1 \\ j \neq i}}^p \alpha_{ij} x_j, \quad i = 1, \dots, p. \quad (2.22)$$

Ak má matica  $\mathbf{X}$  dva stĺpce, medzi ktorými je absolútna hodnota korelačného koeficientu rovná 1, potom medzi nimi je presná lineárna závislosť. Za tejto podmienky matica  $\mathbf{X}$

stráca hodnotu a je matica  $\mathbf{X}^\top \mathbf{X}$  singulárna, a teda:

$$\det(\mathbf{X}^\top \mathbf{X}) = 0. \quad (2.23)$$

Vtedy je aspoň jedno vlastné číslo matice  $\mathbf{X}^\top \mathbf{X}$  rovné nule ( $\lambda_i = 0$ , kde  $i = 1, 2, \dots, p$ ) a inverzná matica  $(\mathbf{X}^\top \mathbf{X})^{-1}$  neexistuje. V takom prípade nemožno použiť nasledovný tvar riešenia metódy najmenších štvorcov:

$$\mathbf{c} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (2.24)$$

V praxi sa však presná lineárna závislosť vyskytuje zriedkavo. Častejším javom je silná korelácia medzi príznakmi, napríklad meranie teploty v susedných bodoch zariadenia. V takom prípade je matica  $\mathbf{X}^\top \mathbf{X}$  regulárna, ale niektoré jej vlastné čísla sú veľmi malé. Maticu  $\mathbf{X}^\top \mathbf{X}$  možno rozložiť pomocou vlastných hodnôt a vlastných vektorov na základe spektrálnej vety:

$$\mathbf{X}^\top \mathbf{X} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top, \quad (2.25)$$

kde

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}, \quad (2.26)$$

pričom

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0, \quad (2.27)$$

a  $\mathbf{V}$  je ortogonálna matica vlastných vektorov. Inverzia matice má potom tvar:

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \mathbf{V} \mathbf{\Lambda}^{-1} \mathbf{V}^\top, \quad (2.28)$$

kde

$$\mathbf{\Lambda}^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\lambda_p} \end{bmatrix}. \quad (2.29)$$

Dosadením do vzťahu pre metódu najmenších štvorcov dostávame

$$\mathbf{c} = \mathbf{V} \mathbf{\Lambda}^{-1} \mathbf{V}^\top \mathbf{X}^\top \mathbf{y}. \quad (2.30)$$

Ak je niektoré vlastné číslo  $\lambda_i$  veľmi malé ( $\lambda_i \approx 0$ ), potom hodnota  $\frac{1}{\lambda_i}$  je veľmi veľká. Tento faktor sa následne prenáša do vektora parametrov  $\mathbf{c}$ . V dôsledku toho aj malé zmeny vstupných dát môžu viesť k výrazným zmenám odhadnutých koeficientov, čo spôsobuje nestabilitu tréningu modelu.

### 2.3.8 Odporúčaný pomer parametrov a počtu pozorovaní

Pri návrhu modelov strojového učenia sa v praxi často používa empirické odporúčanie, podľa ktorého by počet dát mal byť rádovo väčší než počet parametrov modelu. Teoretické zdôvodnenie tohto pozorovania uvádzajú (Baum and Haussler, 1989), ktorí ukazujú, že počet potrebných dát rastie približne ako

$$m = O\left(\frac{W}{\varepsilon} \log \frac{N}{\varepsilon}\right), \quad (2.31)$$

kde  $W$  je počet váh v sieti,  $N$  počet neurónov a  $\varepsilon$  požadovaná chyba generalizácie.

Pri zanedbaní logaritmického člena dostávame praktický odhad

$$m \approx \frac{W}{\varepsilon}. \quad (2.32)$$

Je potrebné zdôrazniť, že ide o heuristické zjednodušenie. Uvedený vzťah nepredstavuje presný teoretický výsledok, ale praktické odporúčanie, podľa ktorého by mal byť počet dát výrazne väčší než počet parametrov modelu, aby sa znížilo riziko pretrénovania a zabezpečila dobrá generalizácia (Goodfellow a kol., 2016). Napríklad, pre hodnotu  $\varepsilon = 0,1$  vychádza, že počet dát by mal byť približne desaťnásobok počtu parametrov modelu. Toto odporúčanie je v uvedenej práci spomínané ako Widrowovo empirické pravidlo (Baum and Haussler, 1989).

Počet parametrov modelu závisí od jeho štruktúry. Napríklad pre lineárnu regresiu s  $p$  príznakmi je počet parametrov  $p + 1$ . Pre neurónovú sieť s jednou skrytou vrstvou,  $p$  vstupmi a  $h$  neurónmi v skrytej vrstve je počet parametrov

$$\text{param} = (p + 1)h + (h + 1) = h(p + 2) + 1. \quad (2.33)$$

Nedodržanie tohto odporúčania však automaticky neznamená pretrénovanie modelu. V praxi môže byť požadovaný počet dát výrazne nižší, najmä pri použití regularizačných techník, ako sú napríklad dropout (viď podkapitola 2.1.8) alebo predčasné zastavenie tréningu (*early stopping*). Predčasné zastavenie tréningu monitoruje hodnotu stratovej funkcie na validačnej množine a ukončí tréning v prípade, že sa táto hodnota nezlepší počas určitého počtu po sebe nasledujúcich epoch (angl. *patience*).

## 2.4 Aplikácia Phiture ako nástroj pre vývoj softvérových senzorov

Phiture je webová aplikácia navrhnutá ako výskumný asistent pre vývoj softvérových senzorov a iných modelov. Zdrojový kód aplikácie je dostupný na platforme GitHub.<sup>1</sup> Hlavným cieľom aplikácie je urýchlenie rešeršnej a prototypovacej fázy pri návrhu modelov. Aplikácia automaticky vyhľadáva relevantné vedecké publikácie z databáz, ako sú arXiv alebo Springer, následne z PDF dokumentov extrahuje matematické výrazy a prevádza ich do spustiteľného Python kódu. Takýto postup umožňuje identifikovať kombinácie fyzikálnych veličín vhodné pre návrh príznakov softvérových senzorov, overiť ich použitie a integrovať ich do vlastnej implementácie modelov.

### 2.4.1 Architektúra a pracovný postup

Aplikácia pozostáva z viacerých modulov, ktoré zabezpečujú jednotlivé fázy spracovania vedeckých publikácií. Používateľ postupne prechádza procesom vyhľadávania, správy stiahnutých dokumentov, extrakcie matematických výrazov a ich konverzie do Python kódu. Výstupom celého procesu je spustiteľný kód pripravený na okamžité použitie.

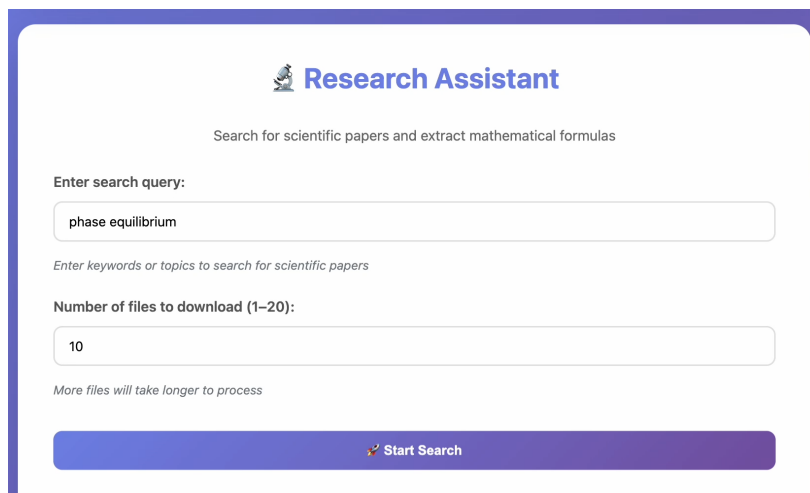
**Vyhľadávanie publikácií.** Používateľ zadá oblasť záujmu a množinu kľúčových slov prostredníctvom formulára na stránke *Search*. Aplikácia následne odošle požiadavky na externé databázy vedeckých publikácií a získa zoznam relevantných článkov podľa zadaných výrazov. Výsledky vyhľadávania slúžia ako vstup pre ďalšie spracovanie. Ukážka modulu vyhľadávania je na obrázku 2.3.

**Sťahovanie a správa PDF dokumentov.** Po výbere relevantných publikácií aplikácia automaticky stiahne príslušné PDF dokumenty. Proces sťahovania prebieha na pozadí a po jeho dokončení sa používateľovi zobrazí zoznam úspešne stiahnutých súborov.

Používateľ môže jednotlivé dokumenty odstrániť, prípadne doplniť ďalšie PDF súbory pred začiatkom spracovania. Týmto spôsobom je možné vytvoriť kolekciu článkov určených na následnú extrakciu matematických výrazov (obrázok 2.4).

---

<sup>1</sup><https://github.com/semyon-suslov/phiture>



**Research Assistant**

Search for scientific papers and extract mathematical formulas

Enter search query:

phase equilibrium

Enter keywords or topics to search for scientific papers

Number of files to download (1–20):

10

More files will take longer to process

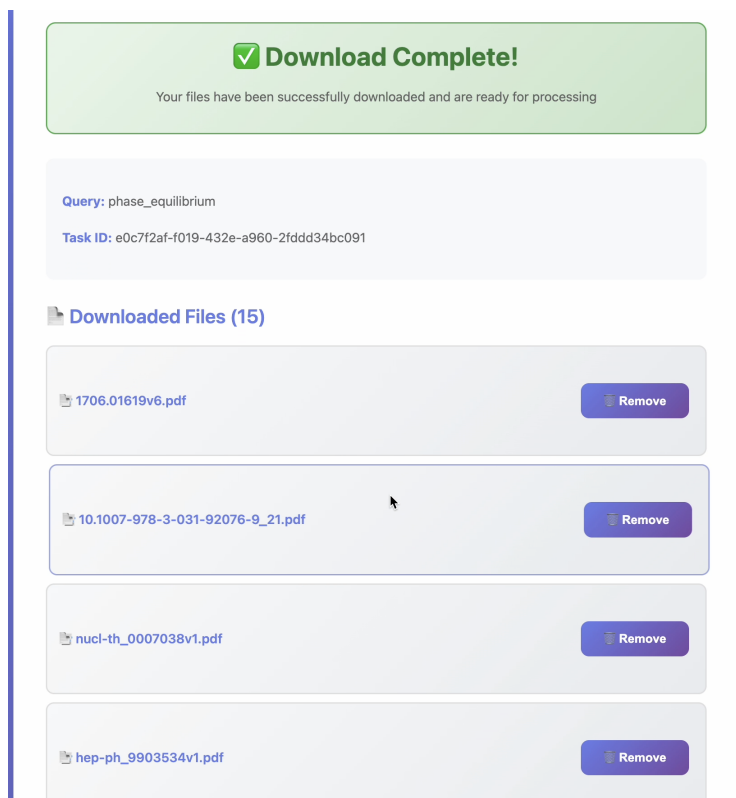
 Start Search

**Obr. 2.3:** Modul vyhľadávania vedeckých publikácií.

**Extrakcia príznakov.** Po výbere dokumentov aplikácia spracuje jednotlivé PDF súbory a pomocou OCR a analýzy dokumentu identifikuje matematické výrazy nachádzajúce sa v texte publikácií. Rozpoznané výrazy sú následne zobrazené v prehľadnom zozname a používateľ môže označiť tie, ktoré považuje za relevantné pre návrh príznakov alebo fyzikálnych modelov softvérového senzora. Výsledok extrakcie matematických výrazov je znázornený na obrázku 2.5.

**Konverzia LaTeX zápisu do Python kódu.** Vybraný matematický výraz je následne odoslaný do modulu `ConverterLatex`, ktorého úlohou je previesť LaTeX zápis na Python kód využívajúci knižnicu `sympy`. Modul najskôr analyzuje štruktúru LaTeX výrazu, identifikuje premenné a matematické operácie a následne vytvorí ekvivalentný symbolický zápis v jazyku Python. Výsledný kód môže byť použitý na symbolické výpočty, numerické simulácie alebo ďalšie úpravy modelu.

Súčasťou modulu je aj možnosť okamžitého spustenia vygenerovaného kódu pomocou tlačidla *Run Python Code*. Týmto spôsobom je možné overiť, či bol matematický výraz korektne prevedený do Python syntaxe a či nevznikli chyby pri parsovaní alebo interpretácii vzorca. Ukážka modulu konverzie je na obrázku 2.6.



**Obr. 2.4:** Stránka po dokončení sťahovania PDF dokumentov. Zobrazuje počet úspešne stiahnutých súborov a umožňuje ich správu pred ďalším spracovaním.

Search

LaTeX Converter

Topic Suggestions

My searches

Logout

Extracted Formulas

10.1007-978-981-95-1487-8\_9.pdf

✓

$$[M + A(\infty)]\ddot{x}_1 + \int_0^t h(t - \tau)\dot{x}_1(\tau)d\tau + D_1x_1 + kx_1 = F_1 \quad (1)$$

○

$$F = \frac{1}{2}\rho C_D D(U - \dot{X})|U - \dot{X}| + \rho C_M \frac{\pi D^2}{4} \frac{\partial U}{\partial t} - \rho C_m \frac{\pi D^2}{4} \ddot{X} \quad (2)$$

✓

$$m\ddot{x}_2 + D_2\dot{x}_2 = (\rho_w V + C_m)a - m\ddot{x}_2 + C_D|u - \dot{x}_2|(u - \dot{x}_2) \quad (3)$$

**Obr. 2.5:** Modul extrakcie matematických výrazov z PDF dokumentov. Používateľ môže označiť relevantné výrazy určené na ďalšie spracovanie.

### LaTeX to Python Converter

Enter LaTeX Formula:

$$m\ddot{x}_2 + D_2\dot{x}_2 = (\rho_w V + C_m)a - m\ddot{x}_2 + C_D|u - \dot{x}_2|(u - \dot{x}_2) \quad (3)$$

✦ Convert to Python

### Converted Python Code

```
from sympy import symbols, Eq, Abs, solve, Function, Derivative

# This expression represents the Morison equation for hydrodynamic force on a body.

# symbolic computation
from sympy import symbols, Eq, solve, Abs

# define variables
# Note: x_2_dot and x_2_ddot represent the first and second time derivatives of x_2
m, D_2, rho_w, V, C_m, a, C_D, u = symbols('m D_2 rho_w V C_m a C_D u')
x_2_dot, x_2_ddot = symbols('x_2_dot x_2_ddot')

# equations and conditions
lhs = m * x_2_ddot + D_2 * x_2_dot
rhs = (rho_w * V + C_m) * a - m * x_2_ddot + C_D * Abs(u - x_2_dot) * (u - x_2_dot)
equation = Eq(lhs, rhs)
```

▶ Run Python Code

Obr. 2.6: Modul konverzie LaTeX zápisu do Python kódu pomocou knižnice `sympy`.

### 2.4.2 Využitie pri návrhu softvérových senzorov

Aplikácia Phiture bola v rámci tejto práce využívaná ako podporný nástroj počas rešeršnej a prototypovacej fázy návrhu softvérových senzorov. Pomocou aplikácie bolo možné efektívne vyhľadávať publikácie z oblasti destilačných kolón, analyzovať publikované fyzikálne modely a identifikovať vhodné kombinácie procesných veličín použiteľné ako vstupné príznaky modelov.

Významným prínosom aplikácie je automatizácia viacerých časovo náročných krokov, ktoré sa pri klasickom postupe vykonávajú manuálne. Ide najmä o vyhľadávanie publikácií, identifikáciu relevantných matematických výrazov a ich prevod do podoby použiteľnej v programovacom jazyku Python. Takýto prístup umožňuje rýchlejšie experimentovanie s rôznymi kombináciami príznakov a urýchľuje implementáciu a testovanie softvérových senzorov.

## Experimentálna časť

---

Táto kapitola popisuje návrh a vyhodnotenie softvérových senzorov na reálnom priemyselnom datasete. Kapitola je rozdelená do piatich etáp. Každá etapa sa líši výberom vstupných príznakov. V každej etape sú hodnotené dva modely: lineárna regresia a neurónová sieť. Cieľom je vybrať vhodné príznaky a architektúru neurónovej siete, posúdiť vplyv výberu príznakov na presnosť predikcie a na základe výsledkov identifikovať najvhodnejší softvérový senzor spomedzi navrhnutých modelov pre daný dataset. Na záver je vykonané vyčerpávajúce prehľadávanie priestoru príznakov, ktoré overuje, či zvolený postup výberu viedol k optimálnemu riešeniu.

### 3.1 Opis datasetu

V rámci práce bol použitý reálny dataset z prevádzky destilačnej kolóny v spoločnosti Sloznaft, a.s., ktorý obsahuje rôzne procesné veličiny a laboratórne merania.

#### 3.1.1 Destilačná kolóna

Destilačná kolóna je technologické zariadenie určené na separáciu zmesí kvapalín na základe rozdielnej prchavosti jednotlivých zložiek. Separácia prebieha prostredníctvom viacnásobného kontaktu parnej a kvapalnej fázy na jednotlivých etážach kolóny, čo umožňuje postupné obohacovanie pár o prchavejšie zložky a kvapaliny o menej prchavé komponenty. Konkrétne ide o depropanizátor, ktorý je súčasťou jednotky katalytického krakovania FCC (angl. *Fluid Catalytic Cracking*). Vstupná zmes pozostáva z deviatich uhľovodíkových zložiek v rozsahu  $C_3 - C_5$ . Kolóna rozdeľuje tento tok na dva produkty: horný produkt (destilát) obohatený o  $C_3$  a dolný produkt obohatený o  $C_4 - C_5$  (Mojto a kol., 2021), ktorého zloženie  $x_B$  je cieľovou premennou tejto práce.

### 3.1.2 Opis premenných

Dataset obsahuje 165 pozorovaní cieľovej premennej, ktorou je molárny zlomok produktu na dne kolóny  $x_B$ . Tieto hodnoty boli získané laboratórnymi meraniami vykonávanými približne každé tri dni. Odbery vzoriek sa realizovali v ranných hodinách, keď sa systém nachádzal v stave blízkom ustálenému stavu, t.j. bez významných zmien v prevádzkových podmienkach počas predchádzajúcich nočných hodín.

V rámci datasetu je cieľová premenná označená ako  $\mathbf{y}$  a je reprezentovaná ako vektor

$$\mathbf{y} = \begin{bmatrix} x_{B_1} \\ x_{B_2} \\ \vdots \\ x_{B_{165}} \end{bmatrix}. \quad (3.1)$$

Okrem cieľovej premennej dataset obsahuje aj maticu vstupných príznakov rozmeru  $38396 \times 16$ , získaných z meraní a prevádzkových veličín destilačnej kolóny. Tieto príznaky reprezentujú procesné veličiny charakterizujúce stav destilačnej kolóny, najmä tlakové, teplotné, prietokové a energetické parametre. Konkrétne:  $p_B$  (tlak na dne kolóny),  $T_B$  (teplota zvyšku),  $T_{c,B}$  (teplota chladiaceho média na dne kolóny),  $p_D$  (tlak na vrchu kolóny),  $T_D$  (teplota destilátu),  $T_{c,D}$  (teplota chladiaceho média na vrchu kolóny),  $F$  (prietok vstupnej suroviny),  $R$  (refluxný pomer),  $F_{B,1}$  (prietok ohrevného média do prvého vyparníka/varáka),  $T_{B,1,in}$  (vstupná teplota v prvom výmenníku),  $T_{B,1,out}$  (výstupná teplota v prvom výmenníku),  $F_{B,2}$  (prietok druhého bočného odberu),  $T_{B,2,in}$  (vstupná teplota v druhom výmenníku),  $T_{B,2,out}$  (výstupná teplota v druhom výmenníku),  $D$  (prietok destilátu) a  $Q_B$  (tepelný výkon privádzaný do varáka).

Vstupné príznaky sú v datasete reprezentované ako matica  $\mathbf{X}$ . Matica vstupných príznakov má tvar:

$$\mathbf{X} = \begin{bmatrix} p_{B,1} & T_{B,1} & \cdots & Q_{B,1} \\ p_{B,2} & T_{B,2} & \cdots & Q_{B,2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{B,165} & T_{B,165} & \cdots & Q_{B,165} \end{bmatrix}. \quad (3.2)$$

Po zlúčení so súborom laboratórných meraní boli z datasetu odstránené všetky záznamy, ktoré časovo nezodpovedali časom odberu vzoriek. Takýto postup bol možný vzhľadom na to, že senzorické dáta predstavujú časovo spriemerované hodnoty za jednotlivé 30-minútové intervaly a čas odberu laboratórných vzoriek bol presne známy. V dôsledku toho bolo možné jednoznačne priradiť jednotlivé merania zo senzorov k

príslušným laboratórnym meraniam, čím vznikol výsledný dataset s korektne zosúladenými vstupnými príznakmi a cieľovou premennou, kde matica  $\mathbf{X}$  má rozmer  $165 \times 16$  a vektor  $\mathbf{y}$  má rozmer  $165 \times 1$ . Hlavným cieľom je využiť informácie z procesných senzorov na čo najpresnejšiu predikciu laboratórných hodnôt.

## 3.2 Implementácia práce

Keďže neexistujú striktne stanovené pravidlá pre návrh architektúry neurónových sietí, v tejto práci boli ako východiskový bod využité empirické odporúčania z odbornej literatúry. Následne boli testované rôzne konfigurácie modelov s cieľom nájsť vhodné riešenie pre daný problém.

### 3.2.1 Použité nástroje a knižnice

Experiment bol realizovaný v dvoch programovacích prostrediach.

Analýza korelácií, tvorba nových príznakov a vizualizácia tepelných máp korelačných matíc boli implementované v prostredí MATLAB. Toto prostredie bolo zvolené vzhľadom na dostupnosť pôvodného datasetu vo formáte `.mat` a na možnosť jednoduchšej vizualizácie výsledkov korelačnej analýzy.

Trénovanie a hodnotenie modelov bolo realizované v jazyku Python. Neurónová sieť bola implementovaná pomocou knižnice `PyTorch`, ktorá poskytuje flexibilné rozhranie pre návrh architektúry, trénovanie a regularizáciu modelov. Lineárna regresia, krížová validácia a škálovanie príznakov boli realizované pomocou knižnice `scikit-learn`. Načítanie datasetu z formátu `.mat` bolo zabezpečené funkciou `scipy.io.loadmat` z knižnice `SciPy`. Na manipuláciu s dátami a ukladanie výsledkov boli použité knižnice `pandas` a `numpy`, prehľadávanie všetkých kombinácií príznakov bolo realizované pomocou modulu `itertools`. Zdrojový kód experimentu je dostupný v repozitári GitHub.<sup>1</sup>

### 3.2.2 Rozdelenie datasetu

Dataset obsahujúci 165 pozorovaní bol rozdelený na trénovaciu a testovaciu množinu v pomere 80/20, čo zodpovedá 132 trénovacím a 33 testovacím pozorovaniam (viď podkapitola 2.3.1). Rozdelenie bolo realizované pomocou funkcie `train_test_split` z

---

<sup>1</sup><https://github.com/semyon-suslov/bachelor-thesis-code>

knižnice `scikit-learn`. Táto funkcia prijíma parameter `random_state` (viď podkapitola 2.3.1).

Rozdelenie bolo zopakované pre 5 rôznych hodnôt generátora rozdelení (34, 42, 7, 99, 123), vďaka čomu vzniká celkovo 125 nezávislých hodnotení na konfiguráciu (viď podkapitola 2.3.5). Hodnoty generátora rozdelení boli zvolené náhodne a následne fixované. Ich fixácia zabezpečuje, že všetky porovnávané modely sú hodnotené na identických rozdeleniach dát, čo je nevyhnutnou podmienkou korektného porovnania.

### 3.2.3 Normalizácia dát

Bola použitá metóda min-max normalizácie (viď podkapitola 2.3.2). Normalizácia bola realizovaná pomocou knižnice `scikit-learn` v jazyku Python, konkrétne s využitím triedy `MinMaxScaler`.

Keďže pôvodné nenormalizované hodnoty neboli dostupné, normalizácia nemohla byť realizovaná výlučne na základe trénovacej množiny. Trénovacia množina tvorí 80 % datasetu a pokrýva väčšinu rozsahu hodnôt, preto je vplyv tohto obmedzenia zanedbateľný (viď podkapitola 2.3.3).

V etapách 1 a 2 pochádzajú všetky príznaky z normalizovaných senzorových dát v rozsahu  $\langle 0, 1 \rangle$ . V etapách 3 a 4 môžu konštruované príznaky nadobúdať hodnoty mimo tohto rozsahu: teplotné rozdiely a logaritmy tlakov môžu byť záporné. Z tohto dôvodu bola pred trénovaním neurónových sietí v etapách 3 a 4 aplikovaná `MinMax` normalizácia príznakov.

### 3.2.4 Implementácia krížovej validácie

Krížová validácia bola realizovaná výlučne na trénovacej množine (132 pozorovaní). Podľa James a kol. (2013) sa pri  $k$ -násobnej krížovej validácii zvyčajne odporúča voliť hodnotu  $k = 5$  alebo  $k = 10$ , ktoré predstavujú štandardnú empirickú voľbu.

Vo všeobecnosti platí, že voľba  $k$  predstavuje kompromis medzi nedostatočným natréňovaním modelu a rozptylom odhadu. Pri vysokých hodnotách  $k$  obsahuje trénovacia množina viac dát, avšak validačná množina je veľmi malá. To znamená, že výsledky validácie sú silne závislé od konkrétneho rozdelenia dát, čo vedie k vyššiemu rozptylu odhadov. Naopak, pri nízkych hodnotách  $k$  je validačná množina väčšia, ale model sa trénuje na menšom počte vzoriek, čo môže viesť k nedostatočnému naučeniu a k podhodnoteniu výkonnosti modelu.

**Tabuľka 3.1:** Porovnanie výsledkov krížovej validácie pre rôzne hodnoty  $k$ 

| $k$ | CV RMSE | CV RMSE STD | CV $R^2$ |
|-----|---------|-------------|----------|
| 5   | 0,1271  | 0,02452     | 0,638    |
| 10  | 0,1254  | 0,03282     | 0,604    |

Pre výber vhodnej hodnoty  $k$  bolo vykonané experimentálne porovnanie pre  $k \in \{5, 10\}$  na neurónovej sieti s jednou skrytou vrstvou a 6 neurónmi. Výsledky sú uvedené v tabuľke 3.1. Rozdiely medzi oboma nastaveniami sú minimálne. Keďže dosiahnuté metriky sú prakticky porovnateľné, bola zvolená hodnota  $k = 5$  ako najjednoduchšia konfigurácia. Táto voľba poskytuje nižšie výpočtové náklady pri zachovaní podobnej kvality odhadu ako pri  $k = 10$  a menší rozptyl odhadov.

Namiesto jednorazovej 5-násobnej krížovej validácie bola použitá opakovaná krížová validácia (viď podkapitola 2.3.5), implementovaná pomocou `RepeatedKfold` z knižnice `scikit-learn`. Počet opakovaní bol zvolený ako 5, čím vzniká celkovo  $5 \times 5 = 25$  hodnotení na jednu hodnotu generátora rozdelení.

### 3.2.5 Výber hlavnej metriky hodnotenia

Na hodnotenie kvality modelov boli použité metriky RMSE (viď podkapitola 2.2.1), koeficient determinácie  $R^2$  (viď podkapitola 2.2.2) a štandardná odchýlka STD (viď podkapitola 2.2.3). Výpočet uvedených metrik bol realizovaný v programovacom jazyku Python s využitím knižnice `scikit-learn`, konkrétne funkcií `mean_squared_error` a `r2_score` z modulu `sklearn.metrics`, štandardná odchýlka bola vypočítaná pomocou funkcie `np.std` z knižnice `NumPy`.

Ako hlavná metrika hodnotenia modelov bola zvolená RMSE krížovej validácie (ďalej len CV RMSE) namiesto testovacej RMSE. Keďže CV RMSE je založená na 125 nezávislých hodnoteniach, čo využíva všetky dostupné dáta a poskytuje štatisticky spoľahlivejší odhad generalizačnej schopnosti modelu.

Testovacia RMSE slúži ako overenie toho, do akej miery sa výkonnosť modelu na testovacej množine zhoduje s odhadom z krížovej validácie a ako by sa model správal pri reálnom nasadení. CV RMSE však zostáva štatisticky spoľahlivejším odhadom, keďže vychádza z väčšieho počtu hodnotení.

### 3.2.6 Implementácia modelov

Lineárna regresia bola implementovaná pomocou triedy `LinearRegression` z knižnice `scikit-learn`. Neurónová sieť bola implementovaná ako dopredná sieť (*feedforward neural network*) pomocou rozhrania `nn.Sequential` z knižnice `PyTorch`. Ako aktivačná funkcia bola použitá funkcia `ReLU` (2.1.6). Model bol trénovaný pomocou optimalizátora `Adam` (2.1.7) s hodnotou rýchlosti učenia  $\eta = 0,001$  a stratovou funkciou `MSELoss` (2.1.5).

Vzhľadom na malý rozsah datasetu a relatívne vysoký počet parametrov neurónových sietí existuje zvýšené riziko pretrénovania. Z tohto dôvodu boli aplikované dve metódy regularizácie: `Dropout` (2.1.8) s pravdepodobnosťou 10 % a predčasné zastavenie trénovania (*early stopping*).

### 3.2.7 Algoritmus výberu príznakov

V každej etape bolo hodnotené maximálne 5 vstupných príznakov (viď podkapitola 3.3.1). Hodnotenie prebieha pre všetky možné kombinácie zvolených príznakov pomocou funkcie `combinations` z knižnice `itertools` v jazyku Python. Výber finálnej množiny príznakov prebiehal iteratívnym algoritmom. Z dostupných kandidátov sa vyberie príznak s najvyššou absolútnou koreláciou s  $x_B$  (viď podkapitola 2.3.6). Príznak sa zaradi do výslednej množiny, ak jeho vzájomná korelácia so všetkými už vybranými príznakmi neprekračuje  $|r| = 0,80$  (viď podkapitola 2.3.7). Postup sa opakuje, kým nie je vybraných 5 príznakov alebo kým nie sú vyčerpaní kandidáti. Tento prístup zabezpečuje, že vybraná množina príznakov maximalizuje koreláciu s cieľovou premennou pri súčasnom minimalizovaní redundancie medzi príznakmi.

### 3.2.8 Vyčerpávajúce prehľadávanie priestoru príznakov

Pre overenie účinnosti algoritmu výberu príznakov bolo implementované globálne vyčerpávajúce prehľadávanie priestoru príznakov. Celkový počet kombinácií pre množinu 15 kandidátnych príznakov s maximálne 5 príznakmi je:

$$\sum_{n=1}^5 \binom{15}{n} = 15 + 105 + 455 + 1365 + 3003 = 4943. \quad (3.3)$$

Z dôvodu výpočtovej náročnosti ( $4943 \times 2$  modelov) bol počet opakovaní krížovej validácie znížený na  $n_{\text{repeats}} = 2$  (50 hodnotení na kombináciu namiesto 125 hodnotení v etapách 1–4). Výsledky sú preto štatisticky menej stabilné a slúžia predovšetkým na identifikáciu sľubných kombinácií, nie na presné porovnanie s inými etapami.

**Tabuľka 3.2:** Porovnanie architektúr neurónových sietí

| Architektúra          | Par. | CV<br>RMSE | CV<br>RMSE std | CV<br>R <sup>2</sup> |
|-----------------------|------|------------|----------------|----------------------|
| 1 vrstva, 4 neuróny   | 29   | 0,1339     | 0,0072         | 0,5995               |
| 1 vrstva, 6 neurónov  | 43   | 0,1271     | 0,0055         | 0,6381               |
| 1 vrstva, 8 neurónov  | 57   | 0,1243     | 0,0068         | 0,6553               |
| 1 vrstva, 12 neurónov | 85   | 0,1224     | 0,0052         | 0,6693               |
| 1 vrstva, 16 neurónov | 113  | 0,1214     | 0,0056         | 0,6747               |
| 2 vrstvy, 4×4         | 57   | 0,1440     | 0,0054         | 0,5332               |
| 2 vrstvy, 8×4         | 101  | 0,1317     | 0,0048         | 0,6108               |

### 3.3 Výber architektúry neurónovej siete

Pred spustením hlavného experimentu bola vykonaná analýza architektúr neurónovej siete s cieľom zvoliť konfiguráciu vhodnú pre daný dataset.

#### 3.3.1 Obmedzenie počtu parametrov

Pri návrhu architektúry neurónovej siete je potrebné zohľadniť veľkosť dostupného datasetu. Ako bolo uvedené v podkapitole 2.3.8, podľa empirického Widrowovho pravidla by mal byť počet trénovacích dát rádovo väčší než počet parametrov modelu (často približne 10-násobok).

Napríklad podľa vzťahu uvedeného v podkapitole 2.3.8 má sieť s jednou skrytou vrstvou,  $p = 5$  príznakmi a  $h = 6$  neurónmi celkovo  $P = 43$  parametrov. Z tohto dôvodu bol počet vstupných príznakov obmedzený na  $p = 5$ , aby sa minimalizoval počet parametrov modelu a tým aj jeho zložitosť.

#### 3.3.2 Porovnanie architektúr

Na výber architektúry neurónovej siete bolo porovnaných päť konfigurácií jednovrstvových sietí s rôznym počtom neurónov (4, 6, 8, 12, 16) a dve dvojevrstvé konfigurácie (4×4 a 8×4 neurónov). Ako vstupné príznaky bola použitá päťica príznakov vybraných v etape 1 (viď sekcia 3.4). Prehľad počtu parametrov a dosiahnutých metrík je uvedený v tabuľke 3.2. Z tabuľky vyplýva, že CV RMSE jednovrstvových sietí monotónne klesá so zvyšujúcim sa počtom neurónov od 0,1339 (4 neuróny) po 0,1214 (16 neurónov), avšak pokles sa postupne zmenšuje. Rozdiel medzi architektúrami so 6 a 8 neurónmi predstavuje len 0,0028, čo zodpovedá menej ako 2,3% zlepšeniu pri náraste počtu

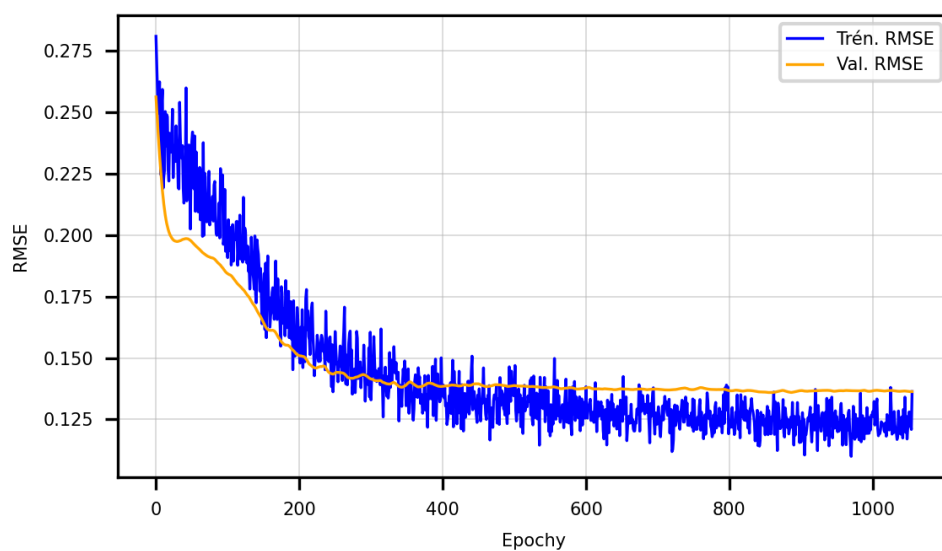
parametrov z 43 na 57. Dvojvrstvové architektúry dosahujú horšie výsledky ako jednovrstvové s porovnateľným počtom parametrov, čo naznačuje, že pre daný dataset nie je hlbšia architektúra výhodná. Pozoruhodným prípadom je architektúra  $4 \times 4$  (57 parametrov), ktorá napriek rovnakému počtu parametrov ako jednovrstvová sieť s 8 neurónmi dosahuje výrazne horšie CV RMSE. Hlbšia sieť je napriek identickému počtu parametrov náročnejšia na optimalizáciu: gradient pri spätnom šírení chýb prechádza väčším počtom vrstiev. Nízka hodnota CV STD = 0,0054 naznačuje, že optimalizátor konzistentne konvergoval do rovnakého suboptimálneho riešenia naprieč všetkými rozdeleniami dát.

Na základe uvedených výsledkov boli ako kandidáti pre ďalšiu analýzu vybrané architektúry so 6 a 8 neurónmi, ktoré dosahovali porovnateľné výsledky. Rozdiel CV RMSE (0,1271 pre 6 neurónov oproti 0,1243 pre 8 neurónov) predstavuje 0,0028, čo naznačuje praktickú ekvivalenciu oboch architektúr.

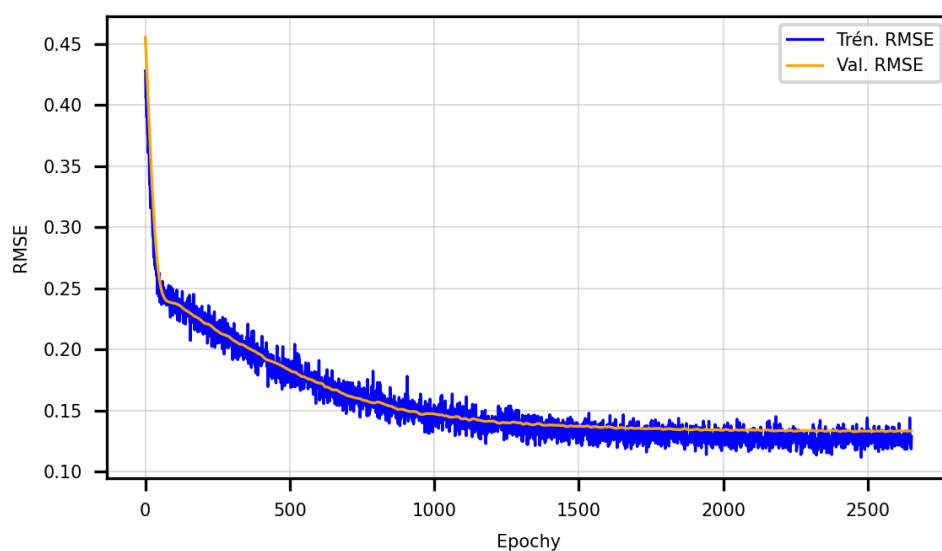
### 3.3.3 Krivky učenia

Pre porovnanie boli vyhodnotené krivky učenia (viď podkapitola 2.1.9) pre architektúry s 8 aj 6 neurónmi. Grafy sú uvedené pre hodnotu generátora rozdelení 34; ostatné hodnoty generátora rozdelení vykazujú podobné správanie. Trénovacie krivky vykazujú výrazný šum spôsobený efektom dropout vrstvy, ktorá počas každej trénovacej epochy náhodne deaktivuje časť neurónov čo spôsobuje kolísanie hodnoty RMSE na trénovacej množine. Validačná krivka je hladšia, keďže počas validácie je dropout neaktívny. Krivky pre architektúru s 8 neurónmi sú znázornené na obrázku 3.1. Na obrázku 3.2 sú krivky učenia pre architektúru so 6 neurónmi. Oba grafy zobrazujú učenie bez výrazného pretrénovania: trénovacia RMSE aj validačná RMSE spolu klesajú a následne sa stabilizujú. Medzi modelmi je však viditeľný jeden rozdiel. Pri architektúre so 6 neurónmi (2600 epoch) sa krivky počas celého tréningu takmer prekrývajú. Pri architektúre s 8 neurónmi (1050 epoch) je validačná RMSE na začiatku tréningu výrazne vyššia než trénovacia RMSE a približne do 200. epochy zostáva nad ním, pričom až následne sa krivky zbiehajú. To môže naznačovať, že model s väčšou kapacitou je v počiatočných fázach učenia menej stabilný.

Výsledné hodnoty RMSE sú pri oboch architektúrach veľmi podobné, čo znamená, že dodatočné neuróny neprinášajú významné zlepšenie presnosti. Model so 6 neurónmi má zároveň menší počet parametrov, čím sa znižuje riziko pretrénovania na nových dátach. Na základe týchto argumentov je architektúra so 6 neurónmi vhodnejšou voľbou.



**Obr. 3.1:** Krivky učenia pre architektúru s 8 neurónmi (hodnota generátora rozdelení 34).



**Obr. 3.2:** Krivky učenia pre architektúru so 6 neurónmi (hodnota generátora rozdelení 34).

### 3.3.4 Absencia výrazného pretrénovania

Napriek tomu, že nominálny počet parametrov väčšiny testovaných architektúr výrazne prekračuje limit odporúčaný Widrowovým pravidlom, žiadna z architektúr nevykazuje výrazné príznaky pretrénovania. CV RMSE monotónne klesá a krivky učenia neodhaľujú výrazný rozchod tréningovej a validačnej RMSE. Tento výsledok je v súlade s tým, že predčasné zastavenie tréningu a dropout spoločne obmedzujú model. Mierny rozchod kriviek pozorovaný pri architektúre s 8 neurónmi (hodnota generátora rozdelení 34) potvrdzuje, že pri malom datasete je riziko pretrénovania vysoké, avšak použitá regularizácia ho udržala na zanedbateľnej úrovni.

### 3.3.5 Voľba architektúry

Na základe výsledkov uvedených v tabuľke 3.2 a analýzy kriviek učenia bola zvolená architektúra s 1 skrytou vrstvou a 6 neurónmi s aktivačnou funkciou ReLU, Dropout (pravdepodobnosť 10 %) a predčasným zastavením tréningu. Hoci architektúra s 8 neurónmi dosiahla veľmi podobné hodnoty RMSE, architektúra so 6 neurónmi bola uprednostnená na základe princípu parsimónie (Occamova britva), keďže obsahuje menší počet parametrov (43 parametrov).

Zvolená architektúra bola fixovaná pre všetky štyri etapy experimentu. Akékoľvek rozdiely v metrikách medzi etapami možno teda pripísať výlučne zmene vstupných príznakov, nie zmene architektúry modelu.

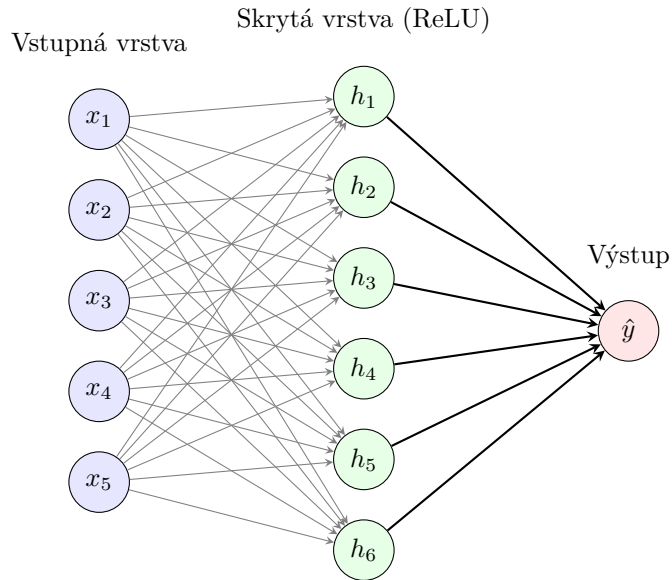
## 3.4 Etapa 1 — Základný výber príznakov

Cieľom prvej etapy je výber príznakov priamo z pôvodných 16 procesných premenných datasetu bez akýchkoľvek transformácií. Tento prístup predstavuje základnú úroveň, voči ktorej sú porovnávané výsledky nasledujúcich etáp.

### 3.4.1 Postup

Na základe algoritmu výberu príznakov (viď podkapitola 3.2.7) boli pre etapu 1 vybrané príznaky uvedené v tabuľke 3.3. Oba modely boli tréňované a hodnotené na vybraných 5 príznakoch v súlade s 3.2. Výsledky pre najlepšie kombinácie príznakov sú uvedené v tabuľke 3.4.

CV RMSE sa s rastúcim  $p$  zlepšuje s postupne klesajúcim prírastkom. Pre LR je najlepší výsledok dosiahnutý pri  $p = 5$ , pričom prechod z  $p = 4$  na  $p = 5$  prináša



**Obr. 3.3:** Architektúra navrhnutej neurónovej siete so šiestimi neurónmi v skrytej vrstve

**Tabuľka 3.3:** Vybrané príznaky etapy 1

| Poradie | Príznak      | Korelácia s $x_B$ |
|---------|--------------|-------------------|
| 1       | $T_{c,B}$    | 0,5672            |
| 2       | $T_B$        | -0,5600           |
| 3       | $T_{c,D}$    | 0,5370            |
| 4       | $T_{B,1,in}$ | 0,4920            |
| 5       | $F_{B,2}$    | -0,3930           |

**Tabuľka 3.4:** Najlepšie výsledky modelov v etape 1

| Model | Príznamy                                          | $p$ | CV<br>RMSE    | CV<br>RMSE std | CV<br>$R^2$   |
|-------|---------------------------------------------------|-----|---------------|----------------|---------------|
| LR    | $T_B$                                             | 1   | 0,1849        | 0,0082         | 0,2606        |
| LR    | $T_B, T_{c,B}$                                    | 2   | 0,1384        | 0,0056         | 0,5784        |
| LR    | $T_B, T_{c,B}, T_{c,D}$                           | 3   | 0,1266        | 0,0048         | 0,6450        |
| LR    | $T_B, T_{c,B}, T_{c,D}, F_{B,2}$                  | 4   | 0,1236        | 0,0050         | 0,6600        |
| LR    | $T_B, T_{c,B}, T_{c,D},$<br>$T_{B,1,in}, F_{B,2}$ | 5   | <b>0,1235</b> | 0,0052         | <b>0,6602</b> |
| NN    | $T_{c,B}$                                         | 1   | 0,1859        | 0,0056         | 0,2501        |
| NN    | $T_B, T_{c,B}$                                    | 2   | 0,1344        | 0,0068         | 0,6019        |
| NN    | $T_B, T_{c,B}, T_{c,D}$                           | 3   | 0,1280        | 0,0062         | 0,6385        |
| NN    | $T_B, T_{c,B}, T_{c,D}, T_{B,1,in}$               | 4   | <b>0,1267</b> | 0,0038         | <b>0,6441</b> |
| NN    | $T_B, T_{c,B}, T_{c,D},$<br>$T_{B,1,in}, F_{B,2}$ | 5   | 0,1273        | 0,0055         | 0,6374        |

Tučným písmom sú označené najlepšie výsledky pre každý model.

**Tabuľka 3.5:** Matica vzájomnej korelácie vybraných príznakov etapy 1

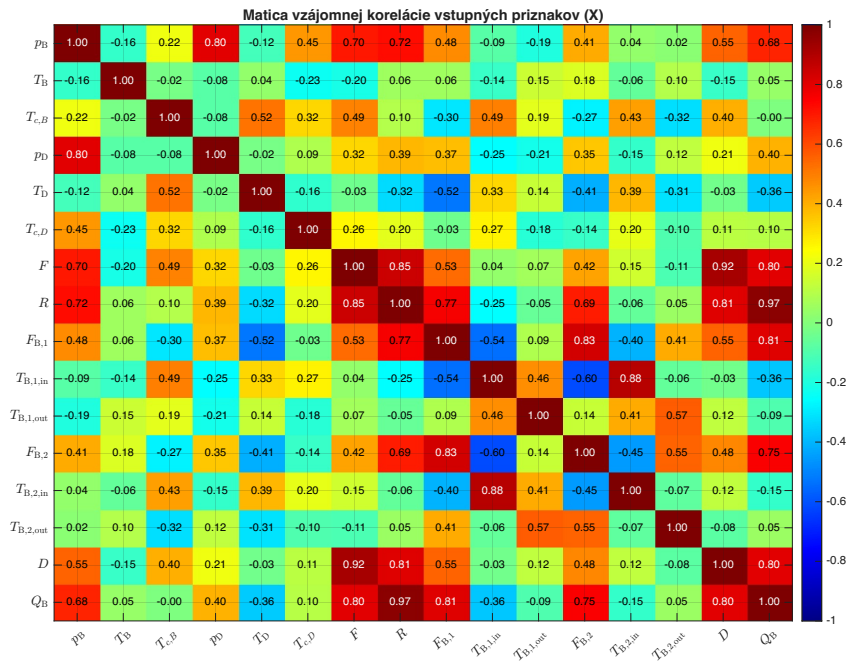
|              | $T_B$ | $T_{c,B}$ | $T_{c,D}$ | $T_{B,1,in}$ | $F_{B,2}$ |
|--------------|-------|-----------|-----------|--------------|-----------|
| $T_B$        | 1,00  | -0,02     | -0,23     | -0,14        | 0,18      |
| $T_{c,B}$    | -0,02 | 1,00      | 0,32      | 0,49         | -0,27     |
| $T_{c,D}$    | -0,23 | 0,32      | 1,00      | 0,27         | -0,14     |
| $T_{B,1,in}$ | -0,14 | 0,49      | 0,27      | 1,00         | -0,60     |
| $F_{B,2}$    | 0,18  | -0,27     | -0,14     | -0,60        | 1,00      |

zlepšenie len o 0,0001. Pre NN je najlepší výsledok pri  $p = 4$ ; pridanie piateho príznaku výkonnosť mierne zhoršilo ( $0,1267 \rightarrow 0,1273$ ).

Pre päť kandidátov s najvyššou absolútnou koreláciou bola vypočítaná matica vzájomných korelácií (obrázok 3.4). Konkrétne hodnoty vzájomných korelácií medzi vybranými kandidátmi sú uvedené v tabuľke 3.5.

### 3.4.2 Výsledky

Lineárna regresia s 5 príznakmi dosiahla CV RMSE = 0,1235 a  $R^2 = 0,660$ , čo predstavuje základnú úroveň výkonnosti pre ďalšie porovnanie. Neurónová sieť dosahuje porovnateľné výsledky (CV RMSE = 0,1267, 4 príznaky), pričom lineárna regresia



Obr. 3.4: Matica vzájomnej korelácie vstupných príznakov v etape 1.

**Tabuľka 3.6:** Test RMSE lineárnej regresie ( $p = 5$ ) v etape 1.

| Hodnota generátora rozdelení | Test RMSE |
|------------------------------|-----------|
| 34                           | 0,1122    |
| 42                           | 0,1373    |
| 7                            | 0,1022    |
| 99                           | 0,1356    |
| 123                          | 0,1492    |

mierne prevyšuje neurónovú sieť. S pribúdajúcim počtom príznakov výkonnosť oboch modelov rastie, avšak prírastok sa postupne znižuje — prechod zo 4 na 5 príznakov prináša zlepšenie CV RMSE len o 0,0001 pre LR, zatiaľ čo u neurónovej siete CV RMSE sa pri 5 príznakoch (0,1273) zväčšilo oproti 4 príznakom (0,1267). Tabuľka 3.6 uvádza Test RMSE najlepšieho modelu (LR,  $p = 5$ ).

Hodnoty Test RMSE najlepšieho modelu (LR,  $p = 5$ ) sa pohybujú v rozmedzí 0,1022 až 0,1492, čo odráža citlivosť na konkrétne rozdelenie dát. Táto variabilita je typickým dôsledkom malého datasetu. Priemerná hodnota Test RMSE ( $\approx 0,1273$ ) je blízka CV RMSE ( $= 0,1235$ ), čo potvrdzuje spoľahlivosť krížovej validácie ako odhadu generalizačnej schopnosti modelu.

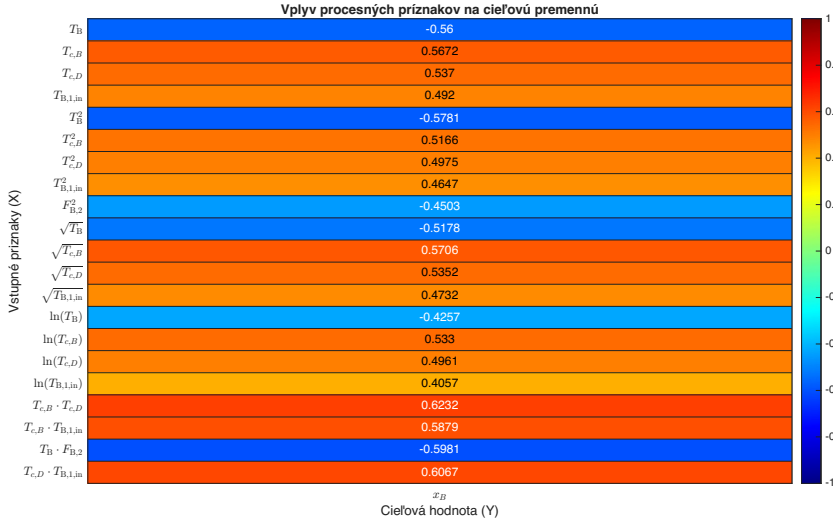
## 3.5 Etapa 2 — Nelineárne transformácie príznakov

Cieľom druhej etapy je rozšírenie priestoru príznakov pomocou nelineárnych matematických transformácií pôvodných premenných. Tento prístup umožňuje lineárnym modelom zachytiť nelineárne závislosti v dátach bez zmeny architektúry modelu.

### 3.5.1 Postup

Pre každý z 16 pôvodných príznakov boli vygenerované nasledujúce transformácie:

- kvadratická transformácia:  $x^2$
- odmocnina:  $\sqrt{x}$
- prirodzený logaritmus:  $\ln(x + \varepsilon)$
- prevrátená hodnota:  $x^{-1}$



**Obr. 3.5:** Korelácia transformovaných príznakov s cieľovou premennou  $x_B$  v etape 2 (zobrazené len príznaky s  $|r| > 0,4$ ).

**Tabuľka 3.7:** Vybrané príznaky etapy 2

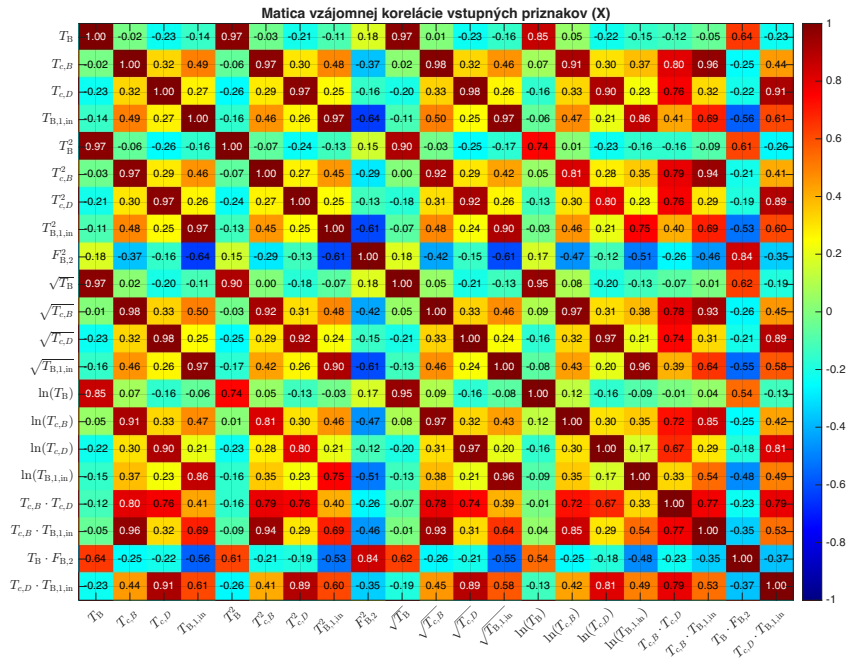
| Poradie | Príznak                    | Korelácia s $x_B$ |
|---------|----------------------------|-------------------|
| 1       | $T_{c,B} \cdot T_{c,D}$    | 0,6232            |
| 2       | $T_{c,D} \cdot T_{B,1,in}$ | 0,6067            |
| 3       | $T_B \cdot F_{B,2}$        | -0,5981           |
| 4       | $T_{c,B} \cdot T_{B,1,in}$ | 0,5879            |
| 5       | $T_B^2$                    | -0,5781           |

kde  $\varepsilon = 0,01$  predstavuje malú konštantu zabráňujúcu numerickej nestabilite pri hodnotách blízkyh nule. Okrem jednotlivých transformácií boli vygenerované aj párové súčiny piatich príznakov vybraných v etape 1, čo zodpovedá 10 kombináciám. Celkovo bolo vygenerovaných 90 kandidátov. Na základe algoritmu výberu príznakov (viď podkapitola 3.2.7) boli pre etapu 2 vybrané príznaky uvedené v tabuľke 3.7. Výsledky korelačnej analýzy sú znázornené na obrázku 3.5. Všetky vybrané príznaky sú párové súčiny alebo kvadratické transformácie teplotných a prietokových veličín.

Pre oba modely CV RMSE klesá monotónne s rastúcim počtom príznakov; prírastok sa postupne znižuje. Pre LR je prechod z  $p = 4$  na  $p = 5$  zanedbateľný ( $0,1299 \rightarrow 0,1298$ ). Pre päť vybraných kandidátov bola vypočítaná matica vzájomných korelácií znázornená na obrázku 3.6. Konkrétne hodnoty vzájomných korelácií medzi vybranými príznakmi

Tabuľka 3.8: Najlepšie výsledky modelov v etape 2

| Model | Príznaky                                                                                                   | $p$ | CV<br>RMSE    | CV<br>RMSE std | CV<br>$R^2$   |
|-------|------------------------------------------------------------------------------------------------------------|-----|---------------|----------------|---------------|
| LR    | $T_{c,B} \cdot T_{c,D}$                                                                                    | 1   | 0,1758        | 0,0052         | 0,3244        |
| LR    | $T_B^2, T_{c,B} \cdot T_{c,D}$                                                                             | 2   | 0,1391        | 0,0045         | 0,5757        |
| LR    | $T_B^2, T_{c,B} \cdot T_{c,D},$<br>$T_B \cdot F_{B,2}$                                                     | 3   | 0,1312        | 0,0043         | 0,6196        |
| LR    | $T_B^2, T_{c,B} \cdot T_{c,D},$<br>$T_{c,B} \cdot T_{B,1,in}, T_B \cdot F_{B,2}$                           | 4   | 0,1299        | 0,0047         | 0,6264        |
| LR    | $T_B^2, T_{c,B} \cdot T_{c,D}, T_{c,B} \cdot T_{B,1,in},$<br>$T_B \cdot F_{B,2}, T_{c,D} \cdot T_{B,1,in}$ | 5   | <b>0,1298</b> | 0,0047         | <b>0,6264</b> |
| NN    | $T_{c,B} \cdot T_{c,D}$                                                                                    | 1   | 0,1743        | 0,0054         | 0,3349        |
| NN    | $T_B^2, T_{c,B} \cdot T_{c,D}$                                                                             | 2   | 0,1307        | 0,0047         | 0,6234        |
| NN    | $T_B^2, T_{c,B} \cdot T_{c,D},$<br>$T_B \cdot F_{B,2}$                                                     | 3   | 0,1288        | 0,0041         | 0,6370        |
| NN    | $T_B^2, T_{c,B} \cdot T_{c,D},$<br>$T_{c,B} \cdot T_{B,1,in}, T_B \cdot F_{B,2}$                           | 4   | 0,1278        | 0,0056         | 0,6372        |
| NN    | $T_B^2, T_{c,B} \cdot T_{c,D}, T_{c,B} \cdot T_{B,1,in},$<br>$T_B \cdot F_{B,2}, T_{c,D} \cdot T_{B,1,in}$ | 5   | <b>0,1273</b> | 0,0047         | <b>0,6412</b> |



Obr. 3.6: Matica vzájomnej korelácie vybraných príznakov etapy 2.

**Tabuľka 3.9:** Matica vzájomnej korelácie vybraných príznakov etapy 2

|                            | $T_{c,B} \cdot T_{c,D}$ | $T_{c,D} \cdot T_{B,1,in}$ | $T_B \cdot F_{B,2}$ | $T_{c,B} \cdot T_{B,1,in}$ | $T_B^2$ |
|----------------------------|-------------------------|----------------------------|---------------------|----------------------------|---------|
| $T_{c,B} \cdot T_{c,D}$    | 1,000                   | 0,786                      | -0,227              | 0,772                      | -0,161  |
| $T_{c,D} \cdot T_{B,1,in}$ | 0,786                   | 1,000                      | -0,373              | 0,525                      | -0,261  |
| $T_B \cdot F_{B,2}$        | -0,227                  | -0,373                     | 1,000               | -0,349                     | 0,615   |
| $T_{c,B} \cdot T_{B,1,in}$ | 0,772                   | 0,525                      | -0,349              | 1,000                      | -0,088  |
| $T_B^2$                    | -0,161                  | -0,261                     | 0,615               | -0,088                     | 1,000   |

**Tabuľka 3.10:** Test RMSE neurónovej siete ( $p = 5$ ) v etape 2.

| Hodnota generátora rozdelení | Test RMSE |
|------------------------------|-----------|
| 34                           | 0,1183    |
| 42                           | 0,1362    |
| 7                            | 0,1107    |
| 99                           | 0,1641    |
| 123                          | 0,1476    |

sú uvedené v tabuľke 3.9.

### 3.5.2 Výsledky

Najlepší výsledok etapy 2 dosiahla neurónová sieť s 5 príznakmi, pričom CV RMSE = 0,1273, čo je mierne horšie ako najlepší výsledok neurónovej siete v etape 1 (CV RMSE = 0,1267 pri 4 príznakoch). V prípade lineárnej regresie dosiahol najlepší model CV RMSE = 0,1298, čo je taktiež horšie ako najlepší výsledok etapy 1 (CV RMSE = 0,1235). Nelineárne transformácie teda nepriniesli zlepšenie oproti pôvodným príznakom ani pre jeden z modelov. Napriek tomu zostáva zaujímavým pozorovaním, že najlepší jednotlivý príznak etapy 2 ( $T_{c,B} \cdot T_{c,D}$ ,  $|r| = 0,623$ ) prekonáva najlepší jednotlivý príznak etapy 1 ( $T_{c,B}$ ,  $|r| = 0,567$ ). Nelineárne interakcie medzi procesnými premennými teda existujú, avšak ich využitie v rámci zvoleného postupu neprinieslo merateľné zlepšenie finálnych modelov. Tabuľka 3.10 uvádza Test RMSE najlepšieho modelu (NN,  $p = 5$ ).

Aj na testovacích dátach najlepšieho modelu (NN,  $p = 5$ ) je viditeľné zhoršenie oproti etape 1: priemerná hodnota Test RMSE ( $\approx 0,1354$ ) je vyššia ako v etape 1 ( $\approx 0,1273$ ). Výsledky teda konzistentne potvrdzujú záver z krížovej validácie.

## 3.6 Etapa 3 — Fyzikálne odôvodnené príznaky

Cieľom tretej etapy je konštrukcia príznakov vychádzajúcich z fyzikálnych zákonitostí destilačného procesu. Na rozdiel od etapy 2, kde boli príznaky generované čisto matematicky, v tejto etape je výber motivovaný termodynamikou procesu.

### 3.6.1 Postup

Pri návrhu fyzikálne odôvodnených príznakov sa vychádzalo z práce Mojto a kol. (2021), ktorá pracuje s rovnakou destilačnou kolónou a vychádza z rovnakých typov meraní, avšak použitý dataset sa mierne líši v počte pozorovaní a množstve dostupných príznakov. Navrhované príznaky sú odvodené z termodynamických vlastností kolóny. V porovnaní s touto prácou však použitý dataset obsahuje iné merania, napríklad samostatné teploty na vstupe a výstupe výmenníkov tepla, čo umožňuje vytvoriť rozšírenú množinu príznakov.

Jednou z kľúčových myšlienok je použitie teploty kompenzovanej tlakom PCT (*Pressure Compensated Temperature*). Práca Mojto a kol. (2021) navrhuje jej výpočet pomocou Clausius-Clapeyronovej rovnice:

$$PCT = \frac{1}{\frac{R}{H_v} \ln \frac{P}{P_{\text{ref}}} + \frac{1}{T}} \quad (3.4)$$

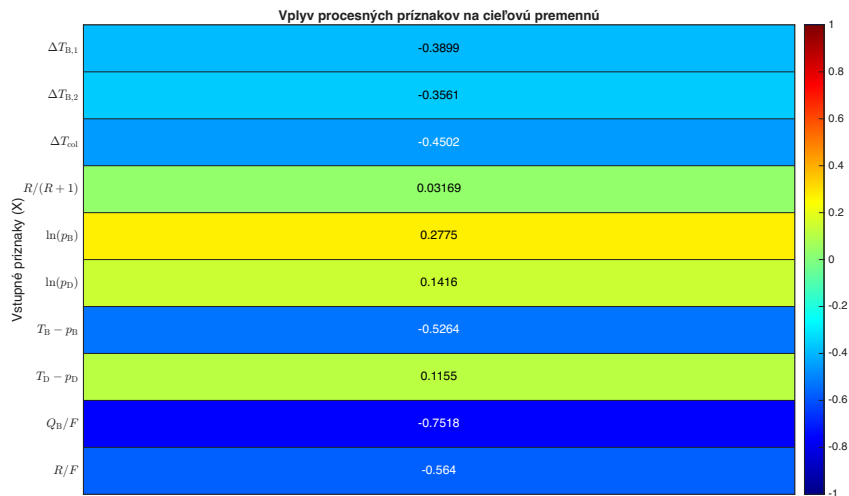
kde  $R$  je univerzálna plynová konštanta,  $H_v$  je výparné teplo,  $P_{\text{ref}}$  je referenčný tlak a  $T$  je teplota v Kelvinoch. Keďže konštanty  $R/H_v$  a  $P_{\text{ref}}$  nie sú pre daný dataset dostupné a teploty nie sú v Kelvinoch, nie je možné vypočítať PCT priamo. Namiesto toho boli komponenty menovateľa ( $\ln(p)$  a  $1/T$ ) použité ako samostatné príznaky; regresné modely dokážu nájsť príslušné váhy a aproximovať tak PCT bez znalosti konštánt.

Pri prehľadávaní literatúry pomocou nástroja Phiture ešte bola nájdená aj linearizovaná forma PCT King (2024):

$$PCT = T_{\text{measured}} - \frac{dT}{dP}(P_{\text{measured}} - P_{\text{ref}}) \quad (3.5)$$

kde  $dT/dP$  predstavuje lokálnu citlivosť teploty na tlak, ktorá je v rámci linearizácie považovaná za konštantnú a  $P_{\text{ref}}$  je referenčný tlak. Keďže tieto konštanty nie sú pre daný dataset dostupné, boli ako aproximácia PCT otestované príznaky tvaru  $T - p$  pre rôzne dvojice teplôt a tlakov v kolóne.

Okrem PCT boli konštruované ďalšie fyzikálne príznaky bežne používané pri analýze



**Obr. 3.7:** Korelácia fyzikálne odôvodnených príznakov s cieľovou premennou  $x_B$  v etape 3.

**Tabuľka 3.11:** Vybrané príznaky etapy 3

| Poradie | Príznak          | Korelácia s $x_B$ |
|---------|------------------|-------------------|
| 1       | $Q_B/F$          | -0,7518           |
| 2       | $R/F$            | -0,5640           |
| 3       | $T_B - p_B$      | -0,5264           |
| 4       | $\Delta T_{col}$ | -0,4502           |
| 5       | $\Delta T_{B,1}$ | -0,3899           |

destilačných kolón: pomer tepelného výkonu varáka k prietoku vstupnej suroviny  $Q_B/F$ , pomer refluxu k prietoku vstupnej suroviny  $R/F$ , teplotné rozdiely v tepelných výmenníkoch  $\Delta T_{B,1}$  a  $\Delta T_{B,2}$  a teplotný spád pozdĺž kolóny  $\Delta T_{col}$ .

Na základe algoritmu výberu príznakov (viď podkapitola 3.2.7) boli pre etapu 3 vybrané príznaky uvedené v tabuľke 3.11. Výsledky korelačnej analýzy sú znázornené na obrázku 3.7.

Pre LR CV RMSE klesá monotónne s rastúcim  $p$ . Pre NN dochádza pri  $p = 4$  k miernemu zhoršeniu oproti  $p = 3$  (0,1282 vs. 0,1273), čo môže súvisieť s vyššou vzájomnou koreláciou príznakov  $Q_B/F$  a  $R/F$ ; najlepší výsledok NN je dosiahnutý pri  $p = 5$ .

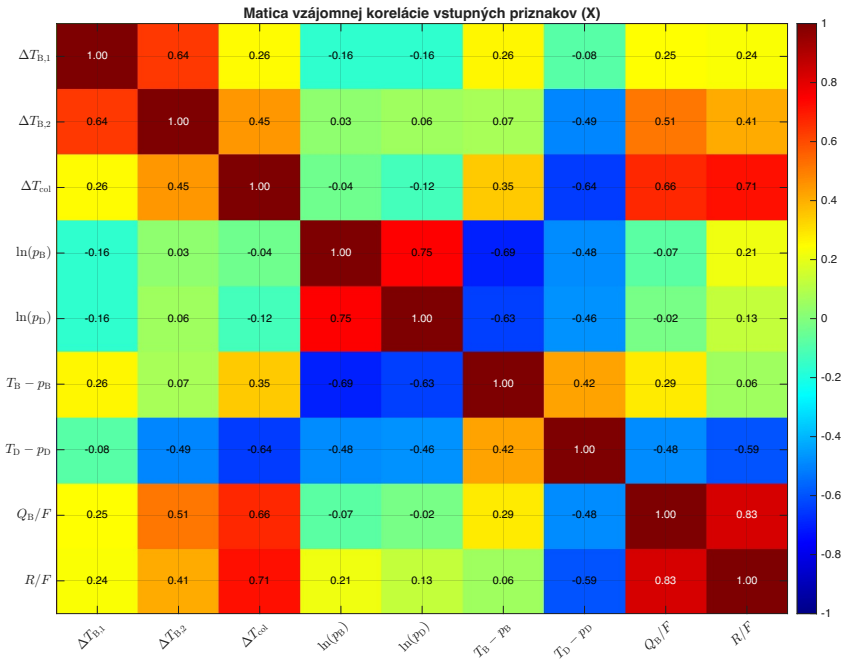
Tabuľka 3.12: Najlepšie výsledky modelov v etape 3

| Model | Príznaky                                                            | $p$ | CV<br>RMSE    | CV<br>RMSE std | CV<br>$R^2$   |
|-------|---------------------------------------------------------------------|-----|---------------|----------------|---------------|
| LR    | $Q_B/F$                                                             | 1   | 0,1485        | 0,0043         | 0,5171        |
| LR    | $T_B - p_B, Q_B/F$                                                  | 2   | 0,1310        | 0,0056         | 0,6207        |
| LR    | $\Delta T_{B,1}, T_B - p_B, Q_B/F$                                  | 3   | 0,1273        | 0,0060         | 0,6413        |
| LR    | $\Delta T_{B,1}, \Delta T_{\text{col}}, T_B - p_B, Q_B/F$           | 4   | 0,1235        | 0,0067         | 0,6617        |
| LR    | $\Delta T_{B,1}, \Delta T_{\text{col}},$<br>$T_B - p_B, Q_B/F, R/F$ | 5   | <b>0,1230</b> | 0,0070         | <b>0,6645</b> |
| NN    | $Q_B/F$                                                             | 1   | 0,1512        | 0,0035         | 0,4996        |
| NN    | $T_B - p_B, Q_B/F$                                                  | 2   | 0,1324        | 0,0071         | 0,6139        |
| NN    | $\Delta T_{B,1}, T_B - p_B, Q_B/F$                                  | 3   | 0,1273        | 0,0068         | 0,6445        |
| NN    | $\Delta T_{B,1}, \Delta T_{\text{col}}, T_B - p_B, Q_B/F$           | 4   | 0,1282        | 0,0091         | 0,6217        |
| NN    | $\Delta T_{B,1}, \Delta T_{\text{col}},$<br>$T_B - p_B, Q_B/F, R/F$ | 5   | <b>0,1262</b> | 0,0084         | <b>0,6463</b> |

Tabuľka 3.13: Matica vzájomnej korelácie vybraných príznakov etapy 3

|                         | $Q_B/F$ | $R/F$ | $T_B - p_B$ | $\Delta T_{\text{col}}$ | $\Delta T_{B,1}$ |
|-------------------------|---------|-------|-------------|-------------------------|------------------|
| $Q_B/F$                 | 1,00    | 0,83  | 0,29        | 0,66                    | 0,25             |
| $R/F$                   | 0,83    | 1,00  | 0,06        | 0,71                    | 0,24             |
| $T_B - p_B$             | 0,29    | 0,06  | 1,00        | 0,35                    | 0,26             |
| $\Delta T_{\text{col}}$ | 0,66    | 0,71  | 0,35        | 1,00                    | 0,26             |
| $\Delta T_{B,1}$        | 0,25    | 0,24  | 0,26        | 0,26                    | 1,00             |

Pre päť vybraných kandidátov bola vypočítaná matica vzájomných korelácií znázornená na obrázku 3.8. Konkrétne hodnoty sú uvedené v tabuľke 3.13. Maximálna vzájomná korelácia dosahuje 0,83 ( $Q_B/F$  a  $R/F$ ), čo mierne prekračuje stanovený prah  $|r| = 0,80$ . Napriek tomu oba príznaky môžu obsahovať potenciálne užitočné informácie o tepelných a hydraulických pomeroch procesu: príznak  $Q_B/F$  súvisí s tepelnou časťou procesu, zatiaľ čo  $R/F$  charakterizuje hydraulické pomery v kolóne. Zároveň  $R/F$  vykazuje relatívne vysokú koreláciu s cieľovou premennou a v rámci tejto etapy nemá priamy ekvivalent medzi ostatnými kandidátmi. Z tohto dôvodu boli oba príznaky ponechané a ich vplyv na kvalitu modelu bol následne overený experimentálne.



Obr. 3.8: Matica vzájomnej korelácie vybraných príznakov etapy 3.

**Tabuľka 3.14:** Test RMSE lineárnej regresie ( $p = 5$ ) v etape 3

| Hodnota generátora rozdelení | Test RMSE |
|------------------------------|-----------|
| 34                           | 0,0986    |
| 42                           | 0,1391    |
| 7                            | 0,0964    |
| 99                           | 0,1314    |
| 123                          | 0,1559    |

**Tabuľka 3.15:** Porovnanie najlepších výsledkov CV RMSE v etapách 1–3

| Etapa                         | LR      |     | NN      |     |
|-------------------------------|---------|-----|---------|-----|
|                               | CV RMSE | $p$ | CV RMSE | $p$ |
| 1 — Pôvodné príznaky          | 0,1235  | 5   | 0,1267  | 4   |
| 2 — Matematické transformácie | 0,1298  | 5   | 0,1273  | 5   |
| 3 — Fyzikálne príznaky        | 0,1230  | 5   | 0,1262  | 5   |

### 3.6.2 Výsledky

Najlepší výsledok etapy 3 dosiahla lineárna regresia s 5 príznakmi, ktorá dosiahla CV RMSE = 0,1230 a  $R^2 = 0,665$ . Tento výsledok predstavuje zlepšenie oproti etape 1 (CV RMSE = 0,1235) aj etape 2 (CV RMSE = 0,1298). V prípade neurónovej siete dosiahol najlepší model s 5 príznakmi CV RMSE = 0,1262, čo predstavuje mierne zlepšenie oproti najlepšiemu výsledku NN v etape 1 (CV RMSE = 0,1267 pri 4 príznakoch). Fyzikálne odôvodnené príznaky teda priniesli merateľné zlepšenie pre oba modely. Tabuľka 3.14 uvádza Test RMSE najlepšieho modelu (LR,  $p = 5$ ).

Napriek variabilite závislej od rozdelenia dát je pri väčšine hodnôt generátora rozdelení viditeľné zlepšenie najlepšieho modelu (LR,  $p = 5$ ) oproti etapám 1 a 2. Priemerná hodnota Test RMSE ( $\approx 0,1243$ ) je nižšia ako v etape 1 ( $\approx 0,1273$ ) a etape 2 ( $\approx 0,1354$ ), čo potvrdzuje, že fyzikálne odôvodnené príznaky priniesli skutočné zlepšenie.

## 3.7 Etapa 4 — Výber príznakov z predchádzajúcich etáp

Cieľom štvrtej etapy je vybrať päť príznakov z množiny 15 kandidátov získaných v etapách 1–3 (viď tabuľka 3.15) a porovnať ich výkonnosť s najlepšími výsledkami predchádzajúcich etáp.

**Tabuľka 3.16:** Súbor 15 kandidátnych príznakov pre etapu 4

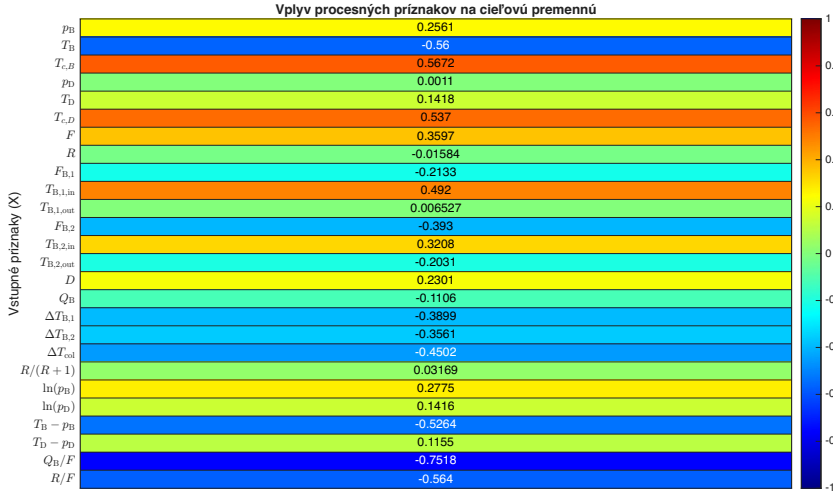
| #  | Príznak                    | Typ                                                  | Zdroj   |
|----|----------------------------|------------------------------------------------------|---------|
| 1  | $T_{c,B}$                  | pôvodný                                              | etapa 1 |
| 2  | $T_B$                      | pôvodný                                              | etapa 1 |
| 3  | $T_{c,D}$                  | pôvodný                                              | etapa 1 |
| 4  | $T_{B,1,in}$               | pôvodný                                              | etapa 1 |
| 5  | $F_{B,2}$                  | pôvodný                                              | etapa 1 |
| 6  | $T_{c,B} \cdot T_{c,D}$    | súčin teplôt                                         | etapa 2 |
| 7  | $T_{c,D} \cdot T_{B,1,in}$ | súčin teplôt                                         | etapa 2 |
| 8  | $T_B \cdot F_{B,2}$        | súčin temp./tok                                      | etapa 2 |
| 9  | $T_{c,B} \cdot T_{B,1,in}$ | súčin teplôt                                         | etapa 2 |
| 10 | $T_B^2$                    | kvadratický                                          | etapa 2 |
| 11 | $Q_B/F$                    | pomer tepelného výkonu<br>varáka a prietoku suroviny | etapa 3 |
| 12 | $R/F$                      | refluxný pomer                                       | etapa 3 |
| 13 | $T_B - p_B$                | lineariz. PCT                                        | etapa 3 |
| 14 | $\Delta T_{col}$           | teplotný spád                                        | etapa 3 |
| 15 | $\Delta T_{B,1}$           | tepl. rozdiel                                        | etapa 3 |

### 3.7.1 Postup

Celkový súbor kandidátov etapy 4 tvorí 15 príznakov: päť príznakov vybraných v etape 1 (pôvodné procesné premenné), päť príznakov vybraných v etape 2 (matematické transformácie) a päť fyzikálne odôvodnených príznakov etapy 3. Kompletný zoznam je uvedený v tabuľke 3.16.

Výber príznakov pre etapu 4 bol založený na analýze korelácií a postupnom výbere kandidátnych príznakov z etáp 1–3. Najskôr boli z výberu vylúčené súčinové príznaky etapy 2. Táto etapa nepreukázala zlepšenie oproti etape 1 (LR CV RMSE 0,1298 vs. 0,1235) a zároveň sa ukázalo, že súčiny zdieľajúce rovnaké premenné sú medzi sebou silne korelované ( $|r| = 0,93\text{--}0,97$ ), čo by viedlo k výraznej multikolinearite.

Následne boli zo zostávajúcich desiatich kandidátov etáp 1 a 3 odstránené fyzikálne nadbytočné príznaky. Pomer  $R/F$  bol vyradený, keďže zdieľa menovateľa  $F$  s  $Q_B/F$  a vykazuje s ním vysokú koreláciu ( $|r| = 0,826$ ), čo naznačuje štruktúrálnu závislosť. Tento príznak bol v etape 3 zaradený skôr z experimentálnych dôvodov, keďže množina dostupných kandidátov bola obmedzená. V etape 4 bol však vyradený v dôsledku multikolinearity a dostupnosti iných, potenciálne informatívnejších príznakov. Zároveň sa na základe výsledkov etapy 3 ukázalo, že jeho zaradenie neprináša výrazné zlepšenie

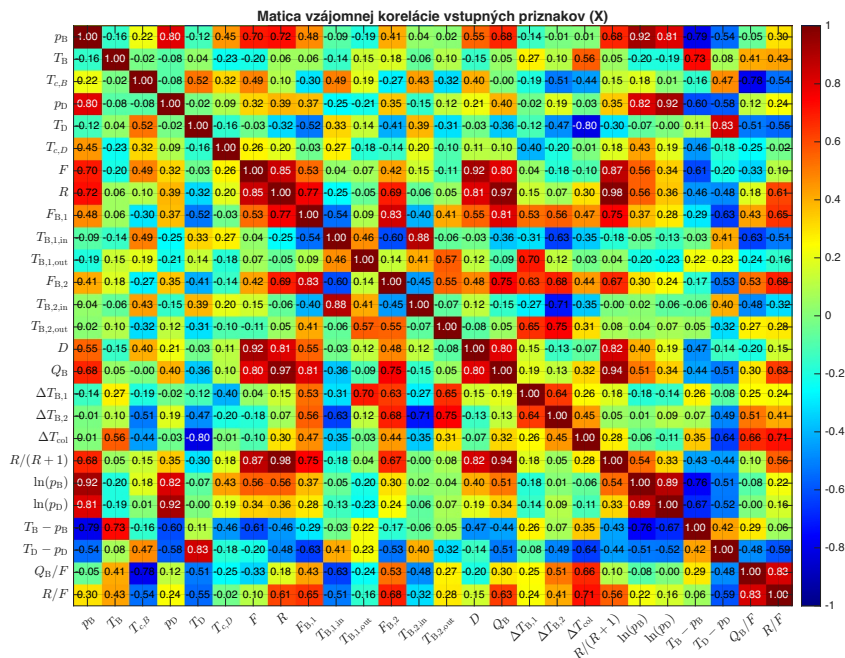
Obr. 3.9: Korelácia vybraných príznakov s cieľovou premennou  $x_B$  v etape 4.

Tabuľka 3.17: Matica vzájomnej korelácie vybraných príznakov etapy 4

|                  | $Q_B/F$ | $T_{c,B}$ | $T_B$ | $T_{c,D}$ | $\Delta T_{col}$ |
|------------------|---------|-----------|-------|-----------|------------------|
| $Q_B/F$          | 1,00    | -0,78     | 0,41  | -0,25     | 0,66             |
| $T_{c,B}$        | -0,78   | 1,00      | -0,02 | 0,32      | -0,44            |
| $T_B$            | 0,41    | -0,02     | 1,00  | -0,23     | 0,56             |
| $T_{c,D}$        | -0,25   | 0,32      | -0,23 | 1,00      | -0,01            |
| $\Delta T_{col}$ | 0,66    | -0,44     | 0,56  | -0,01     | 1,00             |

výkonnosti modelov. Príznak  $T_B - p_B$  predstavuje tlakovo kompenzovanú verziu  $T_B$  a neprináša nezávislú informáciu. Vylúčená bola aj vstupná teplota  $T_{B,1,in}$ , ktorá odráža podmienky výmenníka tepla a nie priamo stav separačného procesu.

Výber bol realizovaný zo siedmich zostávajúcich kandidátov ( $T_B$ ,  $T_{c,B}$ ,  $T_{c,D}$ ,  $F_{B,2}$ ,  $Q_B/F$ ,  $\Delta T_{col}$ ,  $\Delta T_{B,1}$ ) podľa 3.2.7 Obrázok 3.9 znázorňuje korelácie vybraných príznakov s cieľovou premennou  $x_B$ . Pre vybraných 5 príznakov bola vypočítaná matica vzájomných korelácií znázornená na obrázku 3.10. Konkrétne hodnoty sú uvedené v tabuľke 3.17. Výsledkom výberu (popísaného v predchádzajúcej podkapitole) je päťica príznakov výhradne z etáp 1 a 3, uvedená v tabuľke 3.18. Pre oba modely CV RMSE klesá monotónne s rastúcim  $p$ ; prírastok sa postupne znižuje. Prechod z  $p = 4$  na  $p = 5$  prináša pre LR zlepšenie len o 0,0001 (0,1128  $\rightarrow$  0,1127).



Obr. 3.10: Matica vzájomnej korelácie vybraných príznakov etapy 4.

Tabuľka 3.18: Vybrané príznaky etapy 4

| Príznak          | Korelácia s $x_B$ | Zdroj   |
|------------------|-------------------|---------|
| $Q_B/F$          | -0,7518           | etapa 3 |
| $T_{c,B}$        | +0,5672           | etapa 1 |
| $T_B$            | -0,5600           | etapa 1 |
| $T_{c,D}$        | +0,5370           | etapa 1 |
| $\Delta T_{col}$ | -0,4502           | etapa 3 |

Tabuľka 3.19: Najlepšie výsledky v etape 4

| Model | Príznamy                                       | $p$ | CV<br>RMSE    | CV<br>RMSE std | CV<br>$R^2$ |
|-------|------------------------------------------------|-----|---------------|----------------|-------------|
| LR    | $Q_B/F$                                        | 1   | 0,1485        | 0,0043         | 0,5171      |
| LR    | $Q_B/F, T_{c,D}$                               | 2   | 0,1259        | 0,0041         | 0,6520      |
| LR    | $Q_B/F, T_B, T_{c,D}$                          | 3   | 0,1149        | 0,0047         | 0,7076      |
| LR    | $Q_B/F, T_B, T_{c,D}, \Delta T_{col}$          | 4   | 0,1128        | 0,0054         | 0,7185      |
| LR    | $Q_B/F, T_{c,B}, T_B, T_{c,D}, \Delta T_{col}$ | 5   | <b>0,1127</b> | 0,0055         | 0,7185      |
| NN    | $Q_B/F$                                        | 1   | 0,1512        | 0,0035         | 0,4996      |
| NN    | $Q_B/F, T_{c,D}$                               | 2   | 0,1254        | 0,0049         | 0,6518      |
| NN    | $Q_B/F, T_B, T_{c,D}$                          | 3   | 0,1171        | 0,0034         | 0,6976      |
| NN    | $Q_B/F, T_{c,B}, T_B, T_{c,D}$                 | 4   | 0,1169        | 0,0051         | 0,6986      |
| NN    | $Q_B/F, T_{c,B}, T_B, T_{c,D}, \Delta T_{col}$ | 5   | <b>0,1164</b> | 0,0068         | 0,6936      |

Tabuľka 3.20: Test RMSE lineárnej regresie ( $p = 5$ ) v etape 4.

| Hodnota generátora rozdelení | Test RMSE |
|------------------------------|-----------|
| 34                           | 0,1020    |
| 42                           | 0,1202    |
| 7                            | 0,0902    |
| 99                           | 0,1181    |
| 123                          | 0,1470    |

### 3.7.2 Výsledky

Najlepší výsledok etapy 4 dosiahla lineárna regresia s 5 príznakmi, ktorá dosiahla CV RMSE = 0,1127 a  $R^2 = 0,7185$ . Tento výsledok predstavuje výrazné zlepšenie oproti etape 1 (CV RMSE = 0,1235), etape 2 (CV RMSE = 0,1298) aj etape 3 (CV RMSE = 0,1230). V prípade neurónovej siete dosiahol najlepší model s 5 príznakmi CV RMSE = 0,1164, čo predstavuje zlepšenie oproti najlepšiemu výsledku NN v etape 1 (CV RMSE = 0,1267 pri 4 príznakoch) aj etape 3 (CV RMSE = 0,1262). Výsledky naznačujú, že kombinácia príznakov z predchádzajúcich etáp umožnila vytvoriť informatívnejšiu množinu vstupov, ktorá viedla k výraznému zlepšeniu presnosti lineárnej regresie aj neurónovej siete. Tabuľka 3.20 uvádza Test RMSE najlepšieho modelu (LR,  $p = 5$ ).

V porovnaní s etapou 3 je na testovacích dátach najlepšieho modelu (LR,  $p = 5$ ) opäť viditeľné zlepšenie: priemerná hodnota Test RMSE ( $\approx 0,1155$ ) je výrazne nižšia ako v

**Tabuľka 3.21:** Najlepšie výsledky vyčerpávajúceho prehľadávania (etapa 5)

| Model | Príznaky                                                | $p$ | CV RMSE       |
|-------|---------------------------------------------------------|-----|---------------|
| LR    | $T_B, T_{c,D}, Q_B/F, \Delta T_{col}, \Delta T_{B,1}$   | 5   | <b>0,1121</b> |
| LR    | $T_{c,D}, T_B^2, Q_B/F, \Delta T_{col}, \Delta T_{B,1}$ | 5   | 0,1125        |
| LR    | $T_B, T_{c,B}, T_{c,D}, Q_B/F, \Delta T_{col}$          | 5   | 0,1126        |
| LR    | $T_B, T_{c,D}, Q_B/F, \Delta T_{col}$                   | 4   | 0,1127        |
| LR    | $T_B, T_{c,D}, Q_B/F$                                   | 3   | 0,1147        |
| NN    | $T_{c,D}, T_B^2, Q_B/F, \Delta T_{col}$                 | 4   | <b>0,1140</b> |
| NN    | $T_{c,D}, T_B^2, Q_B/F$                                 | 3   | 0,1148        |
| NN    | $T_B, T_{c,D}, Q_B/F$                                   | 3   | 0,1179        |

etape 3 ( $\approx 0,1243$ ). Systematický výber príznakov teda viedol k reálnemu zlepšeniu generalizačnej schopnosti modelu.

## 3.8 Etapa 5 — Vyčerpávajúce prehľadávanie priestoru príznakov

### 3.8.1 Postup

Cieľom piatej etapy je overiť, či výber príznakov v etape 4 identifikoval najlepšie riešenie v rámci prehľadávaného priestoru príznakov, alebo len lokálne optimum. Za týmto účelom boli testované všetky možné kombinácie zo súboru kandidátov (3.2.8).

### 3.8.2 Výsledky

Najlepšie kombinácie z vyčerpávajúceho prehľadávania sú zhrnuté v tabuľke 3.21.

Vyčerpávajúce prehľadávanie odhalilo, že najlepšia nájdená konfigurácia pre LR je  $\{T_B, T_{c,D}, Q_B/F, \Delta T_{col}, \Delta T_{B,1}\}$  s CV RMSE = 0,1121 ( $p = 5$ ).

Pre neurónovú sieť bola najlepšia nájdená konfigurácia  $\{T_{c,D}, T_B^2, Q_B/F, \Delta T_{col}\}$  s CV RMSE = 0,1140 ( $p = 4$ ).

Tabuľka 3.22 uvádza Test RMSE najlepšieho modelu (LR,  $p = 5$ ).

**Tabuľka 3.22:** Test RMSE lineárnej regresie ( $p = 5$ ) v etape 5.

| Hodnota generátora rozdelení | Test RMSE |
|------------------------------|-----------|
| 34                           | 0,0972    |
| 42                           | 0,1289    |
| 7                            | 0,0903    |
| 99                           | 0,1154    |
| 123                          | 0,1438    |

**Tabuľka 3.23:** Celkové porovnanie najlepších výsledkov CV RMSE naprieč etapami

| Etapa                             | LR            |     | NN            |     |
|-----------------------------------|---------------|-----|---------------|-----|
|                                   | CV RMSE       | $p$ | CV RMSE       | $p$ |
| 1 — Pôvodné príznaky              | 0,1235        | 5   | 0,1267        | 4   |
| 2 — Matematické transformácie     | 0,1298        | 5   | 0,1273        | 5   |
| 3 — Fyzikálne príznaky            | 0,1230        | 5   | 0,1262        | 5   |
| 4 — Výber z predchádzajúcich etáp | 0,1127        | 5   | 0,1164        | 5   |
| 5 — Vyčerpávajúce prehľadávanie*  | <b>0,1121</b> | 5   | <b>0,1140</b> | 4   |

\* Etapa 5 použila  $n_{\text{repeats}} = 2$ ; etapy 1–4 použili  $n_{\text{repeats}} = 5$ . Priame číselné porovnanie je preto orientačné.

Hodnoty Test RMSE najlepšieho modelu (LR,  $p = 5$ ) sú porovnateľné s etapou 4: priemerná hodnota ( $\approx 0,1151$ ) sa líši len o 0,0004 oproti etape 4 ( $\approx 0,1155$ ). Väčšina hodnôt generátora rozdelení vykazuje len minimálne rozdiely, čo naznačuje, že vyčerpávajúce prehľadávanie neprinieslo výrazné zlepšenie oproti iteratívnemu výberu príznakov z etapy 4.

### 3.9 Celkové porovnanie etáp 1–5

Tabuľka 3.23 zahŕňa najlepšie dosiahnuté výsledky CV RMSE naprieč všetkými etapami experimentu. Etapa 4 dosiahla CV RMSE = 0,1127 pre LR ( $p = 5$ ), čo predstavuje zlepšenie oproti etapám 1–3. Etapa 5 odhalila najlepšiu nájdenú LR konfiguráciu s CV RMSE = 0,1121 ( $p = 5$ ), pričom výber z etapy 4 skončil na 3. mieste s rovnakou veľkosťou kombinácie a rozdielom len 0,0005 (0,4 %).



Táto kapitola sa zameriava na interpretáciu výsledkov experimentu. Porovnávajú sa jednotlivé etapy z hľadiska výberu príznakov, analyzuje sa správanie lineárnej regresie a neurónovej siete a navrhuje sa finálna konfigurácia softvérového senzora vhodná pre priemyselné nasadenie.

## 4.1 Porovnanie etáp a výberu príznakov

Tabuľka 3.23 v predchádzajúcej kapitole zhrňa najlepšie CV RMSE dosiahnuté v jednotlivých etapách. Etapa 1 slúži ako základný referenčný výsledok, voči ktorému sa hodnotia výsledky nasledujúcich etáp.

Etapa 2 (matematické transformácie) nepriniesla zlepšenie oproti etape 1 a pre LR výsledky dokonca zhoršila. Hlavným dôvodom je vysoká multikolinearita medzi súčinnmi procesných premenných a redundancia informácie. Napriek vysokej individuálnej korelácii s  $x_B$  tieto príznaky neposkytujú modelu nezávislé informácie.

Etapa 3 priniesla mierne zlepšenie vďaka fyzikálne odôvodneným príznakom. Výrazne dominuje príznak  $Q_B/F$  s koreláciou  $|r| = 0,752$ , čo je najvyššia absolútna korelácia spomedzi všetkých príznakov vo všetkých etapách. Samotný tento príznak pri lineárnej regresii dosahuje CV RMSE = 0,1485, čo je lepší výsledok ako pri ktoromkoľvek jednotlivom príznaku z etáp 1 a 2. Fyzikálne to zodpovedá skutočnosti, že tepelný výkon privádzaný do varáka na jednotku vstupného toku priamo určuje energetickú náročnosť separácie, a tým aj čistotu produktu na dne kolóny.

Etapa 4 riešila problém multikolinearity systematickým výberom príznakov. Ten zahŕňal vylúčenie súčinových príznakov etapy 2, fyzikálne motivované odstránenie príznakov ( $R/F$ ,  $T_B - p_B$ ,  $T_{B,1,in}$ ) a následný výber príznakov na základe korelačnej analýzy. Výsledok lineárnej regresie (CV RMSE = 0,1127) je najlepší spomedzi všetkých

iteratívnych prístupov, čo potvrdzuje, že eliminácia redundantných príznakov priamo prispieva k zlepšeniu generalizačnej schopnosti modelu.

Etapa 5 odhalila najlepšiu nájdenú konfiguráciu lineárnej regresie v rámci prehľadávaného priestoru s CV RMSE = 0,1121 pri  $p = 5$ . Použitá bola konfigurácia príznakov:  $\{T_B, T_{c,D}, Q_B/F, \Delta T_{col}, \Delta T_{B,1}\}$ . Globálne optimum<sup>1</sup> neurónovej siete tvorila konfigurácia  $\{T_{c,D}, T_B^2, Q_B/F, \Delta T_{col}\}$  s CV RMSE = 0,1140 pri  $p = 4$ . Pozoruhodné je, že NN optimum využíva  $T_B^2$  (kvadratický príznak z etapy 2), čo naznačuje, že vzťah medzi  $T_B$  a  $x_B$  obsahuje nelinearitu, ktorú neurónová sieť dokáže zachytiť lepšie ako lineárna regresia pri dostatočnom počte dát. Príznačky  $Q_B/F$ ,  $T_{c,D}$  a  $\Delta T_{col}$  sa opakovane objavujú vo väčšine najlepších kombinácií, čo poukazuje na ich vysokú informatívnosť vzhľadom na cieľovú premennú.

Kombinácia príznakov vybraná v etape 4 dosiahla CV RMSE = 0,1126, čo predstavuje rozdiel  $\approx 0,4\%$  oproti najlepšej nájdennej konfigurácii etapy 5. Tento výsledok potvrdzuje účinnosť zvoleného postupu: kombináciou fyzikálne motivovaných kritérií a korelačnej analýzy sa podarilo identifikovať takmer globálne optimálnu kombináciu príznakov bez prehľadávania všetkých 4943 možností.

Zaujímavé výsledky dosahujú aj trojpríznakové modely. Najlepšia trojpríznaková konfigurácia je  $\{T_B, T_{c,D}, Q_B/F\}$  s CV RMSE = 0,1147, čo predstavuje len  $2,3\%$  zhoršenie oproti päťpríznakovej najlepšej nájdennej konfigurácii.

Porovnanie naprieč etapami ukazuje, že samotné hodnoty korelácie príznakov s cieľovou premennou nie sú dostatočným kritériom výberu. Rozhodujúca je aj vzájomná nezávislosť príznakov a ich fyzikálna interpretovateľnosť. To potvrdzuje neúspech etapy 2 napriek vysokým individuálnym koreláciám, ako aj úspech etapy 4 vďaka eliminácii redundantných príznakov.

## 4.2 Porovnanie modelov: lineárna regresia vs. neurónová sieť

Vo všetkých etapách dosiahla lineárna regresia porovnateľné alebo mierne lepšie výsledky ako neurónová sieť. Relatívny rozdiel najlepších CV RMSE nepresahuje  $5\%$ .

Prvou príčinou je malý rozsah datasetu (165 pozorovaní). Neurónová sieť nedokáže

---

<sup>1</sup>Pod pojmom *globálne optimum* sa tu rozumie najlepšia nájdená konfigurácia v rámci prehľadávaného priestoru príznakov a zvolenej metodológie, nie absolútne optimum.

plne využiť svoju nelineárnu kapacitu bez rizika pretrénovania. Regularizačné mechanizmy (dropout, predčasné zastavenie) toto riziko zmiernujú, avšak zároveň obmedzujú schopnosť siete nájsť komplexné vzory. Tento efekt sa prejavil už v etape 1, kde žiadne nelineárne transformácie neboli použité a lineárna regresia stále dosiahla porovnateľné výsledky.

Druhou príčinou, ktorá sa uplatňuje v etapách 3–5, je analytické zachytenie nelineárnych závislostí priamo v príznakoch (poddiely  $Q_B/F$ ,  $R/F$ ; teplotné rozdiely  $\Delta T$ ). Lineárny model pracujúci s takto transformovanými príznakmi dokáže implicitne modelovať nelineárne vzťahy v pôvodnom priestore merania, čím sa výhoda neurónovej siete ďalej znižuje.

Výsledky naznačujú rozdielnú citlivosť modelov na výber príznakov: v etape 4 LR pri  $p = 4$  pridá  $\Delta T_{\text{col}}$ , zatiaľ čo NN preferuje  $T_{c,B}$ .

Z praktického hľadiska má lineárna regresia pre priemyselnú implementáciu viaceré výhody: model je plne interpretovateľný (každý koeficient zodpovedá merateľnému príspevku premennej) a výpočtovo triviálny (jedna maticová operácia). Neurónová sieť môže poskytnúť vyššiu presnosť, ak je k dispozícii výrazne väčší dataset alebo ak sa vstupné dáta menia nelineárne v rôznych prevádzkových režimoch.

## 4.3 Odporúčanie finálnej konfigurácie softvérového senzora

Pri výbere finálnej konfigurácie je potrebné zvážiť nielen štatistickú presnosť, ale aj počet a typ fyzických meraní potrebných na výpočet príznakov, interpretovateľnosť a udržiavateľnosť modelu v prevádzke.

### 4.3.1 Fyzické senzory potrebné pre rôzne konfigurácie

Tabuľka 4.1 porovnáva fyzické merania potrebné pre vybrané konfigurácie. Kompaktná konfigurácia  $\{T_B, T_{c,D}, Q_B/F\}$  vyžaduje štyri fyzické merania: snímač teploty dna  $T_B$ , snímač teploty chladiaceho média  $T_{c,D}$ , merač tepelného výkonu  $Q_B$  a prietokomer napájaného toku  $F$ .  $Q_B$  je vo väčšine destilačných kolón dostupné z merania prietoku pary a jej entalpie. Ide teda o minimálnu a technicky realizovateľnú konfiguráciu.

Globálne LR optimum a iteratívny výber etapy 4 vyžadujú šesť fyzických meraní (pridávajú sa  $T_{c,B}$  resp.  $T_{B,1,\text{in}}$  pre teplotné rozdiely). Zlepšenie CV RMSE oproti

**Tabuľka 4.1:** Počet fyzických meraní pre vybrané konfigurácie softvérového senzora

| Konfigurácia                       | Príznamy                                              | Fyzické merania |
|------------------------------------|-------------------------------------------------------|-----------------|
| Globálne optimum LR (etapa 5)      | $T_B, T_{c,D}, Q_B/F, \Delta T_{col}, \Delta T_{B,1}$ | 6               |
| Kompaktná konfigurácia ( $p = 3$ ) | $T_B, T_{c,D}, Q_B/F$                                 | 4               |
| Iteratívny výber (etapa 4)         | $Q_B/F, T_{c,B}, T_B, T_{c,D}, \Delta T_{col}$        | 6               |

kompaktnej konfigurácii je menej ako 2,3 %, čo pre väčšinu priemyselných aplikácií neodôvodňuje dve dodatočné merania.

### 4.3.2 Odporúčaný model a príznaky

Na základe vyhodnotenia experimentu sa odporúča nasledujúca konfigurácia softvérového senzora:

- **Model:** lineárna regresia
- **Príznamy:**  $T_B, T_{c,D}, Q_B/F$
- **CV RMSE:** 0,1147 (LR, najlepší výsledok pre  $p = 3$ , etapa 5)
- **TEST RMSE:** 0,1037 / 0,1254 / 0,0943 / 0,1189 / 0,1369

Hodnoty Test RMSE sú porovnateľné s ostatnými modelmi experimentu, čo potvrdzuje, že redukcia na tri príznaky nezhoršila generalizačnú schopnosť modelu. Táto konfigurácia dosahuje najlepší kompromis medzi presnosťou, interpretovateľnosťou a nákladmi na implementáciu. Lineárna regresia s tromi fyzikálne motivovanými príznakmi je transparentný, robustný a ľahko laditeľný model, ktorý si nevyžaduje špecializovanú hardvérovú ani softvérovú infraštruktúru.

Pre aplikácie vyžadujúce maximálnu presnosť je možné rozšíriť konfiguráciu na päť príznakov  $\{T_B, T_{c,D}, Q_B/F, \Delta T_{col}, \Delta T_{B,1}\}$  (najlepšia nájdená LR konfigurácia, CV RMSE = 0,1121) za cenu dvoch dodatočných fyzických meraní. Ak je k dispozícii väčšie množstvo historických dát, neurónová sieť s konfiguráciou  $\{T_{c,D}, T_B^2, Q_B/F\}$  (CV RMSE = 0,1148) môže predstavovať vhodnú alternatívu.

## 4.4 Obmedzenia a zdroje neistoty

Výsledky najlepšieho modelu každej etapy na testovacích množinách (tabuľky 3.6–3.22) sú vo väčšine prípadov porovnateľné s hodnotami CV RMSE, čo potvrdzuje, že opakovaná krížová validácia poskytla spoľahlivý odhad výkonnosti modelov.

Napriek tomu je viditeľná variabilita výsledkov naprieč hodnotami generátora rozdelení. V etape 3 (LR) sa Test RMSE pohybuje v rozsahu 0,0964–0,1559, v etape 1 (LR) v rozsahu 0,1022–0,1492. Táto variabilita odráža citlivosť modelov na konkrétne rozloženie trénovacej a testovacej množiny: pri malom datasete môžu rôzne rozdelenia dát priniesť výrazne odlišné testovacie podmienky.

Etapy 4 a 5 (LR) vykazujú konzistentnejšie testové výsledky s nižším rozptylom. To naznačuje, že fyzikálne motivované príznaky s elimináciou multikolinearity vedú nielen k lepšiemu CV RMSE, ale aj k robustnejšiemu správaniu modelu naprieč rôznymi rozdeleniami dát.

Výkonnosť modelu pri reálnom nasadení závisí od distribúcie prevádzkových podmienok kolóny v čase. Ak sa nové prevádzkové dáta výraznejšie odlišujú od rozloženia trénovacích dát, presnosť predikcie môže klesnúť. Model teda nie je navrhnutý na dokonalú predikciu pri ľubovoľných podmienkach, ale na spoľahlivú predikciu v rámci prevádzkového rozsahu zachyteného v datasete.



## Záver

---

Cieľom práce bol návrh a porovnanie modelov softvérového senzora určeného na odhad molárneho zlomku  $x_B$  na dne destilačnej kolóny a hodnotenie ich presnosti pomocou opakovanej krížovej validácie (CV RMSE). Boli porovnané dva typy modelov (lineárna regresia a neurónová sieť) v piatich etapách líšiacich sa prístupom ku konštrukcii vstupných príznakov.

Experiment ukázal, že voľba a konštrukcia príznakov mala rozhodujúci vplyv na presnosť predikcie, zatiaľ čo typ modelu hral druhoradú úlohu. Lineárna regresia dosahovala vo všetkých etapách porovnateľné alebo mierne lepšie výsledky ako neurónová sieť (rozdiel nepresahoval 5%). Fyzikálne odôvodnené príznaky, najmä  $Q_B/F$ , výrazne zlepšili výkonnosť modelov. Eliminácia multikolineárnych príznakov bola nevyhnutnou podmienkou dosiahnutia lepšej generalizácie.

Najlepšiu presnosť spomedzi iteratívnych etáp dosiahla etapa 4: lineárna regresia s kombináciou  $\{Q_B/F, T_{c,B}, T_B, T_{c,D}, \Delta T_{col}\}$  s CV RMSE = 0,1127, čo je zlepšenie o 8,8 % oproti výkonnostnej základni z etapy 1. Najlepšia nájdená konfigurácia lineárnej regresie v rámci prehľadávaného priestoru ( $\{T_B, T_{c,D}, Q_B/F, \Delta T_{col}, \Delta T_{B,1}\}$ , CV RMSE = 0,1121), identifikovaná vyčerpávacím prehľadávaním v etape 5, sa líšila od výsledku etapy 4 o menej ako 0,5 %.

Odporúčanou konfiguráciou pre priemyselné nasadenie je lineárna regresia s trojicou príznakov  $\{T_B, T_{c,D}, Q_B/F\}$  (CV RMSE = 0,1147), ktorá vyžaduje iba štyri fyzikálne merania a dosahuje len 2,3 % horší výsledok ako päťpríznaková najlepšia nájdená konfigurácia.

Testové hodnotenia na nových dátach potvrdili konzistentnosť výsledkov krížovej validácie a ukázali, že modely etáp 4 a 5 vykazujú najnižší rozptyl naprieč rôznymi rozdeleniami dát.

Všetky výsledky boli získané na dátach z jednej konkrétnej destilačnej kolóny. Navrhnutý postup je preto platný v rámci tohto datasetu a jeho priama aplikácia na iné kolóny si vyžaduje overenie. Výber vhodných príznakov a ich fyzikálna interpretácia budú pravdepodobne závisieť od konkrétneho procesu. Ako smer budúceho výskumu sa odporúča overenie navrhnutého postupu na dátach z iných destilačných kolón a rozšírenie datasetu, ktoré by umožnilo plnšie využiť potenciál neurónových sietí.

# Literatúra

Natalia Bakhtadze, Evgeny M. Maximov, and Ravil T. Valiakhmetov. Fuzzy soft sensors for chemical and oil refining processes. *IFAC Proceedings Volumes*, 41(2):46–50, 2008. ISSN 1474-6670. doi: <https://doi.org/10.3182/20080706-5-KR-1001.00007>. URL <https://www.sciencedirect.com/science/article/pii/S1474667016389236>.

Petr Kadlec, Bogdan Gabrys, and Sibylle Strandt. Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4):795–814, 2009. ISSN 0098-1354. doi: <https://doi.org/10.1016/j.compchemeng.2008.12.012>. URL <https://www.sciencedirect.com/science/article/pii/S0098135409000076>.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 1st edition, 2013. ISBN 978-1-4614-7137-0. URL <https://www.statlearning.com/>.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL <https://arxiv.org/abs/1412.6980>.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Damodar N. Gujarati. *Basic Econometrics*. McGraw-Hill, 5th edition, 2009.

Eric B. Baum and David Haussler. What size net gives valid generalization? In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 1 (NIPS 1988)*, pages 81–90. Morgan-Kaufmann, 1989. URL <https://proceedings.neurips.cc/paper/1988/file/1d7f7abc18fcb43975065399b0d1e48e-Paper.pdf>.

- Martin Mojto, Karol Lubušký, Miroslav Fikar, and Radoslav Paulen. Data-based design of inferential sensors for petrochemical industry. *Computers & Chemical Engineering*, 148:107437, 2021. doi: 10.1016/j.compchemeng.2021.107437.
- Myke King. Practical process control part 14: Pressure compensated temperatures. *The Chemical Engineer*, May 2024. URL <https://www.thechemicalengineer.com/features/practical-process-control-part-14-pressure-compensated-temperatures/>.