

# Metode regularizacije s uključenim domenskim znanjem stručnjaka za odabir značajki u linearnom regresijskom modelu

---

Mokriš, Marinela

Doctoral thesis / Disertacija

2025

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Organization and Informatics / Sveučilište u Zagrebu, Fakultet organizacije i informatike**

Permanent link / Trajna poveznica: <https://urn.nsk.hr/urn:nbn:hr:211:096890>

Rights / Prava: [In copyright](#) / [Zaštićeno autorskim pravom](#).

Download date / Datum preuzimanja: **2025-04-01**



Repository / Repozitorij:

[Faculty of Organization and Informatics - Digital Repository](#)





Sveučilište u Zagrebu

Fakultet organizacije i informatike

Marinela Mokriš

**Metode regularizacije s uključenim  
domenskim znanjem stručnjaka za  
odabir značajki u linearnom  
regresijskom modelu**

DOKTORSKI RAD

Varaždin, 2024.



Sveučilište u Zagrebu

Marinela Mokriš

**Metode regularizacije s uključenim  
domenskim znanjem stručnjaka za  
odabir značajki u linearnom  
regresijskom modelu**

DOKTORSKI RAD

Mentori:

prof. dr. sc. Diana Šimić,

doc. dr. sc. Dario Šebalj

Varaždin, 2024.



University of Zagreb

Marinela Mokriš

**Regularization methods that include  
experts' domain knowledge for feature  
selection in a linear regression model**

DOCTORAL THESIS

Supervisors:

Diana Šimić, PhD, full professor,

Dario Šebalj, PhD, assistant professor

Varaždin, 2024.

# ZAHVALA

Hvala Hrvatskoj zakladi za znanost što podupire razvoj mladih istraživača kroz „Projekt razvoja karijera mladih istraživača – izobrazba novih doktora znanosti”, u okviru kojega sam i ja upisala doktorski studij.

Hvala Agenciji za pravni promet i posredovanje nekretninama, Centru za praćenje poslovanja energetskeg sektora i investicija i Fondu za zaštitu okoliša i energetske učinkovitost na ustupljenim podacima.

Hvala stručnjacima na dijeljenju svog domenskog znanja.

Hvala Sveučilišnom računskom centru (Srce) Sveučilišta u Zagrebu koji su ustupili korištenje usluge Naprednog računanja (Padobran) za izradu dijela ovog istraživanja.

Hvala M., zahvaljujući kojoj sam upisala doktorski studij i krenula na ovo putovanje i koja, slijedom životnih okolnosti, nije bila formalna mentorica ovog doktorskog rada, ali je bila mentorica u svakom smislu te riječi.

Hvala mentorima na nesebičnom dijeljenju znanja, svim savjetima i podršci tijekom ovog puta. Bilo ih je zaista mnogo i uložili ste puno svoga vremena u mene. Mnogo toga sam naučila tijekom ovog putovanja zahvaljujući vama i, zato, hvala.

Hvala Povjerenstvu za ocjenu doktorskog rada na njihovim primjedbama za poboljšanje rada.

Hvala kolegama na pruženoj podršci i motivaciji.

Na kraju, hvala mojim bližnjima na pruženoj ljubavi i potpori.

# INFORMACIJE O MENTORIMA

## **Prof. dr. sc. Diana Šimić**

Prof. dr. sc. Diana Šimić rođena je u Zagrebu, gdje je diplomirala matematičku informatiku i statistiku na Matematičkom odjelu Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu. Na istome je fakultetu završila znanstveni magistarski poslijediplomski studij, a doktorirala je na Medicinskom fakultetu Sveučilišta u Zagrebu. Radno je iskustvo stjecala kao programerka u RIZ-u Tvornici odašiljača, kao istraživačica, voditeljica Laboratorija za biomatematiku te članica i predsjednica Upravnog vijeća Instituta za medicinska istraživanja i medicinu rada. Bila je pomoćnica ministra za informatiku na Ministarstvu znanosti i tehnologije te zamjenica državnog tajnika u Središnjem državnom uredu za e-Hrvatsku. Radila je i kao konzultantica za e-javnu upravu i e-poslovanje u Infodomu. Od 2010. godine radi na Fakultetu organizacije i informatike Sveučilišta u Zagrebu (FOI), gdje je 2021. godine izabrana u zvanje redovite profesorice u trajnom zvanju na predmetima vezanim uz statističke, kvantitativne te znanstvene metode u informacijskim znanostima. Od 2011. do 2014. bila je prodekanica za znanstvenoistraživački rad, a od 2016. godine voditeljica je doktorskog studija Informacijske znanosti. Članica je Katedre za kvantitativne metode i Laboratorija za analitiku učenja i akademsku analitiku (koji je vodila od 2018. do 2020.) Sudjelovala je u više od 40 znanstvenih projekata, mentorirala 4 doktorska, 12 diplomskih i 2 završna rada. U bazama platforme Web of Science Core Collection indeksirana su joj 43 rada (499 citata, h-indeks 14), a zajedno s drugim bazama na ovoj platformi ukupno joj je indeksirano 60 radova. Koautorica je i 11 knjiga te 4 poglavlja u knjigama. Bila je pozvani predavač na više od 20 konferencija. Članica je *Association for Computing Machinery* (ACM), *American Statistical Association* (ASA), Hrvatskog biometrijskog društva (HBMD, u kojem je obnašala dužnosti predsjednice, dopredsjednice, tajnice i članice Predsjedništva te članice Međunarodnog programskog i Organizacijskog odbora konferencije BIOSTAT), Hrvatskog statističkog društva i Hrvatskog društva za medicinsku in-

formatiku (HDMI). Od 2016. godine potpredsjednica je Hrvatskog informatičkog zbora (HIZ). Dobitnica je plakete Informatika Hrvatskog informatičkog zbora za doprinos razvoju informacijskog društva u Hrvatskoj (2007.), priznanja za poseban doprinos napretku i prepoznatljivosti Sveučilišnog računskog centra (2011.), *Europe's Open Access Champions* (2016.), posebne zahvale Dekana FOI za doprinos razvoju doktorskog studija (2017.), posebne zahvale Dekana FOI za promicanje vidljivosti fakulteta kroz projekt e-Škole (2017.), posebnog priznanja Dekanice za izuzetno zalaganje i doprinos Fakultetu (2020.) te nagrade za životno djelo Fakulteta organizacije i informatike (2023.)

### **Doc. dr. sc. Dario Šebalj**

Doc. dr. sc. Dario Šebalj rođen je u Vinkovcima, gdje je završio osnovnu i srednju školu. Na Ekonomskom fakultetu u Osijeku završio je preddiplomski i diplomski studij Poslovne ekonomije, smjer Poslovna informatika. Na istom fakultetu je i doktorirao 2019. godine. Na Ekonomskom fakultetu u Osijeku zaposlen je od 2014. godine kao asistent na Katedri za kvantitativne metode i informatiku. U travnju 2020. godine biran je na radno mjesto poslijedoktoranda, a 2021. godine na radno mjesto docenta na istoj katedri na kojoj sudjeluje u izvođenju nastave na više kolegija iz područja poslovne informatike. Od 2023. godine voditelj je smjera Poslovna informatika koji je dio sveučilišnog diplomskog studija Poslovna ekonomija. Od 2014. do 2020. godine obnašao je funkciju urednika novaka, a od 2020. do 2023. godine izvršnog urednika časopisa *Ekonomski vjesnik*. Od 2017. do 2019. godine bio je tajnik udruge Alumni EFOS. Objavio je preko 25 znanstvenih radova u časopisima i zbornicima s međunarodnih skupova te jedno poglavlje u knjizi. Kao član organizacijskog odbora sudjelovao je i u organizaciji međunarodnih znanstvenih konferencija *Business Logistics in Modern Management* (od 2015. do danas), *Interdisciplinary Management Research* (2019. godine) te *Conference on Operational Research KOI* (2016. godine). Bio je i član programskog odbora međunarodne znanstvene konferencije IBIMA (2020. godine), odnosno znanstvenog odbora međunarodne znanstvene konferencije EBES (2020. godine). Bio je istraživač na znanstvenom projektu Hrvatske zaklade za znanost te na dva Erasmus+ projekta. Član je organizacije *Project Management Institute* (PMI) i *International Institute of Business Analysis* (IIBA).

# SAŽETAK

Ovo istraživanje imalo je cilj razviti metodu za odabir značajki u linearnom regresijskom modelu koja u regularizaciju uključuje domensko znanje stručnjaka o utjecaju ulaznih značajki na izlaznu, pri čemu se razlikovalo dva načina uključivanja znanja stručnjaka – deterministički i metaheuristički. Predložena metoda se evaluirala na podacima o troškovima energije i karakteristikama zgrada javnog sektora Republike Hrvatske te na tri scenarija simuliranih podataka iz literature. Za evaluaciju su se koristile mjere uspješnosti predikcije (korigirani koeficijent determinacije, *RMSE*, *MAE* i *SMAPE*) te, u slučaju simuliranih podataka, razlika između stvarnih i procijenjenih regresijskih koeficijenata. Metoda je uspoređena s regularizacijom koja ne uključuje domensko znanje stručnjaka te s regresijskim stablima. Rezultati na stvarnim podacima pokazuju da je uvođenje domenskog znanja stručnjaka metaheurističkim pristupom rezultiralo poboljšanjem uspješnosti predikcije u odnosu na spomenute metode, dok je uvođenje domenskog znanja stručnjaka determinističkim pristupom u nekim slučajevima rezultiralo poboljšanjem uspješnosti predikcije. U okviru simulacijskog eksperimenta, uvođenje domenskog znanja stručnjaka rezultiralo je poboljšanjem nekih mjera uspješnosti predikcije u odnosu na regularizaciju koja ne koristi znanje stručnjaka samo kod metaheurističkog pristupa, dok predložena metoda uvođenja domenskog znanja stručnjaka u regularizaciju nije rezultirala uspješnijom predikcijom od metode regresijskih stabala niti za jedan pristup uvođenja domenskog znanja stručnjaka. Rezultati pokazuju da različiti načini procjene penalizacijskih koeficijenata i sama struktura modela utječu na varijabilnost i pristranost modela. Ujedno, otvaraju daljnja pitanja u smislu kako postići i da za kompleksnije skupove podataka domensko znanje stručnjaka doprinese smanjenju varijabilnosti i / ili pristranosti modela te poboljšanju uspješnosti predikcije.

**Ključne riječi:** odabir značajki, linearna regresija, regularizacija, domensko znanje stručnjaka, metaheuristika, pretraživanje kukavice, regresijska stabla, predikcija, energetski troškovi, zgrade javne namjene, simulacije



# SUMMARY

This research aimed to develop a feature selection method for a linear regression model that incorporates experts' domain knowledge about the impact of input features on the output into the regularization process. Two methods of incorporating experts' knowledge were distinguished: deterministic and metaheuristic. The proposed method was evaluated using data on energy costs and characteristics of public sector buildings in the Republic of Croatia, as well as three scenarios of simulated data from the literature. Evaluation measures included prediction performance metrics (adjusted coefficient of determination, *RMSE*, *MAE*, and *SMAPE*) and, in the case of simulated data, the difference between actual and estimated regression coefficients. The method was compared to regularization that does not include experts' domain knowledge and to regression trees. The results on real data indicate that incorporating experts' domain knowledge through a metaheuristic approach improved prediction performance compared to the mentioned methods, while incorporating domain knowledge through a deterministic approach improved prediction performance in certain cases. In the simulation experiment, introducing experts' domain knowledge improved some prediction performance measures over regularization without experts' knowledge only with the metaheuristic approach, while the proposed method of introducing experts' knowledge into regularization did not outperform the decision tree method in prediction for any approach of introducing experts' knowledge. The results indicate that different ways of estimating penalization coefficients and the model structure itself affect the variability and bias of the model. Additionally, they raise further questions about how to achieve improved prediction performance and reduced variability and / or bias by incorporating experts' domain knowledge in more complex data sets.

**Keywords:** feature selection, linear regression, regularization, experts' domain knowledge, metaheuristics, cuckoo search, regression trees, prediction, energy costs, public buildings, simulations

# EXTENDED SUMMARY

The thesis consists of the following chapters: Introduction, Theoretical Background, Previous Research, Methodology, Research Results, Discussion, and Conclusion, along with four appendices (Appendix 1 - Real Data, Appendix 2 - Scenario 1, Appendix 3 - Scenario 2, and Appendix 4 - Scenario 3).

Firstly, the Introduction (chapter 1) emphasizes the importance of dimensionality reduction, which is justified by the exponential growth of data and the increase in its complexity and dimensionality. Additionally, the importance of experts' domain knowledge as a valuable source of information for decision-making is highlighted and the following research plan is proposed:

Research question: How can the predictive performance of a linear regression model be improved using experts' domain knowledge?

Objective: To develop a feature selection method for a linear regression model that incorporates experts' domain knowledge about the impact of input features on the output into the regularization process.

Hypothesis 1: The proposed method results in better predictive performance than a regularization method that does not include experts' domain knowledge.

Hypothesis 2: The proposed method results in better predictive performance than the decision trees method.

In the following, the research context is outlined as this research is part of a broader project named "Methodological Framework for Efficient Energy Management by Intelligent Data Analytics" (acronym: MERIDA, code: HRZZ IP-2016-06-8350), conducted under the Croatian Science Foundation. Finally, the Introduction chapter ends with the structure of the thesis.

The Theoretical Background (chapter 2) describes the theoretical framework of this research. Firstly, the classification of dimensionality reduction methods into feature extraction methods and feature selection methods is made (section 2.1). As the focus of the thesis is on feature

selection methods due to their characteristic of preserving the interpretability of features, an overview and classification of feature selection methods is provided. They are usually divided into filter methods, wrapper methods, and embedded methods, with sometimes a category of hybrid methods introduced as a combination of the previous types.

Furthermore, the linear regression model is explained (section 2.2) and then the most common types of regularization (lasso, ridge, and elastic net regularization) are listed and compared (section 2.3). Since the proposed method aims to incorporate the domain knowledge of experts, it is important to verify their agreement and association, and therefore the measures used for this purpose are described in the same chapter (section 2.4). In the following part of the chapter (section 2.5), the concept of (meta)heuristics is explained, and the cuckoo search algorithm is described as the metaheuristic algorithm used in this study.

The classification and regression trees method is described in section 2.6 as a method to which the newly proposed method's performances are compared. The chapter concludes by listing the evaluation measures used (section 2.7).

The Previous Research (chapter 3) provides an overview of previous research on feature selection methods (section 3.1), methods that include experts' knowledge (section 3.2), and previous research in the context of the energy and construction sectors (section 3.3). The potential of this field is identified in hybrid methods, especially in conjunction with methods based on metaheuristic algorithms and methods that include experts' domain knowledge.

The research description and methodology are detailed in the Methodology (chapter 4). The methodology for incorporating experts' domain knowledge into regularization is described, outlining two different approaches: deterministic and metaheuristic. The data used for method evaluation is described, including a set of real data on energy costs of public sector buildings in Croatia and three scenarios of simulated data derived from the literature. The design of the simulation experiment, comparing the results of regularization methods (lasso, ridge and elastic net regularization) with and without experts' domain knowledge, as well as classification and regression trees, is presented. Data analysis methods for the simulation experiment, used to test the research hypotheses, are described.

The Research Results (chapter 5) presents the research results for each data set, starting with real data and then each of the three scenarios. Results include experts' opinion, estimates of the penalization coefficient obtained by the metaheuristic, and model coefficient estimates in terms of variability and bias. Finally, the results of regularization methods without and

with experts' knowledge, using both deterministic and metaheuristic approaches, as well as the CART (classification and regression trees) method, are evaluated. For simulated data, the results of testing the hypotheses using analysis of variance are presented. Some results are provided in the appendices for clarity.

The Discussion chapter (chapter 6) compares and synthesizes the analysis results from real data and the three simulation experiments. This includes the penalization coefficient estimates, model coefficient estimates in terms of variability and bias, and hypothesis evaluation results. The chapter concludes by comparing these results with previous research.

The study concludes with the Conclusion (chapter 7), which summarizes the main findings and provides guidelines for future research, along with scientific and social contributions.

Throughout this research, the question of improving the prediction accuracy of the linear regression model using experts' domain knowledge was explored. The objective was to develop a feature selection method for a linear regression model that incorporates experts' domain knowledge into the regularization process. Experts' domain knowledge was included by introducing a penalization coefficient whose value depended on the expert's opinion on the impact of the input feature on the output feature. Four categories of input feature impact on the output were considered: *No impact*, *Weak impact*, *Medium impact*, and *Strong impact*, resulting in four different penalization coefficient values. Two approaches for incorporating experts' knowledge were used — deterministic and metaheuristic, making the proposed method a hybrid feature selection method.

The proposed method was evaluated using real data on energy costs of public sector buildings in Croatia and their characteristics, as well as 50 simulated data sets from various literature scenarios, within two research hypotheses. Results from real data support both hypotheses for regularization methods using the metaheuristic approach to incorporate experts' knowledge and partially support methods using the deterministic approach. In the simulation experiment, Hypothesis 1 was confirmed for regularization methods using the metaheuristic approach for *RMSE* and *SMAPE* measures, while Hypothesis 2 was not confirmed for either approach to incorporating experts' knowledge. Results showed that incorporating domain knowledge has a greater contribution in simpler models with fewer predictors, lower correlations, and smaller errors, such as in Scenario 1. In more complex models, such as those from Scenarios 2 and 3, the contribution of domain knowledge becomes less significant due to increased complexity and model errors. Different methods for estimating penalization coefficients and the model structure

## Summary

---

impact the model's variability and bias, raising further questions about how to achieve reduced variability and / or bias in more complex data sets through experts' domain knowledge.

In addition to the aforementioned points, this research opens several potential directions for further investigation:

- 1) Quality of domain knowledge: Ensuring the quality of experts' domain knowledge used for model training is crucial. Poor-quality knowledge can negatively impact prediction accuracy, so it is necessary to ensure high-quality of domain knowledge. For real data, experts' domain knowledge was available and improved prediction performances when included in the model.
- 2) Alternative methods of incorporating domain knowledge: Exploring different methods for collecting and incorporating experts' domain knowledge into regularization methods and adapting them to the problem structure is important, as the study showed that problem structure significantly affects prediction accuracy.
- 3) Incorporating domain knowledge into other decision support methods: Applying experts' domain knowledge to other decision support methods increases the likelihood of finding an appropriate approach to introducing domain knowledge to a broader class of problems.
- 4) Changes in metaheuristic: Metaheuristics is a rapidly growing field with great potential. It is possible to try different optimization parameters within the same metaheuristic, experiment with other metaheuristics, or even combine them. Potential constraints on penalization coefficient size relationships could also be imposed to make them more intuitive for users.

The scientific contributions of the proposed research are multifold and can be summarized as follows:

- 1) Systematization of previous research on using experts' domain knowledge in decision support methods: Previous studies support the introduction of experts' domain knowledge and emphasize the importance of including it in the model training phase and ensuring the quality of domain knowledge.
- 2) Systematization of previous research on feature selection methods: Previous studies recognize the potential of hybrid feature selection methods, particularly those based on metaheuristic algorithms.

- 3) Proposing a feature selection method for a linear regression model that incorporates expert's domain knowledge into the regularization process: Two approaches for incorporating experts' knowledge were used, deterministic and metaheuristic, that make the proposed method a hybrid feature selection method.
- 4) Evaluating the feature selection method on real data from public sector building energy costs in Croatia: experts' domain knowledge was used to create a predictive model for energy costs in Croatian public sector buildings. Results partially support the research hypotheses for regularization methods using the deterministic approach to incorporating experts' knowledge, while they fully support methods using the metaheuristic approach.
- 5) Availability of expert's opinions as open scientific data: The results of experts' opinions on the impact of building characteristics on energy costs are available as open scientific data in the Zenodo repository.
- 6) Evaluation of the proposed method through a simulation experiment on three scenarios from the literature: The study tested the research hypotheses, showing that incorporating domain knowledge has a greater contribution in simpler models with fewer predictors, lower correlations, and smaller errors, as seen in Scenario 1. In more complex models, the contribution of domain knowledge becomes less significant due to increased complexity and errors. Therefore, Hypothesis 1 is partially confirmed. Hypothesis 2 is not confirmed, meaning the proposed method is not more accurate than the CART method.
- 7) Availability of simulated data and generation scripts as open scientific data: Simulated data and scripts for their generation are available as open scientific data in the Zenodo repository.

Contributions 4 and 5 also represent significant social contributions. Predictive models for energy costs and identifying relevant building features can help decision-makers in planning energy renovation.

# SADRŽAJ

<b>1</b>	<b>Uvod</b>	<b>1</b>
1.1	Istraživački kontekst . . . . .	2
1.2	Prijedlog istraživanja . . . . .	4
1.2.1	Istraživačko pitanje, cilj i hipoteze . . . . .	5
1.3	Struktura rada . . . . .	5
<b>2</b>	<b>Teorijska podloga</b>	<b>6</b>
2.1	Redukcija dimenzionalnosti . . . . .	6
2.1.1	Podjela metoda redukcije dimenzionalnosti . . . . .	7
2.2	Linearni regresijski model . . . . .	9
2.3	Regularizacija . . . . .	10
2.3.1	Usporedba metoda regularizacije . . . . .	12
2.4	Mjere slaganja i povezanosti . . . . .	13
2.5	(Meta)heuristike . . . . .	16
2.5.1	Algoritam pretraživanja kukavice . . . . .	19
2.6	Metoda CART . . . . .	21
2.7	Mjere evaluacije . . . . .	23
<b>3</b>	<b>Prethodna istraživanja</b>	<b>26</b>
3.1	Prethodna istraživanja u kontekstu metoda odabira značajki . . . . .	26
3.2	Prethodna istraživanja u kontekstu uključivanja stručnjaka . . . . .	28
3.3	Prethodna istraživanja u kontekstu energetike . . . . .	33
<b>4</b>	<b>Metodologija</b>	<b>37</b>
4.1	Uvođenje domenskog znanja stručnjaka u regularizaciju . . . . .	37
4.1.1	Deterministička procjena penalizacijskog koeficijenta . . . . .	38

4.1.2	Procjena penalizacijskog koeficijenta metaheurističkim algoritmom . . .	39
4.2	Opis stvarnih podataka . . . . .	40
4.2.1	Domensko znanje stručnjaka u slučaju stvarnih podataka . . . . .	43
4.3	Opis simuliranih podataka . . . . .	44
4.3.1	Domensko znanje stručnjaka u slučaju simuliranih podataka . . . . .	46
4.3.2	Procjena veličine učinka po scenarijima . . . . .	47
4.3.3	Podjela simuliranih podataka na javnom repozitoriju . . . . .	48
4.4	Procjena i evaluacija modela . . . . .	48
4.5	Evaluacija hipoteza . . . . .	50
4.6	Računalna implementacija modela . . . . .	54
<b>5</b>	<b>Rezultati istraživanja</b>	<b>56</b>
5.1	Stvarni podaci . . . . .	56
5.1.1	Deskriptivna analiza podataka . . . . .	56
5.1.2	Mišljenje stručnjaka . . . . .	59
5.1.3	Procjena penalizacijskih koeficijenata algoritmom pretraživanja kukavice	62
5.1.4	Evaluacija modela i hipoteza . . . . .	63
5.2	Scenarij 1 . . . . .	69
5.2.1	Modeliranje mišljenja stručnjaka . . . . .	69
5.2.2	Procjena penalizacijskih koeficijenta algoritmom pretraživanja kukavice	72
5.2.3	Procjena koeficijenata modela . . . . .	74
5.2.4	Evaluacija modela i hipoteza . . . . .	78
5.3	Scenarij 2 . . . . .	91
5.3.1	Modeliranje mišljenja stručnjaka . . . . .	91
5.3.2	Procjena penalizacijskih koeficijenata algoritmom pretraživanja kukavice	94
5.3.3	Procjene koeficijenata modela . . . . .	95
5.3.4	Evaluacije modela i hipoteza . . . . .	99
5.4	Scenarij 3 . . . . .	112
5.4.1	Modeliranje mišljenja stručnjaka . . . . .	112
5.4.2	Procjena penalizacijskih koeficijenata algoritmom pretraživanja kukavice	117
5.4.3	Procjena koeficijenata modela . . . . .	118
5.4.4	Evaluacija modela i hipoteza . . . . .	123



<b>6 Rasprava</b>	<b>137</b>
6.1 Procjena penalizacijskih koeficijenata algoritmom pretraživanja kukavice . . . . .	137
6.1.1 Stvarni podaci . . . . .	137
6.1.2 Scenarij 1 . . . . .	138
6.1.3 Scenarij 2 . . . . .	138
6.1.4 Scenarij 3 . . . . .	139
6.2 Procjena koeficijenata modela . . . . .	139
6.2.1 Scenarij 1 . . . . .	140
6.2.2 Scenarij 2 . . . . .	141
6.2.3 Scenarij 3 . . . . .	142
6.3 Hipoteza 1 . . . . .	145
6.3.1 Stvarni podaci . . . . .	145
6.3.2 Scenarij 1 . . . . .	148
6.3.3 Scenarij 2 . . . . .	150
6.3.4 Scenarij 3 . . . . .	152
6.4 Hipoteza 2 . . . . .	156
6.4.1 Stvarni podaci . . . . .	157
6.4.2 Scenariji . . . . .	158
6.5 Usporedba s literaturom . . . . .	160
<b>7 Zaključak</b>	<b>162</b>
7.1 Doprinos . . . . .	164
<b>Prilog 1 - stvarni podaci</b>	<b>166</b>
<b>Prilog 2 - Scenarij 1</b>	<b>174</b>
<b>Prilog 3 - Scenarij 2</b>	<b>182</b>
<b>Prilog 4 - Scenarij 3</b>	<b>192</b>
<b>Bibliografija</b>	<b>200</b>
<b>Životopis</b>	<b>215</b>

# POPIS TABLICA

2.1	Tumačenja mjera slaganja i povezanosti prema iznosu . . . . .	15
4.1	Vrijednosti penalizacijskog koeficijenta u slučaju determinističke procjene . . . .	39
4.2	Naziv, vrsta i opis značajki . . . . .	40
4.3	Opis iskustva stručnjaka . . . . .	43
4.4	Tumačenje veličine učinka prema apsolutnoj vrijednosti Pearsonovog koeficijenta korelacije $ \rho_{X,Y} $ . . . . .	46
4.5	Formiranje "mišljenja stručnjaka" o utjecaju pojedine ulazne značajke na izlaznu temeljem vrijednosti koeficijenta korelacije $ \rho_{X,Y} $ . . . . .	47
5.1	Deskriptivna analiza kvantitativnih značajki . . . . .	56
5.2	Deskriptivna analiza kvalitativnih značajki . . . . .	58
5.3	Tablica frekvencija pojedine kategorije značajki po stručnjacima . . . . .	59
5.4	Usporedba mjera slaganja i povezanosti stručnjaka kada se promatraju sva četiri stručnjaka zajedno (Slučaj 1) i kada se promatraju zajedno Stručnjak 1, Stručnjak 2 i Stručnjak 4 (Slučaj 2) . . . . .	60
5.5	Tablica frekvencija pojedine kategorije značajki u obje promatrane vrste domenskog znanja . . . . .	60
5.6	Tablica frekvencija pojedine kategorije značajki u obje promatrane vrste domenskog znanja za matrice dizajna . . . . .	61
5.7	Vrijednosti penalizacijskog koeficijenta nakon normalizacije u slučaju determinističke procjene i obje vrste domenskog znanja . . . . .	61
5.8	Procjene vrijednosti parametara dobivene optimizacijom prema pojedinoj vrsti regularizacije . . . . .	62
5.9	Evaluacija modela na stvarnim podacima . . . . .	64

5.10	Vrijeme izvođenja metoda regularizacije s metaheurističkom procjenom penalizacijskog koeficijenta . . . . .	65
5.11	Scenarij 1 - deskriptivna analiza procjene penalizacijskog koeficijenta ovisno o vrsti regularizacije (normalizirane vrijednosti) . . . . .	72
5.12	Scenarij 1 - deskriptivna analiza procjene koeficijenta modela ovisno o vrsti regularizacije i načinu procjene penalizacijskog koeficijenta . . . . .	74
5.13	Prosječno vrijeme izvođenja metoda regularizacije s metaheurističkom procjenom penalizacijskog koeficijenta za skupove iz Scenarija 1. . . . .	78
5.14	Scenarij 2 - deskriptivna analiza procjene penalizacijskog koeficijenta ovisno o vrsti regularizacije (normalizirane vrijednosti) . . . . .	94
5.15	Scenarij 2 - deskriptivna analiza procjene koeficijenta modela ovisno o vrsti regularizacije i načinu procjene penalizacijskog koeficijenta . . . . .	96
5.16	Prosječno vrijeme izvođenja metoda regularizacije s metaheurističkom procjenom penalizacijskog koeficijenta za skupove iz Scenarija 2. . . . .	99
5.17	Scenarij 3 - deskriptivna analiza procjene penalizacijskog koeficijenta ovisno o vrsti regularizacije (normalizirane vrijednosti) . . . . .	117
5.18	Scenarij 3 - deskriptivna analiza procjene koeficijenta modela ovisno o vrsti regularizacije i načinu procjene penalizacijskog koeficijenta . . . . .	119
5.19	Prosječno vrijeme izvođenja metoda regularizacije s metaheurističkom procjenom penalizacijskog koeficijenta za skupove iz Scenarija 3. . . . .	123
6.1	Analiza varijabilnosti i pristranosti procjena koeficijenta modela iz Scenarija 1 - poredak modela različitih načina procjene penalizacijskog koeficijenta od boljeg prema lošijem u smislu varijabilnosti i pristranosti . . . . .	140
6.2	Analiza varijabilnosti i pristranosti procjena koeficijenta modela iz Scenarija 2 - poredak modela različitih načina procjene penalizacijskog koeficijenta od boljeg prema lošijem u smislu varijabilnosti i pristranosti . . . . .	141
6.3	Analiza varijabilnosti i pristranosti procjena koeficijenta modela iz Scenarija 3 - poredak modela različitih načina procjene penalizacijskog koeficijenta od boljeg prema lošijem u smislu varijabilnosti i pristranosti . . . . .	143
6.4	Sinteza rezultata analize Hipoteze 1 za svaku mjeru evaluacije, metodu regularizacije i način procjene penalizacijskih koeficijenta na skupu stvarnih podataka	146

6.5	Sinteza rezultata analize Hipoteze 1 za svaku mjeru evaluacije i način procjene penalizacijskih koeficijenata za svaki scenarij . . . . .	155
6.6	Sinteza rezultata analize Hipoteze 2 za svaku mjeru evaluacije, metodu regulacije i način procjene penalizacijskih koeficijenata na skupu stvarnih podataka	158
6.7	Sinteza rezultata analize Hipoteze 2 za svaku mjeru evaluacije i način procjene penalizacijskih koeficijenata za svaki scenarij . . . . .	159
7.1	Rezultati mišljenja stručnjaka o utjecaju svake ulazne značajke na izlaznu . . .	172
7.2	Utjecaj ulazne značajke u obje promatrane vrste domenskog znanja . . . . .	173
7.3	Scenarij 1, Hipoteza 1 - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . .	175
7.4	Scenarij 1, Hipoteza 1, udaljenost koeficijenata - <i>post hoc</i> analiza Tukey HSD testom . . . . .	176
7.5	Scenarij 1, Hipoteza 2a - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . .	179
7.6	Scenarij 1, Hipoteza 2b - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . .	181
7.7	Scenarij 2, Hipoteza 1 - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . .	183
7.8	Scenarij 2, Hipoteza 1, udaljenost koeficijenata - <i>post hoc</i> analiza Tukey HSD testom . . . . .	184
7.9	Scenarij 2, Hipoteza 1, korigirani koeficijent determinacije - <i>post hoc</i> analiza Tukey HSD testom . . . . .	186
7.10	Scenarij 2, Hipoteza 2a - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . .	189
7.11	Scenarij 2, Hipoteza 2b - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . .	191
7.12	Scenarij 3, Hipoteza 1 - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . .	193
7.13	Scenarij 3, Hipoteza 1, korigirani koeficijent determinacije - <i>post hoc</i> analiza Tukey HSD testom . . . . .	194
7.14	Scenarij 3, Hipoteza 2a - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . .	197

7.15 Scenarij 3, Hipoteza 2b - procjene koeficijenata i intervali pouzdanosti za svaku mjeru evaluacije . . . . . 199

# POPIS SLIKA

4.1	Podjela uzorka u slučaju stvarnih podataka . . . . .	48
4.2	Podjela uzorka u slučaju simuliranih podataka . . . . .	49
5.1	Dijagram procjene penalizacijskih koeficijenata (normalizirane vrijednosti) . .	63
5.2	Stvarni podaci - korigirani koeficijent determinacije prema vrsti domenskog znanja, procjeni penalizacijskog koeficijenta i metodi . . . . .	65
5.3	Stvarni podaci - <i>RMSE</i> prema vrsti domenskog znanja, procjeni penalizacijskog koeficijenta i metodi . . . . .	66
5.4	Stvarni podaci - <i>MAE</i> prema vrsti domenskog znanja, procjeni penalizacijskog koeficijenta i metodi . . . . .	67
5.5	Stvarni podaci - <i>SMAPE</i> prema vrsti domenskog znanja, procjeni penalizacijskog koeficijenta i metodi . . . . .	68
5.6	Kutijasti dijagram distribucija procjene penalizacijskih koeficijenta za skupove iz scenarija 1. . . . .	73
5.7	Violinski dijagram distribucija procjene penalizacijskih koeficijenta s profilima procjena za skupove iz scenarija 1. . . . .	73
5.8	Grafički prikazi distribucija procjene koeficijenta $\beta_0$ za skupove iz Scenarija 1.	76
5.9	Grafički prikazi distribucija procjene koeficijenta $\beta_1$ za skupove iz Scenarija 1.	76
5.10	Grafički prikazi distribucija procjene koeficijenta $\beta_2$ za skupove iz Scenarija 1.	77
5.11	Grafički prikazi distribucija procjene koeficijenta $\beta_5$ za skupove iz Scenarija 1.	77
5.12	Scenarij 1, Hipoteza 1 - Udaljenost koeficijenata prema procjeni penalizacijskog koeficijenta i metodi . . . . .	79
5.13	Scenarij 1, Hipoteza 1 - Korigirani koeficijent determinacije prema procjeni penalizacijskog koeficijenta i metodi . . . . .	80

5.14	Scenarij 1, Hipoteza 1 - <i>RMSE</i> prema procjeni penalizacijskog koeficijenta i metodi . . . . .	81
5.15	Scenarij 1, Hipoteza 1 - <i>MAE</i> prema procjeni penalizacijskog koeficijenta i metodi	83
5.16	Scenarij 1, Hipoteza 1 - <i>SMAPE</i> prema procjeni penalizacijskog koeficijenta i metodi . . . . .	84
5.17	Scenarij 1, Hipoteza 2a - <i>RMSE</i> prema metodi. . . . .	86
5.18	Scenarij 1, Hipoteza 2a - <i>MAE</i> prema metodi. . . . .	87
5.19	Scenarij 1, Hipoteza 2a - <i>SMAPE</i> prema metodi. . . . .	88
5.20	Scenarij 1, Hipoteza 2b - <i>RMSE</i> prema metodi. . . . .	89
5.21	Scenarij 1, Hipoteza 2b - <i>MAE</i> prema metodi. . . . .	90
5.22	Scenarij 1, Hipoteza 2b - <i>SMAPE</i> prema metodi. . . . .	91
5.23	Kutijasti dijagram distribucija procjene penalizacijskih koeficijenta za skupove iz scenarija 2. . . . .	95
5.24	Violinski dijagram distribucija procjene penalizacijskih koeficijenta s profilima procjena za skupove iz scenarija 2. . . . .	95
5.25	Grafički prikazi distribucija procjene koeficijenta $\beta_0$ za skupove iz Scenarija 2.	97
5.26	Grafički prikazi distribucija procjene koeficijenta $\beta_{11}$ za skupove iz Scenarija 2.	98
5.27	Grafički prikazi distribucija procjene koeficijenta $\beta_{31}$ za skupove iz Scenarija 2.	98
5.28	Scenarij 2, Hipoteza 1 - Udaljenost koeficijenata prema procjeni penalizacijskog koeficijenta i metodi . . . . .	100
5.29	Scenarij 2, Hipoteza 1 - Korigirani koeficijent determinacije prema procjeni penalizacijskog koeficijenta i metodi . . . . .	102
5.30	Scenarij 2, Hipoteza 1 - <i>RMSE</i> prema procjeni penalizacijskog koeficijenta i metodi . . . . .	103
5.31	Scenarij 2, Hipoteza 1 - <i>MAE</i> prema procjeni penalizacijskog koeficijenta i metodi	104
5.32	Scenarij 2, Hipoteza 1 - <i>SMAPE</i> prema procjeni penalizacijskog koeficijenta i metodi . . . . .	105
5.33	Scenarij 2, Hipoteza 2a - <i>RMSE</i> prema metodi. . . . .	107
5.34	Scenarij 2, Hipoteza 2a - <i>MAE</i> prema metodi. . . . .	108
5.35	Scenarij 2, Hipoteza 2a - <i>SMAPE</i> prema metodi. . . . .	109
5.36	Scenarij 2, Hipoteza 2b - <i>RMSE</i> prema metodi. . . . .	110
5.37	Scenarij 2, Hipoteza 2b - <i>MAE</i> prema metodi. . . . .	111

5.38	Scenarij 2, Hipoteza 2b - <i>SMAPE</i> prema metodi. . . . .	112
5.39	Kutijasti dijagram distribucija procjene penalizacijskih koeficijenta za skupove iz scenarija 3. . . . .	118
5.40	Violinski dijagram distribucija procjene penalizacijskih koeficijenta s profilima procjena za skupove iz scenarija 3. . . . .	118
5.41	Grafički prikazi distribucija procjene koeficijenta $\beta_0$ za skupove iz Scenarija 3. . . . .	121
5.42	Grafički prikazi distribucija procjene koeficijenta $\beta_1$ za skupove iz Scenarija 3. . . . .	121
5.43	Grafički prikazi distribucija procjene koeficijenta $\beta_6$ za skupove iz Scenarija 3. . . . .	122
5.44	Grafički prikazi distribucija procjene koeficijenta $\beta_{11}$ za skupove iz Scenarija 3. . . . .	122
5.45	Scenarij 3, Hipoteza 1 - Udaljenost koeficijenata prema procjeni penalizacijskog koeficijenta i metodi. . . . .	124
5.46	Scenarij 3, Hipoteza 1 - Korigirani koeficijent determinacije prema procjeni penalizacijskog koeficijenta i metodi . . . . .	125
5.47	Scenarij 3, Hipoteza 1 - <i>RMSE</i> prema procjeni penalizacijskog koeficijenta i metodi . . . . .	127
5.48	Scenarij 3, Hipoteza 1 - <i>MAE</i> prema procjeni penalizacijskog koeficijenta i metodi	128
5.49	Scenarij 3, Hipoteza 1 - <i>SMAPE</i> prema procjeni penalizacijskog koeficijenta i metodi . . . . .	129
5.50	Scenarij 3, Hipoteza 2a - <i>RMSE</i> prema metodi. . . . .	131
5.51	Scenarij 3, Hipoteza 2a - <i>MAE</i> prema metodi. . . . .	132
5.52	Scenarij 3, Hipoteza 2a - <i>SMAPE</i> prema metodi. . . . .	133
5.53	Scenarij 3, Hipoteza 2 b - <i>RMSE</i> prema metodi. . . . .	134
5.54	Scenarij 3, Hipoteza 2b - <i>MAE</i> prema metodi. . . . .	135
5.55	Scenarij 3, Hipoteza 2b - <i>SMAPE</i> prema metodi. . . . .	135
7.1	Grafički prikaz distribucija značajki [1/4] . . . . .	166
7.2	Grafički prikaz distribucija značajki [2/4] . . . . .	167
7.3	Grafički prikaz distribucija značajki [3/4] . . . . .	168
7.4	Grafički prikaz distribucija značajki [4/4] . . . . .	169
7.5	Dijagrami rasipanja i korelacije kvantitativnih nezavisnih značajki i zavisne značajke [1/2] . . . . .	170
7.6	Dijagrami rasipanja i korelacije kvantitativnih nezavisnih značajki i zavisne značajke [2/2] . . . . .	171



---

7.7	Scenarij 1, Hipoteza 1 - dijagnostički dijagrami za svaku mjeru evaluacije . . .	174
7.8	Scenarij 1, Hipoteza 2a - dijagnostički dijagrami za svaku mjeru evaluacije . . .	178
7.9	Scenarij 1, Hipoteza 2b - dijagnostički dijagrami za svaku mjeru evaluacije . . .	180
7.10	Scenarij 2, Hipoteza 1 - dijagnostički dijagrami za svaku mjeru evaluacije . . .	182
7.11	Scenarij 2, Hipoteza 2a - dijagnostički dijagrami za svaku mjeru evaluacije . . .	188
7.12	Scenarij 2, Hipoteza 2b - dijagnostički dijagrami za svaku mjeru evaluacije . . .	190
7.13	Scenarij 3, Hipoteza 1 - dijagnostički dijagrami za svaku mjeru evaluacije . . .	192
7.14	Scenarij 3, Hipoteza 2a - dijagnostički dijagrami za svaku mjeru evaluacije . . .	196
7.15	Scenarij 3, Hipoteza 2b - dijagnostički dijagrami za svaku mjeru evaluacije . . .	198

# POPIS KRATICA

## **Kratika**   **Pojašnjenje**

ANOVA	analiza varijance (eng. <i>analysis of variance</i> )
AUC	površina ispod krivulje (eng. <i>area under the curve</i> )
CART	klasifikacijska i regresijska stabla (eng. <i>classification and regression trees</i> )
CERN	Europsko vijeće za nuklearna istraživanja (fran. <i>Conseil européen pour la recherche nucléaire</i> )
CS	pretraživanje kukavice (eng. <i>cuckoo search</i> )
DGNB	Njemačko društvo za održivu gradnju (njem. <i>Deutschen Gesellschaft für Nachhaltiges Bauen</i> )
EU	Europska unija
FA	faktorska analiza
HRZZ	Hrvatska zaklada za znanost
HPC	računarstvo visokih performansi (eng. <i>high-performance computing</i> )
ICA	analiza nezavisnih komponenti (eng. <i>independent components analysis</i> )
ICC	intraklasni koeficijent korelacije (eng. <i>intraclass correlation coefficient</i> )
lasso	operator najmanjeg apsolutnog skupljanja i odabira (eng. <i>least absolute shrinkage and selection operator</i> )
MAE	srednja apsolutna greška (eng. <i>mean absolute error</i> )
MSE	srednje kvadratna greška (eng. <i>mean square error</i> )
PCA	analiza glavnih komponenti (eng. <i>principal components analysis</i> )
RFE	rekurzivna eliminacija značajki (eng. <i>recursive feature elimination</i> )
RMSE	korijen srednje kvadratne greške (eng. <i>root mean square error</i> )
ROC	radna karakteristika prijemnika (eng. <i>receiver operating characteristic</i> )
SMAPE	simetrična srednja apsolutna postotna greška (eng. <i>symmetric mean absolute percentage error</i> )
SVM	metoda potpornih vektora (eng. <i>support vector machines</i> )
ZB	zetabajt (eng. <i>zettabyte</i> )

# 1. UVOD

Živimo u vremenu kada se, u gotovo svim domenama ljudskog djelovanja, stvaraju ogromne količine podataka razmjerima koji bilježe eksponencijalni rast [1]. U tom kontekstu, popularan je engleski izraz "*Big Data*" (hrv. veliki podaci) koji se koristi za karakterizaciju spomenute eksplozije podataka [1].

Primjerice, Statista [2], globalna platforma za podatke i poslovnu inteligenciju s opsežnom zbirkom statistika, izvješća i uvida, na temelju podataka prikupljenih u periodu od 2010. do 2020. godine, procjenjuje da će u 2024. godini biti generiran volumen podataka od 147 zeta-bajtova (eng. *zettabytes* - ZB), a u 2025. 181 ZB, pri čemu zetabajt predstavlja  $10^{21}$  bajtova. Usporedbe radi, 10 godina ranije, 2014. godine, generirano je 12.5 ZB podataka [2].

Brzi porast količine podataka predstavlja izazov i postavlja važan imperativ – učinkovito upravljanje podacima. Naime, vrlo je važno istaknuti da prikupljeni podaci ostvaruju potencijal tek otkrivanjem znanja iz njih. U tu svrhu, danas se primjenjuju različite metode rudarenja podataka i strojnog učenja, koje se moraju nositi sa skupovima podataka čiji se volumen i složenost značajno povećavaju [3,4].

Osim toga, povećava se i dimenzionalnost skupova podataka. Podaci visoke dimenzionalnosti mogu značajno povećati zahtjeve za pohranom memorije i računalne troškove za analizu podataka [3,4]. Visoka dimenzionalnost potencijalno donosi poteškoće u smislu povećane mogućnosti prenaučivosti, povećanja multikolinearnosti, porasta složenosti problema i kapaciteta potrebnih za rad s takvim skupovima podataka i druge [3,5,6]. Upravo zbog svega navedenog, redukcija dimenzionalnosti je iznimno važna.

I dok je nesporno da metode redukcije dimenzionalnosti imaju značajnu ulogu u nošenju s velikom količinom podataka, ne smije se propustiti istaknuti i uloga čovjeka i njegova znanja u istom kontekstu. Ljudski faktor je iznimno bitan iz više razloga, primjerice u smislu razumijevanja podataka i davanja konteksta istima, osiguravanju kvalitete podataka te njihovu korištenju u svrhu donošenja odluka.

Važnost znanja stručnjaka o domeni problema prepoznali su brojni autori koji u svojim istraživanjima nastoje uključiti navedeno znanje u metode za obradu i modeliranje podataka i sustave za donošenje odluka. U tom smislu, ovo istraživanje ima za cilj razviti metodu za odabir značajki koja u metodu regularizacije uključuje domensko znanje stručnjaka. Detaljnije o tome je navedeno u poglavlju *Metodologija*, dok je u nastavku pojašnjen kontekst istraživanja.

U nastavku uvodnog poglavlja prikazani su neki rezultati dosadašnjih istraživanja koji su poslužili kao motivacija ovom istraživanju te su navedeni ciljevi, hipoteze i istraživačko pitanje. Naposljetku je prikazana struktura rada.

## 1.1. ISTRAŽIVAČKI KONTEKST

Istraživanje je dio istraživačkog projekta Hrvatske zaklade za znanost pod nazivom „Metodološki okvir za učinkovito upravljanje energijom s pomoću inteligentne podatkovne analitike“ (akronim: MERIDA, šifra projekta: HRZZ IP-2016-06-8350) koji je trajao u periodu od 1.1.2017. do 28.2.2021. [7].

Pretjerana potrošnja energije, odnosno "... svih oblika energenata, goriva, toplinske energije, obnovljive energije, električne energije ili bilo kojeg drugog oblika energije..." [8] veliki je problem danas, koji ne samo što uzrokuje velike financijske troškove, nego i onečišćuje okoliš te doprinosi klimatskim promjenama i globalnom zatopljenju [9] i otežava uspostavljanje održivog razvoja.

Europska unija (EU) pridaje veliku važnost energetici, što je vidljivo iz sljedećih nekoliko direktiva. Godine 2012., prema Direktivi o energetske učinkovitosti 2012/27/EU [8], EU je postavila cilj povećanja energetske učinkovitosti i postizanja uštede od 20 % primarne potrošnje energije u Uniji do 2020. u usporedbi s projekcijama iz 2007., dok je u Direktivi 2018/2002 [10] postavila cilj poboljšanja energetske učinkovitosti do 2030. godine smanjenjem potrošnje energije za najmanje 32.5 %.

Nadalje, sektor zgradarstva je s okvirno 40 % ukupne potrošnje energije u EU jedan od najvećih pojedinačnih potrošača energije [11]. Naime, poznato je da otprilike 50 % potrošene konačne energije u EU odlazi na grijanje i hlađenje, od čega 80 % u zgradama, te je stoga postizanje energetske i klimatske ciljeve EU povezano s obnavljanjem fonda zgrada s naglaskom na energetske učinkovitost [12]. S tim u vezi, unutar Direktive 2010/31/EU, postavljen je cilj da sve nove zgrade javnog sektora izgrađene nakon 2018. i stambene zgrade izgrađene nakon

2020. budu zgrade gotovo nulte energije (eng. *nearly zero-energy buildings* - NZEB), odnosno zgrade koje imaju vrlo visoku energetska učinkovitost, a ta približno nulta količina energije trebala bi se u vrlo značajnoj mjeri pokrivati energijom iz obnovljivih izvora koja se proizvodi u krugu ili u blizini zgrade [13]. Štoviše, u Direktivi 2018/844 EU je postavila cilj osigurati obnovu postojećih zgrada u zgrade gotovo nulte energije [12].

Budući da se, prema Direktivi 2012/27/EU [8], sve države članice EU obvezuju od 1.siječnja 2014. godine svake godine obnoviti 3 % ukupne površine poda grijanih i / ili hlađenih zgrada u vlasništvu i uporabi središnje vlasti, može se reći da su od 2014. godine krenule energetske obnove zgrada javnog sektora, čiji su cilj energetske i financijske uštede, ali i bolja kvaliteta korištenja prostora [14].

Ove direktive su istaknute zbog svoje aktualnosti u vrijeme trajanja spomenutog HRZZ istraživačkog projekta, čija je svrha, između ostalog, bila pružiti metodologiju koja bi pomogla u ostvarivanju ciljeva navedenih direktiva EU.

Nastavno na to proizlazi i motivacija za stvaranjem modela za predviđanje energetske troškova zgrada javnog sektora u ovisnosti o različitim karakteristikama zgrade - primjerice građevinskim, energetske i drugim, a koji su jedan od rezultata ovog istraživanja. Spomenuti modeli bi trebali pomoći donositeljima odluke u planiranju mjera energetske obnove zgrada javne namjene. Uzorak koji se koristi u tu svrhu sadrži podatke o 26 značajki za 1421 zgradu javnog sektora Republike Hrvatske te je na njemu potrebno provesti redukciju dimenzionalnosti. Podaci su detaljnije opisani u potpoglavlju 4.2 (Opis stvarnih podataka, str. 40).

Važno je istaknuti jednu veliku prednost ovog istraživanja - mogućnost prikupljanja domenskog znanja stručnjaka. Naime, projektni tim, sastavljen od istraživača iz interdisciplinarnih područja (ekonomija, informacijske znanosti, matematika, građevina i strojarstvo), kao člana je imao i osobu koja se u svojim nastavnim i stručnim aktivnostima bavi građevinarstvom, zgradarstvom i energetske certificiranjem, zbog čega se može smatrati domenskim stručnjakom u ovom kontekstu. Dodatno, preko te osobe su identificirani i ispitani i drugi stručnjaci iz iste domene koji predstavljaju vrijedan izvor informacija i znanja (detaljnije u potpoglavlju 4.2.1 (Domensko znanje stručnjaka u slučaju stvarnih podataka, str. 43)).

## 1.2. PRIJEDLOG ISTRAŽIVANJA

Potencijal uključivanja domenskih stručnjaka u metode za potporu odlučivanju prepoznali su brojni autori [15–36], pritom naglašavajući važnost domenskog znanja stručnjaka kao vrlo vrijednog izvora informacija. Posebno je to značajno u područjima, poput npr. medicine, u kojima se ne raspolaže s puno slučajeva i / ili je skupocjeno doći do novih, kao i u onima u kojima se donose odluke iznimne važnosti.

S druge strane, neki autori osporavaju korištenje isključivo domenskog znanja stručnjaka u sustavima za potporu odlučivanju obrazlažući to najčešće sljedećim argumentima [37–39]: (1) dodatnim troškom u smislu radne snage, (2) nemogućnošću automatizacije procesa, (3) nemogućnošću generalizacije znanja na druge domene i (4) velikom ovisnošću o točnosti znanja stručnjaka. Osim toga, navode da ljudske teorije često sadrže pogreške, a da je zadaća sustava strojnog učenja usavršavati teorije [37].

Slijedom svega navedenog, ovo istraživanje pronalazi potencijal upravo u kombiniranju oba pristupa. Dakle, umjesto korištenja isključivo domenskog znanja stručnjaka ili isključivo metode za potporu odlučivanju, bitna je njihova integracija. Pretpostavka je da će uključivanje znanja stručnjaka o domeni problema u metodu, rezultirati točnijim i, krajnjim korisnicima, pouzdanijim modelima.

Nadalje, može se uočiti da ukoliko se koristi domensko znanje stručnjaka, poneki autori zamjeraju uključivanje istoga samo u fazi predobrade, odnosno pri odabiru značajki koji prethodi izgradnji modela jer se na taj način model trenira s fiksnim podskupom značajki koji nije nužno optimalan u prediktivnom smislu [24]. Umjesto toga predlažu da se istovremeno razmatraju uvođenje domenskog znanja stručnjaka i učenje metode. Ukoliko se smjer učenja metode na neki način odredi domenskim znanjem stručnjaka, primjerice kao [24], očekuje se da će konačni model biti pouzdan i interpretabilan krajnjim korisnicima, ali i da će sama metoda, usmjerena znanjem stručnjaka, ispraviti ili poboljšati domensko znanje, čime se dobiva najbolje od oba pristupa. Brojna prethodna istraživanja također idu u prilog korištenju hibridnih metoda, odnosno kombinacija različitih metoda.

Konačno, upravo iz tog razloga, u ovom se istraživanju predlaže pristup koji uvrštava domensko znanje stručnjaka o utjecaju značajki u metode regularizacije. Pojašnjenje zašto su upravo metode regularizacije u fokusu ovog istraživanja dano je u potpoglavlju 2.1 (Redukcija dimenzionalnosti, str. 6).

### 1.2.1. Istraživačko pitanje, cilj i hipoteze

Nastavno na prethodno identificiran istraživački potencijal, ovo istraživanje ima sljedeći plan:

**Istraživačko pitanje:** Kako poboljšati uspješnost predikcije linearnog regresijskog modela korištenjem domenskog znanja stručnjaka?

**Cilj:** Razviti metodu za odabir značajki u linearnom regresijskom modelu koja u metodu regularizacije uključuje domensko znanje stručnjaka o utjecaju ulaznih značajki na izlaznu.

**Hipoteza 1:** Predložena metoda rezultira uspješnijom predikcijom od metode regularizacije koja ne uključuje domensko znanje stručnjaka.

**Hipoteza 2:** Predložena metoda rezultira uspješnijom predikcijom od metode regresijskih stabala.

## 1.3. STRUKTURA RADA

Rad se sastoji od sljedećih poglavlja: [Uvod](#), [Teorijska podloga](#), zatim [Prethodna istraživanja](#), [Metodologija](#), [Rezultati istraživanja](#), [Rasprava](#) i [Zaključak](#) te četiri priloga.

U radu je najprije u poglavlju [Teorijska podloga](#) ukratko pojašnjena važnost i podjela metoda za redukciju dimenzionalnosti, a zatim i metoda odabira značajki. Nadalje, pojašnjen je linearni regresijski model, a potom su navedene i uspoređene najčešće vrste regularizacije (lasso, hrbatna i regularizacija elastičnom mrežom). Osim toga, opisane su mjere koje će se koristiti u svrhu provjere slaganja i povezanosti mišljenja stručnjaka. U nastavku je pojašnjen koncept (meta)heuristike te je dodatno opisan algoritam pretraživanja kukavice kao metaheuristički algoritam koji će se koristiti u ovom istraživanju. Dodatno, opisana je metoda klasifikacijskih i regresijskih stabala kao metoda s kojom će se usporediti rad novopredložene metode, a poglavlje je zaključeno navođenjem mjera evaluacije koje će se koristiti za usporedbu.

Poglavlje [Prethodna istraživanja](#) nosi pregled dosadašnjih istraživanja koja uključuju znanja stručnjaka, kao i dosadašnjih istraživanja u kontekstu energije i građevinarstva.

Uvođenje domenskog znanja stručnjaka u regularizaciju je opisano u poglavlju [Metodologija](#), kao i podaci (stvarni i simulirani) na kojima će se provesti evaluacija predložene metode.

Nadalje, rezultati su predstavljeni i raspravljani u poglavljima [Rezultati istraživanja](#) i [Rasprava](#). Konačno, u poglavlju [Zaključak](#) su dane implikacije i zaključak ovog istraživanja te su navedeni znanstveni i društveni doprinosi, dok su u priložima navedeni rezultati istraživanja koji su, radi preglednosti rada, izostavljeni iz poglavlja [Rezultati istraživanja](#).

## 2. TEORIJSKA PODLOGA

### 2.1. REDUKCIJA DIMENZIONALNOSTI

Veličina skupa podataka može se mjeriti u dvije dimenzije, broju slučajeva ( $n$ ), koji se odnosi na volumen skupa, i broju značajki ( $p$ ), koji se odnosi na dimenzionalnost skupa [4]. Općenito, i broj slučajeva i broj značajki mogu biti iznimno veliki, a kaže se da je skup visokodimenzionalan (eng. *high-dimensional*) u slučaju kada je broj značajki veći od broja slučajeva ( $p > n$ ) [4]. Unatoč tome, treba napomenuti da redukcija dimenzionalnosti ima značajnu ulogu i u slučajevima kada skup ima veliki broj značajki, iako nije nužno visokodimenzionalan [3]. U nastavku potpoglavlja objašnjena je važnost redukcije dimenzionalnosti i navedene su metode koje se koriste u tu svrhu.

U literaturi se uobičajeno spominju dva pojma vezana uz dimenzionalnost; "blagoslovi i prokletstva dimenzionalnosti" (eng.  *blessings and curses of dimensionality*), koji se smatraju "dvjema stranama istog novčića" [40, 41]. "Blagoslovi dimenzionalnosti" temelje se na fenomenu mjere "koji sugerira da se izjave o vrlo visokodimenzionalnim postavkama mogu dati tamo gdje bi umjerene dimenzije bile previše komplicirane". Naime, Gorban i sur. [40, 42] tvrde da ako je skup podataka visokodimenzionalan, tada se određeni problemi iznenađujuće lakše rješavaju jednostavnim i robusnim metodama.

S druge strane, pojam "prokletstvo dimenzionalnosti" odnosi se na mnoge probleme koji postaju eksponencijalno teški u visokim dimenzijama [42]. Prije svega, u visokodimenzionalnom prostoru podaci postaju rjeđi, što nepovoljno utječe na algoritme dizajnirane za niskodimenzionalni prostor. Složenost funkcija više značajki raste eksponencijalno s dimenzijom, stoga, kako bi bilo moguće procijeniti takve funkcije s točnošću istom kao kod funkcija u nižim dimenzijama, treba eksponencijalno rasti i broj slučajeva u skupu za treniranje [5]. Također, s velikim brojem značajki, modeli imaju tendenciju prenaučivosti (eng. *over-fitting*). Radi se o prekomjernom prilagođavanju modela podacima na način da su rezultati za uzorak prividno



dobri, ali ne daju mogućnost generalizacije, što može uzrokovati slabije performanse na novim podacima na kojima se model nije učio [43].

S obzirom na sve ovo, ne začuđuje tvrdnja Haira i sur. [43] kako uspjeh bilo koje multivarijatne tehnike ovisi o odluci koje će se značajke koristiti kao ulazne (nezavisne) i izlazna (zavisna). Naime, navode tri stvari koje bi istraživači pritom trebali imati na umu: jaku teoriju, pogrešku mjerenja i pogrešku specifikacije. Jaka teorija podrazumijeva da odabir ulaznih i izlaznih značajki treba imati konceptualne ili teorijske temelje. Pogreška mjerenja odnosi se na stupanj do kojeg je značajka točna i dosljedna mjera koncepta koji se proučava te tvrde da, ako izlazna značajka ima značajnu pogrešku mjerenja, tada čak ni najbolje ulazne značajke neće postići prihvatljivu razinu prediktivne točnosti. Ipak, pogreška specifikacije smatra se najproblematičnijom, a radi se o uključivanju irelevantnih ili izostavljanju relevantnih ulaznih značajki. Uključivanje irelevantne značajke ujedno može povećati multikolinearnost. Multikolinearnost se odnosi na opseg u kojem se značajka može objasniti drugim značajkama u analizi; ako je prisutna u većoj mjeri, otežava interpretaciju modela jer je teško utvrditi učinak pojedine značajke, pa utječe na procjenu regresijskih koeficijenata i njihovu statističku značajnost. Osim toga, utječe na prediktivne performanse regresijskog modela; povećanjem multikolinearnosti ukupna objašnjena varijanca se smanjuje, zbog čega multikolinearnost može učiniti testiranje statističke značajnosti ulaznih značajki manje preciznim i smanjiti statistički i praktični značaj analize [43].

Bordens i Abbott te Field [44, 45] također tvrde da pri izgradnji modela treba težiti parsimoniji, odnosno tvrde da su jednostavniji modeli poželjniji u odnosu na one kompliciranije, uz pretpostavku da se time ne narušava točnost u velikoj mjeri. S tim u vezi, u model se trebaju uključiti samo prediktori koji imaju veliku moć objašnjavanja promatrane pojave [45].

Može se zaključiti da je redukcija dimenzionalnosti značajan korak u predobradi podataka i otkrivanju znanja iz istih s brojnim dobrobitima poput poboljšanja performansi modela, smanjenja multikolinearnosti, smanjenja mogućnosti prenaučivosti, olakšavanja vizualizacije i razumijevanja podataka, skraćivanja vremena provođenja algoritma, kao i smanjenje kapaciteta potrebnih za njihovo izvršavanje [3, 5, 6] i druge.

### 2.1.1. Podjela metoda redukcije dimenzionalnosti

Metode redukcije dimenzionalnosti obično se dijele na metode ekstrakcije (eng. *feature extraction*) i metode odabira značajki (eng. *feature selection*). Obje vrste metoda usmjerene su na

poboljšanje izvedbe metoda strojnog učenja korištenjem jednostavnijih modela, obično povećanjem brzine učenja [4].

Metode ekstrakcije transformiraju početni prostor značajki u prostor manjih dimenzija i tako rezultiraju novim značajkama koje su funkcija originalnih značajki, čime se nastoje dobiti nove informacije.

S druge strane, metode odabira značajki, kao što samo ime kaže, iz početnog skupa značajki odabiru odgovarajući podskup značajki. Njihov cilj jest otkriti relevantne značajke, a odbaciti nebitne i suvišne (redundantne) [4]. Uklanjanje takvih značajki smanjuje troškove pohrane i računanja, pri čemu ne uzrokuje gubitak informacija ili pogoršanje performansi [3]. Prednost ovih metoda u odnosu na metode ekstrakcije značajki je što se kod metoda odabira značajki zadržava početna interpretabilnost značajki [3]. To je bitno u kontekstima u kojima je važno održati reprezentativnost problema, a osim toga omogućuje poboljšanje performansi i dobro razumijevanje modela koji se želi dobiti te smanjenje troškova pohrane [4]. Ova vrsta metoda bit će u daljnjem fokusu ovog istraživanja pa je u nastavku rečeno malo više o njoj.

Metode odabira značajki se prema strukturalnom obilježju dijele na metode filtriranja (eng. *filter methods*), metode omotača (eng. *wrapper methods*) i ugrađene metode (eng. *embedded methods*), a ponekad se u literaturi navode dodatno i hibridne metode (eng. *hybrid methods*) koje se smatraju kombinacijom prethodno spomenutih metoda za odabir značajki.

Metode filtriranja odabiru podskup značajki neovisno o nekom algoritmu za učenje, oslanjajući se na karakteristike podataka kako bi odredile važnost značajki temeljem nekog kriterija [4]. Prednost metoda filtriranja je što su obično računalno efikasne, no nedostatak je što odabrani podskup značajki ne mora biti optimalan u odnosu na algoritam za učenje koji se obično kasnije koristi [3].

S druge strane, metode omotača, uzimajući u obzir algoritam za učenje, rješavaju nedostatak metoda filtriranja. Konkretno, ovakve metode odabiru podskup značajki, a potom ga evaluiraju u odnosu na performanse koje algoritam za učenje, izgrađen na tim značajkama, postiže. Taj se postupak provodi dok se ne dođe do nekog zaustavnog kriterija, čime se nastoji pronaći podskup optimalan za taj algoritam [3].

Nadalje, ugrađene metode uključuju odabir značajki u algoritam za učenje. Na taj način uzimaju u obzir vezu značajki i algoritma za učenje, no efikasnije su jer ne evaluiraju svaki podskup značajki pa tako predstavljaju kompromis između metoda filtriranja i metoda omotača. Najčešća ugrađena metoda odabira značajki jest regularizacija koja nastoji minimizirati grešku

procjene algoritma učenja i pritom prisiljava koeficijente uz značajke da budu mali (ili nula), što rezultira i modelom regularizacije i odabirom podskupa značajki [46].

Konačno, u hibridnim su metodama agregiranjem više odabranih podskupova značajki iz različitih metoda rezultati robusniji čime se povećava vjerodostojnost odabranih značajki [46].

Nadalje, Bolón - Canedo i sur. navode još jednu klasifikaciju metoda odabira značajki - prema funkcionalnom obilježju; metode rangiranja (eng. *ranking methods*) koje svakoj značajki dodjeljuju neku težinu na način da se očuva redoslijed koji odgovara njihovoj teorijskoj relevantnosti i metode procjene podskupa značajki (eng. *subset evaluation*) koje prvo koristeći strategiju pretraživanja proizvedu podskup značajki, a zatim ga evaluiraju temeljem funkcije evaluacije koja određuje konačni odabrani podskup značajki [4].

Osim toga, metode mogu biti univarijatne ili multivarijatne, ovisno o tome razmatraju li svaku značajku neovisno o ostalima ili ne [4].

Bolón - Canedo i sur. [4] tvrde da su glavni čimbenici u usporedbi metoda upravo brzina računanja i vjerojatnost prenaučivosti, no, može se zaključiti da svaka od postojećih vrsta metoda odabira značajki imaju svoje prednosti i nedostatke te da općenito ne postoji univerzalno "najbolja" metoda odabira značajki.

## 2.2. LINEARNI REGRESIJSKI MODEL

Višestruki linearni regresijski model je dan izrazom [5, 47, 48]:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon, \quad (2.1)$$

pri čemu je  $Y = [y_1, \dots, y_n]^T$  izlazna značajka,  $X_1, \dots, X_p$  su ulazne značajke (prediktori / regresori),  $X_j = [x_{1j}, \dots, x_{nj}]^T, \forall j = 1, \dots, p$ ,  $\beta_0$  je slobodni koeficijent (eng. *intercept*),  $\beta_1, \dots, \beta_p$  su koeficijenti uz značajke  $X_1, \dots, X_p$  redom, a  $\varepsilon$  je slučajna greška ili šum. Nadalje,  $\beta = [\beta_1, \dots, \beta_p]^T$  je vektor koeficijenata uz značajke, a  $X = [X_1, \dots, X_p]$  je matrica prediktora. Linearni regresijski model pretpostavlja linearne povezanosti nezavisne i zavisne značajke [49].

Ukoliko se u matricu prediktora  $X$  uključi i konstanta 1, odnosno  $X = [1, X_1, \dots, X_p]$ , a  $\beta_0$  se uključi u vektor koeficijenata uz značajke  $\beta$ , odnosno  $\beta = [\beta_0, \beta_1, \dots, \beta_p]^T$ , izraz (2.1) se može zapisati kao:

$$Y = X\beta + \varepsilon. \quad (2.2)$$

Procjene koeficijenata uz značajke  $\hat{\beta}$  se dobiju minimiziranjem izraza [5]:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \right\}. \quad (2.3)$$

Hastie i sur. [5] te Zou i Hastie [50] navode dva razloga zbog kojih obično procjene dobivene metodom najmanjih kvadrata (2.3) ne zadovoljavaju. Prvi je razlog točnost predikcije. Naime, ovakve procjene obično imaju malu pristranost, ali veliku varijancu. Skupljanjem (eng. *shrinking*), tj. pritezanjem koeficijenata ili postavljanjem istih na nulu, čime se potencijalno povećava pristranost radi smanjenja varijance (tzv. eng. *bias-variance trade-off*), što obično rezultira boljom ukupnom točnošću predikcije. Drugi je razlog interpretabilnost. Ukoliko je u modelu prisutan veliki broj prediktora, obično je poželjno odabrati manji podskup značajki koji luči najsnažniji efekt. Dakle, s ciljem dobivanja "šire slike", žrtvuju se manji detalji i treba težiti parsimoniji [5, 44, 45].

## 2.3. REGULARIZACIJA

Procjena koeficijenata uz značajke prilikom regularizacije razlikuje se od (2.3) u dodatnom članu prilikom minimizacije.

Lasso (skraćeno od eng. *least absolute shrinkage and selection operator*) regularizacija [5]:

$$\hat{\beta}^{lasso} = \operatorname{argmin}_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}. \quad (2.4)$$

Parametar  $\lambda$  ( $\lambda \geq 0$ ) je parametar složenosti koji kontrolira količinu skupljanja; što je vrijednost  $\lambda$  veća, veća je količina skupljanja i vrijednosti koeficijenata se smanjuju prema nuli [5]. Uvođenjem ograničenja veličine na koeficijente, ublažava se problem multikolinearnosti značajki.

Nadalje, (2.4) može se promatrati kao problem optimizacije izraza [5, 51]:

$$\hat{\beta}^{lasso} = \operatorname{argmin}_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2, \quad (2.5)$$

uz uvjet  $\sum_{j=1}^p |\beta_j| \leq t$ , za neki  $t$ .

Izraz  $\sum_{j=1}^p |\beta_j|$  naziva se lasso penalizacija i odgovara  $l_1$  normi vektora koeficijenata  $\beta$  ( $\|\beta\|_1$ ).

U izrazima (2.4) i (2.5) postoji 1-1 podudarnost između parametara  $\lambda$  i  $t$ . Naime, kako  $t \rightarrow \infty$ , problem se pretvara u metodu najmanjih kvadrata i  $\lambda$  postaje 0. S druge strane, kako  $t \rightarrow 0$ , svi koeficijenti se smanjuju na 0, a  $\lambda \rightarrow \infty$ .

Nadalje, može se uočiti da penalizacijski izraz ne sadrži slobodni koeficijent  $\beta_0$ . Razlog tome je što penalizacija slobodnog koeficijenta jednostavno čini proceduru ovisnom o ishodištu  $Y$ -a te je, stoga, uobičajeno da se značajke centriraju (tj. svaki  $x_{ij}$  se zamjenjuje s  $x_{ij} - \bar{x}_j$ ), a  $\beta_0$  se onda procjenjuje s  $\beta_0 = \bar{Y}$ . Preostali se koeficijenti potom procjenjuju lasso regularizacijom s centriranim značajkama, a matrica prediktora onda ima  $p$  umjesto  $p + 1$  stupaca [5]. Postupak je analogan i kod hrbatne regularizacije i regularizacije elastičnom mrežom.

Hrbatna (eng. *ridge*) regularizacija [5]:

$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}. \quad (2.6)$$

Analogno kao kod lasso regularizacije, (2.6) može se promatrati kao problem optimizacije izraza [5]:

$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2, \\ \text{uz uvjet } \sum_{j=1}^p \beta_j^2 \leq t, \text{ za neki } t.$$

Izraz  $\sum_{j=1}^p \beta_j^2$  naziva se hrbatna penalizacija i odgovara kvadratu  $l_2$  norme vektora koeficijenata  $\beta$  ( $\|\beta\|_2^2$ ). Drugim riječima,  $\|\beta\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$ .

Regularizacija elastičnom mrežom (eng. *elastic net*) [5]:

$$\hat{\beta}^{elasticnet} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^p \beta_j^2 + \lambda_2 \sum_{j=1}^p |\beta_j| \right\}. \quad (2.7)$$

Ovaj izraz se može smatrati penaliziranom metodom najmanjih kvadrata [50]. Ukoliko se uvede oznaka [5,50]:  $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$ , tada se (2.7) može promatrati kao problem optimizacije izraza:

$$\hat{\beta}^{elasticnet} = \operatorname{argmin}_{\beta} \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2, \\ \text{uz uvjet } (1 - \alpha) \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j| \leq t, \text{ za neki } t.$$

Izraz  $(1 - \alpha) \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j|$  se naziva penalizacija elastične mreže i predstavlja konveksnu kombinaciju lasso i hrbatne penalizacije, pri čemu je  $\alpha \in [0, 1]$ . Specijalni slučajevi su za  $\alpha = 0$ , kada se radi o hrbatnoj regularizaciji i  $\alpha = 1$ , kada se radi o lasso regularizaciji [50].

Parametar  $\lambda$  se obično procjenjuje unakrsnom validacijom (eng. *cross-validation*), a parametar  $\alpha$  unakrsnom validacijom [50–52], generaliziranom unakrsnom validacijom [52], Akaike informacijskim kriterijem (eng. *Akaike information criterion* - AIC) [53], Bayesovim informacijskim kriterijem (eng. *Bayesian information criterion* - BIC) [53], generaliziranim informa-

cijskim kriterijem (eng. *generalized information criterion* - GIC) [53], pretraživanjem mreže (eng. *grid search*) [54] itd.

### 2.3.1. Usporedba metoda regularizacije

Hastie i sur. [5] napominju da je prednost metoda regularizacije u odnosu na metode koje odabiru najbolji podskup značajki (dakle, zadržavaju ili odbacuju značajku) to što su neprekidne i nisu toliko varijabilne. Naime, kod metoda za odabir najboljeg podskupa značajki se zadržavanjem podskupa značajki proizvodi model koji je interpretabilan i koji vjerojatno ima nižu grešku predikcije od punog modela (modela sa svim značajkama), no, s obzirom na to da je zadržavanje / odbacivanje značajki diskretan proces, često pokazuje visoku varijancu i stoga ne smanjuje grešku predikcije punog modela.

Hrbatna regresija, za razliku od procesa odabira podskupa značajki, kontinuirani je proces koji priteže koeficijente i stoga je stabilniji. No, nedostatak je što ne postavlja koeficijente na nulu, odnosno uvijek zadržava sve značajke u modelu i stoga ne rezultira lako interpretabilnim modelima [50, 51].

S druge strane, lasso regularizacija uzima najbolje od hrbatne regresije i odabira podskupa značajki jer istovremeno priteže koeficijente i pritom postavlja neke od njih na nulu [50, 51]. To znači da se radi o kontinuiranom i, stoga, stabilnijem procesu koji istovremeno radi odabir značajki, čime postiže interpretabilnost i jednostavnost. S obzirom na rastući broj skupova velike dimenzionalnosti, lasso je često korišten zbog svoje rijetke reprezentacije, no ima ograničenja u nekim slučajevima [50]:

- u slučaju  $p > n$  odabire najviše  $n$  značajki zbog prirode problema konveksne optimizacije. Osim toga, problem nije dobro definiran ako ograničenje  $t$  na  $l_1$  normu koeficijenata nije manje od određene vrijednosti (problem se pretvara u metodu najmanjih kvadrata za velike  $t$  iz izraza (2.5) na str. 10),
- u slučajevima kada postoji grupa značajki u kojoj su sve značajke međusobno visoko korelirane, lasso nastoji odabrati samo jednu značajku i pritom mu je svejedno koju, te mu na taj način nedostaje sposobnost otkrivanja informacije o grupiranju značajki,
- u slučajevima  $p < n$  i ako postoje visoke korelacije između značajki, Tibshirani [51] je empirijski pokazao da hrbatna regresija bolje funkcionira.

S tim u vezi, Zou i Hastie [50] pokušali su pronaći novu metodu koja funkcionira dobro kao i lasso u uvjetima u kojima on dobro funkcionira, no koja može riješiti njegove prethodno navedene nedostatke i postići bolju prediktivnu moć. Predlažu regularizaciju elastičnom mrežom koja, kao i lasso, istovremeno radi odabir značajki i kontinuirano skupljanje, no za razliku od njega, može odabrati i grupu koreliranih značajki. Empirijski su pokazali da elastična mreža često nadmašuje lasso u smislu prediktivne točnosti.

## 2.4. MJERE SLAGANJA I POVEZANOSTI

S obzirom na to da se u ovom istraživanju nastoji uključiti domensko znanje stručnjaka, bitno je provjeriti slaganje i povezanost njihovih procjena, te su stoga u nastavku opisane mjere koje će se koristiti u tu svrhu.

U literaturi se koriste različite mjere za ispitivanje slaganja (eng. *agreement*) između stručnjaka, a u ovom istraživanju potrebne su mjere koje su prikladne u slučaju više stručnjaka (pri čemu "više" podrazumijeva više od dva) te ordinalnih ljestvica zbog toga što su domenska znanja stručnjaka o značajkama predstavljena s kategorijama *Nema utjecaja*, *Slab utjecaj*, *Srednji utjecaj* i *Jak utjecaj* (detaljnije opisano u potpoglavlju [Domensko znanje stručnjaka u slučaju stvarnih podataka](#)).

Sertdemir i sur. [55] navode kako takvih mjera nema mnogo te u svom istraživanju uspoređuju nekoliko mjera slaganja ocjenjivača na simuliranim podacima te na jednom skupu stvarnih podataka. Upozoravaju kako istraživači trebaju biti oprezni pri odabiru mjera slaganja, osobito ako postoje "pomaci" u ocjenama ocjenjivača koji mogu rezultirati jakim povezanošću, ali ne i slaganjem. Savjetuju primjenu više različitih mjera prije donošenja bilo kakvih zaključaka te da se za ordinalne ljestvice, u kombinaciji s mjerama slaganja, koriste i mjere povezanosti (eng. *association*). Obje vrste mjera daju uvid u snagu odnosa između ocjena ocjenjivača [56].

Nadalje, Mitani i sur. radu [57] navode da su ordinalne ljestvice, više ispitanika (više od dva) u istraživanju te neuravnoteženi podaci izazovi za korištenje tradicionalnih mjera slaganja. Primjerice, Cohenova kapa statistika  $\kappa$  popularna je mjera slaganja za ordinalne ljestvice, no ograničena je na dva ocjenjivača (u ovom slučaju stručnjaka). S obzirom na navedeno, razvijena su neka proširenja Cohenove kapa statistike koje pružaju sumarne mjere slaganja i povezanosti između više ocjenjivača poput Fleissove kape, intraklasnog koeficijenta korelacije i drugih. Unatoč dostupnosti ovih proširenih mjera, istraživanja pokazuju da mogu dovesti do složenosti



u interpretaciji i da su neizvedive u slučajevima s velikim brojem ocjenjivača.

S tim u vezi, Mitani i sur. [57, 58] predlažu korištenje mjera slaganja i povezanosti koje se temelje na modelu i koje, kao takve, pružaju fleksibilniji pristup za izračunavanje slaganja i povezanosti više ispitanika za ordinalne ljestvice te su otporne na nedostajuće i neuravnotežene podatke. Te mjere su kapa koeficijent slaganja temeljem modela (eng. *model-based kappa agreement*)  $\kappa_m$  za ordinalne značajke [56] i kapa koeficijent povezanosti temeljem modela (eng. *model-based kappa association*)  $\kappa_{ma}$  za ordinalne značajke [59]. Kapa statistika temeljenu na modelu  $\kappa_m$  temelji se na generaliziranom linearnom mješovitom modelu (GLMM) i na odgovarajući način ispravlja slučajno slaganje. Poprima vrijednosti između 0 i 1, gdje 0 označava slaganje koje nije bolje od slučajnosti, a 1 označava savršeno slaganje među ispitanicima. Ova mjera izbjegava neke od ključnih nedostataka povezanih s Cohenovom kapom i njezinim proširenjima [56]. Kapa statistika povezanosti temeljena na modelu  $\kappa_{ma}$ , uključuje težine (kvadratne ili linearne) koje stavljaju veći naglasak na kategorije koje su bliže nego kategorije koje su udaljene. Neke od mjera koje se također koriste su pojašnjene u nastavku.

Intraklasni koeficijent korelacije (eng. *intraclass correlation coefficient - ICC*) [60] koristi se kao indeks pouzdanosti u analizi pouzdanosti unutar ocjenjivača (eng. *intrarater*), među ocjenjivačima (eng. *interrater*) i testom i ponovnim testom (eng. *test - retest*) [61], dok ga Mitani i sur. [57] definiraju ICC kao mjeru povezanosti koja se koristi za ordinalnu ljestvicu više ocjenjivača. Postoji nekoliko verzija ICC-a ovisno o dizajnu i svrsi, a u ovom istraživanju koristit će se *ICC1*, *ICC2* i *ICC3* (za detaljnije formule pogledati [57]). *ICC* se izvodi iz komponenti modela analize varijance (ANOVA), a vrijednost intraklasnog koeficijenta korelacije se kreće od 0 (sustav mjerenja nije konzistentniji nego što bi se očekivalo po šansi) do 1 (savršena pouzdanost) [57, 61].

Kendallov koeficijent podudarnosti  $W$  (eng. *Kendall's coefficient of concordance  $W$* ) je neparametarska statistika koja predstavlja indeks podudarnosti među ocjenjivačima za ordinalne podatke. Normalizacija je statistike Friedmanova testa i može se koristiti za procjenu slaganja između više ispitanika [62, 63]. Vrijednost  $W$  se kreće od 0 (nema slaganja) do 1 (potpuno slaganje). Navedena statistika  $W$  ima  $\chi^2$  distribuciju i služi za testiranje hipoteza [64]:

$$\begin{aligned} H_0 : W &= 0 \quad (\text{rangovi su nezavisni}) \\ H_1 : W &\neq 0 \quad (\text{rangovi su zavisni}). \end{aligned} \tag{2.8}$$

Krippendorffov  $\alpha$  koeficijent pouzdanosti (eng. *Krippendorff's  $\alpha$  coefficient*) koeficijent je



pouzdanosti razvijen za mjerenje slaganja između ispitanika, a njegova prednost je što se može primijeniti na nominalne, ordinalne ili kontinuirane ljestvice, te na nepotpune skupove podataka gdje neki ispitanici ne ocjenjuju sve stavke [65]. Vrijednost  $\alpha$  kreće se u sljedećem rasponu:

$$1 \geq \alpha \geq 0 \begin{cases} \pm \text{ pogreška uzorkovanja} \\ - \text{ sustavno neslaganje,} \end{cases}$$

gdje je  $\alpha = 1$  označava savršenu pouzdanost,  $\alpha = 0$  odsustvo pouzdanosti, a  $\alpha$  čak može biti i negativna (-1). Nadalje, pogreška uzorkovanja (eng. *sampling error*) može se dogoditi iz više razloga, među kojima je nedovoljna veličina uzorka kada je dostupno nekoliko mjerenja od kojih svako ima veći utjecaj na  $\alpha$  te opažena odstupanja nisu jednaka očekivanim odstupanjima pa  $\alpha$  oscilira s vrijednostima manjim ili većim od nule. S druge strane, sustavno neslaganje (eng. *systematic disagreement*) nastaje kada se ocjenjivači slažu da se ne slažu ili slijede suprotna tumačenja uputa koje su im dane [65].

U tablici 2.1 su navedena tumačenja spomenutih mjera slaganja i povezanosti koja slijede iz literature.

Tablica 2.1: Tumačenja mjera slaganja i povezanosti prema iznosu

Mjera	Iznos	Tumačenje	Izvor
$\kappa_m$ , $\kappa_{ma}$	< 0.00	siromašno slaganje	[66]
	0.00 - 0.20	slabo slaganje	
	0.21 - 0.40	pošteno slaganje	
	0.41 - 0.60	umjereno slaganje	
	0.61 - 0.80	značajno slaganje	
	0.81 - 1.00	skoro savršeno slaganje	
ICC	< 0.50	slaba pouzdanost	[57, 61]
	0.50 - 0.75	umjerena pouzdanost	
	0.75 - 0.90	dobra pouzdanost	
	> 0.90	izvrsna pouzdanost	
W	Temeljem $p$ -vrijednosti stat. testa (2.8) uz nivo statističke značajnosti 0.05.		[64]
$\alpha$	Sugerira se > 0.80 (ili barem > 0.667).		[65]

Mjera	Iznos	Tumačenje	Izvor
-------	-------	-----------	-------

$\kappa_m$  - kapa koeficijent slaganja temeljem modela,  $\kappa_{ma}$  - kapa koeficijent povezanosti temeljem modela,  $ICC1$  - intraklasni koeficijent korelacije 1,  $ICC2$  - intraklasni koeficijent korelacije 2,  $ICC3$  - intraklasni koeficijent korelacije 3,  $W$  - Kendallov koeficijent podudarnosti  $W$ ,  $\alpha$  - Krippendorffov koeficijent pouzdanosti.

## 2.5. (META)HEURISTIKE

Problemi koji zahtjevaju optimizaciju danas su prisutni u raznim sferama života - privatnim, znanstvenim, praktičnim i drugim. Optimizacija se odvija, primjerice, prilikom minimiziranja troškova i maksimizacije profita, ali općenito svaki proces ima potencijal za optimizaciju [67]. No, pronalazak optimalnog rješenja nije uvijek moguć. Naime, navedeni problemi mogu biti izrazito kompleksni, a do optimalnog rješenja problema se dolazi pretraživanjem cjelokupnog prostora rješenja, što je vrlo zahtjevno u pogledu vremena izvršavanja, računske složenosti i potrebnih računalnih kapaciteta.

Algoritmi koji ne jamče pronalazak optimalnih rješenja, ali pronalaze dovoljno dobra, zadovoljavajuća rješenja u razumnom vremenu izvršavanja i imaju relativno nisku računsku složenost nazivaju se heurističkim metodama ili heuristikama [68] (grč. *heuriskein* - otkriti).

S druge strane, metaheuristika je strategija, odnosno metodologija višeg sloja, za dizajn heuristike, odnosno, skup algoritamskih koncepata koji se mogu koristiti za definiranje heuristika primjenjivih na širok raspon problema [68]. Metaheuristike imaju tri glavne svrhe: (1) brže rješavanje problema, (2) rješavanje problema velike složenosti i (3) dobivanje robusnih algoritama. Štoviše, jednostavne su za dizajn i implementaciju te su vrlo fleksibilne [67].

Spomenuti problemi velike složenosti (eng. *large-scale problems*) odnose se na optimizacijske probleme s velikim brojem značajki, probleme s prostorima rješenja određenim velikim brojem (ne)linearnih uvjeta, kao i probleme s velikom količinom podataka za čiju je analizu nužno koristiti računarstvo visokih performansi (eng. *high-performance computing* - HPC) [67, 69].

Metaheuristike rješavaju probleme za koje se vjeruje da su općenito teški, pretražujući obično veliki prostor rješenja, što se postiže smanjenjem efektivne veličine prostora i učinkovitim pretraživanjem tog prostora. Prilikom definiranja načina pretraživanja prostora rješenja, treba uzeti u obzir dva kriterija koja se međusobno suprotstavljaju [67]:

- diverzifikacija, koja se odnosi na pretraživanje prostora rješenja; osigurava da se sva područja prostora pretraživanja ravnomjerno istraže i da se pretraživanje ne ograniči na manji broj područja,
- intenzifikacija, koja se odnosi na iskorištavanje najboljih pronađenih rješenja; temeljem dobivenih "dobrih" rješenja određuju obećavajuće regije, koje se onda temeljitije istražuju s ciljem pronalaska još boljih rješenja.

U nastavku su nabrojane neke vrste podjela metaheuristika [67]:

- nadahnute prirodom (npr. evolucijski algoritmi, algoritmi inteligencije roja) / nenadahnute prirodom (npr. lokalno pretraživanje),
- s uporabom memorije (koriste memoriju koja sadrži neke informacije dobivene tijekom pretraživanja, npr. tabu pretraga) / bez uporabe memorije (npr. lokalno pretraživanje, simulirano kaljenje),
- determinističke (rješavaju problem optimizacije donošenjem determinističkih odluka te isto početno rješenje uvijek dovodi do istog konačnog rješenja, npr. lokalno pretraživanje, tabu pretraživanje) / stohastičke (primjenjuju se neka nasumična pravila, isto početno rješenje ne mora dovesti do istog konačnog rješenja, npr. simulirano kaljenje, evolucijski algoritmi),
- pretraživanje temeljeno na jednom rješenju (manipuliraju i transformiraju jedno rješenje, skloni intenzifikaciji, npr. lokalno pretraživanje, simulirano kaljenje) / pretraživanje temeljeno na populaciji rješenja (razvijena je cijela populacija rješenja, skloni diverzifikaciji, npr. roj čestica, evolucijski algoritmi),
- iterativni (počinju s cjelovitim rješenjem ili populacijom rješenja koja se prilikom svake iteracije transformira pomoću operatora pretraživanja; većina ih je takvih) / pohlepni (polaze od praznog rješenja i u svako korak se dodjeljuje varijabla odluke problema dok se ne dobije potpuno rješenje).

Nadalje, glavni koncepti prilikom dizajniranja metaheuristike su [67]:

- reprezentacija, odnosno prikaz potencijalnih rješenja ili kodiranje, koja igra glavnu ulogu u učinkovitosti metaheuristike. Mora biti prikladna i relevantna za problem optimizacije koji se rješava. Uobičajene su reprezentacije linearne (binarni vektori, vektori diskretnih

ili realnih vrijednosti, vektori permutacija) i nelinearne (npr. stabla). Reprezentacija mora imati sljedeće karakteristike:

- cjelovitost (eng. *completeness*) - odnosi se na zastupljenost svih rješenja povezanih s problemom; moraju sva biti zastupljena,
- povezanost (eng. *connexity*) - između svaka dva rješenja u prostoru rješenja mora postojati put pretraživanja i svako se rješenje može postići,
- učinkovitost (eng. *efficiency*) - operatori pretraživanja mogu lako manipulirati reprezentacijom, a njihova vremenska i prostorna složenost mora biti smanjena,
- definiranje funkcije cilja koja formulira cilj koji se treba postići. Ona svakom rješenju pridružuje stvarnu vrijednost koja odgovara kvaliteti ili prikladnosti rješenja te se tako omogućuje poredak svih rješenja i usmjerava pretraživanje prema "dobrim" rješenjima.

Prilikom korištenja metaheuristika, potrebna je prilagodba parametara (eng. *parameter tuning*) koji zahtjevaju pažljivu inicijalizaciju i imaju veliki utjecaj na učinkovitost pretraživanja [67]. Nije očito koje parametre odabrati, to uglavnom ovisi o problemu. Postoje dvije različite strategije za podešavanje parametara [67]:

- off-line: vrijednosti parametara su fiksirane prije izvršenja metaheuristike,
- online: parametri se kontroliraju i ažuriraju dinamički ili adaptivno tijekom izvođenja metaheuristike.

Konačno, potrebno je nakon korištenja metaheuristike napraviti korektnu analizu rezultata, što je najčešće nemoguće napraviti teorijski pa se koristi eksperimentalna evaluacija u sljedeća tri koraka [67]:

- dizajn eksperimenta: potrebno definirati ciljeve, instance za evaluaciju i čimbenike ispitivanja,
- mjerenje: potrebno je odrediti mjere za procjenu izvedbe, a zatim se dobiveni rezultati analiziraju statističkom analizom i rezultati se uspoređuju s rezultatima drugih algoritama,
- izvješće: potrebna prezentacija rezultata i njihova analiza u odnosu na definirane ciljeve; bitno je osigurati reproducibilnost eksperimenta.

Korištenje metaheuristike za poboljšanje metoda strojnog učenja postalo je moderno i postiglo je velike uspjehe u nadziranim (klasifikacija i regresija) i nenadziranim (grupiranje i rudarenje pravila) problemima u različitim domenama [70], a njihov doprinos u kontekstu redukcije dimenzionalnosti je već spomenut ranije u dijelu prethodnih istraživanja (primjerice [71–78]).

### 2.5.1. Algoritam pretraživanja kukavice

Budući da je pregledom literature uočeno da su metaheuristički algoritmi, pokazali veliki potencijal u mnogim primjenama koje zahtjevaju optimizaciju, i u ovom će se istraživanju koristiti metaheuristički algoritam za procjenu vrijednosti penalizacijskog koeficijenta koji će u regularizacijskom modelu dodatno stajati uz ulazne značajke.

Velika skupina metaheurističkih algoritama nastala je proučavanjem prirodnih procesa. Jedan od njih je i algoritam pretraživanja kukavice (eng. *cuckoo search* - CS), koji se koristi u ovom istraživanju. Algoritam funkcionira na principu ponašanja nekih vrsta ptica kukavica s karakteristikama poput parazitizma legla [79]. Prema [80], ova metaheuristika pokazala se potencijalno efikasnijom od optimizacije roja čestica (eng. *particle swarm optimization*) i genetskih algoritama (eng. *genetic algorithms*).

#### **Ponašanje ptica kukavica**

Ptice kukavice prepoznatljive su po lijepom zvuku kojeg mogu proizvesti, ali i po "agresivnoj" strategiji razmnožavanja [79]. Naime, neke vrste, kao što su *ani* i *Guira* kukavice, polažu svoja jaja u zajednička gnijezda te mogu ukloniti tuđa jaja kako bi povećali vjerojatnost izlijeganja vlastitih jaja. Nadalje, velik broj vrsta kukavica polaže jaja u gnijezda drugih ptica - ptica domaćina, koje su često druge vrste. Ako ptica domaćin otkrije da jaja nisu njezina, ili će baciti jaja koja nisu njezina ili će napustiti svoje gnijezdo i izgraditi novo negdje drugdje. Neke vrste kukavica, kao *Tapera*, tome su se prilagodile na način da su ženke razvile oponašanje boja i uzorka jaja nekoliko odabranih vrsta domaćina, čime smanjuju vjerojatnost da će njihova jajašca biti odbačena, što pogoduje razmnožavanju. Nadalje, kukavice često biraju gnijezdo u kojem je ptica domaćin pravo položila svoja jaja. S obzirom na to da se jaja kukavice obično izlegnu ranije od jaja domaćina i da mala kukavica potom instiktivno iz gnijezda izbacuje jaja, time će dobiti veći udio hrane koju pribavi ptica domaćin. Dodatno, male kukavice mogu oponašati glasanje ptića domaćina [79, 80].

**Algoritam pretraživanja kukavice**

Radi jednostavnosti, koristit će se tri idealizirana pravila [79]:

- 1) kukavica izliježe jedno po jedno jaje u trenutku i polaže ga u slučajno odabrano gnijezdo,
- 2) najbolja gnijezda s jajima visoke kalitete se prenose u iduću generaciju,
- 3) broj dostupnih gnijezda domaćina je fiksna, a vjerojatnost da ptica domaćin otkrije jaje kukavice je  $p, p \in [0, 1]$ . Ptica domaćin tada odbacuje jaje ili napušta gnijezdo i gradi novo, dakle udio od  $p$  od  $n$  gnijezda će se zamijeniti novim gnijezdima.

Nadalje, može se smatrati da svako jaje u gnijezdu predstavlja rješenje, a jaje kukavice novo rješenje. Cilj optimizacije koristiti je nova i potencijalno bolja rješenja (kukavice) za zamjenu rješenja koja nisu možda dovoljno dobra, pri čemu se kvaliteta ili prikladnost (eng. *fitness*) rješenja može mjeriti vrijednošću proporcionalnom vrijednosti funkcije cilja. Premda se algoritam može proširiti i na slučaj kada svako gnijezdo ima više jaja koje predstavljaju skup rješenja, ovdje se koristi najjednostavniji pristup gdje svako gnijezdo ima jedno jaje. Na temelju ovih pravila, osnovni koraci algoritma su prikazani pseudokodom 1 u nastavku [79, 80].

---

**Algoritam 1** Algoritam pretraživanja kukavice (CS) - pseudokod, izvor: [79]

---

funkcija cilja  $f(x), x = (x_1, \dots, x_d)^T$

generiraj inicijalnu populaciju  $n$  gnijezda domaćina  $x_i, i = 1, \dots, n$

**while**  $t < MaxGeneracija$  ili se dostigne zaustavni kriterij **do**

nasumično odabрати pticu kukavicu Lévyjevim letovima i evaluirati njezinu kvalitetu /prikladnost  $F_i$

odabрати slučajno gnijezdo  $j, j = 1, \dots, n$

**if**  $F_i > F_j$  **then**

zamijeniti gnijezdo  $j$  novim rješenjem

**end if**

udio od  $p$  od  $n$  najgorih gnijezda odbačen je i izgrađena su nova

zadrži najbolja rješenja (tj. gnijezda s najkvalitetnijim rješenjima)

rangiraj rješenja i nađi trenutno najbolje

**end while**

obrada i vizualizacija rezultata

---

Novo rješenje, npr.  $x^{(t+1)}$ , za pticu kukavicu  $i$  dobija se na sljedeći način [79, 80]:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \cdot u, \quad (2.9)$$

gdje je  $\alpha > 0$  veličina koraka koja skalira slučajnu varijablu  $u$  koja dolazi iz Lévyjeve distribucije, tj.  $u \sim t^{-\lambda}$ ,  $\lambda \in (1, 3]$ , gdje je  $t$  redni broj generacije. Izraz (2.9) predstavlja Lévyjev let, odnosno slučajnu šetnju (*random walk*) s duljinama koraka koje slijede Lévyjevu distribuciju.

## 2.6. METODA CART

U ovom je istraživanju, osim višestruke linearne regresije s lasso i hrbatnom regularizacijom te regularizacijom elastičnom mrežom, korištena i metoda klasifikacijskih i regresijskih stabala (eng. *classification and regression trees* - CART) radi kreiranja prediktivnih modela i usporedbe rezultata. S obzirom na to da je izlazna značajka u ovom istraživanju numerička, radi se o regresijskoj varijanti CART metode.

Ova metoda odabrana je zbog jednostavne i intuitivne interpretabilnosti, koja je već istaknuta kao bitna karakteristika kreiranih modela u kontekstu stvarnih podataka, odnosno podataka o troškovima energije zgrada javnog sektora Republike Hrvatske.

Metodu CART osmislili su Breiman i sur. [81]. U nastavku će biti rečeno nešto osnovno o metodi, dok se detaljnije o metodi može pročitati primjerice u [49, 81].

Prostor značajki  $X$ , tj. skup mogućih vrijednosti za  $X_1, \dots, X_p$  se nizom binarnih podjela dijeli u disjunktna područja (particije)  $R_1, \dots, R_J$ , za neki  $J$ . Za svaki slučaj koji pripadne u područje  $R_j$ , procjena vrijednosti izlazne značajke je jednaka i dobiva se kao prosječna vrijednost izlazne značajke za sve slučajeve koji pripadaju području  $R_j$ .

Cilj je pronaći regije  $R_1, \dots, R_J$  koje minimiziraju grešku, odnosno:

$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \quad (2.10)$$

gdje je  $\hat{y}_{R_j}$  prosječna vrijednost izlazne značajke za sve slučajeve koji pripadaju području  $R_j$ . Budući da je generalno računalno neizvedivo provjeriti svaku moguću particiju prostora značajki u  $J$  particija, koristi se pohlepni pristup od gore prema dolje. Pristup od gore prema dolje (eng. *top - down*) znači da se kreće od vrha stabla (korijena), kada su svi slučajevi u jednom području, a potom se prostor značajki dijeli; svaki čvor dijeli stablo na dva nova dijela. Nadalje, pristup je pohlepan (eng. *greedy*) jer u svakom koraku traži najbolju podjelu koja se

može napraviti u tom koraku, umjesto da gleda unaprijed koja podjela bi dovela do boljeg stabla u nekom budućem koraku [49].

S ciljem rekurzivnog binarnog dijeljenja, odabire se značajka  $X_j$  i granična vrijednost  $s$  tako da podjela prostora značajki u regije  $\{X|X_j < s\}$  i  $\{X|X_j \geq s\}$  vodi k najvećoj mogućoj redukciji greške u izrazu 2.10, gdje oznaka  $\{X|X_j < s\}$  znači da je to regija prostora značajki gdje značajka  $X_j$  poprima vrijednosti manje od  $s$ . Uz uvođenje sljedećih oznaka:

$$R_1(j, s) = \{X|X_j < s\},$$

$$R_2(j, s) = \{X|X_j \geq s\},$$

gdje su  $R_1(j, s)$  i  $R_2(j, s)$  parovi poluravnina određenih s  $X_j$  i  $s$ , prethodno se može prevesti i u problem minimiziranja izraza

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

gdje je  $\hat{y}_{R_1}$  prosječna vrijednost izlazne značajke za sve slučajeve koji pripadaju području  $R_1(j, s)$ , a  $\hat{y}_{R_2}$  prosječna vrijednost izlazne značajke za sve slučajeve koja pripadaju području  $R_2(j, s)$ . Proces se tako ponavlja dok se ne dostigne neki zaustavni kriterij (npr. minimalni broj slučajeva u području), s tim da se u svakom novom koraku umjesto dijeljenja početnog prostora značajki u particije, dijeli se jedno od područja dobiveno u prethodnom koraku. Na kraju krajeva, kada se formiraju područja, odnosno particije  $R_1, \dots, R_J$ , za novi slučaj iz skupa za testiranje, prati se putanja od korijenskog do terminalnog čvora stabla i odredi se pripadnost području [82], a potom se predikcija vrijednosti izlazne značajke se dobije kao prosječna vrijednost izlazne značajke svih slučajeva koji su u tom istom području [49].

Osim jednostavne i intuitivne interpretabilnosti, prednosti su CART metode jasna vizualizacija i jednostavno "nošenje" s kvantitativnim značajkama. Osim toga, neki ljudi smatraju da je proces donošenja odluka u stablima odlučivanja sličan ljudskom načinu razmišljanja [49].

S druge strane, kao nedostatak ove metode navodi se tendencija prenaučivosti. S tim u vezi, provodi se obrezivanje (eng. *pruning*) stabla. Naime, najprije se proizvede dovoljno veliko stablo  $T_0$ , a potom se on obrezuje unatrag, čime se dobije podstablo. Cilj je odabrati podstablo koje dovodi najmanjoj grešci na skupu za testiranje korištenjem unakrsne validacije [49].

Općenito, može se zaključiti da u regresijskim problemima i metoda linearne regresije i metoda CART rezultiraju interpretabilnim modelima, a odluka o tome koju koristiti se može



donijeti na temelju strukture podataka. Naime, James i sur. [49] navode da ako su veze između ulaznih i izlazne značajke dobro aproksimirane linearnim modelom, tada će vrlo vjerojatno linearna regresija funkcionirati vrlo dobro i pritom nadmašiti metodu CART. Ukoliko je, s druge strane, veza između ulaznih značajki i izlazne vrlo nelinearna ili kompleksna, tada će vrlo vjerojatno stabla nadmašiti linearnu regresiju.

## 2.7. MJERE EVALUACIJE

Chicco i sur. [83] usporedili su nekoliko mjera ( $R^2$ , te *MAPE*, *SMAPE*, *MAE*, *MSE* i *RMSE* greške) koje se uobičajeno koriste u znanstvenoj literaturi za evaluaciju uspješnosti regresijskih analiza i predlažu korištenje koeficijenta determinacije -  $R^2$  kao standardne mjere za evaluaciju regresijskih analiza.

Koeficijent determinacije (eng. *coefficient of determination* -  $R^2$ ) mjera je koja predstavlja udio varijance izlazne značajke, koji je objašnjen ulaznim značajkama / prediktorima. Koeficijent može varirati između 0 i 1, a što je veća vrijednost, to je veća moć objašnjenja modela [43]. Međutim, zamjerka koeficijentu determinacije je što mu vrijednost raste kako se dodaju novi prediktori u model. Iz tog razloga, često se koristi korigirani koeficijent determinacije (eng. *adjusted coefficient of determination* -  $\tilde{R}^2$ ), kojem vrijednost može pasti ako dodani prediktori imaju malu moć objašnjenja ili ako broj stupnjeva slobode postane premali. Pogodan je za usporedbu modela s različitim brojem prediktora, različitom veličinom uzorka ili oboje jer uzima u obzir stupnjeve slobode, kao i modela na različitim skupovima podataka [43]. U ovom istraživanju koristi se korigirani koeficijent determinacije.

Budući da je u svrhu evaluacije uspješnosti predikcije modela dobro koristiti više različitih mjera, u ovom istraživanju se pored korigiranog koeficijenta determinacije ( $\tilde{R}^2$ ) koriste i sljedeće mjere: srednja apsolutna greška (eng. *mean absolute error* - *MAE*), simetrična srednja apsolutna postotna greška (eng. *symmetric mean absolute percentage error* - *SMAPE*) i korijen srednje kvadratne greške (eng. *root mean square error* - *RMSE*).

Osim navedenih mjera, u slučaju simuliranih podataka gdje su poznate i točne vrijednosti koeficijenata uz značajke, kao dodatna mjera evaluacije koristi se razlika procijenjenih koeficijenata uz ulazne značajke i onih korištenih u simulaciji (u smislu euklidske udaljenosti koeficijenata).

Uz oznake kao ranije, u nastavku slijede formule [83, 84]:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

$$SMAPE = 100 \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} [\%],$$

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

a euklidska udaljenost koeficijenata:

$$d(\beta, \hat{\beta}) = \sqrt{\sum_{j=0}^p (\beta_j - \hat{\beta}_j)^2}.$$

Nadalje, u klasičnoj linearnoj regresiji, korigirani koeficijent determinacije dan je izrazom [49]:

$$\tilde{R}^2 = 1 - (1 - R^2) \frac{(n-1)}{n-p-1},$$

pri čemu je  $R^2$  koeficijent determinacije dan formulom:

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

a  $\bar{Y}$  prosječna vrijednost izlazne značajke. Izraz  $n - p - 1 = n - (p + 1)$  predstavlja broj stupnjeva slobode reziduala modela s  $n$  slučajeva u kojem se procjenjuje  $p + 1$  parametara modela (po jedan regresijski koeficijent uz svaku značajku i slobodni koeficijent). Međutim, bitno je za naglasiti da metode regularizacije linearnog regresijskog modela mogu imati drugačije stupnjeve slobode reziduala modela nego klasična linearna regresija jer se u modelu potencijalno procjenjuje drugačiji broj parametara [5, 85]. Broj parametara modela se za lasso regularizaciju procjenjuje brojem značajki s koeficijentima različitim od 0 (aktivni skup prediktora), za hrbatnu regularizacije se procjenjuje tragom *hat* matrice (matrice  $X(X^T X + \lambda I)^{-1} X^T$ ), a za regularizaciju elastičnom mrežom tragom *hat* matrice za aktivni skup prediktora [5, 85]. Pri korekciji korigiranog koeficijenta determinacije regularizacijskih metoda trebaju se koristiti za njih odgovarajući stupnjevi slobode reziduala koji su jednaki razlici broja slučajeva i broju parametara modela.

Pri evaluaciji, točniji su modeli koji rezultiraju većim korigiranim koeficijentom determinacije  $\tilde{R}^2$ , manjim greškama  $MAE$ ,  $SMAPE$ ,  $RMSE$  i, u slučaju simuliranih podataka, manjom euklidskom udaljenosti procijenjenih i stvarnih koeficijenata  $d(\beta, \hat{\beta})$ .

Međutim, bitno je za naglasiti da metoda CART ne rezultira koeficijentima modela te, stoga, kod simuliranih podataka udaljenost koeficijenata kao mjera evaluacije za tu metodu nije smisljena. Također, korigirani koeficijent determinacije u njegovom klasičnom obliku se ne tumači

jednako kao u nelinearnim modelima poput CART-a. Iako postoje modifikacije koeficijenta determinacije, tada je upitna njihova usporedivost s korigiranim koeficijentom determinacije kod metoda regularizacije pa se iz tog razloga izostavlja iz analize kod metode CART.

## 3. PRETHODNA ISTRAŽIVANJA

U ovom poglavlju prikazana su dosadašnja istraživanja u nekoliko područja interesa ovog istraživanja - metodama odabira značajki, uključivanju znanja stručnjaka te, konačno, u području energetike građevinskog sektora.

### 3.1. PRETHODNA ISTRAŽIVANJA U KONTEKSTU

#### METODA ODABIRA ZNAČAJKI

U prethodnom poglavlju (potpoglavlje *Redukcija dimenzionalnosti*) spomenuto je da se metode redukcije dimenzionalnosti dijele na metode ekstrakcije i odabira značajki. Neki primjeri metoda ekstrakcije značajki su analiza glavnih komponenti (eng. *principal component analysis* - PCA), analiza nezavisnih komponenti (eng. *independent component analysis* - ICA), faktorska analiza (eng. *factor analysis* - FA) [9], no, kako je fokus ovog istraživanja na metodama odabira značajki, u nastavku je rečeno nešto više o prethodnim istraživanjima vezanih uz njih.

Metode filtriranja odabiru podskup značajki temeljem nekog kriterija. U istraživanjima se, u tu svrhu, koriste statistički kriteriji (npr. korelacije i  $\chi^2$  test nezavisnosti [82, 84, 86, 87]), informacijski kriteriji (npr. uzajamna informacija i omjer dobiti [71]), kriteriji bazirani na sličnosti, različitosti, udaljenosti [88] i drugi. Dodatno, odabir se značajki provodi i algoritmima, primjerice algoritmom koji radi odabir temeljem korelacija [6, 46] i Relief-F algoritmom [46], ali i klasteriranjem značajki *k*-sredinama (eng. *k-means*) [89], maksimizacijom očekivanja [90] i drugim metodama.

Metode omotača nastoje odabrati podskup značajki koji je optimalan za algoritam za učenje. Budući da je za skup od  $p$  značajki potrebno isprobati  $2^p$  kombinacija značajki, što zna biti zahtjevno i nepraktično ako je  $p$  veći broj, često se radi sekvencijalna pretraga, koriste se razni algoritmi, među kojima su i metaheuristički algoritmi i sl. [3]. Primjeri korištenja tak-

vih metoda su odabir unaprijed [46], eliminacija unatrag, rekurzivna eliminacija značajki u sklopu metode potpornih vektora (eng. *support vector machine recursive feature elimination* - SVM-RFE) [89], Boruta algoritam [86], rekurzivna eliminacija značajki [87], slučajne šume za odabir značajki [91] i mnogi drugi. Od metaheurističkih algoritama koriste se genetski algoritam, simulirano kaljenje, algoritam pretraživanja sove, optimizacija roja čestica, kukavice, leptira, vretenca, [71–78] i drugi.

Najčešća ugrađena metoda odabira značajki jest regularizacija koja nastoji minimizirati pogrešku procjene algoritma učenja i pritom prisiljava koeficijente uz značajke da budu mali (ili nula), što rezultira i modelom regularizacije i odabirom podskupa značajki [46]. Najpoznatije metode regularizacije su lasso, hrbatna i elastična mreža [50, 51].

Hibridne su metode skupina metoda odabira značajki koja, s obzirom na to da obuhvaća kombinacije prethodno spomenutih metoda, ima jako puno mogućih varijacija. Stoga su u nastavku spomenute neke kombinacije koje su se koristile u dosad spomenutim prethodnim istraživanjima - primjerice, kombinacije heurističkih algoritama za odabir značajki s regularizacijom [74], kombinacija slučajne šume i Relief-F algoritma [91] itd., no zaista je moguće mnogo kombinacija.

Bolón - Canedo i sur. [4] usporedili su u kontroliranom scenariju izvedbe nekoliko suvremenih metoda za odabir značajki. Analizirali su njihovu učinkovitost u rješavanju problema poput redundantnosti i nelinearnosti značajki, grešaka u mjerenju te rada s visokodimenzionalnim skupovima podataka. Ključni čimbenici usporedbe bili su brzina obrade i rizik od prenaučivosti. Uspoređujući tri temeljne vrste metoda za odabir značajki - metode filtriranja, metode omotača i ugrađene metode, Bolón - Canedo i sur. [4] zaključili su sljedeće:

- metode filtriranja su najbrže i nose najmanji rizik prenaučivosti
- metode omotača su najsporije i najviše izložene riziku prenaučivosti
- ugrađene su metode u sredini po pitanju brzine i rizika prenaučivosti.

Dakle, metode filtriranja zapravo su u tom kontekstu vodeće, no njima je zamjerka što odabrani podskup značajki ne mora biti optimalan u odnosu na metodu koja se poslije provodi. Iz toga razloga, u ovom će istraživanju od interesa biti upravo ugrađene metode odabira značajki, odnosno, regularizacijske metode. Osim toga, potencijal je vidljiv i u hibridnim metodama odabira značajki, posebno u smislu uključivanja znanja stručnjaka o domeni problema.

## 3.2. PRETHODNA ISTRAŽIVANJA U KONTEKSTU UKLJUČIVANJA STRUČNJAKA

Ponekad se istraživači nalaze u situaciji kada dobro poznaju domenu problema i posjeduju znanje o tome koju su značajke relevantne. Navedeno je iznimno bitno u kontekstu jake teorije, odnosno konceptualnih ili teorijskih temelja pri izboru značajki, ali i pogreške specifikacije koja se odnosi na isključivanje relevantne ili uključivanje irelevantne značajke u analizi, a koje spominju Hair i sur. [43]. Stoga je smisleno pokušati uključiti domensko znanje u metodu za odabir značajki ili model. U nastavku su navedena neka od takvih istraživanja.

Qazi i Wong [15] ističu kako temeljnim algoritmima strojnog učenja nedostaje način kako integrirati domensko znanje u njih, što je ključno u nekim domenama u kojima se donose odluke s ozbiljnim posljedicama, poput medicine. Smatraju da je takvo uključivanje domenskog znanja izazov koji će u konačnici doprinijeti razvoju učinkovitih modela za bolje donošenje odluka. U svom radu rudarenjem teksta najprije ispituju nestrukturirana izvješća kako bi izvukli asocijacije, a zatim predstavljaju model pretraživanja temeljen na asocijativnom ispitivanju. Model je integriran s klasteriranjem kako bi se razvila interaktivna shema otkrivanja znanja uz pomoć ljudi i rudarenja podataka. Predloženo otkrivanje znanja usmjereno na čovjeka i pristup rudarenja tekstualnih podataka može izvući vjerojatne veze i razne sličnosti u smislu kosinusne, Jacquardove i euklidske udaljenosti. Pristup dopušta analitičaru da dopuni svoje znanje o domeni, uključujući odabir funkcija sličnosti za identificiranje asocijacija, odabir dinamičkih značajki za interaktivno grupiranje i dodjeljivanje težine svakoj komponenti. Utvrđeno je da olakšava ljudsko rasuđivanje.

Micallef i sur. [16] predstavljaju pristup koji poboljšava točnost modela za predviđanje izvlačenjem prešutnog prethodnog znanja domenskih stručnjaka o relevantnosti značajki kroz interaktivnu vizualizaciju. Pritom se kreira korisnički model koji na temelju prethodnog unosa stručnjaka vodi odabir značajki od kojih će se sljedeće izvući korisnikovo znanje, a zatim koristi te informacije kao prethodno znanje za model predviđanja. Tvrde da interaktivno prikupljanje znanja o relevantnosti značajki poboljšava predviđanja u usporedbi s nasumično odabranim značajkama u malim skupovima podataka s brojem značajki većim od broja slučajeva.

Daee i sur. [17] također se dotiču takvih skupova podataka. Navode da je takvih skupova mnogo, primjerice u medicini, te da je u takvim situacijama kada je teško doći do novih po-

dataka, svaka dodatna informacija iznimno bitna. Kao vrijedan izvor informacija prepoznaju upravo domenske stručnjake te smatraju da njihovi potencijali nisu dovoljno iskorišteni. U svom radu do znanja dolaze vjerojatnosnim procesom zaključivanja, gdje se stručno znanje sekvencijalno ispituje s ciljem poboljšanja predviđanja. Nadalje, u specijalnom slučaju rijetke (eng. *sparse*) linearne regresije pretpostavljaju da stručnjak posjeduje znanje o relevantnosti značajki ili vrijednosti regresijskih koeficijenata te predlažu algoritam i računsku aproksimaciju za brzu i učinkovitu interakciju, koja sekvencijalno identificira najinformativnije značajke na temelju kojih se ispituje stručno znanje. Predložena metoda u eksperimentu sa simuliranim i stvarnim podacima pokazuje poboljšanu točnost predviđanja i uz ovakav "mali" napor stručnjaka.

Nadalje, i Akulich i sur. [18] tvrde da je odabir značajki u visokodimenzionalnim skupovima s ograničenim slučajevima ključan za stvaranje moćnog prediktivnog modela koji je ujedno i skalabilan i interpretabilan, odnosno transparentan i pouzdan. Istražuju najčešće korištene metode odabira značajki i širok raspon modela strojnog učenja te dizajniraju tri različita scenarija kako bi osigurali da je procjena modela robusna za primjenu razvijenih metodologija u stvarnom vremenu i kako bi otkrili skriveni učinak šuma (eng. *noise*) na konačni ishod. Procjena se provodi i za puni model i za reducirane modele, a na kraju za metode odabira značajki predlažu metriku ispravnosti za procjenu usklađenosti odabranog podskupa značajki s domenskim stručnjacima. Kao rezultat toga, regresija metodom potpunog vektora daje bolju točnost predviđanja i mogućnost generalizacije.

Visokodimenzionalnim skupovima podataka koje proizvode automatizirani sustavi bave se i Ringsquandl i sur. [19]. Oni, naime, tvrde kako su domenskim stručnjacima već poznate ovisnosti između mnogih slučajeva te predlažu pristupe sa semantičkim modelima koji imaju mogućnost ugradnje takvog domenskog znanja. U svom radu uvode dva semantički vođena pristupa odabiru značajki za različite scenarije podataka u sustavima industrijske automatizacije. Pokazalo se da spomenuti pristupi mogu svojim performansama konkurirati konvencionalnim pristupima, ili čak biti bolji u odnosu na njih.

Vandewiele [20] predlaže uključivanje dostupnog znanja u fazi predobrade, ekstrakcije značajki i odabiru procesa strojnog učenja s ciljem poboljšanja prediktivne performanse i vremena potrebnog za učenje modela. Tvrde da su modeli koji se temelje samo na podacima robusniji na stršće vrijednosti i šumove, no nedostatak im je relativno "skupo" vrijeme treniranja i nezimanje u obzir vrijednog prethodnog znanja. S druge strane, modeli koji su temeljeni samo na znanju, kao ekspertni sustavi, objašnjivi su, ali jako pristrani u odnosu na znanje te nemaju

mogućnost usvajanja novih znanja. U konačnici, potencijal za napredak unutar domene strojnog učenja vide u kombiniranju pristupa temeljenog na podacima s pristupom temeljenim na znanju u tzv. hibridni pristup, koji se ponekad naziva i semantičko meta-rudarenje (eng. *semantic meta-mining*). Smatraju da će takav pristup eliminirati spomenue nedostatke pojedinog pristupa i pritom zadržati njihove prednosti.

Jenul i sur. [21] predlažu Bayesov okvir za izbor značajki vođen korisnikom. To je pristup koji ugrađuje odabir značajki u Bayesov statistički okvir i koji uzima i podatke i domensko znanje kao izvor informacija. Najprije se iz podataka izgradi ansambl odabira značajki, a potom korisnik vodi Bayesov okvir za izbor značajki dodjeljivanjem težina (pondera) značajkama i penaliziranjem blokova ili kombinacija značajki implementiranih putem distribucije Dirichletovog tipa. Takav pristup predstavlja svojevrsni kompromis između znanja korisnika i promatranja podataka te postiže točne i robusne rezultate.

Liu i sur. [22] također u višeslojnu metodu odabira značajki uključuju domensko znanje stručnjaka. Spomenuto znanje kvantificira se pomoću ponderiranog bodovanja i integrira u proces odabira značajki kako bi se spriječio rizik uklanjanja bitnih značajki. Metoda rezultira manjim i razumljivijim skupom značajki, a istovremeno osigurava jednaku ili bolju točnost predviđanja u smislu korijena srednje kvadratne greške.

Domenski stručnjaci su i u radu [23] doprinijeli dodijelivši težine kriterijima i podkriterijima u metodama neizrazitog analitičkog hijerarhijskog procesa (eng. *fuzzy analytical hierarchy process* - FAHP) i adaptivnog neuro-neizrazitog sustava zaključivanja (eng. *adaptive neuro-fuzzy inference system* - ANFIS), što je rezultiralo učinkovitim i djelotvornim sustavom za potporu odlučivanju.

Zhao i sur. [24] kritiziraju klasične modele strojnog učenja koji koriste značajke s najvećom diskriminativnom moći smatrajući kako mogu biti nepouzdana, irelevantna ili neinterpretabilna jer ne uzimaju u obzir krajnjeg korisnika. Naglašavaju važnost prediktivnih modela koji uzimaju u obzir preferencije značajki korisnika, objašnjavajući da modeli trebaju biti pouzdani i interpretabilni korisnicima. Pritom navode da uvođenje preferencija korisnika u odabiru značajki u fazi predobrade nije dovoljno jer se tako model trenira s fiksnim podskupom značajki, što ne jamči očuvanje prediktivne moći metode. Stoga tvrde da bi se ta dva aspekta - uzimanje u obzir preferencija korisnika i učenje modela, trebali istovremeno promatrati te predlažu metodu koja poštuje preferencije krajnjeg korisnika dobivene putem relativnog sustava rangiranja i uvode regularizacijski član koji ih uključuje u funkciju cilja metode klasiifikacijskih i regresijskih stabala



(eng. *classification and regression trees* - CART). Time je poboljšana interpretabilnost metode, a smjer učenja je određen preferencijama korisnika. Metodologija je generička i primjenjiva na više vrsta postojećih modela budući da je u suštini varijacija regularizacije te ju je moguće uključiti u metode strojnog učenja s eksplicitnim objektivnim funkcijama. Regularizirani model potom se može optimizirati metodama numeričke optimizacije bez potrebe za stvaranjem posebno dizajniranih tehnika [24].

Huynh i sur. [25] također kritiziraju odabir značajki tradicionalnim metodama koje maksimiziraju ukupnu točnost klasifikacije. Naglašavaju da tako odabrane značajke nisu uvijek bitne za praktičnu primjenu te stoga predlažu da se uvijek dodatno provjeri s domenskim stručnjacima daju li te odabrane značajke smisleno znanje. U svom su istraživanju kombinirali model važnosti zadovoljstva (eng. *importance-satisfaction*) kao domensko znanje u metodu stabla odlučivanja.

Dostupno je još primjera istraživanja koja se ugradnjom domenskog znanja stručnjaka rezultirala točnijim modelima. Primjerice, Zhou i sur. [26] predlažu hibridno modeliranje kako bi se poboljšala točnost i stabilnost modela. Najprije predlažu metodu za ekstrakciju značajki nakon čega koriste neuronske mreže izgrađene na tim podacima. S druge strane, predlažu i metodu odabira značajki temeljenu na prethodnom poznavanju područja, stručnom iskustvu i rudarenju podataka te je izgrađen hibridni model. Rezultati simulacije pokazuju da uvođenje znanja o domeni i stručnog iskustva u modeliranje može poboljšati interpretabilnost modela, pojednostaviti proces oblikovanja i dodatno poboljšati performanse modela.

Ghiasi i sur. [27] naglašavaju da se ponekad ključne informacije koje predstavljaju stručno znanje ne mogu lako izdvojiti iz podataka, pogotovo ako se radi o malom skupu podataka ili skupu podataka sa šumovima. U svom radu istražuju prikladnost klasifikacije Gausovim procesom (eng. *Gaussian Process Classification*) za implementaciju domenskog znanja u fazi treniranja algoritma te pokazuju da je time dobiveno značajno poboljšanje modela.

Lohaj i sur. [28] ustanovili su da su stručnjaci precizniji u odabiru značajnih značajki od metoda odabira unaprijed (eng. *forward selection*), dok su metode objašnjive umjetne inteligencije (eng. *explainability artificial intelligence*) identificirale gotovo iste značajke kao stručnjaci. Navedeno su ocijenili prateći performanse krajnjih modela pomoću metrika preciznosti, točnosti i površine ispod ROC krivulje (eng. *area under the curve* - AUC).

Krishna i sur. [29] predstavili su novi pristup za učinkovit odabir značajki koji integrira znanje stručnjaka s uzajamnom informacijom (eng. *mutual information*), neizravnom logikom

(eng. *fuzzy logic*) te optimizacijom rojem čestica (eng. *particle swarm optimization*). Simulacijama su pokazali da predloženi pristup rezultira točnijim klasifikacijskim modelom u odnosu na postojeće metode u smislu osjetljivosti, specifičnosti i površine ispod ROC krivulje (AUC).

Nadalje, Pienaar i sur. [30] tvrde kako je pri razvoju modela strojnog učenja ključno integrirati ljudsko znanje i postići kompromis između izvedbe modela, sveobuhvatnosti i istinitosti te praktičnosti korištenja. U svom su istraživanju do domenskog znanja stručnjaka došli pretraživanjem literaure i Delfi metodom (eng. *Delphi method*) te su pokazali kako je to značajno doprinijelo specifikaciji problema i odabiru značajki prije predobrade i razvoja modela.

Nadalje, i Chen i sur. [31] naglašavaju važnost domenskog znanja stručnjaka u poboljšavanju sustava umjetne inteligencije integracijom domenskog znanja, prikupljanjem podataka i odabirom značajki. Tvrde da se time poboljšava otkrivanje znanja iz podataka, a također i interpretabilnost i pouzdanost metode. U svom se istraživanju fokusiraju na prijenos znanja od ljudskih stručnjaka do nadziranog problema učenja učenjem znanja domene kao interpretabilnih značajki i pravila, koja se mogu koristiti za konstruiranje sustava temeljenih na pravilima za podršku donošenju odluka. U svom istraživanju predlažu interpretabilni model neuronske mreže temeljen na pravilima iz znanja domene, čime se poboljšava točnost predviđanja.

Važnost integriranja domenskog znanja stručnjaka u metode strojnog učenja, što može poboljšati performanse modela za predviđanja, ističu i druga istraživanja [32–36].

S druge strane, Burns i Danyluk [37] kritiziraju sustave za potporu odlučivanju, tj. ekspertne sustave, koji se sastoje samo od znanja stručnjaka. Navode kako je njihova točnost jako ovisna o točnosti znanja stručnjaka te da ljudske teorije obično sadrže greške ili su nepotpune. Smatraju da podskup značajki koje odabere stručnjak ne mora biti optimalan te da sustavi strojnog učenja trebaju automatski poboljšati i ispraviti teoriju o domeni. Nadalje, tvrde da kad se sustavi revizije teorije primjenjuju na ekspertne teorije, oni se često koncentriraju na preformulaciju pruženog znanja, a ne na preformulaciju ili odabir ulaznih značajki, stoga u svom radu proučavaju usavršavanje teorije i relativne prednosti primjene odabira značajki u odnosu na opsežniju reformulaciju teorije. Pokazali su da su i odabir značajki i reformulacija teorije učinkoviti u poboljšavanju neuronskih mreža temeljenih na znanju.

Nadalje, Sebastiani [38] uspoređuje automatizirani pristup klasifikacije teksta temeljen na metodama strojnog učenja s pristupom inženjeringa znanja koji se temelji na znanju stručnjaka. Kao prednosti automatiziranog pristupa navode vrlo dobru učinkovitost, mogućnost generalizacije na druge domene te značajnu uštedu u smislu radne snage stručnjaka.

Osim toga, Thiam i sur. [39] tvrde da se standardni inženjering značajki koji se provodi ručno temeljem domenskog znanja može učinkovito zamijeniti pristupima dubokog učenja (eng. *deep learning*) koji integriraju inženjering i odabir značajki te optimizaciju modela u jedan proces učenja. S tim u vezi, predlažu autonomno učenje značajki temeljeno na dubokim neuronskim mrežama i tvrde da rezultati pokazuju relevantnost takvog pristupa, koji ujedno nude veću fleksibilnost u smislu transfernog učenja.

### 3.3. PRETHODNA ISTRAŽIVANJA U KONTEKSTU ENERGETIKE

U nastavku je dan pregled rezultata nekih od radova koji se odnose na energetske troškove, potrošnju ili učinkovitost građevinskog sektora i to u kontekstu njihovog modeliranja i procjene ili predviđanja, kao i u kontekstu redukcije dimenzionalnosti.

Istraživači su se u ovom kontekstu najčešće bavili potrošnjom energije ukupnom ili po pojedinom energentu [92–99], emisijom ugljikovog dioksida [100,101], troškovima energije [84,87] i energetske intenzitetom [102]. Pri tome su promatrali vremenske nizove i autoregresivne vrijednosti ili su u obzir uzeli i veze s drugim značajkama [92, 103, 104]. Osim vremenskih nizova promatrali su se i 'statički' podaci, pritom analizirajući veze značajkama poput temperature, građevinskih podataka, podataka o korištenju i slično [82, 84, 86, 87, 105].

Uočeno je da se najviše istraživanja provodilo metodama poput metode potpunih vektora [92], neuronskih mreža [82, 84, 86, 87, 92, 93, 98, 105, 106], stabla odlučivanja [82, 86, 92, 94, 98], slučajnih šuma [82, 86, 94, 95, 99], višestruke linearne regresije s regularizacijom ili bez [9, 84, 95, 96, 98–100, 102, 104, 107, 108]. Koristili su se i modeli poput stohastičkog utjecaja regresije na stanovništvo, blagostanje i tehnologiju (eng. *Stochastic Impacts by Regression on Population, Affluence, and Technology - STIRPAT*) [100] i druge. Metode su najčešće bile provođene na cijelom uzorku, ali i na uzorku podijeljenom u klastere kao u [103, 106] ili u studijima slučaja. Naime, neki se autori koristili simulacije poput Monte Carlo simulacija [101], ali gotove alate za simulacije raznih scenarija i promatranje njihovih učinaka na potrošnju energije. Radilo se o promatranju učinka promjena u manjem broju značajki za koje je "poznato" da utječu na potrošnju kroz studije slučaja. To su alati *Energy Plus Software* [109–113], *Building Information Modeling* [114–116], *Green Building Studio* [115], *ECO.AP* [115], *Energy*

*Visualization Tool* [93], *Energy Plan* [117] i *Urban building energy modelling* [118]. Osim prediktivnih i simulacijskih modela, koristili su se i razni optimizacijski algoritmi, npr. evolucijski i genetski za optimizaciju višestrukih kriterija [106, 119, 120].

Ono što je bitno za naglasiti jest da je samo manji dio ovih radova provodio redukciju dimenzionalnosti, dok je većina radova koristila po nekoliko pažljivo odabranih značajki. U istraživanjima koja ipak jesu provodila redukciju dimenzionalnosti, koristili su se odabir značajki metodama filtriranja temeljenim na statističkim kriterijima [82, 86, 87, 107], rekurzivna eliminacija značajki u sklopu slučajne šume [87], lasso i hrbatna regularizacija te regularizacija elastičnom mrežom [95, 100] i Boruta algoritam u sklopu slučajnih šuma [86]. Nadalje, koristio se i odabir značajki temeljem znanja stručnjaka [94, 107], ali i metode ekstrakcije značajki poput analize nezavisnih komponenti, analize glavnih komponenti i faktorske analize (npr. u [9, 108]) i drugi. Mjere koje su pritom koristili za vrednovanje i usporedbu dobivenih modela su bile koeficijent determinacije  $R^2$  [93, 95, 100, 102, 104, 107] i razne greške (*SMAPE* [82, 84, 86, 87, 94], *NRMSE* [82, 84, 86, 87, 92], *RMSE* [93], *MSE* [93, 105], *MAE* [95], *MAD* [93], *NMBE* [92]).

### **Dosadašnja istraživanja koja uključuju domensko znanje stručnjaka**

Pojedini radovi također integriraju znanje stručnjaka s ciljem dizajniranja sustava za nadzor energije koji se može konfigurirati jednom, a zatim imati široku primjenu uz malo dodatnih troškova ili truda [121]. U tu svrhu, Hu i sur. [121] predložu metodologiju koja iskorištava postojeće senzorne podatke, kodira stručno znanje o aplikacijskom sustavu za stvaranje 'virtualnih senzora' i primjenjuje statističke i matematičke metode za smanjenje vremena potrebnog za ručne konfiguracije. Pristup predstavlja metodu koja otkriva greške iz izmjerenih energetskih podataka, ali i opisuje metodu angažmana sa stručnjacima u domeni energetske sustava za identifikaciju podataka, integrira domensko znanje s podacima, vrši automatski odabir iz velikog skupa potencijalnih ulaznih podataka i koristi strojno učenje za automatsku izgradnju modela detekcije kvara vođenog podacima. Evaluacija na rashladnom uređaju komercijalne zgrade pokazuje da je samo mali broj virtualnih senzora neophodan za detekciju greške s visokim stopama točnosti.

Današnje zgrade imaju složenu tehničku opremu kojom nije jednostavno upravljati te, iz tog razloga, korisnici zgrada trebaju smjernice u smislu pronalaženja najučinkovitijih strategija djelovanja kojima bi se izbjegle greške i rasipanje energije [122]. Iz tog razloga, Pruvost i sur. [122] predlažu ekspertni sustav koji tijekom faze konfiguracije počinje od semantičkog opisa

energetskog sustava zgrade kako bi identificirao potencijalne rizike učinkovitosti i automatski primijenio generičke funkcije nadzora i pravila za rad zgrade, a zatim u operativnoj fazi generira savjete na temelju kontinuirane analize operativnih podataka zgrade u stvarnom vremenu.

U svrhu poboljšanja energetske fleksibilnosti zgrade, Zhou i sur. [123] predstavljaju novi okvir koji integrira strojno učenje (metodu klasifikacijskog i regresijskog stabla te učenje pojačanjem (eng. *reinforcement learning*) i ekspertni sustav temeljen na pravilima, odnosno domenskom znanju. Učinkovitost predloženog okvira procijenjena je pomoću četverogodišnjih podataka prikupljenih iz stvarne poslovne zgrade s nultom neto energijom, a rezultati su pokazali da je ovakva kombinacija metoda uspješnije smanjila troškove energije u odnosu na korištenje samo ekspertnog sustava temeljenog na pravilima.

Nadalje, Guzhov i sur. [107] radili su procjenu potrebne količine toplinske energije, odnosno energetske bilance, za gradske četvrti s dotrajalim zgradama uz korištenje statističke metode obrade informacija, koja je, među ostalim koristila domensko znanje stručnjaka. Metoda se sastojala od sljedećih koraka: (1) analiza podataka i filtriranje nepouzdanih i nedostajućih podataka, (2) određivanje stvarne i "idealne" proizvodnje toplinske energije za promatrani sustav opskrbe toplinom, (3) određivanje čimbenika koji utječu na vrijednosti potrošnje topline temeljem mišljenja stručnjaka i filtriranje nekih pomoću kriterija kolinearnosti i multikolinearnosti, (4) formulacija matematičke funkcije predikcije ili skupine prediktivnih pravila, (5) određivanje vrijednosti svih ulaznih čimbenika i (6) određivanje prognoziranih vrijednosti potrošnje toplinske energije [107].

Napoli i sur. [124] promiču održive gradove i obnovu građevinskog sektora kao značajnog potrošača energije. Spominju različite aspekte poput energetske učinkovitosti, financijsko - ekonomske izvedivosti i zaštite okoliša kao bitne pri planiranju provedbe mjera obnove i korištenja financijskih sredstava. S tim u vezi, predlažu model za donošenje odluka u slučaju višekriterijskog odlučivanja koji uključuje metodu ELECTRE TRI-nC, a skupina stručnjaka uključena je u kreiranje problema odlučivanja i raspravu o rezultatima u stvarnoj studiji slučaja [124].

Jung i Awad [125] također koriste znanje stručnjaka u sklopu metode višekriterijskog odlučivanja. Oni, naime, u cilju planiranja pohrane obnovljive energije, imaju namjeru identificirati komponente urbanog planiranja, pritom zadržavajući najviše standarde zelene ekonomije i ekološke održivosti. Najprije su iz postojećih istraživanja izdvojili elemente održivog urbanog planiranja te su potom, pomoću stručnjaka, hijerarhijski organizirali te elemente i analizirali rezultate pomoću analitičkog hijerarhijskog procesa (AHP) [125].

Može se zaključiti da postoji mnogo različitih pristupa u modeliranju podataka vezanih za energiju i građevinski sektor, a svrha svih prethodno spomenutih modela i algoritama je bila potpora pri odlučivanju i planiranju provođenja mjera energetske obnove / gradnje građevina i zaštite okoliša. Također, postoji potencijal u uključivanju domenskog znanja stručnjaka u razvoju metodologije za učinkovitije upravljanje energijom i zgradarstvom, pa tako i ovo istraživanje uključuje znanja stručnjaka s ciljem doprinosa, ne samo točnosti, nego i pouzdanosti i interpretabilnosti modela krajnjim korisnicima.

## 4. METODOLOGIJA

Temeljna je ideja ovog istraživanja poboljšanje uspješnosti predikcije linearnog regresijskog modela korištenjem domenskog znanja stručnjaka, stoga je u nastavku rada pojašnjeno kako je to izvedeno.

Općenito, prilikom regularizacije, u već prikazane izraze 2.4, 2.6 i 2.7 u potpoglavlju 2.3 moguće je uvesti penalizacijski koeficijent koji dodatno stoji uz ulazne značajke. Na taj način, specificiranjem penalizacijskog koeficijenta moguće je utjecati na učenje regularizacijske metode, koja osim što rezultira modelom, rezultira i odabranim značajkama. Navedena karakteristika poslužila je kao temeljna ideja koja se u ovom istraživanju koristi u svrhu razvoja metode za odabir značajki u linearnom regresijskom modelu koja u metodu regularizacije uključuje domensko znanje stručnjaka.

U nastavku rada specificirana je metodologija uvođenja domenskog znanja stručnjaka u regularizacijske metode. Osim toga, budući da su u svrhu demonstracije rada i evaluacije rezultata predložene metode korištene dvije vrste podataka: (1) stvarni podaci o troškovima energije zgrada javnog sektora Republike Hrvatske i njihovim karakteristikama te (2) simulirani podaci iz nekoliko scenarija proizašlih iz literature, u nastavku su pojašnjene i obje vrste podataka te način na koji je dobiveno domenskog znanja kod svake od njih.

### 4.1. UVOĐENJE DOMENSKOG ZNANJA STRUČNJAKA U REGULARIZACIJU

U ovom se istraživanju domensko znanje stručnjaka ugradilo u regularizaciju uvođenjem penalizacijskog koeficijenta čija vrijednost zavisi o mišljenju stručnjaka o utjecaju te značajke na izlaznu značajku. Pritom su moguće četiri kategorije utjecaja ulaznih značajki na izlaznu: *Nema utjecaja*, *Slab utjecaj*, *Srednji utjecaj* i *Jak utjecaj* te, posljedično, moguće četiri različite

vrijednosti penalizacijskog koeficijenta.

Pri korištenju funkcija iz paketa **glmnet** u R-u, može se specificirati vrijednost penalizacijskog koeficijenta. Pritom vrijedi sljedeće: 1 je zadana vrijednost, vrijednost 0 znači da za pojedinu značajku nema skupljanja (pritezanja) i da je ona uvijek uključena u model, a vrijednosti penalizacijskog koeficijenta se normaliziraju da im suma bude jednaka broju značajki.

U nastavku se, radi jednostavnosti, uvode oznake  $\gamma_c, c = 1, \dots, 4$ , za penalizacijske koeficijente značajki kategorije utjecaja 1: *Nema utjecaja*, 2: *Slab utjecaj*, 3: *Srednji utjecaj* i 4: *Jak utjecaj*, redom. Osim toga, neka  $\mathcal{C}_c, c = 1, \dots, 4$ , predstavljaju skupove indeksa za značajke u svakoj kategoriji utjecaja (npr. ako su značajke  $X_1$  i  $X_5$  kategorije utjecaja 1, tj. *Nema utjecaja*, tada je  $\mathcal{C}_1 = \{1, 5\}$ ).

Procjena koeficijenata modela u regularizaciji sada se može formalizirati po uzoru na formule (2.4), (2.6) i (2.7) (str. 10 - 11) na sljedeći način:

$$\hat{\beta}^{lassopen.} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \sum_{c=1}^4 \gamma_c \lambda \sum_{j \in \mathcal{C}_c} |\beta_j| \right\}, \quad (4.1)$$

$$\hat{\beta}^{ridgepen.} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \sum_{c=1}^4 \gamma_c \lambda \sum_{j \in \mathcal{C}_c} \beta_j^2 \right\}, \quad (4.2)$$

$$\hat{\beta}^{el.netpen.} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \sum_{c=1}^4 \gamma_c \left( \lambda_1 \sum_{j \in \mathcal{C}_c} \beta_j^2 + \lambda_2 \sum_{j \in \mathcal{C}_c} |\beta_j| \right) \right\}. \quad (4.3)$$

Dakle, svim značajkama koje prema domenskom znanju stručnjaka pripadaju jednakoj kategoriji utjecaja, dodjeljuje se jednaka vrijednost penalizacijskog koeficijenta. Na taj se način želi učenje metode usmjeriti domenskim znanjem stručnjaka s ciljem dobivanja modela koji su interpretabilni i pouzdani krajnjim korisnicima.

No, postavlja se pitanje kako procijeniti vrijednosti penalizacijskog koeficijenta  $\gamma_c$ ? U ovom se istraživanju razlikuju dva načina; najprije su se vrijednosti odredile deterministički, a potom se procjena vrijednosti penalizacijskog koeficijenta promatrala kao optimizacijski problem čija su se rješenja tražila metaheurističkim algoritmom.

#### 4.1.1. Deterministička procjena penalizacijskog koeficijenta

U ovom slučaju, penalizacijski koeficijent određen je na način da su se unaprijed određene vrijednosti i veće vrijednosti koeficijenta dale značajkama za koje su stručnjaci smatrali da



imaju manji utjecaj na izlaznu značajku i obratno. Dakle, najmanju vrijednost imaju ulazne značajke koje po mišljenju stručnjaka pripadaju kategoriji *Jak utjecaj*, nešto veću oni koji imaju *Srednji utjecaj*, još veću oni koji imaju *Slab utjecaj*, a najveću vrijednost oni prediktori koji su kategorije *Nema utjecaja*. Konkretno, tablica 4.1 prikazuje vrijednosti korištene za svaku skupinu ulaznih značajki prema utjecaju koji su im dodjelili stručnjaci.

Tablica 4.1: Vrijednosti penalizacijskog koeficijenta u slučaju determinističke procjene

Koeficijent	Vrijednost
$\gamma_1$	$1 \eta$
$\gamma_2$	$0.5 \eta$
$\gamma_3$	$0.2 \eta$
$\gamma_4$	$0.05 \eta$

$\gamma_1, \gamma_2, \gamma_3$  i  $\gamma_4$  penalizacijski su koeficijenti uz značajke kategorije *Nema utjecaja*, *Slab utjecaj*, *Srednji utjecaj* i *Jak utjecaj*, redom.

Vrijednost  $\eta$  se određuje tako da zbroj penalizacijskih koeficijenata bude jednak broju stupaca matrice dizajna, bez slobodnog koeficijenta  $\beta_0$ .

Može se uočiti da se značajkama koje su kategorije *Jak utjecaj* nije dodijelila vrijednost 0, što bi značilo da su uvijek u modelu, nego im se dodijelila mala vrijednost penalizacijskog koeficijenta (0.05) koja sugerira da se značajka manje priteže i na taj način usmjerava učenje metode, ali metoda ipak može sama napraviti odabir značajki.

#### 4.1.2. Procjena penalizacijskog koeficijenta metaheurističkim algoritmom

Postavlja se pitanje postoji li neki drugi način koji će bolje odabrati te vrijednosti uzevši u obzir karakteristike skupa podataka te prirodu problema. S tim u vezi, problem određivanja vrijednosti penalizacijskog koeficijenta promatrao se kao optimizacijski problem, pri čemu se radi o kontinuiranoj optimizaciji (izrazi (4.1) - (4.3) na str. 38). Metaheuristički algoritam koji se koristio u tu svrhu je algoritam pretraživanja kukavice, opisan u potpoglavlju 2.5 pseudokodom 1. Detaljnije o samim postavkama metaheuristike u R-u je opisano u dijelu 4.6 ovog poglavlja.

## 4.2. OPIS STVARNIH PODATAKA

U ovom dijelu istraživanja, u fokusu su zgrade javnog sektora Republike Hrvatske. Općenito, prema [126], zgrade javnog sektora obuhvaćaju "... zgrade centralne države (ministarstva i tijela državne uprave) te zgrade jedinica lokalne i područne (regionalne) samouprave, odnosno zgrade u vlasništvu javnog sektora, u kojima se obavljaju društvene djelatnosti (odgoja, obrazovanja, prosvjete, znanosti, kulture, sporta, zdravstva i socijalne skrbi), djelatnosti državnih tijela i organizacija kao i tijela i organizacija lokalne i područne (regionalne) samouprave, djelatnosti pravnih osoba s javnim ovlastima, zatim zgrade za stanovanje zajednica, zgrade udruga građana i vjerskih zajednica...".

Podaci su sekundarni; preuzeti od Agencije za pravni promet i posredovanje nekretninama, Centra za praćenje poslovanja energetskog sektora i investicija i Fonda za zaštitu okoliša i energetske učinkovitost i integrirani za potrebe istraživanja projekta HRZZ pod nazivom „Metodološki okvir za učinkovito upravljanje energijom s pomoću inteligentne podatkovne analitike“ [7]. Skup podataka činila je 1421 zgrada javnog sektora Republike Hrvatske, a odnosio se na:

- godišnji trošak energije zgrada javnog sektora Republike Hrvatske (izlazna značajka)
- 25 karakteristika zgrada kao što su korištenje i namjena zgrade, konstrukcijske i energetske karakteristike, karakteristike grijanja i hlađenja, geografske i druge (ulazne značajke).

Godišnji trošak energije zgrada javnog sektora Republike Hrvatske dobiven je kao prosječni godišnji trošak energije za period 2018., 2019. i 2020. godine, a podaci o karakteristikama zgrade preuzeti su 2021. godine. Detaljnije o značajkama navedeno je u tablici 4.2, a deskriptivna analiza značajki dana je u tablicama 5.1 i 5.2 u poglavlju **Rezultati istraživanja**.

Tablica 4.2: Naziv, vrsta i opis značajki

Naziv	Vrsta	Opis
REALCOSTTAX	Kvant.	Godišnji trošak energije [ <i>kn</i> ].
NUMBEROFEMPLOYEES	Kvant.	Broj zaposlenika.
NUMBEROFUSERS	Kvant.	Broj korisnika.

Naziv	Vrsta Opis
HEATEDSURFACEOF-THEBUILDING	Kvant. Oplošje grijanog dijela zgrade A [ $m^2$ ].
HEATEDVOLUMEAREA-OF-THEBUILDING	Kvant. Bruto obujam grijanog dijela zgrade oplošja A [ $m^3$ ].
NUMBEROFFLOORS	Kvant. Broj etaža.
H1TRND	Kvant. Najveći dopušteni koeficijent transmis. topl. gubitka po jedinici oplošja grijanog dijela zgrade, H <tr,nd [<math="">W/m^2K].</tr,nd>
SHAREOFWINDOWSSURFACE	Kvant. Udio ploštine prozora u ukupnoj ploštini pročelja [%].
TOTALPOWERBODYHEATRADIATOR	Kvant. Ukupno instalirana snaga ogrjevnih radijatora [ $kW$ ].
TOTALPOWERBODYOTHER	Kvant. Ukupno instalirana snaga ostalih ogrjevnih tijela [ $kW$ ].
TITHERMALPOWEROFHEATERS	Kvant. Ukupna instalirana toplinska snaga ogrjevnih tijela [ $kW$ ].
TOTALHEATINGPOWER	Kvant. Ukupna toplinska snaga [ $kW$ ].
TICOOLINGPOWEROF-COOLERS	Kvant. Ukupna instalirana rashladna snaga rashladnih tijela [ $kW$ ].
LIGHT_INSTALLED_INT_PWR_KW	Kvant. Ukupna instalirana snaga unutarnje rasvjete [ $kW$ ].
TIPOWEROFFOFFICE-EQUIP	Kvant. Ukupna instalirana električna snaga uredske opreme [ $kW$ ].
TIPOWEROFOTHER-CONSUMERS	Kvant. Ukupna instalirana električna snaga ostalih potrošača [ $kW$ ].
OTHER_TOTAL_INSTALLED_PWR_KW	Kvant. Ukupna instalirana snaga ostalih sustava [ $kW$ ].
OBJ_CON_THICKNESS_D3	Kvant. Prozori - debljina građevnog dijela [ $cm$ ].
OBJ_CON_THICKNESS_D7	Kvant. Vanjski zid - debljina građevnog dijela [ $cm$ ].

Naziv	Vrsta Opis
Q1HNDDOP	Kvant. Najveća dopuštena godišnja potrebna toplinska energija za grijanje nestambene zgrade, $Q^*H_{nd,dop}$ [ $kWh/m^3a$ ].
OBJ_REG	Kvalit. Klimatski podaci; kategorije: 1 (primorska Hrvatska), 2 (kontinentalna Hrvatska).
OBJ_OBJT	Kvalit. Vrsta objekta; kategorije: 1 (administrativne), 2 (kulturne), 3_18 (obrazovne s internom temp.18), 3_20 (obrazovne s internom temp. 20), (opće), 8_22 (socijalne s internom temp. 22), 11 (zdravstvene).
YEAROFCOMPLETION-OFCONSTR	Kvalit. Godina završetka izgradnje; kategorije: 1 (> 2000), 2 (1991-2000), 3 (1981-1990), 4 (1971-1980), 5 (1946-1970), 6 (1919-1945), 7 (< 1919).
YEAROFLASTRESTORATION	Kvalit. Godina zadnje obnove; kategorije: 0 (zgrade koje u periodu od 2014., kada su krenule energetske obnove nisu imale obnovu), 1 (zgrade koje su imale obnovu od 2014. ili nadalje).
MODEOFSUPPLYOFDRINKINGWATER	Kvalit. Način opskrbe pitkom vodom; kategorije: 1 (vodovod / komunalno / vodoopskrba), 2 (ostalo).
OBJ_EXTENDED-_GEO_TYPE	Kvalit. Kategorija objekta; kategorije: 1 (samostalna cjelina), 2 (zgrada u kompleksu), 3 (kompleks), 4 (dio zgrade).

Kratice Kvant. i Kvalit. predstavljaju kvantitativne, odnosno kvalitativne vrste značajki.

Grafički su prikazi distribucija značajki radi preglednosti dani u Prilogu 1 (slike 7.1-7.4, str. 166-169). Na istima je uočljivo kako je distribucija svih kvantitativnih značajki, izuzev Q1HNDDOP, desno iskošena, što je provjereno računanjem koeficijenta asimetrije (eng. *skewness*). Vrijednosti koeficijenta asimetrije svih kvantitativnih značajki prikazane su u tablici 5.1 u poglavlju [Rezultati istraživanja](#) (str. 56). U navedenoj tablici vidljivo je kako su se kretale vrijednosti koeficijenta asimetrije; asimetrija je pozitivna kod svih značajki, a kod dosta značajki radilo se o vrlo jakoj asimetriji. Kod svih kvantitativnih značajki, izuzev Q1HNDDOP, koeficijent asimetrije bio je veći od 1. Nadalje, budući da prema [43] logaritamska transformacija obično dobro funkcionira za pozitivno asimetrične distribucije, navedene su značajke prilikom predobrade transformirane logaritamskom transformacijom, točnije transformacijom  $x \mapsto \log_{10}(1+x)$  zbog prisutstva vrijednosti 0 (vidljivo u tablici 5.1). Grafički prikazi distribu-

cija značajki nakon transformacije također su vidljivi u Prilogu 1 (slike 7.1-7.4, str. 166-169).

#### 4.2.1. Domensko znanje stručnjaka u slučaju stvarnih podataka

S ciljem poboljšanja regularizacijskih metoda koje, osim što rezultiraju modelima – u ovom slučaju za predviđanje godišnjeg troška energije, rezultiraju i odabirom značajki, u ovom istraživanju će se u njih pokušati ugraditi domensko znanje stručnjaka. U ovom kontekstu, stručnjacima su se smatrale osobe građevinske struke s iskustvom u energetske učinkovitosti i zgradarstvu te je, tijekom travnja 2022. godine, ispitano četvero stručnjaka.

Tablica 4.3: Opis iskustva stručnjaka

Stručnjak	Iskustvo
Stručnjak 1	Izvanredni profesor jednog građevinskog fakulteta. Nastavne aktivnosti uključuju rad na kolegijima povezanim sa sektorom zgradarstva i energetske učinkovitosti, dok stručne aktivnosti uključuju energetske preglede i certificiranje zgrada.
Stručnjak 2	Tehnički savjetnik, DGNB konzultant za održivu gradnju.
Stručnjak 3	Asistent jednog građevinskog fakulteta. Završio program osposobljavanja i usavršavanja za energetske preglede i energetske certificiranje zgrada te sudjelovao u preko 20 energetskih pregleda javnih zgrada.
Stručnjak 4	Asistent jednog građevinskog fakulteta.

*Napomena autora:* DGNB je kratica za Njemačko društvo za održivu gradnju (njem. *Deutschen Gesellschaft für Nachhaltiges Bauen*) koje predstavlja najveću europsku mrežu za održivu gradnju [127].

Spomenuto je istraživanje dio istraživačkog projekta Hrvatske zaklade za znanost [7] čiji je projektni tim bio sastavljen od istraživača iz interdisciplinarnih područja (ekonomija, informacijske znanosti, matematika, građevina i strojarstvo). Jedan od članova tima iz područja građevine, koji se u svojim nastavnim i stručnim aktivnostima bavi zgradarstvom i energetskim certificiranjem, može se smatrati domenskim stručnjakom. Nadalje, navedeni član tima zamoljen je za kontakte dodatnih osoba koje su također stručne u tom području te je na taj način dobiveno ukupno osam kontakata. Tih osam osoba potom su zamoljene da popune upitnik te da isti proslijede dodatnim osobama koje se mogu smatrati stručnjacima. Na kraju, na upitnik je

stiglo četiri odgovora stoga se može reći da su za potrebe ovog istraživanja ispitana četiri stručnjaka iz područja energetike i građevinarstva (zgradarstva). U tablici 4.3 pojašnjeno je iskustvo navedenih stručnjaka s tim područjem.

U svrhu dobivanja domenskog znanja stručnjaka, sastavljen je upitnik koji je sadržavao nazive i opise 25 značajki, a stručnjaci su zamoljeni da svakoj značajki dodijele vlastitu procjenu njezinog utjecaja na godišnji trošak energije u okviru jedne od četiri ponuđene kategorije – *Nema utjecaja*, *Slab utjecaj*, *Srednji utjecaj* i *Jak utjecaj*. Iz ovoga je vidljivo kako je mišljenje stručnjaka o utjecaju svake značajke na godišnji trošak energije predstavljeno značajkom s četiri moguće kategorije koja ima ordinalnu ljestvicu.

Konzistentnost mišljenja stručnjaka provjerena je primjenom sljedećih mjera slaganja i povezanosti: kapa koeficijent slaganja temeljem modela za ordinalne značajke ( $\kappa_m$ ), kapa koeficijent povezanosti temeljem modela za ordinalne značajke ( $\kappa_{ma}$ ), intraklasni koeficijent korelacije 1 (*ICC1*), intraklasni koeficijent korelacije 2 (*ICC2*), intraklasni koeficijent korelacije 3 (*ICC3*), Kendallov koeficijent podudarnosti  $W$  i Krippendorffov koeficijent pouzdanosti  $\alpha$ . Mišljenja stručnjaka koja nisu bitno odstupala agregirana su na način da za svaku značajku medijan mišljenja stručnjaka predstavlja domensko znanje, dok se mišljenje stručnjaka koje bitno odstupa uvrštava zasebno.

Podaci su prikupljeni samostalno od strane autora za potrebe ovog istraživanja i radi se o primarnim podacima, a rezultati su prikazani u poglavlju *Rezultati istraživanja* (str. 56).

### 4.3. OPIS SIMULIRANIH PODATAKA

Kako bi se omogućila generalizacija rezultata, predložena je metoda evaluirana pomoću simulacijskog eksperimenta. Sličan pristup korišten je i u drugim istraživanjima [47, 50–52]. Provedena su tri eksperimenta na scenarijima iz literature [50, 51].

Prije svega, pretpostavimo da se polazi od linearnog regresijskog modela sličnog kao u (2.2), odnosno:

$$Y = X\beta + \sigma\varepsilon$$

gdje je  $Y$  izlazna značajka,  $X_1, \dots, X_p$  su ulazne značajke ili prediktori, a izraz  $\sigma\varepsilon$  predstavlja grešku modela pri čemu je  $\varepsilon \sim \mathcal{N}(0, 1)$ , a  $\sigma$  je neki skalar. Scenariji se razlikuju u broju ulaznih značajki  $p$ , koeficijentima  $\beta$ , matrici korelacija ulaznih značajki *corr*, te standardnoj devijaciji greške modela koja iznosi  $\sigma$ . Vrijednosti pojedinih parametara za svaki scenarij su

navedene u nastavku. Za svaki scenarij je generirano 50 skupova podataka s 1000 slučajeva.

Prvi scenarij:

$$\begin{aligned}
 p &= 8, \\
 \beta &= (3, 1.5, 0, 0, 2, 0, 0, 0), \\
 \text{corr}(X_i, X_j) &= 0.5^{|i-j|}, \forall i, j, \\
 \sigma &= 3.
 \end{aligned}$$

Drugi scenarij:

$$\begin{aligned}
 p &= 40, \\
 \beta &= (\underbrace{0, 0, \dots, 0}_{10}, \underbrace{2, 2, \dots, 2}_{10}, \underbrace{0, 0, \dots, 0}_{10}, \underbrace{2, 2, \dots, 2}_{10}), \\
 \text{corr}(X_i, X_j) &= 0.5, \forall i, j, i \neq j \\
 \sigma &= 15.
 \end{aligned}$$

Treći scenarij:

$$\begin{aligned}
 p &= 40, \\
 \beta &= (\underbrace{3, 3, \dots, 3}_{15}, \underbrace{0, 0, \dots, 0}_{25}), \\
 \sigma &= 15, \\
 X_i &= Z_1 + \varepsilon_i^x, i = 1, \dots, 5, \\
 X_i &= Z_2 + \varepsilon_i^x, i = 6, \dots, 10, \\
 X_i &= Z_3 + \varepsilon_i^x, i = 11, \dots, 15, \\
 Z_1, Z_2, Z_3 &\sim \mathcal{N}(0, 1), \\
 \varepsilon_i^x &\sim \mathcal{N}(0, 0.01), i = 1, \dots, 15, \\
 (\varepsilon_i^x \text{ su nezavisne i jednako distribuirane [n.j.d.]}) \\
 X_i &\sim \mathcal{N}(0, 1), \text{ n.j.d., } i = 16, \dots, 40.
 \end{aligned}$$

U trećem scenariju postoje tri podjednako važne grupe od 5 značajki te 25 šumova. Poželjno bi bilo da metoda izdvoji 15 značajki, a koeficijenti uz šumove da budu jednaki nuli [50].

### 4.3.1. Domensko znanje stručnjaka u slučaju simuliranih podataka

U slučaju simuliranih podataka, pri formiranju mišljenja stručnjaka koristi se koncept veličine učinka. Naime, Kolesarić i Tomašić Humer [128] navode da je veličina učinka procjena stupnja u kojem je neki fenomen prisutan u populaciji. Pritom ne samo da daje informaciju postoji li neki učinak, nego daje i informaciju koliki je on, što je ideja koja se koristi u ovom istraživanju.

Nadalje, navode dvije vrste pokazatelja veličine učinka. Prva vrsta pokazatelja temelji se na razlici između statističkih vrijednosti rezultata (npr. aritmetičkih sredina), a jedan od najčešće korištenih pokazatelja je Cohenov  $d$  indeks [128]. S druge strane, druga vrsta pokazatelja veličine učinka temelji se na asocijaciji, odnosno korelaciji značajki, čime se predstavlja procjena veličine zajedničke varijance promatranih značajki. Pokazatelji koji se koriste u tu svrhu su Pearsonov korelacijski koeficijent, Cohenov  $f$  i  $f^2$  i drugi.

U ovom istraživanju koristi se Pearsonov korelacijski koeficijent za procjenu veličine učinka nezavisne značajke. Općenito, Pearsonov korelacijski koeficijent između dviju slučajnih značajki  $X, Y$  označava se s  $\rho_{XY}$  i poprima vrijednosti iz  $[-1,1]$  [64]. Tumačenje veličine učinka prediktora  $X$  prema apsolutnoj vrijednosti Pearsonovog korelacijskog koeficijenta  $\rho_{X,Y}$  prikazano je u tablici 4.4 [128].

Tablica 4.4: Tumačenje veličine učinka prema apsolutnoj vrijednosti Pearsonovog koeficijenta korelacije  $|\rho_{X,Y}|$ , prema [128].

Vrijednost $ \rho_{X,Y} $	Tumačenje
$ \rho_{X,Y}  \in [0.10, 0.30)$	mala
$ \rho_{X,Y}  \in [0.30, 0.50)$	srednja
$ \rho_{X,Y}  \in [0.50, 1]$	velika

Navedena ideja korištena je u ovom istraživanju na sljedeći način: očekivana vrijednost koeficijenta korelacije izvede se iz postavki scenarija (opisano u nastavku), a "mišljenje stručnjaka" o utjecaju pojedine ulazne značajke na izlaznu se formira kao u tablici 4.5.



Tablica 4.5: Formiranje "mišljenja stručnjaka" o utjecaju pojedine ulazne značajke na izlaznu temeljem vrijednosti koeficijenta korelacije  $|\rho_{X,Y}|$

Vrijednost $ \rho_{X,Y} $	Kategorija utjecaja
$ \rho_{X,Y}  \in [0.00, 0.10)$	1
$ \rho_{X,Y}  \in [0.10, 0.30)$	2
$ \rho_{X,Y}  \in [0.30, 0.50)$	3
$ \rho_{X,Y}  \in [0.50, 1]$	4

1: Nema utjecaja, 2: Slab utjecaj, 3: Srednji utjecaj, 4: Jak utjecaj.

#### 4.3.2. Procjena veličine učinka po scenarijima

Korelacijski koeficijent  $\rho_{X,Y}$  dviju slučajnih značajki  $X, Y$  dan je formulom:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Očekivana vrijednost korelacijskog koeficijenta  $\rho_{X,Y}$  dviju slučajnih značajki je:

$$E[\rho_{X,Y}] = E \left[ \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \right], \quad (4.4)$$

pri čemu vrijedi [48]:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])],$$

i

$$Var(X) = E[(X - E[X])^2]. \quad (4.5)$$

Nadalje, u svakom scenariju su prediktori standardizirani (imaju očekivanje 0 i varijancu 1 (tj.  $E[X] = 0$  i  $Var(X) = 1$ )), pa se izraz 4.4 pojednostavljuje na sljedeći način:

$$E[\rho_{X,Y}] = E \left[ \frac{Cov(X,Y)}{\sqrt{Var(Y)}} \right].$$

Na taj način, za izračun očekivanog koeficijenta korelacije za svaki prediktor koji stvarno ulazi u model, preostaje izračunati kovarijance prediktora i izlazne značajke te varijancu izlazne značajke.

### 4.3.3. Podjela simuliranih podataka na javnom repozitoriju

Simulirani podaci su javno podijeljeni na repozitoriju Zenodo [129]. Zenodo je otvoren repozitorij istraživačkih podataka koji nudi CERN (skraćeno od fran. *Conseil européen pour la recherche nucléaire* - Europsko vijeće za nuklearna istraživanja) u svrhu očuvanja i stavljanja na raspolaganje istraživačkog, obrazovnog i informativnog sadržaja [130].

Osim simuliranih podataka korištenih u ovom istraživanju (50 skupova podataka iz svakog scenarija), podijeljena je i R skripta s funkcijama za generiranje i spremanje podataka iz spomenutih scenarija. Dijeljenje podataka omogućava čitateljima rada da provjere rezultate ili da usporede rezultate vlastitih modela s onima postignutim u ovom istraživanju. Uz to, zainteresirani korisnici mogu sami generirati željen broj skupova podataka iz istih scenarija ili mijenjati i prilagođavati parametre scenarija te, na taj način, generirati skupove podataka iz sličnih scenarija.

## 4.4. PROCJENA I EVALUACIJA MODELA

Podaci su podijeljeni na slučajan način u omjeru 80:20 na skup za treniranje i skup za testiranje. Struktura podjele uzorka stvarnih podataka je prikazana detaljnije na slici 4.1, a simuliranih na 4.2.



Slika 4.1: Podjela uzorka u slučaju stvarnih podataka

Metode su se učile na skupu za treniranje. Pri učenju metode za regularizaciju linearnog regresijskog modela, potrebno je specificirati vrijednosti parametara  $\lambda$  koji kontrolira količinu skupljanja u regularizaciji i  $\alpha$  koji govori o odnosu  $l_1$  i  $l_2$  norme u regularizacijskom izrazu



Slika 4.2: Podjela uzorka u slučaju simuliranih podataka

( $\alpha = 0$  za hrbatnu regularizaciju,  $\alpha = 1$  za lasso regularizaciju,  $\alpha \in (0, 1)$  za regularizaciju elastičnom mrežom).

U slučaju kada se ne koristi domensko znanje stručnjaka ili se ono uključuje deterministički, procjena parametra  $\alpha$  vršila se na način da se za svaki  $\alpha$  iz niza  $(0, 0.01, 0.02, \dots, 1)$  metoda učila na skupu za treniranje korištenjem funkcije `cv.glmnet` koja optimizira odabir parametra  $\lambda$ . Pritom se za procjenu parametra  $\lambda$  koristila unakrsna validacija uz  $k = 5$ , što je obuhvaćalo sljedeće korake [5]:

- u prvom izvođenju funkcije najprije se procijeni niz  $\lambda$
- podaci iz skupa za treniranje podijeljeni su nasumično na 5 dijelova podjednake veličine,
- jedan podskup ostavljen je po strani, a na preostala 4 podskupa metoda se uči te se potom greška predikcije validira na podskupu koji je ostavljen po strani,
- postupak se ponavlja 5 puta, pri čemu je svaki put drugačiji podskup ostavljen po strani,
- srednje kvadratne greške predikcije (eng. *mean square error* - *MSE*) dobivene validacijom na ostavljenim podskupovima se uprosječuju,
- $\lambda$  se odabire tako da je prosječna srednje kvadratne greška predikcije najmanja.

Na taj način kreirao se 101 model, a odabran je kao konačan onaj koji minimizira srednje kvadratnu grešku predikcije.

U slučaju kada se domensko znanje stručnjaka uključivalo metaheuristikom, istovremeno su se optimizirali i penalizacijski koeficijenti i parametri  $\alpha$  i  $\lambda$ . Ključan element optimizacije

bila je funkcija cilja kojoj je ciljna vrijednost srednje kvadratna greška predikcije, a prilikom optimizacije tražena je minimalna vrijednost. Pritom se za kreiranje modela regularizacije koristila funkcija `glmnet.fit`. Nadalje, postavljene su sljedeće donje i gornje granice, odnosno rasponi, za parametre koji se optimiziraju:

- $(0, p)$  za penalizacijske koeficijente, gdje je  $p$  broj značajki,
- $(0, 100)$  za parametar  $\lambda$ ,
- $(0, 1)$  za parametar  $\alpha$  u regularizaciji elastičnom mrežom.

Bitno je za naglasiti da se i tu provodila unakrsna validacija uz  $k = 5$  i da je unakrsna validacija za sve metode regularizacije rađena tako da su podaci iz skupa za treniranje raspodijeljeni uvijek na isti način u pet polja (vidljivo na slikama 4.1 i 4.2).

Konačno, skup za testiranje se koristio u prediktivne svrhe. Prilikom usporedbe uspješnosti predikcije između više modela, predikcija je uspješnija u slučaju kada su iznosi *MAE*, *SMAPE* i *RMSE* greški manji, a iznos korigiranog koeficijenta determinacije  $\tilde{R}^2$  veći. U slučaju simuliranih podataka, gdje se koristi i euklidska udaljenost koeficijenata, predikcija je uspješnija kada je spomenuta udaljenost manja.

Uspješnost predikcije metoda koje, na različite načine, uključuju domensko znanje stručnjaka, uspoređena je s uspješnošću predikcije metoda lasso i hrbatne regularizacije, regularizacije elastičnom mrežom koje ne uključuju znanje stručnjaka te klasifikacijskim i regresijskim stablima [5, 24] kao metodom koje omogućuje nelinearne odnose između značajki.

## 4.5. EVALUACIJA HIPOTEZA

Kako bi se omogućila generalizacija rezultata, hipoteze su, osim na stvarnim podacima, evaluirane i pomoću simuliranih podataka, odnosno simulacijskog eksperimenta.

S obzirom na to da se u radu koriste različite mjere evaluacije i različiti skupovi podataka, riječ je o "složenim hipotezama". Prema Offermann i sur. [131], složene hipoteze moguće je raščlaniti na nekoliko manjih koje se mogu lakše odbaciti ili ne odbaciti. U ovom istraživanju se, stoga, za svaku hipotezu zasebno razmatraju stvarni podaci i svaki scenarij simuliranih podataka te se na njima, dodatno, zasebno analizira svaka mjera evaluacije.

U slučaju Hipoteze 1, u fokusu su vrijednosti svih prethodno spomenutih mjera evaluacije između modela regularizacije koje ne uključuju znanje stručnjaka s modelima regularizacije koji

ga uključuju deterministički i metaheuristički. Konkretno, kod stvarnih se podataka uspoređuje po jedna dobivena vrijednost svih mjera evaluacije između te dvije vrste modela.

S druge strane, za simulirane podatke dizajniran je dvofaktorski eksperiment za svaku mjeru evaluacije i za svaki scenarij [132]. Prvi je faktor Procjena penalizacijskog koeficijenta koji ima tri razine: (1) - (nije rađena procjena (nije korišteno domensko znanje)), (2) det (deterministička procjena) i (3) mh (procjena metaheuristikom). Drugi je faktor Metoda regularizacije i također ima tri razine: (1) regularizacija elastičnom mrežom, (2) lasso regularizacija i (3) hrbatna regularizacija. Za svaku od spomenutih mjera evaluacije je u pitanju veliki uzorak od ukupno 450 rezultata koji se dijele na 9 grupa (koliko je mogućih kombinacija razina ova dva faktora) po 50 rezultata. Budući da je jednak broj replikacija u svakoj grupi, riječ je o balansiranom skupu [132]. U konačnici, za simulirane podatke, hipoteze su se provjeravale analizom varijance (eng. *analysis of variance* - ANOVA) [132].

Analiza varijance u svom najopćenitijem obliku testira hipotezu o jednakosti očekivane vrijednosti neke zavisne značajke po različitim razinama faktora koji predstavlja nekakav tretman (tzv. faktor tretmana), no ne daje odgovor koje se razine razlikuju ukoliko faktor tretmana ima više od dvije razine [132]. Međutim, hipoteze o razlikama između razina mogu se provjeravati na dva načina. Jedan je način provjera tzv. *post hoc* analizom kada nije poznato unaprijed koje su razlike u fokusu. Kada to je unaprijed poznato, model analize varijance može se parametrizirati na način da se matrica dizajna postavi tako da kontrasti odgovaraju hipotezama koje želimo testirati. Pritom je kontrast linearna kombinacija efekata [132]. Najveći broj linearno nezavisnih kontrasta efekata odgovara broju stupnjeva slobode i kada je baš jednak broju stupnjeva slobode, radi se o zasićenom skupu procjenjivih kontrasta [132].

Pretpostavke valjanosti analize varijance, koja je specijalni slučaj općeg linearnog modela kada su svi prediktori kvalitativne značajke, konstantnost su varijance greške na svim razinama faktora tretmana te nezavisnost i normalnost distribucije grešaka. Provjera ovih pretpostavki izvršena je grafički, odnosno analizom dijagrama. Riječ je o dijagramu raspršenosti reziduala modela u odnosu na razine faktora, koji može pokazati je li varijabilnost koja se vidi na svakoj razini faktora približno jednaka, dok se normalnost distribucije grešaka može provjeriti izradom dijagrama normalne vjerojatnosti reziduala modela [132, 133].

Slijedom navedenog, u nastavku je pojašnjeno na koji način je analiza varijance korištena za donošenje odluke o Hipotezi 1 za svaki scenarij simuliranih podataka i svaku mjeru evaluacije zasebno.

Najprije se uvode sljedeće oznake:

- prvi faktor (Procjena penalizacijskog koeficijenta):
  - $\mu_1$  : marginalna očekivana vrijednost pojedine mjere evaluacije kada se ne radi procjena penalizacijskih koeficijenata (odnosno ne koristi se domensko znanje stručnjaka)
  - $\mu_2$  : marginalna očekivana vrijednost pojedine mjere evaluacije kada se deterministički procjenjuju penalizacijski koeficijenti
  - $\mu_3$  : marginalna očekivana vrijednost pojedine mjere evaluacije kada se metaheuristikom procjenjuju penalizacijski koeficijenti
- drugi faktor (Metoda regularizacije):
  - $\mu_{.1}$  : marginalna očekivana vrijednost pojedine mjere evaluacije za regularizaciju elastičnom mrežom
  - $\mu_{.2}$  : marginalna očekivana vrijednost pojedine mjere evaluacije za lasso regularizaciju
  - $\mu_{.3}$  : marginalna očekivana vrijednost pojedine mjere evaluacije za hrbatnu regularizaciju.

S obzirom na to da oba faktora imaju tri razine, to znači da je za svaki od njih moguće definirati najviše dva linearno nezavisna kontrasta koji čine zasićeni skup.

U nastavku su navedeni kontrasti i hipoteze koje se njima testiraju:

- prvi faktor (Procjena penalizacijskog koeficijenta):
  - kontrast c11 s koeficijentima  $(1, -\frac{1}{2}, -\frac{1}{2})$ , čime se testira hipoteza  $H_0 : \mu_1 = \frac{1}{2}(\mu_2 + \mu_3)$  (usporedba metoda regularizacije bez procjene penalizacijskih koeficijenata s metodama regularizacije koje penalizacijske koeficijente procjenjuju deterministički ili metaheuristikom)
  - kontrast c12 s koeficijentima  $(0, -1, 1)$ , čime se testira hipoteza  $H_0 : \mu_2 = \mu_3$  (usporedba metoda regularizacije koje penalizacijske koeficijente procjenjuju deterministički s onima koje ih procjenjuju metaheuristikom)
- drugi faktor (Metoda regularizacije):

- kontrast c21 s koeficijentima  $(1, -\frac{1}{2}, -\frac{1}{2})$ , čime se testira hipoteza  $H_0 : \mu_1 = \frac{1}{2}(\mu_2 + \mu_3)$  (usporedba regularizacije elastičnom mrežom s lasso i hrbatnom regularizacijom)
- kontrast c22 s koeficijentima:  $(0, -1, 1)$ , čime se testira hipoteza  $H_0 : \mu_2 = \mu_3$  (usporedba lasso regularizacije s hrbatnom regularizacijom).

Bitno je za naglasiti da se pomoću kontrasta c11 testira upravo Hipoteza 1 ovog istraživanja, dok kontrast c12 uspoređuje međusobno dva načina uvođenja domenskog znanja stručnjaka u metode regularizacije.

Nadalje, u model ANOVE uvrštava se i interakcija dvaju spomenutih faktora. Navedeno je bitno radi interpretacije marginalnih efekata. Ako postoji statistički značajna interakcija, učinak jednog faktor na zavisnu značajku razlikovat će se ovisno o razini drugog faktora.

Princip evaluacije Hipoteze 2 analogan je evaluaciji Hipoteze 1, a mjere evaluacije su *RMSE*, *MAE* i *SMAPE* greške (pojašnjeno u potpoglavlju 2.7). U Hipotezi 2 uspoređuju se vrijednosti mjera evaluacije modela između metode CART i metoda regularizacije koje uključuju domensko znanje stručnjaka. Hipoteza se dijeli na dvije podhipoteze - u jednoj se uspoređuju rezultati mjera evaluacije metode CART s metodama regularizacije koje uključuju domensko znanje stručnjaka deterministički (Hipoteza 2a), a u drugoj se rezultati mjera evaluacije metode CART uspoređuju s metodama regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom (Hipoteza 2b). Unutar svake se podhipoteze za stvarne podatke uspoređuje po jedna dobivena vrijednost tih mjera evaluacije između spomenutih metoda, a za simulirane podatke dizajniran je jednofaktorski eksperiment za svaku mjeru evaluacije i za svaki scenarij [132]. Faktor se odnosio na metodu i imao je četiri razine: (1) CART, (2) regularizacija elastičnom mrežom, (3) lasso regularizacija i (4) hrbatna regularizacija. Za svaku od spomenutih mjera evaluacije u pitanju je veliki uzorak od ukupno 200 rezultata koji se dijele na 4 grupe po 50 rezultata te se radi o balansiranom skupu. U konačnici, za simulirane podatke, odluka o hipotezi donijela se analizom varijance.

Najprije se uvode sljedeće oznake za faktor Metoda:

- $\mu_1$  : očekivana vrijednost pojedine mjere evaluacije za metodu CART
- $\mu_2$  : očekivana vrijednost pojedine mjere evaluacije za regularizaciju elastičnom mrežom
- $\mu_3$  : očekivana vrijednost pojedine mjere evaluacije za lasso regularizaciju

- $\mu_4$  : očekivana vrijednost pojedine mjere evaluacije za hrbatnu regularizaciju.

S obzirom na to da faktor ima četiri razine, moguće je definirati najviše tri linearno nezavisna kontrasta koji čine zasićeni skup. U nastavku su navedeni kontrasti i hipoteze koje se njima testiraju:

- kontrast c31 s koeficijentima  $(1, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})$ , čime se testira hipoteza  $H_0 : \mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$  (usporedba CART metode s metodama regularizacija elastičnom mrežom, lasso i hrbatna regularizacija)
- kontrast c32 s koeficijentima  $(0, 1, -\frac{1}{2}, -\frac{1}{2})$ , čime se testira hipoteza  $H_0 : \mu_2 = \frac{1}{2}(\mu_3 + \mu_4)$  (usporedba regularizacije elastičnom mrežom s lasso i hrbatnom regularizacijom)
- kontrast c33 s koeficijentima  $(0, 0, 1, -1)$ , čime se testira hipoteza  $H_0 : \mu_3 = \mu_4$  (usporedba lasso regularizacije s hrbatnom regularizacijom).

Bitno je za naglasiti da se pomoću kontrasta c31 testira upravo Hipoteza 2 ovog istraživanja.

Ukoliko neke dodatne razlike u pojedinoj situaciji budu od interesa, analiza se izvršava *post hoc* Tukeyjevom metodom (eng. *Tukey's HSD (honestly significant difference) method*) [132, 134]. Odluke se donose na razini statističke značajnosti 0.05.

## 4.6. RAČUNALNA IMPLEMENTACIJA MODELA

Sve metode obrade, modeliranja i evaluacije rađene su u programskom jeziku R (verzija 4.1.2), pri čemu se koristilo više različitih dostupnih paketa, među kojima su **irr** za slaganja stručnjaka, **glmnet** za regularizacije, **metaheuristicOpt** za metaheuristike te **rpart** za regresijska stabla.

U okviru izrade ovog istraživanja korištena je usluga Napredno računanje (Padobran) Sveučilišta u Zagrebu Sveučilišnog računskog centra (Srce) [135]. Navedena usluga korištena je prilikom optimizacije parametara metaheuristikom.

U nastavku su navedeni detalji korištenja ovih funkcija radi osiguranja ponovljivosti istraživanja zainteresiranim osobama.

Pri korištenju funkcija `cv.glmnet / glmnet.fit` iz paketa **glmnet** sve značajke standardizirane su i uključen je slobodan koeficijent  $\beta_0$ , što su standardne postavke. Domensko znanje stručnjaka uključuje se u lasso i hrbatnu regularizaciju te regularizaciju elastičnom mrežom specifikiranjem vrijednosti argumenta "penalty.factor". Pritom vrijedi sljedeće: 1 je zadana vrijednost, vrijednost 0 znači da za pojedinu značajku nema skupljanja (pritezanja) i da je ona uvijek



uključena u model, a vrijednosti penalizacijskog koeficijenta normaliziraju se da im suma bude jednaka broju stupaca matrice dizajna.

U sklopu funkcije `metaOpt` iz paketa **metaheuristicOpt**, algoritam pretraživanja kukavice koristio je sljedeće parametre: udio populacije koji se mijenja (`abandonedFraction`) = 0.7, veličina populacije (`numPopulation`) = 100, maksimalni broj iteracija (`maxIter`) = 1000. Broj varijabli odluke (`numVar`) ovisio je o broju različitih kategorija utjecaja u pojedinom skupu podataka. Dodatne varijable odluke bile su i procjena parametra  $\lambda$  iz regularizacije, kao i parametra  $\alpha$  u slučaju regularizacije elastičnom mrežom. Rasponi vrijednosti za procjenu parametara (`rangeVar`) navedeni su u dijelu (4.4) ovog poglavlja. U istom dijelu pojašnjena je funkcija cilja (FUN). Također, "seed" je postavljen na 2023.

Metoda CART u ovom istraživanju korištena je funkcijom `rpart` iz istoimenog paketa uz unakrsnu validaciju uz  $k = 5$  (argument `xval`). Valja napomenuti da je, u slučaju stvarnih podataka, metoda CART koristila originalne vrijednosti zavisne i nezavisnih značajki (kod metoda regularizacije je većina kvantitativnih značajki transformirana logaritamskom transformacijom, kako je pojašnjeno u dijelu 4.2 ovog poglavlja).

## 5. REZULTATI ISTRAŽIVANJA

### 5.1. STVARNI PODACI

Stvarni podaci o godišnjem trošku energije zgrada javnog sektora Republike Hrvatske i 25 karakteristika zgrada kao što su korištenje i namjena zgrade, konstrukcijske i energetske karakteristike, karakteristike grijanja i hlađenja, geografske i druge radi preglednosti grafički su prikazani u Prilogu 1 (str. 166 - 169), dok je deskriptivna analiza dana u nastavku u tablicama 5.1 i 5.2.

#### 5.1.1. Deskriptivna analiza podataka

Tablica 5.1: Deskriptivna analiza kvantitativnih značajki

Značajka	Min.	Med.	Arit. sredina	Max.	St. dev.	Koef. asim.
REAL_COSTTAX [kn]	66.00	43578.00	121090.00	5811146.00	282438.60	11.9
NUMBEROFEMPLOYEES	1	22.01	34.84	922	52.21	6.52
NUMBEROFUSERS	1	90.20	934.70	840576	22370.64	37.26
HEATEDSURFACEOFTHE-BUILDING [ $m^2$ ]	12.38	823.35	1518.69	30250.00	1986.70	4.13
HEATEDVOLUMEAREAOF-THEBUILDING [ $m^3$ ]	38.70	3000.00	6578.60	153000.00	9926.10	6.02
NUMBEROFFLOORS	0	2	2.21	10	1.28	1.38
H1TRND [ $W/m^2K$ ]	0.44	0.61	0.65	1.65	0.20	1.26
SHAREOFWINDOWSSUR-FACE [%]	0.01	0.33	5.98	97.40	11.59	2.97

Značajka	Min.	Med.	Arit. sredina	Max.	St. dev.	Koef. asim.
TOTALPOWERBODYHEAT-RADIATOR [kW]	0.00	68.20	615.80	599119.00	16036.44	36.71
TOTALPOWERBODYOTHER [kW]	0.00	15.31	123.05	81863.00	2174.42	37.42
TITHERMALPOWEROFHEATERS [kW]	0.80	100.50	733.70	599119.00	16179.69	35.81
TOTALHEATINGPOWER [kW]	2.00	128.00	298.33	9505.00	552.79	8.6
TICOOLINGPOWEROFCOOLERS [kW]	0.00	9.33	104.50	56494.00	2082.20	26.46
LIGHT_INSTALLED_INT_PWR_KW [kW]	0.00	1.50	122.90	30504.00	1683.83	16.26
TIPOWEROFFOFFICEEQUIP [kW]	0.00	6.81	19.61	5850.00	180.38	28.09
TIPOWEROFFOTHERCONSUMERS [kW]	0.00	4.18	18.19	4840.00	133.26	33.49
OTHER_TOTAL_INSTALLED_PWR_KW [kW]	0.10	18.00	54.14	17637.00	509.88	31.01
OBJ_CON_THICKNESS_D3 [cm]	0.00	3.62	4.12	200.90	6.85	17.48
OBJ_CON_THICKNESS_D7 [cm]	7.00	40.00	42.16	130.00	14.44	1.41
Q1HNDDOP [kWh/m <sup>3</sup> a]	16.42	23.54	24.24	30.40	4.38	0.12

Nadalje, budući da linearni regresijski model pretpostavlja linearne povezanosti značajki, u Prilogu 1 prikazani su dijagrami rasipanja kvantitativnih nezavisnih značajke i zavisne značajke, kao i iznos njihove korelacije (Slika 7.5 na str. 170 i Slika 7.6 na str. 171). Navedeni dijagrami ne upućuju na značajna odstupanja od ove pretpostavke pa se smatra valjanim koristiti metodu linearne regresije.

Tablica 5.2: Deskriptivna analiza kvalitativnih značajki

Značajka	Vrijednosti i njihove apsolutne i relativne frekvencije
OBJ_REG	1: 346 (24.35 %) 2: 1075 (75.65 %)
OBJ_OBJT	1: 228 (16.05 %) 2: 24 (1.69 %) 3_18: 49 (3.45 %) 3_20: 876 (61.65 %) 4: 28 (1.97 %) 8_22: 54 (3.80 %) 11: 162 (11.40 %)
YEAROFCOMPLETIONOFCONSTR	1: 159 (11.19 %) 2: 101 (7.11 %) 3: 144 (10.13 %) 4: 196 (13.79 %) 5: 348 (24.49 %) 6: 149 (10.49 %) 7: 324 (22.80 %)
YEAROFLASTRESTORATION	0: 1379 (97.04 %) 1: 42 (2.96 %)
MODEOFSUPPLYOFDRINKINGWATER	1: 1396 (98.24 %) 2: 25 (1.76 %)
OBJ_EXTENDED_GEO_TYPE	1: 1065 (74.95 %) 2: 177 (12.46 %) 3: 37 (2.60 %) 4: 142 (9.99 %)

Pojašnjenja kategorija pogledati u tablici 4.2 (str. 40).

### 5.1.2. Mišljenje stručnjaka

U nastavku, tablica 5.3 prikazuje frekvencije kategorija utjecaja ulaznih značajki prema mišljenju stručnjaka, dok tablica 7.1 u Prilogu 1 (str. 172) prikazuje mišljenje stručnjaka za svaku pojedinu ulaznu značajku. Rezultati ispitivanja stručnjaka ujedno su javno podijeljeni na repozitoriju Zenodo [136].

Tablica 5.3: Tablica frekvencija pojedine kategorije značajki po stručnjacima

	Stručnjak			
Utjecaj	1	2	3	4
1	0	2	17	2
2	6	6	5	1
3	9	12	2	12
4	10	5	1	10

1: Nema utjecaja, 2: Slab utjecaj, 3: Srednji utjecaj, 4: Jak utjecaj.

Iz tablice 5.3 vidljivo je da se Stručnjak 3 prilično razlikuje u svojoj procjeni utjecaja ulaznih značajki na godišnji trošak energije od ostalih stručnjaka stoga je upitno koliko je smisleno ova četiri mišljenja agregirati u jedno.

Iz tog razloga, u nastavku su analizirane mjere slaganja i povezanosti u slučaju kada se sva četiri stručnjaka promatraju zajedno. Slaganja i povezanost sva četiri stručnjaka analizirana su mjerama Cohenov kapa koeficijent slaganja temeljem modela za ordinalne značajke ( $\kappa_m$ ) i Cohenov kapa koeficijent povezanosti temeljem modela za ordinalne značajke ( $\kappa_{ma}$ ), intraklasni koeficijent korelacije 1 ( $ICC1$ ), intraklasni koeficijent korelacije 2 ( $ICC2$ ), intraklasni koeficijent korelacije 3 ( $ICC3$ ), Kendallov koeficijent podudarnosti  $W$  i Krippendorffov koeficijent pouzdanosti  $\alpha$ . Usporedbe radi, odmah je analiziran i slučaj kada se promatraju samo tri stručnjaka (Stručnjak 1, Stručnjak 2 i Stručnjak 4). Rezultati su navedeni u tablici 5.4.

U konačnici, vidljivo je da se slučaj kad su stručnjaci razdvojeni pokazao boljim u smislu ovih mjera od slučaja kada se promatraju svi stručnjaci zajedno. Stoga, pri uključivanju znanja stručnjaka u regularizacijske metode razmatraju se dva različita domenska znanja; u prvom slučaju domensko znanje o utjecaju ulaznih značajki na izlaznu određeno je kao medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4 (Domensko znanje 1), a u drugom slučaju je

predstavljeno mišljenjem Stručnjaka 3 (Domensko znanje 2).

Tablica 5.4: Usporedba mjera slaganja i povezanosti stručnjaka kada se promatraju sva četiri stručnjaka zajedno (Slučaj 1) i kada se promatraju zajedno Stručnjak 1, Stručnjak 2 i Stručnjak 4 (Slučaj 2)

Slučaj	Mjera						
	$\kappa_m$	$\kappa_{ma}$	ICC 1	ICC 2	ICC 3	W	$\alpha$
Slučaj 1	0.090	0.178	0.061	0.172*	0.079	0.524*	0.106
Slučaj 2	0.182	0.33	0.346*	0.355*	0.410	0.681*	0.448

$\kappa_m$  - kapa koeficijent slaganja temeljem modela,  $\kappa_{ma}$  - kapa koeficijent povezanosti temeljem modela, ICC1 - intraklasni koeficijent korelacije 1, ICC2 - intraklasni koeficijent korelacije 2, ICC3 - intraklasni koeficijent korelacije 3, W - Kendallov koeficijent podudarnosti W,  $\alpha$  - Krippendorffov koeficijent pouzdanosti; \*: statistički značajno različit od 0 na razini statističke značajnosti 0.05.

Nadalje, tablica 7.2 u Prilogu 1 (str. 173) prikazuje konačno domensko znanje stručnjaka, odnosno kategoriju utjecaja svake značajke, dok tablica 5.5 prikazuje frekvencije pojedinih kategorija utjecaja značajki u oba promatrana slučaja.

Tablica 5.5: Tablica frekvencija pojedine kategorije značajki u obje promatrane vrste domenskog znanja

Utjecaj	Domensko znanje 1	Domensko znanje 2
1	0	17
2	5	5
3	13	2
4	7	1

Kategorije utjecaja: 1: *Nema utjecaja*, 2: *Slab utjecaj*, 3: *Srednji utjecaj*, 4: *Jak utjecaj*. Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3.

Tablica 5.6: Tablica frekvencija pojedine kategorije značajki u obje promatrane vrste domenskog znanja za matrice dizajna

Utjecaj	Domensko znanje 1	Domensko znanje 2
1	0	19
2	5	10
3	20	7
4	12	1

1: *Nema utjecaja*, 2: *Slab utjecaj*, 3: *Srednji utjecaj*, 4: *Jak utjecaj*. Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3.

Budući da se u sklopu funkcija `cv.glmnet / glmnet.fit` iz paketa **glmnet** u R-u kvalitativne značajke s  $k$  kategorija pretvorene u  $k - 1$  binarnih značajki, skup stvarnih podataka sastoji se od ukupno 37 ulaznih značajki, a tablice frekvencija pojedine kategorije značajki takvog skupa prikazane su u tablici 5.6.

Tablica 5.7: Vrijednosti penalizacijskog koeficijenta nakon normalizacije u slučaju determinističke procjene i obje vrste domenskog znanja

Koeficijent	Domensko znanje 1	Domensko znanje 2
$\gamma_1$	-	1.454
$\gamma_2$	2.606	0.727
$\gamma_3$	1.042	0.291
$\gamma_4$	0.261	0.073

$\gamma_1, \gamma_2, \gamma_3$  i  $\gamma_4$  su penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja*, *Slab utjecaj*, *Srednji utjecaj* i *Jak utjecaj*, redom. Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3.

Budući da se vrijednosti penalizacijskog koeficijenta iz tablice 4.1 prilikom korištenja funkcija za regularizaciju iz paketa **glmnet** u R-u naknadno normaliziraju na način da njihova suma

bude jednaka broju značajki, u tablici 5.7 prikazane su vrijednosti nakon normalizacije za obje promatrane vrste domenskog znanja. Navedeni prikaz koristan je i u smislu usporedivosti vrijednosti penalizacijskog koeficijenta između različitih načina procjenjivanja.

### 5.1.3. Procjena penalizacijskih koeficijenata algoritmom pretraživanja kukavice

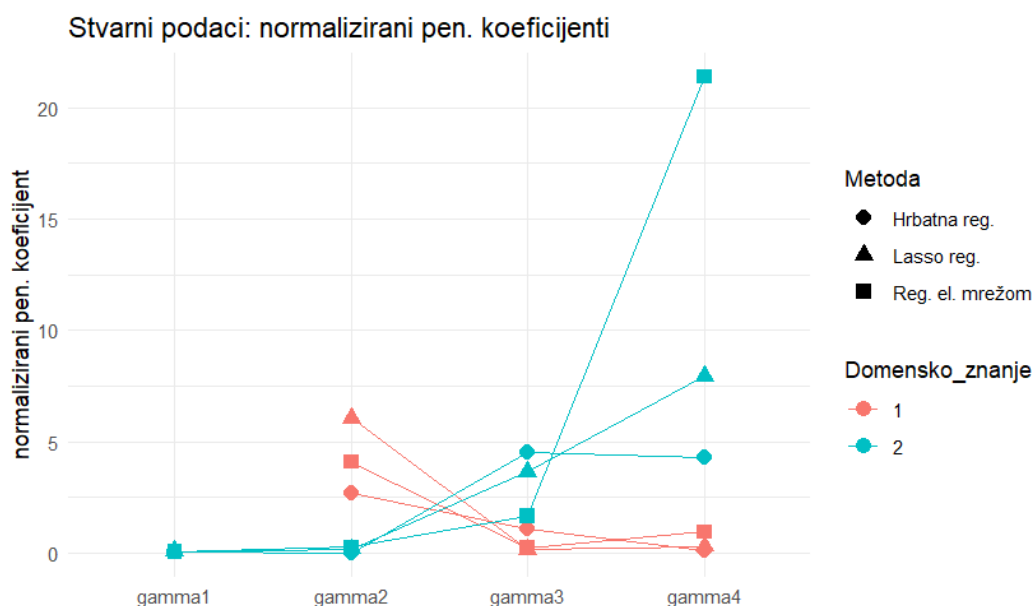
Rezultati procjene vrijednosti penalizacijskih koeficijenata i drugih parametara regularizacije metaheurističkim algoritmom za obje promatrane vrste domenskog znanja stvarnih podataka prikazani su u tablici 5.8 te na slici 5.1.

Tablica 5.8: Procjene vrijednosti parametara dobivene optimizacijom prema pojedinoj vrsti regularizacije

Domensko znanje	Parametar	Vrsta regularizacije		
		lasso	hrbatna	elastična mreža
Domensko znanje 1	$\gamma_2$	13.432 (6.058)	18.994 (2.716)	26.240 (4.094)
	$\gamma_3$	0.344 (0.155)	7.634 (1.091)	1.644 (0.256)
	$\gamma_4$	0.667 (0.301)	0.934 (0.133)	6.091 (0.950)
	$\lambda$	0.822	0.001	0.791
	$\alpha$	1	0	0.914
Domensko znanje 2	$\gamma_1$	0.255 (0.079)	0.429 (0.058)	0.094 (0.063)
	$\gamma_2$	0.603 (0.188)	0.076 (0.010)	0.394 (0.264)
	$\gamma_3$	11.755 (3.668)	33.084 (4.499)	2.505 (1.682)
	$\gamma_4$	25.422 (7.933)	31.577 (4.294)	31.857 (21.386)
	$\lambda$	2.023	99.520	14.377
	$\alpha$	1	0	0.073

Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3.  $\gamma_2, \gamma_3, \gamma_4$ : penalizacijski koeficijenti uz značajke kategorije *Slab utjecaj*, *Srednji utjecaj* i *Jak utjecaj*, redom (vrijednosti u zagradi su vrijednosti nakon normalizacije);  $\lambda$ : parametar složenosti koji kontrolira količinu skupljanja u regularizaciji;  $\alpha$ : parametar regularizacije koji govori o odnosu  $l_1$  i  $l_2$  norme u regularizaciji.





Slika 5.1: Dijagram procjene penalizacijskih koeficijenata (normalizirane vrijednosti)

$\gamma_1$ ,  $\gamma_2$ ,  $\gamma_3$  i  $\gamma_4$  su penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja*, *Slab utjecaj*, *Srednji utjecaj* i *Jak utjecaj*, redom. Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3.

Na istima vidljivo je kako, premda ima odstupanja, u slučaju korištenja Domenskog znanja 1 procijenjena vrijednost penalizacijskog koeficijenta ima tendenciju opadanja što je značajka snažnije kategorije utjecaja, dok u slučaju korištenja Domenskog znanja 2 ima tendenciju povećanja što je značajka snažnije kategorije utjecaja.

#### 5.1.4. Evaluacija modela i hipoteza

U tablici 5.9 u nastavku prikazani su rezultati mjera evaluacije modela na stvarnim podacima. Iz tablice je vidljivo kako je uvođenje domenskog znanja stručnjaka u metode regularizacije u većini slučajeva doprinijelo poboljšanju uspješnosti predikcije, pri čemu je to bilo više pri metaheurističkom pristupu. Nadalje, metode regularizacije s domenskim znanjem stručnjaka su u većini slučajeva bile točnije nego metoda CART.

Tablica 5.9: Evaluacija modela na stvarnim podacima

Metoda	Dom. znanje	Procjena pen. koef.	$\alpha$	$\lambda$	$\tilde{R}^2$	RMSE	MAE	SMAPE
lasso reg.	-	-	1	0.057	0.810	287537	77954	28.23
hrbatna reg.	-	-	0	0.783	0.595	303770	78726	28.81
reg. el. mrežom	-	-	0.18	0.197	0.701	298654	77609	28.33
lasso reg.	1	det	1	0.037	0.833	262162	74721	28.48
hrbatna reg.	1	det	0	1.185	0.690	301665	79647	29.03
reg. el. mrežom	1	det	0.01	0.917	0.716	298001	79331	28.95
lasso reg.	2	det	1	0.031	0.810	274256	75014	28.65
hrbatna reg.	2	det	0	0.632	0.594	302833	77875	28.86
reg. el. mrežom	2	det	0.11	0.197	0.707	291957	76347	28.64
lasso reg.	1	mh	1	0.822	<b>0.869</b>	211738	66172	28.03
hrbatna reg.	1	mh	0	0.001	<b>0.869</b>	<b>210951</b>	<b>65673</b>	<b>28.02</b>
reg. el. mrežom	1	mh	0.914	0.791	0.867	213094	66124	28.03
lasso reg.	2	mh	1	2.023	0.857	215834	66146	28.07
hrbatna reg.	2	mh	0	99.520	0.868	217051	65932	<b>28.02</b>
reg. el. mrežom	2	mh	0.073	14.377	0.854	220073	66878	28.04
CART	-	-	-	-	-	279256	78938	33.60

Najbolji rezultat za svaku mjeru evaluacije označen je podebljano.

Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3. Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

Međutim, bitno je za napomenuti i da su metode regularizacije s metaheurističkim pristupom procjene penalizacijskog koeficijenta imale puno dulje vrijeme izvođenja te je u tu svrhu korištena usluga Naprednog računanja (Padobran) Sveučilišnog računskog centra (Srce) Sveučilišta u Zagrebu. U tablici 5.10 prikazano je vrijeme izvođenja prilikom kojeg se koristilo pet procesora.

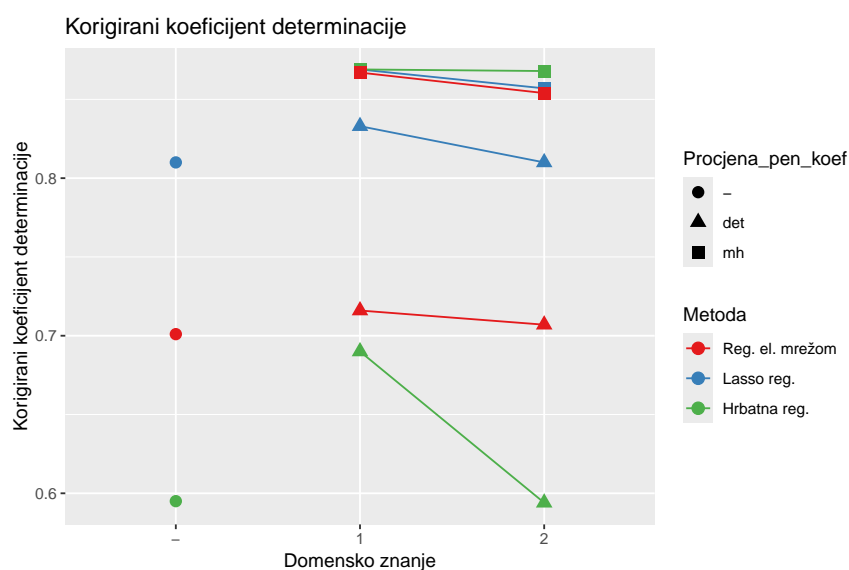
Tablica 5.10: Vrijeme izvođenja metoda regularizacije s metaheurističkom procjenom penalizacijskog koeficijenta

Metoda	Domensko znanje 1	Domensko znanje 2
lasso reg.	11:17	10:48
hrbatna reg.	10:59	10:34
reg. el. mrežom	11:10	11:15

Prilikom izvođenja korišteno je pet procesora, a prikazano vrijeme odnosi se na sate i minute.

Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3.

U nastavku prikazane su vrijednosti pojedine mjere evaluacije prema vrsti domenskog znanja, načinu procjene penalizacijskog koeficijenta i metodi.

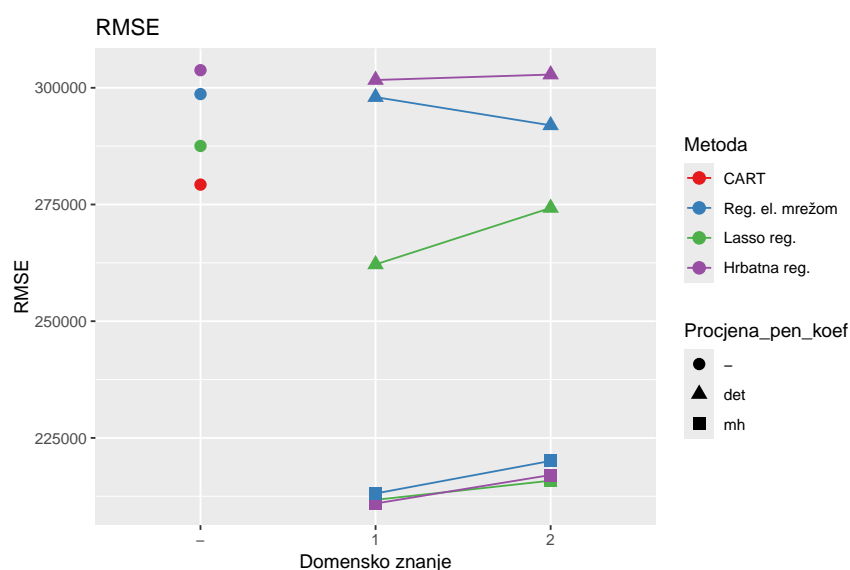


Slika 5.2: Stvarni podaci - korigirani koeficijent determinacije prema vrsti domenskog znanja, procjeni penalizacijskog koeficijenta i metodi

Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3. Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

Na slici 5.2 vidljivo je kako se kretala vrijednost korigiranog koeficijenta determinacije za metode regularizacije, dok metoda CART nije evaluirana ovom mjerom, kako je pojašnjeno u potpoglavlju 2.7, stoga se ova mjera ne tumači u kontekstu Hipoteze 2.

U smislu Hipoteze 1, vidljivo je da je za gotovo svaku metodu regularizacije uvođenje domenskog znanja stručnjaka rezultiralo povećanjem korigiranog koeficijenta determinacije. Izuzetak je uvođenje Domenskog znanja 2 determinističkim pristupom; u hrbatnoj regularizaciji korigirani se koeficijent determinacije smanjio malo ispod razine kada se ne koristi domensko znanje stručnjaka, a u lasso regularizaciji ostao je na istoj razini kao u slučaju kada se ne koristi domensko znanje stručnjaka. Nadalje, može se uočiti da je uvođenje domenskog znanja stručnjaka metaheuristikom rezultiralo većim povećanjem korigiranog koeficijenta determinacije nego uvođenje domenskog znanja stručnjaka determinističkim pristupom. Dodatno, prilikom uvođenja Domenskog znanja 1, vrijednost korigiranog koeficijenta determinacije porasla je više nego li prilikom uvođenja Domenskog znanja 2. Što se tiče metode regularizacije, vidljivo je da u slučajevima kada se ne koristi domensko znanje ili se ono uvodi deterministički, lasso regularizacija ima najviši korigirani koeficijent determinacije, a zatim regularizacija elastičnom mrežom pa hrbatna regularizacija. Kada se domensko znanje uvede metaheuristikom, metode su sličnih iznosa korigiranog koeficijenta determinacije.



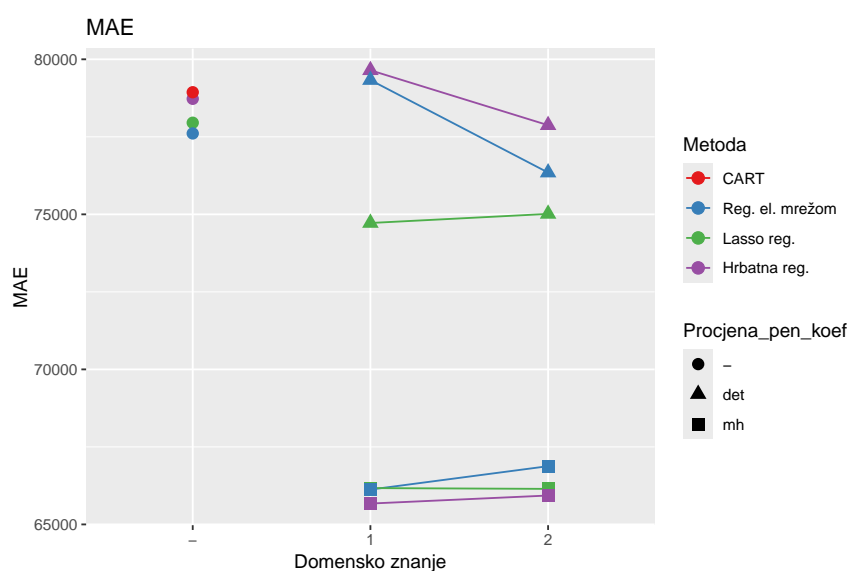
Slika 5.3: Stvarni podaci -  $RMSE$  prema vrsti domenskog znanja, procjeni penalizacijskog koeficijenta i metodi

Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3. Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

Na slici 5.3 vidljivo je kako se kretala vrijednost  $RMSE$  za metode regularizacije i metodu CART. U situaciji kada se ne koristi domensko znanje stručnjaka, vidljivo je da metoda CART rezultira najnižom  $RMSE$ .

U smislu Hipoteze 1, vidljivo je da je za svaku metodu regularizacije uvođenje domenskog znanja stručnjaka rezultiralo smanjenjem  $RMSE$ . Smanjenje je bilo veće kod uvođenja domenskog znanja stručnjaka metaheuristikom nego li kod uvođenja domenskog znanja stručnjaka determinističkim pristupom. Vezano za korišteno domensko znanje, kod determinističkog pristupa je za lasso regularizaciju i hrbatnu regularizaciju bolje Domensko znanje 1, dok je za regularizaciju elastičnom mrežom bolje Domensko znanje 2. Kod uvođenja domenskog znanja stručnjaka metaheuristikom, Domensko znanje 1 rezultiralo je nižom  $RMSE$  za sve tri vrste regularizacije. Što se tiče metode regularizacije, vidljivo je da u slučajevima kada se ne koristi domensko znanje ili se ono uvodi deterministički, lasso regularizacija ima najnižu  $RMSE$ , a zatim regularizacija elastičnom mrežom pa hrbatna regularizacija. Kada se domensko znanje uvede metaheuristikom, hrbatna regularizacija nešto lošija u odnosu na lasso regularizaciju i regularizaciju elastičnom mrežom.

U smislu Hipoteze 2, vidljivo je da je metoda CART rezultirala nižom  $RMSE$  od regularizacija elastičnom mrežom i hrbatne regularizacije s deterministički uvedenim domenskim znanjem stručnjaka. Lasso regularizacija s deterministički uvedenim domenskim znanjem stručnjaka te sve tri regularizacije s metaheuristikom uvedenim domenskim znanjem stručnjaka rezultirale su nižom  $RMSE$  od metode CART.



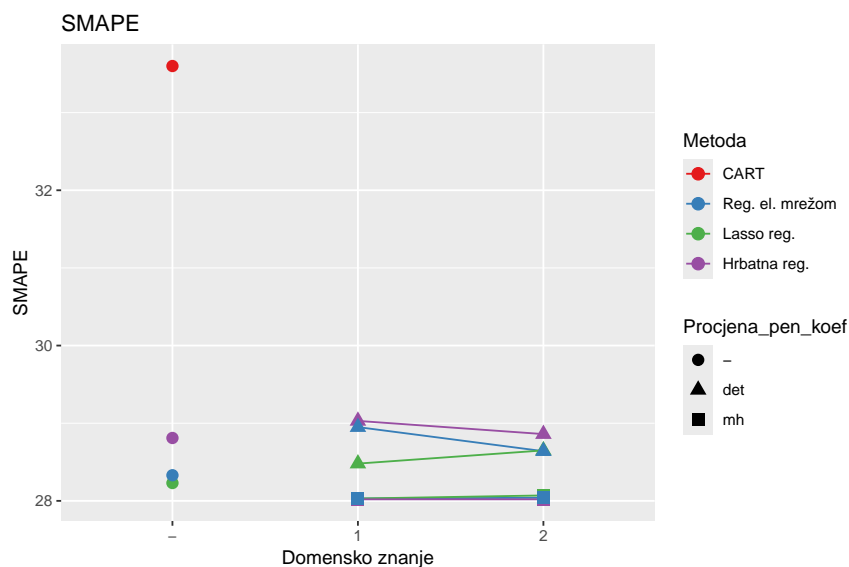
Slika 5.4: Stvarni podaci -  $MAE$  prema vrsti domenskog znanja, procjeni penalizacijskog koeficijenta i metodi

Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3. Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

Na slici 5.4 vidljivo je kako se kretala vrijednost  $MAE$  za metode regularizacije i metodu CART. U situaciji kada se ne koristi domensko znanje stručnjaka, vidljivo je da metoda CART rezultira najvišom  $MAE$ .

U smislu Hipoteze 1, vidljivo je da je uvođenje domenskog znanja stručnjaka u metode regularizacije u gotovo svim slučajevima rezultiralo nižom  $MAE$  te da je to smanjenje bilo veće kod uvođenja domenskog znanja stručnjaka metaheuristikom nego li kod uvođenja domenskog znanja stručnjaka determinističkim pristupom. Izuzetak je regularizacija elastičnom mrežom i hrbatna regularizacija s deterministički uvedenim Domenskim znanjem 1. Vezano za korišteno domensko znanje, za regularizaciju elastičnom mrežom i hrbatnu regularizaciju je kod determinističkog pristupa bolje Domensko znanje 2, a kod metaheuristike Domensko znanje 1. Kod lasso regularizacije vrijedi obratno. Što se tiče metode regularizacije, najveće odstupanje vidljivo je kod deterministički uvedenog Domenskog znanja 1 između lasso regularizacije i preostale dvije vrste regularizacije u korist lasso regularizacije.

U smislu Hipoteze 2, vidljivo je da je metoda CART rezultirala nižom  $MAE$  od regularizacija elastičnom mrežom i hrbatne regularizacije s deterministički uvedenim Domenskim znanjem 1. U svim ostalim su slučajevima metode regularizacije rezultirale nižom  $MAE$  od metode CART.



Slika 5.5: Stvarni podaci -  $SMAPE$  prema vrsti domenskog znanja, procjeni penalizacijskog koeficijenta i metodi

Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3. Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

Na slici 5.5 vidljivo je kako se kretala vrijednost *SMAPE* za metode regularizacije i metodu CART. U situaciji kada se ne koristi domensko znanje stručnjaka, vidljivo je da metoda CART rezultira najvišom *SMAPE*.

U smislu Hipoteze 1, vidljivo je da je za sve metode regularizacije uvođenje domenskog znanja stručnjaka determinističkim pristupom rezultiralo povećanjem *SMAPE*, dok je uvođenje metaheuristikom rezultiralo smanjenjem *SMAPE*. Vezano za korišteno domensko znanje, za regularizaciju elastičnom mrežom i hrbatnu regularizaciju je kod determinističkog pristupa bolje Domensko znanje 2, a kod za lasso regularizacije Domensko znanje 1. Kod metaheurističkog pristupa, za hrbatnu regularizaciju nema razlike u iznosu *SMAPE* greški između Domenskog znanja 1 i Domenskog znanja 2, dok je za lasso regularizaciju i regularizaciju elastičnom mrežom bolje Domensko znanje 1.

U smislu Hipoteze 2, vidljivo je da su metode regularizacije rezultirale nižom *SMAPE* od metode CART u svim slučajevima.

## 5.2. SCENARIJ 1

### 5.2.1. Modeliranje mišljenja stručnjaka

U ovom scenariju poznato je da vrijedi:

$$p = 8$$

$$\beta = (3, 1.5, 0, 0, 2, 0, 0, 0),$$

$$\text{corr}(X_i, X_j) = 0.5^{|i-j|}, \forall i, j,$$

$$\sigma = 3,$$

$$\varepsilon \sim \mathcal{N}(0, 1).$$

Odnosno,

$$Y = 3X_1 + 1.5X_2 + 2X_5 + 3\varepsilon.$$

Budući da je po pretpostavkama regresijskog modela greška modela nezavisna od prediktora  $X_j$ ,  $\text{Var}(Y)$  može se raščlaniti na sljedeći način:

$$\text{Var}(Y) = \text{Var}(3X_1 + 1.5X_2 + 2X_5) + \text{Var}(3\varepsilon),$$

te dalje zbog svojstava varijance i distribucije greške  $\varepsilon$  vrijedi:

$$\begin{aligned}
 \text{Var}(Y) &= \text{Var}(3X_1 + 1.5X_2 + 2X_5) + 3^2\text{Var}(\varepsilon) \\
 &= \text{Var}(3X_1 + 1.5X_2 + 2X_5) + 9 \cdot 1 \\
 &= \text{Var}(3X_1 + 1.5X_2 + 2X_5) + 9.
 \end{aligned}$$

Prema 4.5, vrijedi:

$$\text{Var}(3X_1 + 1.5X_2 + 2X_5) = E \left[ (3X_1 + 1.5X_2 + 2X_5 - E[3X_1 + 1.5X_2 + 2X_5])^2 \right].$$

Nadalje, zbog svojstva da prediktori imaju očekivanje 0, vrijedi:

$$E[3X_1 + 1.5X_2 + 2X_5] = 3E[X_1] + 1.5E[X_2] + 2E[X_5] = 0,$$

pa je:

$$\begin{aligned}
 \text{Var}(3X_1 + 1.5X_2 + 2X_5) &= E \left[ (3X_1 + 1.5X_2 + 2X_5)^2 \right] \\
 &= E \left[ 9X_1^2 + 2.25X_2^2 + 4X_5^2 + 9X_1X_2 + 12X_1X_5 + 6X_2X_5 \right] \\
 &= 9\text{Var}(X_1) + 2.25\text{Var}(X_2) + 4\text{Var}(X_5) + 9\text{Cov}(X_1, X_2) \\
 &\quad + 12\text{Cov}(X_1, X_5) + 6\text{Cov}(X_2, X_5) \\
 &= 9 \cdot 1 + 2.25 \cdot 1 + 4 \cdot 1 + 9 \cdot 0.5 + 12 \cdot 0.5^4 + 6 \cdot 0.5^3 \\
 &= 21.25.
 \end{aligned}$$

Konačno:

$$\text{Var}(Y) = 21.25 + 9 = 30.25.$$

Sada slijedi za svaki prediktor koji ulazi u model, izračunati  $\text{Cov}(X, Y)$  kako bi se moglo doći do očekivane vrijednosti  $\rho_{X, Y}$ .

$$\begin{aligned}
 \text{Cov}_{X_1, Y} &= E \left[ (X_1 - E[X_1])(Y - E[Y]) \right] \\
 &\stackrel{E[X_1]=0}{=} E \left[ X_1(Y - E[Y]) \right] \\
 &= E \left[ X_1Y - X_1E[Y] \right] \\
 &= E \left[ X_1Y \right] - E[X_1]E[Y] \\
 &\stackrel{E[X_1]=0}{=} E \left[ X_1Y \right] \\
 &= E \left[ X_1(3X_1 + 1.5X_2 + 2X_5 + \varepsilon) \right] \\
 &= E \left[ 3X_1^2 + 1.5X_1X_2 + 2X_1X_5 + X_1\varepsilon \right]
 \end{aligned}$$



$$\begin{aligned}
&= 3\text{Var}(X_1) + 1.5\text{Cov}(X_1, X_2) + 2\text{Cov}(X_1, X_5) + \text{Cov}(X_1, \varepsilon) \\
&= 3 \cdot 1 + 1.5 \cdot 0.5 + 2 \cdot 0.5^4 + 0 \\
&= 3.875
\end{aligned}$$

$$E[\rho_{X_1, Y}] = E\left[\frac{\text{Cov}(X_1, Y)}{\sqrt{\text{Var}(Y)}}\right] = E\left[\frac{3.875}{\sqrt{30.25}}\right] = 0.7045.$$

Na analogan način se za  $X_2$  i  $X_5$  dobije sljedeće:

$$\begin{aligned}
\text{Cov}_{X_2, Y} &= E[(X_2 - E[X_2])(Y - E[Y])] \\
&= \dots \\
&= E[X_2 Y] \\
&= E[X_2(3X_1 + 1.5X_2 + 2X_5 + \varepsilon)] \\
&= E[3X_1 X_2 + 1.5X_2^2 + 2X_1 X_5 + X_2 \varepsilon] \\
&= 3\text{Cov}(X_1, X_2) + 1.5\text{Var}(X_2) + 2\text{Cov}(X_2, X_5) + \text{Cov}(X_1, \varepsilon) \\
&= 3 \cdot 0.5 + 1.5 \cdot 1 + 2 \cdot 0.5^3 + 0 \\
&= 3.25
\end{aligned}$$

$$E[\rho_{X_2, Y}] = E\left[\frac{\text{Cov}(X_2, Y)}{\sqrt{\text{Var}(Y)}}\right] = E\left[\frac{3.25}{\sqrt{30.25}}\right] = 0.5909$$

$$\begin{aligned}
\text{Cov}_{X_5, Y} &= E[(X_5 - E[X_5])(Y - E[Y])] \\
&= \dots \\
&= E[X_5 Y] \\
&= E[X_5(3X_1 + 1.5X_2 + 2X_5 + \varepsilon)] \\
&= E[3X_1 X_5 + 1.5X_2 X_5 + 2X_5^2 + X_5 \varepsilon] \\
&= 3\text{Cov}(X_1, X_2) + 1.5\text{Cov}(X_2, X_5) + 2\text{Var}(X_5) + \text{Cov}(X_5, \varepsilon) \\
&= 3 \cdot 0.5^4 + 1.5 \cdot 0.5^3 + 2 \cdot 1 + 0 \\
&= 2.375
\end{aligned}$$

$$E[\rho_{X_5, Y}] = E\left[\frac{\text{Cov}(X_5, Y)}{\sqrt{\text{Var}(Y)}}\right] = E\left[\frac{2.375}{\sqrt{30.25}}\right] = 0.4318.$$

Prema tome, obzirom na tumačenja koja slijede iz tablice 4.5, značajkama  $X_1$  i  $X_2$  dodjeljuje se kategorija *Jak utjecaj*, značajki  $X_5$  kategorija *Srednji utjecaj*, a ostalim značajkama kategorija *Nema utjecaja*.

### 5.2.2. Procjena penalizacijskih koeficijenta algoritmom pretraživanja kukavice

Budući da je u istraživanju simulirano po 50 skupova podataka iz ovog scenarija te da su se parametri regularizacije optimizirali zasebno za svaki od njih i za svaku metodu regularizacije, distribucije normaliziranih penalizacijskih koeficijenta su prikazane grafički (slike 5.6 i 5.7) i deskriptivno (tablica 5.11) u nastavku.

Tablica 5.11: Scenarij 1 - deskriptivna analiza procjene penalizacijskog koeficijenta ovisno o vrsti regularizacije (normalizirane vrijednosti)

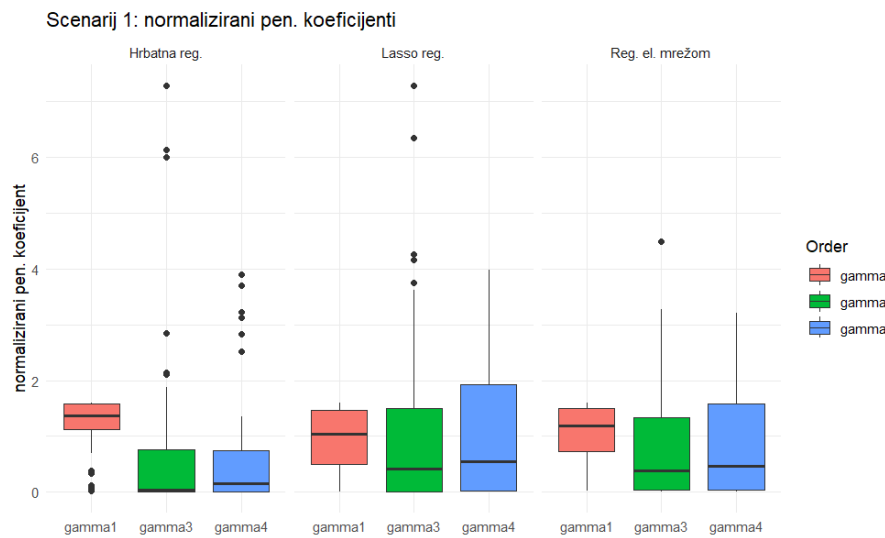
Metoda	Pen. koef.	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
hrbatna reg.	$\gamma_1$	0.026	1.120	1.376	1.179	1.576	1.596	0.520
	$\gamma_3$	0.001	0.008	0.045	0.797	0.763	7.284	1.599
	$\gamma_4$	0.001	0.010	0.148	0.655	0.746	3.891	1.040
lasso reg.	$\gamma_1$	0.004	0.502	1.037	0.948	1.474	1.594	0.537
	$\gamma_3$	0.001	0.014	0.425	1.100	1.503	7.284	1.681
	$\gamma_4$	0.000	0.029	0.555	1.080	1.922	3.972	1.200
reg. el. mrežom	$\gamma_1$	0.026	0.729	1.191	1.083	1.504	1.596	0.445
	$\gamma_3$	0.000	0.035	0.380	0.834	1.337	4.489	1.085
	$\gamma_4$	0.001	0.038	0.461	0.876	1.586	3.209	0.992

$\gamma_1, \gamma_3, \gamma_4$ : penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja*, *Srednji utjecaj* i *Jak utjecaj*, redom.

U tablici 5.11 vidljivo je kako su za svaku metodu regularizacije vrijednosti medijana procjena penalizacijskih koeficijenata nižeg iznosa za  $\gamma_4$  nego za  $\gamma_1$ , dok su za  $\gamma_3$  najmanje.

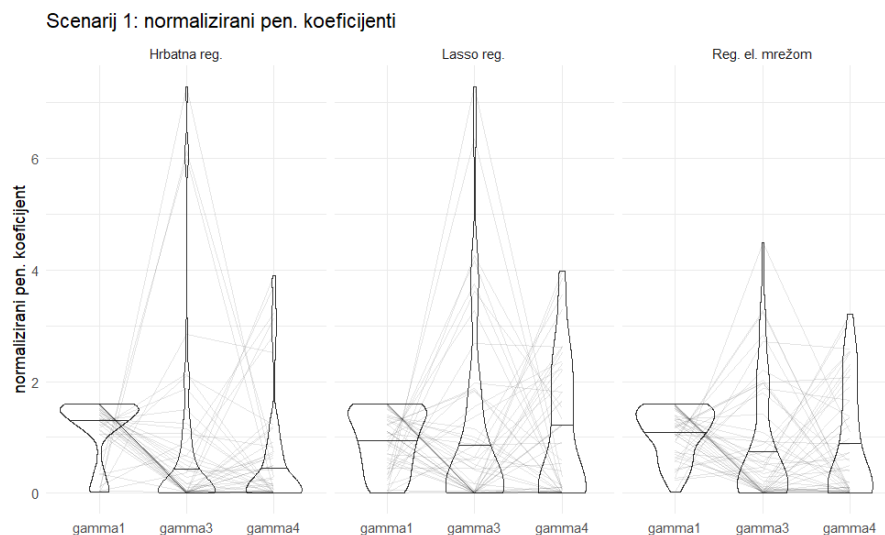
Nadalje, na slici 5.6 prikazani su kutijasti dijagrami penalizacijskih koeficijenata. Vidljivo je da lasso regularizacija rezultira najvećom varijancom. Na slici 5.7 prikazane su distribucije procjena penalizacijskih koeficijenata violinskim dijagramom te, na njima, najčešći profili pe-

nalizacijskih koeficijenta po skupovima, pri čemu jedan profil podrazumijeva procjene svih penalizacijskih koeficijenta za jedan skup podataka iz Scenarija i jednu metodu regularizacije.



Slika 5.6: Kutijasti dijagram distribucija procjene penalizacijskih koeficijenta za skupove iz scenarija 1.

$\gamma_1, \gamma_3, \gamma_4$ : normalizirani penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja*, *Srednji utjecaj* i *Jak utjecaj*, redom.



Slika 5.7: Violinski dijagram distribucija procjene penalizacijskih koeficijenta s profilima procjena za skupove iz scenarija 1.

$\gamma_1, \gamma_3, \gamma_4$ : normalizirani penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja*, *Srednji utjecaj* i *Jak utjecaj*, redom.

Premda je bilo profila gdje su značajkama snažnijeg utjecaja procijenjeni veći penalizacijski koeficijent nego značajkama slabijeg utjecaja, vidljivo je da su to rjeđe situacije. Češća je situacija bila da su procijenjeni manji penalizacijski koeficijenti značajkama snažnijeg utjecaja i obratno.

### 5.2.3. Procjena koeficijena modela

U nastavku su prikazani rezultati procjene koeficijena modela iz Scenarija 1 za slobodni koeficijent te za prediktore s koeficijentima različitim od nule. Za preostale koeficijente su se i distribucije procjena kretale oko nule, ali radi preglednosti njihovi su grafički prikazi i deskriptivna analiza izostavljeni iz rada.

Tablica 5.12: Scenarij 1 - deskriptivna analiza procjene koeficijena modela ovisno o vrsti regularizacije i načinu procjene penalizacijskog koeficijenta

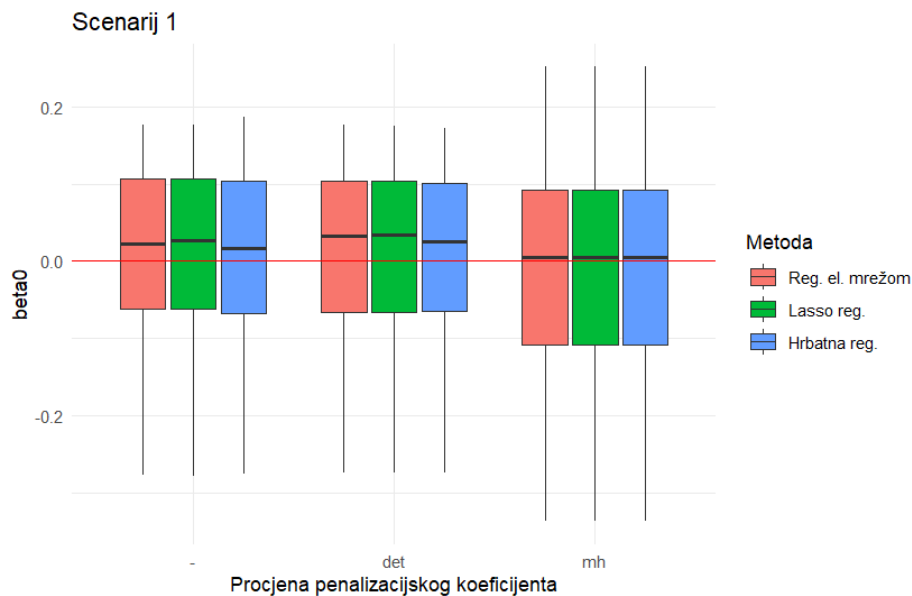
Koef.	Procjena pen. koef.	Metoda	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
$\beta_0 = 0$	-	lasso reg.	-0.28	<b>-0.06</b>	0.02	0.01	0.11	0.18	<b>0.11</b>
		hrbatna reg.	-0.28	<b>-0.06</b>	0.03	0.01	0.11	0.18	<b>0.11</b>
		reg. el. mrežom	-0.28	-0.07	0.02	0.01	0.10	0.19	<b>0.11</b>
	det	lasso reg.	<b>-0.27</b>	-0.07	0.03	0.01	0.10	0.18	<b>0.11</b>
		hrbatna reg.	<b>-0.27</b>	-0.07	0.03	0.01	0.10	<b>0.17</b>	<b>0.11</b>
		reg. el. mrežom	<b>-0.27</b>	<b>-0.06</b>	0.03	0.01	0.10	<b>0.17</b>	<b>0.11</b>
	mh	lasso reg.	-0.34	-0.11	<b>0.00</b>	<b>0.00</b>	<b>0.09</b>	0.25	0.13
		hrbatna reg.	-0.34	-0.11	<b>0.00</b>	<b>0.00</b>	<b>0.09</b>	0.25	0.13
		reg. el. mrežom	-0.34	-0.11	<b>0.00</b>	<b>0.00</b>	<b>0.09</b>	0.25	0.13
$\beta_1 = 3$	-	lasso reg.	2.35	2.60	2.65	2.64	2.71	2.92	0.12
		hrbatna reg.	2.41	2.63	2.72	2.71	2.79	<b>3.00</b>	0.13
		reg. el. mrežom	2.19	2.39	2.46	2.47	2.55	2.80	0.13
	det	lasso reg.	<b>2.60</b>	2.78	2.88	2.86	2.93	3.12	<b>0.11</b>
		hrbatna reg.	<b>2.60</b>	<b>2.83</b>	2.90	2.90	2.96	3.15	0.12
		reg. el. mrežom	2.40	2.58	2.64	2.64	2.69	2.91	<b>0.11</b>
	mh	lasso reg.	2.52	2.80	2.95	2.91	<b>3.00</b>	3.24	0.17
		hrbatna reg.	<b>2.60</b>	<b>2.83</b>	<b>2.96</b>	<b>2.94</b>	3.01	3.25	0.15

Koef.	Procjena pen. koef.	Metoda	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
		reg. el. mrežom	2.51	2.78	2.90	2.90	<b>3.00</b>	3.25	0.18
$\beta_2 = 1.5$	-	lasso reg.	1.07	1.26	1.32	1.33	1.38	1.58	0.11
		hrbatna reg.	0.97	1.17	1.25	1.25	1.32	<b>1.57</b>	0.13
		reg. el. mrežom	1.17	1.32	1.38	1.38	1.42	1.63	<b>0.10</b>
	det	lasso reg.	<b>1.27</b>	1.42	<b>1.50</b>	1.49	1.56	1.74	0.11
		hrbatna reg.	1.18	1.40	1.48	1.47	<b>1.54</b>	1.73	0.12
		reg. el. mrežom	1.26	<b>1.45</b>	1.51	<b>1.50</b>	1.55	1.72	<b>0.10</b>
	mh	lasso reg.	0.99	1.34	1.44	1.43	<b>1.54</b>	1.79	0.15
		hrbatna reg.	0.93	1.32	1.43	1.42	<b>1.54</b>	1.78	0.17
		reg. el. mrežom	1.20	1.40	1.48	1.47	1.56	1.77	0.12
$\beta_5 = 2$	-	lasso reg.	<b>1.39</b>	1.48	1.60	1.61	1.72	<b>2.00</b>	0.15
		hrbatna reg.	1.36	1.51	1.60	1.61	1.70	2.02	0.15
		reg. el. mrežom	1.27	1.38	1.47	1.51	1.58	2.06	0.16
	det	lasso reg.	1.14	1.32	1.42	1.44	1.55	1.91	0.17
		hrbatna reg.	1.10	1.30	1.41	1.42	1.53	1.88	0.17
		reg. el. mrežom	1.03	1.12	1.18	1.21	1.28	1.67	<b>0.13</b>
	mh	lasso reg.	1.23	1.76	1.93	1.88	2.05	2.40	0.21
		hrbatna reg.	1.21	<b>1.77</b>	1.93	<b>1.89</b>	2.06	2.41	0.22
		reg. el. mrežom	1.23	1.69	<b>1.94</b>	1.88	<b>2.03</b>	2.45	0.25

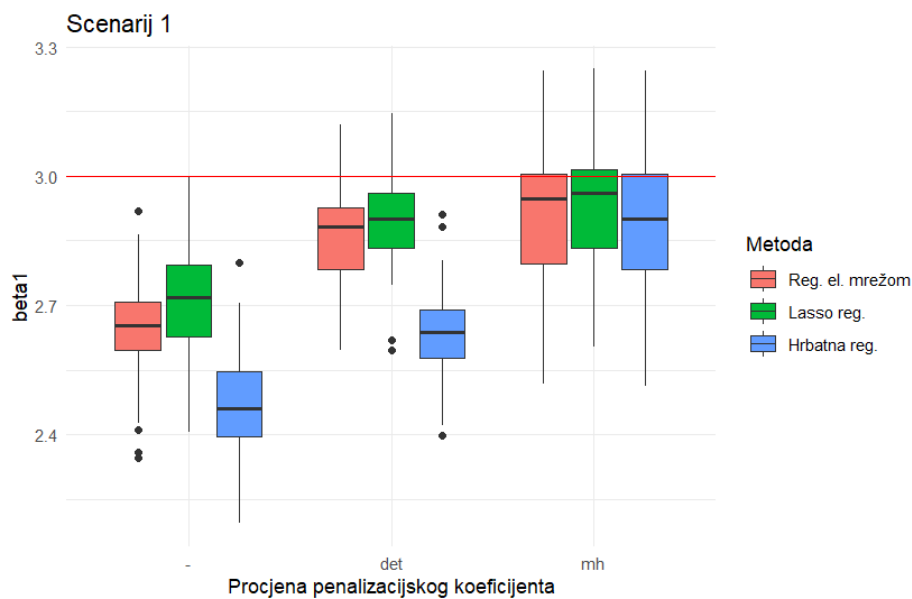
Najtočnija procjena koeficijenta modela za svaku deskriptivnu statistiku označena je podebljano.

Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

U nastavku su grafički prikazane distribucije procjene koeficijenta. Vidljivo je da se za sve metode regularizacije procijenjena vrijednost slobodnog koeficijenta  $\beta_0$  kretala oko nule, dok je u slučaju uvođenja domenskog znanja metaheuristikom imala medijan najbliže nuli (slika 5.8). Nadalje, na slici 5.9 vidljivo je kako je za sve metode regularizacije uvođenje domenskog znanja stručnjaka metaheuristikom rezultiralo procjenom vrijednosti koeficijenta  $\beta_1$  najbližoj stvarnoj vrijednosti (koja iznosi 3), te je potom uvođenje domenskog znanja stručnjaka deterministički bilo točnije od situacije kada se ne koristi domensko znanje.



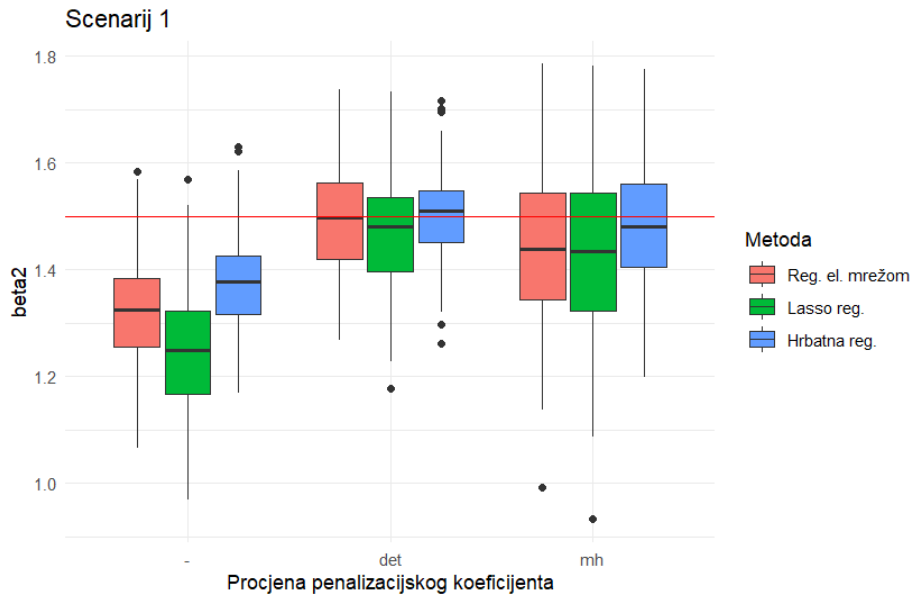
Slika 5.8: Grafički prikazi distribucija procjene koeficijenta  $\beta_0$  za skupove iz Scenarija 1. Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). Crveni pravac predstavlja prave vrijednosti koeficijenata.



Slika 5.9: Grafički prikazi distribucija procjene koeficijenta  $\beta_1$  za skupove iz Scenarija 1. Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). Crveni pravac predstavlja prave vrijednosti koeficijenata.

Za sve metode regularizacije procjena je vrijednosti koeficijenta  $\beta_2$  bila točnija u slučajevima kada se uvodi domensko znanje stručnjaka nego kada se ono ne koristi (slika 5.10). Konačno, uvođenje domenskog znanja metaheuristikom rezultirala je procjenom vrijednosti

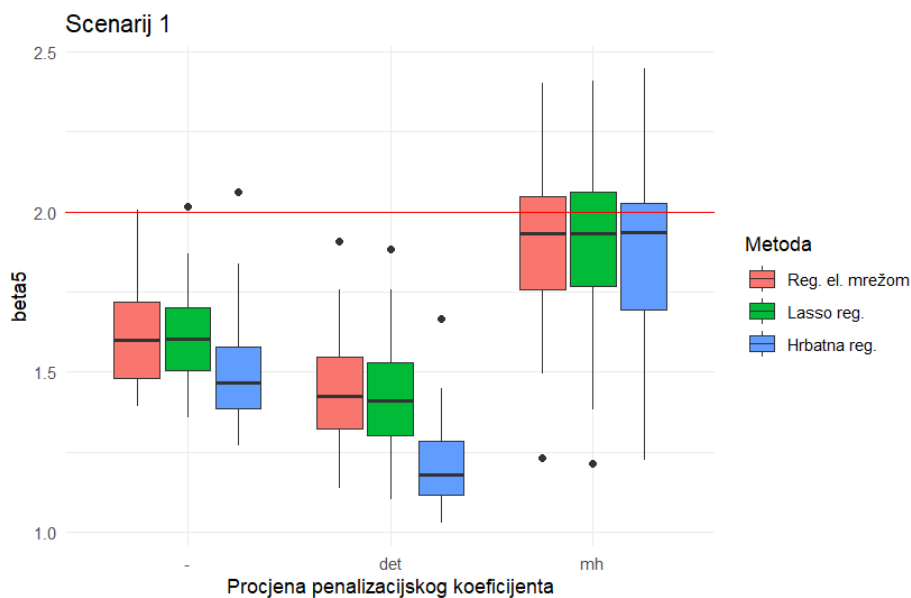
koeficijenta  $\beta_5$  najbližoj stvarnoj vrijednosti (koja iznosi 2), te je potom uvođenje domenskog znanja deterministički bilo manje točno od situacije kada se ne koristi domensko znanje (slika 5.11).



Slika 5.10: Grafički prikazi distribucija procjene koeficijenta  $\beta_2$  za skupove iz Scenarija 1.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Crveni pravac predstavlja prave vrijednosti koeficijenata.



Slika 5.11: Grafički prikazi distribucija procjene koeficijenta  $\beta_5$  za skupove iz Scenarija 1.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Crveni pravac predstavlja prave vrijednosti koeficijenata.

No, kao i kod stvarnih podataka, i na skupovima simuliranim po Scenariju 1 metode regularizacije s metaheurističkim pristupom procjene penalizacijskog koeficijenta imale su dulje vrijeme izvođenja te je u tu svrhu korištena usluga Naprednog računanja (Padobran) Sveučilišnog računskog centra (Sree) Sveučilišta u Zagrebu. U tablici 5.13 u nastavku prikazano je prosječno vrijeme izvođenja za jedan skup prilikom kojeg se koristilo pet procesora.

Tablica 5.13: Prosječno vrijeme izvođenja metoda regularizacije s metaheurističkom procjenom penalizacijskog koeficijenta za skupove iz Scenarija 1.

Metoda	lasso reg.	hrbatna reg.	reg. el. mrežom
Vrijeme	07:26	07:06	06:07

Prilikom izvođenja za jedan skup korišteno je pet procesora, a prikazano vrijeme odnosi se na minute i sekunde.

#### 5.2.4. Evaluacija modela i hipoteza

U potpoglavlju 4.5 (str. 50), objašnjeno kako su metodom ANOVA analizirani rezultati modela svih 50 skupova iz Scenarija 1 s ciljem donošenja odluke o hipotezama. No, prije svega, potrebno je provjeriti zadovoljenost pretpostavki takve analize. Provjera je izvršena analizom dijagrama raspršenosti reziduala modela u odnosu na razine faktora te qq-dijagramom distribucije reziduala u usporedbi sa standardnom normalnom distribucijom za svaku mjeru evaluacije modela unutar pojedine hipoteze. Spomenuti su dijagrami, radi preglednosti, prikazani u Prilogu 2 (slike 7.7, 7.8 i 7.9), a s obzirom na to da nisu upućivali na značajna odstupanja od ovih pretpostavki, smatra se valjanim koristiti analizu varijance. Odluke se donose na razini statističke značajnosti 0.05.

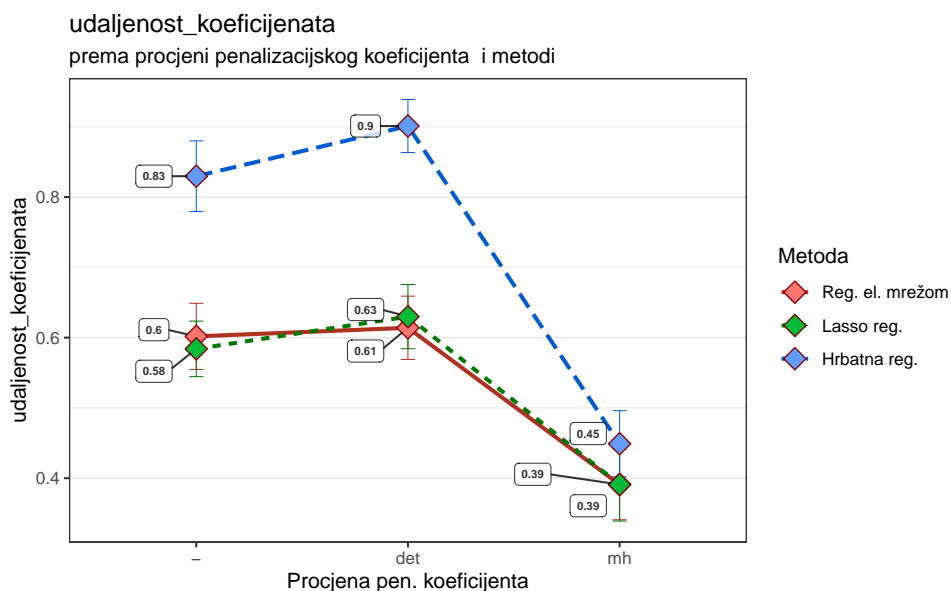
#### Hipoteza 1

U ovoj hipotezi uspoređuju se modeli regularizacije koji ne uključuju znanje stručnjaka s modelima regularizacije koji ga uključuju deterministički i metaheuristički kroz pet različitih mjera evaluacije. Podsjetimo, za evaluaciju hipoteze koristi se dvofaktorska ANOVA s faktorima Procjena penalizacijskog koeficijenta i Metoda regularizacije te njihovom interakcijom. Kontrasti su pojašnjeni u potpoglavlju 4.5 (str. 50), pri čemu se pomoću kontrasta c11 testira upravo Hipoteza 1 ovog istraživanja, dok kontrast c12 uspoređuje međusobno dva načina uvođenja



domenskog znanja stručnjaka u metode regularizacije.

U nastavku su za svaku mjeru evaluacije prikazani interakcijski grafikoni metode regularizacije i različitih načina procjene penalizacijskog koeficijenta.



Slika 5.12: Scenarij 1, Hipoteza 1 - Udaljenost koeficijenata prema procjeni penalizacijskog koeficijenta i metodi.

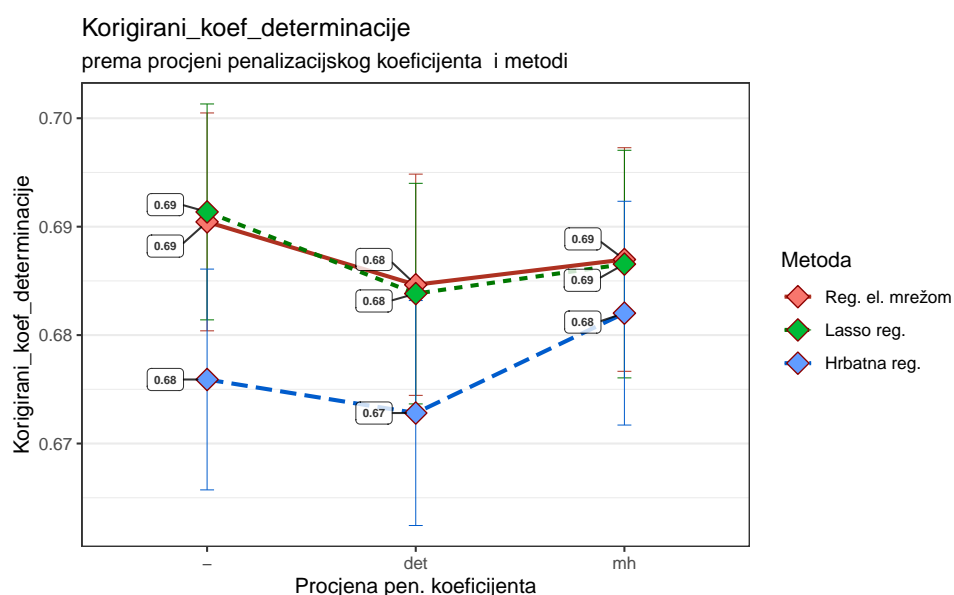
Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Na slici 5.12 prikazane su prosječne vrijednosti udaljenosti koeficijenata prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako za sve metode regularizacije uvođenjem domenskog znanja na deterministički način raste prosječna vrijednost udaljenosti koeficijenata, a onda opadne kada se domensko znanje uvede metaheuristikom (ispod razine slučaja kada se ne koristi domensko znanje stručnjaka). Rezultati se razlikuju i između metode, hrbatna regularizacija ima najvišu prosječnu udaljenost koeficijenata, dok su lasso regularizacija i regularizacija elastičnom mrežom sličnih vrijednosti.

Za mjeru evaluacije udaljenost koeficijenata statistički značajan je interakcijski član ( $F(4, 441) = 8.795$ ,  $p = 7.73 \text{ e-}07$ ), kao i efekti faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 153.84$ ,  $p < 2\text{e-}16$ ) i faktora Metoda regularizacije ( $F(2, 441) = 69.086$ ,  $p < 2\text{e-}16$ ). Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.3 u Prilogu 2 (str. 175).

Budući da je interakcijski član dvaju faktora statistički značajan, efekt pojedinog načina uvođenja domenskog znanja stručnjaka ne može biti generaliziran na sve metode regularizacije.

U interpretaciji rezultata potrebno je koristiti *post hoc* analizu interakcije. U tablici 7.4 u Prilogu 2 (str. 176) prikazan je rezultat *post hoc* analize Tukeyjevim testom HSD te vidljivo je koje su razlike statistički značajno različite od nule na razini statističke značajnosti 0.05. Vidljivo je da je za svaku metodu regularizacije statistički značajna razlika prosječne udaljenosti koeficijenata između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka. Negativan iznos razlike upućuje da je prosječna udaljenost koeficijenata manja kod metaheuristike. Niti za jednu metodu regularizacije nije statistički značajna razlika prosječne udaljenosti koeficijenata između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka. Konačno, kod svake je metode regularizacije statistički značajna razlika prosječne udaljenosti koeficijenata između metaheurističkog pristupa i determinističkog pristupa uvođenja domenskog znanja. Negativan iznos razlike upućuje da je prosječna udaljenost koeficijenata manja kod metaheuristike.



Slika 5.13: Scenarij 1, Hipoteza 1 - Korigirani koeficijent determinacije prema procjeni penalizacijskog koeficijenta i metodi.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

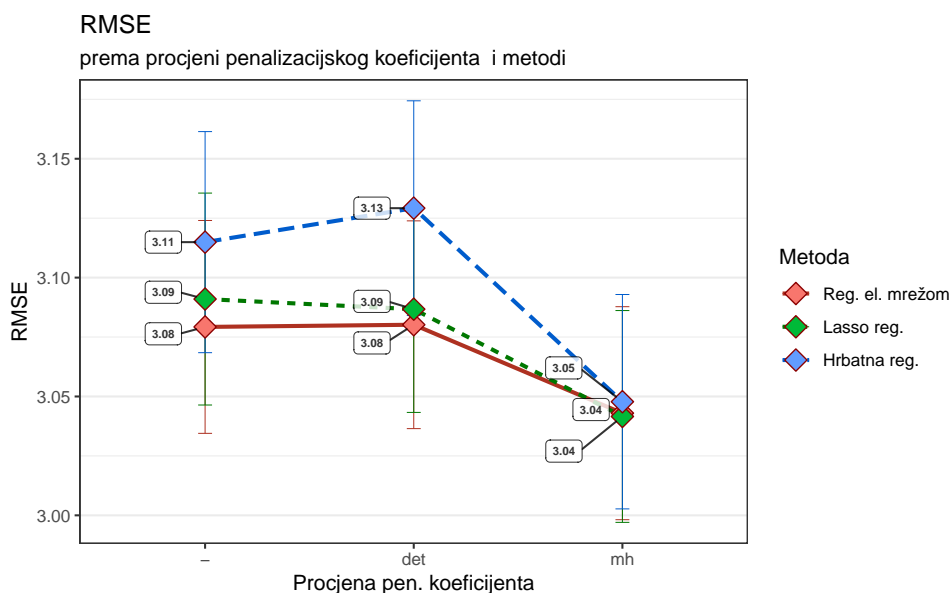
Na slici 5.13 prikazane su prosječne vrijednosti korigiranog koeficijenta determinacije prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako uvođenjem domenskog znanja na deterministički način opada prosječna vrijednost korigiranog koeficijenta determinacije, a onda ponovno raste kada se domensko znanje uvede metaheuristikom, no za lasso regularizaciju i regularizaciju elastičnom mrežom to je ispod razine slučaja kada se ne

koristi domensko znanje stručnjaka, a za hrbatnu regularizaciju iznad. Vidljivo je da, od svih metoda regularizacije, hrbatna regularizacija ima najmanji prosječni korigirani koeficijent determinacije, neovisno o načinu procjene penalizacijskih koeficijenata.

Za mjeru evaluacije korigirani koeficijent determinacije nije statistički značajan interakcijski član ( $F(4, 441) = 0.357, p = 0.8393$ ), kao ni efekt faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 1.024, p = 0.3599$ ), dok efekt faktora Metoda regularizacije statistički je značajan ( $F(2, 441) = 4.162, p = 0.0162$ ).

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.3 u Prilogu 2 (str. 175). U spomenutoj tablici vidljivo je kako jedino koeficijent uz kontrast c22 faktora Metoda ne odstupa statistički značajno od 0. Negativna vrijednost koeficijenta upućuje da je prosječna vrijednost korigiranog koeficijenta determinacije za lasso regularizaciju veća nego za hrbatnu regularizaciju.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika u prosječnom korigiranom koeficijentu determinacije između hrbatne regularizacije i regularizacije elastičnom mrežom negativna i statistički značajna (diff = -0.0104,  $p = 0.0331$ ), dok je razlika između prosječnog korigiranog koeficijenta determinacije lasso regularizacije i regularizacije elastičnom mrežom negativna i nije statistički značajna (diff = -0.0001,  $p = 0.9996$ ).



Slika 5.14: Scenarij 1, Hipoteza 1 - *RMSE* prema procjeni penalizacijskog koeficijenta i metodi. Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

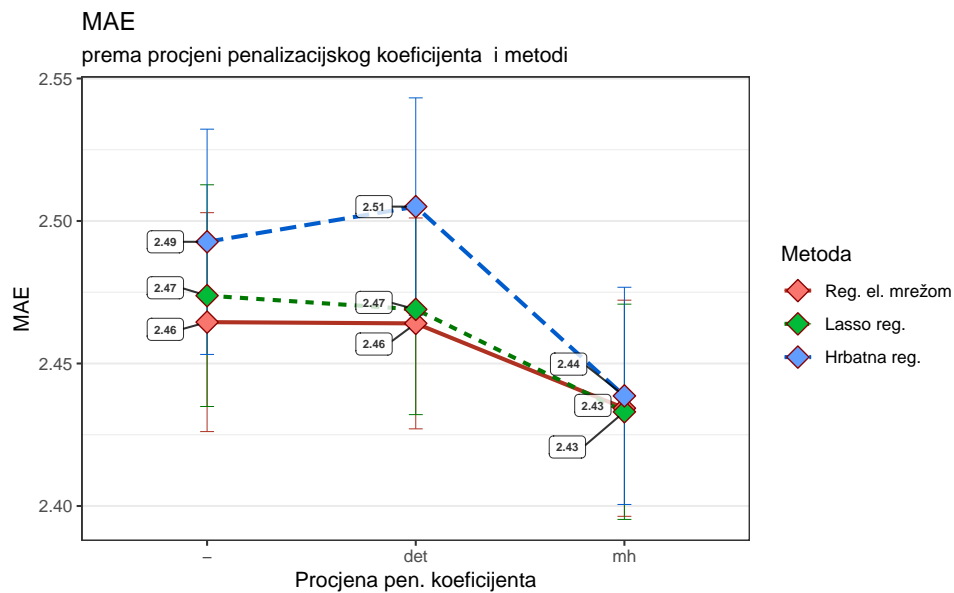
Na slici 5.14 prikazane su prosječne vrijednosti *RMSE* prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako za hrbatnu regularizaciju uvođenjem domenskog znanja stručnjaka na deterministički način raste prosječna vrijednost *RMSE*, dok ona opada uvođenjem domenskog znanja metaheuristikom. Za lasso regularizaciju i regularizaciju elastičnom mrežom riječ je o manjoj promjeni prosječne *RMSE* kod uvođenja domenskog znanja deterministički. No, uvođenjem domenskog znanja metaheuristikom, također opada prosječna vrijednost greške.

Za mjeru evaluacije *RMSE* nije statistički značajan interakcijski član ( $F(4, 441) = 0.298$ ,  $p = 0.87914$ ), kao niti efekt faktora Metoda regularizacije ( $F(2, 441) = 1.522$ ,  $p = 0.21943$ ). Efekt faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 5.636$ ,  $p = 0.00383$ ) je statistički značajan.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.3 u Prilogu 2 (str. 175). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c12 koji govori o razlici *RMSE* između metoda regularizacije koje penalizacijske koeficijente procjenjuju deterministički s onima koje ih procjenjuju metaheuristikom. Negativan iznos koeficijenta upućuje da je prosječna *RMSE* niža u slučaju uvođenja domenskog znanja stručnjaka metaheuristikom, u odnosu na determinističko uvođenje.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *RMSE* između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka pozitivna, ali nije statistički značajna (diff = 0.004,  $p = 0.9780$ ). Razlika prosječnih *RMSE* između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna je i statistički značajna (diff = -0.051,  $p = 0.0146$ ).

Na slici 5.15 prikazane su prosječne vrijednosti *MAE* prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako za hrbatnu regularizaciju uvođenjem domenskog znanja stručnjaka na deterministički način raste prosječna vrijednost *MAE*, dok ona opada uvođenjem domenskog znanja metaheuristikom. Za lasso regularizaciju i regularizaciju elastičnom mrežom riječ je o manjoj promjeni prosječne *MAE* kod uvođenja domenskog znanja deterministički. No, prosječna vrijednost greške opada uvođenjem domenskog znanja metaheuristikom.



Slika 5.15: Scenarij 1, Hipoteza 1 - *MAE* prema procjeni penalizacijskog koeficijenta i metodi. Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

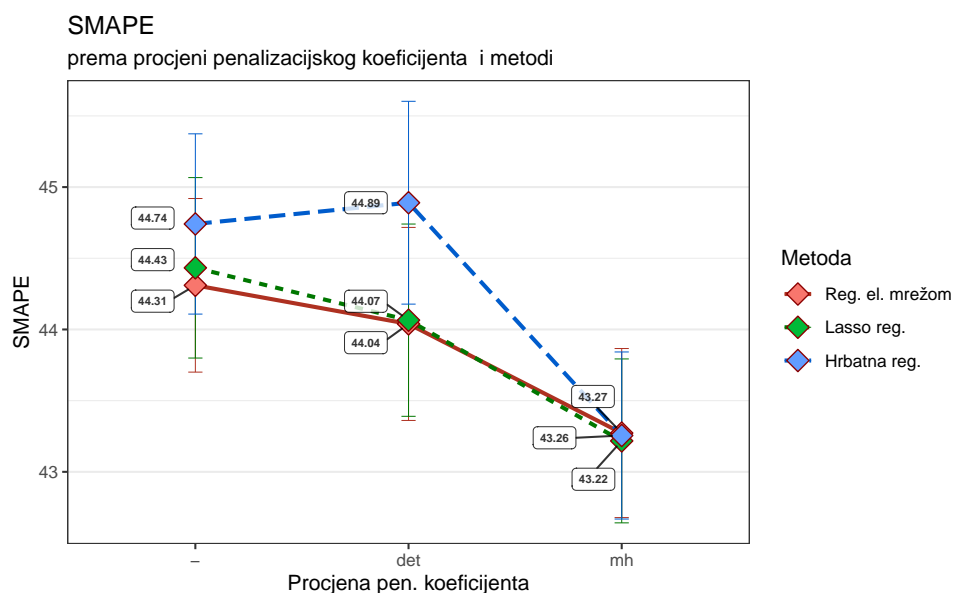
Za mjeru evaluacije *MAE* nisu statistički značajni interakcijski član ( $F(4, 441) = 0.282$ ,  $p = 0.88986$ ) i efekt faktora Metoda regularizacije ( $F(2, 441) = 1.428$ ,  $p = 0.24079$ ), dok efekt faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 5.125$ ,  $p = 0.00631$ ) statistički je značajan.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.3 u Prilogu 2 (str. 175). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c12 koji govori o razlici prosječnih *MAE* između metoda regularizacije koje penalizacijske koeficijente procjenjuju deterministički s onima koje ih procjenjuju metaheuristikom. Negativan iznos koeficijenta upućuje da je prosječna *MAE* niža u slučaju uvođenja domenskog znanja stručnjaka metaheuristikom, u odnosu na determinističko uvođenje.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *MAE* između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka pozitivna, ali nije statistički značajna (diff = 0.002,  $p = 0.9873$ ). Razlika prosječnih *MAE* između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna je i statistički značajna (diff = -0.042,  $p = 0.0201$ ).

Na slici 5.16 prikazane su prosječne vrijednosti *SMAPE* prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako za hrbatnu regularizaciju uvođenjem do-

menskog znanja stručnjaka na deterministički način raste prosječna vrijednost *SMAPE*, dok opada uvođenjem domenskog znanja metaheuristikom. Za lasso regularizaciju i regularizaciju elastičnom mrežom riječ je o smanjenju prosječne *SMAPE* kod uvođenja domenskog znanja deterministički, te o dodatnom smanjenju kod uvođenja domenskog znanja metaheuristikom.



Slika 5.16: Scenarij 1, Hipoteza 1 - *SMAPE* prema procjeni penalizacijskog koeficijenta i metodi.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Za mjeru evaluacije *SMAPE* nisu statistički značajni interakcijski član ( $F(4, 441) = 0.596$ ,  $p = 0.666$ ) i efekt faktora Metoda regularizacije ( $F(2, 441) = 1.660$ ,  $p = 0.191$ ), dok efekt faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 13.802$ ,  $p = 1.53 \text{ e-}06$ ) statistički je značajan.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.3 u Prilogu 2 (str. 175). U spomenutoj tablici vidljivo je kako su statistički značajni jedino koeficijenti uz kontraste c11 i c12.

Kontrast c11 govori o razlici prosječnih *SMAPE* greški između metoda regularizacije koje ne koriste domensko znanje u odnosu na one koje ga uključuju deterministički ili metaheuristikom. Pozitivan iznos koeficijenta upućuje da je prosječna *SMAPE* niža u metodama regularizacije koje koriste domensko znanje stručnjaka. Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *SMAPE* greški između determinističkog pristupa i pris-

tupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna ( $\text{diff} = -0.163$ ,  $p = 0.8019$ ). Razlika prosječnih *SMAPE* greški između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka je negativna i statistički značajna ( $\text{diff} = -1.2464$ ,  $p = 5.5 \text{ e-}6$ ).

Kontrast c12 govori o razlici prosječnih *SMAPE* greški između metoda regularizacije koje penalizacijske koeficijente procjenjuju deterministički s onima koje ih procjenjuju metaheuristikom. Negativan iznos koeficijenta upućuje da je prosječna *SMAPE* niža u slučaju uvođenja domenskog znanja stručnjaka metaheuristikom, u odnosu na determinističko uvođenje.

## Hipoteza 2

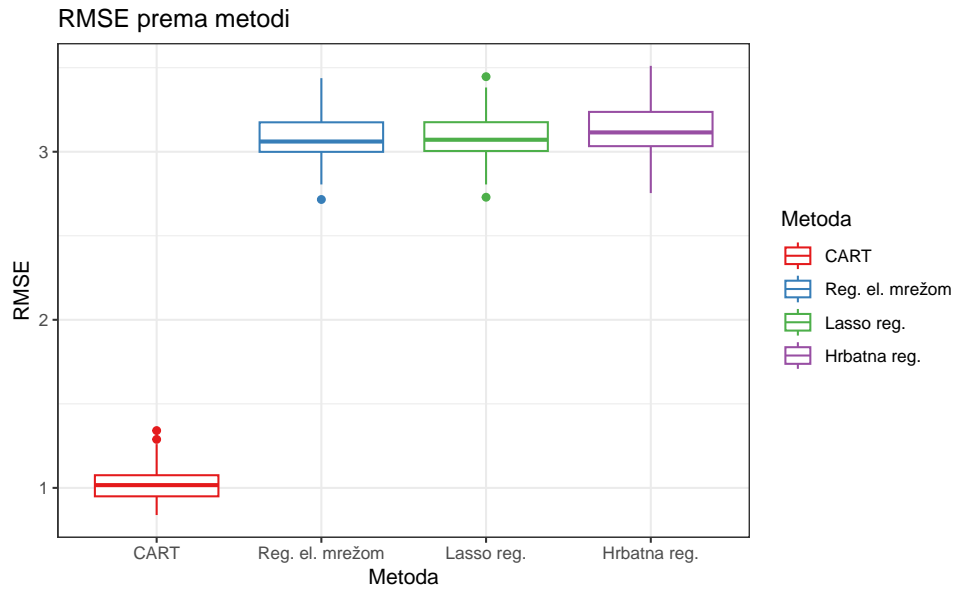
U ovoj se hipotezi kroz tri različite mjere evaluacije uspoređuju metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka. Hipoteza se dijeli na dvije podhipoteze - u jednoj se uspoređuje metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka deterministički (Hipoteza 2a), a u drugoj se metoda CART uspoređuju s metodama regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom (Hipoteza 2b).

### Hipoteza 2a

U ovoj hipotezi uspoređuje se metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka deterministički. Podsjetimo, za evaluaciju hipoteze koristi se jednofaktorska ANOVA s faktorom Metoda. Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2 ovog istraživanja. Odluke se donose na razini statističke značajnosti 0.05. U nastavku su grafički prikazane distribucije svake mjere evaluacije.

Za mjeru evaluacije *RMSE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2593$ ,  $p < 2 \text{ e-}16$ ). Na slici 5.17 vidljivo je kako CART metoda ima nižu prosječnu *RMSE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.5 u Prilogu 2 (str. 179). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *RMSE* od metoda regularizacije koje uvode domensko znanje stručnjaka deterministički.



Slika 5.17: Scenarij 1, Hipoteza 2a - *RMSE* prema metodi.

U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *RMSE* između metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna (diff = - 2.052,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 2.058,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 2.101,  $p = 0.00$ ).

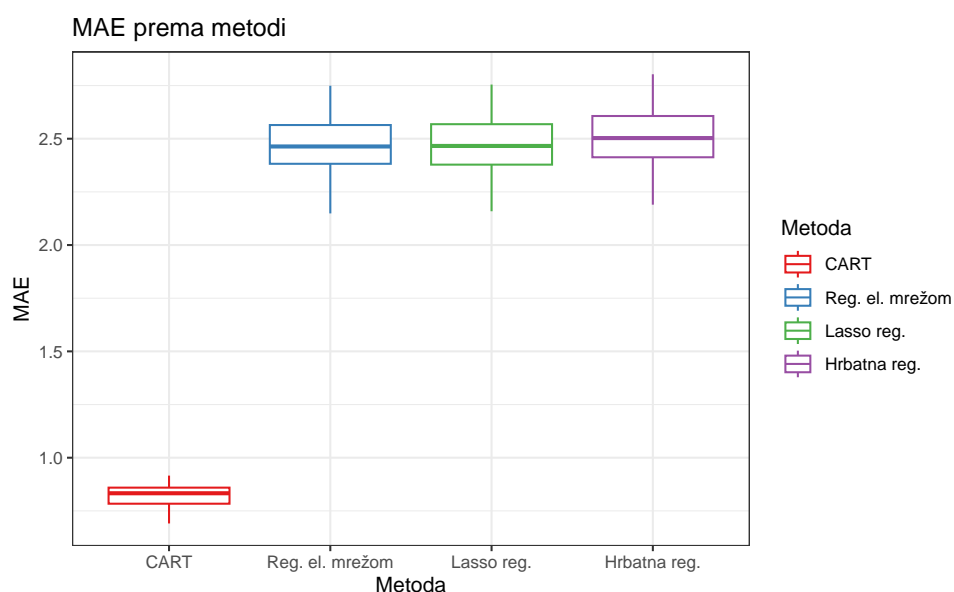
Za mjeru evaluacije *MAE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2515$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.18 vidljivo je kako CART metoda ima nižu prosječnu *MAE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.5 u Prilogu 2 (str. 179). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *MAE* od metoda regularizacije koje uvode domensko znanje stručnjaka deterministički.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *MAE* između metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna (diff = - 1.639,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 1.644,  $p = 0.00$ ) te



metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka ( $\text{diff} = -1.680$ ,  $p = 0.00$ ).



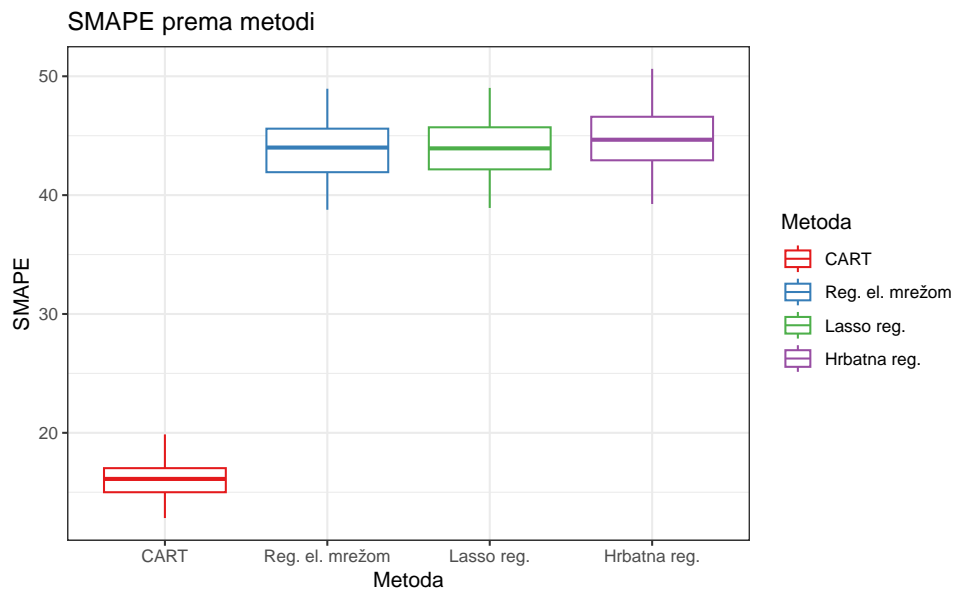
Slika 5.18: Scenarij 1, Hipoteza 2a - MAE prema metodi.

U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Za mjeru evaluacije *SMAPE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1983$ ,  $p < 2 \text{ e-}16$ ). Na slici 5.19 vidljivo je kako CART metoda ima nižu prosječnu *SMAPE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.5 u Prilogu 2 (str. 179). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *SMAPE* od metoda regularizacije koje uvode domensko znanje stručnjaka deterministički.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *SMAPE* greški između metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna ( $\text{diff} = -27.926$ ,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka ( $\text{diff} = -27.952$ ,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka ( $\text{diff} = -28.777$ ,  $p = 0.00$ ).



Slika 5.19: Scenarij 1, Hipoteza 2a - *SMAPE* prema metodi.

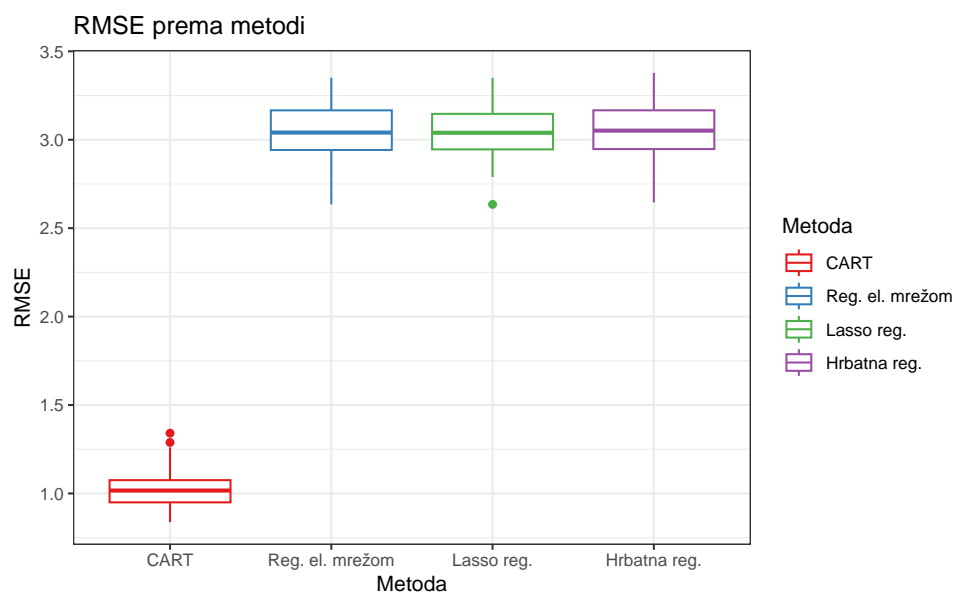
U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

### Hipoteza 2b

U ovoj hipotezi uspoređuje se metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom. Podsjetimo, za evaluaciju hipoteze koristi se jednofaktorska ANOVA s faktorom Metoda. Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2 ovog istraživanja. Odluke se donose na razini statističke značajnosti 0.05. U nastavku su grafički prikazane distribucije svake mjere evaluacije.

Za mjeru evaluacije *RMSE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2389$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.20 vidljivo je kako CART metoda ima nižu prosječnu *RMSE* od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.6 u Prilogu 2 (str. 181). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *RMSE* od metoda regularizacije koje uvode domensko znanje stručnjaka metaheuristikom.



Slika 5.20: Scenarij 1, Hipoteza 2b - *RMSE* prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

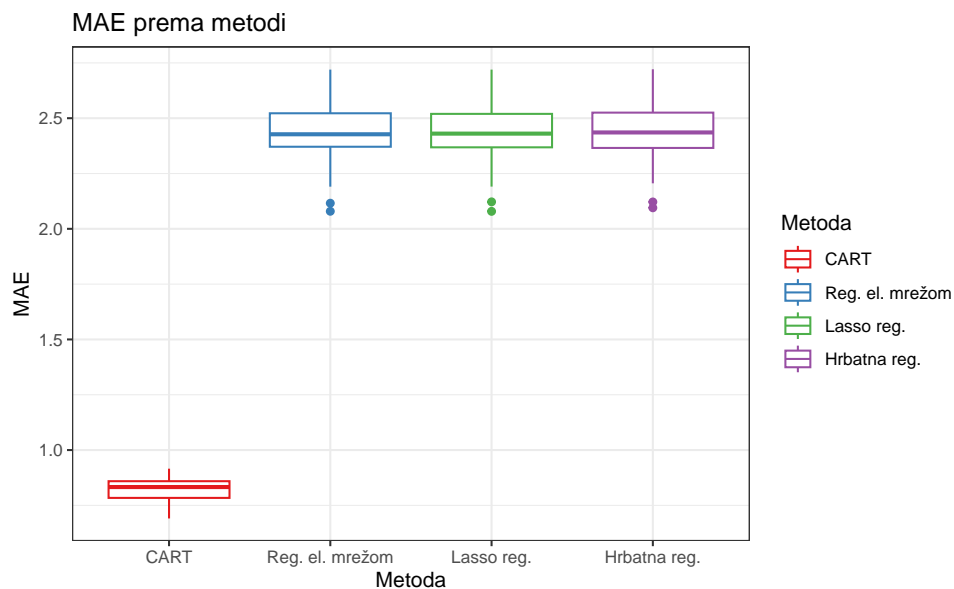
Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *RMSE* između metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 2.015,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 2.013,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 2.019,  $p = 0.00$ ).

Za mjeru evaluacije *MAE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2315$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.21 vidljivo je kako CART metoda ima nižu prosječnu *MAE* grešku od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.6 u Prilogu 2 (str. 181). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *MAE* od metoda regularizacije koje uvode domensko znanje stručnjaka metaheuristikom.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *MAE* između metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 1.61,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 1.608,  $p = 0.00$ )

te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 1.614,  $p = 0.00$ ).



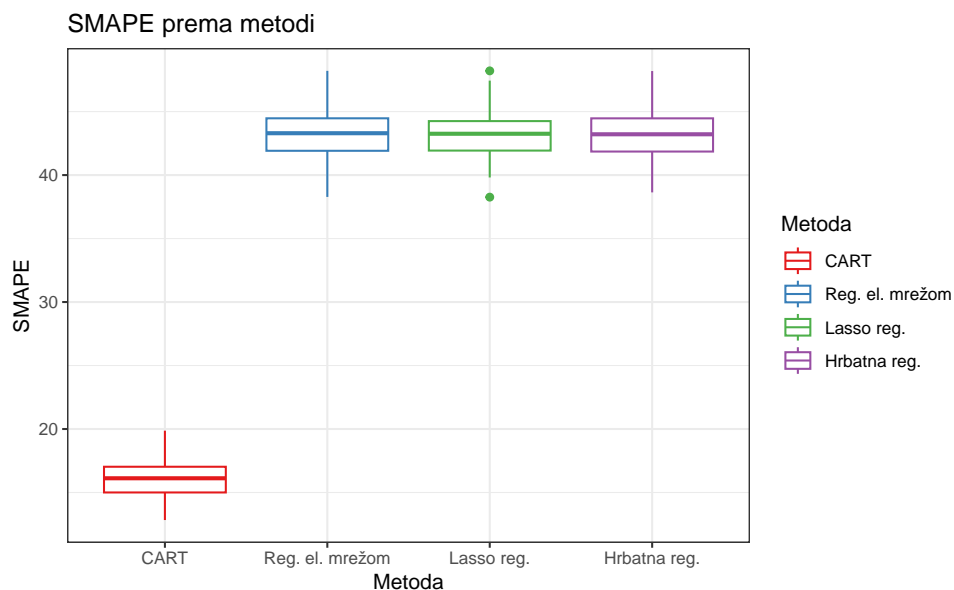
Slika 5.21: Scenarij 1, Hipoteza 2b - MAE prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

Za mjeru evaluacije *SMAPE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2421$ ,  $p < 2 \text{ e-}16$ ). Na slici 5.22 vidljivo je kako CART metoda ima nižu prosječnu *SMAPE* od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.6 u Prilogu 2 (str. 181). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *SMAPE* od metoda regularizacije koje uvode domensko znanje stručnjaka metaheuristikom.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječne *SMAPE* između metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 27.159,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 27.104,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 27.142,  $p = 0.00$ ).



Slika 5.22: Scenarij 1, Hipoteza 2b - *SMAPE* prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

## 5.3. SCENARIJ 2

### 5.3.1. Modeliranje mišljenja stručnjaka

U ovom scenariju poznato je da vrijedi:

$$p = 40,$$

$$\beta = (\underbrace{0, 0, \dots, 0}_{10}, \underbrace{2, 2, \dots, 2}_{10}, \underbrace{0, 0, \dots, 0}_{10}, \underbrace{2, 2, \dots, 2}_{10}),$$

$$\text{corr}(X_i, X_j) = 0.5, \forall i, j, i \neq j$$

$$\sigma = 15,$$

$$\varepsilon \sim \mathcal{N}(0, 1).$$

Odnosno,

$$\begin{aligned} Y &= 2X_{11} + 2X_{12} + \dots + 2X_{20} + 2X_{31} + 2X_{32} + \dots + 2X_{40} + 15\varepsilon \\ &= 2 \sum_{i \in \{11, 12, \dots, 20, 31, 32, \dots, 40\}} X_i + \varepsilon \end{aligned}$$

Radi jednostavnosti, uvodi se oznaka  $I' = \{11, 12, \dots, 20, 31, 32, \dots, 40\}$ , pa se prethodno može zapisati na sljedeći način:

$$Y = 2 \sum_{i \in I'} X_i + 15\varepsilon.$$

Budući da je po pretpostavkama regresijskog modela greška modela nezavisna od prediktora  $X_j$ ,  $Var(Y)$  može se raščlaniti na sljedeći način:

$$Var(Y) = Var\left(2 \sum_{i \in I'} X_i\right) + Var(15\varepsilon),$$

te dalje zbog svojstava varijance i distribucije greške  $\varepsilon$ , vrijedi:

$$\begin{aligned} Var(Y) &= 2^2 Var\left(\sum_{i \in I'} X_i\right) + 15^2 Var(\varepsilon) \\ &= 4 Var\left(\sum_{i \in I'} X_i\right) + 15^2 \cdot 1 \\ &= 4 Var\left(\sum_{i \in I'} X_i\right) + 225. \end{aligned}$$

Prema 4.5, vrijedi:

$$Var\left(\sum_{i \in I'} X_i\right) = E \left[ \left( \sum_{i \in I'} X_i - E\left[\sum_{i \in I'} X_i\right] \right)^2 \right].$$

Nadalje, zbog svojstva da prediktori imaju očekivanje 0, vrijedi:

$$E\left[\sum_{i \in I'} X_i\right] = \sum_{i \in I'} E[X_i] = 0,$$

pa je:

$$\begin{aligned} Var\left(\sum_{i \in I'} X_i\right) &= E \left[ \left( \sum_{i \in I'} X_i \right)^2 \right] \\ &= E \left[ \sum_{i \in I'} X_i^2 + 2 \sum_{\{i,j\} \in I' \times I', i \neq j} X_i X_j \right] \\ &= \sum_{i \in I'} Var(X_i) + 2 \sum_{\{i,j\} \in I' \times I', i \neq j} Cov(X_i, X_j) \end{aligned}$$

U izrazu  $\sum_{\{i,j\} \in I' \times I', i \neq j} Cov(X_i, X_j)$  je poznato da je  $Cov(X_i, X_j) = 0.5$ , no preostaje izračunati koliki je broj dvočlanih kombinacija  $\{i, j\} \in I' \times I', i \neq j$ . Budući da je broj dvočlanih kombinacija 20-članog skupa jednak:

$$\binom{20}{2} = \frac{20!}{2!(20-2)!} = 190,$$

broj dvočlanih kombinacija  $\{i, j\} \in I' \times I'$  jednak je 190.

Nastavno na to, slijedi:

$$\begin{aligned} \text{Var}\left(\sum_{i \in I'} X_i\right) &= \sum_{i \in I'} 1 + 2 \sum_{\{i, j\} \in I' \times I', i \neq j} 0.5 \\ &= 20 \cdot 1 + 2 \cdot 190 \cdot 0.5 \\ &= 210. \end{aligned}$$

Konačno:

$$\text{Var}(Y) = 4 \cdot 210 + 225 = 1065.$$

Sada slijedi za svaki prediktor koji ulazi u model  $(X_i, i \in I')$ , izračunati  $\text{Cov}(X, Y)$  kako bi se moglo doći do očekivane vrijednosti  $\rho_{X, Y}$ . U nastavku je prikazan izračun za  $X_{11}$ , a postupak je analogan za svaki  $X_i, i \in I'$ .

$$\begin{aligned} \text{Cov}_{X_{11}, Y} &= E[(X_{11} - E[X_{11}])(Y - E[Y])] \\ &\stackrel{E[X_{11}]=0}{=} E[X_{11}(Y - E[Y])] \\ &= E[X_{11}Y - X_{11}E[Y]] \\ &= E[X_{11}Y] - E[X_{11}]E[Y] \\ &\stackrel{E[X_{11}]=0}{=} E[X_{11}Y] \\ &= E[X_{11}(2 \sum_{i \in I'} X_i + \varepsilon)] \\ &= E[2X_{11}^2 + 2 \sum_{i \in I' \setminus \{11\}} X_{11}X_i + X_{11}\varepsilon] \\ &= 2\text{Var}(X_{11}) + 2 \sum_{i \in I' \setminus \{11\}} \text{Cov}(X_{11}, X_i) + \text{Cov}(X_{11}, \varepsilon) \\ &= 2 \cdot 1 + 2 \cdot 19 \cdot 0.5 + 0 \\ &= 21 \end{aligned}$$

$$E[\rho_{X_{11}, Y}] = E\left[\frac{\text{Cov}(X_{11}, Y)}{\sqrt{\text{Var}(Y)}}\right] = E\left[\frac{21}{\sqrt{1065}}\right] = 0.6435.$$

S obzirom na tumačenja koja slijede iz tablice 4.5, značajkama  $X_i, i \in I'$  dodjeljuje se kategorija *Jak utjecaj*, a ostalim značajkama kategorija *Nema utjecaja*.

### 5.3.2. Procjena penalizacijskih koeficijenata algoritmom pretraživanja kukavice

Budući da je u istraživanju simulirano po 50 skupova podataka iz ovog scenarija te da su se parametri regularizacije optimizirali zasebno za svaki od njih i za svaku metodu regularizacije, distribucije normaliziranih penalizacijskih koeficijenata prikazane su grafički (slike 5.23 i 5.24) i deskriptivno (tablica 5.14) u nastavku. U tablici je vidljivo kako su za svaku metodu regularizacije vrijednosti prikazanih statistika procjena penalizacijskih koeficijenata nižeg iznosa za  $\gamma_4$  nego za  $\gamma_1$ .

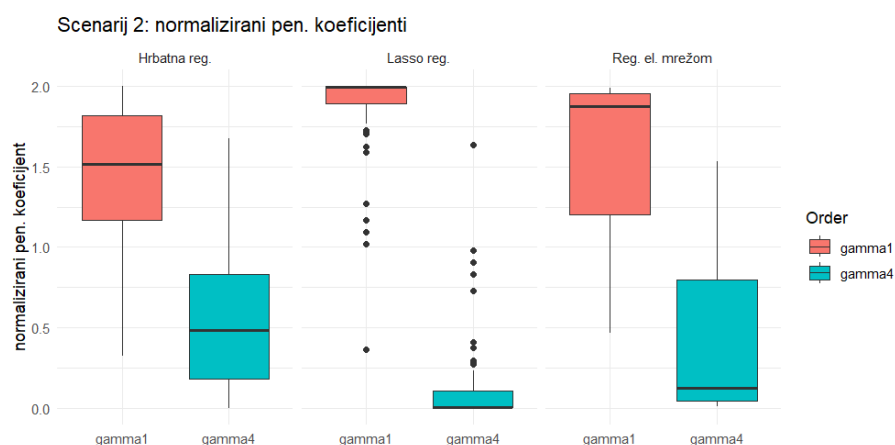
Tablica 5.14: Scenarij 2 - deskriptivna analiza procjene penalizacijskog koeficijenata ovisno o vrsti regularizacije (normalizirane vrijednosti)

Metoda	Pen. koef.	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
hrbatna reg.	$\gamma_1$	0.325	1.166	1.518	1.462	1.817	1.999	0.430
	$\gamma_4$	0.001	0.183	0.482	0.538	0.834	1.675	0.430
lasso reg.	$\gamma_1$	0.364	1.890	1.992	1.847	1.992	1.997	0.325
	$\gamma_4$	0.003	0.008	0.008	0.153	0.110	1.636	0.325
reg. el. mrežom	$\gamma_1$	0.465	1.204	1.873	1.568	1.954	1.987	0.466
	$\gamma_4$	0.013	0.046	0.127	0.432	0.796	1.535	0.466

$\gamma_1, \gamma_4$ : penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja* i *Jak utjecaj*, redom.

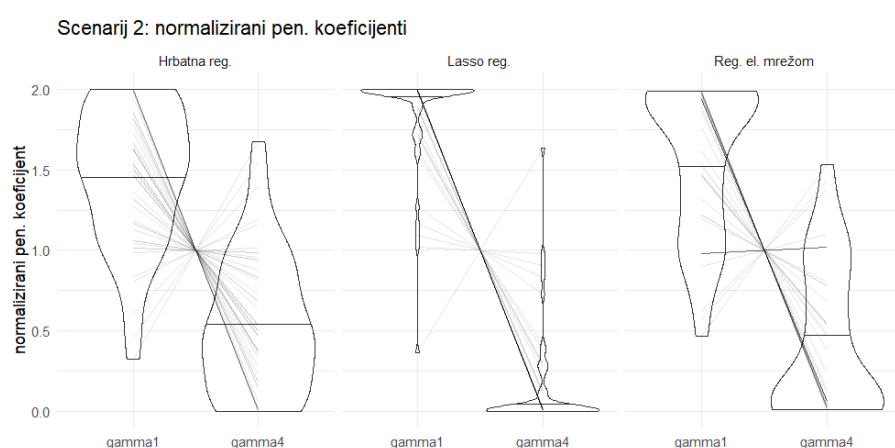
Nadalje, na slici 5.23 prikazani su kutijasti dijagrami penalizacijskih koeficijenata. Vidljiv je pomak u distribucijama procjena penalizacijskih koeficijenata  $\gamma_1$  i  $\gamma_4$  te da lasso regularizacija rezultira najmanjom varijancom. Nadalje, na slici 5.24 prikazane su distribucije procjena penalizacijskih koeficijenata violinskim dijagramom te, na njima, najčešći profili penalizacijskih koeficijenata po skupovima iz Scenarija 2. Premda je bilo profila gdje su značajkama snažnijeg utjecaja procijenjeni veći penalizacijski koeficijent nego značajkama slabijeg utjecaja, vidljivo je da su to rjeđe situacije. Češća situacija je bila da su procijenjeni manji penalizacijski koeficijenti značajkama snažnijeg utjecaja i obratno.





Slika 5.23: Kutijasti dijagram distribucija procjene penalizacijskih koeficijenta za skupove iz scenarija 2.

$\gamma_1, \gamma_4$ : normalizirani penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja* i *Jak utjecaj*, redom.



Slika 5.24: Violinski dijagram distribucija procjene penalizacijskih koeficijenta s profilima procjena za skupove iz scenarija 2.

$\gamma_1, \gamma_4$ : normalizirani penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja* i *Jak utjecaj*, redom

### 5.3.3. Procjene koeficijenta modela

U nastavku su prikazani rezultati procjene koeficijenta modela iz Scenarija 2. S obzirom na to da je u ovom scenariju  $p = 40$ , a od toga je 20 prediktora s koeficijentom različitim od nule, odnosno:

$$\beta = (\underbrace{0, 0, \dots, 0}_{10}, \underbrace{2, 2, \dots, 2}_{10}, \underbrace{0, 0, \dots, 0}_{10}, \underbrace{2, 2, \dots, 2}_{10}),$$

te su korelacije između svih prediktora jednake i iznose 0.5, radi preglednosti prikazani su rezultati za po jedan prediktor iz svakog bloka prediktora s koeficijentima različitim od nule. Prikazani su i rezultati procjene slobodnog koeficijenta. Za preostale su se koeficijente koji su jednaki nuli i distribucije procjena kretale oko nule, ali radi preglednosti njihovi su grafički prikazi i deskriptivna analiza izostavljeni iz rada.

Tablica 5.15: Scenarij 2 - deskriptivna analiza procjene koeficijenata modela ovisno o vrsti regularizacije i načinu procjene penalizacijskog koeficijenta

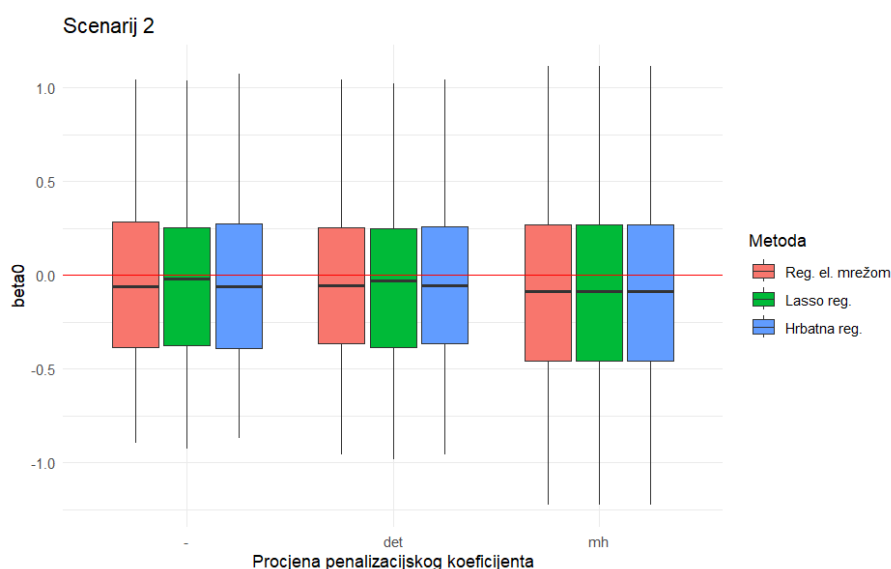
Koef.	Procjena pen. koef.	Metoda	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
$\beta_0 = 0$	-	lasso reg.	-0.90	-0.38	-0.06	<b>-0.03</b>	0.29	1.04	0.47
		hrbatna reg.	-0.93	-0.38	<b>-0.02</b>	-0.04	<b>0.25</b>	1.04	0.47
		reg. el. mrežom	<b>-0.87</b>	-0.39	-0.06	<b>-0.03</b>	0.27	1.07	0.47
	det	lasso reg.	-0.96	<b>-0.37</b>	-0.06	-0.04	<b>0.25</b>	1.04	<b>0.46</b>
		hrbatna reg.	-0.98	-0.39	-0.03	-0.04	<b>0.25</b>	<b>1.02</b>	0.47
		reg. el. mrežom	-0.96	<b>-0.37</b>	-0.06	-0.04	0.26	1.04	0.47
	mh	lasso reg.	-1.23	-0.46	-0.09	-0.06	0.27	1.11	0.55
		hrbatna reg.	-1.23	-0.46	-0.09	-0.06	0.27	1.11	0.55
		reg. el. mrežom	-1.23	-0.46	-0.09	-0.06	0.27	1.11	0.55
$\beta_{11} = 2$	-	lasso reg.	0.33	1.12	1.42	1.48	1.79	2.62	0.50
		hrbatna reg.	0.12	1.23	1.73	1.67	<b>2.10</b>	2.84	0.65
		reg. el. mrežom	0.76	1.04	1.14	1.15	1.26	1.75	0.18
	det	lasso reg.	0.52	1.56	1.74	1.72	1.89	2.39	0.30
		hrbatna reg.	0.18	1.40	1.85	1.81	2.25	2.95	0.65
		reg. el. mrežom	<b>1.22</b>	<b>1.60</b>	1.70	1.68	1.81	<b>2.03</b>	<b>0.17</b>
	mh	lasso reg.	0.63	1.51	2.03	<b>1.97</b>	2.27	3.24	0.62
		hrbatna reg.	0.55	1.54	2.09	2.04	2.48	3.24	0.66
		reg. el. mrežom	0.63	1.51	<b>2.01</b>	1.94	2.27	3.25	0.60

Koef.	Procjena pen. koef.	Metoda	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
$\beta_{31} = 2$	-	lasso reg.	0.16	1.11	1.34	1.43	1.75	2.66	0.56
		hrbatna reg.	0.00	1.11	1.73	1.63	2.16	3.28	0.78
		reg. el. mrežom	0.73	1.00	1.13	1.16	1.28	1.77	0.24
	det	lasso reg.	0.56	1.51	1.72	1.71	<b>1.95</b>	2.48	0.37
		hrbatna reg.	0.08	1.24	1.83	1.78	2.31	3.32	0.75
		reg. el. mrežom	<b>1.25</b>	<b>1.55</b>	1.68	1.68	1.81	<b>2.15</b>	<b>0.20</b>
	mh	lasso reg.	0.38	1.45	2.06	1.93	2.49	3.74	0.75
		hrbatna reg.	0.38	1.44	2.17	<b>2.00</b>	2.53	3.74	0.79
		reg. el. mrežom	0.40	1.45	<b>2.00</b>	1.91	2.48	3.63	0.72

Najtočnija procjena koeficijenta modela za svaku deskriptivnu statistiku označena je podebljano.

Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

Na slikama u nastavku grafički su prikazane distribucije procjena koeficijenta modela na skupovima iz drugog scenarija. Vidljivo je da se za sve metode regularizacije procijenjena vrijednost slobodnog koeficijenta  $\beta_0$  kretala oko nule (slika 5.25).

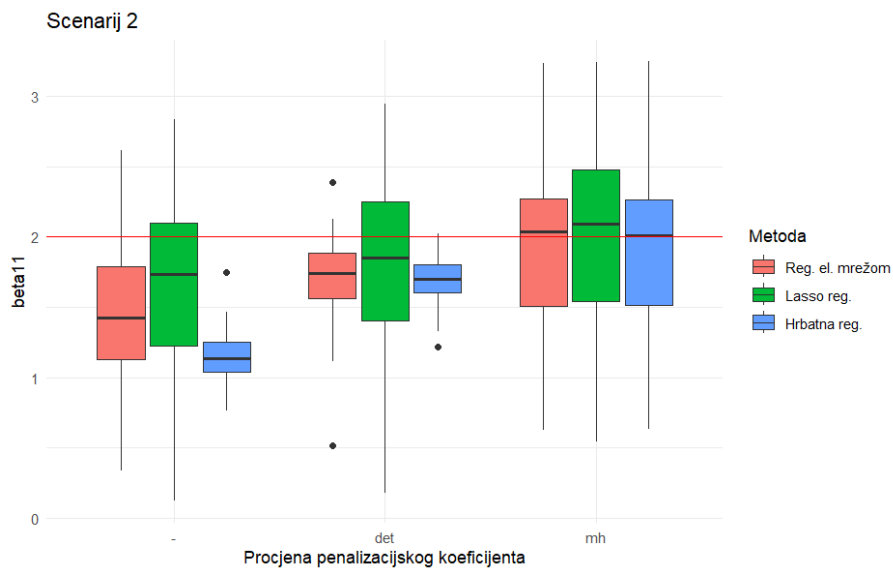


Slika 5.25: Grafički prikazi distribucija procjene koeficijenta  $\beta_0$  za skupove iz Scenarija 2.

Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

Crveni pravac predstavlja prave vrijednosti koeficijenta.

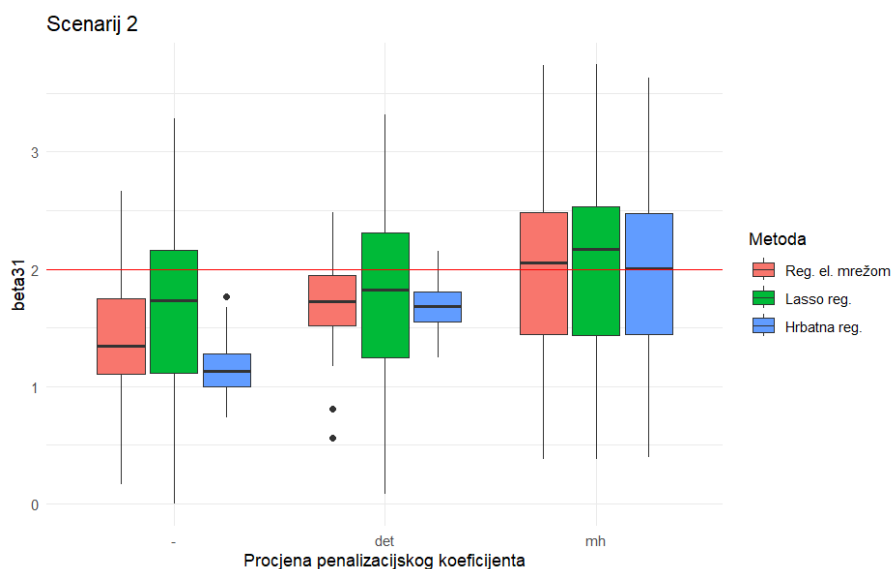
Nadalje, na slikama 5.26 i 5.27 vidljivo je kako je za sve metode regularizacije uvođenje domenskog znanja stručnjaka metaheuristikom rezultiralo procjenom vrijednosti koeficijenta  $\beta_{11}$  i  $\beta_{31}$  najbližoj stvarnoj vrijednosti (koja iznosi 2), te je potom uvođenje domenskog znanja stručnjaka deterministički bilo točnije od situacije kada se ne koristi domensko znanje.



Slika 5.26: Grafički prikazi distribucija procjene koeficijenta  $\beta_{11}$  za skupove iz Scenarija 2.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Crveni pravac predstavlja prave vrijednosti koeficijenata.



Slika 5.27: Grafički prikazi distribucija procjene koeficijenta  $\beta_{31}$  za skupove iz Scenarija 2.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Crveni pravac predstavlja prave vrijednosti koeficijenata.

Kao i u prethodnim slučajevima, i na skupovima simuliranim po Scenariju 2 metode regularizacije s metaheurističkim pristupom procjene penalizacijskog koeficijenta imale su dulje vrijeme izvođenja te je u tu svrhu korištena usluga Naprednog računanja (Padobran) Sveučilišnog računskog centra (Sree) Sveučilišta u Zagrebu. U tablici 5.16 u nastavku prikazano je prosječno vrijeme izvođenja za jedan skup prilikom kojeg se koristilo pet procesora.

Tablica 5.16: Prosječno vrijeme izvođenja metoda regularizacije s metaheurističkom procjenom penalizacijskog koeficijenta za skupove iz Scenarija 2.

Metoda	lasso reg.	hrbatna reg.	reg. el. mrežom
Vrijeme	57:31	47:53	54:59

Prilikom izvođenja za jedan skup korišteno je pet procesora, a prikazano vrijeme odnosi se na minute i sekunde.

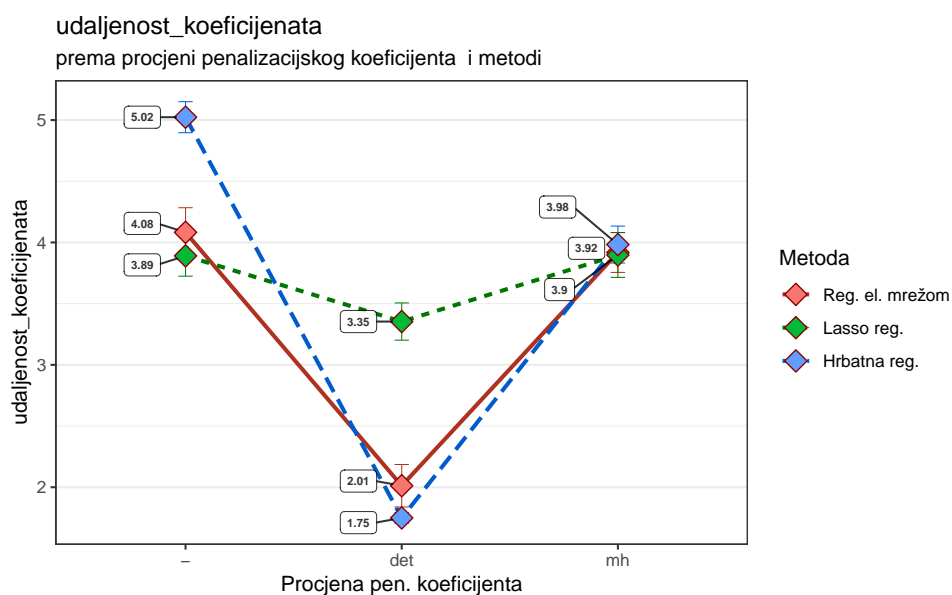
#### 5.3.4. Evaluacije modela i hipoteza

U potpoglavlju 4.5 (str. 50), objašnjeno kako su metodom ANOVA analizirani rezultati modela svih 50 skupova iz Scenarija 2 s ciljem donošenja odluke o hipotezama. Provjera pretpostavki za analizu varijance izvršena je analizom dijagrama raspršenosti reziduala modela u odnosu na razine faktora te qq-dijagramom distribucije reziduala u usporedbi sa standardnom normalnom distribucijom za svaku mjeru evaluacije modela unutar pojedine hipoteze. Spomenuti dijagrami su, radi preglednosti, prikazani u Prilogu 3 (slike 7.10, 7.11 i 7.12), a s obzirom na to da nisu upućivali na značajna odstupanja od ovih pretpostavki, smatra se valjanim koristiti analizu varijance. Odluke se donose na razini statističke značajnosti 0.05.

#### Hipoteza 1

Kao u Scenariju 1, u ovoj hipotezi uspoređuju se modeli regularizacije koji ne uključuju znanje stručnjaka s modelima regularizacije koji ga uključuju deterministički i metaheuristički kroz pet različitih mjera evaluacije. Za evaluaciju hipoteze koristi se dvofaktorska ANOVA s faktorima Procjena penalizacijskog koeficijenta i Metoda regularizacije te njihovom interakcijom. Kontrasti su pojašnjeni u potpoglavlju 4.5 (str. 50), pri čemu se pomoću kontrasta c11 testira upravo Hipoteza 1 ovog istraživanja, dok kontrast c12 uspoređuje međusobno dva načina uvođenja domenskog znanja stručnjaka u metode regularizacije.

U nastavku su za svaku mjeru evaluacije prikazani interakcijski grafikoni metode regularizacije i različitih načina procjene penalizacijskog koeficijenta.



Slika 5.28: Scenarij 2, Hipoteza 1 - Udaljenost koeficijenata prema procjeni penalizacijskog koeficijenta i metodi.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Na slici 5.28 prikazane su prosječne vrijednosti udaljenosti koeficijenata prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako za sve metode regularizacije uvođenjem domenskog znanja na deterministički način opada prosječna vrijednost udaljenosti koeficijenata, a onda ona raste kada se domensko znanje uvede metaheuristikom (za hrbatnu regularizaciju i regularizaciju elastičnom mrežom raste na razinu ispod razine slučaja kada se ne koristi domensko znanje stručnjaka, dok za lasso regularizaciju raste na višu razinu). Rezultati se razlikuju i između metode, lasso regularizacija ima stabilnije prosječne vrijednosti od hrbatne regularizacije i regularizacije elastičnom mrežom.

Za mjeru evaluacije udaljenost koeficijenata statistički su značajni interakcijski član ( $F(4, 441) = 81.89$ ,  $p = 2 \text{ e-}16$ ), kao i efekti faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 527.24$ ,  $p < 2 \text{ e-}16$ ) i faktora Metoda regularizacije ( $F(2, 441) = 17.96$ ,  $p = 3.16 \text{ e-}8$ ). Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.7 u Prilogu 3 (str. 183).

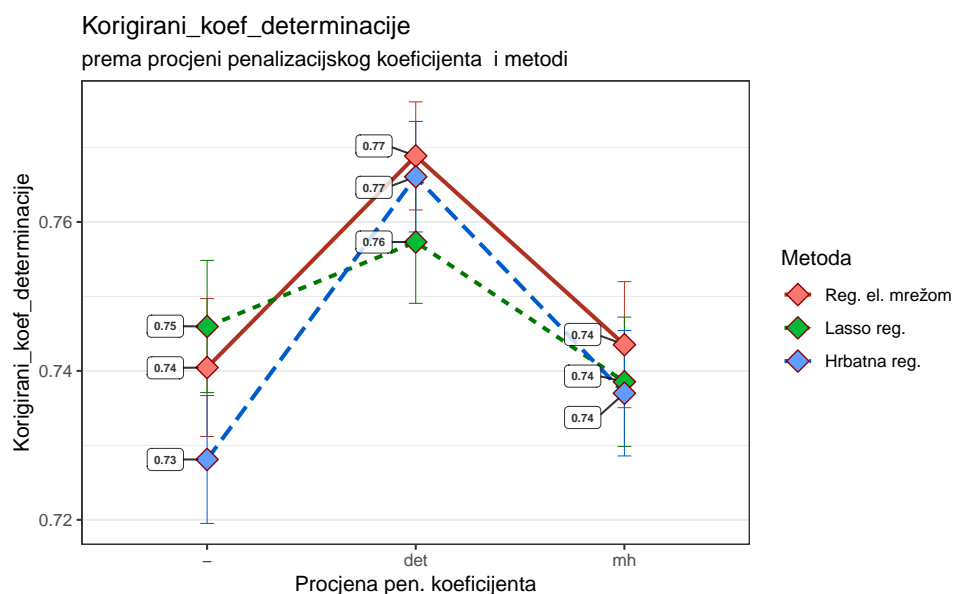
Budući da je interakcijski član dvaju faktora statistički značajan, efekt pojedinog načina uvođenja domenskog znanja stručnjaka na udaljenost koeficijenata ne može biti generaliziran

na sve metode regularizacije. U interpretaciji rezultata potrebno je koristiti *post hoc* analizu interakcije. U tablici 7.8 u Prilogu 3 (str. (str. 184)) prikazan je rezultat *post hoc* analize Tukeyjevim testom HSD te je vidljivo koje su razlike prosječnih vrijednosti udaljenosti koeficijenata statistički značajno različite od nule na razini statističke značajnosti 0.05. Vidljivo je upravo da je za svaku metodu regularizacije statistički značajna razlika prosječnih vrijednosti udaljenosti koeficijenata između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka. Negativan iznos razlike upućuje da je prosječna udaljenost koeficijenata manja kod determinističkog pristupa. Niti za jednu metodu regularizacije nije statistički značajna razlika prosječnih udaljenosti koeficijenata između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka. Konačno, kod svake metode regularizacije statistički je značajna razlika prosječnih udaljenosti koeficijenata između metaheurističkog pristupa i determinističkog pristupa uvođenja domenskog znanja. Pozitivan iznos razlike upućuje da je prosječna udaljenost koeficijenata veća kod metaheuristike.

Na slici 5.29 prikazane su prosječne vrijednosti korigiranog koeficijenta determinacije prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako za sve metode regularizacije uvođenjem domenskog znanja na deterministički način raste prosječna vrijednost korigiranog koeficijenta determinacije, a onda ona pada kada se domensko znanje uvede metaheuristikom. Za regularizaciju elastičnom mrežom i hrbatnu regularizaciju dođe na razinu i dalje višu nego prije uvođenja domenskog znanja stručnjaka, dok za lasso regularizaciju padne na nižu razinu od početne.

Za mjeru evaluacije korigirani koeficijent determinacije statistički su značajni interakcijski član ( $F(4, 441) = 2.663$ ,  $p = 0.0321$ ) i efekt faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 36.503$ ,  $p = 2.14 \cdot 10^{-15}$ ). Efekt faktora Metoda regularizacije ( $F(2, 441) = 2.250$ ,  $p = 0.1067$ ) nije statistički značajan. Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.7 u Prilogu 3 (str. 183).

Budući da je interakcijski član dvaju faktora statistički značajan, efekt pojedinog načina uvođenja domenskog znanja stručnjaka na korigirani koeficijent determinacije ne može biti generaliziran na sve metode regularizacije. U interpretaciji rezultata potrebno je koristiti *post hoc* analizu interakcije. U tablici 7.9 u Prilogu 3 ((str. 186)) prikazan je rezultat *post hoc* analize Tukeyjevim testom HSD te vidljivo je koje su razlike statistički značajno različite od nule na razini statističke značajnosti 0.05.



Slika 5.29: Scenarij 2, Hipoteza 1 - Korigirani koeficijent determinacije prema procjeni penalizacijskog koeficijenta i metodi.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Vidljivo je kako je za lasso regularizaciju razlika u prosječnom korigiranom koeficijentu determinacije između metaheurističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka negativna, no nije statistički značajna, dok je razlika između determinističkog pristupa i pristupa kada se ne koristi znanje stručnjaka pozitivna, ali također nije statistički značajna. Razlika u prosječnom korigiranom koeficijentu determinacije između metaheurističkog pristupa i determinističkog pristupa uvođenja domenskog znanja stručnjaka negativna je i statistički značajna, što znači da je prosječni korigirani koeficijent za lasso regularizaciju bio statistički značajno veći kod determinističkog uvođenja znanja stručnjaka nego kod metaheurističkog.

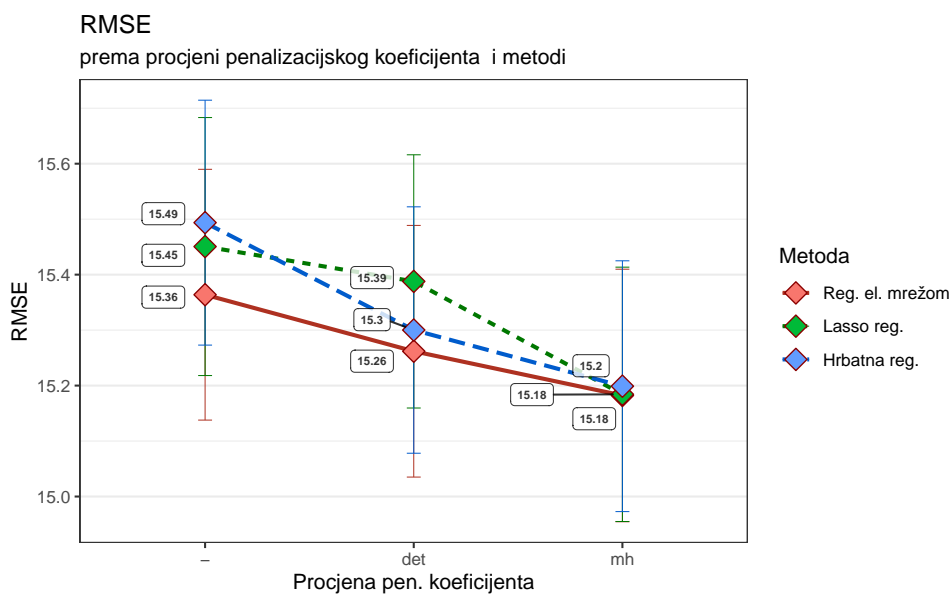
Nadalje, za hrbatnu regularizaciju i regularizaciju elastičnom mrežom razlika je u prosječnom korigiranom koeficijentu determinacije između između metaheurističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka pozitivna, ali nije statistički značajna. S druge strane, razlika u prosječnom korigiranom koeficijentu determinacije između determinističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka pozitivna je i statistički značajna, dok je razlika u prosječnom korigiranom koeficijentu determinacije između metaheurističkog i determinističkog pristupa negativna i statistički značajna.



Dakle, za hrbatnu regularizaciju i regularizaciju elastičnom mrežom uvođenje je domenskog znanja stručnjaka determinističkim pristupom rezultiralo statistički značajnim povećanjem prosječnog korigiranog koeficijenta determinacije, dok za lasso regularizaciju to nije slučaj. Uvođenje domenskog znanja stručnjaka metaheuristikom nije doprinijelo statistički značajnoj razlici prosječnih korigiranih koeficijenta determinacije u odnosu na modele bez znanja stručnjaka niti za jednu regularizaciju.

Na slici 5.30 prikazane prosječne vrijednosti *RMSE* prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako uvođenjem domenskog znanja stručnjaka na deterministički način opada prosječna vrijednost greške, te ona dodatno opada uvođenjem domenskog znanja metaheuristikom. Također, za regularizaciju elastičnom mrežom u svim je slučajevima najniža prosječna *RMSE*.

Za mjeru evaluacije *RMSE* nisu statistički značajni interakcijski član ( $F(4, 441) = 0.162$ ,  $p = 0.9575$ ) i efekt faktora Metoda regularizacije ( $F(2, 441) = 0.355$ ,  $p = 0.7016$ ), dok efekt faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 3.617$ ,  $p = 0.0277$ ) statistički je značajan.

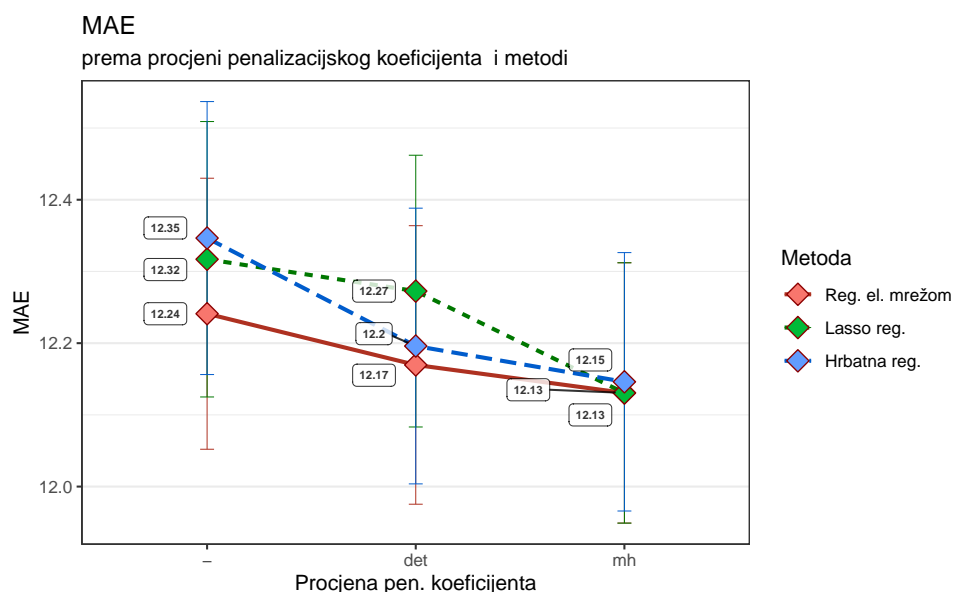


Slika 5.30: Scenarij 2, Hipoteza 1 - *RMSE* prema procjeni penalizacijskog koeficijenta i metodi. Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.7 u Prilogu 3 (str. 183). U spomenutoj tablici vidljivo je da je statistički značajan jedino koeficijent uz kontrast c11

koji govori o razlici prosječnih *RMSE* između metoda regularizacije koje ne koriste domensko znanje u odnosu na one koje ga uključuju deterministički ili metaheuristikom. Pozitivan iznos koeficijenta upućuje da je prosječna *RMSE* niža u metodama regularizacije koje koriste domensko znanje stručnjaka. Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *RMSE* između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna ( $\text{diff} = -0.119$ ,  $p = 0.3975$ ). S druge strane, razlika prosječnih *RMSE* greški između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna je i statistički značajna ( $\text{diff} = -0.248$ ,  $p = 0.0203$ ).

Na slici 5.31 prikazane prosječne vrijednosti *MAE* prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je uvođenjem domenskog znanja stručnjaka na deterministički način opada prosječna vrijednost greške, te ona dodatno opada uvođenjem domenskog znanja metaheuristikom. Također, za regularizaciju elastičnom mrežom u svim je slučajevima najniža prosječna *MAE*.



Slika 5.31: Scenarij 2, Hipoteza 1 - *MAE* prema procjeni penalizacijskog koeficijenta i metodi. Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

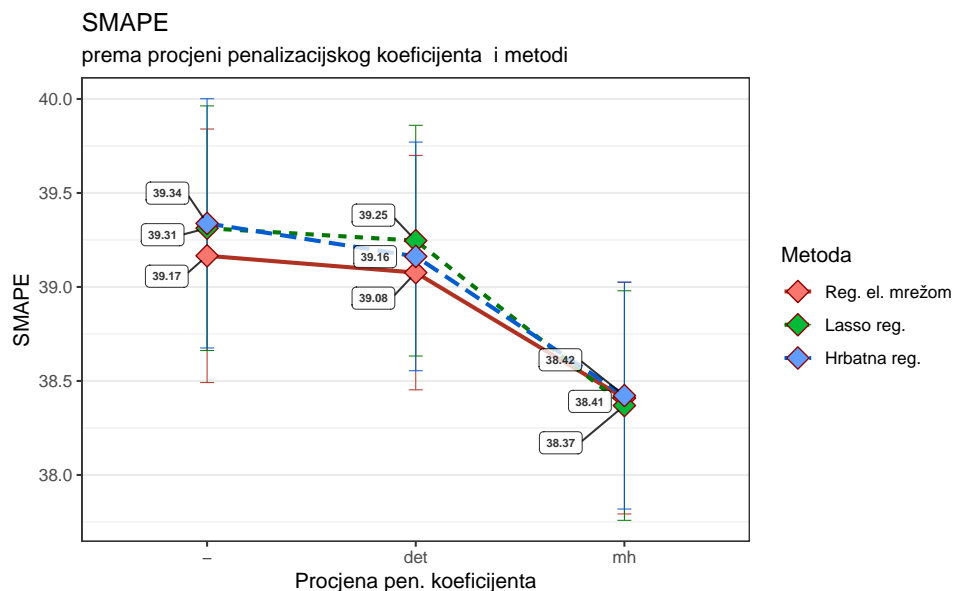
No, za mjeru evaluacije *MAE* nisu statistički značajni interakcijski član ( $F(4, 441) = 0.164$ ,  $p = 0.9566$ ), efekti faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 2.361$ ,  $p = 0.0955$ ) i faktora Metoda regularizacije ( $F(2, 441) = 0.348$ ,  $p = 0.7066$ ).

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.7 u Prilogu 3 (str. 183).

U spomenutoj tablici vidljivo je kako koeficijenti uz kontraste ne odstupaju statistički značajno od 0.

Na slici 5.32 prikazane su prosječne vrijednosti *SMAPE* prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je uvođenjem domenskog znanja stručnjaka na deterministički način opada prosječna vrijednost greške, te ona dodatno opada uvođenjem domenskog znanja metaheuristikom. Također, za regularizaciju elastičnom mrežom u svim je slučajevima najniža prosječna *SMAPE*.

Za mjeru evaluacije *SMAPE* nisu statistički značajni interakcijski član ( $F(4, 441) = 0.042$ ,  $p = 0.99664$ ) i efekt faktora Metoda regularizacije ( $F(2, 441) = 0.085$ ,  $p = 0.91827$ ), dok efekt faktora Procjena penalizacijskog koeficijenta je statistički značajan ( $F(2, 441) = 6.886$ ,  $p = 0.00114$ ).



Slika 5.32: Scenarij 2, Hipoteza 1 - *SMAPE* prema procjeni penalizacijskog koeficijenta i metodi.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.7 u Prilogu 3 (str. 183). U spomenutoj tablici vidljivo je da su statistički značajni jedino koeficijenti uz kontraste c11 i c12.

Kontrast c11 govori o razlici prosječnih *SMAPE* greški između metoda regularizacije koje ne koriste domensko znanje u odnosu na one koje ga uključuju deterministički ili metaheuris-

tikom. Pozitivan iznos koeficijenta upućuje da je prosječna *SMAPE* niža u metodama regularizacije koje koriste domensko znanje stručnjaka. Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *SMAPE* greški između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna (diff = -0.11,  $p = 0.9025$ ). Nadalje, razlika prosječnih *SMAPE* greški između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna je i statistički značajna (diff = -0.872,  $p = 0.0021$ ).

Kontrast c12 govori o razlici prosječnih *SMAPE* greški metoda regularizacije koje penalizacijske koeficijente procjenjuju deterministički s onima koje ih procjenjuju metaheuristikom. Negativan iznos koeficijenta upućuje da je prosječna *SMAPE* niža u slučaju uvođenja domenskog znanja stručnjaka metaheuristikom, u odnosu na determinističko uvođenje.

## Hipoteza 2

U ovoj se hipotezi kroz tri različite mjere evaluacije uspoređuju metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka. Hipoteza se dijeli na dvije podhipoteze - u jednoj se uspoređuje metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka deterministički (Hipoteza 2a), a u drugoj se metoda CART uspoređuju s metodama regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom (Hipoteza 2b).

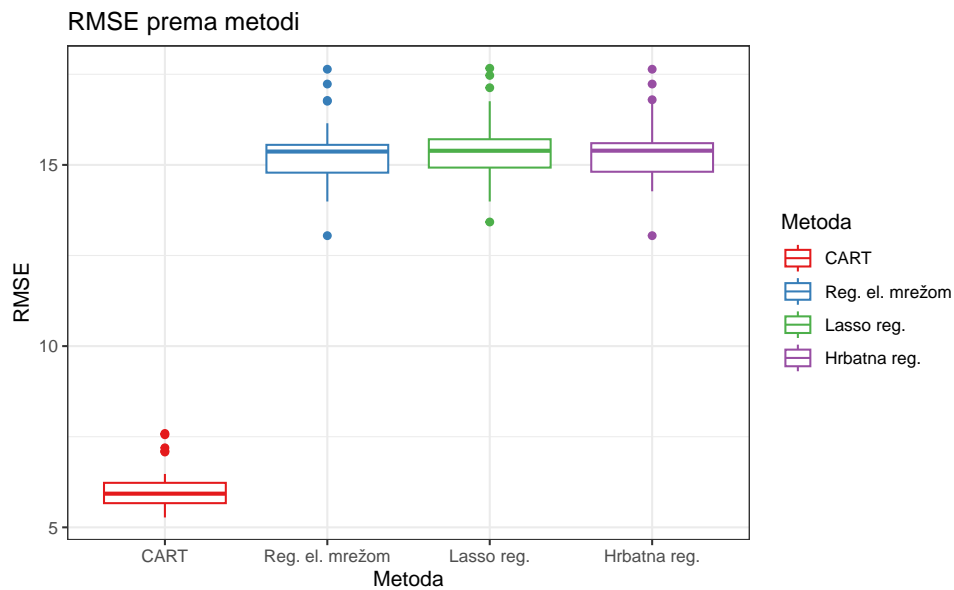
### Hipoteza 2a

U ovoj hipotezi uspoređuje se metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka deterministički. Podsjetimo, za evaluaciju hipoteze koristi se jednofaktorska ANOVA s faktorom Metoda. Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2 ovog istraživanja. Odluke se donose na razini statističke značajnosti 0.05. U nastavku su grafički prikazane distribucije svake mjere evaluacije.

Za mjeru evaluacije *RMSE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1994$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.33 vidljivo je kako CART metoda ima nižu prosječnu *RMSE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.10 u Prilogu 3 (str. 189). U spomenutoj tablici vidljivo je da je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *RMSE* od metoda regularizacije koje uvode domensko znanje stručnjaka determi-

nistički.



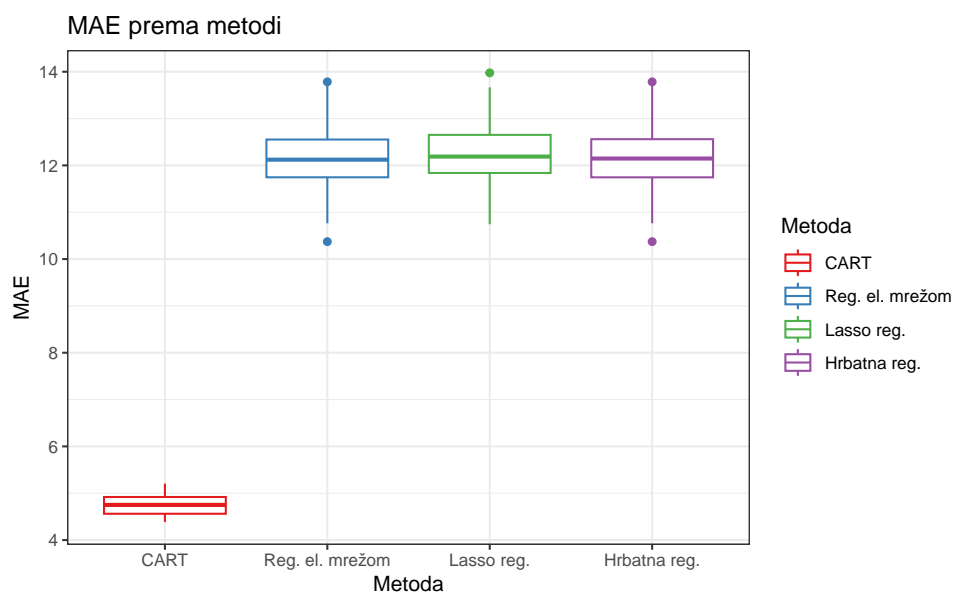
Slika 5.33: Scenarij 2, Hipoteza 2a - *RMSE* prema metodi.

U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih *RMSE* greški metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna (diff = - 9.239,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 9.365,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 9.278,  $p = 0.00$ ).

Za mjeru evaluacije *MAE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1952$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.34 vidljivo je kako CART metoda ima nižu prosječnu *MAE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.10 u Prilogu 3 (str. 189). U spomenutoj tablici vidljivo je da je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *MAE* od metoda regularizacije koje uvode domensko znanje stručnjaka deterministički.



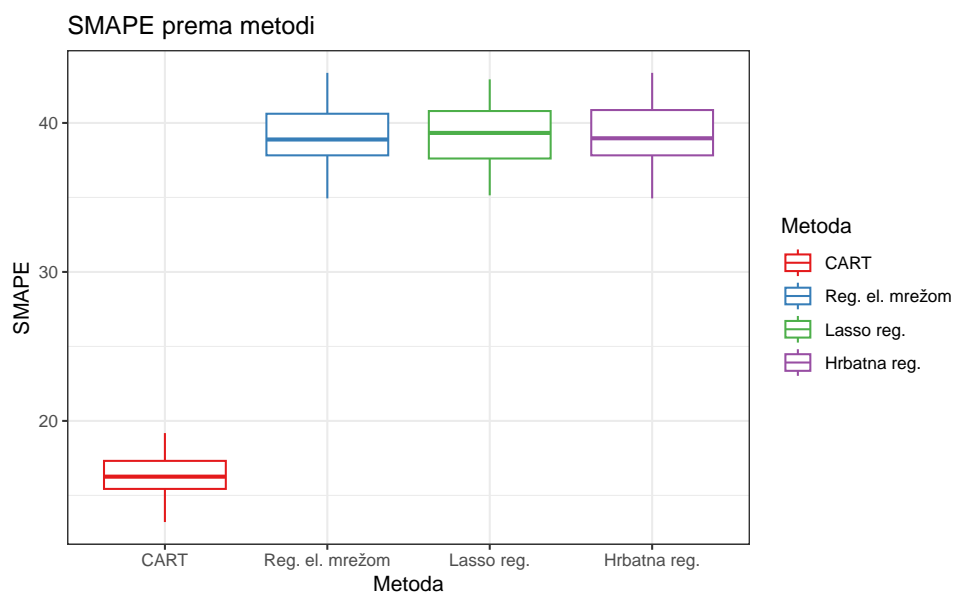
Slika 5.34: Scenarij 2, Hipoteza 2a - MAE prema metodi.

U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih MAE metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna (diff = - 7.411,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 7.514,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 7.437,  $p = 0.00$ ).

Za mjeru evaluacije *SMAPE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1648$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.35 vidljivo je kako CART metoda ima nižu prosječnu *SMAPE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.10 u Prilogu 3 (str. 189). U spomenutoj tablici vidljivo je da je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *SMAPE* od metoda regularizacije koje uvode domensko znanje stručnjaka deterministički.



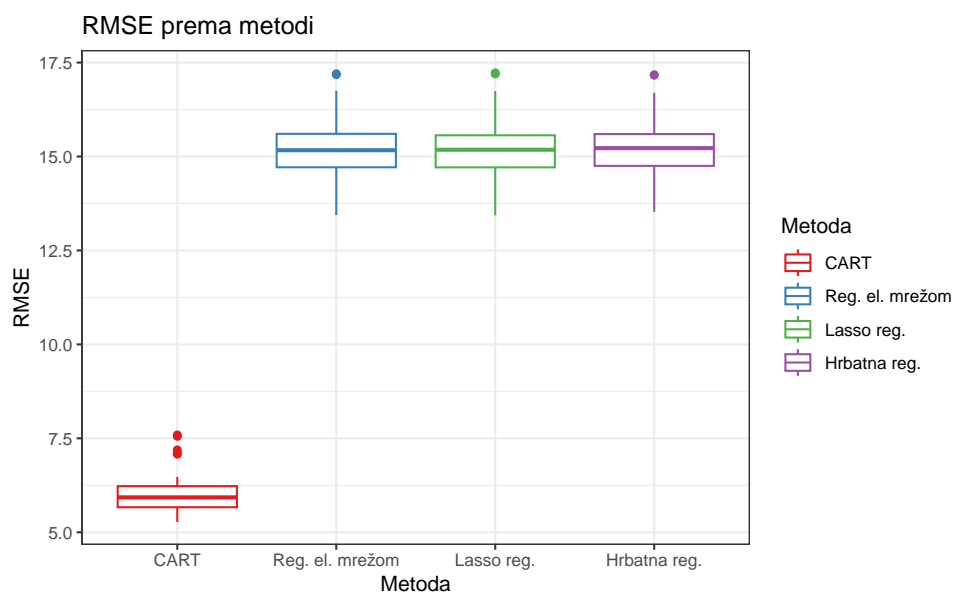
Slika 5.35: Scenarij 2, Hipoteza 2a - *SMAPE* prema metodi.

U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih *SMAPE* greški metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna ( $\text{diff} = -22.784$ ,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka ( $\text{diff} = -22.954$ ,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka ( $\text{diff} = -22.87$ ,  $p = 0.00$ ).

### Hipoteza 2b

U ovoj hipotezi uspoređuje se metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom. Podsjetimo, za evaluaciju hipoteze koristi se jednofaktorska ANOVA s faktorom Metoda. Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2 ovog istraživanja. Odluke se donose na razini statističke značajnosti 0.05. U nastavku su grafički prikazane distribucije svake mjere evaluacije.



Slika 5.36: Scenarij 2, Hipoteza 2b - *RMSE* prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

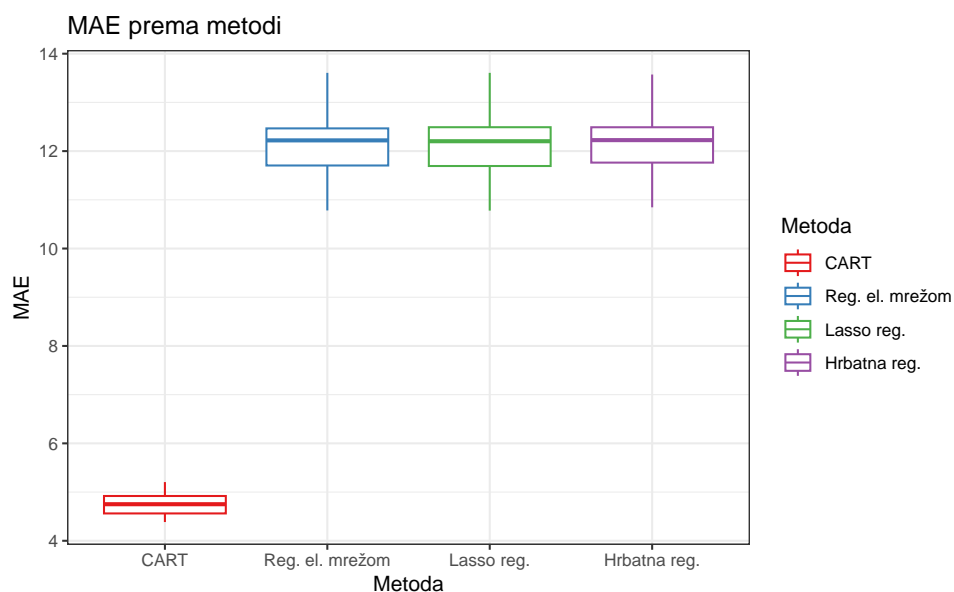
Za mjeru evaluacije *RMSE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1911$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.36 vidljivo je kako CART metoda ima nižu prosječnu *RMSE* od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.11 u Prilogu 3 (str. 191). U spomenutoj tablici vidljivo je da je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *RMSE* od metoda regularizacije koje uvode domensko znanje stručnjaka metaheuristikom.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih *RMSE* metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 9.16  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 9.162,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 9.176,  $p = 0.00$ ).

Za mjeru evaluacije *MAE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2139$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.37 vidljivo je kako CART metoda ima nižu prosječnu *MAE* grešku od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.





Slika 5.37: Scenarij 2, Hipoteza 2b - MAE prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

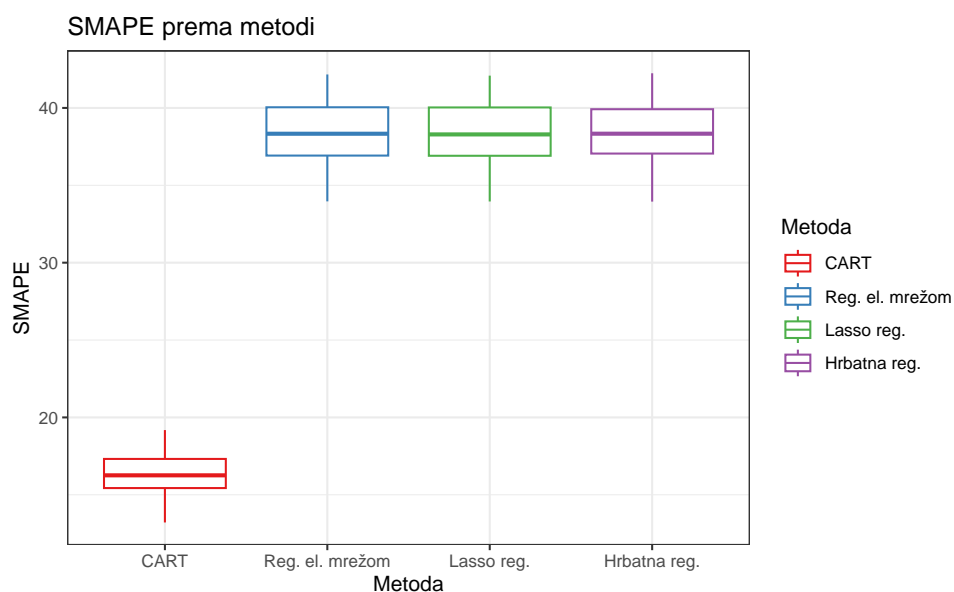
Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.11 u Prilogu 3 (str. 191). U spomenutoj tablici vidljivo je da je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom MAE od metoda regularizacije koje uvode domensko znanje stručnjaka metaheuristikom.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih MAE metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 7.372,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 7.372,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 7.387,  $p = 0.00$ ).

Za mjeru evaluacije SMAPE statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1563$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.38 vidljivo je kako CART metoda ima nižu prosječnu SMAPE od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.11 u Prilogu 3 (str. 191). U spomenutoj tablici vidljivo je da je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom SMAPE od metoda regularizacije koje uvode domensko znanje stručnjaka

metaheuristikom.



Slika 5.38: Scenarij 2, Hipoteza 2b - *SMAPE* prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih *SMAPE* greški metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 22.116,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 22.077,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 22.13,  $p = 0.00$ ).

## 5.4. SCENARIJ 3

### 5.4.1. Modeliranje mišljenja stručnjaka

U ovom scenariju poznato je da vrijedi:

$$\begin{aligned}
 p &= 40, \\
 \beta &= (\underbrace{3, 3, \dots, 3}_{15}, \underbrace{0, 0, \dots, 0}_{25}), \\
 \sigma &= 15,
 \end{aligned}$$

$$\varepsilon \sim \mathcal{N}(0, 1),$$

$$X_i = Z_1 + \varepsilon_i^x, i = 1, \dots, 5,$$

$$X_i = Z_2 + \varepsilon_i^x, i = 6, \dots, 10,$$

$$X_i = Z_3 + \varepsilon_i^x, i = 11, \dots, 15,$$

$$Z_1, Z_2, Z_3 \sim \mathcal{N}(0, 1),$$

$$\varepsilon_i^x \sim \mathcal{N}(0, 0.01), i = 1, \dots, 15,$$

( $\varepsilon_i^x$  su nezavisne i jednako distribuirane [n.j.d.]

$$X_i \sim \mathcal{N}(0, 1), \text{ n.j.d., } i = 16, \dots, 40.$$

Odnosno,

$$\begin{aligned} Y &= 3X_1 + 3X_2 + \dots + 3X_{15} + 15\varepsilon \\ &= 3 \sum_{i \in \{1, \dots, 15\}} X_i + 15\varepsilon \end{aligned}$$

Radi jednostavnosti, uvodi se oznaka  $I'' = \{1, \dots, 15\}$ , pa se prethodno može zapisati na sljedeći način:

$$Y = 3 \sum_{i \in I''} X_i + 15\varepsilon.$$

Nadalje, vrijedi:

$$\text{Var}(Y) = \text{Var}\left(3 \sum_{i \in I''} X_i + 15\varepsilon\right),$$

Budući da je po pretpostavkama regresijskog modela greška modela nezavisna od prediktora  $X_j$ ,  $\text{Var}(Y)$  može se raščlaniti na sljedeći način:

$$\text{Var}(Y) = \text{Var}\left(3 \sum_{i \in I''} X_i\right) + \text{Var}(15\varepsilon),$$

te dalje zbog svojstava varijance i distribucije greške  $\varepsilon$ , vrijedi:

$$\begin{aligned} \text{Var}(Y) &= 3^2 \text{Var}\left(\sum_{i \in I''} X_i\right) + 15^2 \text{Var}(\varepsilon) \\ &= 9 \text{Var}\left(\sum_{i \in I''} X_i\right) + 225 \cdot 1 \\ &= 9 \text{Var}\left(\sum_{i \in I''} X_i\right) + 225. \end{aligned}$$

Prema 4.5, vrijedi:

$$\text{Var}\left(\sum_{i \in I''} X_i\right) = E \left[ \left( \sum_{i \in I''} X_i - E\left[\sum_{i \in I''} X_i\right] \right)^2 \right].$$

Nadalje, zbog svojstva da prediktori imaju očekivanje 0, vrijedi:

$$E\left[\sum_{i \in I''} X_i\right] = \sum_{i \in I''} E[X_i] = 0,$$

pa je:

$$\begin{aligned} \text{Var}\left(\sum_{i \in I''} X_i\right) &= E \left[ \left( \sum_{i \in I''} X_i \right)^2 \right] \\ &= E \left[ \sum_{i \in I''} X_i^2 + 2 \sum_{\{i,j\} \in I' \times I', i \neq j} X_i X_j \right] \\ &= \sum_{i \in I''} \text{Var}(X_i) + 2 \sum_{\{i,j\} \in I' \times I', i \neq j} \text{Cov}(X_i, X_j) \end{aligned}$$

Potrebno je još izračunati  $\text{Var}(X_i)$  i  $\text{Cov}(X_i, X_j)$ .

Izračun  $\text{Var}(X_i)$  je u nastavku prikazan u slučaju kad je  $i = 1$ , a za ostale  $i \in I''$  je analogno.

$$\text{Var}(X_1) = \text{Var}(Z_1 + \varepsilon_1^x)$$

Uz pretpostavku da je greška  $\varepsilon_1^x$  nezavisna od prediktora  $Z_1$ , slijedi:

$$\begin{aligned} \text{Var}(X_1) &= \text{Var}(Z_1) + \text{Var}(\varepsilon_1^x) \\ &= 1 + 0.01 \\ &= 1.01. \end{aligned}$$

U izrazu  $\sum_{\{i,j\} \in I'' \times I'', i \neq j} \text{Cov}(X_i, X_j)$  preostaje izračunati  $\text{Cov}(X_i, X_j)$ , što ovisi o tome jesu li  $X_i$  i  $X_j$  iz iste grupe ili ne.

**1. slučaj:**  $X_i$  i  $X_j$  iz iste grupe, npr.  $X_1$  i  $X_2$ .

$$\begin{aligned} \text{Cov}(X_1, X_2) &= E[(X_1 - E[X_1])(X_2 - E[X_2])] \\ &\stackrel{E[X_1]=E[X_2]=0}{=} E[X_1 X_2] \\ &= E[(Z_1 + \varepsilon_1^x)(Z_1 + \varepsilon_2^x)] \\ &= E[(Z_1^2 + Z_1 \varepsilon_1^x + Z_1 \varepsilon_2^x + \varepsilon_1^x \varepsilon_2^x)] \\ &= E[Z_1^2] + E[Z_1 \varepsilon_1^x] + E[Z_1 \varepsilon_2^x] + E[\varepsilon_1^x \varepsilon_2^x] \end{aligned}$$

Budući da su po pretpostavkama regresijskog modela greške  $\varepsilon_1^x$  i  $\varepsilon_2^x$  nezavisne od prediktora  $Z_1$  te da su  $\varepsilon_1^x$  i  $\varepsilon_2^x$  nezavisne i jednako distribuirane, prethodno se može zapisati:

$$\begin{aligned} \text{Cov}(X_1, X_2) &= \text{Var}(Z_1) + E[Z_1]E[\varepsilon_1^x] + E[Z_1]E[\varepsilon_2^x] + E[\varepsilon_1^x]E[\varepsilon_2^x] \\ &= 1 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 \\ &= 1. \end{aligned}$$

Dakle, kovarijance prediktora iz iste grupe jednake su 1.

**2. slučaj:**  $X_i$  i  $X_j$  nisu iz iste grupe, npr.  $X_1$  i  $X_6$ .

$$\begin{aligned} \text{Cov}(X_1, X_6) &= E[(X_1 - E[X_1])(X_6 - E[X_6])] \\ &\stackrel{E[X_1]=E[X_6]=0}{=} E[X_1 X_6] \\ &= E[(Z_1 + \varepsilon_1^x)(Z_2 + \varepsilon_6^x)] \\ &= E[(Z_1 Z_2 + Z_2 \varepsilon_1^x + Z_1 \varepsilon_6^x + \varepsilon_1^x \varepsilon_6^x)] \\ &= E[Z_1 Z_2] + E[Z_2 \varepsilon_1^x] + E[Z_1 \varepsilon_6^x] + E[\varepsilon_1^x \varepsilon_6^x] \end{aligned}$$

Budući da su  $\varepsilon_1^x$  i  $\varepsilon_6^x$  nezavisne i jednako distribuirane, prethodno se može zapisati:

$$\begin{aligned} \text{Cov}(X_1, X_6) &= \text{Cov}(Z_1, Z_2) + \text{Cov}(Z_2, \varepsilon_1^x) + \text{Cov}(Z_1, \varepsilon_6^x) + E[\varepsilon_1^x]E[\varepsilon_6^x] \\ &= \text{Cov}(Z_1, Z_2) + \text{Cov}(Z_2, \varepsilon_1^x) + \text{Cov}(Z_1, \varepsilon_6^x) \end{aligned}$$

Nadalje, uz pretpostavku nezavisnosti  $Z_1, Z_2$  slijedi:

$$\text{Cov}(X_1, X_6) = 0.$$

Dakle, kovarijance prediktora koji nisu iz iste grupe jednake su 0.

Prema tome, uz oznake  $I_1'' = \{1, 2, \dots, 5\}$ ,  $I_2'' = \{6, 7, \dots, 10\}$ ,  $I_3'' = \{11, 12, \dots, 15\}$ , slijedi:

$$\begin{aligned} \sum_{\{i,j\} \in I'' \times I'', i \neq j} \text{Cov}(X_i, X_j) &= \sum_{\{i,j\} \in I_1'' \times I_1'', i \neq j} \text{Cov}(X_i, X_j) \\ &+ \sum_{\{i,j\} \in I_2'' \times I_2'', i \neq j} \text{Cov}(X_i, X_j) \\ &+ \sum_{\{i,j\} \in I_3'' \times I_3'', i \neq j} \text{Cov}(X_i, X_j) \end{aligned}$$

Budući da je u svakoj grupi od 5 prediktora upravo  $\binom{5}{2} = 10$  dvočlanih kombinacija  $\{X_i, X_j\}, i \neq j$ , slijedi:

$$\sum_{\{i,j\} \in I'' \times I'', i \neq j} Cov(X_i, X_j) = 3 \cdot 10 \cdot 1 = 30.$$

Konačno:

$$Var\left(\sum_{i \in I''} X_i\right) = \sum_{i \in I''} 1.01 + 2 \cdot 30 = 15 \cdot 1.01 + 60 = 75.15,$$

pa je

$$Var(Y) = 9 \cdot 75.15 + 225 = 901.35.$$

Sada slijedi za svaki prediktor koji ulazi u model  $(X_i, i \in I'')$ , izračunati  $Cov(X, Y)$  kako bi se moglo doći do očekivane vrijednosti  $\rho_{X,Y}$ . U nastavku je prikazan izračun za  $X_1$ , a postupak je analogan za svaki  $X_i, i \in I''$ .

$$\begin{aligned} Cov_{X_1, Y} &= E[(X_1 - E[X_1])(Y - E[Y])] \\ &\stackrel{E[X_1]=0}{=} E[X_1(Y - E[Y])] \\ &= E[X_1Y - X_1E[Y]] \\ &= E[X_1Y] - E[X_1]E[Y] \\ &\stackrel{E[X_1]=0}{=} E[X_1Y] \\ &= E[X_1(3 \sum_{i \in I''} X_i + \varepsilon)] \\ &= E[3X_1^2 + 3 \sum_{i \in I'' \setminus \{1\}} X_1X_i + X_1\varepsilon] \\ &= 3Var(X_1) + 3 \sum_{i \in I'' \setminus \{1\}} Cov(X_1, X_i) + Cov(X_1, \varepsilon) \\ &= 3Var(X_1) + 3 \sum_{i \in I'' \setminus \{1\}} Cov(X_1, X_i) + Cov(X_1, \varepsilon) \\ &= 3 \cdot 1.01 + 3 \cdot (4 \cdot 1) + 0 \\ &= 15.03 \end{aligned}$$

$$E[\rho_{X_1, Y}] = E\left[\frac{Cov(X_1, Y)}{\sqrt{Var(Y)}}\right] = E\left[\frac{15.03}{\sqrt{631.35}}\right] = 0.5006.$$

S obzirom na tumačenja koja slijede iz tablice 4.5, značajkama  $X_i, i \in I''$  i dodjeljuje se kategorija *Jak utjecaj*, a ostalim značajkama kategorija *Nema utjecaja*.

#### 5.4.2. Procjena penalizacijskih koeficijenata algoritmom pretraživanja kukavice

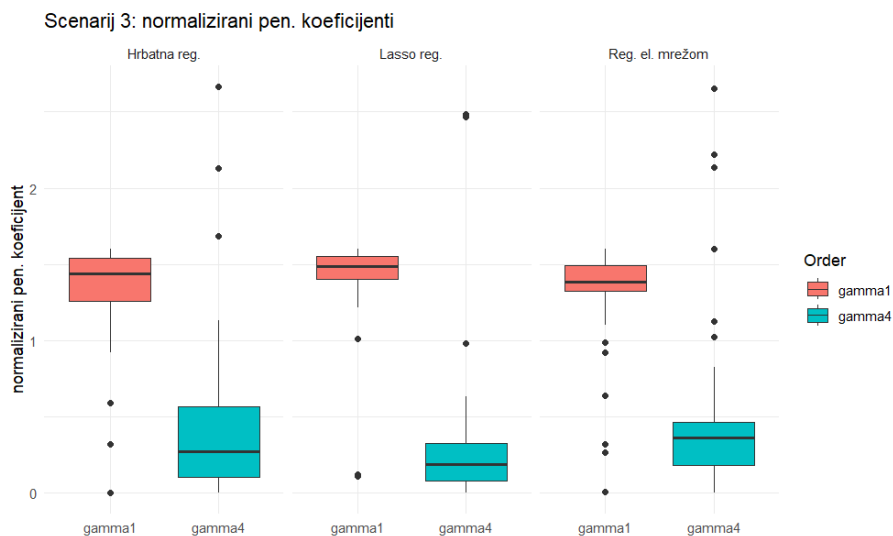
Budući da je u istraživanju simulirano po 50 skupova podataka iz ovog scenarija te da su se parametri regularizacije optimizirali zasebno za svaki od njih i za svaku metodu regularizacije, distribucije normaliziranih penalizacijskih koeficijenata prikazane su grafički (slike 5.39 i 5.40) i deskriptivno (tablica 5.17) u nastavku. U tablici je vidljivo kako su za svaku metodu regularizacije srednje vrijednosti (medijan i aritmetička sredina) procjena penalizacijskih koeficijenata nižeg iznosa za  $\gamma_4$  nego za  $\gamma_1$ .

Tablica 5.17: Scenarij 3 - deskriptivna analiza procjene penalizacijskog koeficijenata ovisno o vrsti regularizacije (normalizirane vrijednosti)

Metoda	Pen. koef.	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
hrbatna reg.	$\gamma_1$	0.000	1.258	1.438	1.332	1.538	1.600	0.326
	$\gamma_4$	0.001	0.103	0.270	0.446	0.569	2.667	0.543
lasso reg.	$\gamma_1$	0.109	1.405	1.486	1.419	1.551	1.598	0.291
	$\gamma_4$	0.004	0.081	0.191	0.302	0.326	2.486	0.484
reg. el. mrežom	$\gamma_1$	0.010	1.323	1.382	1.305	1.491	1.598	0.340
	$\gamma_4$	0.004	0.181	0.363	0.492	0.462	2.650	0.566

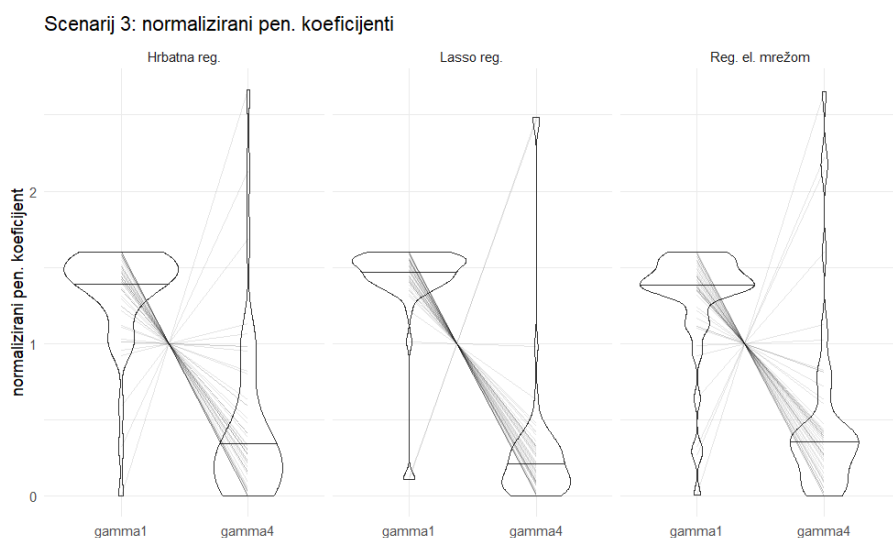
$\gamma_1, \gamma_4$ : penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja* i *Jak utjecaj*, redom.

Nadalje, na slici 5.39 prikazani su kutijasti dijagrami penalizacijskih koeficijenata. Vidljiv je pomak u distribucijama procjena penalizacijskih koeficijenata  $\gamma_1$  i  $\gamma_4$  te da procjena  $\gamma_1$  rezultira manjom varijancom. Nadalje, na slici 5.40 prikazane su distribucije procjena penalizacijskih koeficijenata i najčešći profili penalizacijskih koeficijenata po skupovima. Premda je bilo profila gdje su značajkama snažnijeg utjecaja procijenjeni veći penalizacijski koeficijent nego značajkama slabijeg utjecaja, vidljivo je da su to rjeđe situacije. Češća situacija bila je da su procijenjeni manji penalizacijski koeficijenti značajkama snažnijeg utjecaja i obratno.



Slika 5.39: Kutijasti dijagram distribucija procjene penalizacijskih koeficijenta za skupove iz scenarija 3.

$\gamma_1, \gamma_4$ : normalizirani penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja* i *Jak utjecaj*, redom



Slika 5.40: Violinski dijagram distribucija procjene penalizacijskih koeficijenta s profilima procjena za skupove iz scenarija 3.

$\gamma_1, \gamma_4$ : normalizirani penalizacijski koeficijenti uz značajke kategorije *Nema utjecaja* i *Jak utjecaj*, redom.

### 5.4.3. Procjena koeficijenta modela

U nastavku su prikazani rezultati procjene koeficijenta modela iz Scenarija 3. S obzirom na to da je u ovom scenariju  $p = 40$ , a od toga je 15 prediktora s koeficijentom različitim od nule,



odnosno:

$$\beta = (\underbrace{3, 3, \dots, 3}_{15}, \underbrace{0, 0, \dots, 0}_{25}),$$

te postoje tri podjednako važne grupe od 5 značajki te 25 šumova, radi preglednosti su prikazani rezultati za po jedan prediktor iz svake grupe prediktora s koeficijentima različitim od nule. Prikazani su i rezultati procjene slobodnog koeficijenta. Za preostale su se koeficijente koji su jednaki nuli i distribucije procjena kretale oko nule, ali radi preglednosti njihovi su grafički prikazi i deskriptivna analiza izostavljeni iz rada.

Tablica 5.18: Scenarij 3 - deskriptivna analiza procjene koeficijenata modela ovisno o vrsti regularizacije i načinu procjene penalizacijskog koeficijenta

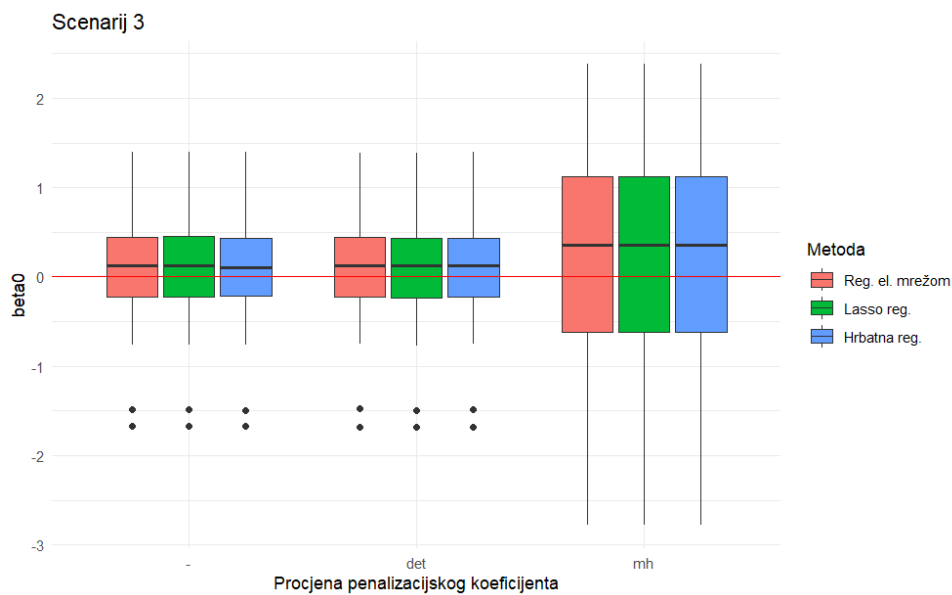
Koef.	Procjena pen. koef.	Metoda	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
$\beta_0 = 0$	-	lasso reg.	-1.68	-0.23	0.12	0.07	0.44	<b>1.39</b>	0.57
		hrbatna reg.	<b>-1.67</b>	-0.23	0.13	0.07	0.45	<b>1.39</b>	0.57
		reg. el. mrežom	<b>-1.67</b>	<b>-0.21</b>	<b>0.10</b>	<b>0.06</b>	<b>0.43</b>	1.40	0.57
	det	lasso reg.	-1.68	-0.23	0.13	0.07	0.45	<b>1.39</b>	<b>0.56</b>
		hrbatna reg.	-1.68	-0.24	0.13	<b>0.06</b>	0.44	<b>1.39</b>	0.57
		reg. el. mrežom	-1.69	-0.22	0.13	0.07	0.44	1.40	0.57
	mh	lasso reg.	-2.78	-0.61	0.35	0.24	1.12	2.38	1.08
		hrbatna reg.	-2.78	-0.61	0.35	0.24	1.12	2.38	1.08
		reg. el. mrežom	-2.78	-0.61	0.35	0.24	1.12	2.38	1.08
$\beta_1 = 3$	-	lasso reg.	0.00	2.23	2.50	2.73	<b>3.02</b>	9.49	1.62
		hrbatna reg.	0.00	0.00	0.26	2.40	5.31	10.10	3.20
		reg. el. mrežom	2.20	2.43	2.53	2.54	2.66	2.94	<b>0.16</b>
	det	lasso reg.	0.00	2.38	2.63	2.85	2.82	10.11	1.63
		hrbatna reg.	0.00	0.00	0.34	2.46	5.29	10.11	3.24
		reg. el. mrežom	<b>2.26</b>	<b>2.48</b>	2.59	2.60	2.71	<b>3.02</b>	0.17
	mh	lasso reg.	-2.20	1.61	2.69	<b>3.05</b>	3.82	10.95	2.58
		hrbatna reg.	-8.55	0.00	1.12	2.84	5.46	11.58	4.11
		reg. el. mrežom	-1.85	2.19	<b>2.84</b>	3.15	3.64	11.71	2.36

Koef.	Procjena pen. koef.	Metoda	Min.	Prvi kvartil	Med.	Arit. sred.	Treći kvartil	Max.	St. dev.
$\beta_6 = 3$	-	lasso reg.	0.00	2.02	2.66	2.59	<b>2.92</b>	8.47	1.56
		hrbatna reg.	0.00	0.00	2.12	2.92	5.26	8.69	2.84
		reg. el. mrežom	2.31	2.45	2.51	2.53	2.62	2.76	<b>0.12</b>
	det	lasso reg.	0.00	2.44	2.67	2.51	2.79	6.90	1.22
		hrbatna reg.	0.00	0.00	2.16	<b>2.97</b>	5.39	8.84	2.86
		reg. el. mrežom	<b>2.32</b>	2.50	2.59	2.59	2.69	<b>2.85</b>	0.13
	mh	lasso reg.	-0.24	2.51	3.08	3.49	4.74	9.22	1.98
		hrbatna reg.	-1.81	0.58	<b>2.96</b>	3.58	6.18	13.57	3.43
		reg. el. mrežom	-0.84	<b>2.59</b>	3.14	3.27	3.84	9.33	1.59
$\beta_{11} = 3$	-	lasso reg.	0.00	2.51	2.68	2.81	3.18	7.82	1.45
		hrbatna reg.	0.00	0.00	<b>3.00</b>	<b>3.08</b>	4.90	10.88	2.93
		reg. el. mrežom	2.35	2.47	2.55	2.56	2.63	2.84	<b>0.11</b>
	det	lasso reg.	0.00	2.57	2.72	2.75	<b>2.88</b>	7.82	1.26
		hrbatna reg.	0.00	0.00	3.05	3.12	4.90	10.88	2.95
		reg. el. mrežom	<b>2.39</b>	2.52	2.60	2.62	2.70	<b>2.90</b>	<b>0.11</b>
	mh	lasso reg.	-1.27	2.55	3.25	3.66	4.34	10.08	2.34
		hrbatna reg.	-9.03	0.46	3.79	3.85	6.82	11.80	4.03
		reg. el. mrežom	-1.29	<b>2.68</b>	3.23	3.67	3.80	9.65	1.99

Najtočnija procjena koeficijena modela za svaku deskriptivnu statistiku označena je podebljano.

Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika).

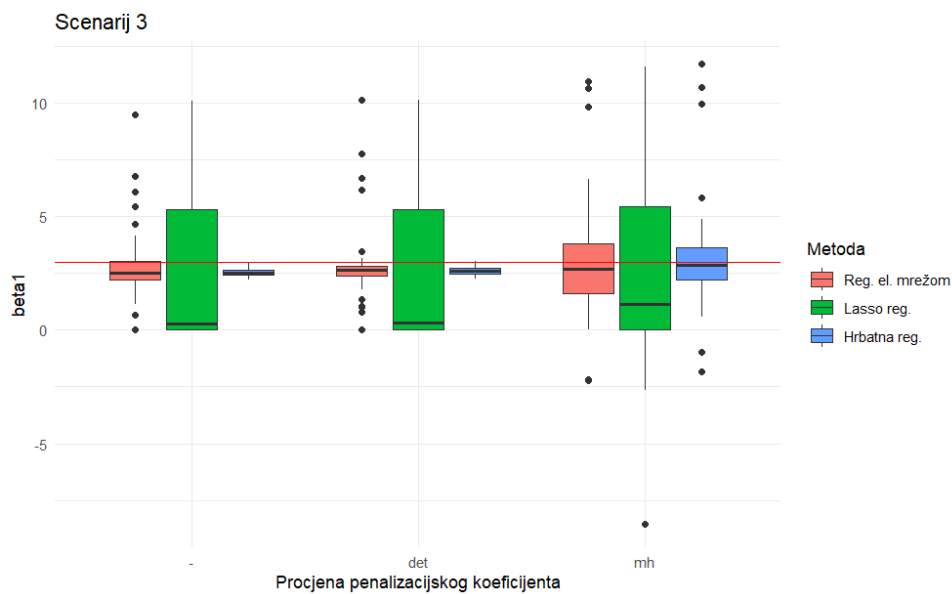
Na slikama u nastavku grafički su prikazane distribucije procjena koeficijena modela na skupovima iz trećeg scenarija. Vidljivo je da se za sve metode regularizacije procijenjena vrijednost slobodnog koeficijenta  $\beta_0$  kretala oko nule, dok je u slučaju uvođenja domenskog znanja stručnjaka metaheuristikom imala veću varijancu, a u slučaju kada se domensko znanje stručnjaka ne koristi ili je uvedeno deterministički, medijalna vrijednost bila je bliža nuli (slika 5.41). Nadalje, na slikama 5.42 i 5.43 vidljivo je kako su za sve metode regularizacije medijalne vrijednosti procjene koeficijenta  $\beta_1$  i  $\beta_6$  najbliže kretale oko stvarne vrijednosti (koja iznosi 3) za uvođenje domenskog znanja metaheuristikom. Vidljivo je i da lasso regularizacija ima najveću varijancu i pristranost u procjeni koeficijenta  $\beta_1, \beta_6, \beta_{11}$  (slike 5.42, 5.43, 5.44).



Slika 5.41: Grafički prikazi distribucija procjene koeficijenta  $\beta_0$  za skupove iz Scenarija 3.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

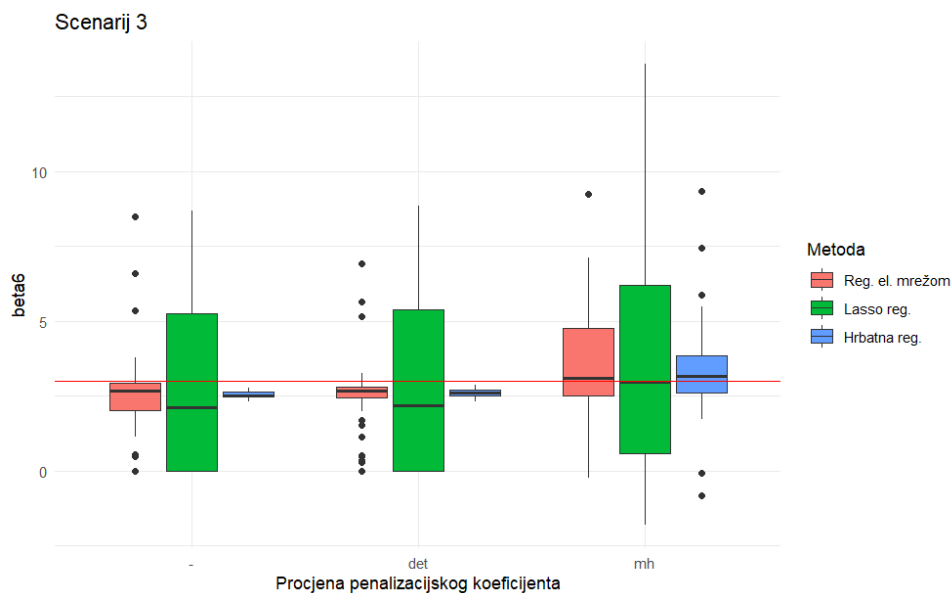
Crveni pravac predstavlja prave vrijednosti koeficijenata.



Slika 5.42: Grafički prikazi distribucija procjene koeficijenta  $\beta_1$  za skupove iz Scenarija 3.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

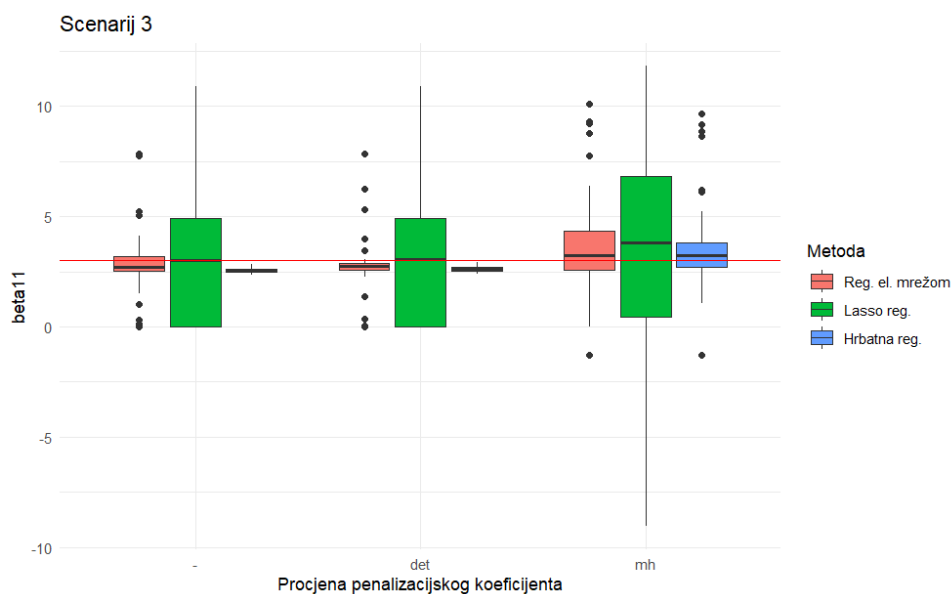
Crveni pravac predstavlja prave vrijednosti koeficijenata.



Slika 5.43: Grafički prikazi distribucija procjene koeficijenta  $\beta_6$  za skupove iz Scenarija 3.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Crveni pravac predstavlja prave vrijednosti koeficijenata.



Slika 5.44: Grafički prikazi distribucija procjene koeficijenta  $\beta_{11}$  za skupove iz Scenarija 3.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Crveni pravac predstavlja prave vrijednosti koeficijenata.

Kao i u prethodnim slučajevima, i na skupovima simuliranim po Scenariju 3 metode regularizacije s metaheurističkim pristupom procjene penalizacijskog koeficijenta imale su dulje vrijeme izvođenja te je u tu svrhu korištena usluga Naprednog računanja (Padobran) Sveuči-

lišnog računskog centra (Srce) Sveučilišta u Zagrebu. U tablici 5.19 u nastavku prikazano je prosječno vrijeme izvođenja za jedan skup prilikom kojeg se koristilo pet procesora.

Tablica 5.19: Prosječno vrijeme izvođenja metoda regularizacije s metaheurističkom procjenom penalizacijskog koeficijenta za skupove iz Scenarija 3.

Metoda	lasso reg.	hrbatna reg.	reg. el. mrežom
Vrijeme	02:43:45	00:49:18	01:38:07

Prilikom izvođenja za jedan skup korišteno je pet procesora, a prikazano vrijeme odnosi se na sate, minute i sekunde.

#### 5.4.4. Evaluacija modela i hipoteza

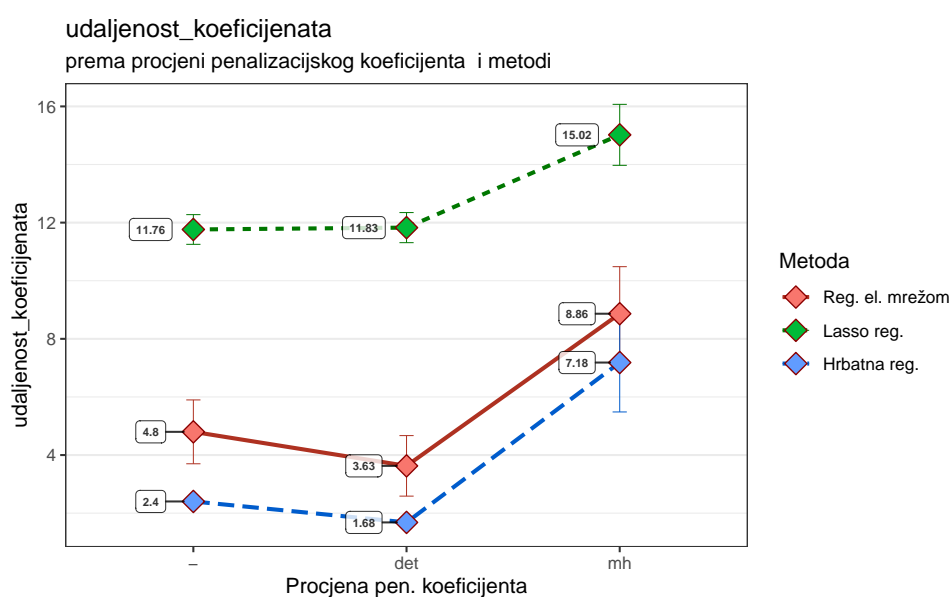
U potpoglavlju 4.5 (str. 50), objašnjeno kako su metodom ANOVA analizirani rezultati modela svih 50 skupova iz Scenarija 3 s ciljem donošenja odluke o hipotezama. Provjera pretpostavki za analizu varijance izvršena je analizom dijagrama raspršenosti reziduala modela u odnosu na razine faktora te qq-dijagramom distribucije reziduala u usporedbi sa standardnom normalnom distribucijom za svaku mjeru evaluacije modela unutar pojedine hipoteze. Spomenuti dijagrami su, radi preglednosti, prikazani u Prilogu 4 (slike 7.13, 7.14 i 7.15), a s obzirom na to da nisu upućivali na značajna odstupanja od ovih pretpostavki, smatra se valjanim koristiti analizu varijance. Odluke se donose na razini statističke značajnosti 0.05.

#### Hipoteza 1

Kao u Scenariju 1 i Scenariju 2, u ovoj se hipotezi uspoređuju se modeli regularizacije koji ne uključuju znanje stručnjaka s modelima regularizacije koji ga uključuju deterministički i metaheuristički kroz pet različitih mjera evaluacije. Za evaluaciju hipoteze koristi se dvofaktorska ANOVA s faktorima Procjena penalizacijskog koeficijenta i Metoda regularizacije te njihovom interakcijom. Kontrasti su pojašnjeni u potpoglavlju 4.5 (str. 50), pri čemu se pomoću kontrasta c11 testira upravo Hipoteza 1 ovog istraživanja, dok kontrast c12 uspoređuje međusobno dva načina uvođenja domenskog znanja stručnjaka u metode regularizacije.

U nastavku su za svaku mjeru evaluacije prikazani interakcijski grafikoni metode regularizacije i različitih načina procjene penalizacijskog koeficijenta.

Na slici 5.45 prikazane su prosječne vrijednosti udaljenosti koeficijenata prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako za lasso regularizaciju prosječna vrijednost udaljenosti koeficijenata poraste uvođenjem znanja stručnjaka determinističkim pristupom, a dodatno poraste kada se isto uvede metaheurističkim pristupom. Za hrbatnu regularizaciju i regularizaciju elastičnom mrežom uvođenjem domenskog znanja na deterministički način opada prosječna vrijednost udaljenosti koeficijenata, a onda ona raste kada se domensko znanje uvede metaheuristikom na razinu iznad početne. Vidljivo je da je za lasso regularizaciju u svim slučajevima veća prosječna udaljenost koeficijenata u odnosu na druge metode regularizacije.



Slika 5.45: Scenarij 3, Hipoteza 1 - Udaljenost koeficijenata prema procjeni penalizacijskog koeficijenta i metodi.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

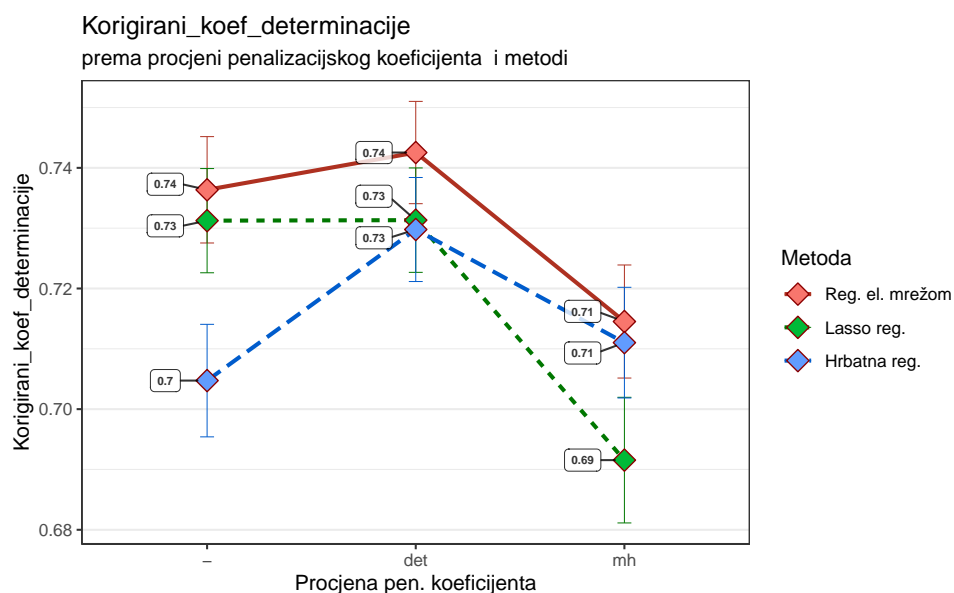
Za mjeru evaluacije udaljenost koeficijenata interakcijski član ( $F(4,441) = 1.644$ ,  $p = 0.162$ ) nije statistički značajan, dok su efekti faktora Procjena penalizacijskog koeficijenta ( $F(2,441) = 73.4$ ,  $p < 2 \cdot 10^{-16}$ ) i faktora Metoda regularizacije ( $F(2,441) = 264.307$ ,  $p < 2 \cdot 10^{-16}$ ) statistički značajni.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.12 u Prilogu 4. U spomenutoj tablici vidljivo je kako su statistički značajni koeficijenti uz sve pojedinačne kontraste (kontrasti uz faktor Penalizacijski koeficijent i faktor Metoda) te koeficijent uz interakciju kon-

trasta c12 i c22.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječne vrijednosti udaljenosti koeficijenata između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna (diff = -0.61,  $p = 0.3098$ ). Razlika prosječne vrijednosti udaljenosti koeficijenata između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka je pozitivna i statistički značajna (diff = 4.035,  $p = 0.00$ ), kao i razlika prosječne vrijednosti udaljenosti koeficijenata između metaheurističkog i determinističkog pristupa (diff = 4.644,  $p = 0.00$ ).

Nadalje, *post hoc* analiza Tukeyjevim testom HSD za faktor Metoda kaže da je razlika između prosječne vrijednosti udaljenosti koeficijenata lasso regularizacije i regularizacije elastičnom mrežom pozitivna i statistički značajna (diff = 7.107,  $p = 0.00$ ). Razlika prosječne vrijednosti udaljenosti koeficijenata između hrbatne regularizacije i regularizacije elastičnom mrežom je negativna i statistički značajna (diff = -2.007,  $p = 5.945 \text{ e-}6$ ), kao i razlika prosječne vrijednosti udaljenosti koeficijenata hrbatne i lasso regularizacije (diff = -9.115,  $p = 0.00$ ).



Slika 5.46: Scenarij 3, Hipoteza 1 - Korigirani koeficijent determinacije prema procjeni penalizacijskog koeficijenta i metodi.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Na slici 5.46 prikazane su prosječne vrijednosti korigiranog koeficijenta determinacije prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako za hrbatnu regula-

rizaciju prosječna vrijednost korigiranog koeficijenta determinacije poraste uvođenjem domenskog znanja stručnjaka, pri čemu više poraste kod determinističkog pristupa. Za regularizaciju elastičnom mrežom prosječna vrijednost korigiranog koeficijenta determinacije poraste uvođenjem domenskog znanja stručnjaka determinističkim pristupom, a opadne kod metaheurističkog pristupa. Za lasso regularizaciju prosječna vrijednost korigiranog koeficijenta determinacije ostane na podjednakoj razini uvođenjem domenskog znanja stručnjaka determinističkim pristupom, dok opadne kod uvođenja domenskog znanja metaheuristikom.

Za mjeru evaluacije korigirani koeficijent determinacije statistički su značajni interakcijski član ( $F(4, 441) = 6.688$ ,  $p = 3.12 \cdot 10^{-5}$ ), kao i efekti faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 31.401$ ,  $p = 1.78 \cdot 10^{-13}$ ) i faktora Metoda regularizacije ( $F(2, 441) = 10.661$ ,  $p = 3.01 \cdot 10^{-5}$ ). Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.12 u Prilogu 4 (str. 193).

Budući da je interakcijski član dvaju faktora statistički značajan, efekt pojedinog načina uvođenja domenskog znanja stručnjaka na korigirani koeficijent determinacije ne može biti generaliziran na sve metode regularizacije. U interpretaciji rezultata potrebno je koristiti *post hoc* analizu interakcije. U tablici 7.13 u Prilogu 4 (str. 194) prikazan je rezultat *post hoc* analize Tukeyjevim testom HSD te vidljivo je koje su razlike statistički značajno različite od nule na razini statističke značajnosti 0.05.

Vidljivo je kako je za hrbatnu regularizaciju razlika u prosječnom korigiranom koeficijentu determinacije između između determinističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka pozitivna i statistički značajna. S druge strane, razlika u prosječnom korigiranom koeficijentu determinacije između metaheurističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka je pozitivna i nije statistički značajna, dok je razlika u prosječnom korigiranom koeficijentu determinacije između metaheurističkog i determinističkog pristupa negativna i statistički značajna.

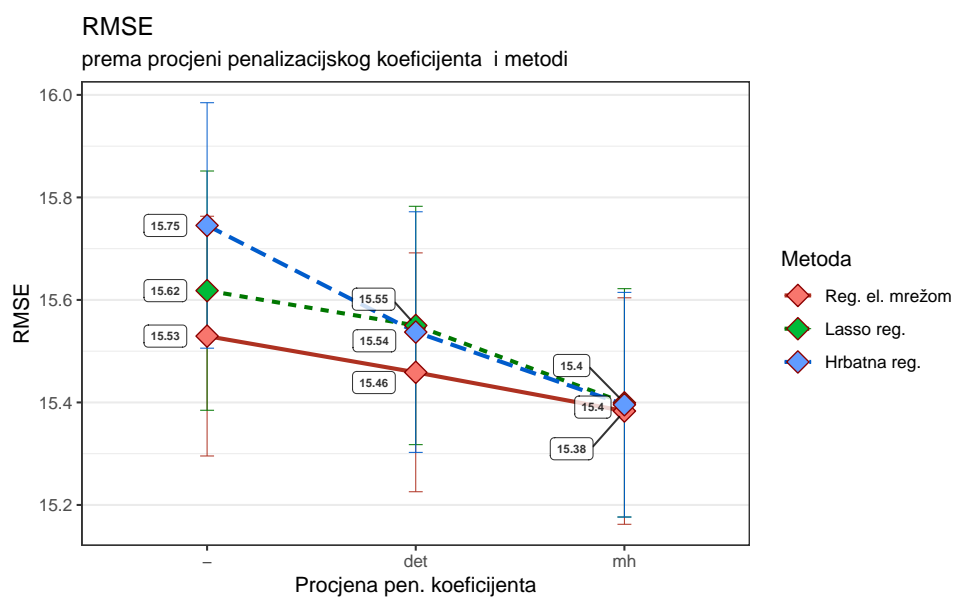
Nadalje, za lasso regularizaciju i regularizaciju elastičnom mrežom vidljivo je kako razlika u prosječnom korigiranom koeficijentu determinacije između između determinističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka nije statistički značajna. S druge strane, razlike u prosječnom korigiranom koeficijentu determinacije između metaheurističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka te između metaheurističkog i determinističkog pristupa su negativne i statistički značajne.

Dakle, uvođenje domenskog znanja stručnjaka doprinijelo je statistički značajnom poveća-



nju prosječnog korigiranog koeficijenta determinacije samo za hrbatnu regularizaciju i deterministički pristup.

Na slici 5.47 prikazane su prosječne vrijednosti *RMSE* greške prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je uvođenjem domenskog znanja stručnjaka na deterministički način opada prosječna vrijednost greške, te ona dodatno opada uvođenjem domenskog znanja metaheuristikom. Vidljivo je da je za regularizaciju elastičnom mrežom u svim slučajevima najniža prosječna *RMSE*.

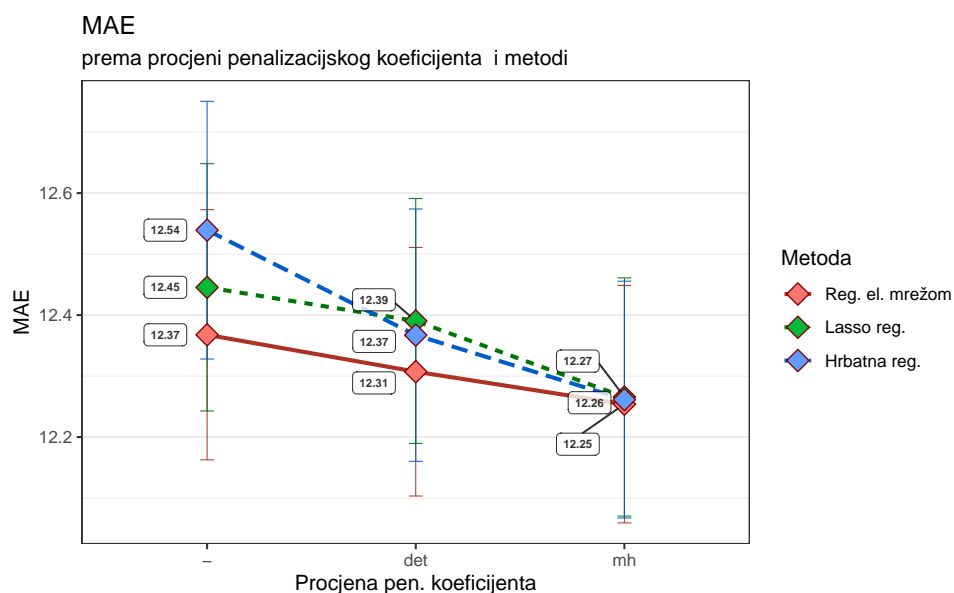


Slika 5.47: Scenarij 3, Hipoteza 1 - *RMSE* prema procjeni penalizacijskog koeficijenta i metodi. Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Za mjeru evaluacije *RMSE* interakcijski član ( $F(4, 441) = 0.238$ ,  $p = 0.9166$ ) i efekt faktora Metoda regularizacije ( $F(2, 441) = 0.614$ ,  $p = 0.5417$ ) nisu statistički značajni, dok je efekt faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 3.250$ ,  $p = 0.0397$ ) statistički značajan.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.12 u Prilogu 4 (str. 193). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c11 koji govori o razlici prosječnih *RMSE* metoda regularizacije koje ne koriste domensko znanje u odnosu na one koje ga uključuju deterministički ili metaheuristikom. Pozitivan iznos koeficijenta upućuje da je prosječna *RMSE* niža u metodama regularizacije koje koriste domensko znanje stručnjaka. Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika

prosječnih *RMSE* između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna ( $\text{diff} = -0.116$ ,  $p = 0.4324$ ), dok je razlika prosječnih *RMSE* između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna ( $\text{diff} = -0.238$ ,  $p = 0.0299$ ).



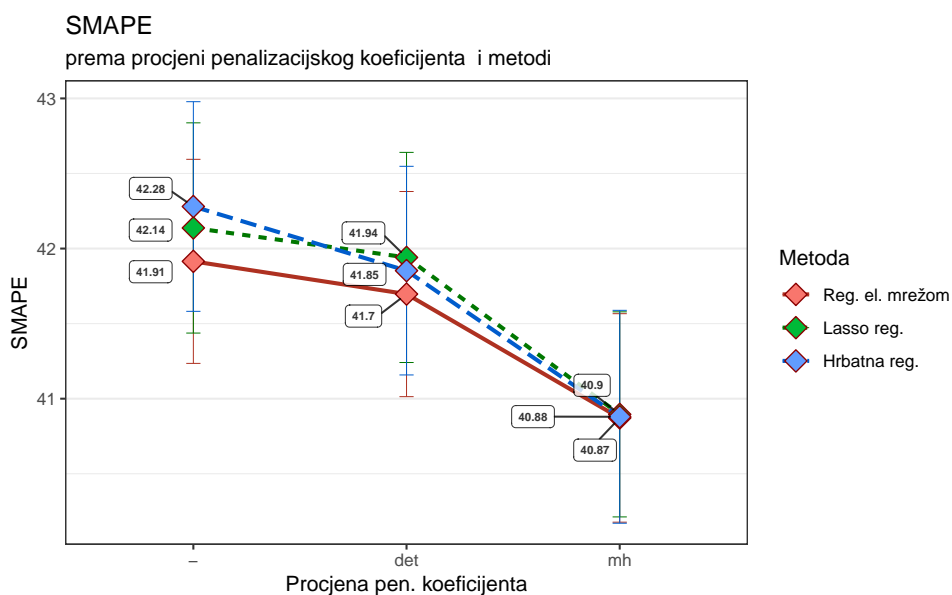
Slika 5.48: Scenarij 3, Hipoteza 1 - *MAE* prema procjeni penalizacijskog koeficijenta i metodi. Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Na slici 5.48 prikazane su prosječne vrijednosti *MAE* prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je kako uvođenjem domenskog znanja stručnjaka na deterministički način opada prosječna vrijednost greške te da ona dodatno opada uvođenjem domenskog znanja metaheuristikom. Vidljivo je da je za regularizaciju elastičnom mrežom u svim slučajevima najniža prosječna *MAE*.

Za mjeru evaluacije *MAE* nisu statistički značajni interakcijski član ( $F(4, 441) = 0.208$ ,  $p = 0.9341$ ), kao niti efekti faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 2.700$ ,  $p = 0.0683$ ) i faktora Metoda regularizacije ( $F(2, 441) = 0.504$ ,  $p = 0.6046$ ).

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.12 u Prilogu 4 (str. 193). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c11 koji govori o razlici prosječnih *MAE* metoda regularizacije koje ne koriste domensko znanje u odnosu na one koje ga uključuju deterministički ili metaheuristikom. Pozitivan iznos koeficijenta upućuje da je prosječna *MAE* niža u metodama regularizacije koje koriste domensko

znanje stručnjaka. Međutim, *post hoc* analiza Tukeyjevim testom HSD pokazuje da nema značajnih efekata.



Slika 5.49: Scenarij 3, Hipoteza 1 - *SMAPE* prema procjeni penalizacijskog koeficijenta i metodi.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Na slici 5.49 prikazane su prosječne vrijednosti *SMAPE* prema procjeni penalizacijskog koeficijenta i metodi regularizacije. Vidljivo je uvođenjem domenskog znanja stručnjaka na deterministički način opada prosječna vrijednost greške, te ona dodatno opada uvođenjem domenskog znanja metaheuristikom. Vidljivo je da je za regularizaciju elastičnom mrežom u svim slučajevima najniža prosječna *SMAPE*.

Za mjeru evaluacije *SMAPE* nisu statistički značajni interakcijski član ( $F(4, 441) = 0.085$ ,  $p = 0.987$ ) i efekt faktora Metoda regularizacije ( $F(2, 441) = 0.243$ ,  $p = 0.784$ ), dok efekt faktora Procjena penalizacijskog koeficijenta ( $F(2, 441) = 10.435$ ,  $p = 3.73 \cdot 10^{-5}$ ) statistički je značajan.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.12 u Prilogu 4 (str. 193). U spomenutoj tablici vidljivo je kako su statistički značajni jedino koeficijenti uz kontraste c11 i c12.

Kontrast c11 govori o razlici prosječnih *SMAPE* greški između metoda regularizacije koje ne koriste domensko znanje u odnosu na one koje ga uključuju deterministički ili metaheuris-

tikom. Pozitivan iznos koeficijenta upućuje da je prosječna *SMAPE* niža u metodama regularizacije koje koriste domensko znanje stručnjaka. *Post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika prosječnih *SMAPE* greški između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna ( $\text{diff} = -0.28$ ,  $p = 0.5802$ ), dok je razlika prosječnih *SMAPE* greški između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna ( $\text{diff} = -1.228$ ,  $p = 4.83 \text{ e-}5$ ).

Kontrast c12 govori o razlici prosječnih *SMAPE* greški metoda regularizacije koje penalizacijske koeficijente procjenjuju deterministički s onima koje ih procjenjuju metaheuristikom. Negativan iznos koeficijenta upućuje da je prosječna *SMAPE* niža u slučaju uvođenja domenskog znanja stručnjaka metaheuristikom, u odnosu na determinističko uvođenje.

## Hipoteza 2

U ovoj se hipotezi kroz tri različite mjere evaluacije uspoređuju metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka. Hipoteza se dijeli na dvije podhipoteze - u jednoj se uspoređuje metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka deterministički (Hipoteza 2a), a u drugoj se metoda CART uspoređuju s metodama regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom (Hipoteza 2b).

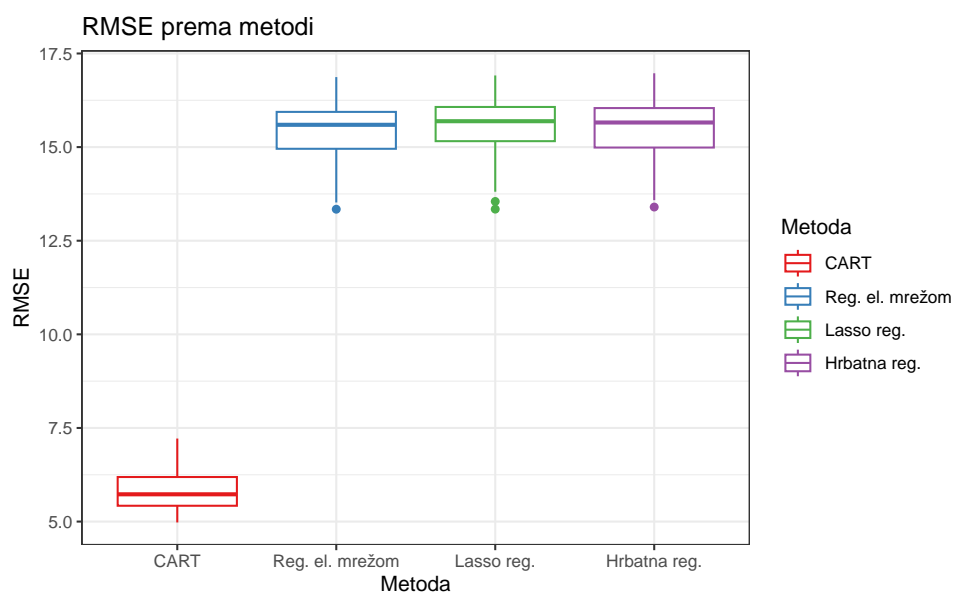
### Hipoteza 2a

U ovoj hipotezi uspoređuje se metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka deterministički. Podsjetimo, za evaluaciju hipoteze koristi se jednofaktorska ANOVA s faktorom Metoda. Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2 ovog istraživanja. Odluke se donose na razini statističke značajnosti 0.05. U nastavku su grafički prikazane distribucije svake mjere evaluacije.

Za mjeru evaluacije *RMSE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2043$ ,  $p < 2 \text{ e-}16$ ). Na slici 5.50 vidljivo je kako CART metoda ima nižu prosječnu *RMSE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.14 u Prilogu 4 (str. 197). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *RMSE* od metoda regularizacije koje uvode domensko znanje stručnjaka determi-

nistički.



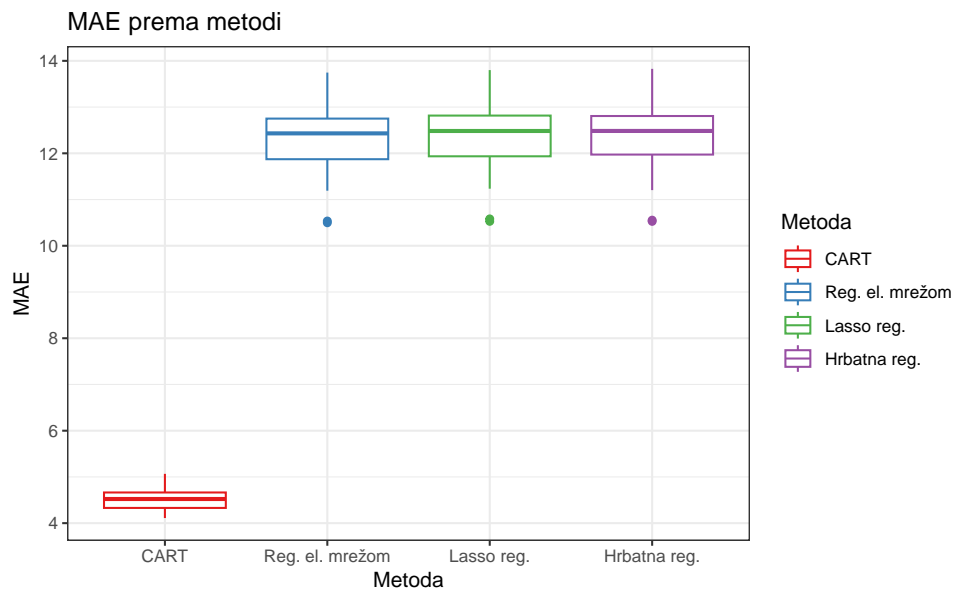
Slika 5.50: Scenarij 3, Hipoteza 2a - *RMSE* prema metodi.

U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih *RMSE* metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna (diff = - 9.642,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 9.734,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 9.721,  $p = 0.00$ ).

Za mjeru evaluacije *MAE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1928$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.51 vidljivo je kako CART metoda ima nižu prosječnu *MAE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.14 u Prilogu 4 (str. 197). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *MAE* od metoda regularizacije koje uvode domensko znanje stručnjaka deterministički.



Slika 5.51: Scenarij 3, Hipoteza 2a - MAE prema metodi.

U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

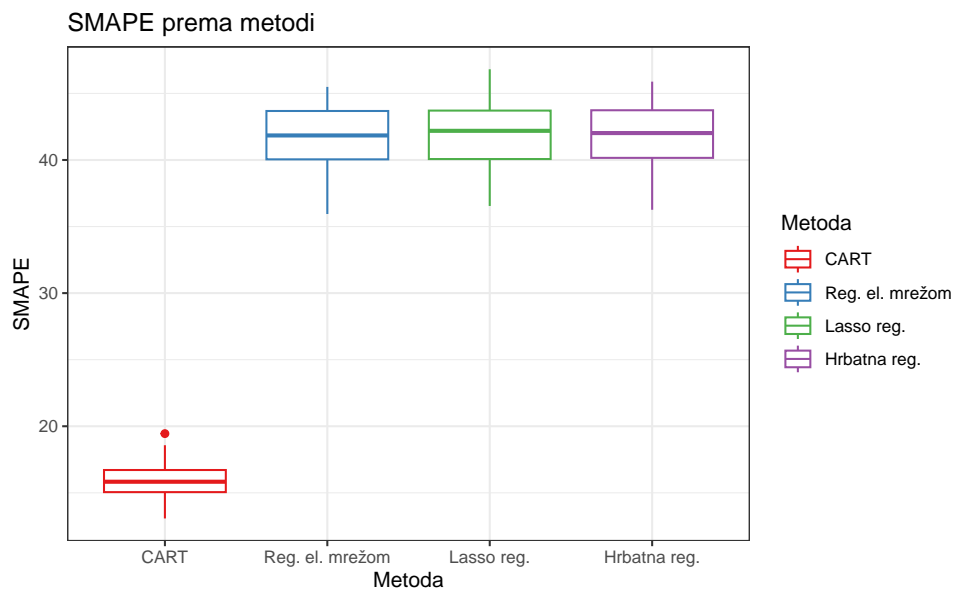
Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih MAE metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna (diff = - 7.781,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 7.864,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 7.841,  $p = 0.00$ ).

Za mjeru evaluacije *SMAPE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1698$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.52 vidljivo je kako CART metoda ima nižu prosječnu *SMAPE* od metoda regularizacije s deterministički uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.14 u Prilogu 4 (str. 197). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *SMAPE* od metoda regularizacije koje uvode domensko znanje stručnjaka deterministički.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih *SMAPE* greški metode CART i regularizacije elastičnom mrežom s deterministički uvedenim znanjem stručnjaka statistički značajna (diff = - 25.786,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 26.03,  $p =$

0.00) te metode CART i hrbatne regularizacije s deterministički uvedenim znanjem stručnjaka (diff = - 25.942,  $p = 0.00$ ).



Slika 5.52: Scenarij 3, Hipoteza 2a - *SMAPE* prema metodi.

U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

### Hipoteza 2b

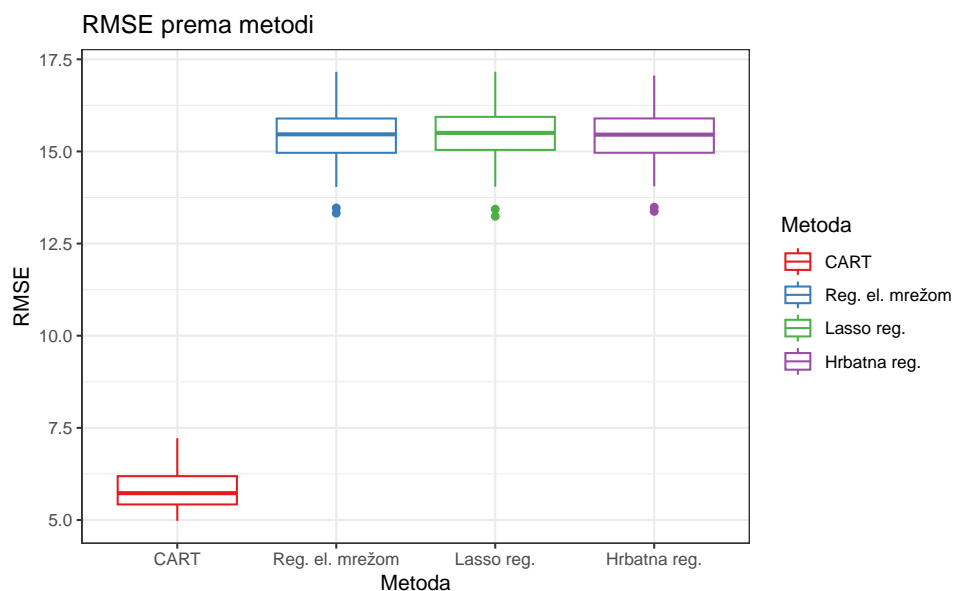
U ovoj hipotezi uspoređuje se metoda CART s metodama regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom. Podsjetimo, za evaluaciju hipoteze koristi se jednofaktorska ANOVA s faktorom Metoda. Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2 ovog istraživanja. Odluke se donose na razini statističke značajnosti 0.05. U nastavku su grafički prikazane distribucije svake mjere evaluacije.

Za mjeru evaluacije *RMSE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2192$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.53 vidljivo je kako CART metoda ima nižu prosječnu *RMSE* od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.15 u Prilogu 4 (str. 199). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *RMSE* od metoda regularizacije koje uvode domensko znanje stručnjaka metaheuristikom.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih

*RMSE* metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 9.567,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 9.583,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 9.579,  $p = 0.00$ ).



Slika 5.53: Scenarij 3, Hipoteza 2 b - *RMSE* prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

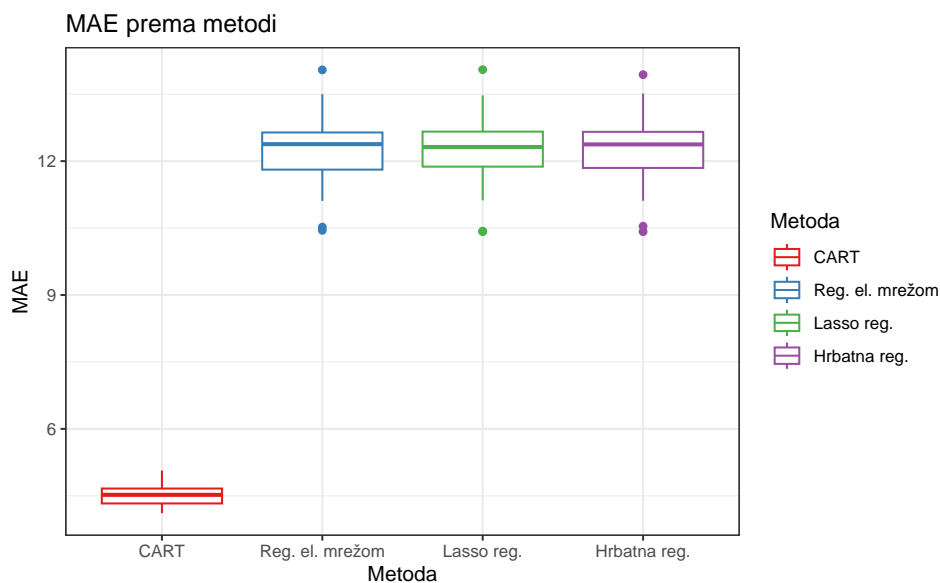
Za mjeru evaluacije *MAE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 2056$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.54 vidljivo je kako CART metoda ima nižu prosječnu *MAE* od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.

Procjene koeficijenata s intervalima pouzdanosti dane su u tablici 7.15 u Prilogu 4 (str. 199). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom prosječnom *MAE* od metoda regularizacije koje uvode domensko znanje stručnjaka metaheuristikom.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između *MAE* metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 7.728,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 7.74,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = -



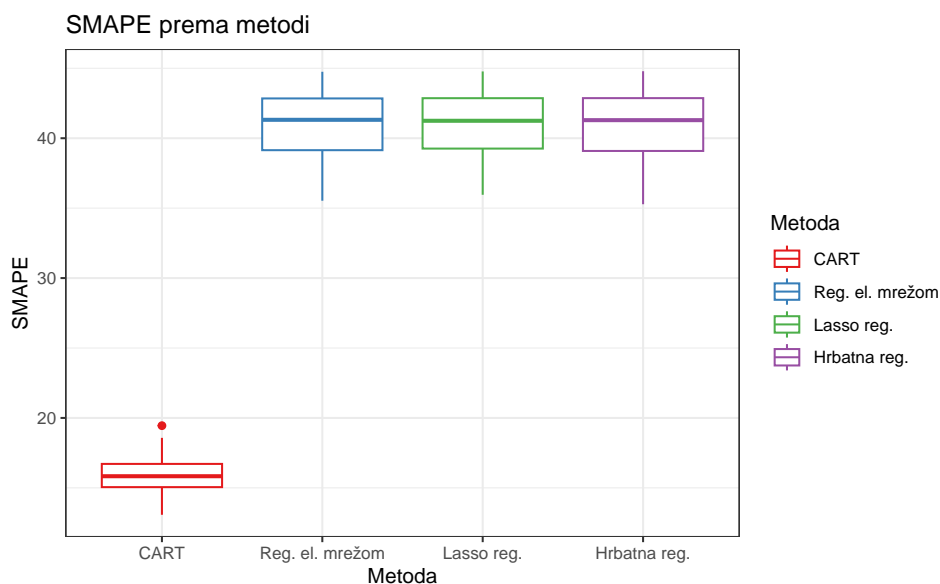
7.736,  $p = 0.00$ ).



Slika 5.54: Scenarij 3, Hipoteza 2b - MAE prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

Za mjeru evaluacije *SMAPE* statistički značajan je efekt faktora Metoda ( $F(3, 196) = 1565$ ,  $p < 2 \cdot 10^{-16}$ ). Na slici 5.55 vidljivo je kako CART metoda ima nižu *SMAPE* od metoda regularizacije s metaheuristikom uvedenim znanjem stručnjaka.



Slika 5.55: Scenarij 3, Hipoteza 2b - SMAPE prema metodi.

U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

Procjene koeficijenta s intervalima pouzdanosti dane su u tablici 7.15 u Prilogu 4 (str. 199). U spomenutoj tablici vidljivo je kako je statistički značajan jedino koeficijent uz kontrast c31. Negativna vrijednost koeficijenta uz kontrast c31 upućuje da metoda CART rezultira nižom *SMAPE* od metoda regularizacije koje uvode domensko znanje stručnjaka metaheuristikom.

Dodatno, *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između *SMAPE* greški metode CART i regularizacije elastičnom mrežom s metaheuristikom uvedenim znanjem stručnjaka statistički značajna (diff = - 24.961,  $p = 0.00$ ), kao i ona između metode CART i lasso regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 24.985,  $p = 0.00$ ) te metode CART i hrbatne regularizacije s metaheuristikom uvedenim znanjem stručnjaka (diff = - 24.969,  $p = 0.00$ ).

## 6. RASPRAVA

U nastavku su najprije za svaku vrstu podataka raspravljani rezultati procjene penalizacijskih koeficijenata metaheuristikom. Korištena metaheuristika jest algoritam pretraživanja kukavice (Algoritam 1, str. 20). Potom su analizirane procjene koeficijenata modela za simulirane skupove podataka. Naposljetku je uslijedila rasprava rezultata istraživanja u kontekstu postavljenih hipoteza te usporedba s literaturom.

### 6.1. PROCJENA PENALIZACIJSKIH KOEFICIJENATA ALGORITMOM PRETRAŽIVANJA KUKAVICE

Kod determinističke procjene, odnosno, određivanja penalizacijskih koeficijenata (tablica 4.1 na str. 39), vidljivo je da je značajkama snažnije kategorije utjecaja dodijeljena manja vrijednost penalizacijskog koeficijenta (i obratno). U nastavku su analizirani rezultati procjene vrijednosti penalizacijskih koeficijenata metaheurističkim algoritmom pretraživanja kukavice posebno za svaku vrstu podataka.

#### 6.1.1. Stvarni podaci

Tablica 7.1 (str. 172) prikazuje rezultate mišljenja stručnjaka o utjecaju pojedine ulazne značajke na izlaznu značajku za stvarni skup podataka pojašnjen na str. 40. U istoj je vidljivo kako se Stručnjak 3 prilično razlikuje u svojoj procjeni utjecaja ulaznih značajki te iz tog razloga njegovo mišljenje nije agregirano s mišljenjem preostala tri stručnjaka. U konačnici su se promatrale dvije vrste domenskih znanja, Domensko znanje 1 kao medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4 te Domensko znanje 2 kao mišljenje Stručnjaka 3. Kategorije utjecaja po značajkama za obje vrste domenskog znanja navedene su u tablici 7.2 (str. 173).

Tablica 5.5 (str. 60) prikazuje frekvencije pojedine kategorije značajki za svaku vrstu do-

menskog znanja za početni skup podataka, dok tablica 5.6 (str. 61) prikazuje isto za matricu dizajna. Vidljivo je da prema Domenskom znanju 1 ne postoji niti jedna značajka kategorije *Nema utjecaja*, vrlo ih je malo kategorije *Slab utjecaj*, te ih je znatno više kategorija *Srednji utjecaj* i *Jak utjecaj*. S druge strane, kod Domenskog znanja 2 gotovo je obratna situacija, malo je značajki snažnijih kategorija utjecaja (*Srednji utjecaj* i *Jak utjecaj*), dok je znatno više značajki kategorija *Nema utjecaja* i *Slab utjecaj*. Vidljivo je da su ove dvije vrste domenskog znanja prilično drugačije.

Za rezultate procjene vrijednosti penalizacijskih koeficijenata metaheurističkim algoritmom (vidljivih u tablici 5.8 (str. 62) te na slici 5.1 (str. 63)) naslućuju se dvije različite tendencije kretanja vrijednosti penalizacijskih koeficijenata. Premda ima odstupanja, u slučaju korištenja Domenskog znanja 1 procijenjena vrijednost penalizacijskog koeficijenta ima tendenciju opadanja što je značajka snažnije kategorije utjecaja, dok u slučaju korištenja Domenskog znanja 2 ima tendenciju povećanja što je značajka snažnije kategorije utjecaja. Ovakva razlika u veličinama i rasporedu procjena penalizacijskih koeficijenata između domenskih znanja vrlo vjerojatno potječe od različitih distribucija kategorija utjecaja značajki unutar njih. Nadalje, moglo bi se reći da Domensko znanje 1 rezultira intuitivnijim, odnosno krajnjim korisnicima smislenijim, penalizacijskim koeficijentima nego li Domensko znanje 2.

### 6.1.2. Scenarij 1

Distribucije procjena penalizacijskih koeficijenata za Scenarij 1 prikazane su na slikama 5.6 (str. 73) i 5.7 (str. 73) te u tablici 5.11 (str. 72). Na istima je za svaki penalizacijski koeficijent vidljiva distribucija procjena te profili procjena penalizacijskih koeficijenata za svih 50 skupova podataka i za svaku vrstu regularizacije (jedan profil čine sve procjene penalizacijskih koeficijenata za jedan skup podataka iz scenarija i jednu metodu regularizacije). Premda je bilo stršila te premda su rasponi vrijednosti nekad bili takvi da je za značajke snažnije kategorije utjecaja procijenjena veća vrijednost penalizacijskog koeficijenta nego obratno, medijalne vrijednosti imaju očekivani, intuitivni poredak. Osim toga, profili penalizacijskih koeficijenata pokazuju da su češće procijenjeni manji penalizacijski koeficijenti značajkama snažnijeg utjecaja i obratno.

### 6.1.3. Scenarij 2

Distribucije procjena penalizacijskih koeficijenata za Scenarij 2 prikazane su na slikama 5.23 (str. 95) i 5.24 (str. 95) te u tablici 5.14 (str. 94). Premda je bilo stršila, vidljiv je pomak u

distribucijama procjene penalizacijskih koeficijenata  $\gamma_1$  i  $\gamma_4$  u smislu da su vrijednosti penalizacijskog koeficijenta  $\gamma_1$  veće od vrijednosti penalizacijskog koeficijenta  $\gamma_4$ , što je i intuitivno. Osim toga, profili penalizacijskih koeficijenata svjedoče isto.

#### 6.1.4. Scenarij 3

Distribucije procjena penalizacijskih koeficijenata za Scenarij 3 prikazane na slikama 5.39 (str. 118) i 5.40 (str. 118) te u tablici 5.17 (str. 117). Kao i kod Scenarija 2, premda je bilo stršila, vidljiv je još izraženiji pomak u distribucijama procjene penalizacijskih koeficijenata  $\gamma_1$  i  $\gamma_4$  u smislu da su vrijednosti penalizacijskog koeficijenta  $\gamma_1$  veće od vrijednosti penalizacijskog koeficijenta  $\gamma_4$ , što je i intuitivno. Profili penalizacijskih koeficijenata svjedoče isto.

Općenito, za sve scenarije, ali i za stvarne podatke, razlozi tome što za neke skupove podataka odnosi veličina procjena penalizacijskih koeficijenata za značajke različite kategorije utjecaja nisu očekivani, odnosno intuitivni, dijelom se mogu pripisati korištenju metaheuristike pri optimizaciji funkcije cilja. Naime, metaheuristike ne jamče pronalazak optimalnog rješenja jer ne pretražuju cjelokupni prostor rješenja, no većinom pronalaze dovoljno dobra, zadovoljavajuća rješenja u razumnom vremenu izvršavanja i imaju relativno nisku računsku složenost. Možda bi povećanje broja iteracija doprinijelo stabilizaciji profila penalizacijskih koeficijenata, što je jedna od mogućnosti za daljnja istraživanja, kao i provjera performansi drugih metaheuristika.

## 6.2. PROCJENA KOEFICIJENATA MODELA

Budući da je za modele na simuliranim skupovima podataka moguće usporediti procjene koeficijenata modela s njihovim pravim vrijednostima, u nastavku su za svaki scenarij raspravljani rezultati procjene koeficijenata modela. Ključni elementi usporedbe su varijabilnost i pristranost. Naime, poželjno je da distribucija procjena koeficijenata bude što manje varijabilna te da procjena bude što je više moguće nepristrana. Informaciju o varijabilnosti procjene može dati standardna devijacija procjene koeficijenata modela prikazana u deskriptivnoj analizi u tablicama 5.12 (str. 74), 5.15 (str. 96) i 5.18 (str. 119). Nadalje, što je očekivana vrijednost procjene bliža pravoj vrijednosti koeficijenta, pristranost je manja (procjena je nepristrana ukoliko je očekivana vrijednost procjene jednaka pravoj vrijednosti parametra koji se procjenjuje). Informaciju o pristranosti može dati aritmetička sredina prikazana u istim tablicama kao procjena

očekivanja. Osim toga, dodatnu informaciju o procjenama koeficijenata daje i udaljenost koeficijenata kao sumarna mjera budući da predstavlja zbroj kvadratnih odstupanja procjena svih koeficijenata modela od prave vrijednosti koeficijenata, a koja se promatrala kao jedna mjera evaluacije modela. Rezultati udaljenosti koeficijenata analizirani su metodom analize varijanci i prikazani na slikama 5.12 (str. 79), 5.28 (str. 100) i 5.45 (str. 124).

### 6.2.1. Scenarij 1

Distribucije procjena koeficijenata modela za Scenarij 1 prikazane su na slikama 5.8, 5.9, 5.10 i 5.11 (str. 76 - 77) i u tablici 5.12 (str. 74).

Radi preglednosti, sinteza rezultata analize varijabilnosti i pristranosti koeficijenata temeljem deskriptivnih statistika (standardna devijacija i aritmetička sredina) prikazana je tablično u nastavku.

Tablica 6.1: Analiza varijabilnosti i pristranosti procjena koeficijenata modela iz Scenarija 1 - poredak modela različitih načina procjene penalizacijskog koeficijenta od boljeg prema lošijem u smislu varijabilnosti i pristranosti

Koeficijent	Varijabilnost			Pristranost		
$\beta_0$	-/det		mh	mh		-/det
$\beta_1$	det	-	mh	mh	det	-
$\beta_2$	det	-	mh	det	mh	-
$\beta_5$	-	det	mh	mh	-	det

Oznaka / podrazumijeva jednake vrijednosti.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Iz tablice je, primjerice, vidljivo kako je procjena koeficijenta  $\beta_0$  kod modela koji domensko znanje stručnjaka uključuju metaheuristikom imala veću varijabilnost u odnosu na modele koji ne koriste domensko znanje stručnjaka i modele koji domensko znanje stručnjaka uključuju deterministički. No, ujedno je za modele koji domensko znanje stručnjaka uključuju metaheuristikom procjena bila manje pristrana. Modeli bez znanja stručnjaka ili s deterministički uvedenim znanjem stručnjaka ponašaju se isto u smislu pristranosti i varijabilnosti. Tumačenja

za procjene ostalih koeficijenata su analogna.

Nadalje, premda se nije radilo o velikim razlikama (vidjeti u tablici 5.12 na str. 74), iz tablice 6.1 vidljivo je da uvođenje domenskog znanja stručnjaka metaheuristikom poboljšava procjene koeficijenata modela jer se smanjuje pristranost, dok deterministički pristup pruža stabilnije procjene u smislu varijabilnosti.

Rezultati analize varijanci i *post hoc* analiza Tukeyjevim testom HSD također upućuju da je za svaku metodu regularizacije uvođenje domenskog znanja metaheuristikom rezultiralo statistički značajnom negativnom razlikom u prosječnoj udaljenosti koeficijenata u odnosu na modele bez domenskog znanja stručnjaka, dok je kod uvođenja domenskog znanja stručnjaka deterministički pristupom ta razlika bila pozitivna, ali ne i statistički značajna.

### 6.2.2. Scenarij 2

Distribucije procjena koeficijenata modela za Scenarij 2 prikazane su na slikama 5.25, 5.26 i 5.27 (str. 97 - 98) i u tablici 5.15 (str. 96).

Radi preglednosti, sinteza rezultata analize varijabilnosti i pristranosti koeficijenata temeljem deskriptivnih statistika (standardna devijacija i aritmetička sredina) prikazana je tablično u nastavku.

Tablica 6.2: Analiza varijabilnosti i pristranosti procjena koeficijenata modela iz Scenarija 2 - poredak modela različitih načina procjene penalizacijskog koeficijenta od boljeg prema lošijem u smislu varijabilnosti i pristranosti

Koeficijent	Varijabilnost	Pristranost
$\beta_0$	det - mh	- det mh
$\beta_{11}$	det - mh	mh det -
$\beta_{31}$	det - mh	mh det -

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Iz tablice je, primjerice, vidljivo kako je varijabilnost procjena koeficijenata  $\beta_{11}$  i  $\beta_{31}$  najveća kod modela koji domensko znanje stručnjaka uključuju metaheuristikom, zatim kod modela koji ne koriste domensko znanje stručnjaka, a najmanja je kod modela koji domensko zna-

nje stručnjaka uključuju deterministički. No, ujedno je vidljivo kako je pristranost kod modela koji domensko znanje stručnjaka uključuju metaheuristikom bila najmanja, pa zatim slijedi ona kod modela koji domensko znanje stručnjaka uključuju deterministički. Konačno, pristranost je najveća kod modela koji ne koriste domensko znanje stručnjaka. Tumačenje za procjenu  $\beta_0$  je analogno.

Nadalje, iz tablice 6.2 vidljivo je kako uvođenje domenskog znanja stručnjaka determinističkim pristupom poboljšava procjene koeficijenata modela jer se za gotovo sve koeficijente smanjuju i varijabilnost i pristranost u odnosu na modele bez znanja stručnjaka. S druge strane, modeli koje domensko znanje stručnjaka uključuju metaheuristikom su imali najveću varijabilnost, ali najčešće i najmanju pristranost (vidjeti u tablici 5.15 na str. 96).

Međutim, rezultati analize varijanci i *post hoc* analiza Tukeyjevim testom HSD upućuju da je za svaku metodu regularizacije uvođenje domenskog znanja deterministički rezultiralo statistički značajnom negativnom razlikom u prosječnoj udaljenosti koeficijenata u odnosu na modele bez domenskog znanja stručnjaka, dok kod uvođenja domenskog znanja stručnjaka metaheuristikom je ta razlika bila pozitivna, ali ne i statistički značajna.

Kod ovog scenarija se ističe i lasso regularizacija kao metoda s najvećom varijabilnosti.

### 6.2.3. Scenarij 3

Distribucije procjena koeficijenata modela za Scenarij 3 prikazane su na slikama 5.41, 5.42, 5.43 i 5.44 (str. 121 - 122) i u tablici 5.18 (str. 119).

Radi preglednosti, sinteza rezultata analize varijabilnosti i pristranosti koeficijenata temeljem deskriptivnih statistika (standardna devijacija i aritmetička sredina) prikazana je tablicom 6.3.

Iz tablice je, primjerice, vidljivo kako je procjena koeficijenta  $\beta_6$  i  $\beta_{11}$  kod modela koji domensko znanje stručnjaka uključuju metaheuristikom imala najveću varijabilnost i pristranost, dok su modeli koji domensko znanje stručnjaka uključuju metaheuristikom imali najmanju varijabilnost i pristranost. Tumačenja su za procjene ostalih koeficijenata analogna.



Tablica 6.3: Analiza varijabilnosti i pristranosti procjena koeficijenata modela iz Scenarija 3 - poredak modela različitih načina procjene penalizacijskog koeficijenta od boljeg prema lošijem u smislu varijabilnosti i pristranosti

Koeficijent	Varijabilnost	Pristranost
$\beta_0$	det - mh	-/det mh
$\beta_1$	- det mh	mh det -
$\beta_6$	det - mh	det - mh
$\beta_{11}$	det - mh	det - mh

Oznaka / podrazumijeva jednake vrijednosti.

Procjena penalizacijskog koeficijenta (- : nije rađena, det: deterministički, mh: metaheuristika).

Za ovaj scenarij djeluje da uvođenje domenskog znanja determinističkim putem poboljšava pristranost i varijabilnost procjena u odnosu na modele koji ne koriste domensko znanje stručnjaka, međutim rezultati analize varijanci i *post hoc* analiza Tukeyjevim testom HSD upućuju da je razlika prosječnih udaljenosti koeficijenata između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna. S druge strane, naslućuje se da domenskog znanja metaheuristikom ne poboljšava procjenu koeficijenata niti u jednom aspektu. Tome u prilog idu rezultati analize varijanci i *post hoc* analiza Tukeyjevim testom HSD prema kojima je razlika prosječnih udaljenosti koeficijenata između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka pozitivna i statistički značajna.

Kod ovog scenarija ističe se i lasso regularizacija kao metoda s najvećom varijabilnosti.

Konačno, analizom rezultata procjene koeficijenata modela između ova tri scenarija slijedi:

- Scenarij 1 - uvođenje domenskog znanja stručnjaka metaheuristikom doprinosi smanjenju pristranosti procjena koeficijenata modela,
- Scenarij 2 - uvođenje domenskog znanja stručnjaka deterministički doprinosi smanjenju pristranosti i varijabilnosti procjena koeficijenata modela,
- Scenarij 3 - uvođenje domenskog znanja stručnjaka ne doprinosi smanjenju pristranosti ili varijabilnosti procjena koeficijenata modela.

Razlog za razlike u (ne)doprinosu domenskog znanja procjenama koeficijenata vjerojatno leži u strukturi svakog scenarija.

- Scenarij 1 - sastoji se od osam prediktora od čega njih troje ulazi u model (tj. po postavkama scenarija imaju koeficijent različit od nule), standardna devijacija greške je 3, a korelacije između prediktora su 0.5 ili manje (točnije  $corr(X_i, X_j) = 0.5^{|i-j|}$ ), prema "mišljenju stručnjaka" bilo je 5 prediktora kategorije *Nema utjecaja*, 1 kategorije *Srednji utjecaj* i 2 kategorije *Jak utjecaj*. U ovom scenariju mali broj prediktora i niske korelacije te relativno niska greška modela omogućavaju metaheuristikama smanjenje pristranosti.
- Scenarij 2 - sastoji se od 40 prediktora od čega njih 20 ulazi u model, standardna devijacija greške je 15, a korelacije između prediktora su 0.5, prema "mišljenju stručnjaka" bilo je 20 prediktora kategorije *Nema utjecaja* i 20 kategorije *Jak utjecaj*. U ovom scenariju broj je prediktora veći, kao i korelacije među njima. Unatoč tome i unatoč većoj greški modela, jasna podjela prediktora na one bez utjecaja i s jakim utjecajem i deterministički pristup uvođenja takvog znanja u modele pomaže stabilnosti procjena koeficijenata modela jer smanjuje i varijabilnost i pristranost procjena u odnosu na procjene modela bez domenskog znanja stručnjaka.
- Scenarij 3 - sastoji se od 40 prediktora od čega njih 15 ulazi u model, standardna devijacija greške je 15, prediktori su grupirani u 3 grupe od 5 prediktora, prema "mišljenju stručnjaka" bilo je 25 prediktora kategorije *Nema utjecaja* i 15 kategorije *Jak utjecaj*. Uvođenje domenskog znanja stručnjaka ne doprinosi značajno procjenama koeficijenata modela, moguće zbog visoke korelacija među prediktorima i postojanja grupa prediktora, što povećava složenost modela.

Na temelju prethodnih rezultata simulacija na 50 skupova iz svakog scenarija, slijedi da je metaheuristički pristup uvođenja domenskog znanja stručnjaka najviše doprinosio skupovima s manjim brojem prediktora i niskim korelacijama. Deterministički pristup uvođenja domenskog znanja stručnjaka bio je bolji za skupove s većim brojem prediktora koji nisu grupirani i umjerenom su korelirani, kao što su skupovi iz Scenarija 2. Skupovi iz Scenarija 3 imaju jednak broj prediktora i grešku modela kao oni iz Scenarija 2, ali su kompleksniji zbog postojanja grupa prediktora.

Dakle, uvođenje domenskog znanja ima veći doprinos u jednostavnijim modelima s manje prediktora, nižim korelacijama i manjim greškama kao što je Scenarij 1. U složenijim mo-

delima, doprinos domenskog znanja postaje manje značajan zbog povećane kompleksnosti i grešaka u modelu. Međutim, treba biti oprezan i ne generalizirati ove rezultate na skupove podataka drugačijih karakteristika. Svakako, ovi rezultati pokazuju kako različiti načini procjene penalizacijskih koeficijenta i sama struktura modela utječu na varijabilnost i pristranost modela.

## 6.3. HIPOTEZA 1

**Hipoteza 1:** Predložena metoda rezultira uspješnijom predikcijom od metode regularizacije koja ne uključuje domensko znanje stručnjaka.

U ovoj hipotezi uspoređuju se metode regularizacije koje ne uključuju domensko znanje stručnjaka s onima koje ga uključuju deterministički i metaheuristički. Mjere evaluacije su korigirani koeficijent determinacije te *RMSE*, *MAE* i *SMAPE* predikcije.

Za simulirane podatke dodatno se kao mjera evaluacije koristi i euklidska udaljenost koeficijenata. Za evaluaciju hipoteze kod simuliranih podataka koristila se dvofaktorska ANOVA s faktorima Procjena penalizacijskog koeficijenta i Metoda regularizacije te njihovom interakcijom. Pomoću kontrasta c11 testirala se upravo Hipoteza 1 ovog istraživanja, dok kontrast c12 uspoređuje međusobno dva načina uvođenja domenskog znanja stručnjaka u metode regularizacije.

### 6.3.1. Stvarni podaci

Za ovaj skup podataka korištene su dvije vrste domenskog znanja; Domensko znanje 1 dobiveno kao medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, te Domensko znanje 2 koje predstavlja mišljenje Stručnjaka 3. Svaka vrsta domenskog znanja zasebno se uvrštavala u regularizaciju. Vrijednosti mjera evaluacije prikazane su u tablici 5.9 (str. 64).

Na slici 5.2 (str. 65) vidljivo je da je za gotovo svaku metodu regularizacije uvođenje domenskog znanja stručnjaka rezultiralo povećanjem korigiranog koeficijenta determinacije. Izuzetak je uvođenje Domenskog znanja 2 determinističkim pristupom kada se korigirani koeficijent determinacije spustio malo ispod razine kada se ne koristi domensko znanje stručnjaka za hrbatnu regularizaciju, dok je za lasso regularizaciju ostao na jednakoj razini.

Na slici 5.3 (str. 66) vidljivo je da je za svaku metodu regularizacije uvođenje domenskog znanja stručnjaka rezultiralo smanjenjem *RMSE*.

Na slici 5.4 (str. 67) vidljivo je da je uvođenje domenskog znanja stručnjaka u metode regu-

larizacije u gotovo svim slučajevima rezultiralo nižom *MAE*. Izuzetak je kod hrbatne regularizacije i regularizacije elastičnom mrežom koje Domensko znanje 1 uključuju deterministički.

Na slici 5.5 (str. 68) vidljivo je da je za sve metode regularizacije uvođenje domenskog znanja stručnjaka determinističkim pristupom rezultiralo povećanjem *SMAPE*, dok je uvođenje metaheuristikom rezultiralo smanjenjem *SMAPE*.

Budući da se analiza rezultata Hipoteze 1 razlikuje prema načinu procjene penalizacijskih koeficijenata, metodi regularizacije i korištenom domenskom znanju, radi lakše preglednosti, sljedeća tablica predstavlja njihovu sintezu.

Tablica 6.4: Sinteza rezultata analize Hipoteze 1 za svaku mjeru evaluacije, metodu regularizacije i način procjene penalizacijskih koeficijenata na skupu stvarnih podataka (+: rezultati podupiru Hipotezu 1, -: rezultati ne podupiru Hipotezu 1, /: nema promjene u rezultatu).

	Mjera evaluacije	det			mh		
		El. mreža	Lasso	Hrbatna	El. mreža	Lasso	Hrbatna
<b>Domensko znanje 1</b>	$\tilde{R}^2$	+	+	+	+	+	+
	<i>RMSE</i>	+	+	+	+	+	+
	<i>MAE</i>	-	+	-	+	+	+
	<i>SMAPE</i>	-	-	-	+	+	+
<b>Domensko znanje 2</b>	$\tilde{R}^2$	+	/	-	+	+	+
	<i>RMSE</i>	+	+	+	+	+	+
	<i>MAE</i>	+	+	+	+	+	+
	<i>SMAPE</i>	-	-	-	+	+	+

Procjena penalizacijskog koeficijenta (det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Za metode regularizacije koje domensko znanje stručnjaka uključuju deterministički, rezultati nekih mjera evaluacije modela podupiru Hipotezu 1, dok ju kod uvođenja domenskog znanja stručnjaka metaheuristikom podupiru u svim mjerama evaluacije. Ta poboljšanja u uspješnosti predikcije u odnosu na modele koji ne koriste domenska znanja stručnjaka su bila znatna, kako je i vidljivo u tablici 5.9 (str. 64).

Može se reći da rezultati na stvarnim podacima djelomično podupiru Hipotezu 1 za metode regularizacije koje domensko znanje stručnjaka uključuju deterministički (u slučajevima navedenim u tablici 6.4). Nadalje, rezultati na stvarnim podacima podupiru Hipotezu 1 za metode regularizacije koje domensko znanje stručnjaka uključuju metaheuristikom.

### Usporedba determinističke procjene penalizacijskih koeficijenata i procjene metaheuristikom

Nadalje, bilo bi zanimljivo usporediti međusobno različite načine procjene penalizacijskih koeficijenata - deterministički i metaheuristikom.

Za korigirani koeficijent determinacije može se uočiti da je uvođenje domenskog znanja stručnjaka metaheuristikom rezultiralo većim povećanjem korigiranog koeficijenta determinacije nego uvođenje domenskog znanja stručnjaka determinističkim pristupom. Za *RMSE* grešku je smanjenje greške uvođenjem domenskog znanja bilo veće kod procjene penalizacijskih koeficijenata metaheuristikom nego li kod determinističke procjene. Smanjenje *MAE* također je bilo veće kod procjene penalizacijskih koeficijenata metaheuristikom nego li kod determinističke procjene. Izuzetak je regularizacija elastičnom mrežom i hrbatna regularizacija s deterministički uvedenim Domenskim znanjem 1. Konačno, uvođenje domenskog znanja stručnjaka determinističkim pristupom rezultiralo je povećanjem *SMAPE*, dok je uvođenje metaheuristikom rezultiralo smanjenjem *SMAPE*.

Može se zaključiti da je u slučaju stvarnih podataka procjena penalizacijskih koeficijenata metaheuristikom češće rezultirala boljim izvedbama modela nego li deterministička procjena, kako je i vidljivo u tablici 6.4 (str. 146).

### Usporedba domenskih znanja

Nadalje, bilo bi zanimljivo usporediti međusobno dvije korištene vrste domenskog znanja.

Prilikom uvođenja Domenskog znanja 1 u modele regularizacije, vrijednost korigiranog koeficijenta determinacije je više porasla nego li prilikom uvođenja Domenskog znanja 2. *RMSE* je u gotovo svim slučajevima bila niža pri korištenju Domenskog znanja 1. Izuzetak je jedino regularizacija elastičnom mrežom i deterministička procjena penalizacijskih koeficijenata za koju je bilo bolje Domensko znanje 2. Kod *MAE* je za regularizaciju elastičnom mrežom i hrbatnu regularizaciju kod determinističkog pristupa bolje Domensko znanje 2, a kod metaheuristike

Domensko znanje 1. Kod lasso regularizacije vrijedi obratno. Za *SMAPE* je za regularizaciju elastičnom mrežom i hrbatnu regularizaciju kod determinističkog pristupa bolje Domensko znanje 2, a za lasso regularizacije Domensko znanje 1. Kod metaheurističkog pristupa, za hrbatnu regularizaciju nema razlike u iznosu *SMAPE* greški između Domenskog znanja 1 i Domenskog znanja 2, dok je za lasso regularizaciju i regularizaciju elastično mrežom bolje Domensko znanje 1.

Premda je malo češće razlika u mjerama evaluacije išla u korist Domenskog znanja 1, teško je zaključiti koje je domensko znanje točnije.

### 6.3.2. Scenarij 1

U nastavku za svaku mjeru evaluacije slijedi rasprava rezultata dvofaktorske analize varijance s faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije.

Prosječne vrijednosti udaljenosti koeficijenata za skupove iz Scenarija 1 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije vidljive su na slici 5.12 (str. 79). Vidljivo je da uvođenje domenskog znanja determinističkim pristupom rezultira većim prosječnim udaljenostima koeficijenata nego kad se domensko znanje stručnjaka ne koristi, dok se prosječna udaljenost koeficijenata smanjuje kod uvođenja domenskog znanja stručnjaka metaheuristikom. Rezultati analize varijance i *post hoc* analize Tukeyjevim testom HSD pokazuju da je za svaku metodu regularizacije uvođenje domenskog znanja metaheuristikom rezultiralo statistički značajnom negativnom razlikom u prosječnoj udaljenosti koeficijenata u odnosu na modele bez domenskog znanja stručnjaka, dok je kod uvođenja domenskog znanja stručnjaka deterministički pristupom ta razlika bila pozitivna, ali ne i statistički značajna. Dodatno, za svaku metodu regularizacije uvođenje domenskog znanja metaheuristikom rezultiralo je statistički značajnom negativnom razlikom u prosječnoj udaljenosti koeficijenata u odnosu na modele koji domensko znanja stručnjaka uvode deterministički. Može se reći da za udaljenost koeficijenata Hipoteza 1 nije potvrđena za metode regularizacije koje domensko znanje stručnjaka uključuju deterministički, dok je potvrđena za one koje ga uključuju metaheuristikom.

Na slici 5.13 (str. 80) prikazane su prosječne vrijednosti korigiranog koeficijenta determinacije za skupove iz Scenarija 1 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je kako uvođenjem domenskog znanja na deterministički način opada prosječna vrijednost korigiranog koeficijenta determinacije, a onda ponovno raste kada se domensko znanje uvede metaheuristikom, no za lasso regularizaciju i regularizaciju elastičnom

mrežom je to ispod razine slučaja kada se ne koristi domensko znanje stručnjaka, a za hrbatnu regularizaciju iznad. Rezultati analize varijance pokazuju kako efekt uvođenja domenskog znanja stručnjaka u metode regularizacije nije statistički značajan. Može se reći da za korigirani koeficijent determinacije Hipoteza 1 nije potvrđena.

Na slici 5.14 str. (81) prikazane su prosječne vrijednosti *RMSE* za skupove iz Scenarija 1 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je kako uvođenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. Izuzetak je jedino uvođenje domenskog znanja determinističkim pristupom u hrbatnu regularizaciju kada prosječna *RMSE* poraste. Rezultati analize varijance i *post hoc* analize Tukeyjevim testom HSD upućuje da je razlika prosječne *RMSE* između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna, dok je razlika prosječne *RMSE* između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka je pozitivna, ali nije statistički značajna. Nadalje, razlika prosječne *RMSE* između metoda regularizacije koje penalizacijske koeficijente procjenjuju deterministički s onima koje ih procjenjuju metaheuristikom je negativna i statistički značajna. Može se reći da je za *RMSE* Hipoteza 1 potvrđena za metode regularizacije koje domensko znanje stručnjaka uključuju metaheuristikom, dok nije potvrđena za one koje ga uključuju deterministički.

Na slici 5.15 (str. 83) prikazane su prosječne vrijednosti *MAE* za skupove iz Scenarija 1 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Kao i kod *RMSE*, vidljivo je kako uvođenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. Izuzetak je jedino uvođenje domenskog znanja determinističkim pristupom u hrbatnu regularizaciju kada prosječna *MAE* poraste. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju da je razlika prosječne *MAE* između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna, dok je razlika prosječne *MAE* između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka je pozitivna, ali nije statistički značajna. Razlika u iznosu prosječne *MAE* između metaheurističkog i determinističkog pristupa negativna je i statistički značajna. Može se reći da je za *MAE* Hipoteza 1 potvrđena za metode regularizacije koje domensko znanje stručnjaka uključuju metaheuristikom, dok nije potvrđena za one koje ga uključuju deterministički.

Na slici 5.16 (str. 84) prikazane su prosječne vrijednosti *SMAPE* za skupove iz Scena-

rija 1 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je kako uvođenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. Izuzetak je jedino uvođenje domenskog znanja determinističkim pristupom u hrbatnu regularizaciju kada prosječna *SMAPE* poraste. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju da je razlika između prosječne *SMAPE* determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna, dok je razlika prosječnih *SMAPE* greški između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna. Razlika u iznosu prosječne *SMAPE* između metaheurističkog i determinističkog pristupa je negativna i statistički značajna. Može se reći da je za *SMAPE* Hipoteza 1 potvrđena za metode regularizacije koje domensko znanje stručnjaka uključuju metaheuristikom, dok nije potvrđena za one koje ga uključuju deterministički.

### 6.3.3. Scenarij 2

U nastavku za svaku mjeru evaluacije slijedi rasprava rezultata dvofaktorske analize varijance s faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije.

Na slici 5.28 (str. 100) prikazane su prosječne vrijednosti udaljenosti koeficijenata za skupove iz Scenarija 2 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Na slici vidljivo je da uvođenje domenskog znanja determinističkim pristupom rezultira manjim prosječnim udaljenostima koeficijenata nego kad se domensko znanje stručnjaka ne koristi, dok se prosječna udaljenost koeficijenata kod uvođenja domenskog znanja stručnjaka metaheuristikom povećava - za hrbatnu regularizaciju i regularizaciju elastičnom mrežom raste na razinu ispod razine slučaja kada se ne koristi domensko znanje stručnjaka, dok za lasso regularizaciju raste na višu razinu). Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju da je za svaku metodu regularizacije uvođenje domenskog znanja deterministički rezultiralo statistički značajnom negativnom razlikom u prosječnoj udaljenosti koeficijenata u odnosu na modele bez domenskog znanja stručnjaka, dok kod uvođenja domenskog znanja stručnjaka metaheuristikom je ta razlika bila pozitivna, ali ne i statistički značajna. Dodatno, za svaku metodu regularizacije uvođenje domenskog znanja metaheuristikom rezultiralo je statistički značajnom pozitivnom razlikom u prosječnoj udaljenosti koeficijenata u odnosu na modele koji domensko znanje stručnjaka uvode deterministički. Može se reći da je za udaljenost



koeficijenta Hipoteza 1 potvrđena za metode regularizacije koje domensko znanje stručnjaka uključuju deterministički, dok nije potvrđena za one koje ga uključuju metaheuristikom.

Na slici 5.29 (str. 102) prikazane su prosječne vrijednosti korigiranog koeficijenta determinacije za skupove iz Scenarija 2 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je kako za sve metode regularizacije uvođenjem domenskog znanja na deterministički način raste prosječna vrijednost korigiranog koeficijenta determinacije, a onda ona pada kada se domensko znanje uvede metaheuristikom. Za regularizaciju elastičnom mrežom i hrbatnu regularizaciju dođe na razinu i dalje višu nego prije uvođenja domenskog znanja stručnjaka, dok za lasso regularizaciju padne na nižu razinu od početne. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju da je za hrbatnu regularizaciju i regularizaciju elastičnom mrežom uvođenje domenskog znanja stručnjaka determinističkim pristupom rezultiralo statistički značajnim povećanjem prosječnog korigiranog koeficijenta determinacije, dok za lasso regularizaciju to nije slučaj. Uvođenje domenskog znanja stručnjaka metaheuristikom nije doprinijelo statistički značajnoj razlici prosječnih korigiranih koeficijenta determinacije u odnosu na modele bez znanja stručnjaka niti za jednu regularizaciju. Može se reći da se za korigirani koeficijent determinacije Hipoteza 1 djelomično potvrđuje za metode regularizacije koje domensko znanje stručnjaka uključuju deterministički (odnosno potvrđuje se za hrbatnu regularizaciju i regularizaciju elastičnom mrežom, a za lasso regularizaciju ne), dok se za one koji ga uključuju metaheuristikom ne potvrđuje.

Na slici 5.30 (str. 103) prikazane su prosječne vrijednosti *RMSE* za skupove iz Scenarija 2 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je uvođenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuje da je razlika između prosječnih *RMSE* determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna, dok je razlika između prosječnih *RMSE* metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna. Može se reći da za *RMSE* grešku Hipoteza 1 potvrđuje za metode regularizacije koje domensko znanje stručnjaka uključuju metaheuristikom, dok se ne potvrđuje za one koje ga uključuju deterministički.

Na slici 5.31 (str. 104) prikazane su prosječne vrijednosti *MAE* za skupove iz Scenarija 2 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je uvo-

đenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. No, rezultati analize varijance pokazuju da nije statistički značajan niti jedan pojedinačni niti interakcijski efekt. Može se reći da se za *MAE* Hipoteza 1 ne potvrđuje.

Na slici 5.32 (str. 105) prikazane su prosječne vrijednosti *SMAPE* za skupove iz Scenarija 2 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je kako uvođenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju da je razlika prosječnih *SMAPE* greški između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna, dok je razlika prosječnih *SMAPE* greški između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna. Osim toga, razlika prosječnih *SMAPE* greški između metaheurističkog i determinističkog pristupa je negativna i statistički značajna. Može se reći da se za *SMAPE* Hipoteza 1 potvrđuje za metode regularizacije koje domensko znanje stručnjaka uključuju metaheuristikom, dok se ne potvrđuje za one koje ga uključuju deterministički.

#### 6.3.4. Scenarij 3

U nastavku za svaku mjeru evaluacije slijedi rasprava rezultata dvofaktorske analize varijance s faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije.

Na slici 5.45 (str. 124) prikazane su prosječne vrijednosti udaljenosti koeficijenata za skupove iz Scenarija 3 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Na slici je vidljivo kako za lasso regularizaciju prosječna vrijednost udaljenosti koeficijenta poraste uvođenjem znanja stručnjaka determinističkim pristupom, a dodatno poraste kada se isto uvede metaheurističkim pristupom. Za hrbatnu regularizaciju i regularizaciju elastičnom mrežom uvođenjem domenskog znanja na deterministički način opada prosječna vrijednost udaljenosti koeficijenata, a onda ona raste kada se domensko znanje uvede metaheuristikom na razinu iznad početne. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju da je razlika prosječnih udaljenosti koeficijenata između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna, dok su razlike prosječnih udaljenosti koeficijenata između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka te između metaheurističkog i determinističkog pristupa pozitivne i statistički

značajne. Osim toga, razlika u prosječnim udaljenosti koeficijenata između lasso regularizacije i regularizacije elastičnom mrežom pozitivna je i statistički značajna, dok su razlike između hrbatne regularizacije i regularizacije elastičnom mrežom te hrbatne i lasso regularizacije negativne i statistički značajne. Može se reći da se za udaljenost koeficijenata Hipoteza 1 ne potvrđuje.

Na slici 5.46 (str. 125) prikazane su prosječne vrijednosti korigiranog koeficijenta determinacije za skupove iz Scenarija 3 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju kako je za hrbatnu regularizaciju razlika u prosječnom korigiranom koeficijentu determinacije između između determinističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka pozitivna i statistički značajna. S druge strane, razlika u prosječnom korigiranom koeficijentu determinacije između metaheurističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka je pozitivna i nije statistički značajna, dok je razlika u prosječnom korigiranom koeficijentu determinacije između metaheurističkog i determinističkog pristupa negativna i statistički značajna. Nadalje, za lasso regularizaciju i regularizaciju elastičnom mrežom vidljivo je kako razlika u prosječnom korigiranom koeficijentu determinacije između između determinističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka nije statistički značajna. S druge strane, razlike u prosječnom korigiranom koeficijentu determinacije između metaheurističkog pristupa i pristupa kada se ne koristi domensko znanje stručnjaka te između metaheurističkog i determinističkog pristupa su negativne i statistički značajne. Može se reći da se za korigirani koeficijent determinacije Hipoteza 1 potvrđuje za hrbatnu metodu regularizacije koja domensko znanje stručnjaka uključuju deterministički, dok se u svim ostalim slučajevima ne potvrđuje.

Na slici 5.47 (str. 127) prikazane su prosječne vrijednosti *RMSE* za skupove iz Scenarija 3 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je uvođenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju da je razlika prosječnih *RMSE* između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna, dok je razlika prosječnih *RMSE* između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna. Može se reći da se za *RMSE* grešku Hipoteza 1 potvrđuje za metode regularizacije koje

domensko znanje stručnjaka uključuju metaheuristikom, dok se ne potvrđuje za one koje ga uključuju deterministički.

Na slici 5.48 (str. 128) prikazane su prosječne vrijednosti *MAE* za skupove iz Scenarija 3 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je kako uvođenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. Rezultati analize varijance pokazuju da nisu statistički značajni pojedinačni niti interakcijski efekti. Može se reći da se za *MAE* Hipoteza 1 ne potvrđuje.

Na slici 5.49 (str. 129) prikazane su prosječne vrijednosti *SMAPE* za skupove iz Scenarija 3 prema faktorima Procjena penalizacijskih koeficijenata i Metoda regularizacije. Vidljivo je kako uvođenjem domenskog znanja stručnjaka opada prosječna vrijednost greške, pri čemu više opadne kod metaheurističkog pristupa uvođenja domenskog znanja stručnjaka nego li kod determinističkog. Rezultati analize varijance i *post hoc* analiza Tukeyjevim testom HSD upućuju da je razlika prosječnih *SMAPE* greški između determinističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna, ali nije statistički značajna, dok je razlika prosječnih *SMAPE* greški između metaheurističkog pristupa i pristupa koji ne koristi znanje stručnjaka negativna i statistički značajna. Osim toga, razlika prosječnih *SMAPE* greški između determinističkog i metaheurističkog pristupa negativna je i statistički značajna. Može se reći da se za *SMAPE* Hipoteza 1 potvrđuje za metode regularizacije koje domensko znanje stručnjaka uključuju metaheuristikom, dok se ne potvrđuje za one koje ga uključuju deterministički.

Konačno, budući da se odluka o Hipotezi 1 razlikuje prema scenariju, mjeri evaluacije i načinu procjene penalizacijskih koeficijenata, radi lakše preglednosti, tablica 6.5 predstavlja sintezu rezultata analize Hipoteze 1 za svaki scenarij.

Vidljivo je kako na simuliranim podacima uvođenje domenskog znanja stručnjaka determinističkim pristupom većinski nije doprinijelo poboljšanju uspješnosti predikcije modela, dok domenskog znanja stručnjaka metaheuristikom većinski doprinijelo je poboljšanju uspješnosti predikcije modela.

**Konačno, može se reći da na podacima simuliranim iz Scenarija 1 Hipoteza 1 nije potvrđena za metode regularizacije koje domensko znanje uključuju deterministički, te da je djelomično potvrđena za metode regularizacije koje domensko znanje uključuju metaheuristikom (u slučajevima navedenim u tablici 6.5). Nadalje, može se reći da se na podacima simuliranim iz Scenarija 2 i Scenarija 3 Hipoteza 1 djelomično potvrđena (u**

slučajevima navedenim u tablici 6.5).

Tablica 6.5: Sinteza rezultata analize Hipoteze 1 za svaku mjeru evaluacije i način procjene penalizacijskih koeficijenata za svaki scenarij (+: Hipoteza 1 je potvrđena, -: Hipoteza 1 nije potvrđena).

Procjena pen. koef.	Mjera evaluacije	Scenarij 1	Scenarij 2	Scenarij 3
det	UK	-	+	-
	$\tilde{R}^2$	-	+ za hrbatnu i el. mrežu	+ za hrbatnu
	<i>RMSE</i>	-	-	-
	<i>MAE</i>	-	-	-
	<i>SMAPE</i>	-	-	-
mh	UK	+	-	-
	$\tilde{R}^2$	-	-	-
	<i>RMSE</i>	+	+	+
	<i>MAE</i>	+	-	-
	<i>SMAPE</i>	+	+	+

Kratica UK označava udaljenost koeficijenata. Procjena penalizacijskog koeficijenta (det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Kao i kod analize procjene koeficijenata modela između različitih načina uvođenja domenskog znanja stručnjaka, i ovdje se pretpostavlja da je razlog razlikama u (ne)doprinosu domenskog znanja poboljšanju predikcije modela vjerojatno u strukturi svakog scenarija, ali i o posebnostima pojedine mjere evaluacije. Podsjetimo na strukturu pojedinog scenarija:

- Scenarij 1 - mali broj prediktora i niske korelacije te relativno niska greška modela,
- Scenarij 2 - veći broj prediktora koji nisu grupirani, umjerene korelacije, viša greška modela,
- Scenarij 3 - veći broj prediktora, postoje grupe prediktora, viša greška modela.

Udaljenost koeficijenata kao mjera evaluacije modela i mogući razlozi u razlikovanju rezultata uvođenja domenskog znanja stručnjaka između scenarija i načina uvođenja domenskog znanja (deterministički i metaheuristički) već su razmatrani u potpoglavlju *Procjena koeficijenata modela* (str. 139). Ukratko - uvođenje domenskog znanja pokazalo je doprinos u smislu pristranosti i varijabilnosti procjena koeficijenata modela za jednostavnije modela kao što su modeli na skupovima iz Scenarija 1. U složenijim modelima, doprinos domenskog znanja bio je manje značajan zbog povećane kompleksnosti i grešaka u modelu. Kod Scenarija 3 gdje postoje grupe prediktora, doprinos domenskog znanja stručnjaka izostao, dok je kod Scenarija 2 gdje je bio veći broj prediktora koji nisu grupirani i umjereno su korelirani, uvođenje domenskog znanja doprinijelo procjeni koeficijenata.

Zanimljivo je da korigirani koeficijent determinacije u jednostavnijim modelima (modeli na skupovima iz Scenarija 1) nije bio bolji uvođenjem domenskog znanja stručnjaka. Za veće skupove i malo složenije modele (modeli na skupovima iz Scenarija 2 i 3), uvođenje domenskog znanja stručnjaka determinističkim pristupom je pojedinim metodama regularizacije doprinijelo povećanju korigiranog koeficijenta determinacije.

Greške predikcije (*RMSE*, *MAE*, *SMAPE*) nisu se smanjile uvođenjem domenskog znanja stručnjaka deterministički. S druge strane, kod metaheurističkog pristupa smanjile su se greške *RMSE* i *SMAPE*, ali ne i *MAE*. U smislu ovih grešaka, vidljivo je da je metaheuristika djelovala bolje na uspješnost predikcije modela.

Ove razlike naglašavaju važnost razmatranja strukture podataka i karakteristika prediktora pri uvođenju domenskog znanja u modele, ali i važnost korištenja različitih mjera evaluacije modela jer svaka analizira uspješnost predikcije u drugačijem aspektu.

## 6.4. HIPOTEZA 2

**Hipoteza 2:** Predložena metoda rezultira uspješnijom predikcijom od metode regresijskih stabala.

U ovoj hipotezi uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka. Mjere evaluacije su *RMSE*, *MAE* i *SMAPE*.

Za evaluaciju hipoteze kod simuliranih podataka koristi se jednofaktorska ANOVA s faktorom Metoda. Pomoću kontrasta c31 testira se Hipoteza 2 ovog istraživanja.

Hipoteza se evaluirala kroz dvije jednostavnije podhipoteze - u jednoj se uspoređuje metoda

CART s metodama regularizacije koje uključuju domensko znanje stručnjaka deterministički (radi jednostavnosti označena kao "Hipoteza 2a"), a u drugoj se metoda CART uspoređuju s metodama regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom (radi jednostavnosti označena kao "Hipoteza 2b").

#### 6.4.1. Stvarni podaci

U smislu Hipoteze 2a, na slici 5.3 (str. 66) i u tablici 5.9 (str. 64) vidljivo je kako je za obje vrste domenskog znanja metoda CART rezultirala nižom *RMSE* od regularizacija elastičnom mrežom i hrbatne regularizacije, a višom od lasso regularizacije. Nadalje, u istoj tablici i na slici 5.4 (str. 67) vidljivo je kako je metoda CART rezultirala višom *MAE* za obje vrste domenskog znanja i sve metode regularizacije, osim u slučaju regularizacije elastičnom mrežom i hrbatne regularizacije kada se koristi Domensko znanje 1. Naposljetku, metode regularizacije rezultirale nižim *SMAPE* od metode CART u svim slučajevima (slika 5.5, str. 68).

Za oba domenska znanja, rezultati analiza za mjeru evaluacije *RMSE* podupiru Hipotezu 2 za lasso regularizaciju, dok za hrbatnu regularizaciju i regularizaciju elastičnom mrežom ne podupiru ju. Rezultati analiza mjere *MAE* podupiru Hipotezu 2 za Domensko znanje 2 za sve vrste regularizacije, dok je za Domensko znanje 1 podupiru samo za lasso regularizaciju. Rezultati analiza za mjeru evaluacije *SMAPE* podupiru Hipotezu 2 za oba domenska znanja i sve vrste regularizacije (tablica 6.6).

U smislu Hipoteze 2b, vidljivo je da su metode regularizacije s domenskim znanjem stručnjaka uvedenim metaheuristikom rezultirale nižim *RMSE*, *MAE* i *SMAPE* od metode CART u svim slučajevima (tablica 5.9 (str. 64) i slike 5.3 (str. 66), 5.4 (str. 67) i 5.5 (str. 68)).

Tablica 6.6 predstavlja sintezu rezultata analize Hipoteze 2 na stvarnim podacima. **Može se reći da rezultati na stvarnim podacima djelomično podupiru Hipotezu 2 za metode regularizacije koje domensko znanje stručnjaka uključuju deterministički** (u slučajevima navedenim u tablici 6.6). **Nadalje, rezultati na stvarnim podacima podupiru Hipotezu 2 za metode regularizacije koje domensko znanje stručnjaka uključuju metaheuristikom.**

Tablica 6.6: Sinteza rezultata analize Hipoteze 2 za svaku mjeru evaluacije, metodu regularizacije i način procjene penalizacijskih koeficijenata na skupu stvarnih podataka (+: rezultati podupiru Hipotezu 2, -: rezultati ne podupiru Hipotezu 2).

	Mjera evaluacije	det			mh		
		El. mreža	Lasso	Hrbatna	El. mreža	Lasso	Hrbatna
<b>Domensko znanje 1</b>	<i>RMSE</i>	-	+	-	+	+	+
	<i>MAE</i>	-	+	-	+	+	+
	<i>SMAPE</i>	+	+	+	+	+	+
<b>Domensko znanje 2</b>	<i>RMSE</i>	-	+	-	+	+	+
	<i>MAE</i>	+	+	+	+	+	+
	<i>SMAPE</i>	+	+	+	+	+	+

Procjena penalizacijskog koeficijenta (det: deterministički, mh: metaheuristika). U Hipotezi 2 uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka.

### 6.4.2. Scenariji

Rezultati mjera evaluacije simuliranih podataka prikazani su sljedećim slikama (Scenarij 1: 5.17, 5.18, 5.19, 5.20, 5.21 i 5.22 (str. 86 - 88); Scenarij 2: 5.33, 5.34, 5.35, 5.36, 5.37 i 5.38 (str. 131 - 133); Scenarij 3: 5.50, 5.51, 5.52, 5.53, 5.54 i 5.55 (str. 131 - 133)). Na njima je vidljivo kako je za svaku mjeru evaluacije CART metoda imala znatno bolje rezultate od metoda regularizacije s uključenim domenskim znanjem stručnjaka.

S obzirom na to da su sljedeća tumačenja jednaka za sve scenarije i oba načina uvođenja domenskog znanja u metode regularizacije (i deterministički i metaheuristikom), odnosno obje podhipoteze, u nastavku slijedi skupno pojašnjenje.

Rezultati jednofaktorskih analiza varijanci s faktorom Metoda za *RMSE*, *MAE* i *SMAPE* pokazuju da je efekt faktora Metoda statistički značajan. U analizi varijance za svaku mjeru evaluacije, statistički je značajno različit od nule koeficijent uz kontrast c31 (vidljivo u tablicama u prilogima rada 7.5 (str. 179), 7.6 (str. 181), 7.10 (str. 189), 7.11 (str. 191), 7.14 (str. 197) i 7.15 (str. 199)). Kod svih je prisutna negativna vrijednost koeficijenta uz kontrast c31



što upućuje da metoda CART rezultira nižom prosječnom greškom od metoda regularizacije koje uključuju domensko znanje stručnjaka. Dodatna *post hoc* analiza Tukeyjevim testom HSD upućuje da je i statistički značajna razlika prosječnih greški između metode CART i svake pojedinačne metode regularizacije u korist metode CART.

Tablica 6.7 predstavlja sintezu rezultata analize Hipoteze 2 na scenarijima. **Može se reći da se na simuliranim podacima Hipoteza 2 nije potvrđena.**

Tablica 6.7: Sinteza rezultata analize Hipoteze 2 za svaku mjeru evaluacije i način procjene penalizacijskih koeficijenata za svaki scenarij (+: Hipoteza 2 je potvrđena, -: Hipoteza 2 nije potvrđena).

Procjena pen. koef.	Mjera evaluacije	Scenarij 1	Scenarij 2	Scenarij 3
det	<i>RMSE</i>	-	-	-
	<i>MAE</i>	-	-	-
	<i>SMAPE</i>	-	-	-
mh	<i>RMSE</i>	-	-	-
	<i>MAE</i>	-	-	-
	<i>SMAPE</i>	-	-	-

Procjena penalizacijskog koeficijenta (det: deterministički, mh: metaheuristika). U Hipotezi 2 uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka.

Pitanje je što je razlog tome da je u sva tri scenarija simuliranih podataka metoda CART toliko uspješnija u predikciji od metoda regularizacije s uključenim domenskim znanjem stručnjaka? S obzirom na to da su scenariji dosta različitih struktura, za pretpostaviti je da struktura scenarija ipak nije glavni razlog tomu. Potencijalno je problem u tome kako se kod simuliranih podataka modeliralo "mišljenje stručnjaka", odnosno domensko znanje. Možda promatranje korelacija te, određivanje kategorije utjecaja ulaznih značajki prema vrijednostima korelacija nije dovoljno. Možda stvarni stručnjaci zaista imaju neka domenska znanja koja nisu vidljiva iz podataka, pa se kao takva niti ne mogu oponašati, makar ne na ovaj način. Navedeno naglašava važnost kvalitete domenskog znanja. Naime, doprinos uvođenja domenskog znanja stručnjaka pri učenju modela jako je ovisan o kvaliteti toga znanja. Ukoliko je kvaliteta narušena, teško da

će takvo domensko znanje doprinijeti poboljšanju uspješnosti predikcije, što ostavlja prostora za daljnja istraživanja, odnosno, istraživanja kako je na simuliranim podacima moguće generirati domensko znanje.

S druge strane, za stvarne podatke, rezultati podupiru Hipotezu 2 za uvođenje domenskog znanja metaheuristikom, dok za determinističko uvođenje domenskog znanja stručnjaka djelomično potvrđuju Hipotezu 2. Premda su ti podaci naizgled kompleksniji, na njima je zaista bilo moguće ispitati stručnjake iz te domene i oni su se pokazali kao vrijedan izvor informacija koji doprinosi poboljšanju performansi modela. Navedeni rezultat je vrlo zanimljiv jer pokazuje da je danas, kada vrlo brzo napreduje i poboljšava se metodologija potpore odlučivanju, i dalje ipak iznimno važan ljudski faktor i razumijevanje podataka koji se analiziraju, odnosno domensko znanje.

## 6.5. USPOREDBA S LITERATUROM

U ovom istraživanju predložena je metoda za odabir značajki u linearnom regresijskom modelu koja u metodu regularizacije uključuje domensko znanje stručnjaka evaluirana pomoću simulacijskog eksperimenta, po uzoru na [50, 51].

U radu [50] uvedena je metoda regularizacije elastičnom mrežom te su, na scenarijima gotovo jednakim onima u ovom istraživanju, analizirane uspješnosti predikcije lasso i hrbatne regularizacije te regularizacije elastičnom mrežom. Uspješnosti predikcije triju metoda regularizacije uspoređene su u smislu srednje kvadratne greške (eng. *mean square error - MSE*). Istraživanje [50] pokazalo je da je regularizacija elastičnom mrežom znatno točnija od lasso regularizacije, čak i kada je lasso regularizacija puno točnija od hrbatne regularizacije, pri čemu je to bilo najizraženije u Scenariju 3, odnosno pod kolinearnošću.

Scenariji korišteni u ovom istraživanju slični su onima u [50], ali nisu potpuno identični. Naime, oba istraživanja su simulirala po 50 skupova podataka iz pojedinog scenarija, ali nisu imala isti broj slučajeva za treniranje, validaciju i testiranje (u ovom istraživanju je broj slučajeva veći). Dodatno se u istraživanju [50] provodilo poduzorkovanje metodom *bootstrapping-a*. S obzirom na to da scenariji nisu potpuno isti pa je time narušena usporedivosti iznosa greške predikcije u kvantitativom smislu, usporedba rezultata ovog istraživanja s prethodnim istraživanjem može se provesti na kvalitativnoj razini usporedbom mjere evaluacije *RMSE* za različite metode regularizacije budući da je ta mjera direktno usporediva s *MSE* koja je korištena u [50].

Rezultati su sljedeći:

- Scenarij 1 - za modele koji ne koriste domensko znanje stručnjaka ili ga uključuju deterministički, najmanju prosječnu *RMSE* ima regularizacija elastičnom mrežom, zatim lasso pa hrbatna regularizacija, što je u skladu s [50]. U slučaju kada se domensko znanje stručnjaka uključuje metaheuristikom, sve metode su imale sličnu prosječnu *RMSE* (slika 5.14 na stranici 81).
- Scenarij 2 - za modele koji ne koriste domensko znanje stručnjaka najmanju prosječnu *RMSE* ima regularizacija elastičnom mrežom, zatim lasso pa hrbatna regularizacija, što je u skladu s [50]. Kada se domensko znanje stručnjaka uključuje deterministički, najmanju prosječnu *RMSE* ima regularizacija elastičnom mrežom, zatim hrbatna pa lasso regularizacija. U slučaju kada se domensko znanje stručnjaka uključuje metaheuristikom, sve metode su imale sličnu prosječnu *RMSE* (slika 5.30 na stranici 103).
- Scenarij 3 - za modele koji ne koriste domensko znanje stručnjaka najmanju prosječnu *RMSE* ima regularizacija elastičnom mrežom, zatim lasso pa hrbatna regularizacija, što je u skladu s [50]. Kada se domensko znanje stručnjaka uključuje deterministički, najmanju prosječnu *RMSE* ima regularizacija elastičnom mrežom, zatim hrbatna pa lasso regularizacija. U slučaju kada se domensko znanje stručnjaka uključuje metaheuristikom, sve metode su imale sličnu prosječnu *RMSE* (slika 5.47 na stranici 127).

Dakle, vidljivo je da su odnosi veličina *RMSE* greški između metoda regularizacije bez domenskog znanja stručnjaka isti kao u [50], dok se odnosi veličina mijenjaju pri uvođenju domenskog znanja stručnjaka. No, bitno je za naglasiti da rezultati simulacijskog eksperimenta unutar kojega se provodila dvofaktorska analiza varijance za mjeru evaluacije *RMSE*, pokazuju da efekti faktora Metoda regularizacije niti za jedan scenariji nisu statistički značajni.

## 7. ZAKLJUČAK

Kroz ovo istraživanje protezalo se pitanje poboljšanja uspješnosti predikcije linearnog regresijskog modela korištenjem domenskog znanja stručnjaka, a istraživanje je imalo za cilj razviti metodu za odabir značajki u linearnom regresijskom modelu koja u metodu regularizacije uključuje domensko znanje stručnjaka. Novopredložena metoda odabira značajki temeljila se na postojećim metodama lasso regularizacije, regularizacije elastičnom mrežom i hrbatne regularizacije u koje se domensko znanje stručnjaka uključilo uvođenjem penalizacijskog koeficijenta. Time se nastojalo učenje metode usmjeriti domenskim znanjem stručnjaka s ciljem dobivanja modela koji su interpretabilni i pouzdani krajnjim korisnicima te uspješniji u predikciji. Vrijednost penalizacijskog koeficijenta zavisila je o mišljenju stručnjaka o utjecaju ulazne značajke na izlaznu značajku, pri čemu su bile moguće četiri kategorije utjecaja ulaznih značajki na izlaznu: *Nema utjecaja*, *Slab utjecaj*, *Srednji utjecaj* i *Jak utjecaj* i procjenjivale su se na dva različita načina - determinističkim i metaheurističkim pristupom. U okviru determinističkog pristupa, penalizacijski koeficijent određen je na način da su se unaprijed određene vrijednosti i veće vrijednosti koeficijenta dale značajkama koje su stručnjaci smatrali manje bitnima i obratno. U okviru metaheurističkog pristupa optimizirale su se vrijednosti penalizacijskih koeficijenata, a ujedno i parametara regularizacije. Predložena metoda može se smatrati hibridnom metodom odabira značajki. Evaluacija predložene metode provela se na stvarnim podacima o troškovima energije zgrada javnog sektora Republike Hrvatske te na 50 skupova podataka simuliranih iz tri scenarija proizašla iz literature, u okviru dvije postavljene istraživačke hipoteze.

**Hipoteza 1:** Predložena metoda rezultira uspješnijom predikcijom od metode regularizacije koja ne uključuje domensko znanje stručnjaka.

Rezultati na stvarnim podacima podupiru djelomično Hipotezu 1 za metode regularizacije koje domensko znanje stručnjaka uvode determinističkim pristupom, a podupiru je za metode regularizacije koje domensko znanje stručnjaka uvode metaheurističkim pristupom. Poboljša-

nja u uspješnosti predikcije uvođenjem domenskog znanja metaheuristikom u odnosu na modele koji ne koriste domenska znanja stručnjaka su bila znatna u kvantitativnom smislu. U okviru simulacijskog eksperimenta, Hipoteza 1 potvrđena je za metode regularizacije koje domensko znanje stručnjaka uvode metaheurističkim pristupom za mjere *RMSE* i *SMAPE*, dok za druge mjere evaluacije nije. Navedeni rezultat naglašava važnost korištenja različitih mjera evaluacije modela jer svaka analizira uspješnost predikcije u drugačijem aspektu.

**Hipoteza 2:** Predložena metoda rezultira uspješnijom predikcijom od metode regresijskih stabala.

Kao i kod Hipoteze 1, rezultati na stvarnim podacima podupiru djelomično Hipotezu 2 za metode regularizacije koje domensko znanje stručnjaka uvode determinističkim pristupom, a podupiru je za metode regularizacije koje domensko znanje stručnjaka uvode metaheurističkim pristupom. U okviru simulacijskog eksperimenta Hipoteza 2 nije potvrđena niti za jedan pristupu uvođenja domenskog znanja stručnjaka. Rezultati su pokazali da je uvođenje domenskog znanja imalo veći doprinos u jednostavnijim modelima s manje prediktora, nižim korelacijama i manjim greškama kao što je u Scenariju 1. U složenijim modelima kao što su oni iz Scenarija 2 i Scenarija 3, doprinos domenskog znanja postaje manje značajan zbog povećane kompleksnosti i grešaka u modelu.

Razlog zašto rezultati analiza skupa stvarnih podataka podupiru postavljene istraživačke hipoteze vjerojatno leži u strukturi tih podataka te u činjenici da je za te podatke zaista bilo moguće ispitati stručnjake iz domene. Oni su se pokazali kao vrijedan izvor informacija koji doprinosi poboljšanju performansi modela. Navedeni rezultat je zanimljiv jer naglašava važnost ljudskog faktora i poznavanje domene problema pri korištenju sustava za potporu odlučivanju.

Konačno, rezultati simulacijskog eksperimenta pokazuju kako različiti načini procjene penalizacijskih koeficijenta i sama struktura modela utječu na varijabilnost i pristranost modela. Ujedno, otvaraju daljnja pitanja u smislu kako postići i da za kompleksnije skupove podataka domensko znanje stručnjaka doprinese smanjenju varijabilnosti i/ili pristranosti modela. Osim navedenog, ovo istraživanje otvara još nekoliko mogućih smjerova za daljnja istraživanja navedenih u nastavku:

- 1) Kvaliteta domenskog znanja: ključno je osigurati kvalitete domenskog znanja stručnjaka koje se koristi za učenje modela. Loša kvaliteta znanja može negativno utjecati na us-

pješnost predikcije, stoga je potrebno osigurati kvalitetno domensko znanje. U ovom je istraživanju za stvarne podatke bilo moguće doći do domenskog znanja stručnjaka, a njihovo uključivanje u model poboljšalo je uspješnost predikcije.

- 2) Drugačiji načini uključivanja domenskog znanja: istražiti drugačije načine prikupljanja i uključivanja domenskog znanja stručnjaka u metode regularizacije i prilagoditi ih strukturi problema, s obzirom na to da je istraživanje pokazalo da je struktura problema od značajne važnosti za ishod predikcije, odnosno njezinu uspješnost.
- 3) Uključivanje domenskog znanja u druge metode potpore odlučivanju: primjena domenskog znanja stručnjaka u drugim metodama potpore odlučivanju povećava vjerojatnost da se pronađe odgovarajući pristup uvođenja domenskog znanja široj klasi problema.
- 4) Promjene u korištenoj metaheuristici: metaheuristike su brzo rastuće područje s velikim potencijalom. Moguće je unutar iste metaheuristike isprobati drugačije parametre optimizacije, kao i isprobati druge metaheuristike ili ih čak kombinirati. Potencijalno postaviti ograničenja na odnose veličina penalizacijskih koeficijenata kako bi bili intuitivni korisnicima.

## 7.1. DOPRINOS

Znanstveni doprinos predloženog istraživanja višestruk je i može se sažeti u sljedećim točkama:

- 1) **Sistematizacija dosadašnjih istraživanja o korištenju domenskog znanja stručnjaka u metodama potpore odlučivanju.**

Dosadašnja istraživanja podupiru uvođenje domenskih znanja stručnjaka i naglašavaju važnost uključivanja istih u fazi učenja modela kao i važnost kvalitete domenskog znanja.

- 2) **Sistematizacija dosadašnjih istraživanja o metodama odabira značajki.**

Dosadašnja istraživanja prepoznaju potencijal u hibridnim metodama odabira značajki, posebice u vezi s metodologijama temeljenim na metaheurističkim algoritmima.

- 3) **Prijedlog metode za odabir značajki u linearnom regresijskom modelu koja u metodu regularizacije uključuje domensko znanje stručnjaka.**

Koriste se dva pristupa uključivanja znanja stručnjaka - deterministički i pomoću metaheurističkog algoritma, a predložena metoda može se smatrati hibridnom metodom odabira značajki.

- 4) **Evaluacija metode za odabir značajki na stvarnom skupu podataka o energetske troškovima zgrada javnog sektora Republike Hrvatske, pri čemu su korištena domenska znanja stručnjaka temeljem kojih je izrađen prediktivni model troškova energija zgrada javnog sektora Republike Hrvatske.**

Rezultati djelomično podupiru postavljene istraživačke hipoteze za metode regularizacije koje koriste deterministički pristup uvođenja domenskog znanja stručnjaka, dok ih podupiru za metode regularizacije koje koriste metaheuristički pristup uvođenja domenskog znanja stručnjaka. Odnosno, uvođenje domenskog znanja stručnjaka metaheurističkim pristupom bilo je uspješnije u predikciji od modela bez domenskog znanja stručnjaka, kao i od metode CART, dok je uvođenje domenskog znanja stručnjaka determinističkim pristupom samo u nekim mjerama evaluacije bilo uspješnije od spomenutih metoda.

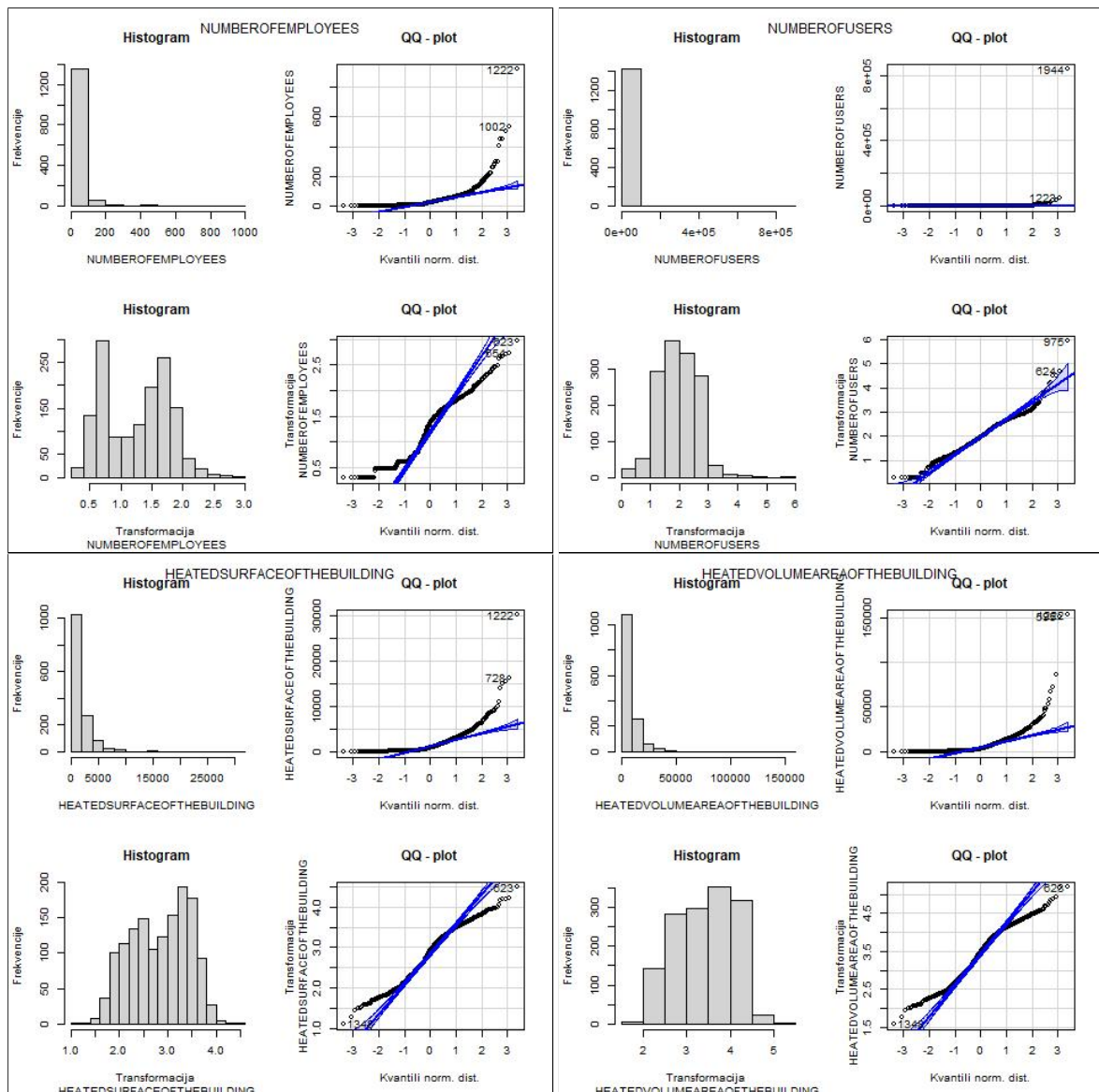
- 5) **Rezultati mišljenja stručnjaka, tj. domenskog znanja, o utjecaju pojedine karakteristike zgrade na troškove energije dostupni su kao otvoreni znanstveni podaci u repozitoriju Zenodo [136].**
- 6) **Predložena metoda evaluirana je i u okviru simulacijskog eksperimenta na tri scenarija iz literature te su provjerene postavljene istraživačke hipoteze.**

Uvođenje domenskog znanja ima veći doprinos u jednostavnijim modelima s manje prediktora, nižim korelacijama i manjim greškama kao što je Scenarij 1. U složenijim modelima, kao što su modeli na skupovima iz Scenarija 2 i 3, doprinos domenskog znanja postaje manje značajan zbog povećane kompleksnosti i grešaka u modelu. Hipoteza 1 je, stoga, djelomično potvrđena. S druge strane Hipoteza 2 nije potvrđena, odnosno predložena metoda nije točnija od metode CART.

- 7) **Simulirani podaci i skripte za njihovo generiranje dostupni su kao otvoreni znanstveni podaci u repozitoriju Zenodo [129].**

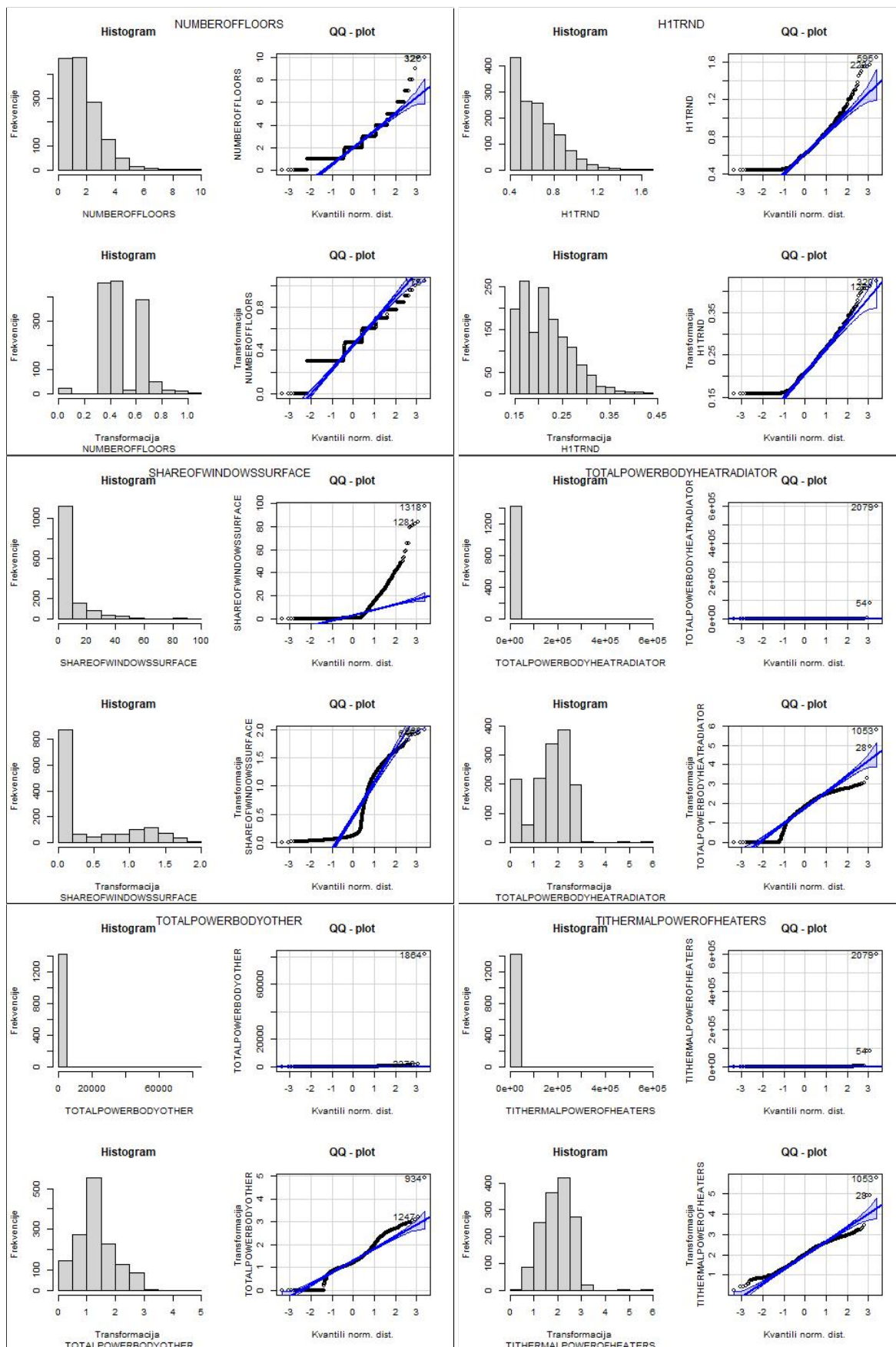
Doprinosi 4) i 5) ujedno predstavljaju značajan društveni doprinos. Naime, modeli za predviđanje troškova energije i identificiranje relevantnih značajki zgrada javnog sektora mogu biti korisni donositeljima odluka u provođenju mjera energetske obnove.

# PRILOG 1 - STVARNI PODACI

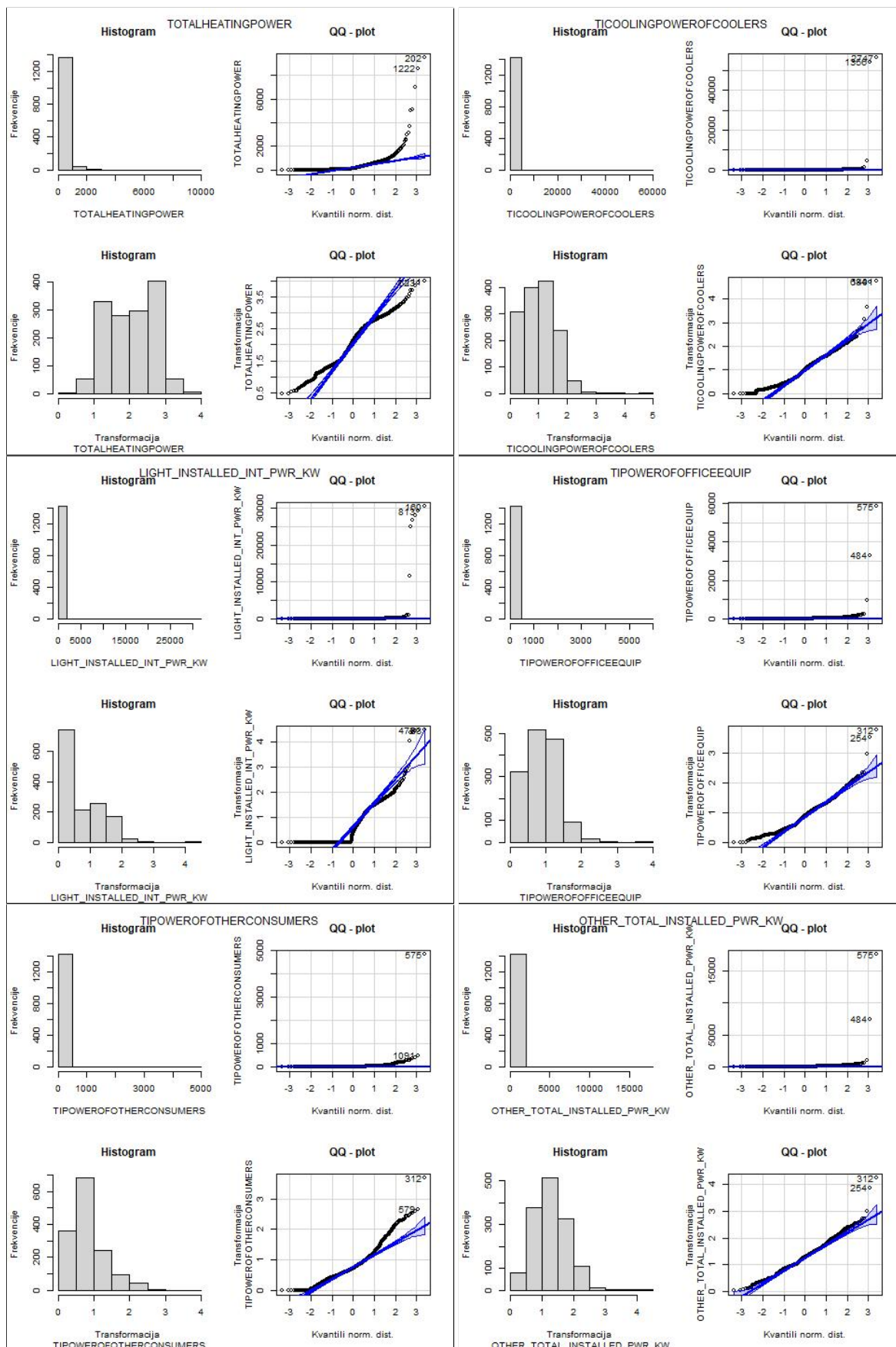


Slika 7.1: Grafički prikaz distribucija značajki [1/4]

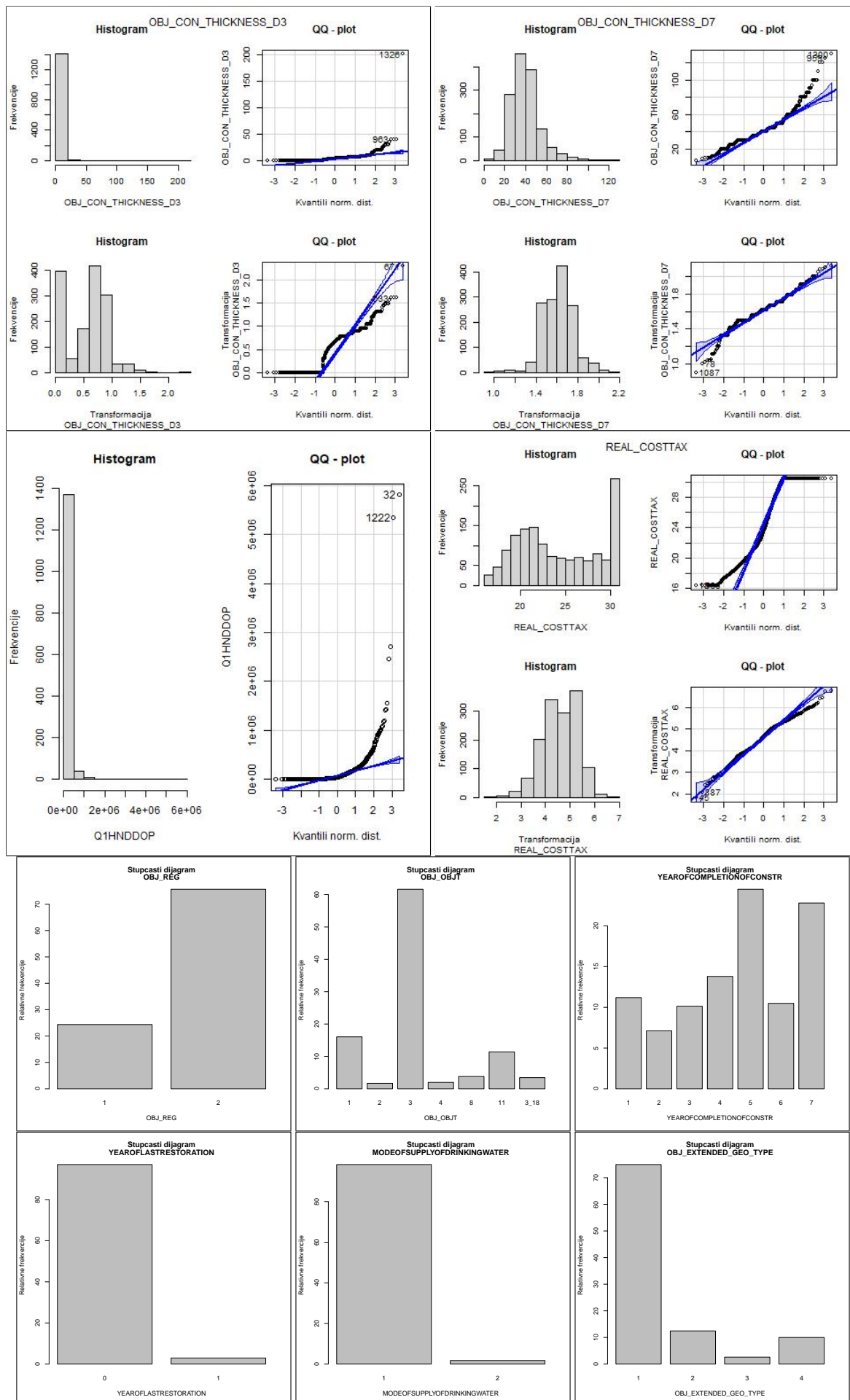




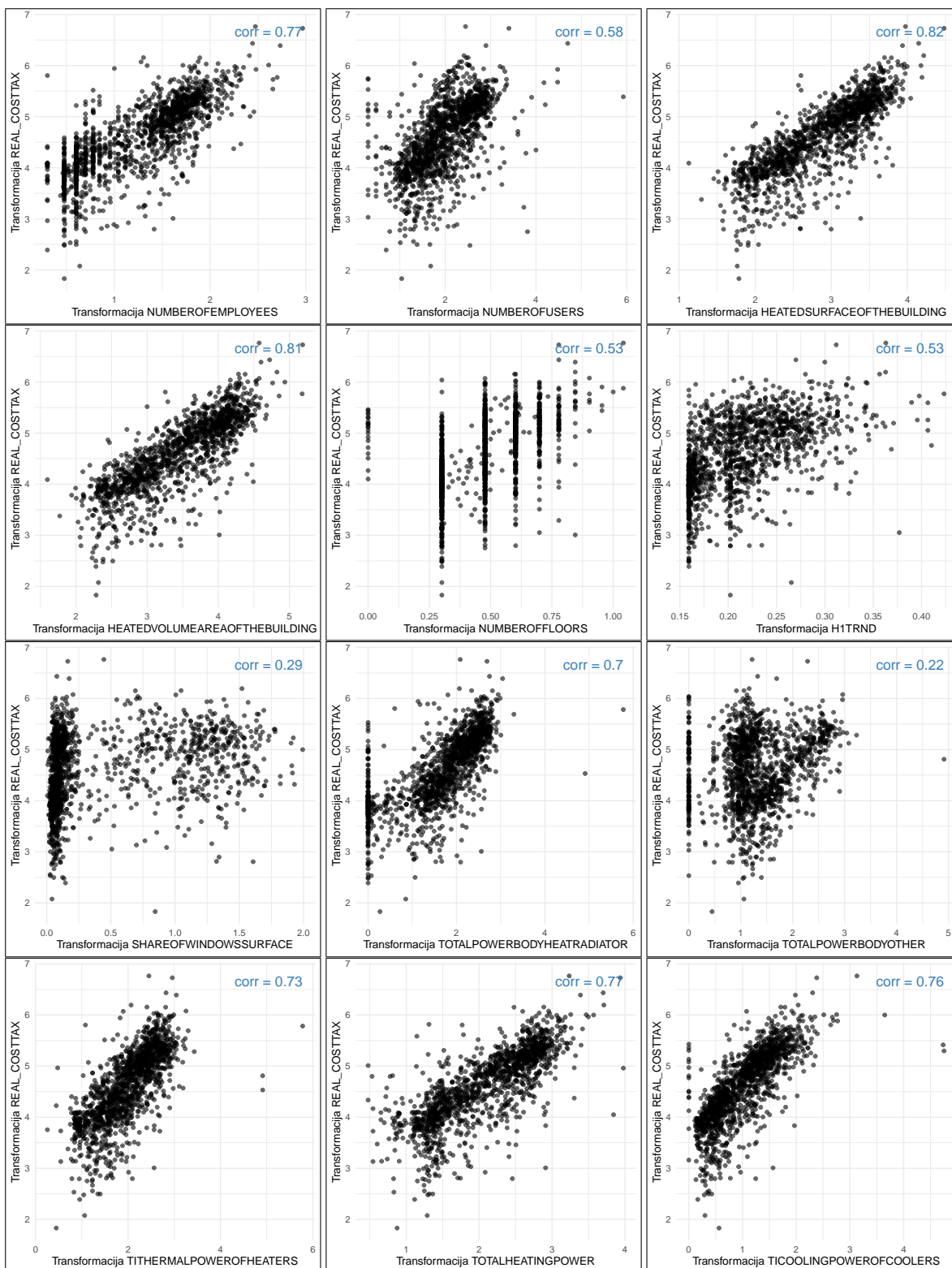
Slika 7.2: Grafički prikaz distribucija značajki [2/4]



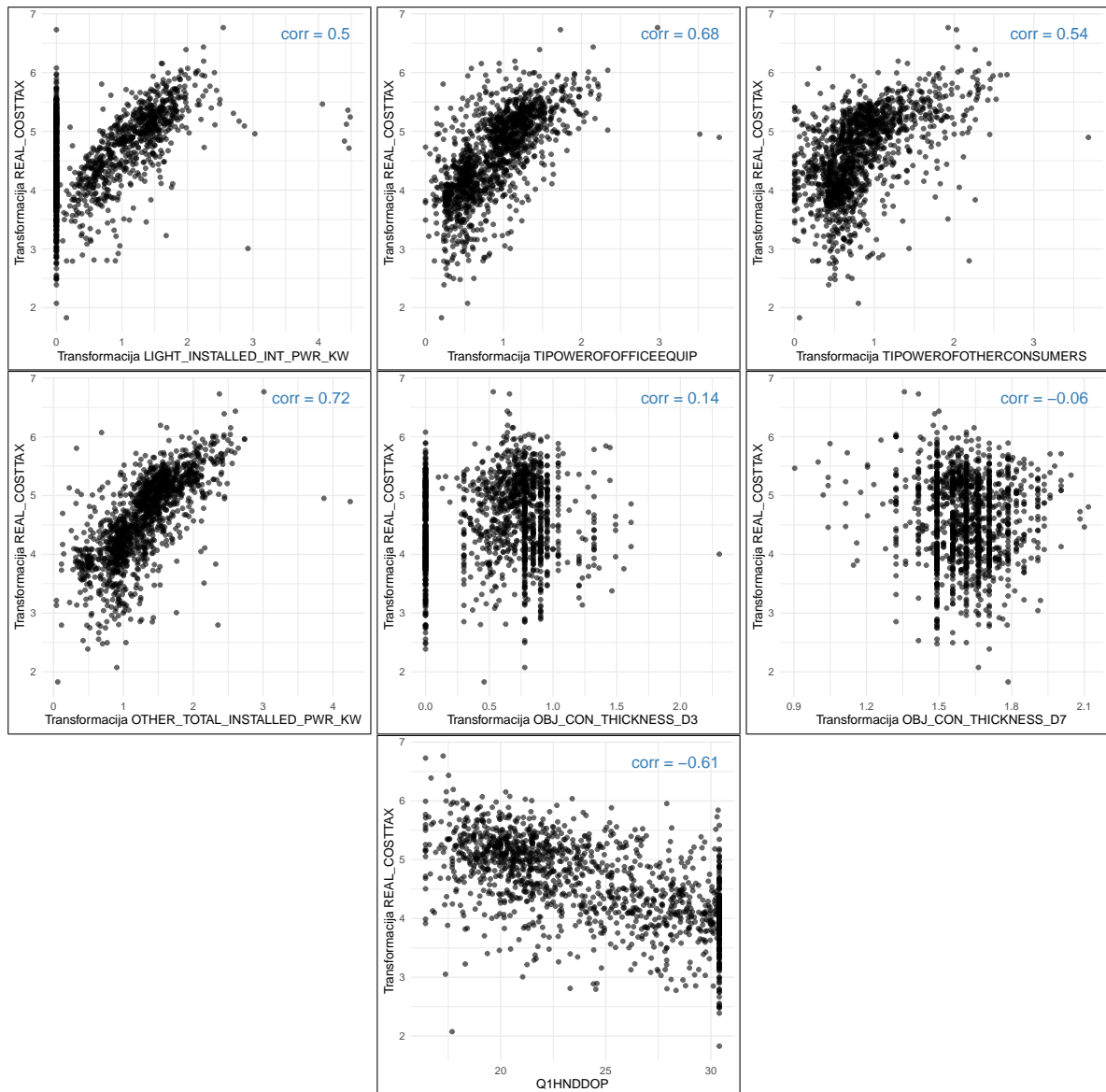
Slika 7.3: Grafički prikaz distribucija značajki [3/4]



Slika 7.4: Grafički prikaz distribucija značajki [4/4]



Slika 7.5: Dijagrami rasipanja i korelacije kvantitativnih nezavisnih značajki i zavisne značajke [1/2]



Slika 7.6: Dijagrami rasipanja i korelacije kvantitativnih nezavisnih značajki i zavisne značajke [2/2]

Tablica 7.1: Rezultati mišljenja stručnjaka o utjecaju svake ulazne značajke na izlaznu

Značajka	Stručnjak			
	1	2	3	4
NUMBEROFEMPLOYEES	3	1	1	3
NUMBEROFUSERS	3	1	1	3
HEATEDSURFACEOFTHEBUILDING	4	3	2	4
HEATEDVOLUMEAREAOFTHEBUILDING	4	4	1	4
NUMBEROFFLOORS	3	2	1	4
H1TRND	4	3	1	1
SHAREOFWINDOWSSURFACE	3	3	1	4
TOTALPOWERBODYHEATRADIATOR	3	3	1	3
TOTALPOWERBODYTHER	3	3	1	3
TITHERMALPOWEROFHEATERS	3	3	1	3
TOTALHEATINGPOWER	3	3	1	3
TICOOLINGPOWEROFCOOLERS	3	3	1	3
LIGHT_INSTALLED_INT_PWR_KW	2	2	1	2
TIPOWEROFFOFFICEEQUIP	2	2	2	3
TIPOWEROFOTHERCONSUMERS	2	2	2	3
OTHER_TOTAL_INSTALLED_PWR_KW	2	2	2	3
OBJ_CON_THICKNESS_D3	4	4	1	4
OBJ_CON_THICKNESS_D7	4	4	1	4
Q1HNDDOP	4	3	1	1
OBJ_REG	4	4	4	4
OBJ_OBJT	4	3	2	3
YEAROFCOMPLETIONOFCONSTR	4	3	3	4
YEAROFLASTRESTORATION	4	4	3	4
MODEOFSUPPLYOFDRINKINGWATER	2	2	1	3
OBJ_EXTENDED_GEO_TYPE	2	3	1	4

1: Nema utjecaja, 2: Slab utjecaj, 3: Srednji utjecaj, 4: Jak utjecaj.

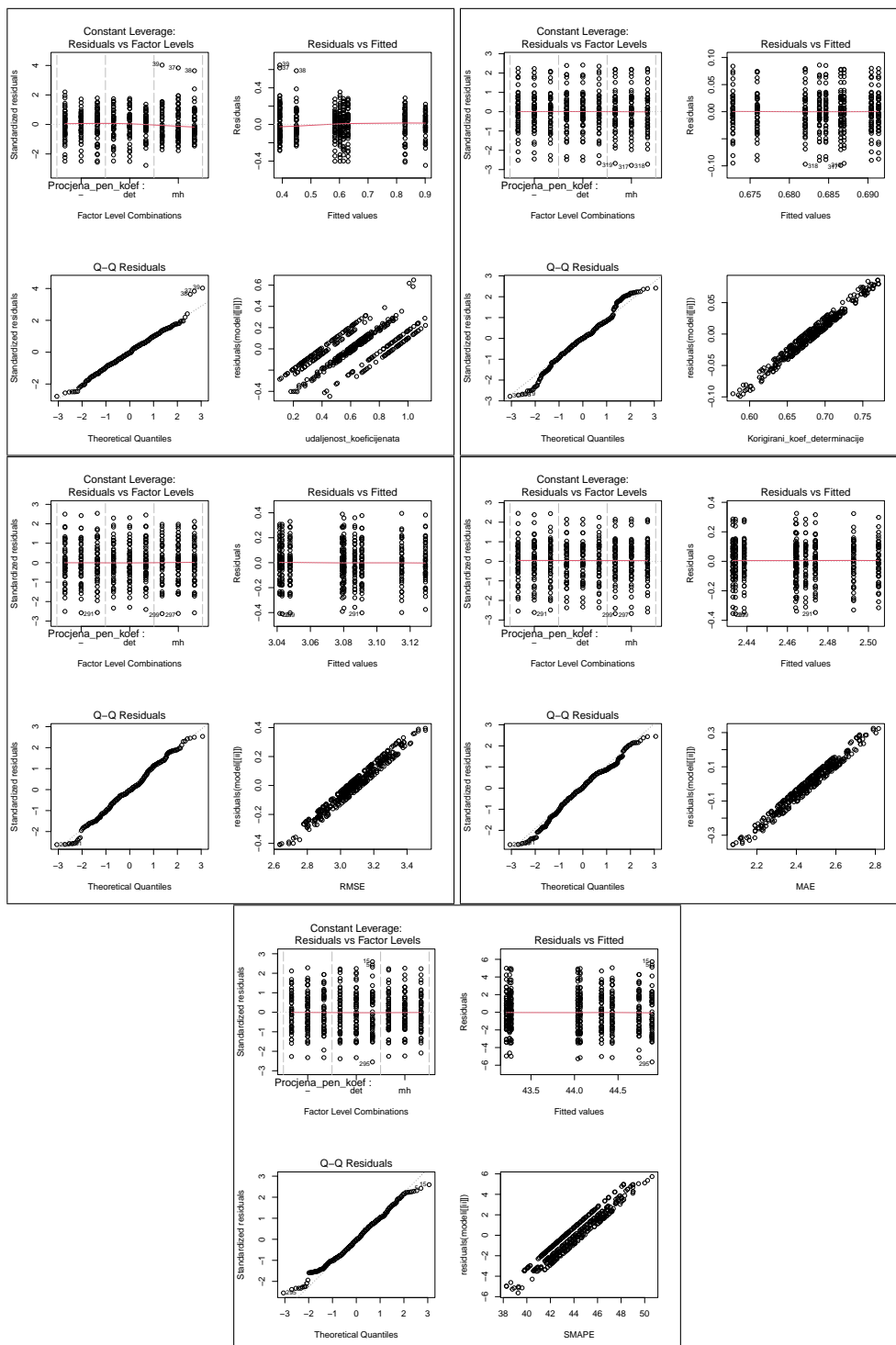
Tablica 7.2: Utjecaj ulazne značajke u obje promatrane vrste domenskog znanja

Značajka	Domensko znanje 1	Domensko znanje 2
NUMBEROFEMPLOYEES	3	1
NUMBEROFUSERS	3	1
HEATEDSURFACEOFTHEBUILDING	4	2
HEATEDVOLUMEAREAOFTHEBUILDING	4	1
NUMBEROFFLOORS	3	1
H1TRND	3	1
SHAREOFWINDOWSSURFACE	3	1
TOTALPOWERBODYHEATRADIATOR	3	1
TOTALPOWERBODYOTHER	3	1
TITHERMALPOWEROFHEATERS	3	1
TOTALHEATINGPOWER	3	1
TICOOLINGPOWEROFCOOLERS	3	1
LIGHT_INSTALLED_INT_PWR_KW	2	1
TIPOWEROFFOFFICEEQUIP	2	2
TIPOWEROFOTHERCONSUMERS	2	2
OTHER_TOTAL_INSTALLED_PWR_KW	2	2
OBJ_CON_THICKNESS_D3	4	1
OBJ_CON_THICKNESS_D7	4	1
Q1HNDDOP	3	1
OBJ_REG	4	4
OBJ_OBJT	3	2
YEAROFCOMPLETIONOFCONSTR	4	3
YEAROFLASTRESTORATION	4	3
MODEOFSUPPLYOFDRINKINGWATER	2	1
OBJ_EXTENDED_GEO_TYPE	3	1

1: Nema utjecaja, 2: Slab utjecaj, 3: Srednji utjecaj, 4: Jak utjecaj. Domensko znanje 1: medijan mišljenja Stručnjaka 1, Stručnjaka 2 i Stručnjaka 4, Domensko znanje 2: mišljenje Stručnjaka 3.



# PRILOG 2 - SCENARIJ 1



Slika 7.7: Scenarij 1, Hipoteza 1 - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.



Tablica 7.3: Scenarij 1, Hipoteza 1 - procjene koeficijenata i intervali pouzdanosti (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije

Mjera evaluacije	Značajka	Kontrast	Koeficijent	2.50%	97.50%
Udaljenost koeficijenata	(Slobodni koeficijent)	-	0.599*	0.584	0.614
	Procjena pen. koef.	c11	0.073*	0.051	0.094
		c12	-0.152*	-0.171	-0.134
	Metoda regularizacije	c21	-0.063*	-0.085	-0.042
		c22	0.096*	0.077	0.114
	Procjena pen. koef.:	c11:c21	-0.007	-0.037	0.024
	Metoda regularizacije	c12:c21	0.041*	0.015	0.067
		c11:c22	0.027*	0.001	0.053
c12:c22	-0.053*	-0.076	-0.031		
Korigirani koeficijent determinacije	(Slobodni koeficijent)	-	0.684*	0.681	0.687
	Procjena pen. koef.	c11	0.002	-0.003	0.007
		c12	0.002	-0.002	0.006
	Metoda regularizacije	c21	0.004	-0.001	0.008
		c22	-0.005*	-0.009	-0.001
	Procjena pen. koef.:	c11:c21	0.001	-0.006	0.008
	Metoda regularizacije	c12:c21	-0.001	-0.007	0.005
		c11:c22	-0.003	-0.008	0.003
c12:c22	0.002	-0.003	0.007		
<i>RMSE</i>	(Slobodni koeficijent)	-	3.0793*	3.0647	3.0939
	Procjena pen. koef.	c11	0.0158	-0.0049	0.0364
		c12	-0.0273*	-0.0452	-0.0094
	Metoda regularizacije	c21	-0.0118	-0.0325	0.0088
		c22	0.0121	-0.0058	0.0300
	Procjena pen. koef.:	c11:c21	-0.0040	-0.0331	0.0252
	Metoda regularizacije	c12:c21	0.0087	-0.0166	0.0339
		c11:c22	-0.0001	-0.0254	0.0251
c12:c22	-0.0091	-0.0309	0.0128		

## Prilog 2 - Scenarij 1

Mjera evaluacije	Značajka	Kontrast	Koeficijent	2.50%	97.50%
<i>MAE</i>	(Slobodni koeficijent)	-	2.4639*	2.4515	2.4763
	Procjena pen. koef.	c11	0.0131	-0.0045	0.0307
		c12	-0.0220*	-0.0372	-0.0068
	Metoda regularizacije	c21	-0.0096	-0.0272	0.0079
		c22	0.0101	-0.0051	0.0253
	Procjena pen. koef.:	c11:c21	-0.0029	-0.0277	0.0219
	Metoda regularizacije	c12:c21	0.0071	-0.0144	0.0286
		c11:c22	-0.0006	-0.0221	0.0209
		c12:c22	-0.0076	-0.0262	0.0110
<i>SMAPE</i>	(Slobodni koeficijent)	-	44.0249*	43.8180	44.2318
	Procjena pen. koef.	c11	0.4699*	0.1773	0.7625
		c12	-0.5416*	-0.7950	-0.2882
	Metoda regularizacije	c21	-0.1510	-0.4436	0.1416
		c22	0.1952	-0.0582	0.4486
	Procjena pen. koef.:	c11:c21	-0.0336	-0.4474	0.3802
	Metoda regularizacije	c12:c21	0.1583	-0.2001	0.5167
		c11:c22	-0.0411	-0.3995	0.3173
		c12:c22	-0.1968	-0.5072	0.1136

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c11 testira se Hipoteza 1 ovog istraživanja, dok kontrast c12 dodatno uspoređuje dva načina uvođenja domenskog znanja stručnjaka u metode regularizacije. U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

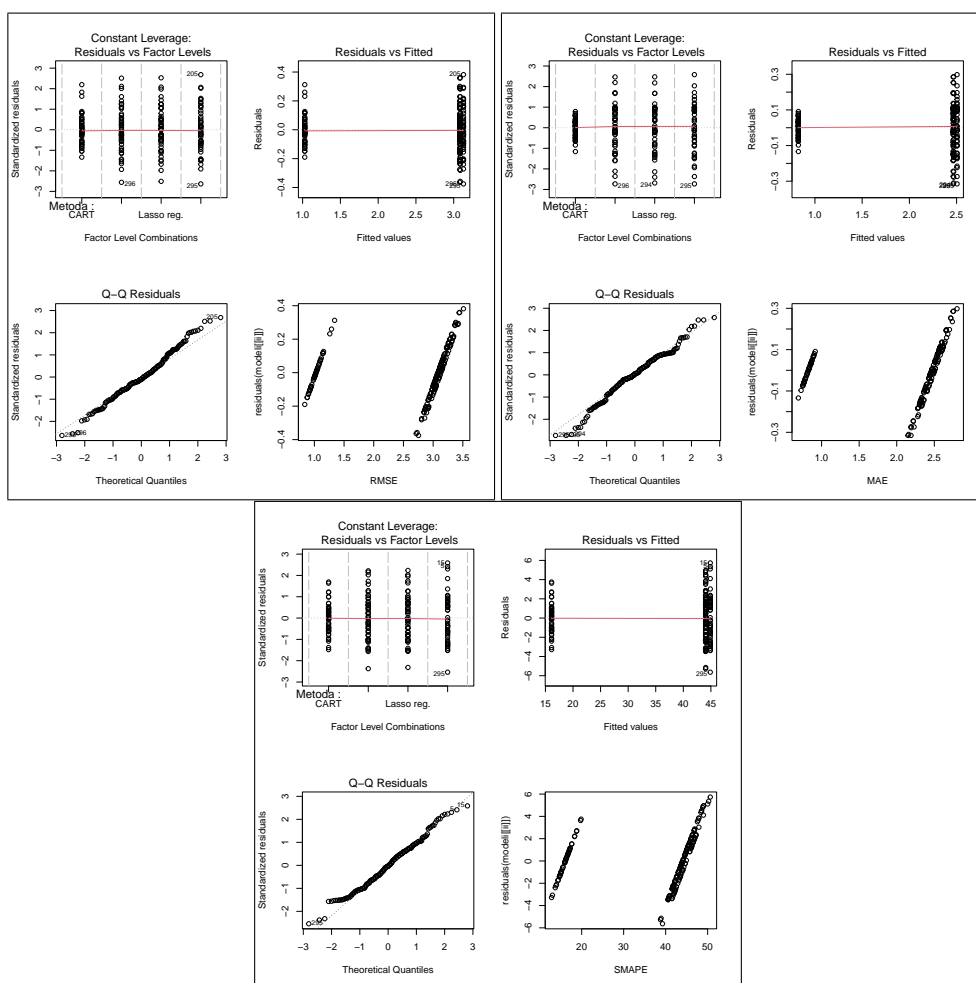
Tablica 7.4: Scenarij 1, Hipoteza 1, udaljenost koeficijenata - *post hoc* analiza Tukey HSD testom (\*: statistički značajno različit od 0 na razini statističke značajnosti 0.05.)

Kombinacije razina faktora	diff	<i>p</i>
det:El. Mreža - :El. Mreža	0.012	1.0000
mh:El. Mreža - :El. Mreža	-0.211*	0.00

Kombinacije razina faktora		diff	<i>p</i>
-:Lasso	-:El. Mreža	-0.018	0.9998
det:Lasso	-:El. Mreža	0.028	0.9946
mh:Lasso	-:El. Mreža	-0.211*	0.00
-:Hrvatna	-:El. Mreža	0.228*	0.00
det:Hrvatna	-:El. Mreža	0.299*	0.00
mh:Hrvatna	-:El. Mreža	-0.153*	0.0001
mh:El. mreža	det:El. mreža	-0.223*	0.00
-:Lasso	det:El. mreža	-0.030	0.9917
det:Lasso	det:El. mreža	0.016	0.9999
mh:Lasso	det:El. mreža	-0.223*	0.00
-:Hrvatna	det:El. mreža	0.216*	0.00
det:Hrvatna	det:El. mreža	0.287*	0.00
mh:Hrvatna	det:El. mreža	-0.165*	0.00
-:Lasso	mh:El. mreža	0.193*	0.00
det:Lasso	mh:El. mreža	0.239*	0.00
mh:Lasso	mh:El. mreža	0.000	1.0000
-:Hrvatna	mh:El. mreža	0.439*	0.00
det:Hrvatna	mh:El. mreža	0.510*	0.00
mh:Hrvatna	mh:El. mreža	0.058	0.6939
det:Lasso	-:Lasso	0.046	0.8938
mh:Lasso	-:Lasso	-0.193*	0.00
-:Hrvatna	-:Lasso	0.246*	0.00
det:Hrvatna	-:Lasso	0.317*	0.00
mh:Hrvatna	-:Lasso	-0.135*	0.0013
mh:Lasso	det:Lasso	-0.239*	0.00
-:Hrvatna	det:Lasso	0.200*	0.00
det:Hrvatna	det:Lasso	0.271*	0.00
mh:Hrvatna	det:Lasso	-0.181*	0.00
-:Hrvatna	mh:Lasso	0.439*	0.00
det:Hrvatna	mh:Lasso	0.510*	0.00
mh:Hrvatna	mh:Lasso	0.058	0.6924

Kombinacije razina faktora		diff	p
det:Hrbatna	-:Hrbatna	0.071	0.4122
mh:Hrbatna	-:Hrbatna	-0.381*	0.00
mh:Hrbatna	det:Hrbatna	-0.452*	0.00

Procjena penalizacijskog koeficijenta(- : nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

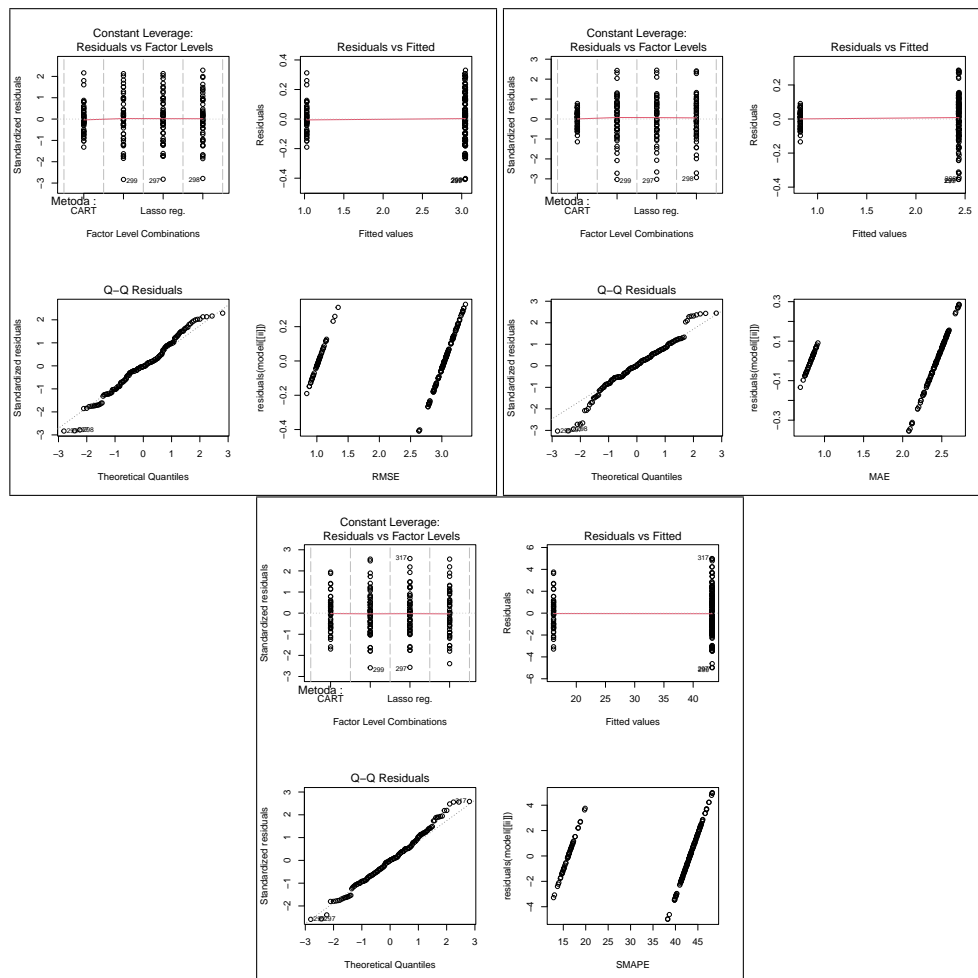


Slika 7.8: Scenarij 1, Hipoteza 2a - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Tablica 7.5: Scenarij 1, Hipoteza 2a - procjene koeficijenta i intervali pouzdanosti (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije

<b>Mjera evaluacije</b>	<b>Značajka</b>	<b>Kontrast</b>	<b>Koeficijent</b>	<b>2.50%</b>	<b>97.50%</b>
<i>RMSE</i>	(Slobodni koeficijent)	-	2.581*	2.561	2.601
	Metoda	c31	-1.553*	-1.587	-1.518
		c32	-0.019	-0.051	0.014
		c33	-0.021	-0.050	0.007
<i>MAE</i>	(Slobodni koeficijent)	-	2.066*	2.049	2.082
	Metoda	c31	-1.241*	-1.269	-1.213
		c32	-0.015	-0.042	0.011
		c33	-0.018	-0.041	0.005
<i>SMAPE</i>	(Slobodni koeficijent)	-	37.277*	36.964	37.589
	Metoda	c31	-21.164*	-21.705	-20.622
		c32	-0.293	-0.803	0.218
		c33	-0.413	-0.855	0.029

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2a. U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.



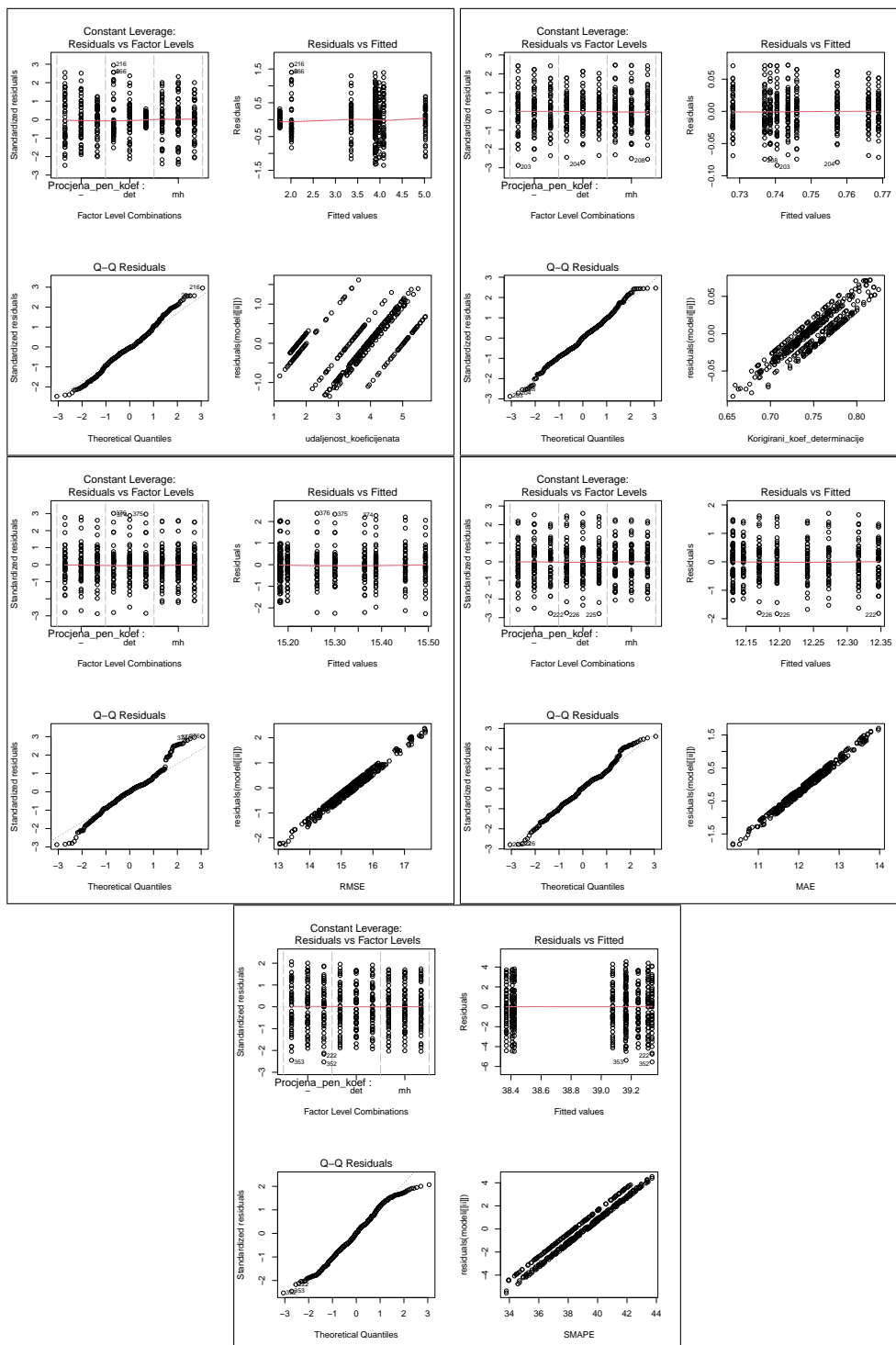
Slika 7.9: Scenarij 1, Hipoteza 2b - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

Tablica 7.6: Scenarij 1, Hipoteza 2b - procjene koeficijenta i intervali pouzdanosti (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije

<b>Mjera evaluacije</b>	<b>Značajka</b>	<b>Kontrast</b>	<b>Koeficijent</b>	<b>2.50%</b>	<b>97.50%</b>
<i>RMSE</i>	(Slobodni koeficijent)	-	2.540*	2.520	2.561
	Metoda	c31	-1.512*	-1.547	-1.477
		c32	-0.001	-0.034	0.032
		c33	-0.003	-0.032	0.026
<i>MAE</i>	(Slobodni koeficijent)	-	2.033*	2.016	2.049
	Metoda	c31	-1.208*	-1.237	-1.179
		c32	-0.001	-0.028	0.026
		c33	-0.003	-0.026	0.021
<i>SMAPE</i>	(Slobodni koeficijent)	-	36.465*	36.193	36.736
	Metoda	c31	-20.351*	-20.822	-19.881
		c32	0.024	-0.420	0.468
		c33	-0.019	-0.403	0.366

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2b. U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

# PRILOG 3 - SCENARIJ 2



Slika 7.10: Scenarij 2, Hipoteza 1 - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.



### Prilog 3 - Scenarij 2

Tablica 7.7: Scenarij 2, Hipoteza 1 - procjene koeficijenata i intervali pouzdanosti (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije za svaku mjeru evaluacije

<b>Mjera evaluacije</b>	<b>Značajka</b>	<b>Kontrast</b>	<b>Koeficijent</b>	<b>2.50%</b>	<b>97.50%</b>
Udaljenost koeficijenata	(Slobodni koeficijent)	-	3.546*	3.494	3.597
	Procjena pen. koeficijenta	c11	0.787*	0.714	0.859
		c12	0.781*	0.718	0.843
	Metoda regularizacije	c21	-0.208*	-0.280	-0.136
		c22	-0.064*	-0.127	-0.002
	Procjena pen. Koeficijenta:	c11:c21	-0.041	-0.144	0.061
	Metoda regularizacije	c12:c21	0.172*	0.084	0.261
		c11:c22	0.631*	0.542	0.720
		c12:c22	0.422*	0.346	0.499
Korigirani koeficijent determinacije	(Slobodni koeficijent)	-	0.747*	0.745	0.750
	Procjena pen. koeficijenta	c11	-0.009*	-0.013	-0.005
		c12	-0.012*	-0.016	-0.009
	Metoda regularizacije	c21	0.004*	0.000	0.007
		c22	-0.002	-0.005	0.002
	Procjena pen. Koeficijenta:	c11:c21	-0.001	-0.007	0.004
	Metoda regularizacije	c12:c21	0.000	-0.005	0.004
		c11:c22	-0.007*	-0.012	-0.002
		c12:c22	-0.003	-0.007	0.002
<i>RMSE</i>	(Slobodni koeficijent)	-	15.314*	15.240	15.388
	Procjena pen. koeficijenta	c11	0.122*	0.018	0.227
		c12	-0.064	-0.155	0.026
	Metoda regularizacije	c21	-0.044	-0.149	0.060
		c22	-0.005	-0.095	0.086
	Procjena pen. Koeficijenta:	c11:c21	-0.028	-0.176	0.120
	Metoda regularizacije	c12:c21	0.024	-0.104	0.152
		c11:c22	0.027	-0.101	0.155

## Prilog 3 - Scenarij 2

Mjera evaluacije	Značajka	Kontrast	Koeficijent	2.50%	97.50%
		c12:c22	0.026	-0.085	0.136
MAE	(Slobodni koeficijent)	-	12.217*	12.155	12.278
	Procjena pen. koeficijenta	c11	0.085	-0.002	0.171
		c12	-0.038	-0.114	0.037
	Metoda regularizacije	c21	-0.036	-0.123	0.050
		c22	-0.005	-0.080	0.070
	Procjena pen. Koeficijenta:	c11:c21	-0.024	-0.147	0.098
	Metoda regularizacije	c12:c21	0.019	-0.087	0.125
		c11:c22	0.020	-0.086	0.126
	c12:c22	0.023	-0.069	0.115	
SMAPE	(Slobodni koeficijent)	-	38.945*	38.739	39.150
	Procjena pen. koeficijenta	c11	0.327*	0.037	0.618
		c12	-0.381*	-0.632	-0.129
	Metoda regularizacije	c21	-0.061	-0.351	0.229
		c22	-0.001	-0.252	0.251
	Procjena pen. Koeficijenta:	c11:c21	-0.045	-0.456	0.365
	Metoda regularizacije	c12:c21	0.047	-0.309	0.403
		c11:c22	0.014	-0.342	0.369
	c12:c22	0.034	-0.274	0.342	

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c11 testira se Hipoteza 1 ovog istraživanja, dok kontrast c12 dodatno uspoređuje dva načina uvođenja domenskog znanja stručnjaka u metode regularizacije. U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

Tablica 7.8: Scenarij 2, Hipoteza 1, udaljenost koeficijenata - *post hoc* analiza Tukey HSD testom (\*: statistički značajno različit od 0 na razini statističke značajnosti 0.05.)

Kombinacije razina faktora	diff	p
det:El. Mreža - :El. Mreža	-2.071	0.00

<b>Kombinacije razina faktora</b>		<b>diff</b>	<b>p</b>
mh:El. Mreža	-:El. Mreža	-0.165	0.8598
-:Lasso	-:El. Mreža	-0.192	0.7211
det:Lasso	-:El. Mreža	-0.729	0.00
mh:Lasso	-:El. Mreža	-0.185	0.7623
-:Hrbatna	-:El. Mreža	0.941	0.00
det:Hrbatna	-:El. Mreža	-2.333	0.00
mh:Hrbatna	-:El. Mreža	-0.100	0.9927
mh:El. mreža	det:El. mreža	1.906	0.00
-:Lasso	det:El. mreža	1.878	0.00
det:Lasso	det:El. mreža	1.341	0.00
mh:Lasso	det:El. mreža	1.886	0.00
-:Hrbatna	det:El. mreža	3.011	0.00
det:Hrbatna	det:El. mreža	-0.263	0.2982
mh:Hrbatna	det:El. mreža	1.971	0.00
-:Lasso	mh:El. mreža	-0.028	1.0000
det:Lasso	mh:El. mreža	-0.564	0.00
mh:Lasso	mh:El. mreža	-0.020	1.0000
-:Hrbatna	mh:El. mreža	1.105	0.00
det:Hrbatna	mh:El. mreža	-2.169	0.00
mh:Hrbatna	mh:El. mreža	0.065	0.9997
det:Lasso	-:Lasso	-0.537	0.0001
mh:Lasso	-:Lasso	0.007	1.0000
-:Hrbatna	-:Lasso	1.133	0.00
det:Hrbatna	-:Lasso	-2.141	0.00
mh:Hrbatna	-:Lasso	0.092	0.9957
mh:Lasso	det:Lasso	0.544	0.00
-:Hrbatna	det:Lasso	1.670	0.00
det:Hrbatna	det:Lasso	-1.604	0.00
mh:Hrbatna	det:Lasso	0.629	0.00
-:Hrbatna	mh:Lasso	1.126	0.00
det:Hrbatna	mh:Lasso	-2.148	0.00

Kombinacije razina faktora		diff	p
mh:Hrbatna	mh:Lasso	0.085	0.9976
det:Hrbatna	-:Hrbatna	-3.274	0.00
mh:Hrbatna	-:Hrbatna	-1.040	0.00
mh:Hrbatna	det:Hrbatna	2.234	0.00

Procjena penalizacijskog koeficijenta ( -: nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

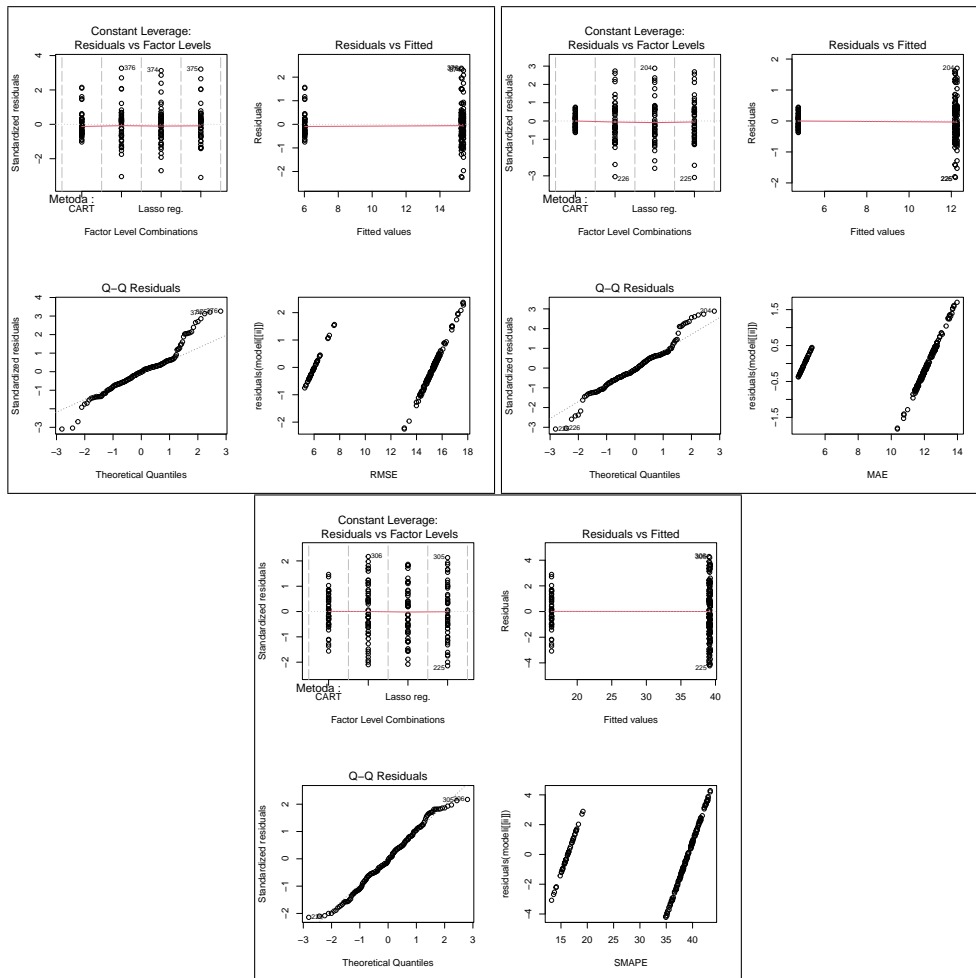
Tablica 7.9: Scenarij 2, Hipoteza 1, korigirani koeficijent determinacije - *post hoc* analiza Tukey HSD testom (\*: statistički značajno različit od 0 na razini statističke značajnosti 0.05.)

Kombinacije razina faktora		diff	p
det:El. Mreža	-:El. Mreža	0.028*	0.0001
mh:El. Mreža	-:El. Mreža	0.003	0.9999
-:Lasso	-:El. Mreža	0.006	0.9908
det:Lasso	-:El. Mreža	0.017	0.1018
mh:Lasso	-:El. Mreža	-0.002	1.0000
-:Hrbatna	-:El. Mreža	-0.012	0.4788
det:Hrbatna	-:El. Mreža	0.026*	0.0006
mh:Hrbatna	-:El. Mreža	-0.003	0.9997
mh:El. mreža	det:El. mreža	-0.025*	0.0007
-:Lasso	det:El. mreža	-0.023*	0.0037
det:Lasso	det:El. mreža	-0.012	0.5713
mh:Lasso	det:El. mreža	-0.030*	0.0000
-:Hrbatna	det:El. mreža	-0.041*	0.0000
det:Hrbatna	det:El. mreža	-0.003	0.9999
mh:Hrbatna	det:El. mreža	-0.032*	0.0000
-:Lasso	mh:El. mreža	0.002	1.0000
det:Lasso	mh:El. mreža	0.014	0.3218

## Prilog 3 - Scenarij 2

Kombinacije razina faktora		diff	<i>p</i>
mh:Lasso	mh:El. mreža	-0.005	0.9954
-:Hrbatna	mh:El. mreža	-0.015	0.1832
det:Hrbatna	mh:El. mreža	0.023*	0.0047
mh:Hrbatna	mh:El. mreža	-0.007	0.9732
det:Lasso	-:Lasso	0.011	0.5975
mh:Lasso	-:Lasso	-0.007	0.9424
-:Hrbatna	-:Lasso	-0.018	0.0643
det:Hrbatna	-:Lasso	0.020*	0.0201
mh:Hrbatna	-:Lasso	-0.009	0.8458
mh:Lasso	det:Lasso	-0.019*	0.0412
-:Hrbatna	det:Lasso	-0.029*	0.0000
det:Hrbatna	det:Lasso	0.009	0.8610
mh:Hrbatna	det:Lasso	-0.020*	0.0180
-:Hrbatna	mh:Lasso	-0.010	0.7018
det:Hrbatna	mh:Lasso	0.028*	0.0001
mh:Hrbatna	mh:Lasso	-0.002	1.0000
det:Hrbatna	-:Hrbatna	0.038*	0.0000
mh:Hrbatna	-:Hrbatna	0.009	0.8511
mh:Hrbatna	det:Hrbatna	-0.029*	0.0000

Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

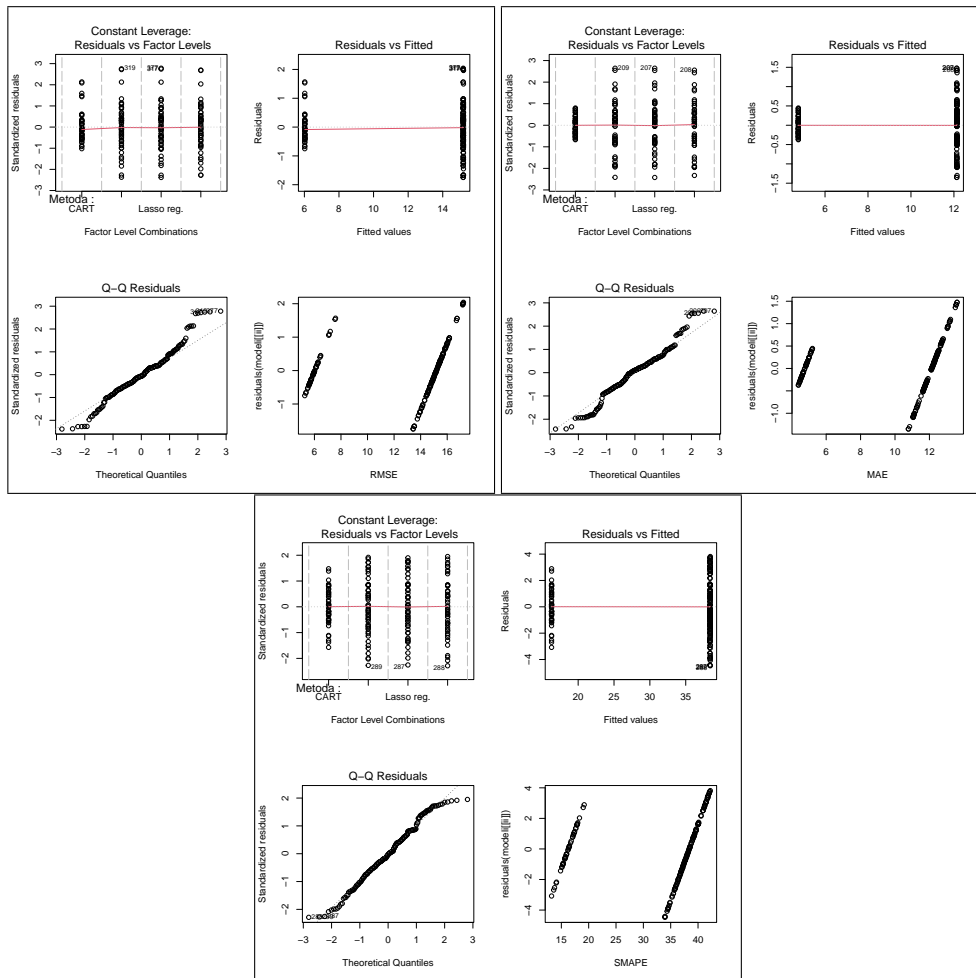


Slika 7.11: Scenarij 2, Hipoteza 2a - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Tablica 7.10: Scenarij 2, Hipoteza 2a - procjene koeficijenata i intervali pouzdanosti (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije za svaku mjeru evaluacije

<b>Mjera evaluacije</b>	<b>Značajka</b>	<b>Kontrast</b>	<b>Koeficijent</b>	<b>2.50%</b>	<b>97.50%</b>
<i>RMSE</i>	(Slobodni koeficijent)	-	12.993*	12.890	13.096
	Metoda	c31	-6.970*	-7.148	-6.793
		c32	-0.055	-0.222	0.113
		c33	0.044	-0.101	0.189
<i>MAE</i>	(Slobodni koeficijent)	-	10.349*	10.266	10.432
	Metoda	c31	-5.590*	-5.734	-5.446
		c32	-0.043	-0.179	0.093
		c33	0.038	-0.079	0.156
<i>SMAPE</i>	(Slobodni koeficijent)	-	33.444*	33.167	33.722
	Metoda	c31	-17.152*	-17.633	-16.671
		c32	-0.085	-0.539	0.368
		c33	0.042	-0.351	0.435

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2a. U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.



Slika 7.12: Scenarij 2, Hipoteza 2b - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

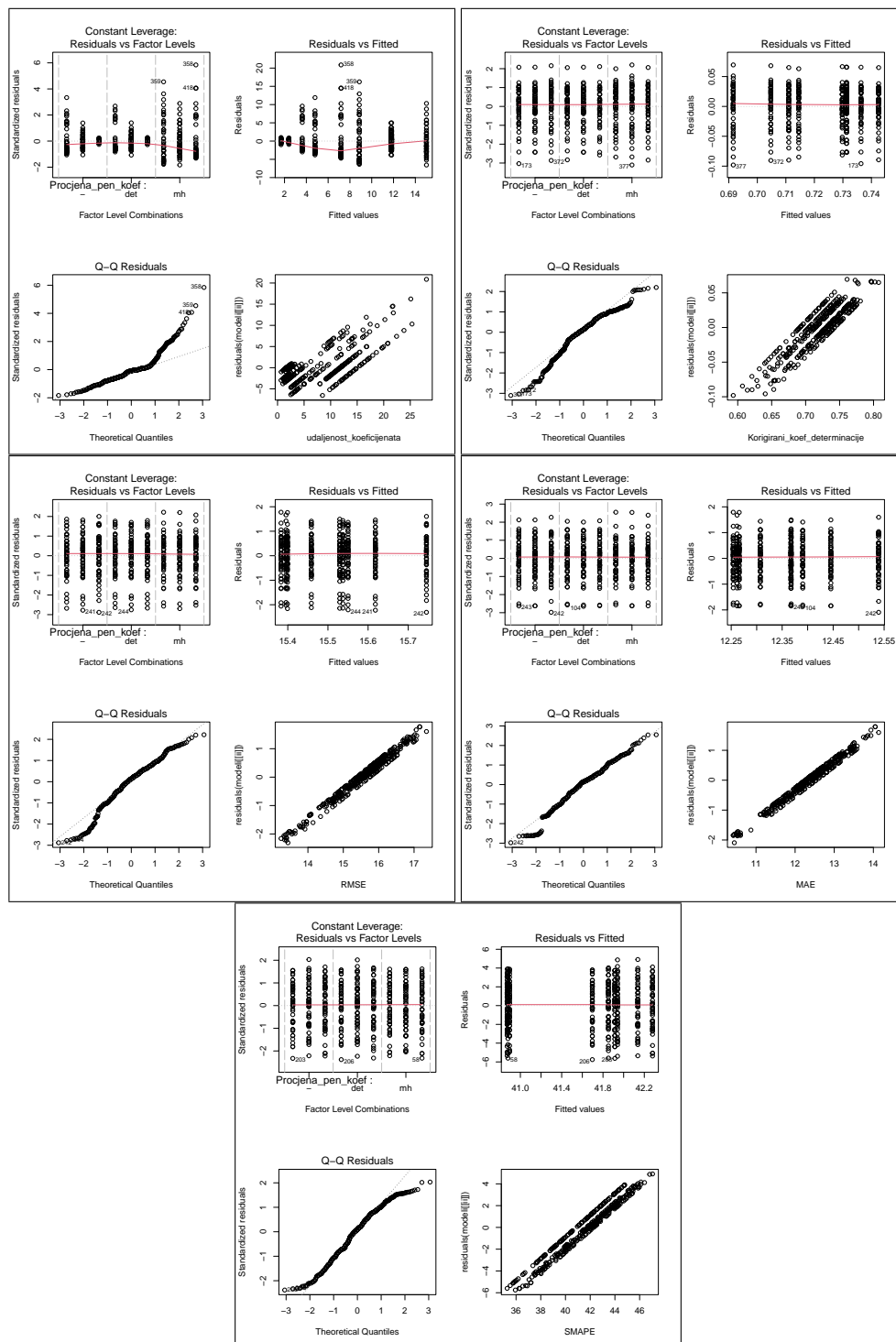


Tablica 7.11: Scenarij 2, Hipoteza 2b - procjene koeficijenata i intervali pouzdanosti (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije za svaku mjeru evaluacije

<b>Mjera evaluacije</b>	<b>Značajka</b>	<b>Kontrast</b>	<b>Koeficijent</b>	<b>2.50%</b>	<b>97.50%</b>
<i>RMSE</i>	(Slobodni koeficijent)	-	12.897*	12.794	13.000
	Metoda	c31	-6.874*	-7.053	-6.695
		c32	-0.006	-0.175	0.163
		c33	-0.007	-0.154	0.139
<i>MAE</i>	(Slobodni koeficijent)	-	10.292*	10.213	10.370
	Metoda	c31	-5.533*	-5.669	-5.396
		c32	-0.005	-0.134	0.123
		c33	-0.008	-0.119	0.103
<i>SMAPE</i>	(Slobodni koeficijent)	-	32.873*	32.597	33.149
	Metoda	c31	-16.581*	-17.058	-16.103
		c32	0.009	-0.442	0.459
		c33	-0.027	-0.417	0.363

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2b. U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

# PRILOG 4 - SCENARIJ 3



Slika 7.13: Scenarij 3, Hipoteza 1 - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

**Prilog 4 - Scenarij 3**

Tablica 7.12: Scenarij 3, Hipoteza 1 - procjene koeficijenata i intervali pouzdanosti  
 (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije

<b>Mjera evaluacije</b>	<b>Značajka</b>	<b>Kontrast</b>	<b>Koeficijent</b>	<b>2.50%</b>	<b>97.50%</b>
Udaljenost koeficijenata	(Slobodni koeficijent)	-	7.463*	7.129	7.798
	Procjena pen. koeficijenta	c11	-1.142*	-1.614	-0.669
		c12	2.322*	1.913	2.731
	Metoda regularizacije	c21	-1.700*	-2.173	-1.227
		c22	-4.557*	-4.967	-4.148
	Procjena pen. Koeficijenta:	c11:c21	0.177	-0.492	0.845
	Metoda regularizacije	c12:c21	0.296	-0.283	0.875
		c11:c22	-0.123	-0.702	0.456
		c12:c22	0.577*	0.076	1.079
Korigirani koeficijent determinacije	(Slobodni koeficijent)	-	0.721*	0.718	0.724
	Procjena pen. koeficijenta	c11	0.003	-0.002	0.007
		c12	-0.014*	-0.018	-0.011
	Metoda regularizacije	c21	0.010*	0.006	0.014
		c22	-0.001	-0.005	0.002
	Procjena pen. Koeficijenta:	c11:c21	0.003	-0.003	0.008
	Metoda regularizacije	c12:c21	0.000	-0.005	0.006
		c11:c22	-0.012*	-0.017	-0.007
		c12:c22	0.005	0.001	0.010
<i>RMSE</i>	(Slobodni koeficijent)	-	15.513*	15.438	15.588
	Procjena pen. koeficijenta	c11	0.118*	0.012	0.224
		c12	-0.061	-0.153	0.030
	Metoda regularizacije	c21	-0.056	-0.162	0.050
		c22	0.018	-0.073	0.110
	Procjena pen. Koeficijenta:	c11:c21	-0.046	-0.196	0.104
	Metoda regularizacije	c12:c21	0.024	-0.106	0.154
		c11:c22	0.045	-0.085	0.175
		c12:c22	0.002	-0.110	0.115

## Prilog 4 - Scenarij 3

Mjera evaluacije	Značajka	Kontrast	Koeficijent	2.50%	97.50%
<i>MAE</i>	(Slobodni koeficijent)	-	12.355*	12.290	12.421
	Procjena pen. koeficijenta	c11	0.095*	0.002	0.188
		c12	-0.047	-0.128	0.033
	Metoda regularizacije	c21	-0.046	-0.139	0.047
		c22	0.011	-0.069	0.092
	Procjena pen. Koeficijenta:	c11:c21	-0.037	-0.169	0.094
	Metoda regularizacije	c12:c21	0.021	-0.093	0.135
		c11:c22	0.036	-0.078	0.150
		c12:c22	0.005	-0.094	0.103
<i>SMAPE</i>	(Slobodni koeficijent)	-	41.608*	41.382	41.834
	Procjena pen. koeficijenta	c11	0.503*	0.183	0.823
		c12	-0.474*	-0.751	-0.197
	Metoda regularizacije	c21	-0.113	-0.433	0.207
		c22	0.006	-0.270	0.283
	Procjena pen. Koeficijenta:	c11:c21	-0.083	-0.535	0.370
	Metoda regularizacije	c12:c21	0.061	-0.330	0.453
		c11:c22	0.065	-0.327	0.457
		c12:c22	0.018	-0.321	0.357

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c11 testira se Hipoteza 1 ovog istraživanja. U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

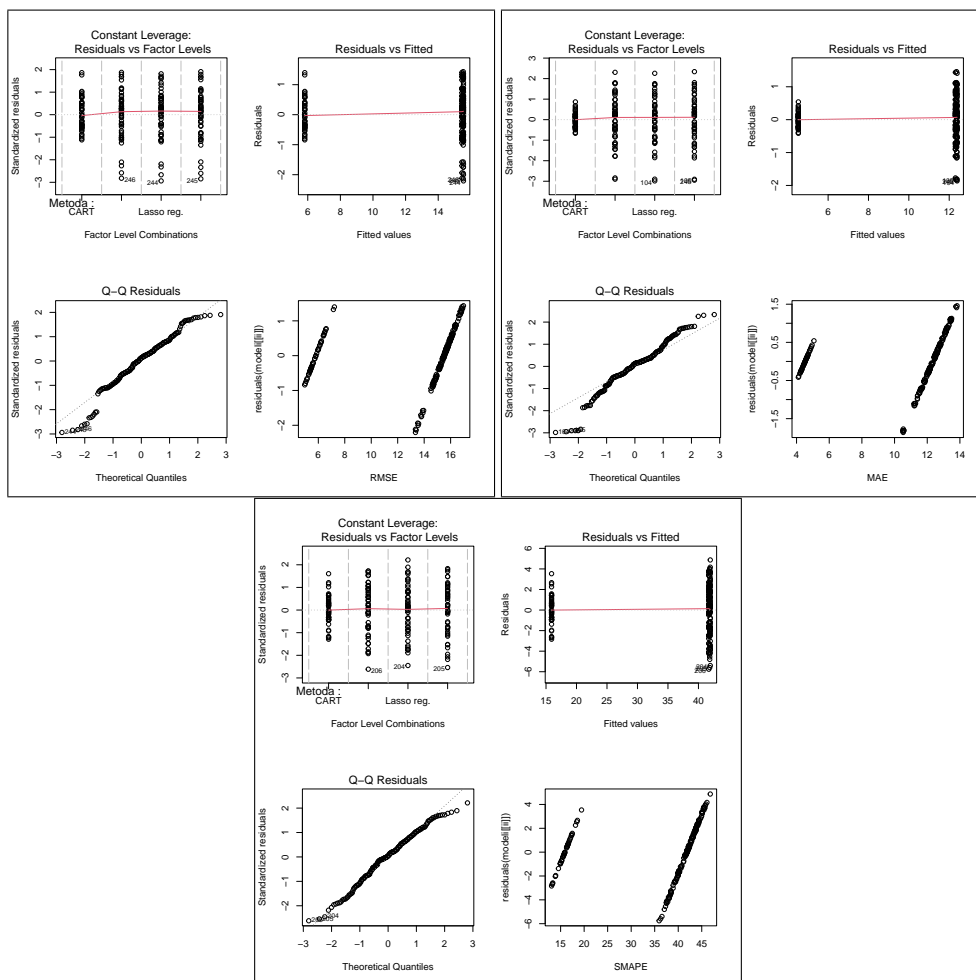
Tablica 7.13: Scenarij 3, Hipoteza 1, korigirani koeficijent determinacije - *post hoc* analiza Tukey HSD testom (\*: statistički značajno različit od 0 na razini statističke značajnosti 0.05.)

Kombinacije razina faktora	diff	<i>p</i>
det:El. Mreža -:El. Mreža	0.006	0.9885
mh:El. Mreža -:El. Mreža	-0.022*	0.0195

<b>Kombinacije razina faktora</b>		<b>diff</b>	<b>p</b>
-:Lasso	-:El. Mreža	-0.005	0.9968
det:Lasso	-:El. Mreža	-0.005	0.9972
mh:Lasso	-:El. Mreža	-0.045*	0.0000
-:Hrvatna	-:El. Mreža	-0.032*	0.0000
det:Hrvatna	-:El. Mreža	-0.007	0.9827
mh:Hrvatna	-:El. Mreža	-0.025*	0.0027
mh:El. mreža	det:El. mreža	-0.028*	0.0005
-:Lasso	det:El. mreža	-0.011	0.7028
det:Lasso	det:El. mreža	-0.011	0.7115
mh:Lasso	det:El. mreža	-0.051*	0.0000
-:Hrvatna	det:El. mreža	-0.038*	0.0000
det:Hrvatna	det:El. mreža	-0.013	0.5452
mh:Hrvatna	det:El. mreža	-0.032*	0.0000
-:Lasso	mh:El. mreža	0.017	0.1819
det:Lasso	mh:El. mreža	0.017	0.1765
mh:Lasso	mh:El. mreža	-0.023*	0.0106
-:Hrvatna	mh:El. mreža	-0.010	0.8391
det:Hrvatna	mh:El. mreža	0.015	0.2928
mh:Hrvatna	mh:El. mreža	-0.004	0.9998
det:Lasso	-:Lasso	0.000	1.0000
mh:Lasso	-:Lasso	-0.040*	0.0000
-:Hrvatna	-:Lasso	-0.027*	0.0013
det:Hrvatna	-:Lasso	-0.001	1.0000
mh:Hrvatna	-:Lasso	-0.020*	0.0432
mh:Lasso	det:Lasso	-0.040*	0.0000
-:Hrvatna	det:Lasso	-0.027*	0.0012
det:Hrvatna	det:Lasso	-0.002	1.0000
mh:Hrvatna	det:Lasso	-0.020*	0.0415
-:Hrvatna	mh:Lasso	0.013	0.4990
det:Hrvatna	mh:Lasso	0.038*	0.0000
mh:Hrvatna	mh:Lasso	0.019	0.0606

Kombinacije razina faktora		diff	<i>p</i>
det:Hrbatna	-:Hrbatna	0.025*	0.0032
mh:Hrbatna	-:Hrbatna	0.006	0.9871
mh:Hrbatna	det:Hrbatna	-0.019	0.0829

Procjena penalizacijskog koeficijenta (-: nije rađena, det: deterministički, mh: metaheuristika). U Hipotezi 1 uspoređuju se metode regularizacije koje uključuju domensko znanje stručnjaka s onima koji ga ne uključuju.

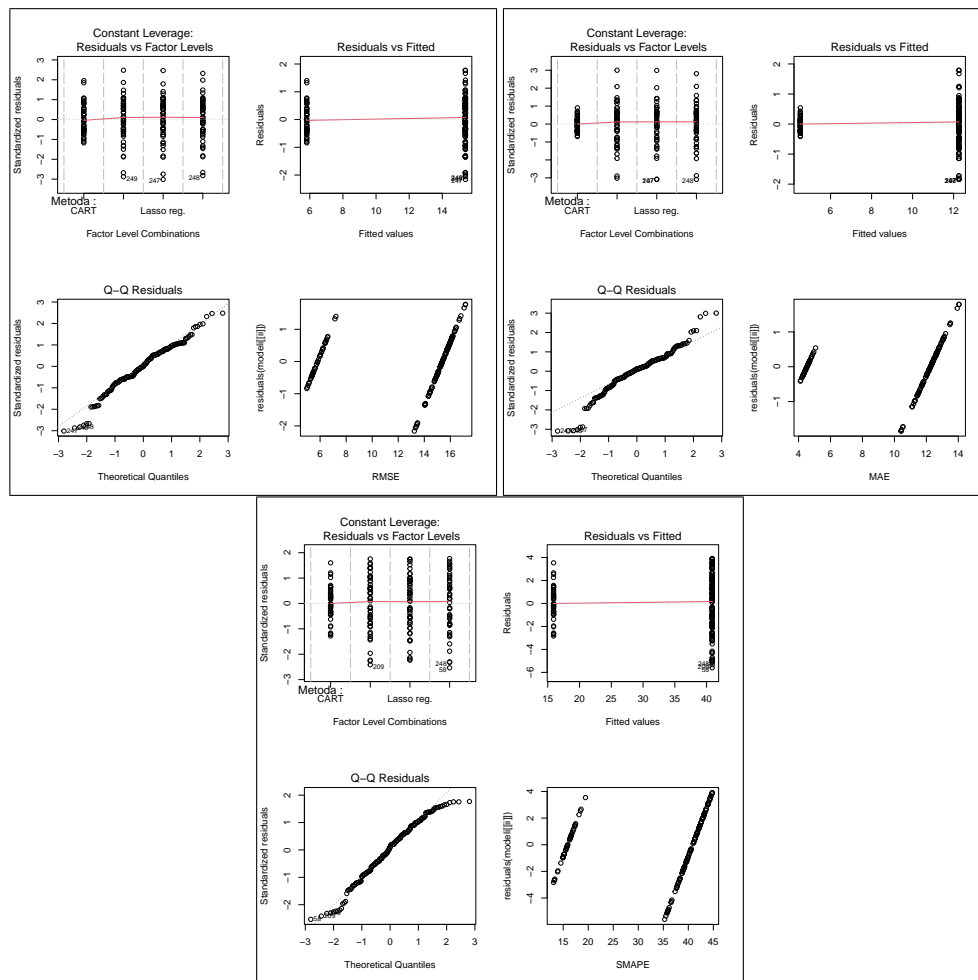


Slika 7.14: Scenarij 3, Hipoteza 2a - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.

Tablica 7.14: Scenarij 3, Hipoteza 2a - procjene koeficijenata i intervali pouzdanosti (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije

Mjera evaluacije	Značajka	Kontrast	Koeficijent	2.50%	97.50%
<i>RMSE</i>	(Slobodni koeficijent)	-	13.091*	12.985	13.196
	Metoda	c31	-7.274*	-7.457	-7.091
		c32	-0.057	-0.229	0.116
		c33	0.006	-0.143	0.156
<i>MAE</i>	(Slobodni koeficijent)	-	10.398*	10.310	10.486
	Metoda	c31	-5.872*	-6.024	-5.719
		c32	-0.048	-0.191	0.096
		c33	0.012	-0.113	0.136
<i>SMAPE</i>	(Slobodni koeficijent)	-	35.350*	35.040	35.660
	Metoda	c31	-19.439*	-19.977	-18.902
		c32	-0.133	-0.640	0.373
		c33	0.044	-0.395	0.482

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2a. U Hipotezi 2a uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka deterministički.



Slika 7.15: Scenarij 3, Hipoteza 2b - dijagnostički dijagrami za svaku mjeru evaluacije U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.



Tablica 7.15: Scenarij 3, Hipoteza 2b - procjene koeficijenata i intervali pouzdanosti (\*: 95% interval pouzdanosti ne sadrži 0) za svaku mjeru evaluacije

<b>Mjera evaluacije</b>	<b>Značajka</b>	<b>Kontrast</b>	<b>Koeficijent</b>	<b>2.50%</b>	<b>97.50%</b>
<i>RMSE</i>	(Slobodni koeficijent)	-	12.999*	12.898	13.099
	Metoda	c31	-7.182*	-7.357	-7.007
		c32	-0.009	-0.174	0.155
		c33	0.002	-0.141	0.144
<i>MAE</i>	(Slobodni koeficijent)	-	10.327*	10.243	10.411
	Metoda	c31	-5.801*	-5.946	-5.655
		c32	-0.006	-0.144	0.131
		c33	0.002	-0.117	0.121
<i>SMAPE</i>	(Slobodni koeficijent)	-	34.639*	34.328	34.951
	Metoda	c31	-18.729*	-19.268	-18.189
		c32	-0.010	-0.519	0.498
		c33	0.008	-0.432	0.448

Pojašnjenja kontrasta dostupna su u potpoglavlju 4.5 (str. 50). Pomoću kontrasta c31 testira se Hipoteza 2b. U Hipotezi 2b uspoređuju se metoda CART i metode regularizacije koje uključuju domensko znanje stručnjaka metaheuristikom.

# BIBLIOGRAFIJA

- [1] D. Moise and D. Shestakov, “Terabyte-scale image similarity search,” in *Handbook of Statistics*, vol. 33, pp. 279–301, Elsevier, 2015. ↑ 1.
- [2] Statista, “Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025.” <https://www.statista.com/statistics/871513/worldwide-data-created/>, 2024. [Datum pristupa: 22.4.2024.]. ↑ 1.
- [3] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, “Feature selection: A data perspective,” *ACM Computing surveys*, vol. 50, no. 6, pp. 1–45, 2017. ↑ 1, 6, 7, 8, 26.
- [4] V. Bolón-Canedo, N. Sánchez-Marroño, and A. Alonso-Betanzos, “Feature selection for high-dimensional data,” *Progress in Artificial Intelligence*, vol. 5, pp. 65–75, 2016. ↑ 1, 6, 8, 9, 27.
- [5] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction (2nd ed.)*. New York: Springer, 2009. ↑ 1, 6, 7, 9, 10, 11, 12, 24, 49, 50.
- [6] L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” in *Proceedings of the 20th International conference on machine learning (ICML-03)*, pp. 856–863, 2003. ↑ 1, 7, 26.
- [7] Ekonomski fakultet u Osijeku, “Metodološki okvir za učinkovito upravljanje energijom s pomoću inteligentne podatkovne analitike.” <http://merida.efos.hr/>, 2017. [Datum pristupa: 6.12.2023.]. ↑ 2, 40, 43.

- [8] Europski parlament i Vijeće, “Direktiva 2012/27/EU Europskog parlamenta i Vijeća od 25. listopada 2012. o energetske učinkovitosti, izmjeni direktiva 2009/125/EZ i 2010/30/EU i stavljanju izvan snage direktiva 2004/8/EZ i 2006/32/EZ,” *Službeni list Europske unije*, vol. 315, pp. 1–56, 2012. ↑ 2, 3.
- [9] M. Mokriš, “The independent component analysis with the linear regression—predicting the energy costs of the public sector buildings in Croatia,” *Croatian Operational Research Review*, vol. 13, no. 2, pp. 173–185, 2022. ↑ 2, 26, 33, 34.
- [10] Europski parlament i Vijeće, “Direktiva (EU) 2018/2002 Europskog parlamenta i Vijeća od 11. prosinca 2018. o izmjeni Direktive 2012/27/EU o energetske učinkovitosti,” *Službeni list Europske unije*, vol. 328, pp. 210–230, 2018. ↑ 2.
- [11] H. Tommerup, J. Rose, and S. Svendsen, “Energy-efficient houses built according to the energy performance requirements introduced in Denmark in 2006,” *Energy and Buildings*, vol. 39, no. 10, pp. 1123–1130, 2007. ↑ 2.
- [12] Europski parlament i Vijeće, “Direktiva (EU) 2018/844 Europskog parlamenta i Vijeća od 30. svibnja 2018. o izmjeni Direktive 2010/31/EU o energetske svojstvima zgrada i Direktive 2012/27/EU o energetske učinkovitosti,” *Službeni list Europske unije*, vol. 156, pp. 75–91, 2018. ↑ 2, 3.
- [13] Europski parlament i Vijeće, “Direktiva 2010/31/EU Europskog parlamenta i Vijeća od 19. svibnja 2010. o energetske učinkovitosti zgrada (preinaka),” *Službeni list Europske unije*, vol. 153, pp. 13–35, 2010. ↑ 3.
- [14] Fond za zaštitu okoliša i energetske učinkovitost, “Energetska obnova javnih zgrada.” <https://www.fzoeu.hr/hr/energetska-obnova-javnih-zgrada/7699>, 2023. [Datum pristupa: 7.12.2023.]. ↑ 3.
- [15] N. Qazi and B. W. Wong, “An interactive human centered data science approach towards crime pattern analysis,” *Information Processing & Management*, vol. 56, no. 6, p. 102066, 2019. ↑ 4, 28.
- [16] L. Micallef, I. Sundin, P. Marttinen, M. Ammad-Ud-Din, T. Peltola, M. Soare, G. Jaccucci, and S. Kaski, “Interactive elicitation of knowledge on feature relevance improves

- predictions in small data sets,” in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pp. 547–552, 2017. ↑ 4, 28.
- [17] P. Daeë, T. Peltola, M. Soare, and S. Kaski, “Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction,” *Machine Learning*, vol. 106, pp. 1599–1620, 2017. ↑ 4, 28.
- [18] F. Akulich, H. Anahideh, M. Sheyyab, and D. Ambre, “Explainable predictive modeling for limited spectral data,” *Chemometrics and Intelligent Laboratory Systems*, vol. 225, p. 104572, 2022. ↑ 4, 29.
- [19] M. Ringsquandl, S. Lamparter, S. Brandt, T. Hubauer, and R. Lepratti, “Semantic-guided feature selection for industrial automation systems,” in *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II 14*, pp. 225–240, Springer, 2015. ↑ 4, 29.
- [20] G. Vandewiele, “Enhancing white-box machine learning processes by incorporating semantic background knowledge,” in *The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28–June 1, 2017, Proceedings, Part II 14*, pp. 267–278, 2017. ↑ 4, 29.
- [21] Jenul, Anna and Schrunner, Stefan and Pilz, Jürgen and Tomic, Oliver, “A user-guided bayesian framework for ensemble feature selection in life science applications (UBayFS),” *Machine Learning*, vol. 111, no. 10, pp. 3897–3923, 2022. ↑ 4, 30.
- [22] Y. Liu, J.-M. Wu, M. Avdeev, and S.-Q. Shi, “Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties,” *Advanced Theory and Simulations*, vol. 3, no. 2, p. 1900215, 2020. ↑ 4, 30.
- [23] S. El-Sappagh, F. Ali, A. Ali, A. Hendawi, F. A. Badria, and D. Y. Suh, “Clinical decision support system for liver fibrosis prediction in hepatitis patients: A case comparison of two soft computing techniques,” *IEEE Access*, vol. 6, pp. 52911–52929, 2018. ↑ 4, 30.
- [24] Y. Zhao, X. Yang, C. Bolnykh, S. Harenberg, N. Korchiev, S. R. Yerramsetty, B. P. Vellanki, R. Kodumagulla, and N. F. Samatova, “Predictive models with end user preference,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 1, pp. 69–82, 2022. ↑ 4, 30, 31, 50.

- [25] T.-T. Huynh-Cam, V. Nalluri, L.-S. Chen, and Y.-Y. Yang, “IS-DT: a new feature selection method for determining the important features in programmatic buying,” *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 118, 2022. ↑ 4, 31.
- [26] H. Zhou, K.-M. Yu, and H.-P. Hsu, “Hybrid modeling method for soft sensing of key process parameters in chemical industry,” *Sensors & Materials*, vol. 33, 2021. ↑ 4, 31.
- [27] S. Ghiasi, A. Patane, A. Greco, L. Laurenti, E. P. Scilingo, and M. Kwiatkowska, “Gaussian processes with physiologically-inspired priors for physical arousal recognition,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 54–57, IEEE, 2020. ↑ 4, 31.
- [28] O. Lohaj, J. Paralič, P. Bednár, Z. Paraličová, and M. Huba, “Unraveling COVID-19 dynamics via machine learning and XAI: Investigating variant influence and prognostic classification,” *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1266–1281, 2023. ↑ 4, 31.
- [29] P. R. Krishna and P. Rajarajeswari, “Improved swarm intelligence optimization model with mutual information estimation for feature selection in microarray bioinformatics datasets for diseases diagnosis,” *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 23, no. 2, pp. 117–129, 2023. ↑ 4, 31.
- [30] M. A. Pienaar, J. B. Sempa, N. Luwes, E. C. George, and S. C. Brown, “Elicitation of domain knowledge for a machine learning model for paediatric critical illness in South Africa,” *Frontiers in Pediatrics*, vol. 11, p. 1005579, 2023. ↑ 4, 32.
- [31] X. Chen, Y. Zeng, S. Kang, and R. Jin, “Inn: An interpretable neural network for ai incubation in manufacturing,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 5, pp. 1–23, 2022. ↑ 4, 32.
- [32] B. Pfeifer, H. Baniecki, A. Saranti, P. Biecek, and A. Holzinger, “Multi-omics disease module detection with an explainable greedy decision forest,” *Scientific Reports*, vol. 12, no. 1, p. 16857, 2022. ↑ 4, 32.
- [33] J. Sun, C. Wu, W. Peng, J. Huang, C. Han, Y. Zhu, and Y. Lyu, “Mining human preference via self-correction causal structure learning,” *Scientific Reports*, vol. 12, no. 1, p. 5051, 2022. ↑ 4, 32.

- [34] A. Livne, E. S. Tov, A. Solomon, A. Elyasaf, B. Shapira, and L. Rokach, “Evolving context-aware recommender systems with users in mind,” *Expert Systems with Applications*, vol. 189, p. 116042, 2022. ↑ 4, 32.
- [35] S. Stein, C. Leng, S. Thornton, and M. Randrianandrasana, “A guided analytics tool for feature selection in steel manufacturing with an application to blast furnace top gas efficiency,” *Computational Materials Science*, vol. 186, p. 110053, 2021. ↑ 4, 32.
- [36] J. Bazan, S. Bazan-Socha, M. Ochab, S. Buregwa-Czuma, T. Nowakowski, and M. Woźniak, “Effective construction of classifiers with the k-NN method supported by a concept ontology,” *Knowledge and Information Systems*, vol. 62, pp. 1497–1510, 2020. ↑ 4, 32.
- [37] B. D. Burns and A. P. Danyluk, “Feature selection vs theory reformulation: A study of genetic refinement of knowledge-based neural networks,” *Machine Learning*, vol. 38, pp. 89–107, 2000. ↑ 4, 32.
- [38] F. Sebastiani, “Machine learning in automated text categorization,” *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002. ↑ 4, 32.
- [39] P. Thiam, P. Bellmann, H. A. Kestler, and F. Schwenker, “Exploring deep physiological models for nociceptive pain recognition,” *Sensors*, vol. 19, no. 20, p. 4503, 2019. ↑ 4, 33.
- [40] A. N. Gorban and I. Y. Tyukin, “Blessing of dimensionality: mathematical foundations of the statistical physics of data,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2118, p. 20170237, 2018. ↑ 6.
- [41] D. L. Donoho, “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS math challenges lecture*, vol. 1, no. 2000, p. 32, 2000. ↑ 6.
- [42] A. N. Gorban, V. A. Makarov, and I. Y. Tyukin, “High-dimensional brain in a high-dimensional world: Blessing of dimensionality,” *Entropy*, vol. 22, no. 1, p. 82, 2020. ↑ 6.
- [43] J. F. Hair Jr., W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis, 7th ed.* Harlow: Pearson, 2014. ↑ 7, 23, 28, 42.

- [44] K. S. Bordens and B. B. Abbott, *Research design and methods: A process approach*. McGraw-Hill, 2002. ↑ 7, 10.
- [45] A. Field, *Discovering statistics using IBM SPSS statistics*. SAGE Publications, 2013. ↑ 7, 10.
- [46] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” *The Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004. ↑ 9, 26, 27.
- [47] H. Zou, “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006. ↑ 9, 44.
- [48] M. Benšić and N. Šuvak, “Uvod u vjerojatnost i statistiku,” *Sveučilište Josipa Jurja Strossmayera, Odjel za matematiku, Osijek*, 2014. ↑ 9, 47.
- [49] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013. ↑ 9, 21, 22, 23, 24.
- [50] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005. ↑ 10, 11, 12, 13, 27, 44, 45, 160, 161.
- [51] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996. ↑ 10, 11, 12, 27, 44, 160.
- [52] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001. ↑ 11, 44.
- [53] Y. Zhang, R. Li, and C.-L. Tsai, “Regularization parameter selections via generalized information criterion,” *Journal of the American statistical Association*, vol. 105, no. 489, pp. 312–323, 2010. ↑ 11, 12.
- [54] L. El Gueddari, C. Giliyar Radhakrishna, E. Chouzenoux, and P. Ciuciu, “Calibration-less multi-coil compressed sensing magnetic resonance image reconstruction based on OSCAR regularization,” *Journal of Imaging*, vol. 7, no. 3, p. 58, 2021. ↑ 12.

- [55] Y. Sertdemir, H. R. Burgut, Z. N. Alparslan, I. Unal, and S. Gunasti, “Comparing the methods of measuring multi-rater agreement on an ordinal rating scale: a simulation study with an application to real data,” *Journal of Applied Statistics*, vol. 40, no. 7, pp. 1506–1519, 2013. ↑ 13.
- [56] K. P. Nelson and D. Edwards, “Measures of agreement between many raters for ordinal classifications,” *Statistics in medicine*, vol. 34, no. 23, pp. 3116–3132, 2015. ↑ 13, 14.
- [57] A. A. Mitani, P. E. Freer, and K. P. Nelson, “Summary measures of agreement and association between many raters’ ordinal classifications,” *Annals of epidemiology*, vol. 27, no. 10, pp. 677–685, 2017. ↑ 13, 14, 15.
- [58] A. Mitani, “AyaMitani/modelkappa: v1.” Zenodo, 2019. doi:10.5281/zenodo.3546381. ↑ 14.
- [59] K. P. Nelson and D. Edwards, “A measure of association for ordered categorical data in population-based studies,” *Statistical methods in medical research*, vol. 27, no. 3, pp. 812–831, 2018. ↑ 14.
- [60] K. A. Hallgren, “Computing inter-rater reliability for observational data: an overview and tutorial,” *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, p. 23, 2012. ↑ 14.
- [61] T. K. Koo and M. Y. Li, “A guideline of selecting and reporting intraclass correlation coefficients for reliability research,” *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016. ↑ 14, 15.
- [62] M. G. Kendall and B. B. Smith, “The problem of  $m$  rankings,” *The Annals of Mathematical Statistics*, vol. 10, no. 3, pp. 275–287, 1939. ↑ 14.
- [63] M. G. Kendall, *Rank correlation methods*. London: Griffin, 1948. ↑ 14.
- [64] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures, 2nd ed.* Boca Raton: Chapman & Hall/CRC, 2000. ↑ 14, 15, 46.
- [65] K. Krippendorff, *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage publications, 1980. ↑ 15.



- [66] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, pp. 159–174, 1977. ↑ 15.
- [67] E.-G. Talbi, *Metaheuristics: from design to implementation*. John Wiley & Sons, 2009. ↑ 16, 17, 18.
- [68] M. Dorigo and T. Stützle, *Ant colony optimization*. Bradford Books, MIT Press, 2004. ↑ 16.
- [69] J. Ocampo, C. Acosta, and H. Millwater, “High performance computing implementation on a risk assessment code,” in *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, p. 181, 2010. ↑ 16.
- [70] M. E. B. J. P. Siarry, *Metaheuristics for Machine Learning: New Advances and Tools*. Springer Nature, 2023. ↑ 19.
- [71] A. M. Usman, U. K. Yusof, and S. Naim, “Filter-based multi-objective feature selection using NSGA III and cuckoo optimization algorithm,” *IEEE Access*, vol. 8, pp. 76333–76356, 2020. ↑ 19, 26, 27.
- [72] K. Liang, W. Dai, and R. Du, “A feature selection method based on improved genetic algorithm,” in *Proceedings of the Global Reliability and Prognostics and Health Management (PHM-Shanghai)*, pp. 1–5, 2020. ↑ 19, 27.
- [73] M. Tubishat, M. Alswaitti, S. Mirjalili, M. A. Al-Garadi, T. A. Rana, *et al.*, “Dynamic butterfly optimization algorithm for feature selection,” *IEEE Access*, vol. 8, pp. 194303–194314, 2020. ↑ 19, 27.
- [74] R. Sawhney and R. Jain, “Modified binary dragonfly algorithm for feature selection in human papillomavirus-mediated disease treatment,” in *Proceedings of the International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pp. 91–95, 2018. ↑ 19, 27.
- [75] K.-C. Ying, S.-W. Lin, C.-Y. Huang, M. Liu, and C.-T. Lin, “Solving the mask data preparation scheduling problem using meta-heuristics,” *IEEE Access*, vol. 7, pp. 24192–24203, 2019. ↑ 19, 27.

- [76] A. K. Mandal, R. Sen, S. Goswami, A. Chakrabarti, and B. Chakraborty, “A new approach for feature subset selection using quantum inspired owl search algorithm,” in *Proceedings of the 10th International Conference on Information Science and Technology (ICIST)*, pp. 266–273, IEEE, 2020. ↑ 19, 27.
- [77] P. Parhi, R. Bisoi, and P. K. Dash, “An integrated nature-inspired algorithm hybridized adaptive broad learning system for disease classification,” *IEEE Access*, vol. 11, pp. 31636–31656, 2023. ↑ 19, 27.
- [78] Z. Ye, Y. Xu, Q. He, M. Wang, W. Bai, and H. Xiao, “Feature selection based on adaptive particle swarm optimization with leadership learning,” *Computational Intelligence and Neuroscience*, vol. 2022, 2022. ↑ 19, 27.
- [79] X.-S. Yang and S. Deb, “Cuckoo search via Lévy flights,” in *2009 World congress on nature & biologically inspired computing (NaBIC)*, pp. 210–214, Ieee, 2009. ↑ 19, 20, 21.
- [80] X.-S. Yang, *Nature-inspired metaheuristic algorithms*. Luniver press, 2nd ed., 2010. ↑ 19, 20, 21.
- [81] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. The Wadsworth statistics / probability series, CRC, 1984. ↑ 21.
- [82] M. Zekić-Sušac, S. Mitrović, and A. Has, “Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities,” *International journal of information management*, vol. 58, p. 102074, 2021. ↑ 22, 26, 33, 34.
- [83] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation,” *PeerJ Computer Science*, vol. 7, p. e623, 2021. ↑ 23, 24.
- [84] M. Zekić-Sušac, M. Knežević, and R. Scitovski, “Modeling the cost of energy in public sector buildings by linear regression and deep learning,” *Central European Journal of Operations Research*, vol. 29, no. 1, pp. 307–322, 2021. ↑ 24, 26, 33, 34.
- [85] H. Zou, T. Hastie, and R. Tibshirani, “On the “degrees of freedom” of the lasso,” *The Annals of Statistics*, vol. 35, no. 5, pp. 2173 – 2192, 2007. ↑ 24.

- [86] M. Zekić-Sušac, A. Has, and M. Knežević, “Predicting energy cost of public buildings by artificial neural networks, cart, and random forest,” *Neurocomputing*, vol. 439, pp. 223–233, 2021. ↑ 26, 27, 33, 34.
- [87] M. Zekić-Sušac, M. Knežević, and R. Scitovski, “Deep learning in modeling energy cost of buildings in the public sector,” in *Proceedings of the 14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019)*, pp. 101–110, Cham:Springer, 2020. ↑ 26, 27, 33, 34.
- [88] A. Jović, K. Brkić, and N. Bogunović, “A review of feature selection methods with applications,” in *Proceedings of the 38th International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 1200–1205, 2015. ↑ 26.
- [89] X. Tang, M. Dong, S. Bi, M. Pei, D. Cao, C. Xie, and S. Chi, “Feature selection algorithm based on k-means clustering,” in *Proceedings of the 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 1522–1527, 2017. ↑ 26, 27.
- [90] J. G. Dy and C. E. Brodley, “Feature selection for unsupervised learning,” *Journal of machine learning research*, vol. 5, no. Aug, pp. 845–889, 2004. ↑ 26.
- [91] G. Hafeez, K. S. Alimgeer, A. B. Qazi, I. Khan, M. Usman, F. A. Khan, and Z. Wadud, “A hybrid approach for energy consumption forecasting with a new feature engineering and optimization framework in smart grid,” *IEEE Access*, vol. 8, pp. 96210–96226, 2020. ↑ 27.
- [92] Y. Fu, Z. Li, H. Zhang, and P. Xu, “Using support vector machine to predict next day electricity load of public buildings with sub-metering devices,” *Procedia Engineering*, vol. 121, pp. 1016–1022, 2015. ↑ 33, 34.
- [93] M. Azaza, A. Eskilsson, and F. Wallin, “An open-source visualization platform for energy flows mapping and enhanced decision making.,” *Energy Procedia*, vol. 158, pp. 3208–3214, 2019. ↑ 33, 34.

- [94] Z. Tonković, S. Mitrović, and M. Zekić-Sušac, “Business intelligence system for managing natural gas consumption of public buildings,” *Economic and Social Development: Book of Proceedings*, pp. 769–778, 2018. ↑ 33, 34.
- [95] W. Zhang, S. Guhathakurta, R. Pendyala, V. Garikapati, and C. Ross, “A generalizable method for estimating household energy by neighborhoods in US urban regions,” *Energy Procedia*, vol. 143, pp. 859–864, 2017. ↑ 33, 34.
- [96] B. Howard, L. Parshall, J. Thompson, S. Hammer, J. Dickinson, and V. Modi, “Spatial distribution of urban building energy consumption by end use,” *Energy and buildings*, vol. 45, pp. 141–151, 2012. ↑ 33.
- [97] F. Biessmann, B. Kamble, and R. Streblow, “An automated machine learning approach towards energy saving estimates in public buildings,” *Energies*, vol. 16, no. 19, p. 6799, 2023. ↑ 33.
- [98] N. Jurišević, D. Gordić, and A. Vukićević, “Assessment of predictive models for the estimation of heat consumption in kindergartens,” *Thermal Science*, 2022. ↑ 33.
- [99] N. Jurišević, D. Gordić, V. Vukašinić, and A. Vukićević, “Assessment of predictive models for estimation of water consumption in public preschool buildings,” *Journal of Engineering Research*, vol. 10, no. 2B, pp. 98–111, 2022. ↑ 33.
- [100] M. Ma, R. Yan, and W. Cai, “An extended STIRPAT model-based methodology for evaluating the driving forces affecting carbon emissions in existing public building sector: evidence from China in 2000–2015,” *Natural Hazards*, vol. 89, pp. 741–756, 2017. ↑ 33, 34.
- [101] H. Li, Y. Zheng, G. Gong, and H. Guo, “A simulation study on peak carbon emission of public buildings—in the case of Henan province, China,” *Sustainability*, vol. 15, no. 11, p. 8638, 2023. ↑ 33.
- [102] V. J. Reina and C. Kontokosta, “Low hanging fruit? Regulations and energy efficiency in subsidized multifamily housing,” *Energy Policy*, vol. 106, pp. 505–513, 2017. ↑ 33, 34.
- [103] S. Maki, S. Ashina, M. Fujii, T. Fujita, N. Yabe, K. Uchida, G. Ginting, R. Boer, and R. Chandran, “Employing electricity-consumption monitoring systems and integrative

- time-series analysis models: A case study in Bogor, Indonesia,” *Frontiers in energy*, vol. 12, pp. 426–439, 2018. ↑ 33.
- [104] A. Summerfield, R. Lowe, and T. Oreszczyn, “Two models for benchmarking UK domestic delivered energy,” *Building Research & Information*, vol. 38, no. 1, pp. 12–24, 2010. ↑ 33, 34.
- [105] L. G. B. Ruiz, M. P. Cuéllar, M. D. Calvo-Flores, and M. D. C. P. Jiménez, “An application of non-linear autoregressive neural networks to predict energy consumption in public buildings,” *Energies*, vol. 9, no. 9, p. 684, 2016. ↑ 33, 34.
- [106] A. Abdelaziz, V. Santos, and M. S. Dias, “Convolutional neural network with genetic algorithm for predicting energy consumption in public buildings,” *IEEE Access*, 2023. ↑ 33, 34.
- [107] S. Guzhov, E. Gasho, and V. Shepel, “Preparation of the forecasted fuel and energy balance of a boiler in an inadequate data condition,” in *IOP Conference Series: Materials Science and Engineering*, vol. 791, p. 012012, IOP Publishing, 2020. ↑ 33, 34, 35.
- [108] E. Perez-Montalvo, M.-E. Zapata-Velasquez, L.-M. Benitez-Vazquez, J.-M. Cermeno-Gonzalez, J. Alejandro-Miranda, M.-A. Martinez-Cabero, and A. de la Puente-Gil, “Model of monthly electricity consumption of healthcare buildings based on climatological variables using PCA and linear regression,” *Energy Reports*, vol. 8, pp. 250–258, 2022. ↑ 33, 34.
- [109] C. T. Cheung, K. W. Mui, and L. T. Wong, “A hybrid simulation approach to predict cooling energy demand for public housing in Hong Kong,” in *Building simulation*, vol. 8, pp. 603–611, Springer, 2015. ↑ 33.
- [110] E. Bánóczy and P. T. Szemes, “Simulation-based optimization in energy efficiency retrofit for office building,” in *2014 IEEE/SICE International Symposium on System Integration*, pp. 222–227, IEEE, 2014. ↑ 33.
- [111] H. Qin and W. Pan, “Energy use of subtropical high-rise public residential buildings and impacts of energy saving measures,” *Journal of Cleaner Production*, vol. 254, p. 120041, 2020. ↑ 33.

- [112] M. Jradi, C. Veje, and B. N. Jørgensen, “Deep energy renovation of the Mærsk office building in Denmark using a holistic design approach,” *Energy and Buildings*, vol. 151, pp. 306–319, 2017. ↑ 33.
- [113] M. Jradi, P. Lecuelle, K. M. H. Madsen, C. Veje, and B. N. Jørgensen, “Dynamic Model-Driven Energy Retrofit of Bøgevangen and Runevej Daycare Centers in Aarhus,” *Energy Procedia*, vol. 132, pp. 975–981, 2017. ↑ 33.
- [114] Z. Ding, J. Niu, S. Liu, H. Wu, and J. Zuo, “An approach integrating geographic information system and building information modelling to assess the building health of commercial buildings,” *Journal of Cleaner Production*, vol. 257, p. 120532, 2020. ↑ 33.
- [115] F. Rodrigues, A. Isayeva, H. Rodrigues, and A. Pinto, “Energy efficiency assessment of a public building resourcing a BIM model,” *Innovative Infrastructure Solutions*, vol. 5, pp. 1–12, 2020. ↑ 33.
- [116] T. Spiegelhalter, “Achieving the net-zero-energy-buildings “2020 and 2030 targets” with the support of parametric 3-D/4-D BIM design tools,” *Journal of Green Building*, vol. 7, no. 2, pp. 74–86, 2012. ↑ 33.
- [117] F. Calise, M. Dentice D’Accadia, C. Barletta, V. Battaglia, A. Pfeifer, and N. Duic, “Detailed modelling of the deep decarbonisation scenarios with demand response technologies in the heating and cooling sector: A case study for Italy,” *Energies*, vol. 10, no. 10, p. 1535, 2017. ↑ 34.
- [118] M. H. Kristensen, A. Brun, and S. Petersen, “Predicting Danish residential heating energy use from publicly available building characteristics,” *Energy and Buildings*, vol. 173, pp. 28–37, 2018. ↑ 34.
- [119] H. Son and C. Kim, “Evolutionary multi-objective optimization in building retrofit planning problem,” *Procedia Engineering*, vol. 145, pp. 565–570, 2016. ↑ 34.
- [120] Y.-K. Juan, Y.-C. Cheng, Y.-H. Perng, and D. Castro-Lacouture, “Optimal decision model for sustainable hospital building renovation—A case study of a vacant school building converting into a community public hospital,” *International journal of environmental research and public health*, vol. 13, no. 7, p. 630, 2016. ↑ 34.

- [121] R. Hu, J. Granderson, D. Auslander, and A. Agogino, "Design of machine learning models with domain experts for automated sensor selection for energy fault detection," *Applied energy*, vol. 235, pp. 117–128, 2019. ↑ 34.
- [122] H. Pruvost, A. Wilde, and O. Enge-Rosenblatt, "Ontology-based expert system for automated monitoring of building energy systems," *Journal of Computing in Civil Engineering*, vol. 37, no. 1, p. 04022054, 2023. ↑ 34.
- [123] X. Zhou, H. Du, Y. Sun, H. Ren, P. Cui, and Z. Ma, "A new framework integrating reinforcement learning, a rule-based expert system, and decision tree analysis to improve building energy flexibility," *Journal of Building Engineering*, vol. 71, p. 106536, 2023. ↑ 35.
- [124] G. Napoli, M. Bottero, G. Ciulla, F. Dell'Anna, J. R. Figueira, and S. Greco, "Supporting public decision process in buildings energy retrofitting operations: The application of a multiple criteria decision aiding model to a case study in southern Italy," *Sustainable Cities and Society*, vol. 60, p. 102214, 2020. ↑ 35.
- [125] C. Jung and J. Awad, "Sharjah sustainable city: An analytic hierarchy process approach to urban planning priorities," *Sustainability*, vol. 15, no. 10, p. 8217, 2023. ↑ 35.
- [126] Vlada Republike Hrvatske, "Odluka o donošenju Programa energetske obnove zgrada javnog sektora u razdoblju od 2016. do 2020.," *Narodne novine*, vol. 22, no. 508, 2017. ↑ 40.
- [127] Deutsche Gesellschaft für Nachhaltiges Bauen, "About DGNB." <https://www.dgnb.de/en/dgnb/about-dgnb>. [Datum pristupa: 13.12.2023.]. ↑ 43.
- [128] V. Kolesarić and J. Tomašić Humer, "Veličina učinka," *Sveučilište Josipa Jurja Strossmayera u Osijeku, Filozofski fakultet*, 2016. ↑ 46.
- [129] M. Mokriš, D. Šimić, and D. Šebalj, "R code with functions for simulating data and simulated datasets." Zenodo, 2024. doi:10.5281/zenodo.11065897. ↑ 48, 165.
- [130] Zenodo, "Terms of use." <https://about.zenodo.org/terms/>, 2021. [Datum pristupa: 2.5.2024.]. ↑ 48.

- [131] P. Offermann, O. Levina, M. Schönherr, and U. Bub, “Outline of a design science research process,” in *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, pp. 1–11, 2009. ↑ 50.
- [132] J. Lawson, *Design and Analysis of Experiments with R*. Boca Raton: CRC press, 2015. ↑ 51, 53, 54.
- [133] G. W. Oehlert, *A first course in design and analysis of experiments*. 2010. ↑ 51.
- [134] J. W. Tukey, “Comparing individual means in the analysis of variance,” *Biometrics*, pp. 99–114, 1949. ↑ 54.
- [135] Sveučilišni računski centar (Srce), “Padobran.” <https://wiki.srce.hr/display/NR/Padobran>, 2023. [Datum pristupa: 13.3.2024.]. ↑ 54.
- [136] M. Mokriš, D. Šimić, and D. Šebalj, “Survey results.” Zenodo, 2024. doi:10.5281/zenodo.11083643. ↑ 59, 165.



# ŽIVOTOPIS

Marinela Mokriš rođena je 1994. godine u Osijeku. Pohađala je I. gimnaziju u Osijeku. Godine 2012. upisuje Preddiplomski sveučilišni studij matematike na Odjelu za matematiku (današnji Fakultet primijenjene matematike i informatike) pri Sveučilištu Josipa Jurja Strossmayera u Osijeku, a 2015. godine Diplomski sveučilišni studij matematike, smjer Financijska matematika i statistika. Tijekom studija obavljala je stručnu praksu u području podatkovne analitike u tvrtci Farmeron d.o.o. Diplomirala je 2018. godine s temom "Modeliranje koncentracije peludi ambrozije vektorskim autoregresivnim modelom" te stekla titulu magistra matematike. Iste godine se zapošljava na Ekonomskom fakultetu u Osijeku kao asistent i doktorand na projektu „Projekt razvoja karijera mladih istraživača – izobrazba novih doktora znanosti” u okviru kojeg radi na istraživačkom projektu Hrvatske zaklade za znanost pod nazivom “Metodološki okvir za učinkovito upravljanje energijom s pomoću inteligentne podatkovne analitike” (MERIDA). Godine 2019. upisuje Poslijediplomski doktorski studij Informacijske znanosti na Sveučilištu u Zagrebu Fakultetu organizacije i informatike. 2022. godine je završila Program pedagoško – psihološke i didaktičko – metodičke izobrazbe na Fakultetu za odgojne i obrazovne znanosti pri Sveučilištu Josipa Jurja Strossmayera u Osijeku, a iste godine pohađala je i doktorandsku školu "*PhD School on Data Driven Decision Making and Optimization*" u organizaciji EURO-a (*Association of European Operational Research Societies*), koja se održavala na Institutu za matematiku pri Sveučilištu u Sevilli u Španjolskoj. Područje Marinelinog interesa različite su metode obrade podatka, odnosno podatkovna analitika i strojno učenje. Dosada je bila (ko)autor 9 znanstvenih radova, jednog sažetka sa skupa i dva repozitorija istraživačkih podataka. Članica je Hrvatskog društva za operacijska istraživanja (HDOI) i Hrvatskog statističkog društva (HSD).

**Popis publikacija:**

1. M. Knežević and A. Has, "Comparison of energy cost due to energy renewal of public sector buildings in Republic of Croatia", in *Proceedings of the 8th International Scientific Symposium: Economy of Eastern Croatia – Vision and Growth*, pp. 609–617, Osijek: Faculty of Economics in Osijek, 2019.
2. M. Knežević, "Independent component analysis as a dimensionality reduction method on public sector buildings energy data", in *Proceedings of the ISCCRO - International Statistical Conference in Croatia* pp. 20-20, Zagreb: Hrvatsko statističko društvo, 2020. (sažetak izlaganja sa skupa)
3. M. Zekić-Sušac, A. Has, and M. Knežević, "The Use of Discrete-Event Simulation for Business Education: Learning by Observing, Simulating and Improving", in *Research Anthology on Business and Technical Education in the Information Era*, pp. 538-555, Hershey (PA): IGI Global, 2021.
4. M. Zekić-Sušac, M. Knežević, and R. Scitovski, "Deep learning in modeling energy cost of buildings in the public sector," in *Proceedings of the 14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019)*, pp. 101–110, Cham: Springer, 2020.
5. M. Zekić-Sušac, A. Has, and M. Knežević, "Predicting energy cost of public buildings by artificial neural networks, CART, and random forest", *Neurocomputing*, vol. 439, pp. 223-233, 2021.
6. M. Zekić-Sušac, M. Knežević, and R. Scitovski, "Modeling the cost of energy in public sector buildings by linear regression and deep learning", *Central European Journal of Operations Research*, vol. 29, no. 1, pp. 307–322, 2021.
7. K. Hodak, A. Has, M. Mokriš, "CLUSTER ANALYSIS FOR PROFILING PUBLIC SECTOR BUILDINGS OF CONTINENTAL CROATIA AS A SUPPORT FOR REGIONAL DEVELOPMENT", in *Proceedings of the 11th International scientific symposium Region, Entrepreneurship, Development (RED 2022)* pp. 271-280, Osijek: Faculty of Economics in Osijek, 2022.

8. M. Mokriš, “The independent component analysis with the linear regression–predicting the energy costs of the public sector buildings in Croatia”, *Croatian Operational Research Review*, vol. 13, no. 2, pp. 173–185, 2022.
9. A. Has, K. Hodak, and M. Mokriš, "DEEP LEARNING FOR PREDICTING CORPORATE FINANCIAL DISTRESS OF CROATIAN IT COMPANIES", in *Proceedings of the 17th International Symposium on Operational Research in Slovenia SOR'23.*, pp. 321-328, Ljubljana: University of Maribor, 2023.
10. M. Mokriš, D. Šimić, and D. Šebalj, “R code with functions for simulating data and simulated datasets.” Zenodo, 2024. doi:10.5281/zenodo.11065897
11. M. Mokriš, D. Šimić, and D. Šebalj, “Survey results.” Zenodo, 2024. doi:10.5281/zenodo.11083643
12. P. Garmaz, P. Juka, M. Mokriš, K. Hodak and A. Has, „DISKRETNA SIMULACIJA RADA SAMOPOSLUŽNOG STUDENTSKOG RESTORANA U SVRHU POBOLJŠANJA UČINKOVITOSTI POSLOVANJA“, in *Zbornik studentskih radova Sveučilišta Jurja Dobrile u Puli*, vol. 3, pp. 37-57, 2024.