Fuzzy logika v predspracovaní údajov a jej vplyv na výkonnosť modelu strojového učenia XGBOOST

Fuzzy logic in data preprocessing and its impact on the performance of the XGBoost machine learning model

Andrej Bednařík[1]

Abstrakt
Fuzzy logika poskytuje efektívny prístup k predspracovaniu číselných údajov v strojovom učení, najmä v regresii. Tento článok skúma vplyv fuzzyfikácie premenných, ako vek a BMI na presnosť predikcie nákladov na zdravotnú starostlivosť. Použitím fuzzy transformácie sme testovali výkon XGBoost regresora pri rôznych variantoch predspracovania datasetu. Výsledky naznačujú, že fuzzy logika môže v niektorých prípadoch zlepšiť presnosť predikcie (nižšie RMSE), najmä pri premenných s nejasnými hranicami. Diskutujeme tiež o situáciách, kde jej aplikácia neprináša zlepšenie, a identifikujeme scenáre, v ktorých je najvhodnejšia.

Kľúčové slová
Fuzzy, Regresia, XGBoost, Dáta

Abstract
Fuzzy logic provides an effective approach to preprocessing numerical data in machine learning, particularly in regression tasks. This paper explores the impact of fuzzification of variables such as age and BMI on the accuracy of healthcare cost prediction. By applying fuzzy transformation, we evaluated the performance of the XGBoost regressor across different dataset preprocessing variants. The results suggest that fuzzy logic can, in some cases, improve prediction accuracy (lower RMSE), especially for variables with unclear boundaries. We also discuss situations where its application does not lead to improvement and identify scenarios where it is most suitable.

Key words
Fuzzy, Regression, XGBoost, Data

JEL classification
C61, C89

1 Úvod

Fuzzy logika predstavuje prístup k spracovaniu neurčitých a nepresných informácií, ktorý je inšpirovaný spôsobom, akým ľudia uvažujú. Na rozdiel od klasickej binárnej logiky, kde hodnoty nadobúdajú len dve možnosti (pravda alebo nepravda, 0 alebo 1), fuzzy logika umožňuje hodnotám existovať v rozmedzí medzi týmito dvoma extrémami. Tento koncept bol prvýkrát predstavený Lotfi Zadehom v roku 1965 ako spôsob modelovania neurčitosti v reálnom svete (Zadeh, 1965). Základným princípom fuzzy logiky je, že premenné nie sú striktne

[1] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra matematiky a aktuárstva, Dolnozemská cesta 1, 852 35 Bratislava, andrej.bednarik@euba.sk.

klasifikované do pevne daných kategórií, ale môžu patriť do viacerých kategórií s rôznou mierou príslušnosti. Napríklad teplota môže byť "nízka", "stredná" alebo "vysoká", pričom jedna konkrétna hodnota môže súčasne patriť do viacerých týchto skupín s rôznou intenzitou (Klir & Yuan, 1995). Táto flexibilita umožňuje efektívnejšie modelovanie systémov, ktoré pracujú s neurčitosťou, a nachádza široké uplatnenie v oblastiach ako riadiace systémy, rozpoznávanie vzorov, umelá inteligencia a spracovanie prirodzeného jazyka (Ross, 2010). V praxi sa fuzzy logika využíva na reprezentáciu neurčitých konceptov prostredníctvom tzv. funkcií členstva. Tieto funkcie určujú mieru príslušnosti konkrétnej hodnoty k danej fuzzy množine. Medzi najčastejšie používané patria trojuholníkové, lichobežníkové a gaussovské funkcie členstva. Napríklad, ak chceme definovať fuzzy množiny pre vek človeka, môžeme zaviesť kategórie "mladý", "stredný vek" a "starý" s plynulými prechodmi medzi nimi (Mendel, 2001). Fuzzy logika je tiež úzko spojená s fuzzy inferenciou, čo je proces odvodenia záverov na základe fuzzy pravidiel. Tieto pravidlá sú často formulované v podobe "Ak – Potom" (napr. "Ak je teplota vysoká, potom ventilátor beží na plný výkon"). Na spracovanie týchto pravidiel sa využívajú metódy ako Mamdaniho a Sugenov model inferencie, ktoré sa aplikujú v rôznych inžinierskych a vedeckých disciplínach (Ross, 2010). V oblasti strojového učenia môže fuzzy logika zlepšiť interpretovateľnosť a robustnosť modelov tým, že umožňuje efektívnejšie spracovanie číselných premenných s neurčitými hranicami. Napríklad, v predikčných modeloch môže byť vek alebo krvný tlak transformovaný do fuzzy premenných, čím sa lepšie zachytia ich skutočné vplyvy na cieľovú premennú (Mendel, 2001). Tento prístup sa osvedčil najmä v prípadoch, kde tradičné metódy spracovania údajov narážajú na problémy so striktným kategorizovaním premenných. Celkovo fuzzy logika predstavuje silný nástroj na modelovanie neurčitosti v rôznych oblastiach. V kombinácii so strojovým učením môže pomôcť pri efektívnejšom spracovaní údajov a zlepšení výkonnosti modelov, čo ju robí atraktívnou vo výskume aj priemyselnej praxi.

2 Fuzzy logika v predspracovaní dát

Predspracovanie dát je kľúčovým krokom v strojovom učení, ktorý ovplyvňuje presnosť a robustnosť modelov. Fuzzy logika ponúka alternatívu k tradičným metódam, najmä pri práci s kontinuálnymi premennými, ktoré nemajú presne definované hranice medzi kategóriami. Použitím fuzzyfikácie môžeme získať flexibilnejšiu reprezentáciu údajov a eliminovať skreslenia spôsobené ostrými prahmi kategorizácie. Dôvody prečo dáta fuzzyfikovať:

- Zachovanie kontinuity informácií: Tradičná kategorizácia (napr. vek rozdelený do skupín 20–30, 30–40) môže spôsobiť stratu informácií. Fuzzy reprezentácia umožňuje plynulý prechod medzi kategóriami.
- Redukcia šumu: Pri premenných, ako je krvný tlak alebo BMI, môže byť meranie ovplyvnené rôznymi faktormi. Fuzzy logika pomáha zmierniť vplyv náhodných odchýlok.

Zlepšenie interpretovateľnosti modelov: Fuzzy premenné môžu byť lepšie pochopiteľné pre analytikov a doménových expertov, najmä v medicíne alebo poisťovníctve. Príklady fuzzyfikácie premenných:

- Vek:
  - Klasický prístup: vek = 35 (jedna číselná hodnota)
  - Fuzzy prístup: vek môže byť súčasne "mladý" (0.2) a "stredný vek" (0.8), čo lepšie vystihuje realitu prechodu medzi skupinami.

- BMI:
  - Klasický prístup: BMI = 28 → kategória "nadváha".
  - Fuzzy prístup: BMI = 28 → príslušnosť k "normálna váha" (0.3) a "nadváha" (0.7), čím sa zachytí plynulý prechod medzi stavmi.

- Krvný tlak:
  - Klasický prístup: hodnota 139 mmHg sa môže považovať za normálnu, ale 140 mmHg už ako vysoký tlak.
  - Fuzzy prístup: Hodnota 139 mmHg môže mať členstvo 0.8 v "normálny" a 0.2 v "vysoký tlak", čím sa predíde ostrým hraniciam.

Fuzzyfikácia premenných sa realizuje pomocou funkcií členstva, ktoré definujú mieru príslušnosti danej hodnoty k jednotlivým fuzzy množinám. Bežne sa používajú:

- Trojuholníkové funkcie – jednoduchá reprezentácia pre plynulé prechody.
- Lichobežníkové funkcie – vhodné pre premenné s rozsiahlymi strednými hodnotami.
- Gaussovské funkcie – používajú sa v prípadoch, kde je potrebné modelovať postupné zmeny bez ostrých hraníc.

V ďalších častiach článku sa zameriame na praktickú implementáciu fuzzyfikácie pri analýze dát a hodnotenie jej vplyvu na výkonnosť predikčných modelov.

3 Metodika a experimentálny dizajn

V tejto časti skúmame vplyv fuzzy logiky na výkonnosť modelov strojového učenia pri predikcii nákladov na zdravotnú starostlivosť. Cieľom je overiť, či fuzzyfikácia niektorých numerických premenných môže viesť k zlepšeniu presnosti predikcie. Zameriavame sa predovšetkým na premenné, ktoré majú neostré alebo subjektívne definovateľné hranice – ako napríklad BMI či vek.

Obr. 1: Náhľad datasetu

age	sex	bmi	children	smoker	region	charges
18	male	33.77	1	no	southeast	1725.5523
19	male	24.6	1	no	southwest	1837.237
18	male	34.1	0	no	southeast	1137.011
18	female	26.315	0	no	northeast	2198.18985
19	female	28.6	5	no	southwest	1728.897
19	male	20.425	0	no	northwest	1625.43375
18	female	38.665	2	no	northeast	1728.897
18	female	35.625	0	no	northeast	2211.13075
19	female	28.9	0	no	southwest	1743.214
18	female	30.115	0	no	northeast	1728.897

Zdroj: Vlastné spracovanie

Metodika pozostáva z nasledujúcich krokov:

1. **Fuzzyfikácia vybraných numerických premenných** – pomocou preddefinovaných fuzzy množín transformujeme pôvodné hodnoty (napr. BMI) na fuzzy reprezentácie.

2. **Tréning modelov pomocou algoritmu XGBoost** – tento výkonný boostingový algoritmus aplikujeme na rôzne verzie datasetu, aby sme mohli porovnať vplyv fuzzy predspracovania.
3. **Porovnanie výsledkov** – hodnotíme výkonnosť modelov pomocou metriky RMSE (root mean squared error), pričom analyzujeme, v ktorých prípadoch fuzzy logika prispela k zlepšeniu a kedy nie.

Testujeme tri varianty modelov:

1. Model bez fuzzyfikácie – štandardný model pracujúci s pôvodnými numerickými hodnotami bez akéhokoľvek rozmazania hraníc medzi kategóriami.
2. Model s fuzzyfikáciou BMI – v tomto prípade je premenná BMI transformovaná do fuzzy priestoru pomocou jazykových hodnôt ako „nízky“, „normálny“, „vysoký“ a „veľmi vysoký“.

Obr. 2: Fuzzyfikácia premennej Bmi

```
def fuzzify_bmi(row):
    bmi = row['bmi']
    underweight, normal, overweight, obesity = 0, 0, 0, 0

    if bmi < 18.5:
        underweight = 1
    elif 18.5 <= bmi < 24.9:
        normal = (bmi - 18.5) / (24.9 - 18.5)  # Increasing from 0 to 1
        underweight = 1 - normal               # Decreasing from 1 to 0
    elif 24.9 <= bmi < 29.9:
        overweight = (bmi - 24.9) / (29.9 - 24.9)  # Increasing from 0 to 1
        normal = 1 - overweight                    # Decreasing from 1 to 0
    elif bmi >= 30:
        obesity = 1
```

Zdroj: Vlastné spracovanie

3. Model s fuzzyfikáciou veku – kde vek bude transformovaný do fuzzy premenných

Obr. 3: Fuzzyfikácia premennej vek

```
def fuzzify_age(row):
    age = row['age']
    young, middle, old = 0, 0, 0

    if age <= 25:
        young = 1
    elif 25 < age <= 30:
        young = 1 - (age - 25) / 5  # Decreasing from 1 to 0
        middle = (age - 25) / 5     # Increasing from 0 to 1
    elif 30 < age <= 50:
        middle = 1
    elif 50 < age <= 55:
        middle = 1 - (age - 50) / 5  # Decreasing from 1 to 0
        old = (age - 50) / 5         # Increasing from 0 to 1
    elif age > 55:
        old = 1
```

Zdroj: Vlastné spracovanie

Použitý dataset obsahuje premenné ako vek, pohlavie, BMI, počet detí, fajčenie, región a cieľovú premennú `charges` (náklady na zdravotnú starostlivosť). Modely budú vyhodnotené pomocou metriky RMSE na testovacej množine. Na vytvorenie a trénovanie modelu použijeme XGBoost pre regresiu, ktorý umožňuje efektívne pracovať s nelineárnymi vzťahmi v dátach. Modely budú vyhodnotené pomocou metriky RMSE na testovacej množine. Okrem toho budeme

využívať K- Fold cross-validáciu s náhodným premiešavaním dát (shuffle), aby sme zabezpečili robustnejšie vyhodnotenie výkonu modelov a minimalizovali vplyv konkrétneho rozdelenia datasetu. Výsledky experimentov nám umožnia zhodnotiť, v ktorých situáciách fuzzy logika prispieva k lepšiemu modelovaniu reálnych, často neostro definovaných javov, ako sú zdravotné riziká spojené s nadváhou alebo vekom.

Nastavenie hyperparametrov XGBoost modelu:

- **n_estimators:** [5, 10, 20, 30, 40, 60, 70, 100] – Počet stromov v modeli, experimentálne testujeme viacero hodnôt.
- **max_depth:** 3 – Maximálna hĺbka stromov, zabraňuje nadmernému prispôsobeniu (overfittingu).
- **learning_rate:** 0.1 – Rýchlosť učenia, nižšia hodnota umožňuje postupnejšie optimalizovanie váh.
- **subsample:** 0.9 – Podiel náhodne vybraných vzoriek pri trénovaní každého stromu, zlepšuje generalizáciu.
- **random_state:** 42 – Zabezpečuje reprodukovateľnosť výsledkov.
- **colsample_bytree:** 1 – Použitie všetkých premenných pri konštrukcii jednotlivých stromov.

3 Experimentálne výsledky a hodnotenie modelov

Ako prvý sme natrénovali model bez fuzzyfikácie, ktorý pracoval s pôvodnými numerickými hodnotami všetkých premenných. Tento model slúžil ako referenčná základňa na porovnanie s ďalšími verziami, kde sme aplikovali fuzzy úpravy vybraných premenných. Následne sme vytvorili modely s fuzzyfikáciou veku a BMI, aby sme analyzovali ich vplyv na výkonnosť modelu a presnosť predikcií.

Obr. 4: Výsledky modelu bez fuzzyfikácie

```
Best number of estimators: 70
Best RMSE (cross-validated): 1906.30
```

Zdroj: Vlastné spracovanie

Výsledky na obrázku 4 ukazujú, že optimálny počet stromov (n_estimators) je 70, čo znamená, že model dosiahol najlepšiu výkonnosť pri trénovaní s týmto počtom stromov. Menej stromov by mohlo viesť k underfittingu, zatiaľ čo viac stromov by nemuselo priniesť výrazné zlepšenie a mohlo by viesť k miernemu overfittingu.Hodnota Root Mean Squared Error (RMSE) = 1906.30 vyjadruje priemernú chybu predikcie v jednotkách cieľovej premennej, teda v tomto prípade v eurách. To znamená, že priemerná odchýlka medzi skutočnými a predikovanými nákladmi na zdravotnú starostlivosť je približne 1906 eur. Nižšia hodnota RMSE znamená presnejší model, preto je dôležité porovnať tento výsledok s modelom bez fuzzy logiky a s fuzzyfikáciou iných premenných, aby sme vyhodnotili prínos tohto prístupu.

Obr. 5: Náhľad datasetu s fuzzyfikáciou (BMI)

	age	sex	bmi	children	smoker	region	charges	underweight	normal	overweight	obesity
1	18	male	33.77	1	no	southeast	1725.5523	0.0	0.0	0.0	1.0
2	19	male	24.6	1	no	southwest	1837.237	0.046874999999999556	0.9531250000000004	0.0	0.0
3	18	male	34.1	0	no	southeast	1137.011	0.0	0.0	0.0	1.0
4	18	female	26.315	0	no	northeast	2198.18985	0.0	0.7169999999999994	0.28300000000000053	0.0
5	19	female	28.6	5	no	southwest	1728.897	0.0	0.25999999999999945	0.7400000000000005	0.0
6	19	male	20.425	0	no	northwest	1625.43375	0.6992187499999998	0.30078125000000017	0.0	0.0
7	18	female	38.665	2	no	northeast	1728.897	0.0	0.0	0.0	1.0
8	18	female	35.625	0	no	northeast	2211.13075	0.0	0.0	0.0	1.0
9	19	female	28.9	0	no	southwest	1743.214	0.0	0.19999999999999996	0.8	0.0
10	18	female	30.115	0	no	northeast	1728.897	0.0	0.0	0.0	1.0
11	19	female	28.4	1	no	southwest	2331.519	0.0	0.30000000000000004	0.7	0.0
12	18	male	23.75	0	no	northeast	1705.6245	0.17968749999999978	0.8203125000000002	0.0	0.0

Zdroj: Vlastné spracovanie

Po fuzzyfikácii datasetu otestujeme, či v modeli XGBoost fuzzyfikácia BMI zlepší presnosť predikcií. Skúmame dva prístupy: (1) úplné nahradenie pôvodnej premennej BMI fuzzy príslušnosťami alebo (2) zachovanie pôvodnej hodnoty BMI spolu s fuzzy reprezentáciou. Cieľom je zistiť, ktorá z týchto stratégií vedie k presnejšiemu modelu.

```
Best number of estimators: 70
Best RMSE (cross-validated): 2102.04
```

Zdroj: Vlastné spracovanie

Obr. 6: Výsledky experiment BMI prístup 1

Na základe uvedených výsledkov na obrázku 6 môžeme pozorovať, že počet optimálnych stromov (n_estimators) ostal rovnaký na hodnote 70, čo znamená, že zmena v predspracovaní údajov (fuzzyfikácia) neovplyvnila potrebu vyššej ani nižšej modelovej kapacity. Hodnota RMSE (cross- validated) sa však zvýšila o 195.74 bodov na 2102.04, čo predstavuje zhoršenie oproti pôvodnému modelu bez fuzzyfikácie, ktorý dosiahol RMSE 1906.30.

Obr. 7: Výsledky experiment BMI prístup 2

```
Best number of estimators: 70
Best RMSE (cross-validated): 1901.91
```

Zdroj: Vlastné spracovanie

Na základe výsledkov na obrázku 7 môžeme povedať, že počet optimálnych stromov (n_estimators) ostal rovnaký na hodnote 70. Hodnota RMSE (cross-validated) sa však v tomto prístupe znížila o 4.39 bodov na 1901.91, čo predstavuje zlepšenie predikcii oproti pôvodnému modelu bez fuzzyfikácie, ktorý dosiahol RMSE 1906.30.

Výsledky experimentu naznačujú, že prítomnosť pôvodného stĺpca spolu s jeho „fuzzy“ verziou modelu neškodí a zároveň jeho vynechanie nijako model výrazne nezlepšuje no naopak zhoršuje. Na základe tohto zistenia pre premennú vek vykonáme len experiment s prístupom 2 .

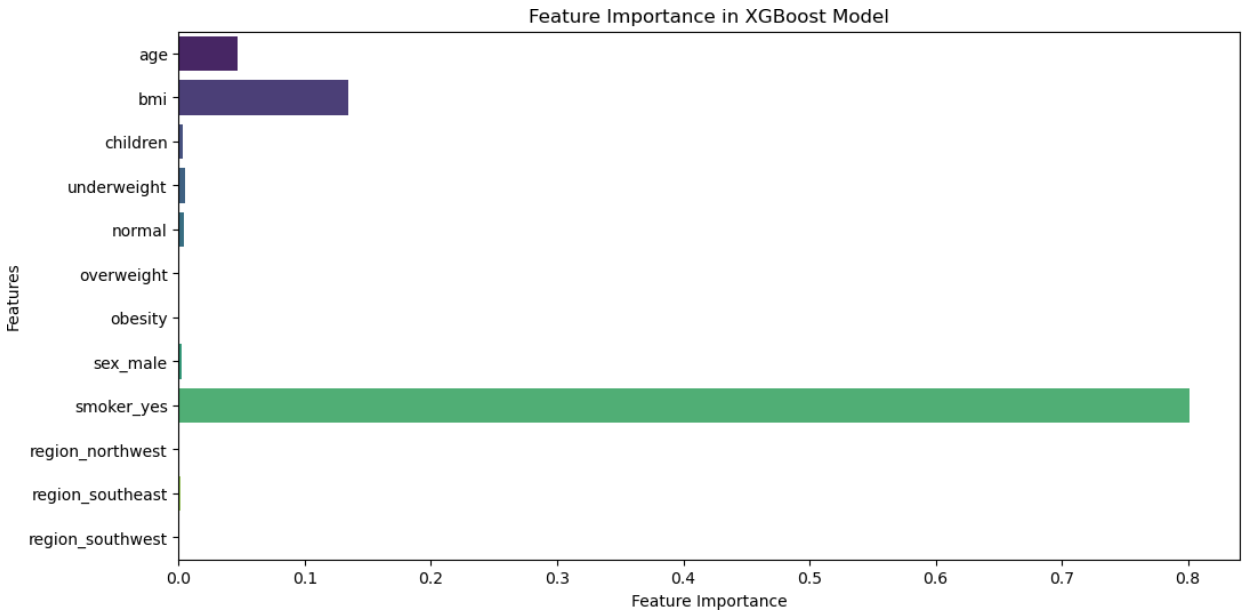
Obr. 8: Výsledky experiment Vek

```
Best number of estimators: 70
Best RMSE (cross-validated): 1908.01
```

Zdroj: Vlastné spracovanie

Na základe výsledkov na obrázku 8 môžeme povedať, že počet optimálnych stromov (n_estimators) ostal rovnaký na hodnote 70. Hodnota RMSE (cross-validated) sa však v tomto prístupe zvýšila o 1.71 bodu na 1908.01, čo predstavuje zhoršenie predikcii oproti pôvodnému modelu bez fuzzyfikácie, ktorý dosiahol RMSE 1906.30.

Na základe obrázku 9 môžeme povedať, že premenná "smoker_yes" má najvyššiu dôležitosť – to znamená, že najväčšiu časť variability v predikcii nákladov na zdravotnú starostlivosť vysvetľuje práve informácia o tom, či je osoba fajčiar alebo nie. V tomto prípade fuzzyfikácia iných premenných nemusí mať zásadný dopad, pretože hlavný prediktor je už jednoznačný a binárny (0/1). Premenná "BMI" je druhá najvýznamnejšia – keďže BMI má stále relatívne veľký vplyv na predikované hodnoty, fuzzyfikácia mohla pomôcť modelu lepšie zachytiť jeho vzťah k cieľovej premennej, čo viedlo k zlepšeniu RMSE. Premenná "age" má relatívne nízku dôležitosť – keďže vek je menej významný v porovnaní s BMI a fajčením, fuzzyfikácia veku by mohla priniesť len minimálne zlepšenie alebo zhoršenie ako aj vyplynulo z našich experimentov, pretože samotná premenná nie je kľúčovým faktorom v predikcii. Ostatné premenné (children, región, podkategórie BMI) majú veľmi nízku dôležitosť, čo naznačuje, že ich fuzzyfikácia by pravdepodobne nemala významný vplyv na zlepšenie modelu.



Zdroj: Vlastné spracovanie

Obr. 9: Feature importance

4 Záver

Výsledky experimentov ukázali, že fuzzyfikácia premenných môže v niektorých prípadoch prispieť k zlepšeniu presnosti modelu, avšak jej efektivita závisí od viacerých faktorov. V našej analýze sa ukázalo, že fuzzyfikácia veku nepriniesla zlepšenie no naopak zhoršenie, zatiaľ čo fuzzyfikácia BMI v kombinácii s ponechaním pôvodnej hodnoty mierne znížila RMSE, čím sa zvýšila presnosť predikcií. Naopak, úplné odstránenie pôvodných hodnôt pri fuzzyfikácii spôsobilo výrazné zhoršenie výkonu modelu vytvoreného pomocou XGBOOST.Jedným z kľúčových poznatkov je, že významnosť premenných hrá zásadnú úlohu v tom, či fuzzyfikácia prinesie zlepšenie. V našom datasete bola najvýznamnejšou premennou binárna kategória „smoker_yes“, ktorá rozhodujúcim spôsobom ovplyvňovala predikované náklady na zdravotnú starostlivosť. To naznačuje, že v datasetoch, kde dominujú diskrétne alebo binárne kategórie, nemusí fuzzyfikácia číselných premenných výrazne pomôcť. Okrem presnosti modelu je dôležité zohľadniť aj jeho komplexnosť a výpočtovú náročnosť. Pridanie nových fuzzy premenných zväčšuje dimenzionalitu datasetu, čo môže viesť k vyššej výpočtovej záťaži a potenciálnym problémom s interpretáciou modelu. Preto je nutné vždy zvážiť, či mierne zvýšenie presnosti modelu kompenzuje túto dodatočnú zložitosť. Z uvedených zistení vyplýva, že fuzzyfikácia môže byť užitočná pri vhodných premenných, ale nie vždy je prínosná. V prípade datasetov, kde sú hlavné prediktory binárne alebo kategorické, môže byť jej efekt minimálny. Preto by sa mala fuzzyfikácia aplikovať selektívne, s ohľadom na významnosť premenných a celkový dopad na model. V ďalších výskumoch by bolo vhodné preskúmať, ako fuzzyfikácia ovplyvňuje modely pri iných typoch datasetov, kde by mohla mať väčší prínos – napríklad v prípadoch, kde sú všetky hlavné prediktory kontinuálne a majú nejednoznačné hranice medzi kategóriami. Najdôležitejším aspektom je správne rozdelenie fuzzy intervalov, pretože nesprávne zvolená fuzzy reprezentácia môže skresliť vzťahy medzi premennými a viesť k strate relevantných informácií. Optimálne nastavenie fuzzy intervalov je preto kľúčové pre dosiahnutie maximálneho prínosu tejto metódy.

Tento príspevok vznikol v rámci výskumného projektu
VEGA 1/0497/25, *Implementácia inovatívnych prístupov v oblasti riadenia a modelovania rizík v rámci interných modelov poisťovní.*
VEGA č. 1/0377/25 Inovatívne metódy Enterprise risk managementu a ich využitie v riadení a modelovaní rizík.

Literatúra

1. Klir, G. J., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic: Theory and applications*. Prentice Hall.
2. Mendel, J. M. (2001). *Uncertain rule-based fuzzy logic systems: Introduction and new directions*. Prentice Hall.
3. Ross, T. J. (2010). *Fuzzy logic with engineering applications* (3rd ed.). John Wiley & Sons.
4. Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X