VILNIAUS UNIVERSITETAS
ŠIAULIŲ AKADEMIJA

INFORMACINIŲ TECHNOLOGIJŲ VALDYMO MAGISTRO STUDIJŲ PROGRAMA

IRMANTAS PILYPAS

Magistro studijų baigiamasis darbas

MAŠININIO MOKYMOSI MODELIŲ TAIKYMAS PROGNOZUOJANT REGIONŲ EKONOMINIUS RODIKLIUS NUTS 2 LYGMENIU

Darbo vadovė: doc. dr. Irma Šileikienė

Šiauliai, 2026

Studijuojančiojo, teikiančio baigiamąjį darbą, GARANTIJA

WARRANTY of Final Thesis

Vardas, pavardė *Name, Surname*	**Irmantas Pilypas**
Padalinys *Faculty*	**Šiaulių akademija** ***Šiauliai Academy***
Studijų programa *Study Programme*	**Informacinių technologijų valdymas** ***Information technology management***
Darbo pavadinimas *Thesis topic*	**Mašininio mokymosi modelių taikymas prognozuojant regionų ekonominius rodiklius NUTS 2 lygmeniu** ***Application of Machine Learning Models for Predicting Regional Economic Indicators at the NUTS 2 Level***
Darbo tipas *Thesis type*	**Baigiamasis darbas** ***Final Thesis***

Garantuoju, kad mano baigiamasis darbas yra parengtas sąžiningai ir savarankiškai, kitų asmenų indėlio į parengtą darbą nėra. Jokių neteisėtų mokėjimų už šį darbą niekam nesu mokėjęs. Šiame darbe tiesiogiai ar netiesiogiai panaudotos kitų šaltinių citatos yra pažymėtos literatūros nuorodose.

I guarantee that my thesis is prepared in good faith and independently, there is no contribution to this work from other individuals. I have not made any illegal payments related to this work. Quotes from other sources directly or indirectly used in this thesis, are indicated in literature references.

Aš, Irmantas Pilypas, pateikdamas (-a) šį darbą, patvirtinu (*pažymėti*) ☒

Embargo laikotarpis
Embargo Period

Prašau nustatyti šiam baigiamajam darbui toliau nurodytos trukmės embargo laikotarpį:
I am requesting an embargo of this thesis for the period indicated below:

☐ _____ mėnesių / *months*
(embargo laikotarpis negali viršyti 60 mėn. / *an embargo period shall not exceed 60 months*).

☒ Embargo laikotarpis nereikalingas / *no embargo requested.*

Embargo laikotarpio nustatymo priežastis / *Reason for embargo period*:

VILNIAUS UNIVERSITETO ŠIAULIŲ AKADEMIJOS DIRBTINIO INTELEKTO ĮRANKIŲ NAUDOJIMO RAŠTO DARBUOSE DEKLARACIJA

Vardas, pavardė: Irmantas Pilypas.
Studijų programa, kursas: Informacinių technologijų valdymas, 2 kursas.
Rašto darbo pavadinimas: Mašininio mokymosi modelių taikymas prognozuojant regionų ekonominius rodiklius NUTS 2 lygmeniu.
Rašto darbo tipas: Magistro baigiamasis darbas.
Deklaracijos pateikimo data: 2025-12-13

Patvirtinu, kad vertinimui pateiktas savarankiškai atliktas rašto darbas yra parengtas laikantis Vilniaus universiteto Akademinės etikos kodekso[1], Vilniaus universiteto studijuojančiųjų rašto darbų rengimo, gynimo ir kaupimo nuostatai[2], Vilniaus universiteto Šiaulių akademijos rašto darbų rengimo ir gynimo tvarkos[3] ir Dirbtinio intelekto naudojimo Vilniaus universitete gairių[4].
Pažymiu, kad rašto darbo rengimui *buvo naudojami* / *nebuvo naudojami (pabraukti tinkamą variantą)* generatyvinio dirbtinio intelekto (toliau - DI) įrankiai.
Tuo atveju, jei darbo rengimui buvo pasitelkti DI įrankiai, pateikiu tikslią ir išsamią informaciją apie jų naudojimo tikslus ir apimtį:

1. DI įrankiai:

- Naudoti įrankiai (pvz.: Connected Papers, Midjourney, GitHub Copilot ar kiti, *įrašyti*):

  ChatGPT (OpenAI), Anthropic (Claude).

- Naudojimo tikslas (duomenų rinkimas, redagavimas, vertimas, priedų generavimas ar kitas, *įrašyti*):
  Santraukų teksto vertimui ir tobulinimui, 17 priedo generavimui iš duomenų rinkinio pirmo stulpelio [geo].

2. Esminiai DI panaudojimo atvejai:

Darbo dalis/vieta (puslapis, pastraipa)	Naudojimo tikslas	Užklausų pavyzdžiai
Summary (11 psl.)	Santraukos vertimas į anglų kalbą.	*"Translate the abstract into formal academic English while preserving the original meaning and terminology."*
Santrauka ir Summary (10-11 psl.)	Santraukų akademinio stiliaus tobulinimas ir terminijos suderinimas lietuvių ir anglų kalbomis.	*"Improve the academic style of this abstract without changing its meaning."*
17 priedas (127-131 psl.)	17 priedo sudarymas iš įvesties duomenų failo	*"Extract unique NUTS 2 region codes from the [geo] column and*

[1] Vilniaus universiteto akademinės etikos kodeksas, https://www.vu.lt/site_files/Akademines_etikos_kodeksas_suvestine_redakcija.pdf
[2] Vilniaus universiteto studijuojančiųjų rašto darbų rengimo, gynimo ir kaupimo nuostatai, 2024-05-21_Studijuojančiųjų_RD_rengimo_gynimo_ir_kaupimo_nuostatai.pdf
[3] Vilniaus universiteto Vilniaus universiteto Šiaulių akademijos rašto darbų rengimo ir gynimo tvarka, https://www.sa.vu.lt/external/sa/files/Studiju_skyrius/VUSA_Rasto_darb%C5%B3_rengimo_ir_gynimo_tvarka_2024.pdf
[4]Dirbtinio intelekto naudojimo Vilniaus universitete gairės, https://www.vu.lt/site_files/SPN-54_2024_priedas.pdf

	„*Įvesties_duomenų_nuts2_rodikliai_2025-11-04.csv*“, išgaunant NUTS 2 regionų kodus iš pirmojo stulpelio [geo] ir priskiriant atitinkamas šalis pagal oficialią NUTS klasifikaciją, nekeičiant ekonominių duomenų ir negeneruojant naujų analitinių reikšmių.	*assign corresponding countries based on official NUTS classification, without modifying economic data or generating new analytical values.”*

3. Savarankiškumas ir autorystė:

   - Patvirtinu, kad esu atsakingas(-a) už rašto darbo savarankiškumą ir autorystę: visos DI sugeneruotos rašto darbo dalys mano atidžiai patikrintos, patikslintos, pataisytos ir pažymėtos literatūros nuorodose. Darbas parengtas užtikrinant jame pateikiamų DI įrankiais sugeneruotų teksto dalių ir (ar) priedų tikslumą, patikimumą ir korektiškumą;

   - Patvirtinu, kad DI įrankiai nenaudoti darbe pateikiamų duomenų, informacijos, autorių minčių klastojimui, iškraipymui ar kitokio pobūdžio klaidinančiam pateikimui.

4. Akademinė etika:

   - Patvirtinu, kad pateikta informacija apie DI įrankių naudojimą yra tiksli ir išsami;

   - Patvirtinu, kad DI įrankiai naudoti savarankiško <u>duomenų rinkimo</u>, tyrimo, analizės, <u>teksto</u> ar priedų (*pabraukti tinkamą; jei reikia, įrašyti papildomą informaciją*_____) (pa)pildymui, o ne kaip savarankiško darbo pakaitalas;

   - Suprantu, kad nenurodytas DI įrankių naudojimas gali būti laikomas draudimo plagijuoti pažeidimu, kuris numatytas Vilniaus universiteto Akademinės etikos kodekse.

TURINYS

PAVEIKSLŲ SĄRAŠAS

LENTELIŲ SĄRAŠAS

PRIEDŲ SĄRAŠAS

SANTRUMPOS

AI (*angl. Artificial Intelligence*) - dirbtinis intelektas;

BPCA (*angl. Bayesian Principal Component Analysis*) - Bajeso pagrindinių komponenčių analizė;

CPU (*angl. Central Processing Unit*) - centrinis procesorius;

CSV (*angl. Comma-Separated Values*) - kableliu atskirtų reikšmių failas;

ES (*angl. European Union*) - Europos Sąjunga;

GDP (*angl. Gross Domestic Product*) - bendrasis vidaus produktas;

GPU (*angl. Graphics processing unit*) - grafikos procesorius;

KNN (*angl. k-Nearest Neighbors*) - artimiausių kaimynų metodas;

LGBM (*angl. Light Gradient Boosting Machine*) - lengvasis gradientinis stiprinimas;

MAE (*angl. Mean Absolute Error*) - vidutinė absoliutinė paklaida;

MAR (*angl. Missing At Random*) - atsitiktinai trūkstami duomenys;

MCAR (*angl. Missing Completely At Random*) - visiškai atsitiktinai trūkstami duomenys;

MICE (*angl. Multiple Imputation by Chained Equations*) - daugiamatis duomenų užpildymas;

MNAR (*angl. Missing Not At Random*) - neatsitiktinai trūkstami duomenys;

ML (*angl. Machine Learning*) - mašininis mokymasis;

NUTS (*angl. Nomenclature of Territorial Units for Statistics*) - teritorinių statistinių vienetų nomenklatūra;

PRISMA (*angl. Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) - įrodymais pagrįsta sisteminių literatūros apžvalgų ir meta analizių rengimo metodika;

$\boldsymbol{R^2}$ (*angl. Coefficient of Determination*) - determinacijos koeficientas;

RF (*angl. Random Forest*) - atsitiktinio miško algoritmas;

RMSE (*angl. Root Mean Squared Error*) - vidutinės kvadratinės paklaidos šaknis;

sMAPE (*angl. Symmetric Mean Absolute Percentage Error*) - simetrinė vidutinė absoliučioji procentinė paklaida;

SVM (*angl. Support Vector Machine*) - atraminių vektorių metodas;

XGB (*angl. Extreme Gradient Boosting*) - ekstremalaus gradientinio stiprinimo algoritmas.

SANTRAUKA

Magistro baigiamajame darbe nagrinėjamas mašininio mokymosi modelių taikymas trūkstamoms NUTS 2 lygmens regionų ekonominių rodiklių reikšmėms prognozuoti. NUTS 2 regionų ekonominiai rodikliai yra viena pagrindinių Europos Sąjungos struktūrinių fondų paskirstymo ir regioninės politikos formavimo priemonių, tačiau sistemingai pasitaikančios trūkstamos reikšmės duomenų rinkiniuose apsunkina objektyvų regionų palyginimą ir pagrįstų politinių sprendimų priėmimą. Darbo tikslas - pritaikyti ir optimizuoti Random Forest ir XGBoost mašininio mokymosi modelius ekonominių rodiklių prognozavimui, įvertinant jų veiksmingumą esant skirtingiems duomenų trūkstamumo scenarijams.

Tyrime taikyta sisteminė literatūros analizė pagal *PRISMA* metodologiją, trūkstamų duomenų mechanizmo identifikavimas naudojant *Little MCAR* testą bei eksperimentinis tyrimas, paremtas sintetiniu *hold-out* testavimo metodu su trimis trūkstamumo lygiais (10 %, 20 % ir 30 %). Modelių hiperparametrų optimizavimui naudotas *RandomizedSearchCV* metodas. Suformuotas duomenų rinkinys apima 78 ekonominius rodiklius, 244 *NUTS* 2 regionus ir 25 metų laikotarpį, leidžiantį įvertinti modelių stabilumą ir tikslumą heterogeniškų trūkstamų duomenų sąlygomis.

Eksperimentinio tyrimo rezultatai parodė, kad *XGBoost* modelis nuosekliai pranoksta Random Forest visose taikytose vertinimo metrikose (n*RMSE*, n*MAE*, R^2 ir *sMAPE*). Hiperparametrų optimizavimas ženkliai pagerino *Random Forest* modelio stabilumą ir prognozavimo tikslumą, tačiau *XGBoost* modelis jau su baziniais parametrais demonstravo aukštą veikimo ir tikslumo lygį visais trūkstamumo scenarijais.

Gauti rezultatai patvirtina, kad optimizuotas *XGBoost* modelis yra tinkamiausias *NUTS* 2 regionų ekonominių rodiklių užpildymui, užtikrinantis aukštą prognozavimo tikslumą ir rezultatų stabilumą. Darbo rezultatai turi praktinę vertę Europos Sąjungos statistikos institucijoms ir regioninės politikos formuotojams, nes sudaro prielaidas patikimesniam regionų ekonominės būklės ir atsparumo vertinimui.

SUMMARY

The master's thesis analyzes the application of machine learning models for imputing missing values in economic indicators at the NUTS 2 regional level. NUTS 2 regional economic indicators serve as a fundamental instrument for the allocation of European Union structural funds and the formulation of regional policy. However, the systematic occurrence of missing values in statistical datasets impedes objective cross-regional comparisons and undermines evidence-based policy-making. The objective of this study is to implement and optimise Random Forest and XGBoost machine learning models for economic indicator imputation and to evaluate their performance across varying levels of data missingness.

The research methodology comprises a systematic literature review conducted in accordance with the PRISMA guidelines, statistical identification of the missing-data mechanism using Little's MCAR test, and an experimental framework employing a synthetic hold-out approach with three levels of missingness (10%, 20%, and 30%). Hyperparameter optimisation was performed using RandomizedSearchCV. The compiled dataset encompasses 78 economic indicators across 244 NUTS 2 regions over a 25-year period, enabling a comprehensive assessment of model stability and predictive accuracy under heterogeneous missing-data conditions.

The experimental results demonstrate that the XGBoost model consistently outperforms the Random Forest model across all evaluation metrics, including normalised Root Mean Square Error (nRMSE), normalised Mean Absolute Error (nMAE), coefficient of determination (R^2), and symmetric Mean Absolute Percentage Error (sMAPE). Although hyperparameter optimisation substantially enhances the stability and predictive accuracy of the Random Forest model, the XGBoost model achieves superior performance even with baseline parameter configurations across all missingness scenarios.

The results confirm that the optimised XGBoost model constitutes the most appropriate solution for imputing missing economic indicators at the NUTS 2 regional level, ensuring both high predictive accuracy and robust performance. These results carry practical implications for European Union statistical agencies and regional policy-makers, providing a methodological foundation for more reliable assessments of regional economic conditions and resilience.

ĮVADAS

Temos aktualumas. Regioninių ekonominių duomenų kokybė tiesiogiai lemia ekonominės politikos formavimo pagrįstumą ir regioninės raidos analizės tikslumą. NUTS 2 lygmens ekonominiai rodikliai plačiai naudojami vertinant Europos Sąjungos regionų ekonominę būklę, jų raidos dinamiką bei atsparumo pokyčius, tačiau šiuos vertinimus dažnai riboja sistemingai pasitaikančios trūkstamos reikšmės duomenų rinkiniuose. Informatikos mokslų pažanga, ypač duomenų analizės ir mašininio mokymosi srityse, sudaro prielaidas šią problemą spręsti kuriant metodus, leidžiančius prognuozuoti pilnesnius ir analitiškai patikimesnius duomenų rinkinius. Tokie duomenys yra būtini tiksliam teritorinių vienetų palyginimui, ekonominės raidos vertinimui ir atsparumo analizės pagrįstumui.

Temos mokslinis ištirtumas ir tyrimo naujumas. Nors trūkstamų ekonominių duomenų problematika yra plačiai nagrinėta mokslinėje literatūroje, ansamblinių mašininio mokymosi algoritmų taikymas NUTS 2 lygmens regionų ekonominiams rodikliams, atliekant sistemingą hiperparametrų optimizavimą ir vertinant kelis duomenų trūkstamumo scenarijus, iki šiol išlieka ribotai ištirta. Šio darbo mokslinis naujumas grindžiamas kompleksiniu metodologiniu požiūriu, apimančiu eksperimentinį mašininio mokymosi modelių vertinimą, skirtingų duomenų trūkstamumo lygių įtakos modelių tikslumui analizę bei praktinį šių metodų taikymą regioninių ekonominių rodiklių prognozavimo uždaviniui spręsti.

Tiriamoji problema. NUTS 2 regionų ekonominių rodiklių duomenų rinkiniuose sisteminis trūkstamų reikšmių buvimas trukdo tiksliai vertinti regionų ekonominę būklę, analizuoti jų ekonominį atsparumą ir atlikti patikimą lyginamąją analizę, o šią problemą dar labiau gilina empirinių tyrimų stoka, vertinančių mašininio mokymosi metodų efektyvumą užpildant įvairių tipų trūkstamus duomenis.

Tyrimo objektas - mašininio mokymosi modelių ir trūkstamų NUTS 2 regionų ekonominių rodiklių reikšmių prognozavimo tarpusavio sąryšis.

Darbo tikslas - pritaikyti ir optimizuoti mašininio mokymosi modelius trūkstamoms NUTS 2 regionų ekonominių rodiklių reikšmėms prognozuoti, įvertinant modelių tikslumą ir praktinio taikymo galimybes.

Uždaviniai tikslui pasiekti:

1. Išanalizuoti moksliniuose tyrimuose taikytus trūkstamų reikšmių užpildymo metodus.
2. Išanalizuoti NUTS 2 regionų ekonominių rodiklių duomenų rinkinius ir parengti duomenų rinkinį mašininio mokymosi modelių treniravimui.

3. Eksperimentiškai įvertinti pasirinktų mašininio mokymosi modelių veikimą skirtinguose duomenų trūkstamumo scenarijuose, taikant standartines prognozavimo tikslumo metrikas.
4. Atlikti pasirinktų mašininio mokymosi modelių hiperparametrų optimizavimą viename trūkstamumo scenarijuje ir įvertinti optimizuotų modelių veikimą taikant standartines vertinimo metrikas.

Tyrimo metodai - tyrime taikomi kiekybiniai tyrimo metodai: sisteminė literatūros analizė, statistinė trūkstamų duomenų analizė, eksperimentinis tyrimas su kontroliuojamais duomenų trūkstamumo scenarijais, mašininio mokymosi modelių treniravimas, testavimas ir validavimas.

Praktinė vertė - tyrimo rezultatai turi praktinę vertę institucijoms, analizuojančioms regionų ekonomikos raidą ir atsparumą, nes sudaro patikimą pagrindą naudoti pilnus ir analitiškai nuoseklius NUTS 2 ekonominių rodiklių duomenų rinkinius. Sukurtas metodinis sprendimas gali būti integruotas į regionų ekonomikos atsparumo tyrimus ir projektus, sudarant sąlygas tiksliau vertinti atsparumą lemiančius veiksnius ir priimti duomenimis grįstus regioninės politikos sprendimus.

Darbo rezultatų aprobavimas.

Magistro baigiamasis darbas rengiamas dalyvaujant LMT mokslininkų grupių projekte „*Regionų ekonomikos atsparumo prognozavimas taikant mašininio mokymosi algoritmus*“ (Nr. S-MIP-24-35), vykdant su projektu susijusius taikomuosius mokslinius tyrimus.

Magistrinio darbo rezultatai pristatyti mokslinėse konferencijose, skaitant mokslinius pranešimus:

1. *Didžiųjų duomenų kokybės gerinimas: atsitiktinio miško modelio taikymas trūkstamų reikšmių užpildymui*. Irmantas Pilypas. Vadovė asist. dr. Irma Šileikienė. Jaunųjų tyrėjų tarptautinė mokslinė konferencija „Jaunasis tyrėjas išmaniajai visuomenei“. Vilniaus universiteto Šiaulių akademija, 2025-05-08. [18 priedas]
2. Irmantas Pilypas, Irma Šileikienė (VU ŠA) *„Random Forest modelio taikymas užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 lygmeniu“*. Lietuvos magistrantų informatikos ir IT tyrimai. Lietuvos mokslų akademija, Vilnius, 2025-05-13. [19 priedas]
3. *Applying machine learning to solve big data quality problems: random forest model for filling missing values / Mašininio mokymosi taikymas didžiųjų duomenų kokybės problemoms spręsti: Random Forest modelis trūkstamų reikšmių užpildymui*. Irmantas Pilypas. Vad. doc. dr. Irma Šileikienė. Tarptautinė studentų mokslinė - praktinė konferencija „Verslas, naujos technologijos ir sumani visuomenė“. Šiaulių valstybinė kolegija, 2025-05-15. [20 priedas]

4. *XGBoost ir Random Forest modelių palyginimas trūkstamų ekonominių rodiklių verčių įrašymui NUTS 2 lygiu*. Irmantas Pilypas, Irma Šileikienė. „Kompiuterininkų dienos - 2025“, Vilniaus universiteto Šiaulių akademija, 2025-09-25.

Eksperimentiniai tyrimo rezultatai paskelbti recenzuojamame mokslo žurnale:

1. Pilypas, I., ir Šileikienė, I. (2025). *Atsitiktinio miško modelio taikymas užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 lygmeniu*. 167-181. https://doi.org/10.15388/LMITT.2025.20

Eksperimentiniai tyrimo rezultatai pateikti recenzuojamam mokslo žurnalui:

2. Pilypas, I., Šileikienė, I. (2025). *Comparison of XGBoost and Random Forest Models for Imputing Missing Economic Indicator Values at NUTS 2 Level.* Information Technology and Control.

Darbo struktūra. Magistro baigiamasis darbas sudarytas iš įvado, trijų skyrių, išvadų, literatūros sąrašo ir priedų.

Pirmajame skyriuje atliekama trūkstamų duomenų problematikos teorinė analizė, apimanti sisteminę literatūros apžvalgą, trūkstamų duomenų klasifikaciją ir mašininio mokymosi algoritmų taikymo teorinį pagrindimą.

Antrajame skyriuje pristatoma tyrimo metodologija, apimanti NUTS 2 klasifikacijos sistemos aprašymą, tyrimui pasirinktų ekonominių rodiklių atranką, duomenų paruošimo procesą ir trūkstamų reikšmių analizę.

Trečiajame skyriuje pateikiamas eksperimentinis tyrimas, apimantis skirtingus duomenų trūkstamumo scenarijus, mašininio mokymosi modelių hiperparametrų optimizavimą ir modelių vertinimą taikant prognozavimo tikslumo metrikas.

1. TRŪKSTAMŲ DUOMENŲ IR JŲ UŽPILDYMO METODŲ TEORINĖ ANALIZĖ

1.1 Dirbtinio intelekto ir mašininio mokymosi vaidmuo ekonominių rodiklių prognozavime

Dirbtinis intelektas ir mašininis mokymasis tampa vis svarbesni ekonominių rodiklių prognozavimo srityje. Damaševičius teigia [1], kad dirbtinis intelektas leidžia kompiuterinėms sistemoms analizuoti didelius duomenų kiekius, mokytis iš jų ir priimti sprendimus, kurie imituoja žmogaus intelektinius gebėjimus. Mašininis mokymasis, kaip dirbtinio intelekto posritis, leidžia sistemoms automatiškai tobulėti iš patirties, neprogramuojant konkrečių taisyklių [2]. Tačiau šių metodų efektyvumas tiesiogiai priklauso nuo duomenų kokybės ir pilnumo, todėl trūkstamų reikšmių problema tampa vienu svarbiausių iššūkių ekonominių rodiklių prognozavime.

Ekonominių rodiklių prognozavimo kontekste, mašininio mokymosi modeliai demonstruoja išskirtinį efektyvumą dėl gebėjimo apdoroti didelius duomenų kiekius ir aptikti sudėtingus ryšius. Liu tyrime [3], publikuotame „Expert Systems with Applications", pabrėžiama, kad yra trys pagrindinės mašininio mokymosi kategorijos ekonominėje analizėje: prižiūrimas mokymasis (angl. *supervised learning*), neprižiūrimas mokymasis (angl. *unsupervised learning*) ir skatinamasis mokymasis (angl. *reinforcement learning*) [4].

Mašininio mokymosi modelių pritaikymas ekonominių rodiklių prognozavimui reikalauja specifinio požiūrio į duomenų paruošimą ir modelių parinkimą. Reznikov ir Turlakova, savo publikacijoje [5], „Data science methods and models in modern economy" išskiria pagrindinius iššūkius, su kuriais susiduriama prognozuojant ekonominius rodiklius:

1. Duomenų sezoniškumas ir trendai;
2. Ekonominių rodiklių tarpusavio priklausomybės;
3. Išorinių veiksnių įtaka;
4. Prognozavimo metodo pasirinkimas [6].

Mašininio mokymosi privalumai ekonominių rodiklių prognozavime, lyginant su tradiciniais statistiniais metodais, pasireiškia per:

- Gebėjimą apdoroti netiesinius ryšius;
- Automatinį reikšmingų požymių išskyrimą;
- Prisitaikymą prie besikeičiančių sąlygų [7].

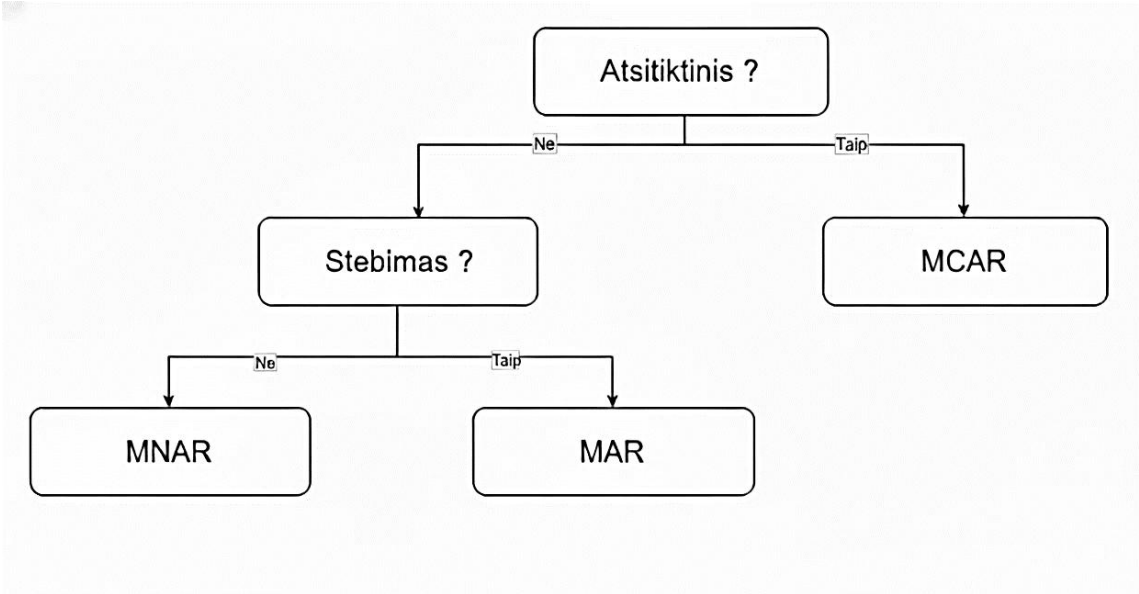
Kaip teigia Khashimova ir Buranova [8], mašininio mokymosi modelių parinkimas priklauso nuo:

- Prognozuojamų rodiklių tipo;

- Duomenų prieinamumo;
- Reikalaujamo tikslumo;
- Skaičiavimo resursų [9].

1.2 Trūkstamų duomenų samprata, klasifikacija ir reikšmė analizei

Vienas iš pagrindinių iššūkių, su kuriais susiduriama analizuojant NUTS 2 regionų ekonominius duomenis, yra trūkstamų reikšmių problema duomenų rinkiniuose. Autoriai Little ir Rubin tyrime [10] klasifikuoja trūkstamus duomenis į tris kategorijas: visiškai atsitiktinai trūkstami (*angl. MCAR - Missing Completely at Random*), atsitiktinai trūkstami (*angl. MAR - Missing at Random*) ir neatsitiktinai trūkstami (*angl. MNAR - Missing Not at Random*).

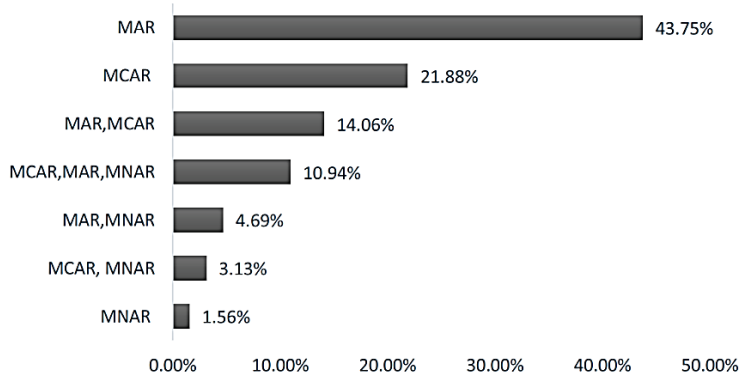


Šaltinis: sudaryta autoriaus remiantis *[11]*.

1 pav. Trūkstamų duomenų tipai

Kaip matyti 1 paveiksle, trūkstamų duomenų tipų klasifikacija priklauso nuo dviejų pagrindinių kriterijų: ar duomenų trūkumas yra atsitiktinis, ir ar jis priklauso nuo stebimų kintamųjų. MCAR atveju duomenų trūkumas yra visiškai atsitiktinis ir nepriklauso nuo jokių kitų kintamųjų. MAR tipo atveju trūkstamų duomenų tikimybė priklauso nuo stebimų kintamųjų, o MNAR atveju - nuo nestebimų kintamųjų ar pačių trūkstamų reikšmių [11].

Sisteminga mašininio mokymosi metodų analizė atskleidžia, kaip šie trūkstamų duomenų tipai yra sprendžiami praktikoje. Thomas ir Rajabi atlikta 117 tyrimų apžvalga [12] parodė, kad dauguma tyrimų (43,75%) koncentruojasi į MAR tipo duomenis, kurie yra dažniausiai pasitaikantys realiuose duomenų rinkiniuose (žiūrėti 2 pav.).



Šaltinis: Thomas & Rajabi: *[12]*

2 pav. Tyrimų pasiskirstymas pagal nagrinėjamus trūkstamų duomenų tipus

Regioninių ekonominių duomenų kontekste dažniausiai ir susiduriama su MAR tipo trūkstamais duomenimis, kai tikimybė, kad reikšmė trūks, priklauso nuo kitų stebimų kintamųjų [13]. NUTS 2 regionų duomenų rinkiniuose ši problema yra ypač aktuali dėl nevienodo duomenų rinkimo metodologijų taikymo skirtingose šalyse narėse ir regionuose. Pavyzdžiui, mažesnių ar ekonomiškai silpnesnių regionų statistinės institucijos dažnai neturi pakankamai resursų surinkti visus reikalingus duomenis, todėl susidaro sisteminis duomenų trūkumas, kuris koreliuoja su regiono ekonominio išsivystymo lygiu.

Šiame magistro darbe bus sprendžiama trūkstamų ekonominių rodiklių užpildymo problema, ieškant tinkamiausio mašininio mokymosi modelio, kuris galėtų efektyviai prognozuoti trūkstamas reikšmes atsižvelgiant į kitų ekonominių rodiklių tarpusavio ryšius ir priklausomybes. Tikimasi, kad tyrimo rezultatai leis sukurti modifikuotą mašininio mokymosi modelį, optimizuotą NUTS 2 regionų ekonominių duomenų specifikai, kuris galėtų užtikrinti aukštą prognozavimo tikslumą net esant nepilniems duomenų rinkiniams.

1.3 Mašininio mokymosi modelių taikymo trūkstamų duomenų problemai spręsti literatūros apžvalga

1.3.1 Sisteminės literatūros paieškos ir atrankos metodologija

Šiame magistro darbe mašininio mokymosi modelių taikymo trūkstamų duomenų problemai spręsti analizė grindžiama sistemine literatūros apžvalga, atlikta laikantis PRISMA (*angl. Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) metodologijos principų. PRISMA metodologija pasirinkta siekiant užtikrinti literatūros paieškos proceso skaidrumą, nuoseklumą ir pakartojamumą bei sumažinti subjektyvumo riziką atliekant mokslinių šaltinių atranką.

Literatūros analizės procesas buvo vykdomas pagal iš anksto apibrėžtą protokolą, kuriame nustatyti tyrimo klausimai, paieškos strategija, įtraukimo ir atmetimo kriterijai bei analizės eiga. Šis

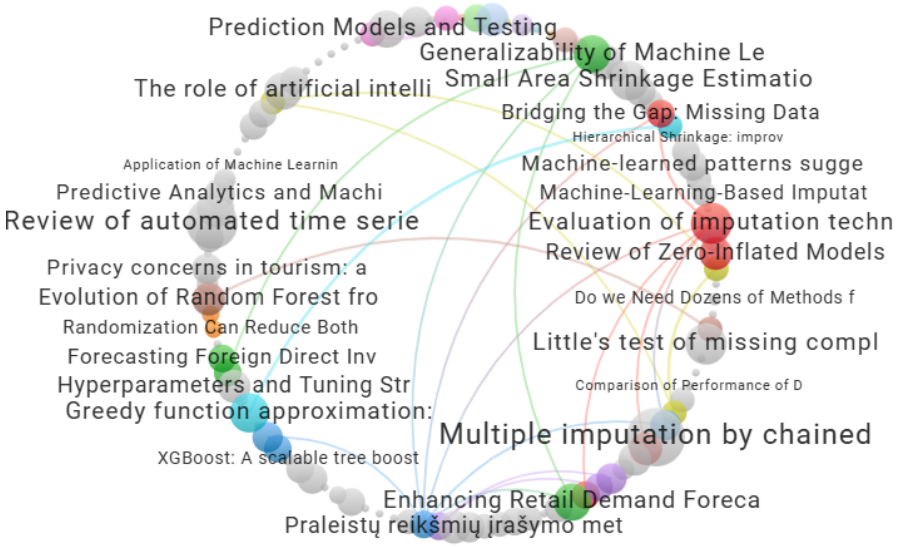
protokolas pateikiamas 1 priede , kuriame detaliai aprašyti paieškos etapai, naudotos duomenų bazės, publikacijų laikotarpis, kalbiniai apribojimai ir tematinės ribos.

Mokslinių publikacijų paieška buvo vykdoma pagrindinėse tarptautinėse mokslinėse duomenų bazėse, orientuotose į informatikos, duomenų analizės ir taikomųjų mokslų sritis. Paieškos užklausos buvo formuojamos derinant raktažodžius, susijusius su trūkstamais duomenimis (*angl. missing data, missing values, predicting, imputation*), mašininio mokymosi metodais (*machine learning, Random Forest, XGBoost, MICE* ir kt.) bei taikymo kontekstu. Detalūs paieškos rezultatai, gauti kiekvienoje duomenų bazėje ir susisteminti pagal PRISMA reikalavimus, pateikiami 2 priede.

Publikacijų atranka vyko keliais etapais: pradžioje buvo šalinami dublikatai, vėliau - atliekama pirminė atranka pagal pavadinimus ir santraukas, o galutiniame etape vertinamas pilnas straipsnių tekstas, atsižvelgiant į jų metodologinę kokybę ir atitikimą tyrimo tematikai. Galutinė į sisteminę analizę įtrauktų publikacijų apžvalga pateikiama 3 priede. Visas publikacijų atrankos procesas grafiškai apibendrintas 16 priede, mokslinių publikacijų atrankos schemoje, (*angl. PRISMA flow diagram*)[14], kurioje vizualiai atvaizduoti visi PRISMA etapai - nuo pradinės identifikacijos iki galutinio įtraukimo.

Mokslinių straipsnių kaupimui buvo naudojama bibliografinių nuorodų tvarkymo programa *Mendeley*, kuri užtikrino nuoseklų citavimo procesą viso darbo rengimo metu.

Papildomai literatūros analizės rezultatų struktūrai ir teminėms tendencijoms vizualizuoti buvo pasitelktas VOSviewer įrankis [15]. Remiantis visos sukauptos ir naudotos literatūros sąrašu, buvo sudarytas mokslinių publikacijų autorių ir raktažodžių tinklas, leidžiantis identifikuoti dominuojančias tyrimų kryptis, dažniausiai pasitaikančius terminus bei ryšius tarp autorių. Ši vizualizacija pateikiama 3 paveiksle.



Šaltinis: sudaryta autoriaus

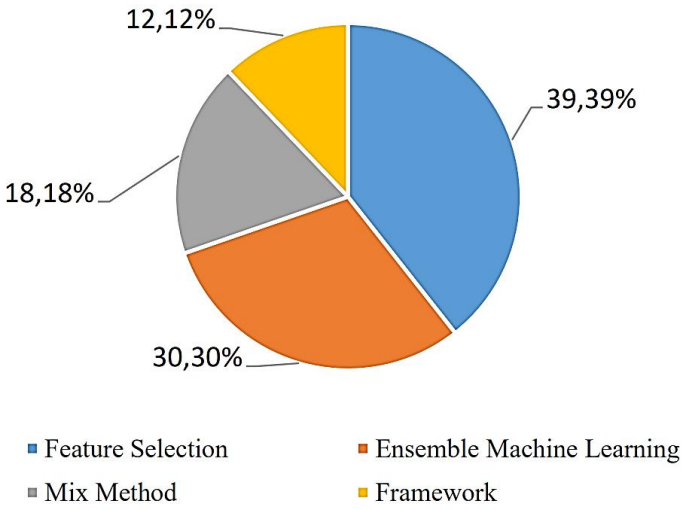
3 pav. Mokslinių publikacijų autorių ir raktažodžių ryšių tinklo vizualizacija

Analizuojamas 3 paveikslas sudarytas remiantis visa šiame darbe naudota literatūra, iš kurios buvo generuoti autorių ir raktažodžių tinklai. Toks vizualizavimo metodas leidžia ne tik identifikuoti pagrindinius trūkstamų duomenų ir mašininio mokymosi tyrimuose vyraujančius raktažodžius, bet ir atskleisti bendradarbiavimo struktūras tarp tyrėjų bei teminių klasterių formavimąsi. Pažymėtina, kad *VOSviewer* taip pat suteikia galimybę sukurti publikacijų tinklą, kuriame mazgai atitiktų atskirus straipsnius (su jų pavadinimais), o ryšiai būtų apibrėžiami bendrais autoriais arba bendrais raktažodžiais, taip dar labiau išplečiant literatūros analizės interpretacines galimybes.

Apibendrinant galima teigti, kad taikyta sisteminės literatūros analizės metodologija, paremta PRISMA gairėmis, struktūrizuotais priedais ir bibliometrine vizualizacija, sudarė patikimą pagrindą mašininio mokymosi metodų taikymo trūkstamų duomenų problemos analizei ir moksliškai pagrįstai metodų atrankai šiame magistro darbe.

1.3.2 Sisteminės literatūros analizės rezultatai ir taikomų metodų apžvalga

Atlikta išsami 2 priede pateikta mokslinių tyrimų analizė atskleidė platų mašininio mokymosi metodų spektrą, taikomą trūkstamų duomenų problemai spręsti. Po nuoseklaus publikacijų vertinimo ir atrankos proceso buvo identifikuotos vyraujančios metodologinės kryptys ir algoritmų taikymo tendencijos.



Šaltinis: [16]

4 pav. Siūlomų metodų pasiskirstymas trūkstamų duomenų tvarkymui

Kaip matyti 4 paveiksle, požymių atranka (*angl. Feature Selection*) dominuoja tarp siūlomų trūkstamų reikšmių užpildymo metodų, apimanti 39,39 % visų analizuotų sprendimų. Ansambliniai mašininio mokymosi metodai (*angl. Ensemble Machine Learning*) sudaro 30,30 %, mišrūs metodai (*angl. Mix Method*) - 18,18%, o atvirojo kodo karkasai (*angl. Framework*) - 12,12 %. Šie duomenys skritulinėje diagramoje gauti iš Setiawan ir kt. [16] atliktos sisteminės literatūros analizės, apėmusios 40 mokslinių publikacijų iš *ScienceDirect* duomenų bazės 2010 - 2023 metų laikotarpiu. Tyrimas

analizavo trūkstamų reikšmių tvarkymo metodus įvairiose srityse - socialiniuose moksluose, medicinoje, ekonomikoje kas svarbu šiame tyrime ir kitose disciplinose. Nors požymių atranka išlieka populiariausia vieninga metodologija (39,39 %), integruoti sprendimai (ansambliniai metodai, mišrūs metodai ir atvirojo kodo karkasai) kartu sudaro 60,61% visų siūlomų metodų, o tai rodo aiškią tendenciją pereiti prie hibridinių sprendimų, derinančių kelias skirtingas technikas siekiant aukštesnio trūkstamų reikšmių užpildymo tikslumo.

Ši literatūroje identifikuota tendencija pereiti prie sudėtingesnių ir integruotų trūkstamų reikšmių užpildymo sprendimų kelia papildomą praktinį iššūkį - didesnį skaičiavimo resursų poreikį. Todėl praktiniame taikyme svarbu ne tik pasiekti aukštą trūkstamų reikšmių užpildymo tikslumą, bet ir įvertinti skaičiavimo resursų poreikį, ypač dirbant su didelės apimties duomenų rinkiniais. Silkauskaitė [17] savo darbe atliko išsamų eksperimentinį vienolikos trūkstamų reikšmių užpildymo metodų palyginamąjį tyrimą, apimantį 9900 eksperimentų su skirtingais duomenų dydžiais (100 - 5000 stebėjimų), trūkstamumo lygiais (10 % - 30 %) ir duomenų trūkstamumo tipais (MCAR, MAR, MNAR). Tyrime analizuoti tiek tradiciniai daugiamačiai duomenų užpildymo metodai (*angl. Multiple Imputation by Chained Equations*) variantai su skirtingais įverčiais *(angl. estimators*), tiek sprendimų medžiais ir ansambliais grįsti metodai, tokie kaip Random Forest, XGBoost ir LightGBM, taip pat ir paprastesnis KNN metodas (*angl. k-Nearest Neighbors*).

1 lentelė. Modelių vertinimas naudojant skirtingus trūkstamų reikšmių užpildymo metodus

Metodas	MSE	R^2	Užpildymo laikas
Iterative Imputer (MICE)	0,12	0,90	1,46
Linear Regression Estimator (MICE)	0,11	0,90	0,79
Lasso Estimator (MICE)	0,12	0,89	1,28
Decision Tree Estimator (MICE)	0,26	0,77	16,62
KNN Imputation	0,25	0,78	35,98
Random Forest Imputer	0,26	0,77	22,54
Random Forest Estimator (MICE)	0,26	0,77	1050,72
XGBoost Estimator (MICE)	0,26	0,76	131,72
XGBoost Imputer	0,27	0,76	2,88
Gradient Boosting Estimator (MICE)	0,26	0,76	384,76
LGBM Estimator (MICE)	0,27	0,75	193,90

Šaltinis: sudaryta autoriaus remiantis [17].

1 lentelė atskleidžia svarbų aspektą [17] tyrime - kompromisą tarp trūkstamų reikšmių užpildymo tikslumo ir skaičiavimo efektyvumo. Daugiamatis duomenų užpildymo metodas MICE, naudojantis iteracinį trūkstamų reikšmių užpildymo principą, kai trūkstamos reikšmės pakartotinai užpildomos taikant kelis modelius, bei tiesinės regresijos įverčių metodas, pagrįstas tiesinės priklausomybės tarp kintamųjų modeliavimu (*angl. Linear Regression Estimator*), demonstruoja

aukščiausius R^2 rezultatus (0,90). Tuo tarpu ansamblinis Random Forest Estimator pasižymi didesniu skaičiavimo laiko poreikiu (1050,72 sekundžių), kuris yra būdingas sudėtingesniems, nelinijines sąsajas fiksuojantiems modeliams. Šie rezultatai atskleidžia kompromisą tarp skaičiavimo efektyvumo ir modelio struktūrinio sudėtingumo, pabrėžiant, kad didesnių skaičiavimo resursų reikalaujantys metodai gali būti ypač vertingi tais atvejais, kai siekiama geriau užfiksuoti sudėtingas duomenų priklausomybes [18].

1.3.3 Trūkstamų reikšmių užpildymo metodų įtaka prognozavimo modelių tikslumui

Trūkstamų duomenų problema ekonominiuose tyrimuose reikalauja sisteminio požiūrio į trūkstamų reikšmių užpildymo metodų parinkimą ir vertinimą, ypač vertinant jų poveikį galutinių prognozavimo modelių tikslumui, nes užpildymo kokybė tiesiogiai veikia galutinių prognozavimo modelių tikslumą. Melnyk ir Rozora [19] atliko išsamų palyginamąjį tyrimą, kuriame analizavo keturių trūkstamų reikšmių užpildymo metodų - trūkstamų reikšmių šalinimo (*angl. dropping*), vidurkio pagrindu atliekamo užpildymo (*angl. simple mean imputation*), artimiausių kaimynų metodo (*angl. nearest neighbor imputation*) ir daugiareikšmio trūkstamų reikšmių užpildymo (*angl. multiple imputation*) - poveikį skirtingų prognozavimo modelių (polinominės regresijos, Random Forest ir XGBoost Gradient Boosting) veiksmingumui.

2 lentelė. RMSE reikšmės modelių ir trūkstamų reikšmių užpildymo metodų kombinacijoms

Trūkstamų reikšmių užpildymo metodas	**Prognozavimo modelis**		
	Polynomial Regression	**Random Forest**	**XGBoost Gradient Boosting**
Dropping	204 173,64	162 626,13	164 587,21
Simple Mean Imputation	198 360,98	164 199,44	168 550,60
Nearest Neighbor Imputation	186 717,50	162 181,68	170 628,48
Multiple Imputation	172 011,71	158 998,93	160 595,07

Šaltinis: sudaryta autoriaus remiantis [19].

3 lentelė. MAE reikšmės modelių ir trūkstamų reikšmių užpildymo metodų kombinacijoms

Trūkstamų reikšmių užpildymo metodas	**Prognozavimo modelis**		
	Polynomial Regression	**Random Forest**	**XGBoost Gradient Boosting**
Dropping	204 173,64	162 626,13	164 587,21
Simple Mean Imputation	114 831,57	85 802,97	94 920,79
Nearest Neighbor Imputation	110 972,54	86 450,22	92 138,34
Multiple Imputation	108 525,78	82 961,17	87 345,78

Šaltinis: sudaryta autoriaus remiantis [19].

Tyrimo [19] rezultatai 2 ir 3 lentelėse atskleidžia aiškią hierarchiją tarp skirtingų trūkstamų reikšmių užpildymo metodų. Pašalinimo metodas (*angl. dropping*), kuris paprasčiausiai pašalina visus stebėjimus su trūkstamomis reikšmėmis, demonstruoja prasčiausius rezultatus visose modelių

kombinacijose. Tai patvirtina fundamentalią problemą - duomenų praradimas ne tik sumažina imties dydį, bet ir gali lemti sisteminį poslinkį, jei trūkstamos reikšmės pasiskirsto neatsitiktinai ir iškreipia analizuojamos imties reprezentatyvumą.

Vidurkio (*angl. simple mean*) metodas, užpildantis trūkstamas reikšmes kintamojo vidurkiu, parodo nedidelį pagerėjimą lyginant su duomenų pašalinimu, tačiau vis dar nepasiekia optimalių rezultatų. Artimiausių kaimynų algoritmas (*angl. nearest neighbors imputation algorithm*) demonstruoja tolesnį tikslumo augimą, naudodamas panašių stebėjimų informaciją trūkstamoms reikšmėms užpildyti.

Remiantis 2 ir 3 lentelėse pateiktais rezultatais, daugiareikšmis trūkstamų reikšmių užpildymo metodas (*angl. Multiple Imputation*) nuosekliai pasižymi mažiausiomis RMSE ir MAE reikšmėmis visose nagrinėtose prognozavimo modelių ir užpildymo metodų kombinacijose. Polinominės regresijos modelyje, vertinant pagal RMSE metriką (2 lentelė), paklaida sumažėja nuo 204 173,64 trūkstamų reikšmių šalinimo atveju iki 172 011,71 taikant daugiareikšmį užpildymą, kas atitinka maždaug 15,8 % prognozavimo tikslumo pagerėjimą. Random Forest modelyje daugiareikšmis trūkstamų reikšmių užpildymas pasiekia mažiausias RMSE (158 998,93; 2 lentelė) ir MAE (82 961,17; 3 lentelė) reikšmes, o XGBoost Gradient Boosting modelyje - atitinkamai RMSE = 160 595,07 (2 lentelė) ir MAE = 87 345,78 (3 lentelė).

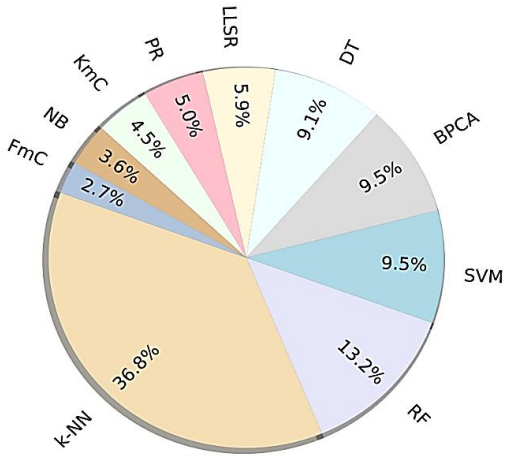
Įdomu pastebėti, kad skirtingi prognozavimo modeliai nevienodai reaguoja į trūkstamų reikšmių užpildymo metodus - Random Forest demonstruoja didžiausią atsparumą trūkstamiems duomenims net su paprastesniu užpildymu, tuo tarpu Polynomial Regression yra jautriausias užpildymo kokybei.

Apibendrinant iki šiol aptartus empirinius rezultatus ir jų interpretaciją, vertinant trūkstamų reikšmių užpildymo metodus, literatūroje taikomi du pagrindiniai trūkstamų reikšmių užpildymo kokybės vertinimo požiūriai - tiesioginis ir netiesioginis. Tiesioginis vertinimas remiasi užpildytų reikšmių palyginimu su faktinėmis (tikrosiomis) reikšmėmis, kurios dažniausiai gaunamos dirbtinai sukuriant trūkstamumo scenarijus pilnuose duomenų rinkiniuose. Šis metodas leidžia tiksliai įvertinti užpildymo paklaidą, tačiau realių duomenų analizėje jis yra riboto pritaikomumo, nes tikrosios trūkstamos reikšmės paprastai nėra žinomos.

Dėl šios priežasties praktiniuose taikymuose vis didesnę reikšmę įgauna netiesioginis užpildymo vertinimo metodas, kai užpildymo kokybė vertinama per jos poveikį tolesnėms analitinėms užduotims, tokioms kaip prognozavimas ar klasifikavimas. Šiuo atveju trūkstamų reikšmių užpildymas laikomas sėkmingas tuomet, kai galutiniai mašininio mokymosi modeliai, apmokyti su užpildytais duomenimis, pasiekia aukštą prognozavimo tikslumą ir stabilumą.

Būtent šis netiesioginis vertinimo požiūris dominuoja mokslinėje literatūroje, kurioje analizuojama, kokie mašininio mokymosi modeliai dažniausiai pasitelkiami užpildymo poveikiui

įvertinti. Šią tendenciją sistemingai nagrinėja Hasan ir kt. tyrime [20], kuriame, remiantis taip pat PRISMA metodologija, apžvelgiami 191 tyrimai ir analizuojami dažniausiai netiesioginiam trūkstamų reikšmių užpildymo vertinimui taikomi modeliai bei vertinimo metrikos. Toliau šiame poskyryje detaliau aptariami minėto tyrimo rezultatai ir jų reikšmė atliekamam magistriniam darbui.



Šaltinis: [20]

5 pav. Mašininio mokymosi modeliai netiesioginiam trūkstamų reikšmių užpildymo vertinimui

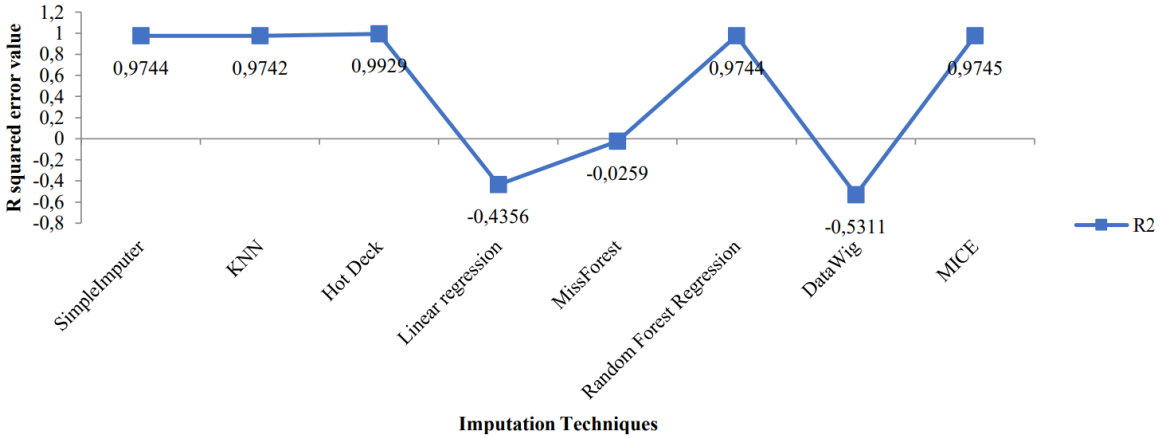
Remiantis tyrimu [20], 5 paveikslėlyje pateikta mašininio mokymosi modelių, naudojamų netiesioginiam užpildymo metodų vertinimui, pasiskirstymo skritulinė diagrama. Artimiausių kaimynų algoritmas K-NN (*angl. k-Nearest Neighbors*) dominuoja su 36,8 %, po jo seka atsitiktinio miško algoritmas (*angl. Random Forest*) su 13,2 %, atraminių vektorių algoritmas SVM (*angl. Support Vector Machine*) - su 9,5 %, bei Bajeso pagrindinių komponenčių analizės metodas BPCA (*angl. Bayesian Principal Component Analysis*) - taip pat su 9,5 %. Šie modeliai literatūroje dažniausiai naudojami kaip netiesioginiai trūkstamų reikšmių užpildymo kokybės indikatoriai, vertinant galutinių prognozavimo modelių tikslumą: kuo geresnius prognozavimo rezultatus modelis pasiekia dirbdamas su užpildytais duomenimis, tuo aukštesnė laikoma taikyto užpildymo metodo kokybė.

1.3.4 Trūkstamų reikšmių užpildymo metodų veiksmingumo palyginimas literatūroje

Trūkstamų reikšmių užpildymo metodų veiksmingumo vertinimas literatūroje rodo, kad vien statistinių tikslumo rodiklių nepakanka universaliam metodo pranašumui nustatyti, nes skirtingų metodų rezultatai reikšmingai priklauso nuo duomenų struktūros, trūkstamumo pobūdžio bei taikymo konteksto. Dėl šios priežasties vis daugiau tyrimų siekia vertinti duomenų užpildymo metodus ne tik pagal teorines savybes, bet ir pagal jų praktinį veiksmingumą realiuose duomenų rinkiniuose.

Toks taikomasis požiūris pasirinktas ir Chehal ir kt. tyrime [21], kuriame sistemingai lyginami aštuoni skirtingi trūkstamų reikšmių užpildymo metodai, taikomi realiam didelės apimties duomenų rinkiniui su trūkstamomis reikšmėmis. Straipsnyje analizuojamas įvairių trūkstamų reikšmių

užpildymo technikų poveikis mašininio mokymosi modelių prognozavimo tikslumui, o metodų veiksmingumas vertinamas pasitelkiant R^2, MSE ir MAE metrikas. Toliau šiame poskyryje aptariami minėto tyrimo metodiniai sprendimai, gauti rezultatai ir jų reikšmė trūkstamų reikšmių užpildymo metodų pasirinkimui praktiniuose taikymuose.



Šaltinis: [21].

6 pav. Trūkstamų reikšmių užpildymo metodų veiksmingumas pagal R^2 vertinimo metrikos kriterijų

Remiantis tyrimu [21], 6 paveikslėlyje pateikti rezultatai atskleidžia reikšmingą R^2 reikšmių variaciją tarp skirtingų užpildymo metodų, parodančią jų skirtingą veiksmingumą nagrinėjamame duomenų rinkinyje. Paprasti trūkstamų reikšmių užpildymo metodai, tokie kaip *SimpleImputer*, kuris realizuoja vidurkio, medianos ar modos pagrindu veikiančias strategijas, demonstruoja stabiliai aukštą prognozavimo tikslumą ($R^2 > 0,97$). Panašų veiksmingumo lygį pasiekia ir *k-Nearest Neighbors* (*KNN*) bei *Hot Deck* metodai, kurie trūkstamas reikšmes užpildo remdamiesi stebėjimų panašumu.

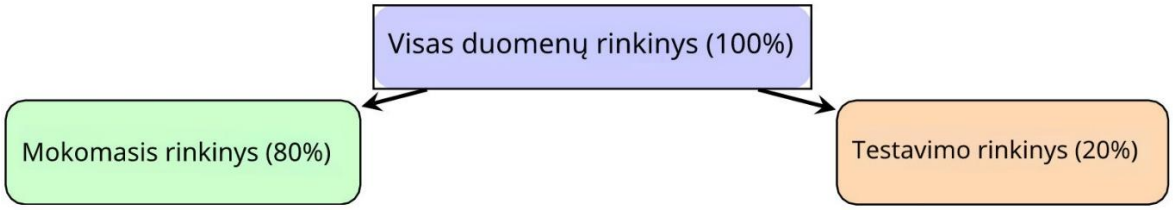
Tuo tarpu *Linear Regression* užpildymo metodas pasižymi neigiama R^2 reikšme (-0,4356), o *MissForest* metodas - artima nuliui, tačiau taip pat neigiama R^2 reikšme (-0,0259). Tokios reikšmės rodo, kad šių metodų taikymas nagrinėjamame kontekste neužtikrina prognozavimo pranašumo lyginant su paprastomis bazinėmis strategijomis. Šie rezultatai leidžia daryti išvadą, kad sudėtingesni užpildymo metodai gali būti jautrūs duomenų struktūrai ir trūkstamumo pobūdžiui, o jų veiksmingumas nėra universalus.

Priešingai, *Random Forest Regression* ir *MICE* metodai išlaiko aukštas R^2 reikšmes (apie 0,97), patvirtindami, kad tinkamai suderinti ansambliniai ir iteratyvūs užpildymo metodai gali užtikrinti aukštą prognozavimo tikslumą. Gauti rezultatai pabrėžia literatūroje akcentuojamą išvadą, jog trūkstamų reikšmių metodo pasirinkimas turi būti grindžiamas konkretaus duomenų rinkinio savybėmis, o ne vien metodo teoriniu sudėtingumu [21].

Toliau literatūroje akcentuojama, kad trūkstamų reikšmių metodų veiksmingumo vertinimas negali būti atskirtas nuo duomenų skaidymo į mokymo ir testavimo rinkinius strategijos, ypač tais

atvejais, kai analizuojami laiko eilučių duomenys. Šį aspektą išsamiai nagrinėja Akhil ir kt. tyrimas [22], kuriame lyginamas skirtingų mašininio ir giliojo mokymosi modelių veikimas, taikant chronologiškai struktūruotą duomenų skaidymą.

Autoriai pabrėžia, kad standartinė 80 / 20 proporcijos duomenų skaidymo schema, išlaikanti laiko seką, leidžia objektyviai įvertinti modelių gebėjimą prognozuoti naujus, anksčiau nematytus stebėjimus ir kartu išvengti duomenų nutekėjimo (*angl. data leakage*) problemos. Remiantis tyrimu [22], 7 paveiksle pateikta autorių taikyta duomenų skaidymo schema, iliustruojanti chronologinės struktūros išsaugojimą vertinant prognozavimo modelių veiksmingumą.



Šaltinis: sudaryta autoriaus remiantis [22]

7 pav. Mokymo ir testavimo duomenų skaidymo schema laiko eilutėms

Tinkamas duomenų skaidymas į mokymo ir testavimo rinkinius yra esminė prielaida patikimam trūkstamų reikšmių metodų veiksmingumo vertinimui. Tik išlaikant chronologinę seką ir užtikrinant, kad testavimo duomenys neatkartotų mokymo imties informacijos, galima objektyviai įvertinti, kaip duomenų užpildymo kokybė kinta esant skirtingam trūkstamų reikšmių lygiui. Ši sąsaja leidžia pagrįstai lyginti įvairių metodų (pavyzdžiui, MICE ar KNN) tikslumą ir nustatyti ribas, kuriose duomenų trūkumas pradeda reikšmingai bloginti modelio prognozavimo gebėjimus.

Būtent siekiant sistemiškai įvertinti, kaip trūkstamų reikšmių metodų veiksmingumas kinta didėjant trūkstamų duomenų proporcijai, literatūroje vis dažniau taikomos specializuotos analizės ir simuliaciniai tyrimai. Tokį požiūrį savo darbe taiko Shrivastava ir Shukla [23], kurie pristato daugiapakopę trūkstamų reikšmių užpildymo metodų vertinimo sistemą, leidžiančią nuosekliai analizuoti įvairių algoritmų tikslumą esant skirtingiems trūkstamumo lygiams.

Autoriai, naudodami modifikuotą duomenų rinkinį, atliko sistemingą vertinimą, taikydami tokias metodikas kaip *Multiple Imputation by Chained Equations* (*MICE*), *k-Nearest Neighbors (KNN)*, atsitiktinį duomenų užpildymą ir kitus tradicinius metodus. Siūlomas *Impute-VSS* įrankis leidžia stebėti ne tik užpildytų trūkstamų reikšmių pasiskirstymo pokyčius, bet ir kiekybiškai įvertinti jų efektyvumą pagal vidutinės absoliučios paklaidos (MAE), vidutinės kvadratinės paklaidos (MSE) ir kvadratinės paklaidos šaknies (RMSE) metrikas. Tokia metodinė prieiga suteikia galimybę detaliai įvertinti, kaip didėjantis duomenų trūkumas veikia tiek modelių prognozavimo tikslumą, tiek bendrą duomenų atkūrimo kokybę.

4 lentelė. MICE metodo vertinimo metrikų palyginimas su originaliu duomenų rinkiniu

Trūkstamų duomenų procentas	MAE	MSE	RMSE
20 %	0,9031	8,2249	2,8679
40 %	2,1097	20,8808	4,5695
60 %	3,156	29,3567	5,4182
80 %	4,1762	41,1747	6,4168

Šaltinis: sudaryta autoriaus remiantis [23].

Remiantis tyrimo šaltiniu [23], 4 lentelėje pateikti rezultatai rodo, kad taikant MICE metodą trūkstamų reikšmių užpildymo tikslumas nuosekliai mažėja didėjant trūkstamų duomenų kiekiui. Esant 20 % trūkstamų reikšmių, vidutinė absoliuti paklaida (MAE = 0,9031), vidutinė kvadratinė paklaida (MSE = 8,2249) ir jos šaknis (RMSE = 2,8679) išlieka santykinai mažos, o tai leidžia teigti, kad MICE metodas efektyviai atkuria prarastas reikšmes esant ribotam duomenų trūkumui.

Didėjant trūkstamų reikšmių proporcijai iki 80 %, visų vertinimo metrikų reikšmės ženkliai išauga - MAE pasiekia 4,1762, MSE - 41,1747, o RMSE - 6,4168. Ši tendencija atskleidžia spartų trūkstamų reikšmių užildymo paklaidų augimą, didėjant duomenų trūkumui, ir leidžia identifikuoti praktines MICE metodo taikymo ribas. Gauti rezultatai sutampa su kitų tyrimų išvadomis, kurios pabrėžia, kad net pažangūs iteratyvūs trūkstamų reikšmių užpildymo metodai praranda tikslumą esant dideliam trūkstamų reikšmių lygiui, todėl duomenų pilnumo ir kokybės užtikrinimas išlieka kritinis veiksnys patikimam modelių taikymui.

1.3.5 Literatūros analizės apibendrinimas ir rekomenduojamų metodų atranka

Atlikta sisteminė literatūros analizė leidžia suformuluoti apibendrinančias išvadas ir praktines rekomendacijas dėl mašininio mokymosi metodų taikymo trūkstamų duomenų problemai spręsti. Apžvelgti tyrimai patvirtina, kad nėra vieno universaliai tinkamo trūkstamų reikšmių užpildymo metodo, o optimalus sprendimas priklauso nuo duomenų savybių, taikymo konteksto bei techninių apribojimų [24], [12].

Literatūroje nuosekliai pabrėžiama, kad trūkstamų reikšmių užpildymo metodų veiksmingumas glaudžiai susijęs su tinkama duomenų skaidymo į mokymo ir testavimo rinkinius strategija. Chronologinės sekos išsaugojimas ir informacijos nutekėjimo prevencija leidžia objektyviau įvertinti, kaip skirtingi trūkstamų reikšmių užpildymo metodai veikia prognozavimo modelių tikslumą, ypač dirbant su laiko eilučių ar sekvenciniais duomenimis [25].

Shrivastava ir Shukla [23] pateiktoje *Impute-VSS* analizėje parodyta, kad didėjant trūkstamų duomenų proporcijai, trūkstamų reikšmių užpildymo paklaidos sparčiai didėja. Esant 20 % trūkstamų reikšmių, paklaidų rodikliai išlieka santykinai maži (MAE = 0,9031; RMSE = 2,8679), tačiau esant 80 % trūkstamumo jų reikšmės padidėja kelis kartus (MAE = 4,1762; RMSE = 6,4168), kas aiškiai

atskleidžia praktines net ir pažangių metodų taikymo ribas. Ši tendencija sutampa su kitų autorių išvadomis, jog esant dideliam duomenų trūkumui mažėja tiek trūkstamų reikšmių užpildymo metodų, tiek galutinių prognozavimo modelių stabilumas [26].

Apžvelgtoje literatūroje dažniausiai išskiriamos trys pagrindinės trūkstamų reikšmių užpildymo metodų grupės: paprastieji užpildymo metodai, tradiciniai statistiniai sprendimai ir pažangūs mašininio mokymosi algoritmai. Paprastieji užpildymo metodai, tokie kaip *SimpleImputer*, pasižymi skaičiavimo efektyvumu ir paprastu įgyvendinimu, tačiau ignoruoja požymių tarpusavio sąveikas ir gali iškreipti duomenų dispersiją, todėl rekomenduojami tik esant nedideliam trūkstamų reikšmių kiekiui arba kaip bazinis palyginimo metodas [27], [28], [29].

Pažangesni metodai, tokie kaip Random Forest, XGBoost, k-Nearest Neighbors ir Multiple Imputation by Chained Equations (MICE), literatūroje vertinami kaip tinkamesni sudėtingesniems ir didesnės apimties duomenų rinkiniams. Random Forest ir XGBoost metodai pasižymi gebėjimu modeliuoti nelinearinius ryšius ir dažnai užtikrina aukštą trūkstamų reikšmių bei prognozavimo tikslumą, tuo tarpu MICE metodas išsiskiria gebėjimu efektyviai atkurti sisteminį trūkstamumą, nors reikalauja didesnių skaičiavimo išteklių [12], [23].

Apibendrinant, literatūros analizė leidžia pagrįstai identifikuoti Random Forest, XGBoost, K-NN, MICE ir SimpleImputer kaip dažniausiai taikomus ir praktiniuose tyrimuose efektyviausius trūkstamų reikšmių užpildymo metodus. Atsižvelgiant į jų skirtingą sudėtingumo lygį, interpretavimo galimybes ir skaičiavimo sąnaudas, šie metodai toliau bus detaliau nagrinėjami teorinės dalies 1.4 skyriuje.

1.4 Trūkstamų reikšmių užpildymo algoritmų teorinė analizė

1.4.1 Random Forest algoritmas

Random Forest algoritmas yra „ansamblio“ (angl. *ensemble*) mokymosi metodas, kuris ekonominiuose [30] prognozavimo uždaviniuose užima svarbią vietą dėl savo aukšto tikslumo ir patikimumo. Šis algoritmas remiasi sprendimų medžių (angl. *decision tree*) ansamblio principais, derinant *bagging* metodą su atsitiktinių požymių atranka kiekviename mazge [31].

Algoritmo teoriniai pagrindai siejasi su Leo Breiman sukurta Bootstrap Aggregating (bagging) koncepcija, kuri buvo išplėtota į Forest-RI (*Random Input selection*) versiją. Algoritmo esmė - sukurti daug nepriklausomų sprendimo medžių, kur kiekvienas medis mokomas skirtingu duomenų poaibiu ir naudoja tik dalį prieinamų požymių, o galutinė prognozė formuojama visų medžių sprendimų vidurkiu (regresija) arba balsavimu (klasifikacija) [32].

Matematinis modelis grindžiamas tarpusavio koreliacijos ir generalizavimo klaidos sąveika. Algoritmo efektyvumas priklauso nuo dviejų pagrindinių komponentų: atskirų sprendimų medžių

stiprumo ir jų tarpusavio koreliacijos lygmens. Mažesnis koreliacijos lygis tarp medžių lemia geresnį generalizavimo gebėjimą, tuo tarpu aukštesnis medžių stiprumas didina prognozavimo tikslumą [33].

Šie teoriniai principai realizuojami per ansamblio agregavimo mechanizmą, kai galutinė prognozė formuojama agregacijos būdu. Pagrindinė algoritmo lygtis pateikiama formule (1):

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i\,(x), \tag{1}$$

kur:

- $\hat{y}$ yra galutinė Random Forest modelio prognozė įvesties duomenims x,
- N yra bendras sprendimų medžių skaičius „ansamblyje“,
- $f_i(x)$ yra i-ojo sprendimų medžio prognozė,
- (x) - įvesties požymių vektorius.

Random Forest modelio matematinis pagrindas atskleidžia algoritmo esmę - kaip parodyta formulėje (1), galutinė prognozė, formuojama kaip visų nepriklausomai apmokytų sprendimų medžių prognozių aritmetinis vidurkis. Kiekvienas medis $f_i(x)$ mokomas skirtingu *bootstrap* mėginiu, sukurtu atsitiktiniu būdu pasirenkant stebėjimus su pakartojimais iš pradinio duomenų rinkinio, o kiekviename sprendimo mazge vertinamas tik atsitiktinai pasirinktas požymių poaibis. Šis agregavimo principas, grindžiamas ansamblio mokymosi teorija, užtikrina, kad atsitiktinės individualių medžių paklaidos kompensuojasi galutinėje prognozėje, kas lemia ženkliai stabilesnį ir tikslesnį rezultatą nei bet kurio atskiro medžio prognozė. Formulės paprastumas maskuoja sudėtingą stochastinį procesą, kuris automatiškai balansuoja modelio šališkumą (angl. *bias*) ir dispersiją (angl. *variance*), užtikrinant optimalų generalizavimo gebėjimą ekonominių duomenų prognozavimo užduotims [34]. Verta paminėti, jog „Random Forest“ algoritmas gali sumažinti tiek šališkumą, tiek dispersiją, ypač esant aukštam signal-to-noise ratio (SNR), ir dažnai pranoksta tradicinį „bagging“ metodą, kai duomenyse egzistuoja struktūriniai šablonai, kuriuos gali užfiksuoti atsitiktinė požymių atranka [32].

Trūkstamų reikšmių užpildymo kontekste Random Forest algoritmas gali būti taikomas darbui su nepilnais duomenų rinkiniais, naudojant specialius trūkstamų reikšmių apdorojimo mechanizmus sprendimų medžiuose. Kai kintamasis su trūkstama reikšme pasiekia skaidymo tašką, stebėjimas gali būti nukreipiamas į abi medžio šakas proporcingai, atsižvelgiant į kitų stebėjimų pasiskirstymą, arba taikant pakaitinius (*angl. surrogate*) skaidymus. Toks trūkstamų reikšmių apdorojimas leidžia sumažinti informacijos praradimą mokymosi metu ir sudaro prielaidas taikyti Random Forest algoritmą trūkstamų reikšmių užpildymo uždaviniuose, kai jis naudojamas prognozuojant dirbtinai užmaskuotas reikšmes [33].

Random Forest algoritmo pagrindu sukurtas neparametrinis trūkstamų reikšmių nustatymo būdas (*angl. MissForest)* metodas taiko iteracinę procedūrą trūkstamoms reikšmėms užpildyti. Pirmajame žingsnyje visos trūkstamos reikšmės užpildomos preliminariai (pavyzdžiui, kiekybiniams kintamiesiems - medianos reikšme, o kategoriniams - dažniausiai pasitaikančia reikšme). Toliau kiekvienas kintamasis su trūkstamomis reikšmėmis modeliuojamas kaip priklausomas kintamasis, naudojant visus kitus kintamuosius kaip prediktorius. Random Forest modelis mokomas pilnais duomenimis ir naudojamas prognozuoti trūkstamas reikšmes. Procesas kartojamas iteratyviai, kol konverguojama arba pasiekiamas maksimalus iteracijų skaičius [23].

Skirtingai nuo paprastų vieno kintamojo trūkstamų reikšmių užpildymo metodų (tokių kaip vidurkio ar medianos užpildymas), Random Forest algoritmas atsižvelgia į visų požymių tarpusavio sąveikas. Kiekvienas sprendimų medis mokosi skirtingo duomenų poaibio ir vertina skirtingą požymių kombinaciją, o ansamblio agregacija užtikrina, kad galutinė užpildyta reikšmė atspindėtų daugiamačius duomenų ryšius. Tai ypač svarbu ekonominių rodiklių kontekste, kur kintamieji dažnai yra tarpusavyje susiję per fundamentalius ekonominius mechanizmus.

Empiriniai tyrimai patvirtina Random Forest efektyvumą trūkstamų reikšmių užpildymo uždaviniuose. Literatūroje pateikta analizė (žr. 1.3.2 poskyrio, 1 lentelę) rodo, kad Random Forest Imputer pasiekia R^2 balą 0,77 ir MSE 0,26, nors užpildymo laikas (22,54 sekundės) yra žymiai trumpesnis nei kai kurių kitų metodų [19]. Tyrimai taip pat atskleidė, kad MissForest metodas gali generuoti neigiamą R^2 reikšmę (-0,0259) specifiniuose duomenų kontekstuose, kas rodo, jog algoritmo efektyvumas priklauso nuo duomenų struktūros suderinamumo [21].

NUTS 2 regionų ekonominių rodiklių kontekste Random Forest algoritmas yra ypač tinkamas dėl kelių priežasčių. Pirma, regioniniai duomenys pasižymi sudėtingais tarpregioniniais ryšiais ir erdvine autokoreliacija, kuriuos algoritmas gali efektyviai modeliuoti. Antra, ekonominiai rodikliai turi nevienalytes skales ir pasiskirstymus, o Random Forest yra atsparus tokiai heterogeniškumo problemai. Trečia, regioninių ekonominių rodiklių duomenys NUTS 2 lygmenyje dažnai pasižymi dideliu trūkstamų reikšmių paplitimu, todėl tikslinga taikyti ansamblinius metodus, kurie, derinami su trūkstamų duomenų apdorojimo strategijomis, leidžia stabiliai modeliuoti duomenis esant aukštam trūkstamumo lygiui [35].

1.4.2 XGBoost algoritmas

Gradient Boosting algoritmas yra ansamblio mokymosi metodas, kuris iteratyviai konstruoja sprendimų medžių seką, kur kiekvienas naujas medis mokosi iš ankstesnių medžių klaidų. Šis metodas, ypač pažangi šio metodo versija XGBoost (angl. *Extreme Gradient Boosting*), plačiai taikomas struktūrizuotų duomenų analizėje ir ekonominių rodiklių prognozavime [36].

Algoritmo teoriniai pagrindai remiasi gradientinio stiprinimo (*angl. gradient boosting*) principais, kur modelis konstruojamas adityviu būdu, kiekviename etape pridedant naują bazinį mokytoją (sprendimų medį), kuris optimizuojamas pagal nuostolių funkcijos gradientą. Formaliai, prognozė po K iteracijų išreiškiama matematine formule (2) taip:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \tag{2}$$

kur:

- $\hat{y}_i$ - galutinė prognozė i-tajam stebėjimui,
- K - bendras iteracijų (medžių) skaičius,
- f_k - k-tasis sprendimų medis,
- x_i - įvesties požymių vektorius.

XGBoost algoritmas optimizuoja tikslią objekto funkciją, kuri apjungia nuostolių funkciją ir reguliarizacijos komponentą. Bendroji optimizavimo funkcija išreiškiama formule (3):

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

kur L yra diferencijuojama nuostolių funkcija, matuojanti skirtumą tarp tikrosios reikšmės y_i ir prognozės $\hat{y}_i$, o $\Omega(f_k)$ - reguliarizacijos narys, kontroliuojantis modelio sudėtingumą ir apsaugantis nuo persimokymo.

Matematinis modelio sudarymas vyksta per antrojo laipsnio Teiloro (*angl. Taylor*) aproksimaciją, kas leidžia efektyviai optimizuoti tikslią objekto funkciją. Kiekviename stiprinimo žingsnyje t, naujas medis f_t parenkamas minimizuojant aproksimuotą tikslinę funkciją, pateikiamą formule (4):

$$\sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \tag{4}$$

kur g_i ir h_i yra atitinkamai pirmosios ir antrosios eilės nuostolių funkcijos dalinės išvestinės pagal prognozę [37].

Duomenų apdorojimo ir optimizavimo galimybės išskiria XGBoost algoritmą ekonominių rodiklių prognozavimo kontekste. Algoritmo gebėjimas mokymosi metu apdoroti trūkstamas reikšmes, nustatant optimalias skaidymo kryptis sprendimų medžiuose, leidžia stabiliai treniruoti modelį net esant neišsamiems duomenims. Tai ypač svarbu dirbant su realiais ekonominiais duomenimis, kurie dažnai pasižymi trūkstamais stebėjimais [38].

Trūkstamų duomenų kontekste XGBoost pasižymi integruotu trūkstamų reikšmių apdorojimo mechanizmu mokymosi metu, kuris skiriasi nuo daugelio kitų mašininio mokymosi metodų. Kiekviename sprendimų medžio mazge algoritmas automatiškai nustato optimalią kryptį stebėjimams su trūkstamomis reikšmėmis, remdamasis duomenų struktūra [36].

XGBoost algoritme trūkstamų reikšmių apdorojimas grindžiamas vadinamuoju „numatytosios krypties“ (*angl. default direction*) principu. Mokymo metu, susidūrus su trūkstama reikšme konkrečiame požymyje, stebėjimai laikinai nukreipiami į skirtingas medžio šakas, o optimali kryptis parenkama taip, kad maksimaliai sumažintų nuostolių funkciją. Ši mokymosi metu nustatyta kryptis išsaugoma kiekviename skaidymo mazge ir taikoma prognozavimo metu [39]. Toks požiūris leidžia algoritmui toleruoti trūkstamas reikšmes ir išnaudoti galimus trūkstamumo dėsningumus, sudarant prielaidas XGBoost taikyti trūkstamų reikšmių užpildymo uždaviniams, kai atkuriamos dirbtinai užmaskuotos reikšmės.

Naudojant XGBoost trūkstamų reikšmių užpildymo užduotims, dažniausiai taikomas iteracinis metodas, panašus į MICE principą. Kiekvienas kintamasis su trūkstamomis reikšmėmis modeliuojamas kaip tikslinė reikšmė (*angl. target*), naudojant visus kitus kintamuosius kaip prediktorius. XGBoost modelis mokomas pilnais duomenimis ir generuoja prognozes trūkstamoms reikšmėms. Procesas kartojamas iteratyviai visiems kintamiesiems, kol užpildytų reikšmių pokyčiai stabilizuojasi [19].

Empiriniai tyrimai atskleidžia, kad XGBoost Estimator MICE metodologijoje reikalauja žymiai ilgesnio skaičiavimo laiko (131,72 sekundės) lyginant su baziniais metodais, tačiau pasiekia palyginamą R^2 balą (0,76) ir MSE (0,26) su kitais ansambliniais metodais [19]. Tai rodo kompromisą tarp tikslumo ir skaičiavimo efektyvumo. XGBoost Imputer, naudojamas kaip atskiras trūkstamų reikšmių užpildymo įrankis, demonstruoja šiek tiek žemesnius rezultatus ($R^2 = 0,76$, MSE = 0,27), bet žymiai trumpesnį apdorojimo laiką (2,88 sekundės) [19].

Be to, algoritmas taiko du reguliarizacijos tipus - L1 (Lasso) ir L2 (Ridge), kurie kontroliuojami per α ir λ parametrus, užtikrinant modelio apibendrinimo gebėjimus [38].

Hiperparametrų optimizavimas yra svarbus XGBoost algoritmo veikimo aspektas. Pagrindiniai parametrai apima sprendimų medžių skaičių (*angl. n_estimators*), mokymosi greitį (angl. learning_rate), maksimalų medžio gylį (*angl. max_depth*), duomenų imties dalį kiekvienam medžiui (*angl. subsample*) ir požymių dalį kiekvienam medžiui (*angl. colsample_bytree*). Optimalioms hiperparametrų kombinacijoms nustatyti taikomi Bajeso optimizacijos metodai arba atsitiktinė paieška, kurie efektyviau nei tinklelio paieška tyrinėja didelę parametrų erdvę [40].

Ekonominių rodiklių prognozavimo tyrimuose XGBoost modelis su optimizuotais hiperparametrais demonstruoja aukštą tikslumą - BVP prognozavimo uždaviniuose determinacijos

koeficientas gali siekti ($R^2 > 0,98$), o vidutinė absoliuti paklaida (MAE < 0,05), kaip parodyta ekonomikos analizės tyrime su (1970 - 2022) metų duomenimis [41].

Skaičiavimo efektyvumas ir našumas išsiskiria kaip reikšmingas XGBoost privalumas. Algoritmas naudoja pažangias duomenų struktūras ir lygiagretinimo technikas, įskaitant stulpelių blokų struktūrą sparčiai skaidymo taškų paieškai ir išretintų matricų palaikymą. Tai leidžia efektyviai apdoroti didelius duomenų rinkinius su tūkstančiais požymių, kas būdinga regioninių ekonominių rodiklių analizei NUTS 2 lygmeniu. Algoritmas taip pat palaiko GPU akseleraciją, kas gali pagreitinti mokymosi procesą iki 10 kartų lyginant su CPU implementacija [36].

Požymių svarbos vertinimas XGBoost algoritme suteikia vertingų įžvalgų ekonominių procesų analizei. Algoritmas skaičiuoja tris požymių svarbos metrikas: vidutinis tikslumo pagerinimas (*angl. gain*), vidutinis stebėjimų skaičius (*angl. cover*) ir panaudojimo dažnis medžiuose (*angl. frequency*). Šie rodikliai leidžia identifikuoti svarbiausius ekonominius veiksnius, įtakojančius prognozuojamus rodiklius, o tai ypač svarbu formuojant ekonominę politiką regioniniu lygmeniu [42].

NUTS 2 regionų ekonominių rodiklių kontekste XGBoost yra ypač vertingas dėl kelių priežasčių. Pirma, algoritmo gebėjimas mokymosi metu apdoroti trūkstamas reikšmes, nustatant optimalias jų skaidymo kryptis sprendimų medžiuose, yra ypač svarbus dirbant su regioniniais ekonominiais duomenimis, kuriems būdingas didelis trūkstamumo lygis. Antra, reguliarizacijos mechanizmai (L1 ir L2) apsaugo nuo persimokymo net esant aukštam požymių skaičiui ir ribotam stebėjimų kiekiui laiko eilutėse. Trečia, požymių svarbos vertinimas leidžia identifikuoti, kurie ekonominiai rodikliai yra svarbiausi prognozuojant kitus rodiklius, kas suteikia papildomų įžvalgų regioninių ekonominių procesų supratimui [43].

1.4.3 KNN

KNN (*angl. k-Nearest Neighbors*) algoritmas yra neparametrinis taikomas metodas trūkstamų reikšmių užpildymui, kai kiekviena stebėjimo eilutė su trūkstama reikšme imama kaip taškas daugiamačiame požymių erdvėje, o trūkstama reikšmė užpildoma pagal k artimiausių kaimynų reikšmes. Beretta ir Santaniello tyrime [44] pabrėžia, kad KNN metodas leidžia „naudoti pagalbos informaciją iš x - kintamųjų“ ir taip išsaugoti duomenų struktūrą.

Taikant KNN metodą trūkstamų reikšmių užpildymui, dažniausiai naudojama atstumo metrika yra Euklido atstumas, kuris apibrėžiamas formule (5):

$$\mathrm{Dist}_{xy} = \sqrt{\sum_{j=1}^{m} \left(X_{ij} - X_{kj}\right)^2}, \tag{5}$$

kur m - požymių skaičius, X_{ij} - i-tosios eilutės j-to požymio reikšmė,o X_{kj} kaimyno k-tos eilutės atitinkamai [45].

Apskaičiavus atstumus tarp stebėjimų ir identifikavus k artimiausių kaimynų rinkinį, trūkstamos reikšmės užpildomos naudojant svertinį kaimynų reikšmių vidurkį, pateikiamą formule (6):

$$\hat{X}_i = \frac{\sum_{j=1}^{k} w_j \, v_j}{\sum_{j=1}^{k} w_j}, \tag{6}$$

kur v_j - kaimyno reikšmė, w_j - svoris (pavyzdžiui, atstumo inversas).

Trūkstamų ekonominių rodiklių kontekste KNN metodas gali būti efektyvus, kai tarp regioninių rodiklių vystosi panašios tendencijos (tokiu būdu galima rasti „artimus“ kaimynus), o trūkstamų reikšmių dalis nėra pernelyg didelė. Memon savo tyrime [24] pažymi, kad trūkstamų reikšmių užpildymas su KNN algoritmu, duomenų rinkiniuose su kategoriniais požymiais buvo vienas iš efektyviausių pasirinkimų. Tačiau KNN algoritmas reikalauja, kad kintamieji būtų tinkamai paruošti (pavyzdžiui standartizuoti, kategoriniai koduoti), ir kad pasirinktas k bei atstumo metrika būtų pritaikyti duomenų struktūrai.

KNN Privalumai:

- Išlaiko požymių tarpusavio struktūrą ir nelinearius ryšius, todėl dažnai užtikrina tikslesnį trūkstamų reikšmių užpildymą nei vieno požymio metodai.
- Nereikalauja prielaidų apie duomenų pasiskirstymą, todėl gali būti taikomas gana plačiai įvairių tipų duomenims [44].
- Paprasta taikyti esant vidutiniškai dideliam duomenų rinkiniui ir neaukštam trūkstamumo lygiui.

KNN Trūkumai ir apribojimai

- Didelis skaičiavimo sudėtingumas: atstumų skaičiavimas tarp visų stebėjimų sukelia reikšmingą laiko ir atminties sąnaudą dideliuose duomenų rinkiniuose [45].
- Kai trūkstamų reikšmių dalis yra labai didelė arba duomenys yra itin retai užpildyti, kaimynų atranka gali tapti neinformatyvi arba šališka, dėl ko reikšmingai sumažėja trūkstamų reikšmių užpildymo kokybė.
- Jei trūkstamumo mechanizmas priklauso nuo nestebimų kintamųjų (MNAR), KNN metodas dažnai neatitinka reikiamų prielaidų [45].

1.4.4 MICE

Daugiareikšmis trūkstamų duomenų užpildymo metodas MICE (*angl. Multiple Imputation by Chained Equations*) yra pažangus trūkstamų reikšmių tvarkymo būdas, plačiai taikomas statistikoje ir mašininio mokymosi srityje. Šis metodas grindžiamas daugkartinio užpildymo principu, kai kiekviena trūkstama reikšmė užpildoma ne vieną, o kelis kartus, suformuojant kelis alternatyvius užpildytus duomenų rinkinius. Tai leidžia įvertinti užpildymo neapibrėžtumą ir sumažinti sprendimų šališkumą.

MICE metodas remiasi iteracinių regresijų grandine (*angl. fully conditional specification, FCS*), kai kiekvienam kintamajam su trūkstamomis reikšmėmis sudaromas atskiras regresinis modelis, naudojant kitus kintamuosius kaip prediktorius. Pagrindiniai žingsniai:

1. Tarkime, kad duomenų rinkinys $X = (X_1, \dots, X_p)$, kur kai kuriuose X_j yra trūkstamų reikšmių.
2. Pradiniame žingsnyje trūkstamos reikšmės užpildomos preliminariai (pavyzdžiui vidurkiu, mediana ar atsitiktine reikšme).
3. Tuomet už kiekvieną kintamąjį X_j iteratyviai taikomas regresijos modelis, kuris aprašomas formule (7):

$$X_j^{(t)} = f_j\left(X_{-j}^{(t)}\right) + \varepsilon_j, \tag{7}$$

kur

- $X_j^{(t)}$- kintamasis, kuriam atliekamas užpildymas t - osios iteracijos metu;
- $X_{-j}^{(t)}$- visi kiti kintamieji (išskyrus X_j) tuo metu atitinkamai užpildyti arba stebimi;
- $f_j(\cdot)$ - pasirinktas regresijos modelis (pavyzdžiui tiesinė regresija, logistinė, atsitiktinių miškų, priklausomai nuo kintamojo tipo);
- ε_j - atsitiktinis (klaidos) komponentas.

Šis procesas kartojamas iteratyviai visiems kintamiesiems tol, kol pasiekiamas konvergencijos kriterijus arba nustatytas iteracijų skaičius, o galutiniai rezultatai gaunami apjungiant kelių užpildytų duomenų rinkinių įverčius [46].

Kai sukurti m komplektai užpildytų duomenų, pagal klasikinę daugkartinio užpildymo metodiką *(angl. multiple imputation, MI*) tvarką Donald B. Rubin taisyklėmis [47] jie sujungiami į vieną bendrą įvertį, kuris apskaičiuojamas pagal formulę (8):

$$\bar{Q} = \frac{1}{m}\sum_{k=1}^{m} Q_k\,, \quad T = \bar{U} + \left(1 + \frac{1}{m}\right)B, \tag{8}$$

kur Q_k - analizės rezultatas iš k - tojo užpildyto rinkinio, $\bar{U}$ - vidutinė vidinės variacijos įverčio reikšmė, B - tarprinkinių dispersija ir T bendras įverčio (po užpildymo) neapibrėžtumas [48]. Formulėje (8) $\bar{Q}$ nusako galutinį taškinį analizės rezultatą, o T - šio įverčio bendrąjį neapibrėžtumą, apjungiantį tiek vidinę, tiek tarpinę (po trūkstamų reikšmių užpildymo) variaciją.

MICE Privalumai:

- MICE leidžia išsaugoti kintamųjų tarpusavio ryšius, nes kiekvieno kintamojo trūkstamos reikšmės užpildomos remiantis kitais kintamaisiais, todėl užpildymo struktūra atspindi duomenų tarpusavio sąryšius.
- Metodo lankstumas: skirtingiems kintamiesiems gali būti taikomi skirtingi modeliai (kiekybiniai, kategoriniai), galima naudoti specializuotus trūkstamų reikšmių užpildymo būdus (pavyzdžiui *„passive imputation“*) [46].
- Įtraukiamas trūkstamų reikšmių užpildymo neapibrėžtumas: generuojant kelis užpildytus duomenų rinkinius ir juos sujungiant pagal Rubin taisykles, įvertinama ne tik užpildytų reikšmių centrinė tendencija, bet ir su užpildymu susijęs neapibrėžtumas. [48].

MICE Trūkumai ir apribojimai

- Didelis skaičiavimo sudėtingumas - MICE metodas gali būti imlus skaičiavimams, kai duomenų rinkinys didelis, kintamųjų daug, o trūkstamųjų duomenų struktūra sudėtinga.
- Modelio priklausomumas - rezultatai priklauso nuo pasirinkto regresinio modelio kiekvienam kintamajam ir nuo prielaidų, ar tas modelis atitinka tikrąją duomenų generavimo struktūrą.
- Konvergencijos problema - ne visuomet garantuojama, kad iteracijos greitai stabilizuosis ar bus pasiekta globali konvergencija [46].

1.4.5 SimpleImputer

Paprastasis trūkstamų reikšmių užpildymo metodas (*angl. Simple Imputation*) yra vienas iš bazinių algoritmų, taikomų tiek statistinėje analizėje, tiek mašininio mokymosi duomenų paruošimo etapuose. Šis metodas priskiriamas vienkartinio užpildymo metodų grupei (*angl. single imputation methods*) ir dažniausiai naudojamas kaip atskaitos taškas, vertinant sudėtingesnių trūkstamų reikšmių užpildymo algoritmų veiksmingumą.

SimpleImputer principas grindžiamas prielaida, kad trūkstamas reikšmes galima pakankamai tiksliai įvertinti naudojant kintamojo pasiskirstymo centrines tendencijas arba dažniausiai pasitaikančias reikšmes. SimpleImputer metodo veikimo principas formaliai aprašomas formule (9):

$$X_{ij}^{*} = \begin{cases} X_{ij}, & \text{jei } X_{ij} \text{ nėra trūkstama,} \\ \hat{\mu}_j, & \text{jei } X_{ij} \text{ yra trūkstama,} \end{cases} \quad (9)$$

kur X_{ij} - pradinė i- tosios eilutės ir j - tojo požymio reikšmė, o $\hat{\mu}_j$ - pasirinktas trūkstamų reikšmių užpildymo įverčio operatorius (vidurkis, mediana ar moda) [27].

Šis metodas yra įgyvendintas daugelyje duomenų analizės bibliotekų, pavyzdžiui, Python „scikit-learn“ bibliotekos aplinkoje, kur vartotojas gali nurodyti trūkstamų reikšmių užpildymo strategiją:

- mean - trūkstamos reikšmės pakeičiamos kintamojo aritmetiniu vidurkiu;
- median - trūkstamos reikšmės pakeičiamos medianos reikšme;
- most_frequent - trūkstamos reikšmės pakeičiamos dažniausiai pasitaikančia reikšme (naudojama kategoriniams duomenims);
- constant - trūkstamos reikšmės pakeičiamos fiksuota konstanta, kurią apibrėžia vartotojas [49].

SimpleImputer metodas išsiskiria savo skaičiavimo paprastumu ir greitumu, todėl dažnai taikomas kaip bazinis palyginamasis metodas (*angl. baseline imputation*) prieš taikant sudėtingesnius algoritmus užpildymui. Nors šis metodas neužtikrina tarpinių kintamųjų sąryšių išsaugojimo, jis tinkamas pradiniam duomenų apdorojimui, kai trūkstamų reikšmių kiekis nedidelis (paprastai iki 5 - 10 % visų duomenų) [10].

Pagrindinis metodo privalumas - nepriklausomumas nuo duomenų tipo ir pasiskirstymo formos, kas leidžia jį taikyti tiek kiekybiniams, tiek kategoriniams kintamiesiems. Be to, metodas yra stabilus ir nesukelia konvergencijos problemų, priešingai nei kai kurie iteraciniai algoritmai [28].

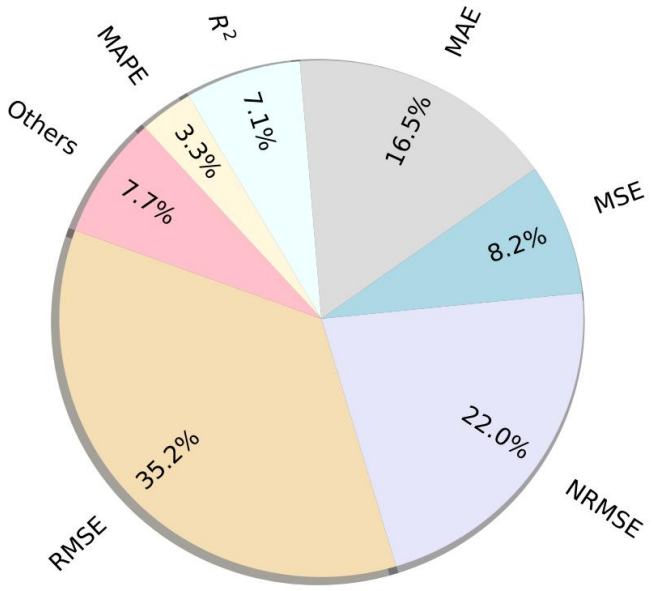
Vis dėlto, *SimpleImputer* turi ir esminių apribojimų. Kadangi kiekviena trūkstama reikšmė pakeičiama tuo pačiu įverčiu, duomenų variacija sumažėja, o tai gali lemti šališką dispersijos įvertį bei iškreiptas koreliacijas tarp kintamųjų [46]. Ši problema ypač aktuali, kai trūkstamos reikšmės nėra visiškai atsitiktinės MNAR (*angl. not missing completely at random*). Dėl šių priežasčių *SimpleImputer* rekomenduojama naudoti tik kaip pradinį duomenų paruošimo etapą arba kaip etaloninį metodą, lyginant sudėtingesnius trūkstamų reikšmių užpildymo sprendimus [29].

Praktiniuose tyrimuose pastebima, kad *SimpleImputer* metodas, taikomas kartu su standartizavimo arba normalizavimo etapais, gali duoti pakankamai gerus rezultatus, kai duomenys pasižymi mažu triukšmu ir nedideliu trūkstamumo lygiu.

Apibendrinant galima teigti, kad *SimpleImputer* yra efektyvus, paprastas ir greitas metodas trūkstamų reikšmių užpildymui, tačiau dėl savo supaprastinto pobūdžio netinka situacijoms, kai duomenų struktūra sudėtinga arba trūkstamumo mechanizmas nėra atsitiktinis. Tokiais atvejais rekomenduojama taikyti iteracinius arba ansamblinius trūkstamų reikšmių užpildymo metodus [50].

1.5 Trūkstamų reikšmių užpildymo metodų vertinimo metrikų analizė ir pagrindimas

Trūkstamų reikšmių užpildymo metodų veiksmingumo vertinimas yra svarbus siekiant parinkti tinkamiausią užpildymo metodą konkrečiam duomenų rinkiniui. Norint užtikrinti, kad užpildyti duomenys išlaikytų pradinių duomenų struktūrinį vientisumą ir statistinį patikimumą, būtina taikyti tinkamas vertinimo metrikas [51]. Vertinimo metrikos skirstomos į dvi pagrindines kategorijas: regresijos metrikas, taikomas kiekybiniams duomenims, ir klasifikacijos metrikas, naudojamas kategoriniams duomenims. Regresijos ir klasifikacijos metrikų matematinės išraiškos bei jų formalūs apibrėžimai pateikiami 5 lentelėje.



Šaltinis: [20]

8 pav. Trūkstamų reikšmių užpildymo metodų vertinimo metrikų pasiskirstymas atliktuose tyrimuose

Kaip matyti 8 paveiksle, tyrimų [20] analizė atskleidžia dažniausiai naudojamų tiesioginių vertinimo metrikų pasiskirstymą. RMSE (*angl. Root Mean Squared Error*) yra plačiausiai taikoma metrika, apimanti 35,2 % visų analizuotų tyrimų. Po jos seka NRMSE (*angl. Normalized RMSE*) su 22,0 %, MSE (*angl. Mean Squared Error*) su 15.9 %, MAE (*angl. Mean Absolute Error*) su 11,7 %, R^2 determinacijos koeficientas su 8,2 % ir MAPE (*angl. Mean Absolute Percentage Error*) su 3,3 %. Kitos metrikos sudaro tik 7,7 % visų taikomų vertinimo metodų [20].

5 lentelė. Regresijos ir klasifikacijos metrikų matematinės išraiškos

Metrikos tipas	
Regresijos metrikos	**Klasifikacijos metrikos**
$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (10)	$Tikslumas = \frac{TP + TN}{TP + TN + FP + FN}$ (14)

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	(11)	$Preciziškumas = \frac{TP}{TP + FP}$	(15)
$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	(12)	$Atpažinimas = \frac{TP}{TP + FN}$	(16)
$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \times 100\ \%$	(13)	$F_1 = 2 \times \frac{Preciziškumas \times Atpažinimas}{Preciziškumas + Atpažinimas}$	(17)

Šaltinis: sudaryta autoriaus remiantis *[51]*.

Regresijos metrikos naudojamos vertinti kiekybinių duomenų užpildymo tikslumą. Pagrindinės metrikos apima:

Vidutinė kvadratinė paklaida MSE (*angl. Mean Squared Error*) matuoja vidurkį kvadratinių skirtumų tarp faktinių ir užpildytų reikšmių. MSE apskaičiuojama pagal formulę (10), kur y_i yra faktinė reikšmė $\hat{y}_i$ yra prognozuota reikšmė, o n yra stebėjimų skaičius. Mažesnis MSE rodo geresnę užpildymo kokybę [51].

Šakninė vidutinė kvadratinė paklaida RMSE (*angl. Root Mean Squared Error*), parodyta formulėje (11), yra MSE kvadratinė šaknis. RMSE pranašumas yra tas, kad ji pateikia paklaidų matavimus tais pačiais vienetais kaip ir tikslinė reikšmė, kas palengvina interpretaciją. Be to, RMSE stipriau baudžia didesnes paklaidas ir yra mažiau jautri išskirtims lyginant su MSE [51]. Tyrimuose RMSE yra viena iš dažniausiai naudojamų metrikų - net 27 % analizuotų tyrimų naudojo šią metriką [12].

Vidutinė absoliuti paklaida MAE (*angl. Mean Absolute Error*), apibrėžta formulėje (12), skaičiuoja vidutinį absoliutų skirtumą tarp faktinių ir prognozuotų reikšmių. MAE pranašumas yra jos paprastumas ir intuityvus interpretavimas. Apie 19 % mašininio mokymosi tyrimų, skirtų trūkstamų reikšmių užpildymui, MAE metriką naudoja kaip pagrindinį vertinimo kriterijų [12].

Vidutinė absoliuti procentinė paklaida MAPE (*angl. Mean Absolute Percentage Error*), pateikta formulėje (13), išreiškia paklaidą procentais nuo faktinės reikšmės. MAPE yra ypač naudinga lyginant modelius su skirtingais duomenų masteliais, nes ji normalizuoja paklaidas procentine išraiška.

Klasifikacijos metrikos taikomos vertinant kategorinius duomenis po užpildymo:

Tikslumas (*angl. Accuracy*), apibrėžtas formulėje (14), matuoja teisingai klasifikuotų stebėjimų proporciją. Šioje formulėje TP (*angl. True Positive*) ir TN (*angl. True Negative*) reiškia teisingai klasifikuotus stebėjimus, o FP (*angl. False Positive*) ir FN (*angl. False Negative*) - klaidingai klasifikuotus. Tyrimuose PCP (*angl. Percentage of Correct Prediction*), kuris yra ekvivalentiškas tikslumo metrikai, buvo naudojamas 27 % analizuotų tyrimų [12].

Preciziskumas (*angl. Precision*), parodytas formulėje (15), vertina kokią dalį teigiamų prognozių sudaro tikrai teigiami atvejai. Ši metrika yra labai svarbi, kai klaidingai teigiamų rezultatų pasekmės yra reikšmingos [51].

Atpažinimas (*angl. Recall arba Sensitivity*), apibrėžtas formulėje (16), matuoja kokią dalį faktiškai teigiamų atvejų modelis teisingai identifikuoja. Ši metrika yra svarbi, kai reikia identifikuoti visus teigiamus atvejus [51].

F_1 balas (angl. F_1-Score), pateiktas formulėje (17), yra harmoninis vidurkis tarp preciziškumo ir atpažinimo. F_1 balas suteikia subalansuotą modelio veikimo įvertinimą, ypač kai duomenų klasės yra nesubalansuotos [51].

Vertinimo metrikų pasirinkimas priklauso nuo konkretaus trūkstamų reikšmių užpildymo tikslo ir analizuojamų duomenų pobūdžio. Daugelis tyrimų taiko kelias metrikas vienu metu, siekdami gauti išsamesnį trūkstamų reikšmių užpildymo kokybės įvertinimą. Be to, svarbu pažymėti, kad didelė dalis tyrimų (apie 70 %) eksperimentams naudoja dirbtinai sugeneruotus trūkstamus duomenis, nes tai leidžia tiksliai įvertinti taikomų užpildymo metodų veiksmingumą, lyginant rezultatus su žinomomis pradinėmis reikšmėmis [12].

1.6 Skyriaus apibendrinimas

Pirmojo skyriaus teorinė analizė atskleidė mašininio mokymosi modelių svarbą ir plačias taikymo galimybes ekonominių rodiklių prognozavimo srityje, ypač NUTS 2 regionų kontekste. Sisteminga literatūros analizė ir algoritminių sprendimų tyrimas formuoja tvirtą teorinį pagrindą empiriniam tyrimui.

Dirbtinio intelekto ir mašininio mokymosi reikšmė ekonominių rodiklių prognozavime išsiskyrė kaip sparčiai auganti sritis su dideliu praktiniu potencialu. Analizė atskleidė, kad mašininio mokymosi modeliai pranoksta tradicinius statistinius metodus gebėjimu apdoroti sudėtingus netiesinius ryšius, automatiškai išskirti reikšmingus požymius ir prisitaikyti prie kintančių ekonominių sąlygų.

Sisteminės literatūros analizės rezultatai patvirtino mašininio mokymosi metodų efektyvumą ekonominiams prognozavimams. Atliktos sisteminės literatūros tyrimų analizė atskleidė svarbiausias tendencijas: hibridiniai modeliai dažnai pranoksta atskirų algoritmų rezultatus, automatizuoti laiko eilučių prognozavimo metodai reikalauja išsamios sistemos architektūros, o Bajeso ir neuroninių tinklų metodai demonstruoja aukštą tikslumą specifinėse srityse.

Atlikta Random Forest algoritmo analizė parodė pagrindinius šio metodo privalumus ekonominių prognozių sudaryme: ansamblio mokymosi principai užtikrina stabilų modelio veikimą, algoritmas geba efektyviai tvarkyti trūkstamas reikšmes ir kategorinių kintamųjų derinius, o tinkamai parinkti hiperparametrai leidžia pasiekti aukštą prognozavimo tikslumą. Matematinis modelis,

grindžiamas sprendimų medžių agregavimo principu, formuoja patikimą teorinį pagrindą praktiniam taikymui.

Tyrimo rezultatai atskleidė, kad XGBoost algoritmas pasižymi išskirtiniu efektyvumu dėl pažangių gradientinio stiprinimo principų taikymo. Algoritmo gebėjimas optimizuoti tikslią objekto funkciją su reguliarizacijos komponentais, automatiškai tvarkyti trūkstamus duomenis ir suteikti požymių svarbos vertinimus daro jį ypač tinkamą regioninės ekonomikos analizei.

Pagrindinės teorinio tyrimo išvados:

1. Algoritmų tinkamumas - Random Forest ir XGBoost metodai išsiskyrė kaip patikimiausi sprendimai ekonominių rodiklių prognozavimui dėl aukšto tikslumo ir duomenų tvarkymo galimybių.
2. Metodologinis pagrindimas - sistemina literatūros analizė patvirtino šių algoritmų pranašumus prieš tradicinius statistinius metodus ekonominės analizės kontekste.
3. Hibridinių modelių potencialas - tyrimai atskleidė, kad algoritmų deriniai gali pasiekti dar aukštesnį prognozavimo tikslumą nei atskiri metodai.
4. Duomenų kokybės svarba - visi analizuoti algoritmai reikalauja kvalifikuoto duomenų paruošimo ir hiperparametrų optimizavimo optimaliam veikimui.
5. Regioninės ekonomikos specifika - NUTS 2 lygmens ekonominių rodiklių ypatumai reikalauja specializuotų metodologinių sprendimų, kuriuos gali pateikti pažangūs mašininio mokymosi algoritmai.

Teorinės analizės rezultatai formuoja tvirtą pagrindą tolesniam empiriniam tyrimui, kuriame bus praktiškai testuojami Random Forest ir XGBoost algoritmai NUTS 2 regionų ekonominių rodiklių prognozavimo uždaviniams. Algoritminių sprendimų teorinis pagrindimas užtikrina, kad praktinis tyrimas bus metodologiškai pagrįstas ir orientuotas į efektyviausių sprendimų identifikavimą regioninės ekonomikos analizės kontekste.

2. NUTS 2 REGIONŲ EKONOMINIŲ RODIKLIŲ TYRIMO METODOLOGIJA

2.1 NUTS 2 regionų klasifikacijos sistema ir charakteristikos

NUTS (angl. *Nomenclature of Territorial Units for Statistics*) arba lietuviškai - Statistinių teritorinių vienetų nomenklatūra, yra hierarchinė sistema, sukurta Europos Sąjungoje, siekiant standartizuoti regioninę statistiką ir palengvinti regioninės politikos formavimą bei įgyvendinimą. Ši klasifikacija leidžia objektyviai palyginti regionus tarpusavyje, stebėti jų ekonominę raidą ir tikslingai paskirstyti ES struktūrinius fondus [52].

2.1.1 Eurostat duomenų šaltiniai ir jų apribojimai

Eurostat duomenų bazė yra pagrindinis šio tyrimo duomenų šaltinis. Eurostat, būdamas Europos Komisijos statistikos tarnyba, renka, apdoroja ir standartizuoja statistinius duomenis iš visų ES valstybių narių [53]. NUTS 2 lygmens duomenys renkami bendradarbiaujant su nacionalinėmis statistikos institucijomis, kurios pateikia duomenis pagal suderintus Eurostat metodologinius reikalavimus.

Duomenų rinkimo procesas grindžiamas keletą teisinių ir metodologinių dokumentų. ESA 2010 (*angl. European System of National and Regional Accounts*) reglamentas (ES) Nr. 549/2013 nustato bendrus principus nacionalinių ir regioninių sąskaitų sudarymui. Šis reglamentas užtikrina, kad BVP, pridėtinė vertė pagal sektorius ir kiti makroekonominiai rodikliai skaičiuojami pagal vienodus principus visose valstybėse narėse [54].

NUTS klasifikacijos reglamentas (ES) 2016/2066 apibrėžia regioninį detalizavimo lygį ir užtikrina, kad duomenys rinkti vienodame teritoriniame lygmenyje. NUTS 2 regionai apibrėžiami taip, kad jų gyventojų skaičius svyruotų nuo 800 tūkstančių iki 3 milijonų, kas užtikrina palyginamumą tarp regionų [55], [56].

Tačiau Eurostat duomenų šaltinis turi ir tam tikrų apribojimų, kuriuos svarbu pripažinti ir įvertinti jų įtaką tyrimo rezultatams. Pirmasis apribojimas yra duomenų pilnumo problema. Kaip matyti 6 priede, daugelis specializuotų rodiklių turi didelį trūkstamų reikšmių kiekį. Tai atsiranda dėl kelių priežasčių: ne visos valstybės narės renka visus rodiklius, kai kurių rodiklių rinkimas prasidėjo vėliau nei kitų, mažesnės ar ekonomiškai silpnesnės valstybės narės ne visada turi pakankamai resursų surinkti visus reikalaujamus duomenis.

Antrasis apribojimas yra duomenų peržiūros ir koregavimo ciklas. Eurostat duomenys nėra galutiniai - jie peržiūrimi ir tikslinami po pirminio paskelbimo. Tai reiškia, kad duomenys, naudojami

šiame tyrime, gali būti patikslinti ateityje, kas teoriškai galėtų šiek tiek pakeisti tyrimo rezultatus, jei tyrimas būtų kartojamas su naujomis duomenų versijomis.

Trečiasis apribojimas yra laiko delsimas (*angl. time lag*). Regioniniai duomenys dažnai paskelbami su didesniu vėlavimu nei nacionaliniai duomenys. Pavyzdžiui, BVP regioniniu lygmeniu paprastai paskelbamas su maždaug dvejų metų vėlavimu, kai tuo tarpu nacionalinis BVP skelbiamas su kelių mėnesių vėlavimu. Tai reiškia, kad pats naujausias duomenų laikotarpis šiame tyrime yra 2022 - 2023 metai, nors tyrimas atliekamas 2025 metais.

Ketvirtasis apribojimas yra metodologinių skirtumų problema. Nors Eurostat nustato bendrus duomenų rinkimo principus, nacionalinės statistikos institucijos turi tam tikrą laisvę interpretuoti šiuos principus. Tai gali lemti nedidelius metodologinius skirtumus tarp valstybių narių, kas teoriškai gali įtakoti regionų palyginamumą.

Penktasis apribojimas yra duomenų agregavimo lygmuo. Kai kurie rodikliai pateikiami tik NUTS 1 arba nacionaliniu lygmeniu, o ne NUTS 2 lygmeniu. Tokiais atvejais buvo priimtas sprendimas neįtraukti tokių rodiklių į tyrimą, kad būtų išlaikytas vienodas regioninis detalumas visame duomenų rinkinyje.

Nepaisant šių apribojimų, Eurostat duomenų bazė išlieka geriausiu prieinamu šaltiniu regioninių ekonominių duomenų tyrimams ES kontekste. Duomenų standartizacija, platumas ir prieinamumas daro Eurostat nepakeičiamą šaltiniu akademiniams tyrimams regioninės ekonomikos srityje.

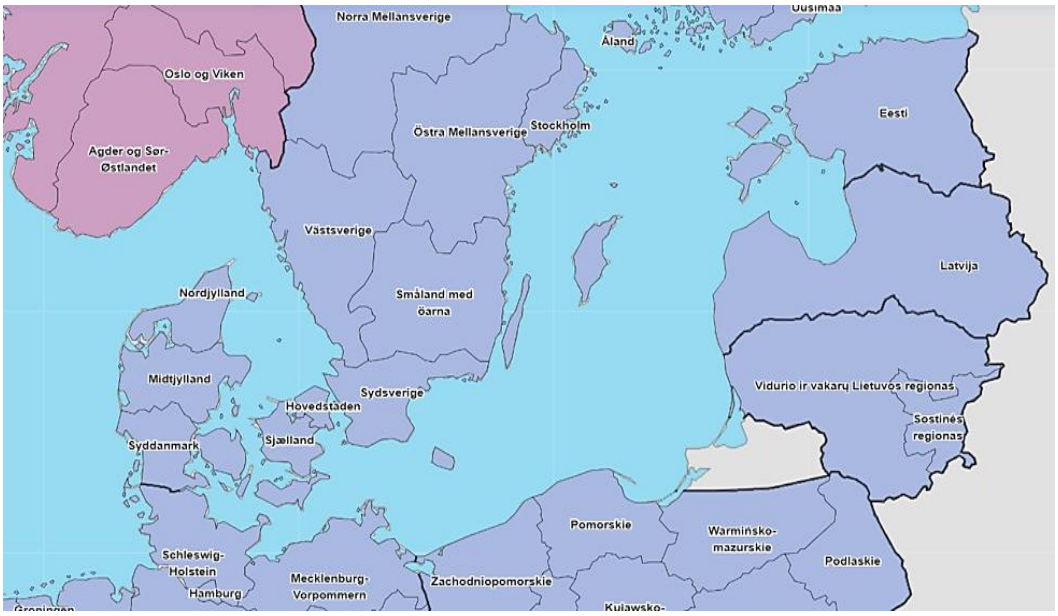
2.1.2 NUTS klasifikacijos struktūra ir lygmenys

NUTS sistema skirsto ES šalių narių teritorijas į tris pagrindinius hierarchinius lygmenis: NUTS 1, NUTS 2 ir NUTS 3. Kiekvienas lygmuo atspindi skirtingo dydžio administracinius ar statistinius regionus, o jų dydis nustatomas pagal gyventojų skaičiaus kriterijus [52]. NUTS 1 lygmuo apima didžiausius regionus, kurių gyventojų skaičius svyruoja nuo 3 iki 7 milijonų. 2024 m. duomenimis, ES teritorijoje yra 92 NUTS 1 regionai [52]. Šie regionai dažniausiai atitinka stambiausius administracinius vienetus arba statistines teritorijas nacionaliniu lygmeniu. Pavyzdžiui, Vokietijoje NUTS 1 lygmenį sudaro 16 federalinių žemių (Bundesländer), o Prancūzijoje - makroregionai [57].

NUTS 2 lygmuo, kuris yra šio magistro darbo tyrimo objektas, apima vidutinio dydžio regionus su gyventojų skaičiumi nuo 800 tūkstančių iki 3 milijonų. 2024 m. duomenimis, ES teritorijoje yra 244 NUTS 2 regionai [52]. Šis lygmuo yra ypač svarbus ES regioninei politikai, nes būtent NUTS 2 regionai yra pagrindiniai ES sanglaudos politikos ir struktūrinių fondų skirstymo vienetai [58]. NUTS 2 regionai dažnai atitinka administracinius vienetus, tokius kaip provincijos ar

apskritys. Lietuvoje NUTS 2 lygmenį sudaro du regionai - Sostinės regionas ir Vidurio bei vakarų Lietuvos regionas.

NUTS 3 lygmuo apima smulkiausius regionus, kurių gyventojų skaičius svyruoja nuo 150 tūkstančių iki 800 tūkstančių. 2024 m. sąjungoje yra 1 165 NUTS 3 regionai [52]. Šie regionai naudojami detalesniems regioninės raidos tyrimams. Lietuvoje NUTS 3 lygmenį sudaro 10 apskričių.



Šaltinis: [59]

9 pav. NUTS 2 regionai Europos Sąjungoje (2024 m.)

Kaip matyti 9 paveiksle, NUTS 2 regionai apima visą Europos Sąjungos teritoriją, o jų dydis ir konfigūracija skiriasi priklausomai nuo šalies dydžio ir administracinės struktūros. Aiškiai matomas regionų skaičiaus skirtumas tarp didesnių šalių (pavyzdžiui Lenkijos, Švedijos, Danijos) ir mažesnių valstybių narių.

2.1.3 NUTS 2 regionų ekonominis heterogeniškumas ir taikymas mašininio mokymosi tyrimuose

Vienas iš pagrindinių NUTS 2 regionų ypatumų - jų didelis ekonominis heterogeniškumas. Kai kurie NUTS 2 regionai pasižymi labai aukštu BVP vienam gyventojui ir išsivysčiusia ekonomine struktūra, tuo tarpu kiti regionai, ypač Rytų ir Pietų Europos šalyse, vis dar susiduria su ekonominės plėtros iššūkiais [60]. Ši ekonominė įvairovė lemia skirtingus regionų atsparumo ekonominėms krizėms lygius.

2008 - 2009 metų finansinė krizė parodė, kad pramoniniai regionai, priklausomi nuo eksporto, patyrė žymiai didesnius ekonominio nuosmukio padarinius nei paslaugų sektoriui orientuoti regionai [61]. COVID-19 pandemija taip pat atskleidė regionų ekonominio atsparumo skirtumus - NUTS 2

lygmens tyrimas parodė, kad regionai su aukštesniu ekonominiu išsivystymu ir diversifikuota ekonomika greičiau atsigavo [62].

NUTS 2 lygmens duomenų standartizacija ir prieinamumas per Eurostat duomenų bazę sudaro palankias sąlygas taikyti mašininio mokymosi modelius regioninių ekonominių rodiklių prognozavimui. Standartizuoti duomenys leidžia palyginti regionus tarpusavyje, identifikuoti bendrus ekonominės raidos modelius ir nustatyti veiksnius, turinčius didžiausią įtaką regionų ekonominiam atsparumui.

Tačiau NUTS 2 duomenų specifika - santykinai nedidelis stebėjimų skaičius laiko eilutėse (metiniai duomenys) ir didelis požymių skaičius - kelia specifinius reikalavimus mašininio mokymosi modelių pasirinkimui ir hiperparametrų optimizavimui. Random Forest ir XGBoost algoritmai, kurie bus analizuojami šiame magistro darbe, pasižymi gebėjimu efektyviai dirbti su tokio pobūdžio duomenimis, automatiškai identifikuoti svarbiausius požymius ir užtikrinti aukštą prognozavimo tikslumą net ir su ribotais istoriniais duomenimis.

2.2 Analizuojamų ekonominių rodiklių atranka ir pagrindimas

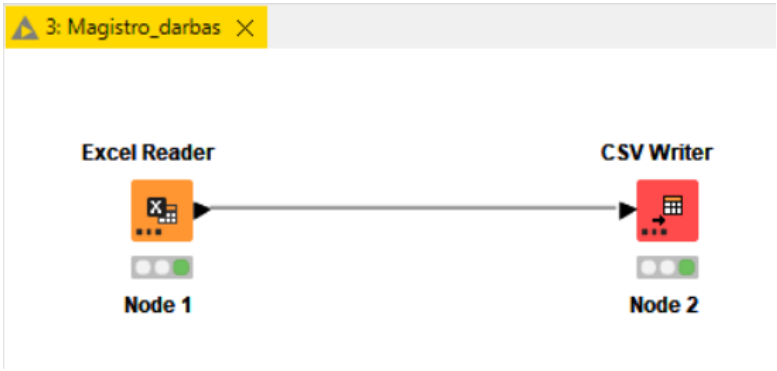
Regioninių ekonominių rodiklių tyrimas NUTS 2 lygmeniu reikalauja sistemingo požiūrio į duomenų rinkimą, paruošimą ir atrankos kriterijų taikymą. Šiame skyriuje aprašoma visa duomenų paruošimo eiga - nuo pradinio duomenų surinkimo iš Eurostat duomenų bazės iki galutinio analizei tinkamo duomenų rinkinio formavimo. Tyrimo metodologija apima kelias nuoseklias stadijas: duomenų agregavimą, formato konvertavimą, rodiklių atrankos pagrindimą bei trūkstamų reikšmių analizę.

Duomenų paruošimo procesas prasideda Eurostat duomenų bazėje publikuojamų regioninių ekonominių rodiklių rinkinių atranka ir jų atsisiuntimu, pasirenkant NUTS 2 lygmens duomenis. Eurostat, būdamas pagrindiniu ES statistikos šaltiniu, teikia standartizuotus regioninius duomenis NUTS 2 lygmeniu, kas užtikrina duomenų palyginamumą tarp skirtingų regionų ir laiko periodų. Tyrimo tikslams pasiekti buvo surinkti duomenys apimantys devynias pagrindines ekonominių rodiklių kategorijas: užimtumo rodiklius, pramonės struktūros rodiklius, įmonių demografijos rodiklius, gyventojų charakteristikas, žmogiškojo kapitalo rodiklius, infrastruktūros rodiklius, inovacijų rodiklius bei institucinio valdymo kokybės rodiklius.

Kaip matyti 4 priede pateiktame rodiklių sąraše, pradiniame etape buvo identifikuoti 97 potencialūs ekonominiai rodikliai, apimantys platų ekonominės veiklos spektrą. Tačiau pradinis duomenų rinkinys turėjo ženklų nevienalytiškumą - kai kurių rodiklių trūkstamų reikšmių dalis siekė net 91,49 procento (pavyzdžiui, rodiklio SCH_FS_18 atveju). Tokia situacija yra tipiška regioninių duomenų analizei, kur skirtingi rodikliai renkami nevienodu intensyvumu, o dalies rodiklių duomenų rinkimas prasidėjo vėlesniais metais nei kitų.

Duomenų agregavimo procesas buvo atliekamas naudojant Power BI platformą, kuri leido efektyviai sujungti duomenis iš įvairių Eurostat lentelių į vieną integruotą duomenų rinkinį. Agregavimas buvo būtinas dėl kelių priežasčių. Pirma, skirtingi ekonominiai rodikliai Eurostat duomenų bazėje yra saugomi atskirose lentelėse su skirtingomis struktūromis ir formatavimu. Antra, reikėjo užtikrinti NUTS 2 regioninio detalumo išsaugojimą visame duomenų rinkinyje. Trečia, agregavimo metu buvo galima standartizuoti laiko periodus ir užtikrinti vienodą duomenų reprezentavimą. Agregavimo rezultatas - duomenų failas *Irmantui.xlsx*, kuriame visi ekonominiai rodikliai susieti su konkrečiais NUTS 2 regionais ir metais.

Tolesnio duomenų apdorojimo etape buvo taikomas KNIME (*angl. Konstanz Information Miner*) - atvirojo kodo duomenų analizės įrankis. KNIME aplinkoje buvo sukurta paprasta, bet efektyvi darbo eiga (*angl. workflow*), susidedanti iš dviejų pagrindinių mazgų: Excel Reader ir CSV Writer (žiūrėti 10 paveikslėlį). Šis konvertavimo procesas buvo būtinas, nes Python mašininio mokymosi bibliotekos efektyviau dirba su CSV formato failais nei su Excel failais. CSV formatas užtikrina greitesnį duomenų įkėlimą, mažesnį atminties naudojimą ir geresnį suderinamumą su įvairiomis duomenų analizės bibliotekomis. Konvertavimo rezultatas - duomenų failas *Įvesties_duomenų_nuts2_rodikliai_2025-11-04.csv*, kuris tapo pagrindiniu duomenų šaltiniu visai tolesnei analizei.



Šaltinis: sudaryta autoriaus

10 pav. Duomenų konvertavimas KNIME aplinkoje

Kaip matyti 7 priede pateiktoje duomenų paruošimo metodinėje eigoje, procesas apima septynias nuoseklias stadijas: duomenų surinkimą iš Eurostat duomenų bazės, rodiklių atranką pagal NUTS 2 lygį, duomenų agregavimą naudojant Power BI, duomenų failo konvertavimą iš .xlsx į .csv naudojant KNIME, duomenų valymą naudojant Python, trūkstamų reikšmių analizę ir koreliacijų analizę. Kiekviena stadija atlieka specifinę funkciją bendroje duomenų paruošimo architektūroje ir prisideda prie galutinio, analizei tinkamo, duomenų rinkinio formavimo.

Python programavimo kalba buvo naudojama duomenų valymo etape, kurio metu buvo identifikuoti ir pašalinti rodikliai, neatitinkantys tyrimo kriterijų. Buvo identifikuota ir pašalinta 19

rodiklių, neatitinkančių tyrimo kriterijų, daugiausia dėl jų metodologinio nesuderinamumo su NUTS 2 lygmens duomenimis. Šie pašalinti rodikliai pažymėti raudona spalva 4 priedo lentelėje.

Šių rodiklių pašalinimas buvo metodologiškai pagrįstas keliais argumentais. Pirma, EU serijos indikatoriai (pavyzdžiui, SCH_IS_EU_01 - SCH_IS_EU_15) neužtikrina nuoseklaus NUTS 2 regioninio detalumo, nes tai išvestiniai rodikliai skaičiuojami kaip santykiniai rodikliai lyginant su ES atitinkamo rodiklio vidurkiu. Antra, daugelis šių rodiklių turėjo labai didelį trūkstamų duomenų procentą, kas galėtų destabilizuoti trūkstamų reikšmių užpildymo procesą. Trečia, dalis jų dubliavo kitus rodiklius arba buvo išvestiniai rodikliai, apskaičiuojami iš jau esamų bazinių rodiklių. Ketvirta, jų įtraukimas būtų iškraipęs koreliacijų matricą, trūkstamų reikšmių užpildymo rezultatus ir rodiklių sinergijos analizę, kadangi šie rodikliai atspindi ne absoliučias regionų charakteristikas, o jų santykinę padėtį Europos Sąjungos kontekste.

2.2.1 Metodologinis rodiklių atrankos pagrindimas

Ekonominių rodiklių atranka mašininio mokymosi modelių treniravimui yra svarbus etapas, lemiantis galutinių prognozių kokybę ir patikimumą. Tyrimo tikslams pasiekti buvo taikomi keturi pagrindiniai atrankos kriterijai: duomenų prieinamumas ir kokybė, regioninis detalumas, ekonominė reikšmė bei tarpusavio suderinamumas.

Visi pasirinkti rodikliai yra oficialiai renkami ir publikuojami Eurostat duomenų bazėje NUTS 2 lygmeniu, atitinka ESA 2010 reglamento (ES) Nr. 549/2013 reikalavimus ir užtikrina duomenų standartizaciją visose ES šalyse [63]. Šis teisinis pagrindas garantuoja, kad duomenys renkami pagal vienodus metodologinius principus, kas užtikrina jų palyginamumą tarp skirtingų regionų ir laiko periodų.

Kaip matyti 4 priede pateiktame rodiklių sąraše, galutiniam tyrimui buvo atrinkti 78 ekonominių rodiklių, apimančių devynias pagrindines kategorijas. Pirmoji kategorija - užimtumo rodikliai (SAN_ED_01) - apibūdina bazines regiono darbo rinkos charakteristikas, susijusias su gyventojų dalyvavimu ekonominėje veikloje ir darbo rinkos struktūra. Šie rodikliai leidžia įvertinti regiono darbo rinkos būklę ir jos stabilumą ekonominių svyravimų kontekste [64].

Antroji kategorija - pramonės struktūros rodikliai (SCH_IS_01 - SCH_IS_15) - atskleidžia regiono ekonominės veiklos sudėtį pagal NACE sektorius. Kaip matyti 4 priede, šie rodikliai matuoja įvairių ekonomikos sektorių (žemės ūkio, pramonės, statybos, paslaugų) sukuriamos pridėtinės vertės dalį. Ekonominės struktūros įvairovė yra svarbus veiksnys, lemiantis regiono atsparumą - diversifikuota ekonomika geriau priešinasi sektoriniams šokams nei koncentruota į vieną ar kelis sektorius [65].

Trečioji kategorija - įmonių demografijos rodikliai (SCH_FS_01 - SCH_FS_28) - apibūdina verslo dinamiką regione. Šie rodikliai apima veikiančių įmonių skaičių, naujai įsteigtų įmonių

rodiklius, veiklą nutraukusių įmonių statistiką bei įmonių išgyvenamumą. Verslo demografija yra svarbus inovacijų ir ekonominio dinamiškumo indikatorius - regionai su aktyvia įmonių kūrimosi ir raidos aplinka dažnai demonstruoja didesnį ekonominį atsparumą [66].

Ketvirtoji kategorija - gyventojų charakteristikos (SAN_SZ_01, SAN_SZ_02, SAN_PD_01, SAN_UR_01 ir SAN_UR_02) - apima demografinius ir erdvinius regiono parametrus. Gyventojų skaičius, tankis ir urbanizacijos lygis yra fundamentalūs regioninio potencialo rodikliai, lemiantys rinkos dydį, darbo jėgos prieinamumą ir aglomeracijos efektus [67], [68].

Penktoji kategorija - žmogiškojo kapitalo rodikliai (HC(Lf)Q_EDUC_01 - HC(Lf)Q_EDUC_08) - matuoja regiono gyventojų išsilavinimo lygį ir švietimo sistemos parametrus. Aukštasis išsilavinimas, dalyvavimas švietimo programose ir ankstyvas mokyklos baigimas yra svarbiausieji žmogiškojo kapitalo kokybės rodikliai, tiesiogiai susiję su regiono gebėjimu pritraukti investicijas ir kurti aukštos pridėtinės vertės produkciją [69], [70], [71], [72], [73], [74], [75].

Šeštoji kategorija - infrastruktūros rodikliai (INFRA_TR_01, INFRA_TR_05, INFRA_TO_01, INFRA_TO_03, INFRA_HLTH_01, INFRA_HLTH_03, INFRA_ICT_01 ir INFRA_ICT_02) - apibūdina regiono fizinės ir skaitmeninės infrastruktūros išsivystymo lygį bei jos panaudojimo intensyvumą. Transporto infrastruktūros rodikliai apima tiek fizinį pajėgumą (automagistralių ilgį kilometrais), tiek infrastruktūros naudojimo rezultatus (oro transportu pervežtų keleivių skaičių), kurie kartu atspindi transporto sistemos prieinamumą ir efektyvumą. Sveikatos priežiūros prieinamumo, turizmo infrastruktūros ir interneto skvarbos rodikliai leidžia įvertinti viešųjų ir komercinių paslaugų infrastruktūros kokybę, turinčią reikšmingą poveikį regiono konkurencingumui, gyventojų mobilumui ir investicijų patrauklumui [76], [77], [78], [79], [80], [81], [82], [83], [84], [85].

Septintoji kategorija - inovacijų rodikliai (INNO_RD_01) - matuoja mokslinių tyrimų ir eksperimentinės plėtros intensyvumą regione. MTEP išlaidos yra vienas svarbiausių ilgalaikio ekonominio augimo veiksnių, nes jos tiesiogiai susijusios su technologijų diegimu, produktyvumo augimu ir ekonominės struktūros modernizavimu [86].

Aštuntoji kategorija - institucinio valdymo kokybės rodikliai (IQG_01 - IQG_06) - įvertina regiono valdžios institucijų veiklos kokybę, nešališkumą ir korupcijos lygį. Europos kokybės indeksas (EQI) ir kiti institucinio valdymo rodikliai yra vis labiau pripažįstami kaip svarbūs ekonominės plėtros veiksniai, nes jie veikia investicijų klimatą, verslo sąnaudas ir visuomenės pasitikėjimą institucijomis [87].

Devintoji kategorija - rezultatiniai darbo rinkos ir ekonominės struktūros rodikliai (Y_01 ir Y_02) - atspindi apibendrintus regionų (NUTS 2) ekonominės veiklos ir užimtumo rezultatus. Y_01 rodiklis žymi darbuotojų dalį technologijų ir žinių intensyviuose sektoriuose pagal NACE Rev. 2 klasifikaciją [88], išreikštą procentine dalimi nuo viso užimtumo, ir parodo regiono orientaciją į

aukštos pridėtinės vertės veiklas bei inovacinį potencialą. Y_02 nusako bendrą užimtumo lygį 15 - 74 metų gyventojų grupėje ir atspindi darbo rinkos aktyvumą bei bendrą ekonominės veiklos intensyvumą. Šie rodikliai interpretuojami kaip integruoti regiono ekonominio pajėgumo ir struktūrinės raidos rezultatai [89], [90], [91].

Rodiklių atrankos metu buvo atsižvelgta ir į praktinį jų tinkamumą mašininio mokymosi modeliams. Pirma, visi atrinkti rodikliai yra kiekybiniai, kas palengvina jų apdorojimą algoritmuose. Antra, rodikliai turi pakankamą laiko eilučių ilgį, leidžiantį identifikuoti tendencijas ir ciklišKumą. Trečia, jų skalės ir pasiskirstymai yra pakankamai homogeniški tarp regionų, kas sumažina būtinybę taikyti sudėtingas transformacijas.

2.2.2 Rodiklių tarpusavio papildomumas ir sinergija

Ekonominiai rodikliai retai veikia izoliuotai - jų tarpusavio sąveika formuoja kompleksinę regiono ekonominę sistemą. Rodiklių tarpusavio papildomumas ir sinergija buvo analizuojami naudojant koreliacijų matricą su hierarchiniu klasterizavimu (žr. 5 priedą). Ši metodologija leidžia ne tik identifikuoti statistiškai reikšmingus ryšius tarp rodiklių, bet ir grupuoti juos į homogeniškas klases pagal jų elgesio panašumą [92].

Kaip matyti 5 priede pateiktoje koreliacijų matricoje, aiškiai identifikuojamos kelios pagrindinės rodiklių grupės su aukšta tarpusavio koreliacija. Pirmoji grupė apima pramonės struktūros rodiklius (SCH_IS_02, SCH_IS_03, SCH_IS_05, SCH_IS_08, SCH_IS_11 ir SCH_IS_12), kuriems būdinga didelė ir labai didelė tarpusavio koreliacija (dažnai viršijanti 0,7) [93]. Tai rodo, kad skirtingų ekonominės veiklos sektorių pridėtinės vertės dalys yra glaudžiai susijusios - pavyzdžiui, regionai su stipria apdirbamąja pramone (SCH_IS_03) dažnai pasižymi ir gerai išvystytais prekybos bei transporto sektoriais (SCH_IS_05).

Antroji aiškiai identifikuojama grupė apima įmonių demografijos rodiklius (SCH_FS_13, SCH_FS_14, SCH_FS_15, SCH_FS_17, SCH_FS_21 ir SCH_FS_23). Šių rodiklių tarpusavio koreliacija taip pat yra labai aukšta, kas rodo, kad verslo dinamika regione yra sisteminis reiškinys - regionai su aukštu įmonių steigimo rodikliu dažnai pasižymi ir didesniu įmonių demografijos intensyvumu bei užimtumo augimu naujose įmonėse.

Trečioji grupė apima žmogiškojo kapitalo rodiklius (HC(Lf)Q_EDUC_01, HC(Lf)Q_EDUC_02, HC(Lf)Q_EDUC_05 ir HC(Lf)Q_EDUC_06). Aukšto išsilavinimo lygio regionai dažnai siejasi su didesniu dalyvavimo švietimo programose lygiu ir mažesniu ankstyvos mokyklos baigimo rodikliu. Tai patvirtina švietimo sistemos kokybės svarbą, formuojant ilgalaikį žmogiškojo kapitalo potencialą.

Ketvirtoji grupė apima infrastruktūros rodiklius (INFRA_TR_01, INFRA_TR_02, INFRA_TR_03, INFRA_TO_01, INFRA_TO_02 ir INFRA_TO_03). Transporto infrastruktūros

išsivystymas (automagistralių tinklas, oro transporto keleivių srautai) glaudžiai koreliuoja su turizmo intensyvumu, kas patvirtina transporto prieinamumo svarbą regiono patrauklumui turistams ir investuotojams.

Itin reikšmingos yra neigiamos koreliacijos tarp kai kurių rodiklių. Pavyzdžiui, žemės ūkio sektoriaus pridėtinės vertės dalis (SCH_IS_01) pasižymi neigiama koreliacija su daugeliu modernių paslaugų sektorių rodiklių, kas atspindi ilgalaikes ekonominės struktūros transformacijos tendencijas - palaipsnį perėjimą nuo agrarinės prie paslaugų ir žiniomis grįstos ekonomikos.

Rodiklių sinergijos analizė atskleidžia, kad tam tikrų rodiklių kombinacijos sukuria stipresnius efektus nei jų atskira įtaka. Pavyzdžiui, aukštas žmogiškojo kapitalo lygis (HC(Lf)Q_EDUC_02) kartu su intensyviomis mokslinių tyrimų ir eksperimentinės plėtros išlaidomis (INNO_RD_01) sudaro palankią aplinką technologijų ir inovacijų sektoriaus plėtrai. Panašiai, gera transporto infrastruktūra (INFRA_TR_03) kartu su aukštu urbanizacijos lygiu (SAN_UR_01) prisideda prie aglomeracijos efektų stiprėjimo ir regioninio konkurencingumo didėjimo.

Koreliacijų analizė taip pat atskleidė keletą rodiklių, kurie pasižymi santykinai silpnu ryšiu su dauguma kitų analizuojamų rodiklių. Pavyzdžiui, institucinio valdymo kokybės rodikliai (IQG_01 - IQG_06) demonstruoja vidutiniškai silpną koreliaciją su daugeliu ekonominės struktūros rodiklių, kas rodo, kad valdymo kokybė yra santykinai nepriklausomas veiksnys, veikiantis ekonominę raidą per kitus, netiesioginius kanalus, o ne tiesiogiai per ekonominės struktūros parametrus.

Hierarchinis klasterizavimas (žr. 5 priedo dendrogramą) patvirtina, kad ekonominiai rodikliai gali būti grupuojami į keturis pagrindinius klasterius [94]: (1) ekonominės struktūros klasteris, apimantis pramonės ir paslaugų sektorių rodiklius; (2) verslo dinamikos klasteris, apimantis įmonių demografijos rodiklius; (3) žmogiškojo kapitalo ir inovacijų klasteris; (4) infrastruktūros ir urbanizacijos klasteris. Ši klasterizacija atspindi fundamentalias regioninės ekonomikos dimensijas ir patvirtina rodiklių atrankos logiką.

2.3 Duomenų paruošimas ir apdorojimas trūkstamų reikšmių analizei

Duomenų paruošimas yra svarbus etapas, užtikrinantis, kad mašininio mokymosi modeliai galės efektyviai mokytis iš pateiktų duomenų. Python programavimo kalba su *pandas*, *numpy* ir *scikit-learn* bibliotekomis buvo naudojama visame duomenų paruošimo procese. Pradinis duomenų įkėlimas buvo atliekamas naudojant *pandas* biblioteką (žr. 11 pav.).

```
11  duomenu_failo_kelias = 'uploads/Įvesties_duomenų_nuts2_rodikliai_2025-11-04.csv'
12  df = pd.read_csv(duomenu_failo_kelias)
```

Šaltinis: sudaryta autoriaus

11 pav. Duomenų įkėlimo *python* kodo fragmentas

Po įkėlimo buvo atliktas preliminarus duomenų tyrimas (*angl. Exploratory Data Analysis, EDA*), įvertinant duomenų rinkinį statistiškai ir vizualiai. Buvo identifikuoti duomenų tipai, patikrintos minimalios ir maksimalios reikšmės, apskaičiuoti pagrindiniai aprašomosios statistikos rodikliai (vidurkis, mediana, standartinis nuokrypis, kvartiliai). Šis etapas leido identifikuoti potencialias anomalijas ir suprasti duomenų pasiskirstymo ypatumus.

Duomenų valymo etape buvo pašalinti rodikliai su per dideliu trūkstamų reikšmių kiekiu. Pradinis duomenų rinkinys apėmė 97 ekonominius rodiklius. Kaip minėta anksčiau, buvo pašalinti EU serijos indikatoriai ir kiti rodikliai, neatitinkantys tyrimo kriterijų. Po rodiklių analizės liko 78 ekonominiai rodikliai, kurių trūkstamų reikšmių pasiskirstymas pateiktas 6 priede. Trūkstamų reikšmių procentas buvo apskaičiuojamas kiekvienam rodikliui (žr. 12 pav.).

```
5  # Skaičiuojame trūkstamų reikšmių procentus
6  trukstamu_proc = df.isnull().mean() * 100
7  trukstamu_proc = trukstamu_proc.sort_values(ascending=False)
```

Šaltinis: sudaryta autoriaus

12 pav. Trūkstamų reikšmių procento apskaičiavimo kodo fragmentas

Kaip matyti 6 priede pateiktoje trūkstamų reikšmių procentinėje analizėje, rodiklių trūkstamumo lygis svyruoja nuo 0 iki 91,2 procento. Dauguma geografinių ir laiko identifikatorių (geo, year) neturi trūkstamų reikšmių, kas užtikrina, kad kiekvienas stebėjimas gali būti tiksliai priskirtas konkrečiam regionui ir metams. Tuo tarpu kai kurie specializuoti rodikliai, tokie kaip institucinio valdymo kokybės rodikliai (IQG_05, IQG_06) ir įmonių demografijos rodikliai (SCH_FS_19), turi labai didelį trūkstamumo lygį.

Duomenų rinkinio apimtis po valymo: 6100 stebėjimų (eilučių) × 80 požymių (stulpelių), iš kurių 2 yra identifikatoriai (geo, year), o likę 78 - ekonominiai rodikliai. Stebėjimų skaičius teoriškai atitinka 244 NUTS 2 regionų duomenis per 25 metų laikotarpį (2000 - 2024), tačiau ne visiems regionams duomenys prieinami visais metais. Analizuojamų NUTS 2 regionų kodų ir jų priskyrimo atitinkamoms šalims sąrašas pateikiamas 16 priede.

Duomenų tipų transformacijos šiame etape nebuvo atliekamos, nes visi ekonominiai rodikliai pradiniame rinkinyje jau buvo pateikti skaitine forma (float64 - 64 bitų slankiojo kablelio skaičiai), metų kintamasis year išsaugotas kaip sveikasis skaičius (int64), o kategorinis kintamasis geo išlaikytas kaip tekstinis regiono identifikatorius (object). Kategorinių kintamųjų transformavimas į skaitines reikšmes atliekamas vėlesniame tyrimo etape (žr. 3 skyrių), kai jie naudojami kaip įvesties požymiai modeliavimo procedūrose.

2.4 Trūkstamų reikšmių analizė

Trūkstamų reikšmių mechanizmo nustatymas yra svarbus etapas prieš pradedant trūkstamų reikšmių užpildymo procedūras, nes skirtingi trūkstamumo mechanizmai reikalauja skirtingų metodologinių sprendimų. Little ir Rubin [95] klasifikacija, išskiriantis MCAR (*angl. Missing Completely At Random*), MAR (*angl. Missing At Random*) ir MNAR (*angl. Missing Not At Random*) mechanizmus, suteikia teorinį pagrindą trūkstamų duomenų problemos analizei. Tinkamas trūkstamų reikšmių mechanizmo identifikavimas leidžia parinkti tinkamiausius trūkstamų reikšmių užpildymo metodus ir įvertinti galimą rezultatų šališkumą galutiniuose analizės rezultatuose.

Little's MCAR testas yra plačiai pripažintas statistinis metodas, skirtas patikrinti, ar duomenų trūkstamumas atitinka MCAR (visiškai atsitiktinio trūkstamumo) mechanizmą. Šis testas, pasiūlytas Roderick J. A. Little 1988 metais, grindžiamas daugiamačių normalių pasiskirstymų teorija ir leidžia formaliai įvertinti, ar trūkstamų reikšmių pasiskirstymas yra nepriklausomas nuo duomenų reikšmių [96].

Testo pagrindas yra palyginimas tarp stebimų vidurkių skirtinguose trūkstamumo šablonuose (*angl. missingness patterns*). Jei duomenys yra MCAR, tuomet kiekvienas trūkstamumo šablonas turėtų turėti panašius vidurkius stebimose reikšmėse. Little's testas formuluoja nulinę hipotezę H_0, kad duomenys yra MCAR, ir naudoja chi-kvadrato statistiką šiai hipotezei patikrinti [97].

Testo statistika apskaičiuojama pagal (18) formulę:

$$d^2 = \sum_{j=1}^{J} n_j \left(\bar{y}_j - \hat{\mu}\right)^T \hat{\Sigma}_{jj}^{-1} \left(\bar{y}_j - \hat{\mu}\right), \tag{18}$$

kur: n_j - stebėjimų skaičius j-ajame trūkstamumo šablone, $\bar{y}_j$ - stebimų reikšmių vidurkis tame šablone, $\hat{\mu}$ - bendrasis įvertintas vidurkis, $\hat{\Sigma}_{jj}$ - kovariacinė matrica stebimoms reikšmėms j-ajame šablone.

Python programavimo kalboje *Little's MCAR* testas buvo implementuotas naudojant specializuotą biblioteką, kuri iteratyviai modeliuoja kiekvieną kintamąjį su trūkstamomis reikšmėmis atsižvelgiant į kitų kintamųjų reikšmes. Šiame tyrime testas buvo taikomas tik skaitinių kintamųjų poaibei, sudarančiam 78 ekonominius rodiklius, turinčius bent vieną trūkstamą reikšmę duomenų rinkinyje.

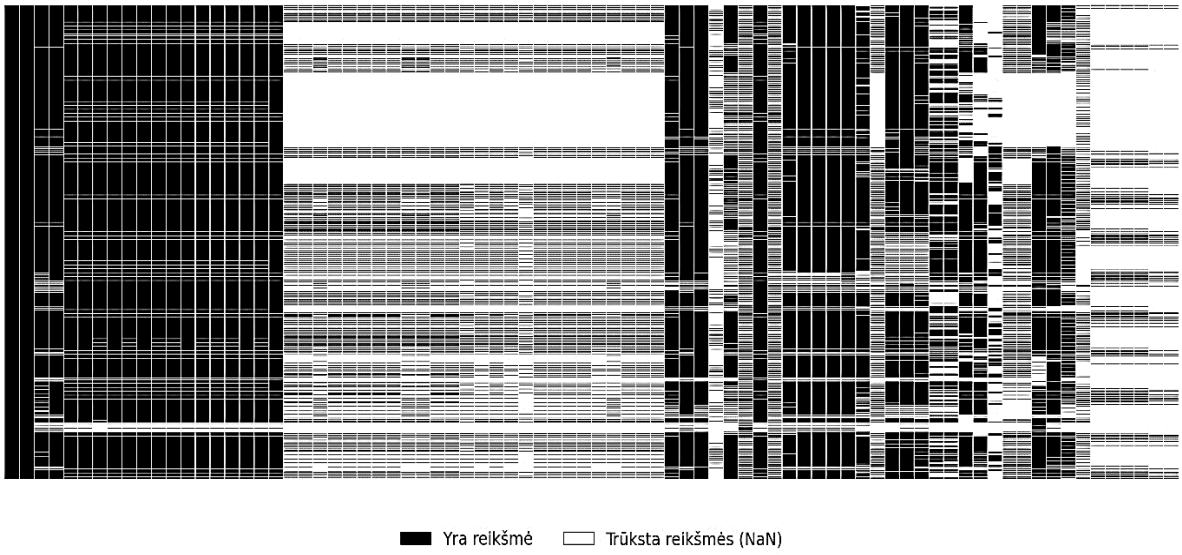
Testo rezultatai pateikti žemiau:

- Chi-kvadrato statistika (*angl. chi-square statistic*) $\chi^2 = 41962{,}019$;
- laisvės laipsniai (*angl. degrees of freedom*): $df = 17919$;
- $p < 0{,}0001$.

Gauta p reikšmė yra žymiai mažesnė nei įprastas statistinio reikšmingumo lygis ($\alpha = 0,05$), todėl nulinė hipotezė apie MCAR mechanizmą yra atmesta [96]. Tai reiškia, kad tikimybė gauti tokius arba dar ekstremalius rezultatus, jei duomenys tikrai būtų MCAR, yra statistiškai nereikšmingai maža. Kitaip tariant, egzistuoja statistiškai reikšmingas ryšys tarp trūkstamų reikšmių pasiskirstymo ir stebimų duomenų reikšmių, todėl rekomenduojama taikyti MAR prielaidą atitinkančius metodus.

Šie rezultatai leidžia daryti išvadą, kad NUTS 2 regionų ekonominių rodiklių duomenų rinkinyje trūkstamos reikšmės nėra visiškai atsitiktinės. Tikėtina MAR (atsitiktinai trūkstančių) struktūra reiškia, kad trūkstamų reikšmių tikimybė priklauso nuo kitų stebimų kintamųjų, bet ne nuo pačių trūkstamų reikšmių. Ekonominių duomenų kontekste tai gali reikšti, pavyzdžiui, kad mažesnių ar ekonomiškai silpnesnių regionų statistinės institucijos dažniau nesurenka tam tikrų specializuotų rodiklių duomenų, o šis trūkumas koreliuoja su regiono ekonominio išsivystymo lygiu, kurį galima stebėti per kitus rodiklius [98]. Atsižvelgiant į *Little MCAR* testo rezultatus, galima teigti, kad trūkstamų duomenų struktūra nėra visiškai atsitiktinė, todėl pažangūs mašininio mokymosi pagrįsti trūkstamų reikšmių užpildymo metodai šiame tyrime yra metodologiškai pagrįstas pasirinkimas.

Norint geriau suprasti šių trūkstamų reikšmių pasiskirstymo struktūrą, buvo atlikta ir vizuali jų analizė. Trūkstamų reikšmių vizuali analizė buvo atlikta naudojant Python *missingno* biblioteką, kuri suteikia intuityvius įrankius trūkstamumo šablonų identifikavimui. Pagrindinė vizualizacija - trūkstamumo matrica (*angl. missingness matrix*) - pateikta 13 paveiksle.



Šaltinis: sudaryta autoriaus

13 pav. Trūkstamų duomenų matrica NUTS 2 regionų ekonominiams rodikliams

Kaip matyti 12 paveiksle, trūkstamumo matrica vizualizuoja visą duomenų rinkinio struktūrą, kur juoda spalva žymi esamas reikšmes, o balta - trūkstamas reikšmes. Pirmieji du stulpeliai (geo ir year) yra visiškai užpildyti, kas patvirtina, kad kiekvienas stebėjimas turi tinkamą geografinį ir laiko

identifikatorių. Tai užtikrina, kad kiekvienas duomenų įrašas gali būti tiksliai priskirtas konkrečiam NUTS 2 regionui ir metams.

Likusių rodiklių trūkstamumo struktūra vizualizacijoje atskleidžia aiškią hierarchiją. Kairėje matricos pusėje koncentruojasi rodikliai su mažesniu trūkstamumo lygiu (tamsesnės spalvos dominuoja), tai daugiausia fundamentalūs ekonominiai rodikliai kaip BVP, užimtumas ir pramonės struktūros rodikliai. Dešinėje matricos pusėje vyrauja šviesesnės spalvos, atspindinčios aukštą trūkstamumo lygį - tai specializuoti rodikliai kaip institucinio valdymo kokybės indeksai (IQG serija) ir įmonių demografijos rodikliai (SCH_FS serija).

Svarbu pažymėti, kad trūkstamų reikšmių pasiskirstymas nėra visiškai atsitiktinis: jis aiškiai siejasi su tam tikrais laiko periodais ir regionais. Tai matyti iš šviesesnių horizontalių juostų, kurios žymi sistemiškai didesnį duomenų trūkumą konkrečiais laikotarpiais ar regionų grupėse. Tokia sistemiška trūkstamų reikšmių koncentracija tam tikruose laikotarpiuose ir regionų grupėse leidžia daryti prielaidą, kad trūkstamumo mechanizmas gali būti susijęs ne tik su stebimais kintamaisiais, bet ir su pačių trūkstamų reikšmių ar nestebimų veiksnių poveikiu, kas sudaro pagrindą neatmesti ir MNAR (*angl. Missing Not At Random*) scenarijaus.

Trūkstamumo dinamikos analizė laiko atžvilgiu pateikta 8 priede, kuris parodo vidutinę trūkstamų reikšmių dalį kiekvienais metais.

Kaip matyti 8 priede, trūkstamų reikšmių procentas laiko eilutėje pasižymi nelinijine dinamika. Ankstyvuoju laikotarpiu (2000 - 2007 m.) trūkstamumo lygis buvo aukštas ir stabilizavosi apie 57 - 61 %, kas iš dalies gali būti siejama su tuo, kad Eurostat duomenų rinkimo ir harmonizavimo procedūros šiame etape dar nebuvo visiškai standartizuotos visose valstybėse narėse. Po 2008 metų pastebimas nuoseklus trūkstamumo lygio mažėjimas iki minimumo 2017 - 2018 metais (apie 21 - 25 %), kai Eurostat duomenų rinkimo metodologijos ir kokybės standartai jau buvo pilnai įdiegti Europos statistikos sistemoje [99].

Tačiau nuo 2019 metų trūkstamumo lygis vėl pradeda augti, pasiekdamas 70,3% 2024 metais. Šis staigus augimas gali būti paaiškinamas keliais veiksniais. Pirma, naujausių metų duomenys vis dar yra preliminarūs ir nevisiškai surinkti publikavimo metu. Antra, COVID-19 pandemija 2020-2021 metais sutrikdė įprastus statistinių duomenų rinkimo procesus daugelyje šalių. Trečia, kai kurie specializuoti rodikliai (pavyzdžiui, institucinio valdymo kokybės indeksai) renkami tik periodiniais tyrimais, o ne kasmet.

Regioninė trūkstamumo analizė pateikta 9 priede, kuriame vaizduojama vidutinė trūkstamų reikšmių dalis pagal NUTS 2 regionus. Šis pateiktas grafikas leidžia identifikuoti regionus, kuriuose fiksuojamas didžiausias sisteminis statistinių duomenų trūkumas.

Kaip matyti 9 priede, trūkstamumo lygis tarp analizuotų regionų svyruoja nuo 59,2 % iki 90,2 %. Didžiausias trūkstamumo lygis nustatytas Portugalijos regionuose - PT1C, PT1B, PT1D, PT1G ir

PT1A, kuriuose trūkstamų reikšmių dalis siekia 89,9 - 90,2 %. Šie rezultatai gali būti siejami su nacionalinių statistikos sistemų struktūriniais ypatumais ir duomenų rinkimo praktikos skirtumais.

Santykinai aukštas trūkstamumo lygis taip pat pastebimas Nyderlandų regionuose NL35 ir NL36 (apie 84 %), kurie išsiskiria iš bendro tendencijų konteksto.

Vidutinio lygio trūkstamumas (apie 62 - 66 %) fiksuojamas keliuose Vidurio ir Vakarų Europos regionuose, įskaitant Kroatijos regionus HR05, HR02 ir HR06, Airijos regioną IE04, taip pat Vokietijos regioną DED4 ir kelis kitus regionus (IE05, IE06, DED5, DEB2, PL92). Tai rodo gana tolygų, tačiau vis dar reikšmingą duomenų trūkumo paplitimą šiose teritorijose.

Mažiausias trūkstamumo lygis nustatytas Vokietijos regionuose DE23 ir DEB3 bei Prancūzijos regione FRY3, kur trūkstamų reikšmių dalis siekia apie 59,2 - 59,3 %. Tai gali atspindėti aukštesnį statistinių sistemų brandos lygį ir nuoseklesnes duomenų rinkimo praktikas šiose šalyse.

Little's MCAR testo rezultatai, parodę MAR struktūrą vietoje MCAR, turi svarbias praktines implikacijas trūkstamų reikšmių užpildymo metodų pasirinkimui. MAR mechanizmas reiškia, kad paprasti užpildymo metodai, tokie kaip vidurkio ar medianos užpildymas, gali sukelti šališkus rezultatus, nes jos ignoruoja sistemingus ryšius tarp trūkstamų ir stebimų duomenų.

Tinkamiausi metodai MAR struktūros duomenims yra:

1. Random Forest pagrindu atliekamas trūkstamų reikšmių užpildymas - mašininio mokymosi metodas, leidžiantis automatiškai identifikuoti sudėtingus netiesinius ryšius tarp kintamųjų ir efektyviai išnaudoti jų tarpusavio priklausomybes [100].
2. XGBoost pagrindu atliekamas trūkstamų reikšmių užpildymas - gradientinio stiprinimo algoritmas, gebantis tvarkyti trūkstamas reikšmes ir modeliuoti sudėtingas priklausomybes tarp kintamųjų, ypač esant nevienalyčiams duomenims[36].

Abu metodai geba adaptuotis prie MAR struktūros, naudodami informaciją iš stebimų kintamųjų trūkstamoms reikšmėms prognozuoti. Tai ypač svarbu NUTS 2 regionų kontekste, kur ekonominiai rodikliai yra glaudžiai susiję per fundamentalius ekonominius mechanizmus.

2.5 Skyriaus apibendrinimas

Atlikta trūkstamų reikšmių analizė parodė, kad NUTS 2 regionų ekonominių rodiklių duomenų rinkinys pasižymi sistemiška ir nevienalyte trūkstamumo struktūra, kuri neatitinka visiškai atsitiktinio trūkstamumo (MCAR) prielaidos. Little’s MCAR testo rezultatai ($\chi^2 = 41962{,}019$; $df = 17919$; $p < 0{,}0001$) statistiškai reikšmingai atmetė hipotezę, kad duomenys yra visiškai atsitiktinai trūkstantys, todėl nustatyta, kad duomenų struktūra atitinka MAR (*angl. Missing At Random*) mechanizmo prielaidą.

Vizualinė analizė papildomai patvirtino testinių rezultatų išvadas. Trūkstamumo matrica atskleidė aiškią hierarchiją tarp kintamųjų grupių: fundamentalūs ekonominiai rodikliai pasižymi

didesniu duomenų pilnumu, tuo tarpu specializuoti rodikliai, tokie kaip institucinio valdymo ir įmonių demografijos indikatoriai, turi sistemingai didesnį trūkstamumo lygį. Laiko analizė parodė ryškią trūkstamumo dinamiką - didesnį trūkstamumą ankstyvaisiais laikotarpiais, reikšmingą sumažėjimą iki 2017 - 2018 metų ir pakartotinį augimą vėlesniais metais. Regioninė analizė identifikavo sistemingus skirtumus tarp šalių, atspindinčius nevienodą statistinių sistemų brandą ir duomenų rinkimo praktiką.

Identifikuotas MAR mechanizmas turi esminę reikšmę tolesnei tyrimo eigai, nes jis lemia trūkstamų reikšmių užpildymo metodų pasirinkimą. Atsižvelgiant į skaičiavimo sudėtingumą, mastelio apribojimus ir literatūroje nustatytus veiksmingumo skirtumus, eksperimentiniam tyrimui pasirinkti ansambliniai mašininio mokymosi modeliai - Random Forest ir XGBoost. Paprasti trūkstamų reikšmių užpildymo metodai, tokie kaip vidurkio ar medianos taikymas, šiuo atveju laikomi nepakankamais dėl galimo sisteminio šališkumo. Todėl pasirenkami pažangesni sprendimai, pagrįsti Random Forest ir XGBoost algoritmais, kurie leidžia išnaudoti stebimų kintamųjų tarpusavio ryšius trūkstamoms reikšmėms prognozuoti.

Šio skyriaus išvados sudaro metodologinį pagrindą trečiajam darbo skyriui, kuriame bus atliekamas skirtingų mašininio mokymosi modelių palyginimas, vertinant jų efektyvumą trūkstamų reikšmių prognozavimui ir užpildymui NUTS 2 regionų ekonominiuose duomenyse. Galiausiai bus analizuojama, kaip skirtingi trūkstamų reikšmių užpildymo metodai prisideda prie tikslesnių ir patikimesnių regionų ekonominės analizės rezultatų užtikrinimo.

3. EKSPERIMENTINIS TRŪKSTAMŲ REIKŠMIŲ PROGNOZAVIMO TYRIMAS NUTS 2 REGIONŲ DUOMENYSE

3.1 Eksperimentinio tyrimo metodika

Šio eksperimentinio tyrimo tikslas yra įvertinti mašininio mokymosi modelių - Random Forest ir XGBoost - efektyvumą prognozuojant trūkstamas ekonominių rodiklių reikšmes NUTS 2 regionų duomenų rinkinyje. Tyrimas orientuotas į metodologinį pagrindimą, kokybinį rodiklių vertinimą ir geriausio modelio identifikavimą skirtinguose trūkstamumo scenarijuose.

Tyrimo eiga pateikiama 10 priede. Schema vizualiai reprezentuoja visą procesą nuo duomenų paruošimo iki galutinio trūkstamų reikšmių užpildymo ir tolesnės ekonominių rodiklių analizės. Kiekvienas šios schemos blokas atspindi specifinę tyrimo fazę, užtikrinančią sisteminį ir metodologiškai pagrįstą požiūrį į trūkstamų reikšmių problemą.

Pirmasis etapas apima NUTS 2 ekonominių rodiklių duomenų rinkinio paruošimą, kuris išsamiai aprašytas antrajame darbo skyriuje. Duomenų rinkinys apima 78 ekonominius rodiklius iš 244 NUTS 2 regionų per 25 metų laikotarpį (2000 - 2024 m.), suformuojant 6100 stebėjimų masyvą su heterogeniška trūkstamumo struktūra.

Antrasis etapas - bazinių Random Forest ir XGBoost modelių kūrimas bei testavimas su trimis skirtingais atsitiktinio trūkstamumo scenarijais: 10 proc., 20 proc. ir 30 proc. Šie scenarijai sukuriami dirbtinai pašalinant atsitiktinai pasirinktą žinomų reikšmių dalį iš duomenų rinkinio, kas leidžia objektyviai įvertinti modelių prognozavimo tikslumą lyginant su faktinėmis reikšmėmis. Tokia metodologija atitinka plačiai literatūroje taikomą sintetinių trūkstamų duomenų generavimo principą, užtikrinantį kontroliuojamą eksperimentinę aplinką [101], [102].

Trečiasis etapas apima bazinių modelių tikslumo metrikų vertinimą naudojant keturias pagrindines metrikas: determinacijos koeficientą (R^2), normalizuotą vidutinę kvadratinę paklaidą (nRMSE), normalizuotą vidutinę absoliučią paklaidą (nMAE) ir simetrinę vidutinę absoliučią procentinę paklaidą (sMAPE). Šios metrikos parinktos dėl jų gebėjimo kompleksiškai įvertinti prognozavimo kokybę tiek absoliutaus tikslumo, tiek santykinės paklaidos aspektais [103].

Ketvirtasis etapas - požymių svarbos analizė, kuri atskleidžia, kurie ekonominiai rodikliai turi didžiausią įtaką prognozuojant trūkstamas reikšmes. Random Forest ir XGBoost algoritmai automatiškai apskaičiuoja kiekvieno požymio svarbą (*angl. feature importance*), remdamiesi to požymio įnašu į modelio sprendimus [4], [5]. Ši analizė suteikia svarbių įžvalgų apie ekonominių rodiklių tarpusavio priklausomybes ir leidžia identifikuoti fundamentalius veiksnius, formuojančius regioninę ekonominę struktūrą.

Penktasis etapas - kernelio tankio įvertinimas (KDE) originalių ir užpildytų reikšmių pasiskirstymams palyginti. KDE analizė leidžia vizualiai ir kiekybiškai įvertinti, kaip gerai užpildytos reikšmės atitinka originalių duomenų statistinį pasiskirstymą [104]. Šis vertinimas yra svarbus siekiant užtikrinti, kad trūkstamų reikšmių užpildymas ne tik kompensuotų duomenų trūkumą, bet ir išsaugotų pagrindines duomenų rinkinio statistines savybes.

Šeštasis etapas - sprendimo priėmimas dėl modelio tobulinimo būtinybės. Remiantis bazinių modelių rezultatais, įvertinama, ar reikalingas hiperparametrų optimizavimas. Jei modelio veikimas neatitinka tikslumo reikalavimų arba pastebimi ženklūs skirtumai tarp skirtingų trūkstamumo scenarijų, atliekamas hiperparametrų derinimas naudojant kryžminį validavimą. Optimizuoti modeliai pakartotinai testuojami ir jų rezultatai lyginami su baziniais modeliais.

Septintasis etapas - galutinis trūkstamų reikšmių užpildymas ir duomenų rinkinio formavimas. Pasirinktas geriausiai veikianti modelis (pagal vertinimo metrikų rezultatus) taikomas visoms tikrai trūkstamoms reikšmėms originalame duomenų rinkinyje. Modelis generuoja prognozes kiekvienam trūkstamam stebėjimui, po to šios reikšmės integruojamos į pradinę duomenų struktūrą, pakeičiant NaN (angl. *Not a Number*) žymes konkrečiomis skaitinėmis reikšmėmis. Rezultatas - pilnas NUTS 2 ekonominių rodiklių duomenų rinkinys be trūkstamų reikšmių.

Devintasis, finalinis etapas - tolimesnė ekonominių rodiklių analizė, kuri nebėra šio magistro darbo tiesioginis objektas, bet sudaro pagrindą būsimiems tyrimams. Pilnas duomenų rinkinys gali būti naudojamas regionų klasterizacijai, ekonominio atsparumo modeliavimui, laiko eilučių prognozavimui ir kitoms analitinėms užduotims.

Visa ši procedūra buvo įgyvendinta naudojant Python 3.13.9 programavimo kalbą, Jupyter Notebook aplinką ir Visual Studio Code redaktorių. Skaičiavimai buvo atlikti darbo stotyje su AMD Ryzen 5 PRO 4650G procesoriumi, bazinis greitis (3,70 GHz) ir 32 GB operatyviosios atminties, užtikrinančiais pakankamą našumą didesnių modelių treniravimui. Programinė įranga veikė Windows 11 Pro operacinėje sistemoje. Tyrime buvo naudojamos šios pagrindinės Python bibliotekos: *pandas* (duomenų apdorojimui), *NumPy* (skaitinėms operacijoms), *scikit-learn* (mašininio mokymosi algoritmams) [105], *xgboost* (XGBoost algoritmo realizacijai) [41], *matplotlib* ir *seaborn* (duomenų vizualizacijai).

3.2 Duomenų reikalingų eksperimentui paruošimas

Eksperimentinių duomenų paruošimas yra svarbus etapas, užtikrinantis tyrimo rezultatų patikimumą ir palyginamumą. Kaip išsamiai aprašyta antrajame darbo skyriuje, pradinis NUTS 2 regionų ekonominių rodiklių duomenų rinkinys buvo suformuotas iš Eurostat duomenų bazės, agregavimo ir standartizavimo procedūrų metu. Šiame poskyryje dėmesys sutelkiamas į specifinius eksperimentui reikalingus duomenų transformavimo žingsnius.

Duomenų valymo principai grindžiami keliais fundamentaliais kriterijais. Pirma, iš duomenų rinkinio buvo pašalinti visi netinkamą regioninį detalumą turintys rodikliai (EU serijos indikatoriai), kadangi jie matuoja ne absoliučias regionines charakteristikas, o santykinę padėtį ES kontekste. Antra, pašalinti rodikliai su ekstremaliai dideliu trūkstamumo lygiu (>92 proc.), nes tokiais atvejais trūkstamų reikšmių užpildymo patikimumas itin mažas. Trečia, išlaikyti tik tie rodikliai, kurie turi aiškų ekonominį interpretavimą ir yra sistemingai renkami per visą analizuojamą laikotarpį. Po šių procedūrų galutinis duomenų rinkinys apėmė 78 ekonominius rodiklius su trūkstamumo lygiu nuo 0 iki 91,2 proc.

Struktūrinių nulių taisyklės yra esminės užtikrinant, kad modelio treniravimas būtų korektiškas ir neiškraipytų duomenų pasiskirstymo. Ekonominiuose duomenyse struktūriniai nuliai - tai reikšmės, kurios yra lygios nuliui ne dėl trūkstamumo, bet dėl ekonominės prasmės (pavyzdžiui, kai tam tikras sektorius neegzistuoja konkrečiame regione). Tokios reikšmės gali būti naudojamos modelio treniravimui (train set), nes jos suteikia vertingos informacijos apie ekonominę struktūrą. Tačiau struktūriniai nuliai negali būti naudojami kaip testavimo duomenys (*test set*), nes jų prognozavimas nereprezentuoja modelio gebėjimo spręsti tikrąją trūkstamų reikšmių problemą [106].

Taikomas *holdout based synthetic test* metodas [107], [108]: kiekvienam kintamajam, kurio trūkstamos reikšmės užpildomos, atsitiktinai atrenkama 10 proc., 20 proc. arba 30 proc. žinomų stebėjimų, kurie neįtraukiami į modelio treniravimą. Šios reikšmės nėra fiziškai pašalinamos iš duomenų rinkinio ir išlieka kaip tikrosios reikšmės testavimo metu. Modeliui pateikiami tik atitinkamų eilučių prediktoriai, o prognozuotos reikšmės lyginamos su originaliomis, siekiant apskaičiuoti tikslumo metrikas (RMSE, MAE, R^2 ir sMAPE). Ši metodologija plačiai pripažinta trūkstamų duomenų užpildymo tyrimuose kaip standartinis būdas objektyviai įvertinti užpildymo algoritmų efektyvumą kontroliuojamoje aplinkoje [109], [110].

Pasirinkti trys trūkumo scenarijai - 10 proc., 20 proc. ir 30 proc. - reprezentuoja skirtingo intensyvumo duomenų užpildymo iššūkius. 10 proc. scenarijus atitinka gana lengvą trūkstamumo atvejį, kuris dažnai pasitaiko gerai prižiūrimuose duomenų rinkiniuose. 20 proc. scenarijus reprezentuoja vidutinio sudėtingumo situaciją, tipišką regioniniams ekonominiams duomenims. 30 proc. scenarijus simbolizuoja aukštą trūkstamumo lygį, artimą probleminiams duomenų rinkiniams, kuriuose užpildymo kokybė gali reikšmingai kristi [23].

Išsamiai ekonominio rodiklio Y_01 analizei pasirinktas 20 proc. trūkumo scenarijus kaip vidutinio sudėtingumo atvejis, leidžiantis aiškiausiai parodyti modelių skirtumus. 10 proc. scenarijus yra per lengvas - modeliai pasiekia labai aukštą tikslumą, todėl skirtumai tarp jų tampa statistiškai nereikšmingi. 30 proc. scenarijus yra per ekstremalus - abu modeliai susiduria su duomenų nepakankamumu, todėl jų prognozės tampa mažiau stabilios ir labiau variacijos paveiktos. 20 proc.

scenarijus suteikia optimalų balansą tarp duomenų pakankamumo treniravimui ir reprezentatyvaus testavimo rinkinio dydžio.

Tyrime naudoti baziniai modelių hiperparametrai buvo parinkti remiantis standartinėmis scikit-learn ir xgboost bibliotekų implementacijomis. Random Forest modeliui taikyti šie hiperparametrai: n_estimators=100 (sprendimų medžių skaičius ansamblyje), max_depth=None (maksimalus medžio gylis neribojamas), min_samples_split=2 (minimalus stebėjimų skaičius mazgo padalijimui), min_samples_leaf=1 (minimalus stebėjimų skaičius lapiniame mazge), max_features='sqrt' (atsitiktinai atrenkamų požymių skaičius kiekviename mazge lygus požymių skaičiaus kvadratinei šakniai), bootstrap=True (naudojamas bootstrap atrankos metodas), random_state=42 (atsitiktinių skaičių generatoriaus fiksavimas rezultatų atkuriamumui) [105]. XGBoost modeliui taikyti šie hiperparametrai: n_estimators=100 (sprendimų medžių skaičius), max_depth=6 (maksimalus medžio gylis), learning_rate=0.3 (mokymosi greitis), subsample=1.0 (duomenų imties dalis kiekvienam medžiui), colsample_bytree=1.0 (požymių dalis kiekvienam medžiui), reg_alpha=0 (L1 reguliarizacijos parametras), reg_lambda=1 (L2 reguliarizacijos parametras), random_state=42 (rezultatų atkuriamumas) [36]. Šie baziniai parametrai buvo pasirinkti kaip atskaitos taškas modelių palyginimui, prieš atliekant galimą hiperparametrų optimizavimą.

3.3 Modelių veikimo palyginimas pagal vertinimo metrikas skirtinguose trūkstamumo scenarijuose

Modelių vertinimo metrikos pagrindimas prasideda nuo normalizacijos būtinybės. RMSE ir MAE normalizacija buvo atlikta dėl skirtingų ekonominių rodiklių mastelių heterogeniškumo. Rodikliai NUTS 2 duomenų rinkinyje svyruoja nuo procentinių dalių (0 - 100) iki didelių absoliučių skaičių (pavyzdžiui BVP milijonais eurų). Normalizuojama pagal kiekvieno rodiklio maksimalios ir minimalios reikšmių skirtumą (max - min), kas leidžia palyginti skirtingų mastelių rodiklių užpildymo tikslumą vienodomis sąlygomis [111], [112]. RMSE ir MAE normalizuotos metrikos apskaičiuojamos pagal (19) ir (20) formules:

$$\text{nRMSE} = \frac{\text{RMSE}}{y_{\max} - y_{\min}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}}, \tag{19}$$

$$\text{nMAE} = \frac{\text{MAE}}{y_{\max} - y_{\min}} = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{y_{\max} - y_{\min}}, \tag{20}$$

kur y_i - faktinė reikšmė, $\hat{y}_i$ - prognozuota reikšmė, n - stebėjimų skaičius, $y_{\max}$ - maksimali rodiklio reikšmė duomenų rinkinyje, $y_{\min}$ - minimali rodiklio reikšmė duomenų rinkinyje.

sMAPE (*angl. Symmetric Mean Absolute Percentage Error*) metrikos pasirinkimas grindžiamas keliais svarbiais argumentais. Pirma, sMAPE yra neproblematiška su nulinėmis reikšmėmis, nes vardiklis niekada netampa lygus nuliui - formulėje naudojamas faktinių ir prognozuotų reikšmių vidurkis. Antra, metrika yra simetriška - vienodo dydžio teigiamos ir neigiamos paklaidos vertinamos vienodai, kas skiriasi nuo tradicinės MAPE. Trečia, sMAPE mažiau pažeidžiama išskirčių, nes santykinė paklaida ribojama teorine 200 proc. riba. Ketvirta, metrika ypač tinka procentiniams ir indeksiniams rodikliams, kurių gausu NUTS 2 duomenų rinkinyje (ekonominės struktūros dalys, užimtumo lygiai) [13]. sMAPE apskaičiuojama pagal (21) formulę:

$$\text{sMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\left(\frac{|y_i| + |\hat{y}_i|}{2}\right)}, \tag{21}$$

kur y_i - faktinė reikšmė,$\hat{y}_i$ - prognozuota reikšmė, n - stebėjimų skaičius.

Determinacijos koeficientas R^2 matuoja modelio paaiškintą dispersijos dalį ir apskaičiuojamas pagal (22) formulę:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{22}$$

kur $\bar{y}$ - faktinių reikšmių vidurkis [113].

6 lentelė. Modelių vertinimo metrikų rezultatai skirtinguose trūkstamumo scenarijuose

Metrika	**10 % trūkumas**		**20 % trūkumas**		**30 % trūkumas**	
	Random Forest	XGBoost	Random Forest	XGBoost	Random Forest	XGBoost
R^2	0,9303	0,9462	0,9171	0,9330	0,9125	0,9411
nRMSE	0,0381	0,0286	0,0367	0,0287	0,0365	0,0275
nMAE	0,0232	0,0166	0,0218	0,0157	0,0218	0,0154
sMAPE	15,12	11,77	15,76	12,33	16,33	12,58

Šaltinis: sudaryta darbo autoriaus.

6 lentelėje pateikta metrikų analizė atskleidžia aiškius modelių veikimo skirtumus ir tendencijas. R^2 rodiklis, matuojantis modelio paaiškintos dispersijos dalį, visuose trūkstamumo scenarijuose išlieka aukštas abiem modeliams ($R^2 > 0,91$). Visuose scenarijuose XGBoost nuosekliai lenkia Random Forest: 10 % trūkumo scenarijuje XGBoost R^2 yra 0,016 didesnis (0,9462 - 0,9303), 20 % scenarijuje skirtumas siekia 0,016, o 30 % scenarijuje - 0,029. Šie rezultatai rodo, kad didėjant duomenų trūkumui XGBoost išlaiko stabilesnę prognozavimo kokybę ir geriau prisitaiko prie informacijos praradimo nei Random Forest.

Normalizuotos RMSE reikšmės patvirtina XGBoost pranašumą visuose trūkstamumo scenarijuose. Random Forest nRMSE svyruoja 0,0365 - 0,0381 intervale, tuo tarpu XGBoost pasiekia ženkliai mažesnes reikšmes - 0,0275 - 0,0287. Tai reiškia, kad XGBoost vidutinė normalizuota kvadratinė paklaida yra maždaug 22 - 25 % mažesnė nei Random Forest. Pastebėtina, kad XGBoost

modelio nRMSE reikšmės skirtinguose trūkstamumo lygiuose išlieka labai stabilios (0,0275 - 0,0287), o bendrame 10 - 30 % intervale stebimi tik nežymūs svyravimai, kas rodo modelio atsparumą didėjančiam duomenų trūkumui.

sMAPE metrika atskleidžia ekonomiškai interpretuojamus prognozavimo tikslumo skirtumus. Random Forest sMAPE didėja nuo 15,12 % (10 % trūkumo scenarijuje) iki 16,33 % (30 % scenarijuje), rodydama nuoseklų tikslumo blogėjimą augant trūkstamų duomenų kiekiui. Tuo tarpu XGBoost sMAPE reikšmės svyruoja siauresniame 11,77 - 12,58 % intervale, pasižymėdamos ženkliai stabilesniu veikimu. Šie skirtumai turi praktinę reikšmę: XGBoost prognozės vidutiniškai nukrypsta nuo faktinių reikšmių apie 12 %, o Random Forest - apie 15 - 16 %, kas realiame ekonominės analizės kontekste gali lemti tikslesnį ir patikimesnį sprendimų priėmimą.

Tarpusavio metrikų palyginimas atskleidžia dar vieną svarbų aspektą - modelių stabilumą skirtinguose trūkstamumo scenarijuose. Random Forest metrikų variacija yra didesnė: R^2 reikšmės kinta nuo 0,9303 iki 0,9125, t. y. svyravimas siekia 0,0178. Tuo tarpu XGBoost modelio R^2 kinta siauresnėse ribose - nuo 0,9462 iki 0,9411 (svyravimas 0,0051), kas rodo žymiai stabilesnį veikimą. Šie rezultatai leidžia teigti, kad XGBoost algoritmas yra mažiau jautrus duomenų trūkumo pokyčiams ir geriau generalizuoja skirtingo sudėtingumo sąlygomis.

Apibendrinus 6 lentelėje pateiktų modelių vertinimo metrikų analizę, gauti rezultatai leidžia palyginti taikytų ansamblinių mašininio mokymosi modelių veikimą su literatūroje aprašytais daugiamačio trūkstamų reikšmių užpildymo metodo *MICE* rezultatais. *MICE* metodo vertinimas, pateiktas sisteminėje literatūros analizėje, 4 lentelėje (žr. 1.3.4 poskyrį), rodo, kad didėjant trūkstamų duomenų daliai nuo 20 % iki 80 %, MAE, MSE ir RMSE metrikų reikšmės nuosekliai ir reikšmingai didėja, atskleidžiant spartų užpildymo tikslumo mažėjimą esant dideliam duomenų trūkumui. Tai leidžia identifikuoti praktines iteratyvių statistinių trūkstamų reikšmių užpildymo metodų taikymo ribas, ypač esant aukštam informacijos praradimo lygiui.

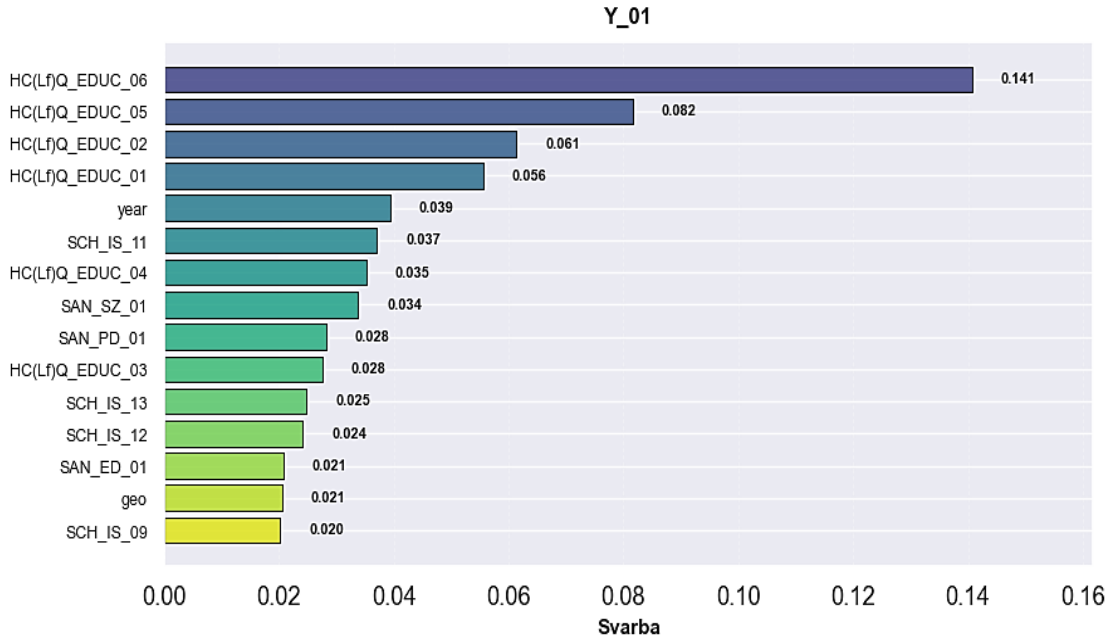
Lyginant šias tendencijas su šiame tyrime gautais Random Forest ir XGBoost modelių rezultatais, matyti, kad ansambliniai mašininio mokymosi modeliai pasižymi didesniu atsparumu didėjančiam trūkstamumui. XGBoost modelis visais nagrinėtais trūkstamumo scenarijais išlaiko stabilesnes nRMSE, nMAE ir sMAPE reikšmes bei aukštesnį determinacijos koeficientą, palyginti ne tik su Random Forest, bet ir su literatūroje pateikiamais MICE metodo rezultatais. Tai rodo, kad nelinijinius ryšius efektyviai modeliuojantys ansambliniai metodai geba geriau kompensuoti informacijos praradimą nei klasikiniai iteratyvūs trūkstamų reikšmių užpildymo algoritmai.

Tokiu būdu gauti empiriniai rezultatai ne tik papildo 4 lentelėje pateiktas sisteminės literatūros analizės išvadas apie MICE metodo jautrumą dideliam trūkstamumo lygiui, bet ir patvirtina XGBoost modelio pranašumą kaip stabilesnį ir universalesnį sprendimą trūkstamų reikšmių užpildymui NUTS 2 lygmens regionų ekonominių rodiklių duomenų rinkinyje.

3.4 Ekonominio rodiklio Y_01 išsami analizė naudojant 20 % trūkumo scenarijų

3.4.1 Požymių svarbos analizė

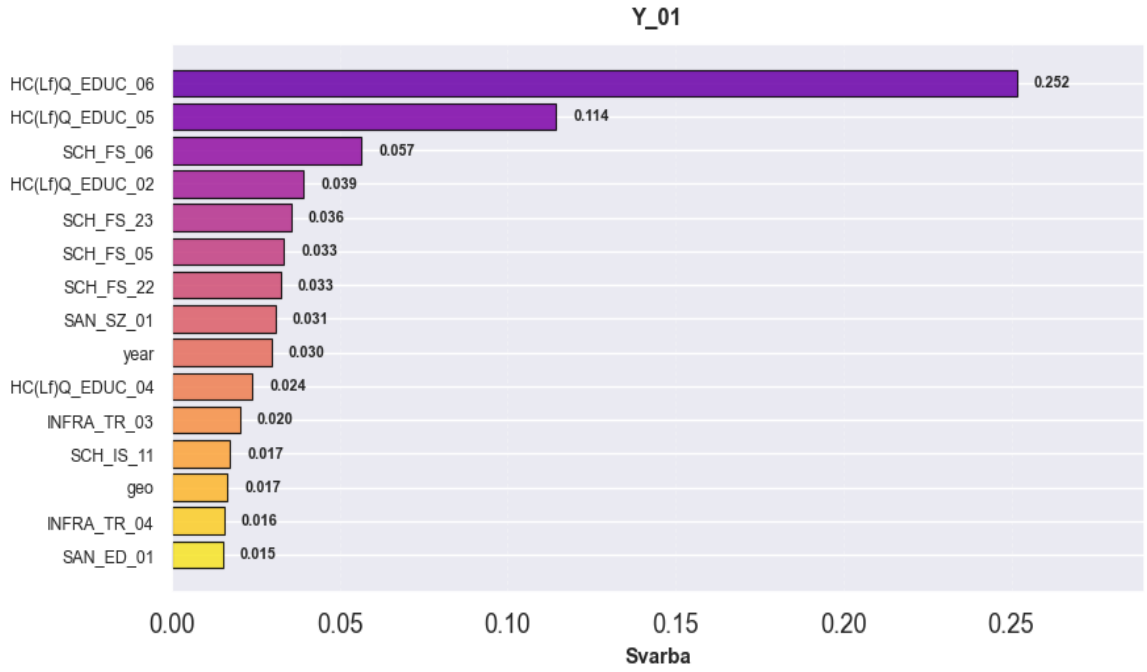
Požymių svarbos analizė leidžia identifikuoti pagrindinius ekonominius rodiklius, darančius didžiausią įtaką Y_01 rodiklio reikšmių užpildymui esant 20 % duomenų trūkumo scenarijui. Analizė suteikia įžvalgų apie modelio sprendimų logiką ir leidžia įvertinti, ar nustatyti svarbiausi požymiai atitinka ekonominę teoriją bei ankstesnių tyrimų rezultatus.



Šaltinis: sudaryta darbo autoriaus.

14 pav. Random Forest požymių svarba Y_01 rodikliui

14 paveikslėlyje pastebima, kad svarbiausias prediktorius yra HC(Lf)Q_EDUC_06 (svarba 0,141), matuojantis žmogiškojo kapitalo potencialą per asmenų, turinčių aukštąjį išsilavinimą ir (arba) dirbančių mokslo bei technologijų srityse, dalį. Antroje pozicijoje yra HC(Lf)Q_EDUC_05 (svarba 0,082), nusakantis aukštąjį išsilavinimą turinčių gyventojų dalį. Šie rezultatai ekonomiškai intuityvūs - technologijų sektorių plėtra natūraliai koreliuoja su kvalifikuotos darbo jėgos prieinamumu regione [114].



Šaltinis: sudaryta darbo autoriaus.

15 pav. XGBoost požymių svarba Y_01 rodikliui

XGBoost požymių svarba Y_01 rodikliui 15 paveikslėlyje demonstruoja žymiai aiškesnę hierarchiją. HC(Lf)Q_EDUC_06 dominuoja su svarba 0,252, o HC(Lf)Q_EDUC_05 užima antrą poziciją su svarba 0,114. Trečioje pozicijoje yra SCH_FS_06 (svarba 0,057), reprezentuojantis samdomųjų darbuotojų skaičių veikiančiose įmonėse, kas indikuoja regiono ekonominio aktyvumo lygį.

Lyginant abiejų modelių požymių svarbos pasiskirstymą, matyti, kad XGBoost pasižymi ryškesne požymių svarbos diferenciacija. Svarbiausias požymis - HC(Lf)Q_EDUC_06 - XGBoost modelyje turi 0,252 svarbą ir yra maždaug 1,79 karto reikšmingesnis nei Random Forest modelyje (0,141). Tai rodo, kad gradientinio stiprinimo mechanizmas efektyviau išryškina dominuojančius veiksnius ir mažiau remiasi mažos svarbos požymiais, taip sutelkiant modelio dėmesį į labiausiai prognozę lemiančius kintamuosius [36].

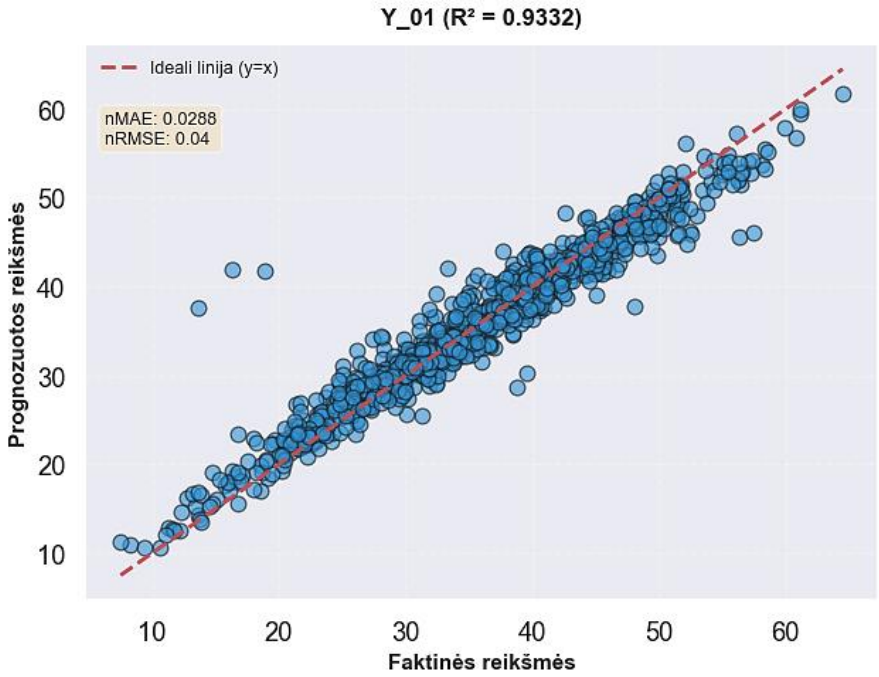
Didžiausią prognozinį reikšmingumą turintys kintamieji HC(Lf)Q_EDUC_06 ir HC(Lf)Q_EDUC_05 rodo, kad žmogiškojo kapitalo lygis regione yra esminis veiksnys aiškinant Y_01 rodiklio kitimą. Ekonominė šio rezultato interpretacija yra nuosekli: regionai, kuriuose darbuotojai pasižymi aukštesniu išsilavinimu bei kvalifikacija, tampa patrauklesni technologijų sektoriaus įmonėms, o tai skatina tiek darbo jėgos pritraukimą, tiek jos išlaikymą. Tokiu būdu susiformuoja savilaikė augimo dinamika: aukštas išsilavinimo lygis lemia spartesnę technologijų sektoriaus plėtrą, kuri generuoja didesnes pajamas ir darbo užmokestį, didina regiono patrauklumą kvalifikuotiems specialistams ir dar labiau stiprina sektoriaus augimą [115]. Ši kaupiamojo efekto

logika paaiškina, kodėl švietimo rodikliai pasižymi itin didele prognozine galia modeliuojant Y_01 rodiklį.

Verta paminėti, kad geografinis kintamasis (geo) ir laikas (year) turi nedidelę svarbą abiejuose modeliuose (apie 0,017 - 0,030). Tai rodo, kad svarbiausi modelių identifikuojami požymiai atspindi esminius regionų ekonominės struktūros aspektus ir lemia trūkstamų reikšmių prognozavimo rezultatus. Gauti požymių svarbos analizės rezultatai atitinka Hasan ir kt. [20] pateiktas išvadas, kuriose pabrėžiama, kad taikant mašininio mokymosi metodus trūkstamų reikšmių užpildymo kontekste modeliai linkę išryškinti struktūrinius ir makro lygio rodiklius, turinčius didžiausią įtaką duomenų generavimo procesui, o šių požymių dominavimas prisideda prie stabilesnių ir nuoseklesnių trūkstamų reikšmių užpildymo rezultatų, ypač analizuojant regioninio lygmens socialinius ir ekonominius duomenis.

3.4.2 Faktinių ir prognozuotų reikšmių palyginimas

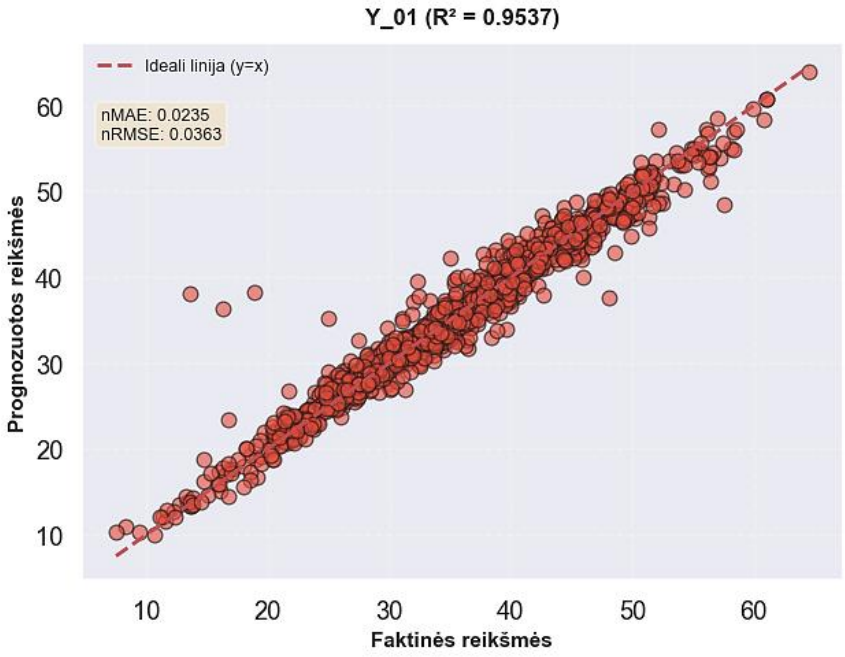
Random Forest faktinių ir prognozuotų reikšmių palyginimas Y_01 rodikliui parodo stiprų tiesinį ryšį tarp faktinių ir prognozuotų reikšmių (R^2 = 0,9332). 16 paveiksle matyti, kad dauguma stebėjimų koncentruojasi apie idealią (y = x) ašies liniją, tačiau pastebimi sisteminiai nukrypimai: žemų reikšmių intervale (20 - 35) prognozės dažniau yra didesnės nei faktinės, o aukštų reikšmių intervale (55 - 65) - mažesnės. Tai rodo, kad Random Forest modelis pasižymi lengva susitraukimo į vidurkį (angl. shrinkage) tendencija, kai prognozės švelniai traukiamos link centrinės reikšmės. Šie nukrypimai yra santykinai maži (nMAE = 0,0288, nRMSE = 0,04).



Šaltinis: sudaryta darbo autoriaus.

16 pav. Random Forest faktinių ir prognozuotų reikšmių palyginimas Y_01 rodikliui

XGBoost faktinių ir prognozuotų reikšmių palyginimas 17 paveikslėlio grafike parodo dar stipresnį ryšį ($R^2 = 0,9537$) ir griežtesnį taškų prigludimą prie idealios linijos. Prognozuotų reikšmių sklaida aplink (y = x) liniją yra mažesnė nei Random Forest atveju (nMAE = 0,0235, nRMSE = 0,0363). XGBoost taip pat demonstruoja žymiai tikslesnes prognozes kraštinėse reikšmių srityse, kuriose Random Forest linkęs daryti didžiausias sistemines paklaidas.



Šaltinis: sudaryta darbo autoriaus.

17 pav. XGBoost faktinių ir prognozuotų reikšmių palyginimas Y_01 rodikliui

Ryšio glaudumas abiejuose modeliuose yra aukštas, tačiau XGBoost nuosekliai pasižymi geresniu veikimu. Nors R^2 skirtumas (0,9537 - 0,9332) nėra didelis, XGBoost sumažina vidutinę absoliučią paklaidą apie 18 %, o normalizuotą vidutinę kvadratinę paklaidą - apie 9 % lyginant su Random Forest. Tai rodo praktiškai reikšmingą prognozavimo tikslumo pagerėjimą.

Anomalijų analizė rodo, kad abu modeliai retkarčiais sukuria prognozes, pastebimai nutolstančias nuo faktinių reikšmių. Šie išsibarstę taškai greičiausiai susiję su specifinėmis regionų struktūrinėmis savybėmis, kurios nėra pilnai atspindėtos turimuose aiškinamuosiuose rodikliuose, arba su staigiais ekonominiais pokyčiais.

Sisteminės paklaidos analizė patvirtina, kad Random Forest pasižymi lengva susitraukimo į vidurkį (*angl. shrinkage*) tendencija: mažesnės faktinės reikšmės dažniau pervertinamos, o didesnės - nuvertinamos. Tuo tarpu XGBoost, pasitelkdamas gradientinio stiprinimo mechanizmą ir griežtesnę reguliarizaciją, geriau prisitaiko prie nelinearių duomenų struktūrų ir veiksmingiau išvengia šio *shrinkage* efekto, kas atitinka literatūroje aprašytą ansamblinių metodų elgseną, pabrėžiančią, kad gradientinio stiprinimo modeliai dėl iteratyvaus klaidų koregavimo pasižymi mažesniu sisteminiu

nuokrypiu ir stabilesnėmis prognozėmis, lyginant su nepriklausomų sprendimų medžių agregavimu pagrįstais metodais [36], [116], [117].

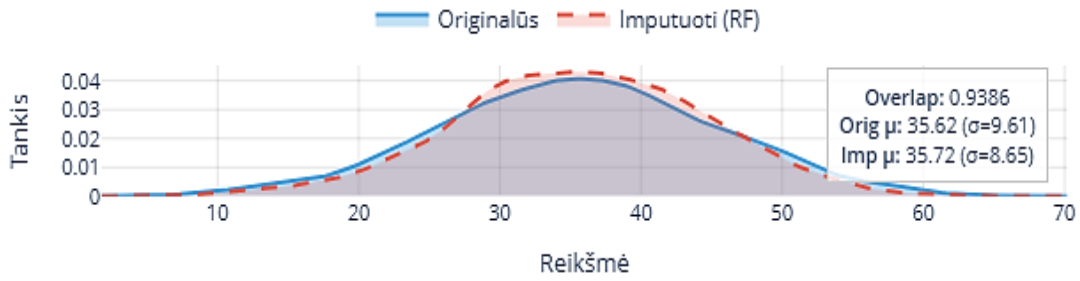
Dėl iteratyvaus mokymosi iš paklaidų XGBoost prognozės griežčiau priglunda prie (y = x) linijos: kiekvienas naujas medis koreguoja ankstesnių medžių klaidas, taip nuosekliai mažindamas likutines paklaidas. Random Forest, priešingai, konstruoja nepriklausomus medžius, kurių prognozės vėliau agreguojamos (apskaičiuojamas jų vidurkis), todėl modelis negali adaptyviai mokytis iš savo klaidų [118].

3.4.3 KDE pagrįsta pasiskirstymo analizė

Branduolio (*angl. kernel*) tankio įvertinimas, dar vadinamas kernelio tankio įvertinimu (*angl. Kernel Density Estimation, KDE*), yra neparametrinis tikimybės tankio funkcijos aproksimavimo metodas, leidžiantis iš esamo duomenų mėginio atkurti jo pasiskirstymo formą. KDE pasižymi tuo, kad vietoje histogramų naudojamų stačiakampių intervalų sukuria glotnią tankio kreivę, kuri geriau atspindi realią duomenų struktūrą - ypač tada, kai svarbu įvertinti pasiskirstymo simetriškumą, uodegų ilgį ar dispersijos pokyčius.

Šiame magistriniame darbe KDE skaičiavimui naudojama „*SciPy*“ biblioteka, o tiksliau funkcija *scipy.stats.gaussian_kde*, skirta neparametriniam tikimybinio tankio įvertinimui. KDE vizualizacijai taikoma „*Plotly*“ biblioteka, kuri suteikia galimybę analitiškai palyginti originalių ir užpildytų reikšmių KDE kreives bei apskaičiuoti jų persidengimą (*angl. overlap*). *Overlap* rodiklis apskaičiuojamas integruojant mažiausią originalios ir užpildytos KDE funkcijų reikšmę per visą analizės intervalą, naudojant „*NumPy*“ trapecijos taisyklės integravimo funkcijas (*np.trapezoid* arba *np.trapz*). Tai leidžia tiksliai nustatyti dviejų pasiskirstymų ploto sutapimo laipsnį.

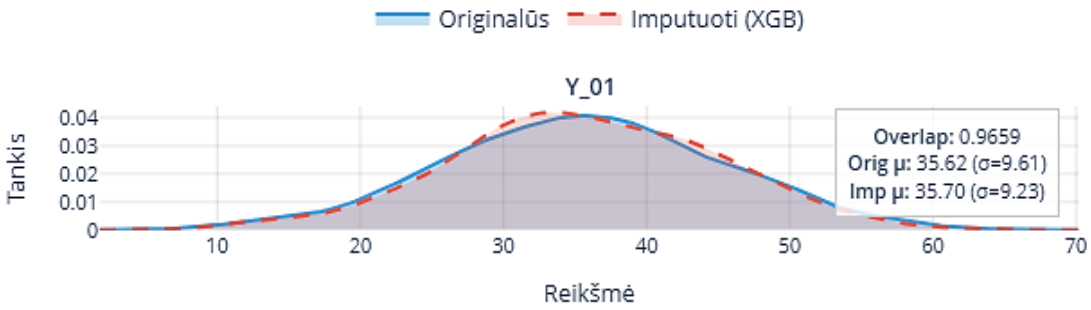
Random Forest modelio, užpildytų ir originalių Y_01 rodikliui, reikšmių KDE palyginimas, pateiktas 17 paveikslėlyje, leidžia įvertinti, kiek užpildytas pasiskirstymas atkartoja pradinę duomenų struktūrą. Šiame kontekste naudojamas pasiskirstymų persidengimo rodiklis (*angl. overlap*), kuris parodo, kokia dalis dviejų pasiskirstymų plotų sutampa. Vertė 1 reiškia tobulą sutapimą, o 0 - visišką nesutapimą. Kaip matyti iš grafiko (žr. 18 pav.), Random Forest atveju gautas pasiskirstymų persidengimo rodiklis (overlap) siekia 0,9386, t. y. apie 93,86 % užpildytų reikšmių pasiskirstymo ploto sutampa su originalių duomenų pasiskirstymu. Taip pat pateikiami vidurkiai ir standartiniai nuokrypiai; σ (standartinis nuokrypis) nusako, kaip plačiai reikšmės išsisklaido aplink vidurkį. Originalių reikšmių vidurkis yra 35,62 (σ = 9,61), o užpildytų - 35,72 (σ = 8,65). Matyti, kad nors vidurkiai beveik identiški, užpildytos reikšmės šiek tiek mažiau išsisklaidžiusios aplink vidurkį. Tai atitinka Random Forest modeliui būdingą savybę mažinti dispersiją dėl prognozių agregavimo [119].



Šaltinis: sudaryta darbo autoriaus.

18 pav. Random Forest užpildytų ir originalių Y_01 reikšmių pasiskirstymo (KDE) palyginimas

XGBoost modelio užpildytų ir originalių Y_01 rodiklio reikšmių pasiskirstymo palyginimas, pateiktas 19 paveikslėlyje, rodo dar tikslesnį užpildymo rezultatą. Šiame grafike matyti, kad persidengimo rodiklis (*angl. overlap*) siekia 0,9659, tai yra XGBoost beveik tobulai atkuria pradinį duomenų tankio pasiskirstymą. Originalių reikšmių vidurkis yra 35,62 (σ = 9,61), o užpildytų - 35,70 (σ = 9,23), todėl galima teigti, kad modelis tiksliai atkuria ne tik centrinę tendenciją, bet ir pradinę dispersijos struktūrą. 18 paveikslėlyje pateiktos KDE kreivės praktiškai sutampa visame reikšmių intervale, o tai rodo, kad XGBoost geba itin tiksliai atkurti statistinę duomenų struktūrą.



Šaltinis: sudaryta darbo autoriaus.

19 pav. XGBoost užpildytų ir originalių Y_01 reikšmių pasiskirstymo (KDE) palyginimas

Overlap rodiklių skirtumas - 0,9659, palyginti su 0,9386 Random Forest atveju - reiškia 2,73 procentinio punkto pagerėjimą XGBoost naudai. Nors skirtumas nėra didelis absoliučia reikšme, jis parodo, kad XGBoost sistemingai geriau išlaiko pradinę pasiskirstymo struktūrą. Random Forest turi tendenciją šiek tiek sumažinti dispersiją dėl agregavimo mechanizmo, todėl užpildytų reikšmių KDE kreivė tampa šiek tiek siauresnė.

XGBoost pranašumas pasireiškia tuo, kad šis modelis, optimizuodamas tikslinę funkciją naudodamas antrojo laipsnio Teiloro aproksimaciją bei reguliarizaciją, geba tiksliau modeliuoti ne tik vidurkių, bet ir dispersijos struktūrą [36]. Todėl užpildytos reikšmės išlaiko platesnį, artimesnį originaliam duomenų pasiskirstymą, o tai atsispindi beveik idealiai sutampančiose KDE kreivėse, pateiktose 18 paveikslėlyje. Dėl šių savybių XGBoost gali būti laikomas pažangesniu ir statistiškai

tinkamesniu metodu trūkstamų reikšmių užpildymui situacijose, kur svarbu išsaugoti pilną duomenų struktūrą.

Gauti KDE pagrįstos pasiskirstymo analizės rezultatai atitinka literatūroje taikomą praktiką, kurioje pabrėžiama, kad trūkstamų reikšmių užpildymo metodų vertinimui svarbu analizuoti ne tik agreguotas paklaidų metrikas, bet ir užpildytų bei originalių duomenų pasiskirstymų atitikimą. Shrivastava ir Shukla siūlomoje „*Impute-VSS*“ sistemoje KDE persidengimas naudojamas kaip vienas pagrindinių kriterijų įvertinti, ar užpildymo metodas išsaugo pradinę duomenų struktūrą [23]. Analogiškai šiame tyrime nustatyta, kad XGBoost modelio užpildytų reikšmių pasiskirstymas KDE analizėje yra artimesnis faktinių reikšmių tankio formai nei Random Forest atveju, ypač esant didesniam trūkstamumo lygiui.

3.5 Hiperparametrų optimizavimas

Hiperparametrų optimizavimas yra svarbus etapas mašininio mokymosi modelių efektyvumo gerinimui, ypač kai siekiama maksimalaus prognozavimo tikslumo sudėtinguose trūkstamų reikšmių užpildymo užduotyse. Kaip parodyta ankstesniuose poskyriuose (žiūrėti 3.3 ir 3.4), baziniai Random Forest ir XGBoost modeliai su standartiniais parametrais jau demonstruoja aukštą prognozavimo tikslumą NUTS 2 regionų ekonominių rodiklių duomenyse. Vis dėlto bazinių modelių rezultatuose pastebėtas kai kurių metrikų svyravimas (standartinis nuokrypis) rodo, kad modelio veikimas nėra visiškai stabilus. Tai leidžia daryti išvadą, kad siekiant sumažinti rezultatų dispersiją ir užtikrinti modelio patikimumą, būtina atlikti sistemingą hiperparametrų parinkimą. Toks optimizavimo procesas ne tik padeda rasti parametrų kombinacijas, kurios užtikrina geresnį tikslumo ir bendrinimo gebėjimų balansą, bet ir padeda išvengti pertreniravimo *(angl. overfitting*) bei per didelės paklaidos dėl išankstinio šališkumo (*angl. bias*). Hiperparametrų derinimas leidžia modeliams išlaikyti aukštą prognozavimo kokybę įvairiuose duomenų pogrupiuose, taip didinant jų stabilumą ir bendrą analitinį patikimumą ekonominių rodiklių prognozavimo kontekste.

Šiame tyrime hiperparametrų optimizavimui taikyta *RandomizedSearchCV* funkcija iš *scikit-learn* bibliotekos [120]. Priešingai nei *GridSearchCV*, kuri išsamiai peržiūri visas parametrų kombinacijas, *RandomizedSearchCV* efektyviau tyrinėja hiperparametrų erdvę, atsitiktinai atrinkdama iš nurodytų parametrų pasiskirstymų fiksuotą kombinacijų skaičių. Toks metodas yra ypač naudingas, kai hiperparametrų erdvė yra plati ir visų galimų kombinacijų išbandymas reikalautų labai daug skaičiavimo išteklių, todėl ne visuomet yra praktiškai įgyvendinamas [38].

Optimizavimo procedūra buvo atliekama naudojant tą patį 20 % trūkstamumo scenarijų, kaip ir bazinių modelių vertinime, siekiant užtikrinti rezultatų palyginamumą. Baziniai modeliai buvo treniruojami su 5-fold kryžminiu validavimu (CV=5), naudojant standartines hiperparametrų reikšmes. Optimizuoti modeliai taip pat buvo validuojami su 5-fold kryžminiu validavimu, nors pati

hiperparametrų paieška buvo atliekama su 3-fold kryžminiu validavimu, siekiant sumažinti bendrą skaičiavimo laiką optimizavimo metu.

Random Forest algoritmo hiperparametrų paieškos erdvė apėmė penkis pagrindinius parametrus (žiūrėti. 20 pav.):

- n_estimators: randint(50, 301) - sprendimų medžių skaičius ansamblyje, atrenkamas iš intervalo [50, 300];
- max_depth: [5, 10, 15, 20, 25, 30, None] - maksimalus medžio gylis;
- min_samples_split: randint(2, 21) - minimalus stebėjimų skaičius, reikalingas mazgui padalinti;
- min_samples_leaf: randint(1, 11) - minimalus stebėjimų skaičius lapiniame mazge;
- max_features: ['sqrt', 'log2', 0.3, 0.5, 0.7] - atsitiktinai atrenkamų požymių skaičius kiekviename mazge.

```
return {
    'n_estimators': randint(50, 301),  # 50-300
    'max_depth': [5, 10, 15, 20, 25, 30, None],
    'min_samples_split': randint(2, 21),  # 2-20
    'min_samples_leaf': randint(1, 11),  # 1-10
    'max_features': ['sqrt', 'log2', 0.3, 0.5, 0.7]
}
```

Šaltinis: sudaryta darbo autoriaus.

20 pav. Random Forest hiperparametrų paieškos erdvė RandomizedSearchCV metodui

XGBoost algoritmo hiperparametrų paieškos erdvė apėmė aštuonis pagrindinius parametrus (žr. 21 pav.):

- n_estimators: randint(50, 301) - sprendimų medžių skaičius;
- learning_rate: uniform(0.01, 0.29) - mokymosi greitis, atrenkamas iš intervalo [0.01, 0.3];
- max_depth: randint(3, 16) - maksimalus medžio gylis;
- subsample: uniform(0.6, 0.4) - duomenų imties dalis kiekvienam medžiui, [0.6, 1.0];
- colsample_bytree: uniform(0.6, 0.4) - požymių dalis kiekvienam medžiui, [0.6, 1.0];
- min_child_weight: randint(1, 11) - minimalus lapinio mazgo svoris;
- reg_alpha: uniform(0, 1) - L1 reguliarizacijos parametras;
- reg_lambda: uniform(0, 1) - L2 reguliarizacijos parametras.

```
return {
    'n_estimators': randint(50, 301),  # 50-300
    'learning_rate': uniform(0.01, 0.29),  # 0.01-0.3
    'max_depth': randint(3, 16),  # 3-15
    'subsample': uniform(0.6, 0.4),  # 0.6-1.0
    'colsample_bytree': uniform(0.6, 0.4),  # 0.6-1.0
    'min_child_weight': randint(1, 11),  # 1-10
    'reg_alpha': uniform(0, 1),  # 0-1
    'reg_lambda': uniform(0, 1)  # 0-1
}
```

Šaltinis: sudaryta darbo autoriaus.

21 pav. XGBoost hiperparametrų paieškos erdvė RandomizedSearchCV metodui

RandomizedSearchCV procedūra kiekvienam algoritmui atliko 100 atsitiktinių hiperparametrų kombinacijų testavimą su 3-fold kryžminiu validavimu (*angl. 3-fold cross-validation*). Geriausios parametrų kombinacijos buvo atrinktos pagal R^2 (determinacijos koeficiento) metriką, kuri matuoja modelio paaiškintos dispersijos dalį. Po optimizavimo procedūros, geriausi modeliai buvo validuojami su 5-fold kryžminiu validavimu, kad būtų galima objektyviai palyginti su baziniais modeliais.

Bazinių ir optimizuotų modelių palyginimas pateiktas 7 lentelėje. Lentelėje matyti, kad optimizavimo įtaka skirtingiems algoritmams buvo nevienoda.

7 lentelė Bazinių ir optimizuotų modelių palyginamieji rezultatai su 5-fold kryžminiu validavimu

Modelis	**Parametrai**	**R^2 (mean ± std)**	**nRMSE (mean ± std)**	**nMAE (mean ± std)**	**sMAPE (%) (mean ± std)**
RF bazinis	numatytieji	0,7942 ± 0,7455	0,0423 ± 0,0344	0,0241 ± 0,0191	17,81 ± 19,99
RF optimizuotas	optimizuoti	0,8868 ± 0,2269	0,0359 ± 0,0285	0,0202 ± 0,0192	15,22 ± 19,59
XGB bazinis	numatytieji	0,9083 ± 0,3110	0,0301 ± 0,0295	0,0159 ± 0,0156	13,27 ± 18,50
XGB optimizuotas	optimizuoti	0,9046 ± 0,2342	0,0313 ± 0,0281	0,0175 ± 0,0166	15,39 ± 19,65

Šaltinis: sudaryta darbo autoriaus.

Kaip matyti 7 lentelėje, Random Forest modelis pasiekė reikšmingą pagerėjimą po hiperparametrų optimizavimo. R^2 rodiklis padidėjo nuo 0,7942 iki 0,8868, o svarbiausia - standartinis nuokrypis žymiai sumažėjo nuo 0,7455 iki 0,2269, kas rodo gerokai stabilesnį modelio veikimą tarp skirtingų kryžminio validavimo iteracijų. Normalizuota vidutinė kvadratinė paklaida (nRMSE) sumažėjo nuo 0,0423 iki 0,0359, o normalizuota vidutinė absoliuti paklaida (nMAE) - nuo 0,0241 iki 0,0202. sMAPE metrika taip pat pagerėjo, sumažėjusi nuo 17,81 % iki 15,22 %. Šie pokyčiai rodo, kad optimalių hiperparametrų parinkimas leido Random Forest algoritmui geriau prisitaikyti prie regioninių ekonominių rodiklių struktūros ir sumažinti rezultatų svyravimus.

XGBoost modelio atveju optimizavimo įtaka buvo mažiau akivaizdi. R^2 rodiklis sumažėjo nežymiai - nuo 0,9083 iki 0,9046, tačiau standartinis nuokrypis taip pat sumažėjo - nuo 0,3110 iki 0,2342, kas indikuoja stabilesnį modelio veikimą skirtingose duomenų atrankose. nRMSE metrika šiek tiek padidėjo - nuo 0,0301 iki 0,0313, o nMAE šiek tiek suprastėjo - nuo 0,0159 iki 0,0175. sMAPE metrika taip pat pablogėjo - nuo 13,27 % iki 15,39 %. Šie rezultatai rodo, kad XGBoost algoritmas jau su baziniais parametrais pasiekė labai aukštą veikimą, todėl papildoma hiperparametrų paieška reikšmingo tikslumo augimo nesuteikė, nors ir šiek tiek padidino modelio stabilumą.

Skaičiavimo efektyvumo požiūriu, hiperparametrų optimizavimo procedūra yra gerokai intensyvesnė nei bazinių modelių treniravimas, kadangi ji reikalauja pakartotinių modelio suderinimų su skirtingomis parametrų kombinacijomis. Atlikti eksperimentai parodė, kad RandomizedSearchCV su 100 iteracijų ir 3-fold kryžmine validacija pareikalavo reikšmingo skaičiavimo laiko: Random Forest modelio optimizavimo procesas truko 3595,99 sekundės (apie 59,93 min.), o XGBoost optimizavimas - 5893,86 sekundės (apie 98,23 min.). Optimizavimas buvo vykdomas naudojant parametrą *n_jobs=10*, reiškiantį, kad hiperparametrų paieškos metu buvo lygiagrečiai paleista iki dešimties nepriklausomų skaičiavimo gijų. Šis nustatymas buvo parinktas atsižvelgiant į naudojamo procesoriaus - AMD Ryzen 5 PRO 4650G - architektūrą, kuri turi 6 fizinius branduolius ir 12 gijų. Todėl 10 gijų buvo skirta RandomizedSearchCV procesams, o likusios 2 gijos paliktos operacinei sistemai bei foninėms programoms, siekiant išvengti visiško sistemos apkrovimo ir užtikrinant stabilų kompiuterio veikimą.

Gauti rezultatai atskleidžia svarbią įžvalgą apie hiperparametrų optimizavimo poveikį skirtingiems algoritmams. Random Forest, būdamas santykinai paprastas ansamblio metodas, akivaizdžiai pasinaudojo sistemingu hiperparametrų derinimu - bazinių parametrų atveju jo rezultatų dispersija buvo itin didelė (*std = 0,7455*), o tai rodė nestabilų modelio veikimą. Optimizavimas šį svyravimą sumažino iki 0,2269, t. y. daugiau nei tris kartus, taip ženkliai padidindamas modelio patikimumą.

Tuo tarpu XGBoost algoritmas, kaip pažangus gradientinio stiprinimo metodas su įtaisytais reguliarizacijos mechanizmais, jau su numatytaisiais parametrais pasiekė aukštą ir santykinai stabilų veikimą. Nors optimizavimas nesuteikė didelio tikslumo pagerėjimo, jis sumažino rezultatų dispersiją (nuo 0,3110 iki 0,2342), o tai rodo, kad modelis tapo stabilesnis skirtingose duomenų atrankose.

Išsamesnė optimizuotų modelių metrikų analizė pagal atskirus ekonominius rodiklius pateikta 12 - 15 prieduose. Kadangi tyrimo duomenų rinkinys apima 78 ekonominius rodiklius, o išsami kiekvieno rodiklio analizė būtų per didelės apimties pagrindiniam tekstui, prieduose pateikiama 19 ekonominių rodiklių detali metrikų palyginimo vizualizacija stulpelinių diagramų forma. Šie grafikai leidžia įvertinti, kaip optimizuoti Random Forest ir XGBoost modeliai veikia skirtinguose rodiklių kontekstuose: 12 priede pateikiamas R^2 metrikų palyginimas, atspindintis kiekvieno modelio

gebėjimą paaiškinti rodiklio dispersiją; 13 priede - nRMSE palyginimas, rodantis normalizuotą vidutinę kvadratinę paklaidą; 14 priede - nMAE palyginimas, iliustruojantis normalizuotą vidutinę absoliučią paklaidą; 15 priede - sMAPE palyginimas, atspindintis simetrinę vidutinę absoliučią procentinę paklaidą. Vizualizacijose aiškiai matoma, kad XGBoost modelis daugumoje rodiklių pasiekia aukštesnes R^2 reikšmes (artimesnes 1) ir žemesnes paklaidos metrikas, kas patvirtina 7 lentelėje pateiktus agreguotus rezultatus ir rodo modelio pranašumą tiek vidutiniškai, tiek individualiems rodikliams.

Atsižvelgiant į atliktą analizę, matyti, kad hiperparametrų optimizavimas darė skirtingą poveikį abiems algoritmams. Random Forest modelis po optimizavimo tapo stabilesnis ir sumažino rezultatų dispersiją, tačiau, net ir optimizavus, nepasiekė tokio aukšto tikslumo kaip XGBoost bazinis modelis. Tuo tarpu XGBoost, nors ir neparodė ryškaus tikslumo pagerėjimo po optimizavimo, jau pradiniame variante demonstravo geriausius R^2, nRMSE, nMAE ir sMAPE rodiklius. Todėl, vertinant bendrą užpildymo kokybę, XGBoost galima laikyti tinkamesniu pasirinkimu NUTS 2 regionų ekonominių rodiklių trūkstamų reikšmių užpildymui, o optimizuotas Random Forest gali būti laikomas naudinga alternatyva, pasižyminčia didesniu rezultatų stabilumu.

3.6 Random Forest ir XGBoost modelių prognozių korekcija taikant empirinio Bajeso sutraukimo (shrinkage) metodą

Atliekant trūkstamų rodiklių reikšmių užpildymą naudojant du nepriklausomus mašininio mokymosi modelius - Random Forest ir XGBoost - nustatyta, kad abu modeliai turėjo tą pačią sisteminę problemą: dalis užpildytų prognozių kai kuriuose regionuose tapo nerealistiškai didelės, palyginti su ilgalaikiais regioniniais vidurkiais. Tai ypač išryškėjo AT11 regione rodiklio SCH_IS_02 atveju, kai modelis sugeneravo prognozę lygią 7905,859331, nors 2000 - 2023 m. AT11 vidurkis tėra 1267,96. Tokie atvejai rodo perteklinės sklaidos (*angl. over-dispersion*) efektą, kai modelio prognozės pernelyg nutolsta nuo realaus ekonominio konteksto, kaip pažymima literatūroje [121], [122].

Abiejų modelių implementacijos tai patvirtina, nes jų prognozės remiasi tais pačiais duomenų su trūkumais apdorojimo principais ir vienoda regioninių statistinių požymių inžinerija, todėl tiek Random Forest, tiek XGBoost iš esmės paveldi tą pačią struktūrinę riziką: regionuose su mažai istorinių duomenų prognozės gali tapti nestabilios.

Atsižvelgiant į identifikuotas problemas ir aiškiai išreikštą anomalijų pobūdį, šiame tyrime buvo papildomai taikomas *Empirical Bayes Shrinkage* metodas, kurio tikslas - stabilizuoti modelio prognozes ir pritraukti jas link ilgalaikių regiono vidurkių, taip išvengiant ekonomiškai neįmanomų šuolių. Šis metodas remiasi prielaida, kad patikimiausia išankstinė tikimybė (*angl. prior*) yra regiono

istorinių stebėjimų vidurkis, o modelio prognozė atitinka a posteriori tikimybę (*angl. posterior*), kuri turi būti koreguojama pagal dispersijų santykį.

Šiame tyrime *Empirical Bayes Shrinkage* metodas taikomas kaip papildoma priemonė suderinti modelio prognozes su ilgalaikėmis regiono istorinėmis tendencijomis, siekiant sumažinti triukšmo ir riboto duomenų kiekio lemiamą neapibrėžtumą. Metodas leidžia adaptuoti kiekvieną prognozę pagal regiono stabilumą: kuo mažiau informatyvūs istoriniai duomenys arba kuo didesnė modelio dispersija, tuo stipresnė korekcija. Toliau pateikiamos dvi formulės, kurios matematiškai apibrėžia šio proceso veikimą.

Empirical Bayes Shrinkage metodo taikymui naudojama (23) formulė:

$$\hat{y}_{EB} = (1 - \lambda)\,\hat{y}_{ML} + \lambda\,\bar{y}_{region}, \tag{23}$$

kur:

- $\hat{y}_{EB}$ - galutinė shrinkage pakoreguota prognozė,
- $\hat{y}_{ML}$ - pirminė modelio (RF arba XGB) prognozė,
- $\bar{y}_{region}$ - regiono 2000 - 2024 m. istorinis vidurkis,
- λ - shrinkage koeficientas nuo 0 iki 1.

Shrinkage svorio λ apskaičiuojamas pagal (24) formulę:

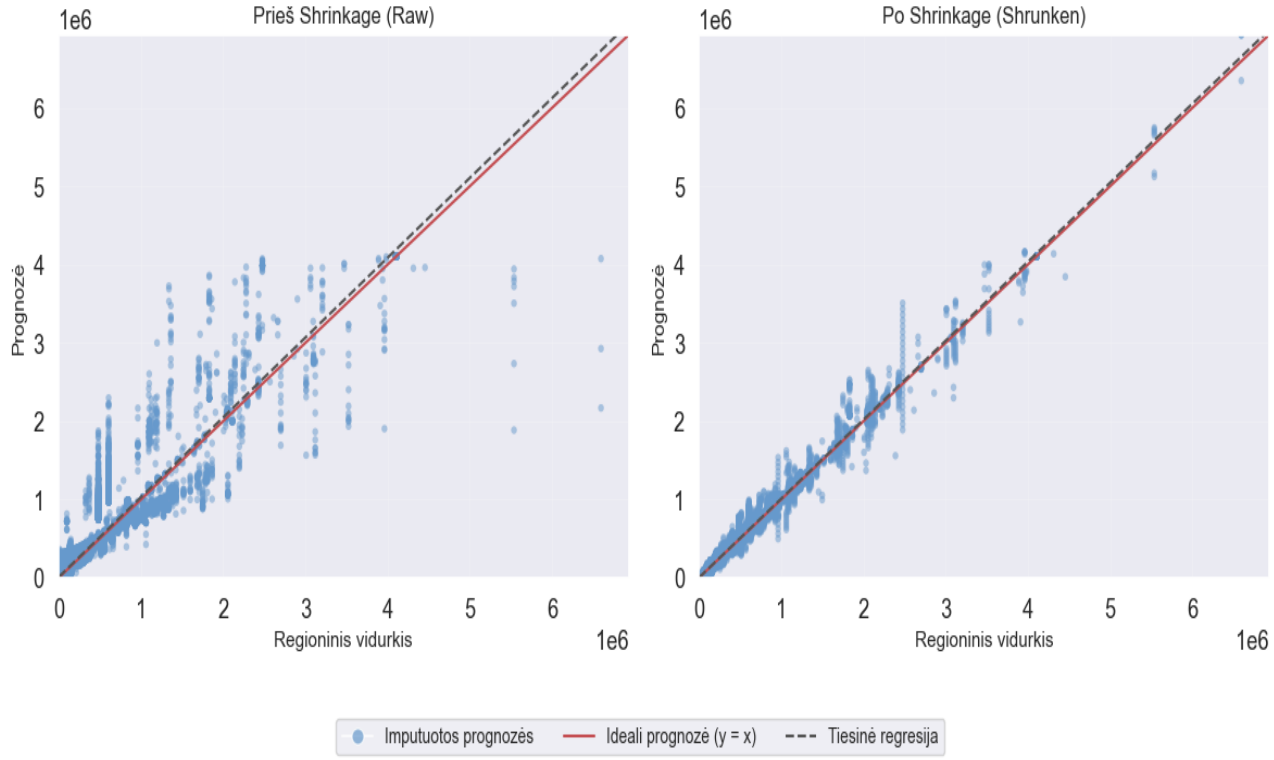
$$\lambda = \frac{\sigma^2_{ML}}{\sigma^2_{ML} + \sigma^2_{region}}, \tag{24}$$

kur:

- σ^2_{ML} - modelio prognozių dispersija regionui,
- σ^2_{region} - regiono istorinių duomenų dispersija.

Literatūroje pabrėžiama, kad kuo mažesnė regiono dispersija arba kuo didesnis modelio nestabilumas, tuo labiau shrinkage „traukia“ prognozę link vidurkio [123], [124].

Ši teorinė logika leidžia tikėtis, kad korekcija sumažins pernelyg išsisklaidžiusias užpildytas reikšmes ir sugrąžins jas arčiau ekonominio konteksto. Tačiau siekiant įvertinti realų metodo poveikį būtina ne tik matematinė, bet ir vizualinė analizė, leidžianti stebėti, kaip keičiasi prognozių pasiskirstymas po shrinkage taikymo. Vizualizacija taip pat suteikia galimybę aiškiai atskleisti, ar sumažėjo ekstremalių reikšmių skaičius ir ar prognozės tapo nuoseklesnės regionų mastu.



Šaltinis: sudaryta darbo autoriaus.

22 pav. Empirical Bayes Shrinkage poveikis Random Forest užpildytų prognozių dispersijai

Pateikta 22 paveikslo vizualizacija aiškiai parodo Empirical Bayes Shrinkage metodo efektyvumą stabilizuojant Random Forest užpildytas prognozes: pradiniame variante (Raw) matomas ryškus „uodegos“ efektas, kai daugelio regionų prognozės ženkliai viršija jų ilgalaikius vidurkius, o tai rodo padidėjusią dispersiją ir modelio jautrumą triukšmui. Po shrinkage taikymo užpildytų verčių pasiskirstymas tampa gerokai kompaktiškesnis, o taškai priartėja prie tapatybės linijos (y = x), kas liudija apie padidėjusį nuoseklumą ir geresnį atitikimą regionų istorinei struktūrai. Be to, tiesinės regresijos kreivė po korekcijos beveik sutampa su idealia linija, todėl galima teigti, kad prognozių sisteminis nuokrypis sumažėjo, o ekstremalių išskirtinių reikšmių (*angl. outlier*) taškų kiekis pastebimai sumažėjo. Šie rezultatai visiškai atitinka Empirical Bayes literatūroje aprašytą tikslą - mažų imčių ir didelio triukšmo atveju prognozes „traukti“ link patikimesnės išankstinės tikimybės (*angl. prior*), taip sumažinant dispersiją ir pagerinant prognozių stabilumą [125].

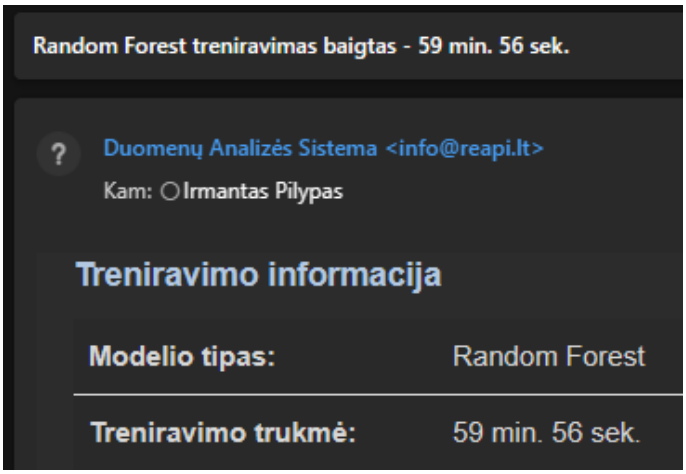
Nors abu modeliai patyrė tą pačią pirminę problemą - prognozių nutolimą nuo regioninių vidurkių - *Empirical Bayes Shrinkage* metodas veikė skirtingai efektyviai, o Random Forest atveju suteikė žymiai stabilesnius rezultatus. Tai galima paaiškinti keliais veiksniais: Random Forest prognozavimo mechanizmas natūraliai mažiau jautrus ekstremalioms reikšmėms, nes medžiai generuoja prognozes vidurkio principu, taip slopindami triukšmą, tuo tarpu XGBoost gradiento stiprinimo metodu (*angl. boosting*) procese linkęs jį sustiprinti [36]. Be to, Random Forest struktūra geriau suderinama su regioninėmis statistinėmis savybėmis, todėl šis modelis išlaiko didesnį

atsparumą mažoms imtims ir regionams, turintiems ribotą istorinę informaciją. Dėl to shrinkage korekcija čia tampa veiksmingesnė - prognozių dispersija sumažėja nuosekliau, o užpildytos reikšmės geriau atitinka realų ekonominį kontekstą.

3.7 Tyrimo rezultatų taikymas ir sistemos diegimo architektūra

Pagrindinis šio tyrimo objektas yra *Tyrimas_Ekonominiu_Rodikliu_Imputacija.ipynb* Jupyter Notebook failas, kuriame įgyvendinta visa eksperimentinė metodologija - nuo duomenų paruošimo iki galutinių užpildytų reikšmių generavimo. Jupyter Notebook aplinka pasirinkta dėl jos interaktyvumo, vizualizacijos galimybių ir gebėjimo dokumentuoti tyrimo eigą kartu su kodu, kas yra ypač svarbu mokslinių tyrimų atkuriamumui [126]. Tyrimo proceso automatizavimas apima ne tik modelio treniravimą ir vertinimą, bet ir sistemingą rezultatų failų generavimą bei pranešimų sistemą, užtikrinančią tyrėjo informuotumą apie ilgai trunkančių skaičiavimo procesų būklę.

Modelio treniravimo procesas, ypač dirbant su dideliais duomenų rinkiniais ir sudėtingais hiperparametrų optimizavimo algoritmais, gali užtrukti nuo kelių minučių iki kelių valandų. Siekiant efektyviai valdyti tyrimo procesą ir informuoti tyrėją apie treniravimo progresą, į tyrimo kodą integruota automatinė el. pašto pranešimų sistema. Šiai funkcijai realizuoti panaudotos Python *smtplib* ir *email* bibliotekos, leidžiančios siųsti el. laiškus tiesiogiai iš Python scenarijų [127], [128].

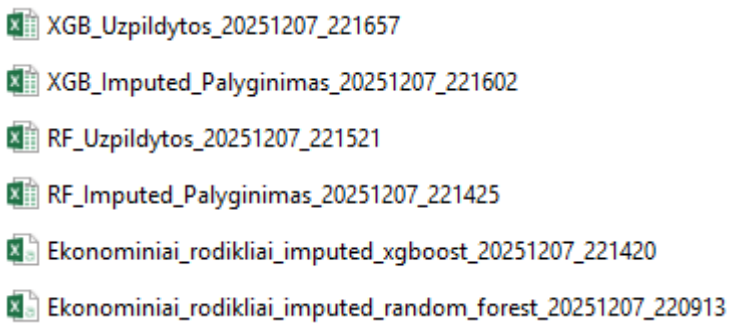


Šaltinis: sudaryta autoriaus

23 pav. Automatinis el. pašto pranešimas apie modelio treniravimo pabaigą

Kaip matyti 23 paveikslėlyje, pranešimas informuoja apie modelio tipą (šiuo atveju Random Forest) ir tikslų treniravimosi laiką (59 minutės 56 sekundės). Tokia funkcionalumo integracija leidžia tyrėjui vykdyti kitus darbus, o apie pasibaigusį treniravimą jis yra nedelsiant informuojamas. Ši implementacija užtikrina, kad kiekvienas esminis modelio treniravimo etapas būtų dokumentuojamas realiu laiku, kas yra kritiškai svarbu atliekant kelių valandų trukmės eksperimentus.

Tyrimo metu Jupyter Notebook aplinkoje generuojami šeši pagrindiniai rezultatų failai, kurie sistemingai dokumentuoja kiekvieną eksperimentinio tyrimo etapą. Pateiktas 24 paveikslėlis, kuriame matyti rezultatų failų sąrašas su jų sugeneravimo laiku ir formato specifikacijomis.



Šaltinis: sudaryta autoriaus

24 pav. Tyrimo rezultatų failų struktūra

Failų generavimas atliekamas naudojant kelias pagrindines Python bibliotekas: *pandas* (versija 2.0.3) duomenų manipuliacijai ir eksportavimui, *openpyxl* (versija 3.1.2) Excel formato failų kūrimui ir *numpy* (versija 1.24.4) skaitinėms operacijoms [129]. Kiekvienas failas atlieka specifinę funkciją tyrimo dokumentavime:

1. XGB_Uzpildytos_[data]_[laikas].xlsx - XGBoost modelio užpildytos ekonominių rodiklių reikšmės su vizualiniu formatavimu. Šiame faile visos modelio užpildytos reikšmės pažymėtos geltonu fonu (hex: #FFEB9C) su rudos spalvos šriftu (hex: #9C6500), kas leidžia identifikuoti, kurios reikšmės buvo užpildytos, o kurios yra originalios. Failas apima du lakštus: „Užpildytos Reikšmės“ su pilnu duomenų rinkiniu ir „Legenda“ su spalvų paaiškinimais.
2. XGB_Imputed_Palyginimas_[data]_[laikas].xlsx - XGBoost modelio prognozių palyginimas su faktinėmis reikšmėmis testavimo rinkinyje, naudojant pažangų *Rich Text* formatavimą. Šis failas unikalioje vizualizacijoje rodo 20 % sintetiškai pašalintų žinomų reikšmių palyginimą: tikroji reikšmė pateikiama žalia spalva (hex: #006100), o modelio prognozė - ruda spalva (hex: #9C6500), atskirtos pasvirojo brūkšnio simboliu („/“). Pavyzdžiui, langelyje matomas įrašas „26,2 / 29,204504“ reiškia, kad tikroji reikšmė buvo 26,2, o XGBoost modelis prognozavo 29,204504. Toks vizualus palyginimas leidžia greitai įvertinti modelio tikslumo pasiskirstymą visame duomenų rinkinyje. Failas taip pat apima „Test Set Comparison“ lakštą su visais duomenimis ir „Legenda“ lakštą su detaliu paaiškinimu apie duomenų rinkinį.
3. RF_Uzpildytos_[data]_[laikas].xlsx - Random Forest modelio užpildytos reikšmės su analogišku vizualiniu formatavimu kaip XGBoost versijoje. Visos užpildytos reikšmės

pažymėtos geltonos spalvos fonu su rudu šriftu, o originalios reikšmės lieka be formatavimo. Struktūra identiška XGBoost failui: „Užpildytos Reikšmės“ lakštas ir „Legenda“ lakštas.

4. RF_Imputed_Palyginimas_[data]_[laikas].xlsx - Random Forest modelio prognozių palyginimas su faktinėmis reikšmėmis testavimo rinkinyje, naudojant *Rich Text* formatavimą. Kiekviename langelyje, kuriame atliktas sintetinis testavimas, pateikiama dviejų spalvų reikšmė: žaliai - tikroji reikšmė, ruda - modelio prognozuota reikšmė. Šis failas leidžia vizualiai palyginti Random Forest modelio prognozavimo tikslumą ir identifikuoti regionus ar rodiklius, kuriuose modelis veikia geriau ar prasčiau.

5. Ekonominiai_rodikliai_imputed_xgboost_[data]_[laikas].csv - galutinis pilnai užpildytas NUTS 2 regionų ekonominių rodiklių duomenų rinkinys CSV formatu, parengtas taikant optimizuotą XGBoost modelį trūkstamų reikšmių užpildymui. Duomenų failas apima 78 ekonominių rodiklių stulpelius bei identifikacinius kintamuosius (geo ir year) ir pateikia duomenis kiekvienam iš 244 NUTS 2 regionų per 25 metų laikotarpį (2000 - 2024 m.), suformuojant 6100 pilnai užpildytų stebėjimų. Duomenų rinkinys eksportuojamas UTF-8 kodavimu ir yra tinkamas tolesnei ekonominių rodiklių analizei, regionų klasterizacijai, laiko eilučių modeliavimui bei kitoms analitinėms užduotims statistinėse aplinkose R, SPSS, Stata ir Python.

6. Ekonominiai_rodikliai_imputed_random_forest_[data]_[laikas].csv - galutinis pilnai užpildytas NUTS 2 regionų ekonominių rodiklių duomenų rinkinys CSV formatu, naudojant optimizuotą Random Forest modelį. Struktūra ir formatas identiškas XGBoost versijai, tačiau visos trūkstamos reikšmės užpildytos Random Forest algoritmu. Šis failas leidžia tyrėjui palyginti dviejų skirtingų algoritmų užpildymo rezultatus ir pasirinkti tinkamesnius duomenis konkrečiai analizei.

Failų generavimas atliekamas naudojant *pandas.DataFrame.to_excel()* metodą, kuris užtikrina formatavimo išsaugojimą ir duomenų tipo suderinamumą su Microsoft Excel programa.

Siekiant objektyviai įvertinti modelio prognozavimo tikslumą, atliekamas sintetinis testavimas, kai 20 procentų žinomų reikšmių dirbtinai pašalinamos iš duomenų rinkinio.

Pateiktame 25 paveikslėlyje, galima matyti, XGBoost modelio testavimo rezultatų fragmentą su realiomis ir prognozuojamomis reikšmėmis.

	A	B	C	D
1	geo	year	Y_01	Y_02
2	AT11	2000	23,2	59
3	AT11	2001	25	57,1
4	AT11	2002	26,4	58,1
5	AT11	2003	26	59,3
6	AT11	2004	28,4	56,8
7	AT11	2005	25,9	59,7
8	AT11	2006	27	60,3
9	AT11	2007	25,6	62
10	AT11	2008	26,2 / 29,204504	62,3 / 61,555321
11	AT11	2009	35	61,9
12	AT11	2010	37,3	62,2
13	AT11	2011	38	61,7
14	AT11	2012	36,5 / 36,422932	61,7 / 60,950394

Šaltinis: sudaryta autoriaus

25 pav. XGBoost modelio sintetinio testavimo rezultatai esant 20 % duomenų trūkumui

Kaip matyti 25 paveikslėlyje, pateikiamas testavimo duomenų failo iškarpos fragmentas, kuriame kiekvienam stebėjimui nurodomas geografinis regionas (geo), metai (year) bei du prognozuojami ekonominiai rodikliai (Y_01 ir Y_02). Abu šie rodikliai testavimo metu buvo dirbtinai užmaskuoti, o vėliau jų reikšmės atkuriamos taikant XGBoost modelį, siekiant įvertinti prognozavimo tikslumą. Pavyzdžiui, AT11 regiono 2008 metų duomenims Y_01 rodiklio faktinė reikšmė buvo 26,2, o modelio prognozuota reikšmė - 29,20450, kas atitinka maždaug 3 procentinių punktų paklaidą. Tokia užmaskuotų reikšmių prognozavimo metodologija leidžia kiekybiškai įvertinti modelio gebėjimą atkurti trūkstamus duomenis ir yra plačiai taikoma mašininio mokymosi tyrimų praktikoje [130], [131], [132].

Atlikus visus testavimo ir validavimo etapus, geriausiai veikiantys modeliai (optimizuotas Random Forest su R^2 = 0,8868 ir optimizuotas XGBoost su R^2 = 0,9046) taikomi visoms tikrai trūkstamoms reikšmėms originaliame duomenų rinkinyje, generuojant du galutinius užpildytus duomenų rinkinius. Pateiktame 26 paveikslėlyje, galima matyti, galutinio užpildyto duomenų failo fragmentą su užpildytomis ekonominių rodiklių reikšmėmis.

	A	B	C	D	E
1	geo,year,Y_01,Y_02,SCH_IS_01,SCH_IS_02,SCH_IS_03,S				
2	AT11,2000,23.2,59.0,211.0,1045.0,888.0,398.0,866.0,941				
3	AT11,2001,25.0,57.1,242.0,1096.0,917.0,388.0,874.0,958				
4	AT11,2002,26.4,58.1,231.0,1121.0,927.0,389.0,920.0,101				
5	AT11,2003,26.0,59.3,254.0,1043.0,857.0,426.0,945.0,103				
6	AT11,2004,28.4,56.8,275.0,1068.0,847.0,431.0,961.0,106				
7	AT11,2005,25.9,59.7,203.0,1013.0,801.0,452.0,991.0,111				
8	AT11,2006,27.0,60.3,227.0,1054.0,851.0,436.0,1043.0,11				
9	AT11,2007,25.6,62.0,257.0,1113.0,905.0,497.0,1116.0,12				
10	AT11,2008,26.2,62.3,226.0,1075.0,854.0,521.0,1071.0,11				
11	AT11,2009,35.0,61.9,194.0,1133.0,907.0,520.0,1114.0,12				
12	AT11,2010,37.3,62.2,221.0,1236.0,986.0,537.0,1175.0,12				
13	AT11,2011,38.0,61.7,256.0,1247.0,1012.0,574.0,1264.0,1				

Šaltinis: sudaryta autoriaus

26 pav. Galutinio užpildyto ekonominių rodiklių duomenų .csv failo struktūros iškarpos fragmentas

Pateiktame galutinio užpildyto ekonominių rodiklių duomenų failo iškarpos fragmente pavaizduota CSV failo struktūra, apimanti visus 78 ekonominių rodiklių stulpelius (SCH_IS_01, SCH_IS_02, SCH_IS_03 ir t. t.) su užpildytomis reikšmėmis kiekvienam NUTS 2 regionui ir metams. Šis galutinis duomenų rinkinys sudaro pagrindą tolesnei ekonominių rodiklių analizei, įskaitant regionų atsparumo prognozavimą [133], ekonominių tendencijų modeliavimą ir klasterizacijos analizes. Eksportuojamas CSV formato failas užtikrina platų suderinamumą su įvairiomis duomenų analizės platformomis ir statistiniais paketais [134].

Greta pagrindinio *Jupyter Notebook* tyrimo, buvo sukurta papildoma prototipinė žiniatinklio aplikacija, naudojant Flask karkasą (versija 2.3.3), skirta užpildytų duomenų vizualizacijai ir interaktyviam tyrimui [135]. Flask pasirinktas dėl jo paprastumo, moduliškumo ir integravimosi suderinamumu su Python mašininio mokymosi bibliotekomis. Aplikacijos struktūrinė schema - diegimo diagrama (*angl. Deployment diagram*) - pateikta 11 priede ir iliustruoja pagrindinius sistemos komponentus bei jų tarpusavio ryšius.

Flask aplikacija veikia kaip ekspertinė sistema ir duomenų vizualizacijos įrankis, leidžianti interaktyviai tyrinėti užpildytus duomenis, generuoti dinamines vizualizacijas naudojant *Plotly* biblioteką (versija 5.15.0) ir atlikti prognozavimus be *Jupyter Notebook* aplinkos [23]. Aplikacijoje integruotas prisijungimas prie nutolusios *MySQL* duomenų bazės (naudojant mysql-connector-python biblioteką, kurios versija 9.4.0) ilgalaikiam duomenų saugojimui ir efektyviam užklausų vykdymui [136]. Duomenų bazės schema pateikta 21 priede. Papildomai naudojama reportlab biblioteka (versija 4.2.0), skirta automatiniam PDF formato ataskaitų generavimui, kuriose pateikiamos trūkstamų reikšmių užpildymo tikslumo metrikų rezultatų santraukos [137].

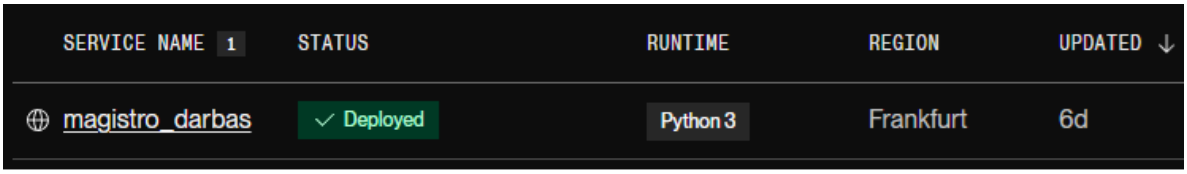
Svarbu pažymėti, kad Flask aplikacija yra tik prototipinė versija, skirta koncepcijos įrodymui ir duomenų tyrimui, o ne pilnai funkcionalus galutinis sprendimas. Pagrindinis tyrimo objektas išlieka

Tyrimas_Ekonominiu_Rodikliu_Imputacija.ipynb failas, kuriame atliekamas visas eksperimentinis darbas.

Prototipinės Flask aplikacijos diegimas į produkcinę aplinką (nutolusį *WEB* serverį) įgyvendintas naudojant modernius DevOps principus ir automatizuoto diegimo (*angl. CI/CD - Continuous Integration / Continuous Deployment*) metodologiją [138]. Bet kokie pakeitimai, atliekami Visual Studio Code aplinkoje, po paskutinių kodo modifikacijų automatiškai siunčiami į GitHub platformą naudojant Git versijų kontrolės sistemą [139]. GitHub repozitorijoje adresu (https://github.com/Pilypas/magistro_darbas) laikomas visas projekto kodas, įskaitant pagrindinį tyrimo *Jupyter Notebook* .ipynb tyrimo failą, *Flask* aplikacijos failus, priklausomybių sąrašą (*requirements.txt*) ir konfigūracijos failus.

GitHub platforma integruota su *Render.com* - moderniu debesijos talpinimo servisu, kuris automatiškai aptinka kodo pakeitimus *GitHub* repozitorijoje ir inicijuoja automatinį diegimo procesą [140]. Šis procesas žinomas kaip automatinis diegimas (*angl. Continuous Deployment*) iš Git repozitorijos. Kai tik naujas pakeitimas (*angl. commit*) įkeliamas į GitHub pagrindinę (main arba master) šaką, Render.com automatiškai:

1. aptinka pakeitimus per GitHub įvykio iškvietimo (*angl. webhook*) mechanizmą;
2. klonuoja naujausią kodo versiją iš repozitorijos;
3. instaliuoja visas priklausomybes iš (*requirements.txt*) failo;
4. atlieka aplikacijos build procesą;
5. įdiegia (*angl. deploy*) naują aplikacijos versiją į serverius;
6. atnaujina veikiančią aplikacijos instanciją be prastovų (*angl. zero-downtime deployment*) [141].



Šaltinis: sudaryta autoriaus

27 pav. Automatinio Flask aplikacijos diegimo į Render.com platformą rezultatas

Kaip matyti 27 paveiksle, aplikacija sėkmingai įdiegta ir jai priskirtas *Deployed* statusas, veikia *Python* 3 aplinkoje ir pasiekiama per *Frankfurt* regioną [142]. Diegimo rezultate galima pastebėti, kad aplikacija atnaujinta prieš 6 dienas ir veikia stabiliai. Toks automatizuotas diegimo procesas užtikrina, kad bet kokie kodo patobulinimai ar klaidų taisymai yra nedelsiant prieinami galutiniams vartotojams be žmogaus įsikišimo. Render.com platforma automatiškai tvarko infrastruktūrą, taip pat užtikrina HTTPS saugumą, automatinį mastelio keitimą (*angl. scaling*) [143] ir atsarginių kopijų kūrimą [144]. Diegimo konfigūracija apibrėžta (*render.yaml*) faile [76].

Aplikacijos priklausomybės apibrėžtos (*requirements.txt)* faile ir apima visas būtinas bibliotekas nuo Flask karkaso iki mašininio mokymosi bibliotekų (*scikit-learn, xgboost*) ir duomenų vizualizacijos įrankių (*matplotlib, seaborn, plotly*).

3.8 Tyrimo rezultatų išvados ir rekomendacijos

Atliktas eksperimentinis tyrimas patvirtino, kad mašininio mokymosi metodai gali būti efektyviai taikomi trūkstamų NUTS 2 lygmens regionų ekonominių rodiklių reikšmių prognozavimui. Analizuojant tris dirbtinai suformuotus trūkstamumo scenarijus (10 %, 20 % ir 30 %), nustatyta, jog didėjant trūkstamų duomenų proporcijai modelių prognozavimo tikslumas nuosekliai mažėja, tačiau pažangūs ansambliniai algoritmai išlaiko pakankamą stabilumą ir praktinį pritaikomumą.

Tyrimo rezultatai parodė, kad XGBoost algoritmas nuosekliai pranoksta Random Forest modelį pagal visas taikytas vertinimo metrikas (nRMSE, nMAE, R^2 ir sMAPE) visuose nagrinėtuose trūkstamumo scenarijuose. Tai leidžia daryti išvadą, kad gradientinio stiprinimo principu paremtas XGBoost modelis yra geriau pritaikytas fiksuoti sudėtingus nelinearinius ryšius tarp ekonominių rodiklių, būdingus regioniniams duomenims NUTS 2 lygmeniu.

Detali ekonominio rodiklio Y_01 analizė, 20 % trūkstamumo scenarijuje, parodė, kad XGBoost modelis ne tik tiksliau atkuria trūkstamas reikšmes, bet ir geriau išlaiko pirminį duomenų pasiskirstymą, kas buvo patvirtinta faktinių ir prognozuotų reikšmių palyginimu bei kernelio tankio įvertinimo (*KDE*) analizės rezultatais. Tai ypač svarbu regioninių ekonominių tyrimų kontekste, kur duomenų pasiskirstymo iškraipymai gali lemti klaidingas interpretacijas ir netikslius duomenimis grįstus sprendimus.

Hiperparametrų optimizavimo rezultatai atskleidė reikšmingą modelių veikimo pagerėjimą, ypač Random Forest atveju, kur optimizavus parametrus modelio stabilumas ir prognozavimo tikslumas padidėjo daugiau nei tris kartus. Tuo tarpu XGBoost modelis jau su baziniais parametrais pasiekė aukštą veikimo lygį, o optimizavimas suteikė papildomą, tačiau santykinai mažesnį tikslumo prieaugį. Šis rezultatas patvirtina XGBoost algoritmo atsparumą ir mažesnį jautrumą pradinių parametrų parinkimui.

Apibendrinant galima teigti, kad tyrimo metu pasiektas darbo tikslas - identifikuotas ir empiriškai pagrįstas optimalus mašininio mokymosi modelis trūkstamų NUTS 2 regionų ekonominių rodiklių reikšmių užpildymui. Gauti rezultatai patvirtina, jog optimizuotas XGBoost modelis yra tinkamiausias sprendimas, užtikrinantis aukštą prognozavimo tikslumą, stabilumą ir praktinį pritaikomumą regioninių ekonominių duomenų analizėje.

Rekomendacijos.

Remiantis atlikto tyrimo rezultatais, rekomenduojama NUTS 2 lygmens regionų ekonominių rodiklių duomenų rinkiniuose trūkstamų reikšmių užpildymui prioritetą teikti XGBoost algoritmui, ypač tais atvejais, kai duomenys pasižymi sudėtingomis tarpusavio priklausomybėmis ir dideliu duomenų trūkumo lygiu. Šis modelis gali būti sėkmingai integruojamas į oficialių statistikos institucijų ar tyrimų centrų duomenų apdorojimo procesus.

Praktiniame taikyme taip pat rekomenduojama taikyti hiperparametrų optimizavimo procedūras, ypač naudojant Random Forest modelį, kadangi baziniai parametro nustatymai gali neužtikrinti pakankamo stabilumo dirbant su didelės apimties ir heterogeniškais ekonominiais duomenimis. Optimizavimo procesas turėtų būti derinamas su skaičiavimo išteklių vertinimu, siekiant subalansuoti tikslumą ir skaičiavimo efektyvumą.

Ateities tyrimuose tikslinga išplėsti analizuojamų metodų spektrą, įtraukiant kitus pažangius ansamblinius ir giliojo mokymosi modelius, tokius kaip lengvasis gradientų stiprininimo modelis *LightGBM* (*angl. Light Gradient Boosting Machine)* ar neuroniniais tinklais grįstas trūkstamų reikšmių užpildymo strategijas [145]. Taip pat rekomenduojama nagrinėti ne tik dirbtinai generuotus, bet ir realius, struktūriškai sudėtingus trūkstamumo atvejus, siekiant dar labiau padidinti rezultatų išorinį validumą.

Galiausiai, rekomenduojama taikyti šiame darbe pasiūlytą metodinį sprendimą kaip pradinį etapą platesniuose regioninės ekonomikos atsparumo, sanglaudos politikos ar investicijų poveikio vertinimo tyrimuose, užtikrinant patikimesnius, pilnesnius ir analitiškai nuoseklius NUTS 2 regionų ekonominių rodiklių duomenų rinkinius.

IŠVADOS

1. Išanalizavus mokslinę literatūrą paaiškėjo, kad pažangūs mašininio mokymosi ir ansambliniai metodai yra tinkamesni regioninių ekonominių rodiklių trūkstamų reikšmių užpildymui nei tradicinės statistinės strategijos. Jie geba efektyviau modeliuoti sudėtingus nelinearinius ryšius tarp kintamųjų ir prisitaikyti prie sisteminio trūkstamumo, būdingo NUTS 2 lygmens duomenims. Literatūros analizė parodė, kad Random Forest ir XGBoost metodai dažniausiai pasižymi didesniu prognozavimo tikslumu ir stabilumu, tačiau jų veiksmingumas priklauso nuo duomenų struktūros ir trūkstamumo pobūdžio.
2. Išanalizavus NUTS 2 regionų ekonominių rodiklių duomenų rinkinius nustatyta, kad šie duomenys pasižymi dideliu heterogeniškumu, ilga laiko dimensija ir reikšmingu trūkstamų reikšmių kiekiu, kas apsunkina tiesioginį statistinių metodų taikymą. Parengus struktūrizuotą ir nuoseklų ekonominių rodiklių duomenų rinkinį, buvo sudarytos prielaidos taikyti mašininio mokymosi modelius trūkstamų reikšmių prognozavimui, išlaikant laiko eilučių struktūrą ir užtikrinant objektyvų modelių vertinimą.
3. Eksperimentiškai įvertinus modelių veikimą 10 %, 20 % ir 30 % duomenų trūkstamumo scenarijuose, tyrimo rezultatai parodė, kad didėjant trūkstamų reikšmių daliai prognozavimo tikslumas mažėja. Random Forest modelio R^2 sumažėjo nuo 0,9303 iki 0,9125 (-0,0178), o sMAPE padidėjo nuo 15,12 % iki 16,33 % (+1,21 procentinio punkto), tuo tarpu XGBoost modelio R^2 sumažėjo tik nuo 0,9462 iki 0,9411 (-0,0051), o sMAPE išliko 11,77 - 12,58 % intervale. XGBoost visais nagrinėtais trūkstamumo scenarijais pasiekė mažesnes paklaidas - jo nRMSE (0,0275 - 0,0287) buvo apie 22 - 25 % mažesnė, o nMAE (0,0154 - 0,0166) apie 28 - 29 % mažesnė nei Random Forest, patvirtinant didesnį XGBoost prognozavimo tikslumą ir stabilumą didėjant duomenų trūkstamumo lygiui.
4. Atlikus mašininio mokymosi modelių hiperparametrų optimizavimą taikant 20 % duomenų trūkstamumo scenarijų nustatyta, kad optimizavimas turėjo reikšmingą poveikį modelių tikslumui ir stabilumui, ypač Random Forest atveju. Random Forest modelio R^2 padidėjo nuo 0,7942 iki 0,8868 (+0,0926), o rezultatų dispersija sumažėjo daugiau nei tris kartus (standartinis nuokrypis nuo 0,7455 iki 0,2269). Tuo tarpu XGBoost modelis jau su baziniais parametrais pasiekė aukštą tikslumą (R^2 = 0,9083), o optimizavimas tik nežymiai pakeitė R^2 (-0,0037), tačiau sumažino R^2 rezultatų dispersiją apie 25 % (nuo 0,3110 iki 0,2342), kas patvirtina didesnį XGBoost atsparumą bazinių parametrų pasirinkimui ir leidžia jį laikyti tinkamiausiu sprendimu NUTS 2 regionų ekonominių rodiklių duomenų užpildymui.

Tolimesni darbai.

1. Užpildytas duomenų failas su prognozuotomis NUTS 2 regionų ekonominių rodiklių reikšmėmis gali būti toliau naudojamas regionų ekonomikos atsparumo prognozavimo instrumento sudedamųjų veiksnių svarbos analizei, taikant mašininio mokymosi algoritmus.
2. Tolesniuose tyrimuose numatoma plėsti taikomų mašininio mokymosi metodų spektrą ir integruoti daugiamates bei tarpdisciplinines veiksnių grupes, siekiant tiksliau modeliuoti regionų ekonomikos atsparumo pokyčius skirtingų krizių scenarijuose.
3. Remiantis gautais rezultatais, planuojama tobulinti regionų ekonomikos atsparumo prognozavimo instrumentą ir pritaikyti jį duomenimis grįstų rekomendacijų rengimui politikos formuotojams bei regionų plėtros institucijoms.

LITERATŪRA

[1] R. Damaševičius, „Regional Economic Development in the AI Era: Methods, Opportunities, and Challenges“, *Journal of Regional Economics*, t. 2, nr. 2, spal. 2023, doi: 10.58567/jre02020001.

[2] H. Zheng, J. Wu, R. Song, L. Guo, ir Z. Xu, „Predicting Financial Enterprise Stocks and Economic Data Trends Using Machine Learning Time Series Analysis“, liep. 2024, doi: 10.20944/PREPRINTS202407.0895.V1.

[3] Z. Liu, „Review on the influence of machine learning methods and data science on the economics“, *Applied and Computational Engineering*, t. 22, nr. 1, p. 137–141, spal. 2023, doi: 10.54254/2755-2721/22/20231208.

[4] R. Avacharmal, A. Balakrishnan, P. Ranjan, M. K. Vandanapu, S. Mulukuntla, ir P. Preethi, „Leveraging reinforcement learning for advanced financial planning for effective personalization in economic forecasting and savings strategies“, *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*, 2024, doi: 10.1109/ICCCNT61001.2024.10725494.

[5] R. Reznikov ir S. Turlakova, „DATA SCIENCE METHODS AND MODELS IN MODERN ECONOMY“, *Економічний простір*, nr. 191, p. 104–113, liep. 2024, doi: 10.32782/2224-6282/191-19.

[6] E. B. Tirkolaee, S. Sadeghi, F. M. Mooseloo, H. R. Vandchali, ir S. Aeini, „Application of Machine Learning in Supply Chain Management: A Comprehensive Overview of the Main Areas“, 2021 m., *Hindawi Limited*. doi: 10.1155/2021/1476043.

[7] S. Czech, „Evaluating the Role of Machine Learning in Economics: A Cutting-Edge Addition or Rhetorical Device?“, *Studies in Logic, Grammar and Rhetoric*, t. 68, nr. 1, p. 279–293, gruodž. 2023, doi: 10.2478/SLGR-2023-0014.

[8] N. Khashimova ir M. Buranova, „Comparative Analysis of Machine Learning Algorithms for Inflation Rate Classification and Economic Trend Forecasting“, *ACM International Conference Proceeding Series*, Association for Computing Machinery, gruodž. 2023, p. 274–282. doi: 10.1145/3644713.3644749.

[9] A. Aljohani, „Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility“, *Sustainability 2023, Vol. 15, Page 15088*, t. 15, nr. 20, p. 15088, spal. 2023, doi: 10.3390/SU152015088.

[10] R. J. A. Little ir D. B. Rubin, „Statistical analysis with missing data“, *Statistical Analysis with Missing Data*, p. 1–449, saus. 2019, doi: 10.1002/9781119482260.

[11] R. Sadegh, A. Mohameden, M. L. Salihi, ir M. F. Nanne, „A survey of missing data imputation techniques: statistical methods, machine learning models, and GAN-based approaches“, *IAES International Journal of Artificial Intelligence*, t. 14, nr. 4, p. 2876–2888, rugpj. 2025, doi: 10.11591/IJAI.V14.I4.PP2876-2888.

[12] T. Thomas ir E. Rajabi, „A systematic review of machine learning-based missing value imputation techniques“, *Data Technologies and Applications*, t. 55, nr. 4, p. 558–585, 2021, doi: 10.1108/DTA-12-2020-0298.

[13] J. L. Schafer ir J. W. Graham, „Missing data: Our view of the state of the art“, *Psychol Methods*, t. 7, nr. 2, p. 147–177, 2002, doi: 10.1037/1082-989X.7.2.147.

[14] „PRISMA 2020 flow diagram — PRISMA statement“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://www.prisma-statement.org/prisma-2020-flow-diagram

[15] „VOSviewer Online“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://app.vosviewer.com/

[16] I. Setiawan, R. Gernowo, ir B. Warsito, „A Systematic Literature Review On Missing Values: Research Trends, Datasets, Methods and Frameworks“, *E3S Web of Conferences*, t. 448, p. 02020, lapkr. 2023, doi: 10.1051/E3SCONF/202344802020.
[17] I. Silkauskaitė Supervisor ir A. Viktor Skorniakov, „Methods of multiple imputation“, 2025 m.
[18] C. Chairperson, „The Thesis of Abhijeet R Murthy is approved“.
[19] A. O. Melnyk ir I. V. Rozora, „Comparative analysis of imputation methods in machine learning models“, *Science Technologies Innovation*, nr. 3(35), p. 61, spal. 2025, doi: 10.35668/2520-6524-2025-3-07.
[20] M. K. Hasan, M. A. Alam, S. Roy, A. Dutta, M. T. Jawad, ir S. Das, „Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)“, *Inform Med Unlocked*, t. 27, p. 100799, saus. 2021, doi: 10.1016/J.IMU.2021.100799.
[21] D. Chehal, P. Gupta, P. Gulati, ir T. Gupta, „Comparative Study of Missing Value Imputation Techniques on E-Commerce Product Ratings“, *Informatica*, t. 47, nr. 3, p. 373–382, birž. 2023, doi: 10.31449/INF.V47I3.4156.
[22] A. Balusani, P. Jai Ram Chowdary, ir B. V. N. P. Paruchuri, „Enhancing Retail Demand Forecasting with Xgboost: A Comparative Study with Machine Learning and Deep Learning Models“, 2025, doi: 10.2139/SSRN.5274510.
[23] V. Shrivastava ir S. Shukla, „Impute-VSS: A comprehensive web-based visualization and simulation suite for comparative data imputation and statistical evaluation“, *SoftwareX*, t. 30, p. 102130, geg. 2025, doi: 10.1016/J.SOFTX.2025.102130.
[24] S. M. Memon, R. Wamala, ir I. H. Kabano, „A comparison of imputation methods for categorical data“, *Inform Med Unlocked*, t. 42, p. 101382, saus. 2023, doi: 10.1016/J.IMU.2023.101382.
[25] A. B. Ghimire *ir kt.*, „Impacts of Missing Data Imputation on Resilience Evaluation for Water Distribution System“, *Urban Science 2024, Vol. 8*, t. 8, nr. 4, spal. 2024, doi: 10.3390/URBANSCI8040177.
[26] D. Varma, C. S. Yajnik, A. Thorave, ir N. Sharma, „Handling Missing Data in Longitudinal Anthropometric Data Using Multiple Imputation Method“, *Lecture Notes in Networks and Systems*, t. 997 LNNS, p. 273–287, 2024, doi: 10.1007/978-981-97-3242-5_19.
[27] M. Hashemi ir H. A. Karimi, „Improved Mean Methods of Imputation“, *Statistics, Optimization & Information Computing*, t. 6, nr. 4, p. 526–535, lapkr. 2018, doi: 10.19139/SOIC.V6I4.281.
[28] Z. Zhang, „Missing data imputation: focusing on single imputation“, *Ann Transl Med*, t. 4, nr. 1, p. 9–9, saus. 2016, doi: 10.3978/J.ISSN.2305-5839.2015.12.38.
[29] A. Jadhav, D. Pramod, ir K. Ramanathan, „Comparison of Performance of Data Imputation Methods for Numeric Dataset“, *Applied Artificial Intelligence*, t. 33, nr. 10, p. 913–933, rugpj. 2019, doi: 10.1080/08839514.2019.1637138;PAGE:STRING:ARTICLE/CHAPTER.
[30] M. Messaoudi, „Random Forest Modeling For Forecasting Economic Growth:, A Machine Learning Approach In The Case Of Algeria ,1964-2022‘“, p. 391, 2024, doi: 10.36539/1427-014-001-019.
[31] J. He, „Analysis and Forecasting of Energy Data and GDP Based on Pearson Correlation Analysis-Random Forest Algorithm“, *Advances in Economics, Management and Political Sciences*, t. 80, nr. 1, p. 209–215, geg. 2024, doi: 10.54254/2754-1169/80/20241835.
[32] M. Ibrahim, „Evolution of Random Forest from Decision Tree and Bagging: A Bias-Variance Perspective“, *Dhaka University Journal of Applied Science and Engineering* , t. 7, nr. 1, p. 66–71, vas. 2022, doi: 10.3329/DUJASE.V7I1.62888.
[33] B. Liu, R. Mazumder, ir F. D’alche-Buc, „Randomization Can Reduce Both Bias and Variance: A Case Study in Random Forests“, *Journal of Machine Learning Research*, t. 26, p. 1–49, 2025, Žiūrėta: 2025 m. rugpjūčio 21 d. [Interaktyvus]. Adresas: http://jmlr.org/papers/v26/24-0255.html.

[34] B. O. Agadagba, „Predicting real estate prices with machine learning“, *Vilniaus universitetas*, 2025.
[35] M. M. Islam, A. Jannat, K. Aruga, ir M. M. Rashid, „Forecasting Foreign Direct Investment Inflow to Bangladesh: Using an Autoregressive Integrated Moving Average and a Machine Learning-Based Random Forest Approach“, *Journal of Risk and Financial Management 2024, Vol. 17, Page 451*, t. 17, nr. 10, p. 451, spal. 2024, doi: 10.3390/JRFM17100451.
[36] T. Chen ir C. Guestrin, „XGBoost: A scalable tree boosting system“, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, t. 13-17-August-2016, p. 785–794, rugpj. 2016, doi: 10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.
[37] J. H. Friedman, „Greedy function approximation: A gradient boosting machine.“, *https://doi.org/10.1214/aos/1013203451*, t. 29, nr. 5, p. 1189–1232, spal. 2001, doi: 10.1214/AOS/1013203451.
[38] J. Snoek, H. Larochelle, ir R. P. Adams, „Practical Bayesian Optimization of Machine Learning Algorithms“, *Adv Neural Inf Process Syst*, t. 4, p. 2951–2959, birž. 2012, Žiūrėta: 2025 m. rugsėjo 16 d. [Interaktyvus]. Adresas: https://arxiv.org/pdf/1206.2944
[39] K. Oono ir T. Suzuki, „Optimization and Generalization Analysis of Transduction through Gradient Boosting and Application to Multi-scale Graph Neural Networks“, Žiūrėta: 2025 m. rugsėjo 16 d. [Interaktyvus]. Adresas: https://github.com/delta2323/GB-GNN
[40] H. Jallow, A. Gibba, R. W. Mwangi, ir H. Imboga, „GDP prediction of The Gambia using generative adversarial networks“, *Front Artif Intell*, t. 8, p. 1546398, kovo 2025, doi: 10.3389/FRAI.2025.1546398/BIBTEX.
[41] R. Mitchell, A. Adinets, T. Rao, ir E. Frank, „XGBoost: Scalable GPU Accelerated Learning“, 2018, Žiūrėta: 2025 m. rugsėjo 16 d. [Interaktyvus]. Adresas: https://github.com/dmlc/xgboost
[42] „Feature importance — xgb.importance • xgboost“. Žiūrėta: 2025 m. lapkričio 11 d. [Interaktyvus]. Adresas: https://xgboost.readthedocs.io/en/latest/r_docs/R-package/docs/reference/xgb.importance.html
[43] „Feature Importance and Feature Selection With XGBoost in Python - MachineLearningMastery.com“. Žiūrėta: 2025 m. lapkričio 11 d. [Interaktyvus]. Adresas: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/
[44] L. Beretta ir A. Santaniello, „Nearest neighbor imputation algorithms: a critical evaluation“, *BMC Medical Informatics and Decision Making 2016 16:3*, t. 16, nr. 3, p. 197–208, liep. 2016, doi: 10.1186/S12911-016-0318-Z.
[45] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, ir O. Tabona, „A survey on missing data in machine learning“, *Journal of Big Data 2021 8:1*, t. 8, nr. 1, p. 1–37, spal. 2021, doi: 10.1186/S40537-021-00516-9.
[46] S. van Buuren ir K. Groothuis-Oudshoorn, „mice: Multivariate imputation by chained equations in R“, *J Stat Softw*, t. 45, nr. 3, p. 1–67, gruodž. 2011, doi: 10.18637/JSS.V045.I03.
[47] J. S. Murray, „Submitted to Statistical Science Multiple Imputation: A Review of Practical and Theoretical Findings“.
[48] M. J. Azur, E. A. Stuart, C. Frangakis, ir P. J. Leaf, „Multiple imputation by chained equations: what is it and how does it work?“, *Int J Methods Psychiatr Res*, t. 20, nr. 1, p. 40–49, kovo 2011, doi: 10.1002/MPR.329.
[49] F. Pedregosa *ir kt.*, „Scikit-learn: Machine Learning in Python“, *Journal of Machine Learning Research*, t. 12, p. 2825–2830, saus. 2012, Žiūrėta: 2025 m. lapkričio 8 d. [Interaktyvus]. Adresas: https://arxiv.org/pdf/1201.0490
[50] D. Bertsimas, A. Delarue, ir J. Pauphilet, „Simple Imputation Rules for Prediction with Missing Data: Contrasting Theoretical Guarantees with Empirical Performance“.

[51] M. Alwateer, E.-S. Atlam, M. Mohammed, A. El-Raouf, O. A. Ghoneim, ir I. Gad, „Missing Data Imputation: A Comprehensive Review“, *Journal of Computer and Communications*, t. 12, p. 53–75, 2024, doi: 10.4236/jcc.2024.1211004.
[52] „Regions in the European Union Nomenclature of territorial units for statistics (NUTS) – 2024 edition - Manuals and guidelines - Eurostat“. Žiūrėta: 2025 m. spalio 20 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/w/ks-gq-23-010
[53] „Statistics | Eurostat“. Žiūrėta: 2025 m. lapkričio 22 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/explore/all/all_themes
[54] „Reglamentas - 549/2013 - EN - EUR-Lex“. Žiūrėta: 2025 m. gruodžio 10 d. [Interaktyvus]. Adresas: https://eur-lex.europa.eu/legal-content/LT/ALL/?uri=celex:32013R0549
[55] „Reglamentas - 2016/2066 - EN - EUR-Lex“. Žiūrėta: 2025 m. gruodžio 10 d. [Interaktyvus]. Adresas: https://eur-lex.europa.eu/legal-content/LT/ALL/?uri=CELEX:32016R2066
[56] „Regions in the European Union“, doi: 10.2785/573744.
[57] „Overview - NUTS - Nomenclature of territorial units for statistics - Eurostat“. Žiūrėta: 2025 m. spalio 20 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/web/nuts
[58] „Cohesion policy indicators - Statistics Explained - Eurostat“. Žiūrėta: 2025 m. spalio 20 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cohesion_policy_indicators
[59] „Eurostat - Statistical Atlas“. Žiūrėta: 2025 m. spalio 20 d. [Interaktyvus]. Adresas: https://ec.europa.eu/statistical-atlas/viewer/?config=typologies.json&mids=BKGCNT,NUTS2024L2,CNTOVL&o=1,1,0.7&ch=NUTS&center=51.39817,28.04278,4&
[60] „Common classification of territorial units for statistics (NUTS) | Fact Sheets on the European Union | European Parliament“. Žiūrėta: 2025 m. spalio 20 d. [Interaktyvus]. Adresas: https://www.europarl.europa.eu/factsheets/en/sheet/99/common-classification-of-territorial-units-for-statistics-nuts-
[61] A. Ariu, „Crisis-proof services: Why trade in services did not suffer during the 2008-2009 collapse“, *J Int Econ*, t. 98, p. 138–149, saus. 2016, doi: 10.1016/J.JINTECO.2015.09.002.
[62] J. C. Brada, P. Gajewski, ir A. M. Kutan, „Economic resiliency and recovery, lessons from the financial crisis for the COVID-19 pandemic: A regional perspective from Central and Eastern Europe“, *International Review of Financial Analysis*, t. 74, kovo 2021, doi: 10.1016/J.IRFA.2021.101658.
[63] „European system of accounts. ESA 2010“, doi: 10.2785/16644.
[64] „[lfst_r_lfe2emp] Employed persons by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/lfst_r_lfe2emp/default/table?lang=en
[65] „Regional economic accounts (reg_eco10)“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/cache/metadata/en/reg_eco10_esms.htm#shortdata_descrDisseminated
[66] „[bd_size_r3] Business demography by size class and NUTS 3 region (2008-2020)“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/bd_size_r3__custom_18676118/default/table
[67] „[reg_area3] Area by NUTS 3 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/reg_area3__custom_18706973/default/table
[68] „[demo_r_d3area] Area by NUTS 3 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/demo_r_d3area__custom_18706952/default/table

[69] „[tgs00109] Persons aged 25-64 with tertiary educational attainment level by sex and NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tgs00109__custom_18420806/default/table
[70] „[trng_lfse_04] Participation rate in education and training (last 4 weeks) by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/trng_lfse_04__custom_18420942/default/table
[71] „RPDTN - Y25-64-TOTAL-PC-ED5-8 | Ardeco Viewer“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://territorial.ec.europa.eu/ardeco/viewer/258593?jdvfys=asc&jdvfc=eu&jdvfnl=2%2C3&acc=nutsLabel%2CvariableDimensionAge%2CvariableDimensionSex%2CvariableDimensionUnit
[72] „RPDTN - Y25-64-TOTAL-PC-ED5-8 | Ardeco Viewer“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://territorial.ec.europa.eu/ardeco/viewer/258593?jdvfys=asc&jdvfc=eu&jdvfnl=2%2C3&acc=nutsLabel%2CvariableDimensionAge%2CvariableDimensionSex%2CvariableDimensionUnit
[73] „RPDTN - Y25-64-TOTAL-PC-ED5-8 | Ardeco Viewer“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://territorial.ec.europa.eu/ardeco/viewer/258593?jdvfys=asc&jdvfc=eu&jdvfnl=2%2C3&acc=nutsLabel%2CvariableDimensionAge%2CvariableDimensionSex%2CvariableDimensionUnit
[74] „[hrst_st_rcat] Human resources in science and technology (HRST) by category and NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/hrst_st_rcat__custom_18515496/default/table
[75] „[edat_lfse_16] Early leavers from education and training by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/edat_lfse_16__custom_18708210/default/table
[76] „[tour_occ_nin2] Nights spent at tourist accommodation establishments by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tour_occ_nin2__custom_18516195/default/table
[77] „[tour_occ_nin2] Nights spent at tourist accommodation establishments by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tour_occ_nin2__custom_18516195/default/table
[78] „[tour_occ_nin2] Nights spent at tourist accommodation establishments by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tour_occ_nin2__custom_18516195/default/table
[79] „[tran_r_avpa_nm] Air transport of passengers by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tran_r_avpa_nm__custom_18516746/default/table
[80] „[tran_r_avgo_nm] Air transport of freight by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tran_r_avgo_nm/default/table?lang=en
[81] „[tran_r_net] Road, rail and navigable inland waterways networks by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tran_r_net__custom_18710531/default/table

[82] „[tran_r_net] Road, rail and navigable inland waterways networks by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tran_r_net__custom_18710531/default/table
[83] „[tran_r_mapa_nm] Maritime transport of passengers by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tran_r_mapa_nm__custom_18516843/default/table
[84] „[tgs00047] Households that have internet access at home by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tgs00047__custom_18517336/default/table
[85] „[tgs00050] Individuals regularly using the internet by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tgs00050/default/table?lang=en
[86] „[tgs00042] Intramural R&D expenditure (GERD) by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/tgs00042__custom_18518342/default/table
[87] N. Charron, V. Lapuente, ir M. Bauhr, „THE EUROPEAN QUALITY OF GOVERNMENT INDEX 2024“, Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: http://www.qog.pol.gu.se
[88] „NACE Rev. 2 - Statistical classification of economic activities - Products Manuals and Guidelines - Eurostat“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-ra-07-015
[89] „[htec_emp_reg2] Employed persons in technology and knowledge-intensive sectors by NACE Rev. 2 activity and NUTS 2 region (2008-2026)“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/htec_emp_reg2__custom_18518787/default/table
[90] „[htec_emp_reg] Employed persons in technology and knowledge-intensive sectors by NACE Rev. 1.1 activity and NUTS 2 region (1999-2008)“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/htec_emp_reg__custom_18518917/default/table
[91] „[lfst_r_lfe2emprt] Employment rates by NUTS 2 region“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/databrowser/view/lfst_r_lfe2emprt__custom_18676448/default/table
[92] M. Tumminello, F. Lillo, ir R. N. Mantegna, „Correlation, hierarchies, and networks in financial markets“, 2008.
[93] „Ekonometrika“. Žiūrėta: 2025 m. gruodžio 13 d. [Interaktyvus]. Adresas: http://www.ilab.lt/stabingiene/sk2_1.html
[94] R. Metulini, G. Gnecco, F. Biancalani, ir M. Riccaboni, „Hierarchical clustering and matrix completion for the reconstruction of world input-output tables“, 123po Kr., doi: 10.1007/s10182-022-00448-6.
[95] R. J. A. Little ir D. B. Rubin, „Statistical analysis with missing data“, *Statistical Analysis with Missing Data*, p. 1–449, saus. 2019, doi: 10.1002/9781119482260.
[96] R. J. A. Little, „A Test of Missing Completely at Random for Multivariate Data with Missing Values“, *J Am Stat Assoc*, t. 83, nr. 404, p. 1198, gruodž. 1988, doi: 10.2307/2290157.
[97] C. Li, „Little’s test of missing completely at random“, *Stata Journal*, t. 13, nr. 4, p. 795–809, gruodž. 2013, doi: 10.1177/1536867X1301300407;WEBSITE:WEBSITE:SAGE;ISSUE:ISSUE:DOI.
[98] „(PDF) Recommended practices for editing and imputation in cross-sectional business surveys“. Žiūrėta: 2025 m. lapkričio 25 d. [Interaktyvus]. Adresas:

https://www.researchgate.net/publication/254974672_Recommended_practices_for_editing_and_imputation_in_cross-sectional_business_surveys
[99] „Our history - Eurostat“. Žiūrėta: 2025 m. lapkričio 25 d. [Interaktyvus]. Adresas: https://ec.europa.eu/eurostat/about-us/history
[100] B. Liu, R. Mazumder, ir F. D’alche-Buc, „Randomization Can Reduce Both Bias and Variance: A Case Study in Random Forests“, *Journal of Machine Learning Research*, t. 26, nr. 150, p. 1–49, 2025, Žiūrėta: 2025 m. lapkričio 29 d. [Interaktyvus]. Adresas: http://jmlr.org/papers/v26/24-0255.html
[101] D. J. Stekhoven ir P. Bühlmann, „MissForest—non-parametric missing value imputation for mixed-type data“, *Bioinformatics*, t. 28, nr. 1, p. 112–118, saus. 2012, doi: 10.1093/BIOINFORMATICS/BTR597.
[102] Y. Liu ir A. C. Constantinou, „IMPROVING THE IMPUTATION OF MISSING DATA WITH MARKOV BLANKET DISCOVERY“.
[103] K. Grzesiak, C. Muller, J. Josse, ir J. Näf, „Do we Need Dozens of Methods for Real World Missing Value Imputation?“, lapkr. 2025, Žiūrėta: 2025 m. gruodžio 2 d. [Interaktyvus]. Adresas: https://arxiv.org/pdf/2511.04833v1
[104] B. W. Silverman, „Density estimation: For statistics and data analysis“, *Density Estimation: For Statistics and Data Analysis*, p. 1–175, saus. 2018, doi: 10.1201/9781315140919/DENSITY-ESTIMATION-STATISTICS-DATA-ANALYSIS-BERNARD-SILVERMAN/RIGHTS-AND-PERMISSIONS.
[105] F. Pedregosa *ir kt.*, „Scikit-learn: Machine Learning in Python“, *Journal of Machine Learning Research*, t. 12, p. 2825–2830, saus. 2012, Žiūrėta: 2025 m. gruodžio 2 d. [Interaktyvus]. Adresas: https://arxiv.org/pdf/1201.0490
[106] T. M. Lukusa, S.-M. Lee, ir C.-S. Li, „Review of Zero-Inflated Models with Missing Data“, 2017, doi: 10.3844/amjbsp.2017.1.12.
[107] M. Platzer ir T. Reutterer, „Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data“, *Front Big Data*, t. 4, p. 679939, birž. 2021, doi: 10.3389/FDATA.2021.679939/BIBTEX.
[108] D. Bertsimas, A. Delarue, ir J. Pauphilet, „Simple Imputation Rules for Prediction with Missing Data: Contrasting Theoretical Guarantees with Empirical Performance“.
[109] D. Bertsimas, C. Pawlowski, ir Y. D. Zhuo, „From Predictive Methods to Missing Data Imputation: An Optimization Approach“, *Journal of Machine Learning Research*, t. 18, p. 1–39, 2018, Žiūrėta: 2025 m. gruodžio 2 d. [Interaktyvus]. Adresas: http://jmlr.org/papers/v18/17-073.html.
[110] S. Sangari ir H. E. Ray, „Evaluation of imputation techniques with varying percentage of missing data“, rugs. 2021, Žiūrėta: 2025 m. gruodžio 2 d. [Interaktyvus]. Adresas: https://arxiv.org/pdf/2109.04227
[111] „How to normalize the RMSE“. Žiūrėta: 2025 m. gruodžio 2 d. [Interaktyvus]. Adresas: https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse
[112] C. John, E. J. Ekpenyong, ir C. C. Nworu, „Imputation of Missing Values in Economic and Financial Time Series Data Using Five Principal Component Analysis Approaches“, *CBN Journal of Applied Statistics*, t. 10, nr. 1, p. 51–73, 2019, doi: 10.33429/Cjas.10119.3/6.
[113] C. Li, X. Ren, ir G. Zhao, „Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data“, *Algorithms 2023, Vol. 16, Page 422*, t. 16, nr. 9, p. 422, rugs. 2023, doi: 10.3390/A16090422.
[114] Y. Wang, „Analysis of Feature Importance Based on Random Forest for Stock Selection“, 2025, doi: 10.5220/0013233500004568.
[115] C. D. Brummitt, A. Gómez-Liévano, R. Hausmann, ir M. H. Bonds, „Machine-learned patterns suggest that diversification drives economic development“, *J R Soc Interface*, t. 17, nr. 162, gruodž. 2018, doi: 10.1098/rsif.2019.0283.

[116] L. Mentch ir S. Zhou, „Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success“, *Journal of Machine Learning Research*, t. 21, p. 1–36, lapkr. 2019, Žiūrėta: 2025 m. gruodžio 5 d. [Interaktyvus]. Adresas: https://arxiv.org/pdf/1911.00190
[117] A. Agarwal, Y. S. Tan, O. Ronen, C. Singh, ir B. Yu, „Hierarchical Shrinkage: improving the accuracy and interpretability of tree-based methods“, *Proc Mach Learn Res*, t. 162, p. 111–135, vas. 2022, Žiūrėta: 2025 m. gruodžio 6 d. [Interaktyvus]. Adresas: https://arxiv.org/pdf/2202.00858
[118] J. H. Friedman, „Greedy function approximation: A gradient boosting machine“, *Ann Stat*, t. 29, nr. 5, p. 1189–1232, 2001, doi: 10.1214/AOS/1013203451.
[119] F. Aracri, M. G. Bianco, A. Quattrone, ir A. Sarica, „Bridging the Gap: Missing Data Imputation Methods and Their Effect on Dementia Classification Performance“, *Brain Sciences 2025, Vol. 15, Page 639*, t. 15, nr. 6, p. 639, birž. 2025, doi: 10.3390/BRAINSCI15060639.
[120] „RandomizedSearchCV — scikit-learn 1.7.2 documentation“. Žiūrėta: 2025 m. gruodžio 7 d. [Interaktyvus]. Adresas: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
[121] „Data Analysis Using Regression and Multilevel/Hierarchical Models | Cambridge University Press & Assessment“. Žiūrėta: 2025 m. gruodžio 8 d. [Interaktyvus]. Adresas: https://www.cambridge.org/lt/universitypress/subjects/statistics-probability/statistical-theory-and-methods/data-analysis-using-regression-and-multilevelhierarchical-models?format=HB&isbn=9780521867061
[122] H. Wu, C. Wang, ir Z. Wu, „A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data“, *Biostatistics*, t. 14, nr. 2, p. 232–243, 2013, doi: 10.1093/biostatistics/kxs033.
[123] J. Willwerscheid, P. Carbonetto, ir M. Stephens, „ebnm: An R Package for Solving the Empirical Bayes Normal Means Problem Using a Variety of Prior Families“, *J Stat Softw*, t. 114, nr. 3, 2025, doi: 10.18637/JSS.V114.I03.
[124] G. Datta ir M. Ghosh, „Small Area Shrinkage Estimation“, *Statistical Science*, t. 27, nr. 1, p. 95–114, 2012, doi: 10.1214/11-STS374.
[125] C. Profssa Rosaria Romano Dott Luca Attolico Chiarmo Jamus Jerome Lim COORDINATORE Chiarma Profssa Margherita Scoppola, „Opening the Black Box: Nowcasting Singapore’s GDP Growth and its Explainability“, gruodž. 2025, Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://arxiv.org/pdf/2512.02092v1
[126] „Jupyter Notebooks in VS Code“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://code.visualstudio.com/docs/datascience/jupyter-notebooks
[127] „smtplib — SMTP protocol client — Python 3.14.2 documentation“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://docs.python.org/3/library/smtplib.html
[128] „email: Examples — Python 3.14.2 documentation“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://docs.python.org/3/library/email.examples.html
[129] „openpyxl · PyPI“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://pypi.org/project/openpyxl/
[130] V. Teodorescu ir L. Obreja Brașoveanu, „Assessing the Validity of k-Fold Cross-Validation for Model Selection: Evidence from Bankruptcy Prediction Using Random Forest and XGBoost“, *Computation 2025, Vol. 13, Page 127*, t. 13, nr. 5, p. 127, geg. 2025, doi: 10.3390/COMPUTATION13050127.
[131] D. Saha, A. Aggarwal, ir S. Hans, „Data Synthesis for Testing Black-Box Machine Learning Models; Data Synthesis for Testing Black-Box Machine Learning Models“.
[132] F. Maleki, K. Ovens, R. Gupta, C. Reinhold, A. Spatz, ir R. Forghani, „Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls“, *Radiol Artif Intell*, t. 5, nr. 1, saus. 2023, doi: 10.1148/RYAI.220028.

[133] S. Du, Y. Xu, ir L. Wang, „Predicting economic resilience: A machine learning approach to rural development“, *Alexandria Engineering Journal*, t. 121, p. 193–200, geg. 2025, doi: 10.1016/J.AEJ.2025.02.049.
[134] W. McKinney, „Data Structures for Statistical Computing in Python“, *SciPy 2010*, p. 56–61, 2010, doi: 10.25080/MAJORA-92BF1922-00A.
[135] „Welcome to Flask — Flask Documentation (3.1.x)“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://flask.palletsprojects.com/en/stable/
[136] „mysql-connector-python · PyPI“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://pypi.org/project/mysql-connector-python/
[137] „reportlab · PyPI“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://pypi.org/project/reportlab/
[138] Y. Jani, „Implementing Continuous Integration and Continuous Deployment (CI/CD) in Modern Software Development“, *International Journal of Science and Research (IJSR)*, t. 12, nr. 6, p. 2984–2987, birž. 2023, doi: 10.21275/SR24716120535.
[139] „(PDF) Git Branching and Release Strategies“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://www.researchgate.net/publication/388531640_Git_Branching_and_Release_Strategies
[140] „Deploying on Render – Render Docs“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://render.com/docs/deploys
[141] „Zero-Ops Backend Hosting for Web Apps | Render“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://render.com/articles/zero-ops-backend-hosting-for-web-apps
[142] „Regions – Render Docs“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://render.com/docs/regions
[143] T. Visockytė, „Debesų paslaugų programavimo sprendimų tyrimas naudojant Google App Engine aplinką“, *Vilniaus universitetas*, 2025.
[144] „Render Postgres Recovery and Backups – Render Docs“. Žiūrėta: 2025 m. gruodžio 9 d. [Interaktyvus]. Adresas: https://render.com/docs/postgresql-backups
[145] G. Huang, „Missing data filling method based on linear interpolation and lightgbm“, *J Phys Conf Ser*, t. 1754, nr. 1, vas. 2021, doi: 10.1088/1742-6596/1754/1/012187.
[146] M. Muth, M. Lingenfelder, ir G. Nufer, „The application of machine learning for demand prediction under macroeconomic volatility: a systematic literature review“, *Management Review Quarterly 2024*, p. 1–44, birž. 2024, doi: 10.1007/S11301-024-00447-8.
[147] S. Meisenbacher *ir kt.*, „Review of automated time series forecasting pipelines“, *Wiley Interdiscip Rev Data Min Knowl Discov*, t. 12, nr. 6, vas. 2022, doi: 10.1002/widm.1475.
[148] A. Al-Azzawi, F. T. Mora, C. Lim, ir Y. Shang, „An Artificial Intelligent Methodology-based Bayesian Belief Networks Constructing for Big Data Economic Indicators Prediction“, *International Journal of Advanced Computer Science and Applications*, t. 14, nr. 5, p. 829–838, 2023, doi: 10.14569/IJACSA.2023.0140588.
[149] R. Seifipour ir A. Mehrabian, „Application of Artificial Neural Networks in Economic and Financial Sciences“, *Research and Applications of Digital Signal Processing [Working Title]*, saus. 2025, doi: 10.5772/INTECHOPEN.1007604.
[150] N. Charron, V. Lapuente, ir M. Bauhr, „THE EUROPEAN QUALITY OF GOVERNMENT INDEX 2024“, Žiūrėta: 2025 m. lapkričio 22 d. [Interaktyvus]. Adresas: http://www.qog.pol.gu.se
[151] „Claude“. Žiūrėta: 2025 m. gruodžio 14 d. [Interaktyvus]. Adresas: https://claude.ai/new

PRIEDAI

1 priedas. Literatūros analizės protokolas

Tema	Mašininio mokymosi modelių taikymas prognozuojant regionų ekonominius rodiklius NUTS 2 lygmeniu.
Mokslinė problema	NUTS 2 regionų ekonominių rodiklių duomenų rinkiniuose sistemingai pasitaikančios trūkstamos reikšmės riboja patikimą regionų ekonominės būklės, atsparumo ir prognozių analizę. Trūksta empirinių tyrimų, kurie sistemingai vertintų pažangių mašininio mokymosi metodų veiksmingumą užpildant trūkstamas reikšmes skirtinguose trūkstamumo scenarijuose.
Raktiniai žodžiai (EN)	Systematic literature review, machine learning, missing data, imputation, economic indicators, regional analysis, forecasting, Random Forest, XGBoost, KNN, MICE, NUTS 2 regions
Raktiniai žodžiai (LT)	Sisteminė literatūros analizė, mašininis mokymasis, trūkstami duomenys, imputacija, ekonominiai rodikliai, regioninė analizė, prognozavimas, Random Forest, XGBoost, MICE, NUTS 2 regionai.
Paieškos užklausos elementai	*Žr. 2 priedo 9 lentelę* (mokslinių duomenų bazių paieškos užklausų rezultatai)
Apribojimai	• Publikacijos laikotarpis: 2015 - 2025; • Kalba: anglų, lietuvių; • Prieiga: atviros prieigos straipsniai; • Mokslinė sritis: informatika, duomenų mokslas, mašininis mokymasis, ekonometrika; • Dokumentų tipai: moksliniai straipsniai, konferencijų medžiaga.
Straipsnių atrankos priėmimo kriterijai (IC)	1. Tyrimai, nagrinėjantys trūkstamų reikšmių užpildymą naudojant mašininio mokymosi metodus; 2. Straipsniai, kuriuose vertinamas trūkstamų reikšmių užpildymo metodų poveikis prognozavimo tikslumui (RMSE, MAE, R^2 ir kt.); 3. Tyrimai, taikomi regioniniams, ekonominiams arba laiko eilučių duomenims; 4. Publikacijos, kuriose aiškiai aprašyta tyrimo metodologija.

Šaltinis: sudaryta darbo autoriaus.

2 priedas. Mokslinių duomenų bazių paieškos užklausų rezultatai pagal PRISMA metodologiją

Nr.	Duomenų bazė	Paieškos užklausa	Rastų rezultatų skaičius	Atrinktos publikacijos (pirminė atranka)	Pilna paieškos užklausos nuoroda
1	Webofscience.com	("NUTS 2 region*" or "economic* indicator*" or "economic* resilienc*") and (AI or "Artific* intellige*" or "neural network*" or "machin* learn*")	529	25	https://www.webofscience.com/wos/woscc/summary/f48819fd-50d8-4527-84b6-0307eaf2d3b7-01454939f1/relevance/1
2	Webofscience.com	'literature review' AND 'Machine learning' AND 'time series forecasting' AND 'economic'	13	2	https://www.webofscience.com/wos/woscc/summary/913ddb17-9d87-47e8-b00a-14184ad0314d-01454b7500/date-descending/1
3	Webofscience.com	("NUTS 2 region*" or "economic* indicator*" or "economic* resilienc*") and (AI or "Artific* intellige*" or "neural network*" or "machin* learn*") and ("Systematic" or "literature")	43	2	https://www.webofscience.com/wos/woscc/summary/047e5829-d0a2-4658-8cee-154b6f88e384-01454cb91e/date-descending/2
4	Researchgate.net	("NUTS 2 region*" or "economic* indicator*" or "economic* resilienc*") and (AI or "Artific* intellige*" or "neural network*" or "machin* learn*")	9851	16	https://www.researchgate.net/search.Search.html?query=%28%22NUTS+2+region*%22+or+%22economic*+indicator*%22+or+%22economic*+resilienc*%22%29+and+%28AI+or+%22Artific*+intellige*%22+or+%22neural+network*%22+or+%22machin*+learn*%22+%29+&type=publication
5	Webofscience.com	("machin* learn*") and ("imputation")	161	9	https://www.webofscience.com/wos/woscc/summary/23d68136-6c10-48cf-90ae-7a32fa38c2ec-018533a4a4/relevance/1

Šaltinis: sudaryta darbo autoriaus. Priedo tęsinys 97 puslapyje.

2 priedo tęsinys. Mokslinių duomenų bazių paieškos užklausų rezultatai pagal PRISMA metodologiją

Nr.	Duomenų bazė	Paieškos užklausa	Rastų rezultatų skaičius	Atrinktos publikacijos (pirminė atranka)	Pilna paieškos užklausos nuoroda
6	Researchgate.net	("missing*" or "dat* im*" or "imputation") and ("Comprehensive") and ("review")	7808	10	https://www.researchgate.net/search.Search.html?query=%28%22missing*%22+or+%22dat*+im*%22+or+%22imputation%22%29+and+%28%22Comprehensive%22%29+and+%28%22review%22%29&type=publication
7	Researchgate.net	("missing*" or "dat* im*" or "imputation") and ("Comprehensive") and ("review") and ("machine learning")	1278	6	https://www.researchgate.net/search.Search.html?query=%28%22missing*%22+or+%22dat*+im*%22+or+%22imputation%22%29+and+%28%22Comprehensive%22%29+and+%28%22review%22%29+and+%28%22machine+learning%22%29&type=publication

3 priedas. Į sisteminę literatūros analizę įtrauktų straipsnių apžvalga

Užklausa	Straipsnio pavadinimas (nuoroda)	Analizuojami duomenų rinkiniai (*angl. datasets*)	Pagrindinis tiriamas klausimas / problema	Naudojamas metodas	Tyrimo dalykinė sritis	Prognozavimo algoritmai ir metodai	Algoritmų efektyvumo vertinimas	Ekonominių rodiklių tipai	Rezultatai
1	The application of machine learning for demand prediction under macroeconomic volatility: a systematic literature review[146]	64 straipsnių apžvalga, apimanti įvairius makroekonominius duomenų rinkinius, įskaitant laiko eilučių duomenis ir rinkos nepastovumo rodiklius.	Kaip efektyviai pritaikyti mašininio mokymosi modelius prognozuoti paklausos pokyčius makroekonominio nepastovumo sąlygomis.	Sisteminis literatūros apžvalgos metodas, pagrįstas analizės etapais: paieška, atranka, kokybės vertinimas, sintezė.	Ekonomika, rinkos analizė, prognozavimas.	LSTM, Random Forest, Gradient Boosting Trees.	RMSE, MAE, tikslumo analizė.	Makroekonominis nepastovumas, infliacija, vartotojų paklausos pokyčiai.	ML modeliai gali tiksliau prognozuoti ekonominį nepastovumą, bet deriniai su kitais metodais gali būti dar veiksmingesni.
2	Review of automated time series forecasting pipelines [147]	Laiko eilučių duomenys iš įvairių sričių, pvz., energetikos, ekonomikos.	Kaip automatizuoti laiko eilučių prognozavimo procesą ?	Automatizuota laiko eilučių prognozavimo metodo struktūra, įskaitant duomenų apdorojimą, hiperparametrų optimizavimą,modelio pasirinkimą ir prognozių derinimą.	Laiko eilučių prognozavimas.	ARIMA, SARIMA, LSTM, GRU, Random Forest, Gradient Boosting Machines, Hibridiniai modeliai.	RMSE, MAPE, MAE, ir kitų klaidų vertinimo rodikliai naudojami modelių efektyvumui palyginti.	Ekonominiai rodikliai, energijos suvartojimas.	Tyrimai turėtų būti orientuoti į visapusišką prognozavimo sistemų automatizavimą,integruojant visas procesų sekos sudedamąsias dalis - nuo duomenų apdorojimo ir modelio parinkimo iki rezultatų interpretavimo.

Šaltinis: sudaryta darbo autoriaus. Priedo tęsinys 99 puslapyje.

3 priedo tęsinys. Į sisteminę literatūros analizę įtrauktų straipsnių apžvalga

Užklausa	**Straipsnio pavadinimas (nuoroda)**	**Analizuojami duomenų rinkiniai (*angl. datasets*)**	**Pagrindinis tiriamas klausimas / problema**	**Naudojamas metodas**	**Tyrimo dalykinė sritis**	**Prognozavimo algoritmai ir metodai**	**Algoritmų efektyvumo vertinimas**	**Ekonominių rodiklių tipai**	**Rezultatai**
3	An Artificial Intelligent Methodology-based Bayesian Belief Networks Constructing for Big Data Economic Indicators Prediction [148]	Pasaulinio banko duomenų rinkinys: BVP vienam gyventojui (PPP), BVP augimas, pramonės pridėtinė vertė, mokslo straipsniai, viešasis vartojimas.	Kaip Bajeso tinklai gali būti naudojami prognozuojant ekonominius rodiklius iš didelių duomenų rinkinių?	Bajeso tinklų konstravimo metodologija, apimanti priklausomybių nustatymą tarp kintamųjų, remiantis domeno žiniomis ir kvadratinės klaidos minimizavimo principais.	Ekonominiai rodikliai, didieji duomenys.	Bajeso tinklai, regresijos modeliai (STE).	-	BVP vienam gyventojui, BVP augimas.	Bajeso tinklo pagrindu sukurti modeliai pasižymi aukštesniu tikslumu nei kiti to tipo ML modeliai, tačiau kai kuriems rodikliams reikia daugiau duomenų.
4	Application of Artificial Neural Networks in Economic and Financial Sciences [149]	Įvairūs ekonominiai rodikliai, įskaitant BVP, infliaciją, akcijų kainas, valiutų kursus ir aplinkosaugos rodiklius.	Kaip dirbtiniai neuroniniai tinklai gali būti pritaikyti prognozuojant ekonominius ir finansinius rodiklius kompleksiškose situacijose.	Dirbtiniai neuroniniai tinklai (ANN), įskaitant daugiasluoksnius perceptroninius tinklus (MLP) ir pasikartojančius neuroninius tinklus (RNN).	Ekonomika, finansų mokslai.	ANN, MLP, RNN, ANFIS	Tikslumas buvo aukštas prognozuojant valiutų kursus, akcijų kainų indeksą ir bankrotą, tačiau buvo pastebėta tinklo struktūros parinkimo sudėtingumas.	BVP, valiutų kursai, akcijų kainos, aplinkosaugos rodikliai.	Neuroniniai tinklai suteikė reikšmingai tikslesnes prognozes nei tradiciniai metodai.

Šaltinis: sudaryta darbo autoriaus. Priedo tęsinys 100 puslapyje.

3 priedo tęsinys. Į sisteminę literatūros analizę įtrauktų straipsnių apžvalga

Užklausa	Straipsnio pavadinimas (nuoroda)	Analizuojami duomenų rinkiniai (*angl. datasets*)	Pagrindinis tiriamas klausimas / problema	Naudojamas metodas	Tyrimo dalykinė sritis	Prognozavimo algoritmai ir metodai	Algoritmų efektyvumo vertinimas	Ekonominių rodiklių tipai	Rezultatai
5	Impute-VSS: A comprehensive web-based visualization and simulation suite for comparative data imputation and statistical evaluation [23]	Duomenų rinkinys su dirbtinai įvestais trūkstamais duomenimis (20%, 40%, 60%, 80% trūkstamų reikšmių).	Kaip sukurti intuityvia internetinę platformą, kuri leistų praktikams ir tyrėjams efektyviai palyginti ir vizualizuoti įvairius trūkstamų duomenų užpildymo (imputacijos) metodus realiuoju laiku ?	Palyginamoji simuliacijos metodologija, integruojanti vienmačius (vidurkis, mediana, moda, atsitiktinis užpildymas), mašininio mokymosi (Gradient Boosting, Random Forest Regressor, Support Vector Regression) ir daugiavariacinius (MICE, KNN) imputacijos metodus su realaus laiko vizualizacija.	Duomenų mokslas, informatika, trūkstamų duomenų analizė, mašininis mokymasis.	Gradient Boosting, Random Forest Regressor, Support Vector Regression, MICE (Multiple Imputation by Chained Equations), KNN, vidurkio/medianos/modos imputacija.	MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), KDE persidengimai, asimetrija, ekscesas, Kolmogorovo-Smirnovo statistika ir p-reikšmė, Kullback-Leibler divergencija	Nors sistema pritaikoma bet kokio tipo duomenims, demonstravime naudojami sveikatos rodikliai.	Sukurta Impute-VSS platforma leidžia palyginti kelių imputacijos metodų efektyvumą su realaus laiko vizualizacija. MICE metodas parodė geriausią tikslumą esant 20% trūkstamų duomenų (MAE=0.9031, RMSE=2.8679), bet tikslumas mažėja didėjant trūkstamų reikšmių kiekiui.

Šaltinis: sudaryta darbo autoriaus. Priedo tęsinys 101 puslapyje.

3 priedo tęsinys. Į sisteminę literatūros analizę įtrauktų straipsnių apžvalga

Užklausa	Straipsnio pavadinimas (nuoroda)	Analizuojami duomenų rinkiniai (*angl. datasets*)	Pagrindinis tiriamas klausimas / problema	Naudojamas metodas	Tyrimo dalykinė sritis	Prognozavimo algoritmai ir metodai	Algoritmų efektyvumo vertinimas	Ekonominių rodiklių tipai	Rezultatai
6	Missing Data Imputation: A Comprehensive Review [51]	Įvairūs duomenų rinkiniai iš skirtingų sričių: švietimo duomenys, hidrometeologiniai duomenys, laiko eilučių duomenys iš finansų ir aplinkosaugos sektorių.	Kaip efektyviai spręsti trūkstamų duomenų problemą naudojant tradicinius, statistinius ir mašininio mokymosi metodus, užtikrinant analizės patikimumą ir tikslumą ?	Išsami literatūros apžvalga, apimanti deterministinius metodus (vidurkio/medianos užpildymas), tikimybinius modelius (EM algoritmas, daugiareikšmis užpildymas), mašininio mokymosi algoritmus (Random Forest, KNN, MICE, GAN) ir giliojo mokymosi strategijas.	Statistinė analizė, mašininis mokymasis, duomenų mokslas.	Mean/Median imputacija, MissForest, KNN, MICE, Random Forest, SVM, Expectation-Maximization (EM), Bayesian imputacija, GAN (Generative Adversarial Networks), gilieji neuroniniai tinklai.	MSE (Mean Squared Error), RMSE, MAE, MAPE (Mean Absolute Percentage Error), R^2, koreliacijos koeficientai, Kullback-Leibler divergencija, statistinių pasiskirstymų išsaugojimas.	Metodai taikomi įvairiems rodikliams.	Random Forest ir GAN metodai demonstruoja aukščiausią tikslumą sudėtingiems duomenų šablonams, MICE efektyviai tvarko mišrius duomenų tipus, gilaus mokymosi metodai pranašesni dideliems duomenų rinkiniams. Metodų pasirinkimas priklauso nuo trūkstamų duomenų mechanizmo tipo (MCAR, MAR, MNAR).

Šaltinis: sudaryta darbo autoriaus. Priedo tęsinys 102 puslapyje.

3 priedo tęsinys. Į sisteminę literatūros analizę įtrauktų straipsnių apžvalga

Užklausa	Straipsnio pavadinimas (nuoroda)	Analizuojami duomenų rinkiniai (*angl. datasets*)	Pagrindinis tiriamas klausimas / problema	Naudojamas metodas	Tyrimo dalykinė sritis	Prognozavimo algoritmai ir metodai	Algoritmų efektyvumo vertinimas	Ekonominių rodiklių tipai	Rezultatai
7	A survey of missing data imputation techniques: statistical methods, machine learning models, and GAN-based approaches [11]	Apžvalga apima įvairius duomenų tipus: medicininius elektroninių sveikatos įrašų duomenis, ekonomikos duomenis, energetikos sektorių, aplinkosaugos duomenis, laiko eilučių duomenis, vaizdo duomenis.	Kaip efektyviai spręsti trūkstamų duomenų problemą naudojant tradicinius statistinius metodus, mašininio mokymosi modelius ir generatyvinius priešpriešinius tinklus (GAN)?	Lyginamoji apžvalga trijų kategorijų metodų: statistinių metodų (hot-deck, cold-deck, vidurkio/medianos, MICE), mašininio mokymosi (linijinė/logistinė regresija, KNN, sprendimų medžiai), ir GAN pagrįstų modelių (GAIN, VIGAN, SolarGAN).	Duomenų mokslas, dirbtinis intelektas, statistinė analizė	Hot-deck, Cold-deck, Mean/Median imputacija, MICE, linijinė regresija, logistinė regresija, KNN, sprendimų medžiai, GAIN (Generative Adversarial Imputation Network), VIGAN (View Imputation GAN), SolarGAN	MSE, RMSE, MAE, MAPE (regresijos metrikos), Accuracy, Precision, Recall, F1-Score (klasifikacijos metrikos), duomenų pasiskirstymo išsaugojimas.	Nors apžvalga nėra specializuota ekonomikai, aptariami metodai yra universalūs ir taikytini įvairių tipų duomenims, įskaitant kiekybinius nominalinius ir ranginius rodiklius.	GAN modeliai išsiskiria gebėjimu tvarkyti sudėtingus duomenų rinkinius ir netiesinius ryšius, pranokdami tradicinius metodus. GAIN, VIGAN ir SolarGAN efektyviai sprendžia MCAR, MAR ir MNAR tipo trūkstamų duomenų problemas.

Šaltinis: sudaryta darbo autoriaus. Priedo tęsinys 103 puslapyje.

3 priedo tęsinys. Į sisteminę literatūros analizę įtrauktų straipsnių apžvalga

Užklausa	Straipsnio pavadinimas (nuoroda)	Analizuojami duomenų rinkiniai (*angl. datasets*)	Pagrindinis tiriamas klausimas / problema	Naudojamas metodas	Tyrimo dalykinė sritis	Prognozavimo algoritmai ir metodai	Algoritmų efektyvumo vertinimas	Ekonominių rodiklių tipai	Rezultatai
7	A systematic review of machine learning techniques for missing value imputation [12]	Duomenų dydis svyravo nuo 6 iki 10 milijonų įrašų.	Kokie mašininio mokymosi metodai buvo sukurti ir taikomi trūkstamų duomenų imputacijai.	Sisteminė literatūros apžvalga, analizuojant ML imputacijos metodus pagal algoritmo kategoriją, eksperimento parametrus, duomenų charakteristikas, trūkstamų duomenų lygį ir vertinimo metrikas.	Mašininis mokymasis, duomenų mokslas.	Klasterizavimas, Instance-based metodai, genetiniai / hibridiniai algoritmai, ansambliai, neuroniniai tinklai, regresija, sprendimų medžiai, gilieji tinklai, Bajeso metodai.	PCP (Percentage of Correct Prediction), RMSE, MAE, MSE, NRMSE, klasifikacijos tikslumas, imputacijos klaida.	Įvairūs duomenų tipai.	KNN ir klasterizavimo algoritmai populiariausi dėl paprastumo ir efektyvumo. 70 % studijų dirbtinai įvedė trūkstamus duomenis testavimui. MAR mechanizmo duomenys buvo dažniausiai nagrinėti.

Šaltinis: sudaryta darbo autoriaus.

4 priedas. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Raudona spalva pažymėti rodikliai žymi ekonominius kintamuosius, kurie duomenų valymo etape buvo pašalinti iš analizės dėl metodologinio nesuderinamumo su NUTS 2 lygmens duomenimis.

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_FS_18	Nutraukusių veiklą įmonių pasiskirstymas pagal dydžio klases procentais	519	5581	91,49
IQG_05	Korupcijos subjektyvusis patyrimo rodiklis regiono lygmeniu	535	5565	91,23
IQG_06	Korupcijos subjektyvusis lūkesčių rodiklis regiono lygmeniu	535	5565	91,23
SCH_FS_16	Naujai įregistruotų įmonių pasiskirstymas pagal dydžio klases procentais	645	5545	89,43
SCH_FS_19	Prieš 3 metus įkurtų ir išlikusių įmonių dalies rodiklis procentais.	852	5248	86,03
IQG_03	Europos kokybės indeksas (EQI) - bendrasis institucinio valdymo kokybės rodiklis	905	5195	85,16
IQG_01	Institucinio valdymo kokybės rodiklis	905	5195	85,16
IQG_02	Institucinio valdymo nešališkumo rodiklis	905	5195	85,16
IQG_04	Korupcijos suvokimo rodiklis regiono lygmeniu	905	5195	85,16
SAN_UR_02	Dirbtinių paviršių dengta žemė pagal NUTS 2 regioną - žmogaus sukurtais statiniais, infrastruktūra ar kitais ne natūraliais paviršiais užimtos teritorijos dalis regione	1000	5100	83,61
SCH_FS_13	Grynasis įmonių populiacijos augimas procentais - bendras įmonių skaičiaus pokytis per ataskaitinius metus, įvertinant naujai įregistruotų ir veiklą nutraukusių įmonių skirtumą	1198	4902	80,36
SCH_FS_14	Bendras įmonių skaičiaus kaitos rodiklis, apskaičiuojamas sudedant naujai įregistruotų ir veiklą nutraukusių įmonių santykinius pokyčius	1216	4884	80,07

Šaltinis: sudaryta darbo autoriaus remiantis [53], [150]. Priedo tęsinys 105 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_FS_17	Veiklą nutraukusių įmonių skaičiaus ir visų veikiančių įmonių skaičiaus santykis, išreikštas procentais.	1216	4884	80,07
SCH_FS_24	Procentinė darbuotojų dalis, dirbusi įmonėse, kurios per ataskaitinius metus nutraukė veiklą, palyginti su visų veikiančių įmonių darbuotojų skaičiumi	1216	4884	80,07
SCH_FS_25	Bendras per metus veiklą nutraukusių įmonių darbuotojų skaičius, padalytas iš veiklą nutraukusių įmonių skaičiaus.	1308	4792	78,56
SCH_FS_20	Procentinė visų veikiančių įmonių dalis, sudaryta iš įmonių, kurioms nuo įsteigimo yra praėję treji metai	1374	4726	77,48
SCH_FS_26	Procentinis darbuotojų skaičiaus pokytis įmonėse, kurios įsteigtos prieš trejus metus ir išliko veikiančios iki ataskaitinių metų	1374	4726	77,48
SCH_FS_27	Procentinė darbuotojų dalis, dirbanti trejus metus veikiančiose įmonėse, palyginti su visų veikiančių įmonių darbuotojų skaičiumi.	1374	4726	77,48
SCH_FS_28	Darbuotojų skaičius įmonėse, įsteigtose prieš trejus metus ir išlikusios veikiančios, padalytas iš tokių įmonių skaičiaus.	1374	4726	77,48
SCH_FS_23	Procentinė visų veikiančių įmonių darbuotojų dalis, dirbanti įmonėse, kurios buvo įsteigtos tais pačiais ataskaitiniais metais.	1392	4708	77,18
SCH_FS_15	Įmonių steigimo rodiklis - tai procentinė dalis, gaunama naujai įregistruotų įmonių skaičių per ataskaitinį laikotarpį padalijus iš tų metų veikiančių įmonių skaičiaus.	1393	4707	77,16

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 106 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_FS_21	Įmonių steigimo užimtumo dalis - tai procentinė darbuotojų dalis, dirbanti tais metais naujai įsteigtose įmonėse, palyginti su visų veikiančių įmonių darbuotojų skaičiumi.	1393	4707	77,16
SCH_FS_22	Naujai įsteigtų įmonių vidutinis rodiklis - tai bendras darbuotojų skaičius įmonėse, įsteigtose ataskaitiniais metais, padalytas iš tais metais įsteigtų įmonių skaičiaus.	1393	4707	77,16
SCH_FS_03	Bendras įmonių, kurios per ataskaitinius metus nutraukė veiklą, skaičius.	1527	4573	74,97
SCH_FS_09	Darbuotojų skaičius įmonėse, kurios per ataskaitinius metus nutraukė veiklą - tai bendras tokių įmonių darbuotojų skaičius.	1527	4573	74,97
SCH_FS_10	Darbuotojų (samdomųjų) skaičius įmonėse, kurios per ataskaitinius metus nutraukė veiklą - tai bendras tokių įmonių samdomųjų darbuotojų skaičius.	1527	4573	74,97
INNO_RD_01	Vidaus mokslinių tyrimų ir eksperimentinės plėtros (MTEP) išlaidos - tai bendrosios regiono organizacijų patirtos MTEP sąnaudos, apskaičiuotos pagal NUTS 2 regioną.	1553	4547	74,54
SCH_FS_11	Darbuotojų skaičius įmonėse, kurios buvo įsteigtos prieš trejus metus ir išliko veikiančios iki ataskaitinių metų, - tai bendras tokių įmonių dirbančiųjų skaičius.	1668	4432	72,66
SCH_FS_12	Bendras dirbančiųjų skaičius įmonėse, kurios buvo įsteigtos tais pačiais ataskaitiniais metais.	1669	4431	72,64

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 107 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_FS_04	Įmonių, įsteigtų prieš trejus metus ir išlikusių veikiančiomis iki ataskaitinių metų, skaičius - tai bendras tokių trejų metų senumo veikiančių įmonių sudarantis skaičius	1670	4430	72,62
SCH_FS_08	Samdomųjų darbuotojų skaičius įmonėse, kurios ataskaitiniais metais buvo naujai įsteigtos - tai bendras tokių įmonių samdomųjų darbuotojų skaičius.	1688	4412	72,33
SCH_FS_01	Veikiančių įmonių skaičius - tai bendras įmonių, kurios ataskaitiniais metais vykdė ekonominę veiklą, skaičius.	1689	4411	72,31
SCH_FS_02	Naujai įsteigtų įmonių skaičius - tai bendras įmonių, kurios ataskaitiniais metais pradėjo veiklą, skaičius.	1689	4411	72,31
SCH_FS_05	Darbuotojų skaičius veikiančiose įmonėse - tai bendras dirbančiųjų skaičius įmonėse, kurios ataskaitiniais metais vykdė ekonominę veiklą.	1689	4411	72,31
SCH_FS_06	Samdomųjų darbuotojų skaičius veikiančiose įmonėse - tai bendras samdomųjų darbuotojų skaičius įmonėse, kurios ataskaitiniais metais vykdė ekonominę veiklą.	1689	4411	72,31
SCH_FS_07	Darbuotojų skaičius naujai įsteigtose įmonėse - tai bendras dirbančiųjų skaičius įmonėse, kurios ataskaitiniais metais pradėjo veiklą	1689	4411	72,31
INFRA_TR_05	Keleivių pervežimas pagal NUTS 2 regioną - tai ataskaitiniais metais jūrų transportu pervežtų keleivių skaičius, susietas su konkrečiu NUTS 2 regionu.	1854	4246	69,61

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 108 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
INFRA_ICT_01	Namų ūkių, turinčių interneto prieigą namuose, dalis pagal NUTS 2 regioną - tai procentinė namų ūkių dalis, kuri turi interneto ryšį savo gyvenamojoje vietoje.	1902	4198	68,82
INFRA_ICT_02	Asmenų, reguliariai besinaudojantys internetu, dalis pagal NUTS 2 regioną - tai procentinė gyventojų dalis, kuri internetą naudoja įprastai ir pakankamai dažnai.	1905	4195	68,77
HC(Lf)Q_EDUC_08	Aukštojo mokslo studentų dalies santykis su atitinkamo regiono gyventojų dalimi - tai rodiklis, lyginantis studentų, studijuojančių aukštosiose mokyklose, proporciją su bendro gyventojų skaičiaus proporcija NUTS 2 regionuose.	2085	4015	65,82
HC(Lf)Q_EDUC_01	Aukštojo išsilavinimo rodiklis - tai gyventojų, turinčių aukštąjį išsilavinimą, dalis pagal NUTS regioną.	2853	3247	53,23
SAN_SZ_EU_01	Regiono teritorijos ploto rodiklis, pateikiamas ES mastu, naudojamas palyginti regiono žemės plotą su Europos Sąjungos lygiu.	2928	3172	52
INFRA_HLTH_03	Sveikatos priežiūros specialistų skaičius pagal NUTS 2 regioną, parodantis regione dirbančių sveikatos sektoriaus darbuotojų skaičių.	3534	2566	42,07
INFRA_TR_02	Oro transportu gabenamų krovinių kiekis pagal NUTS 2 regioną - tai ataskaitiniais metais oro transportu pervežtų krovinių apimtis, priskiriama konkrečiam NUTS 2 regionui.	3704	2396	39,28

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 109 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
INFRA_TR_01	Keleivių vežimas oro transportu pagal NUTS 2 regioną – tai ataskaitiniais metais oro transportu pervežtų keleivių skaičius, priskirtas konkrečiam NUTS 2 regionui	3741	2359	38,67
INFRA_TR_04	Geležinkelio linijų bendras ilgis - tai visų regione esančių geležinkelio linijų ilgis, pateikiamas pagal NUTS 2 regioną.	3772	2328	38,16
SAN_SZ_01	Regiono žemės plotas - tai bendras konkretaus NUTS 2 regiono sausumos teritorijos plotas.	4105	1995	32,7
INFRA_HLTH_01	Laisvų (esamų) ligoninių lovų skaičius pagal NUTS 2 regioną - tai bendras regione esančių ligoninių turimų lovų skaičius.	4282	1818	29,8
INFRA_TR_03	Automagistralių ilgis - tai bendras regione esančių automagistralių ilgis, pateikiamas pagal NUTS 2 regioną.	4353	1 747	28,64
INFRA_HLTH_02	Rodiklis, parodantis gydytojų skaičių 100 tūkst. gyventojų regione.	4413	1687	27,66
INFRA_TO_03	Rodiklis, parodantis turistų nakvynių skaičių apgyvendinimo įstaigose, tenkantį tūkstančiui regiono gyventojų.	4767	1333	21,85
INFRA_TO_01	Rodiklis, parodantis turistų nakvynių skaičių apgyvendinimo įstaigose, tenkantį vienam kvadratiniam kilometrui regiono ploto.	4910	1190	19,51
Y_EU_01	Rodiklis, parodantis technologijų ir žinių - intensyviuose sektoriuose dirbančių asmenų dalį nuo viso regiono užimtųjų skaičiaus (proc.).	4984	1116	18,3

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 110 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
INFRA_TO_02	Rodiklis, parodantis apgyvendinimo įstaigų vietų (lovų) skaičių, tenkantį tūkstančiui regiono gyventojų.	5022	1078	17,67
HC(Lf)Q_EDUC_07	Parodo, kokia dalis 18-24 metų asmenų regione anksčiau nei numatyta pasitraukia iš švietimo ir mokymo sistemos (proc.).	5045	1055	17,3
Y_EU_02	Nusako regiono ekonomikos struktūrą, matuojant pažangių technologijų ir žinių-intensyvių paslaugų sektorių sukuriamos pridėtinės vertės dalį nuo visos pridėtinės vertės (proc.).	5260	840	13,77
SAN_PD_01	Apibūdina regiono visuomenės saugumą, fiksuodamas naujai užregistruotų nusikalstamų veikų skaičių, tenkantį tūkstančiui gyventojų.	5392	708	11,61
HC(Lf)Q_EDUC_02	Įvardija regiono švietimo lygį, matuodamas aukštąjį išsilavinimą turinčių 25-64 metų gyventojų dalį (proc.).	5449	651	10,67
Y_01	Darbuotojų dalis technologijų ir žinių intensyviuose sektoriuose, pagal NACE Rev. 2 veiklos klasifikaciją.	5535	565	9,26
SCH_IS_03	Nusako regiono ekonominę struktūrą, rodydamas apdirbamosios gamybos (sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5563	537	8,8
SCH_IS_EU_03	Žymi regiono ekonominę specializaciją, vertinant apdirbamosios gamybos (sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5563	537	8,8

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 111 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_IS_08	Apibrėžia regiono ekonominę struktūrą, vertindamas finansų ir draudimo veiklos (sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5 578	522	8,56
SCH_IS_12	Parodo sektorių (viešosios administracijos, švietimo, sveikatos apsaugos ir socialinio darbo) sukuriamos bendrosios pridėtinės vertės dalį regione (proc.).	5 578	522	8,56
SCH_IS_EU_08	Nusako regiono ekonominę struktūrą, vertinant finansų ir draudimo veiklos (sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5 578	522	8,56
SCH_IS_EU_12	Įvertina regiono ekonominę specializaciją, matuojant sektorių (viešosios administracijos, švietimo, sveikatos apsaugos ir socialinio darbo) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5578	522	8,56
SCH_IS_05	Apibūdina regiono ekonominę struktūrą, vertindamas sektorių (didmeninė ir mažmeninė prekyba, transportas, apgyvendinimas ir maitinimas) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5580	520	8,52
SCH_IS_07	Parodo regiono ekonominę struktūrą, vertinant informacijos ir ryšių sektoriaus (sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5580	520	8,52

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 112 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_IS_10	Įvardija regiono ekonominę struktūrą, matuojant nekilnojamojo turto veiklos (sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5580	520	8,52
SCH_IS_11	Atspindi regiono ekonominę struktūrą, vertindamas profesinių, mokslinių, techninių bei administracinių paslaugų veiklų (sektorių) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_14	Nusako regiono ekonominę struktūrą, vertinant sektorių (meno, pramogų, poilsio, kitų paslaugų veiklų) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_EU_05	Pažymi regiono ekonominę specializaciją, vertinant sektorių (prekybos, transporto, apgyvendinimo ir maitinimo paslaugų) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_EU_07	Parodo regiono ekonominę specializaciją, vertinant informacijos ir ryšių sektoriaus sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_EU_10	Įvertina regiono ekonominę specializaciją, matuojant nekilnojamojo turto veiklos (sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 113 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_IS_EU_11	Parodo regiono ekonominę specializaciją, vertinant profesinių, mokslinių, techninių ir administracinių paslaugų veiklų (sektorių) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5580	520	8,52
SCH_IS_EU_14	Nusako regiono ekonominę specializaciją, vertinant sektorių (meno, pramogų, poilsio ir kitų paslaugų veiklų) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5580	520	8,52
HC(Lf)Q_EDUC_03	Apibūdina regiono švietimo dalyvavimą, matuojant pirminio (pradinio) ugdymo programose dalyvaujančių gyventojų dalį (proc.).	5596	504	8,26
HC(Lf)Q_EDUC_04	Nusako regiono švietimo lygį, matuojant vidutinį (antrinį) išsilavinimą turinčių 25-64 metų gyventojų dalį (proc.).	5598	502	8,23
HC(Lf)Q_EDUC_05	Parodo regiono švietimo lygį, vertinant aukštąjį išsilavinimą turinčių 25-64 metų gyventojų dalį (proc.).			
SCH_IS_01	Apibūdina regiono ekonominę struktūrą, vertindamas žemės ūkio, miškininkystės ir žuvininkystės veiklų (sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5601	499	8,18
SCH_IS_02	Parodo regiono ekonominę struktūrą, matuojant pramonės (sektorių, išskyrus statybą) sukuriamos bendrosios pridėtinės vertės dalį (proc.)			

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 114 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_IS_04	Nusako regiono ekonominę struktūrą, vertinant statybos sektoriaus sukuriamos bendrosios pridėtinės vertės dalį (proc.)	5601	499	8,18
SCH_IS_06	Įvardija regiono ekonominę struktūrą, vertinant didmeninės ir mažmeninės prekybos bei susijusių paslaugų (sektorių) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_09	Atspindi regiono ekonominę struktūrą, vertinant finansų, draudimo ir verslo paslaugų (sektorių) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_13	Nusako regiono ekonominę struktūrą, matuojant viešosios administracijos, gynybos, švietimo, sveikatos apsaugos ir socialinio darbo veiklų (sektorių) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_EU_01	Įvertina regiono ekonominę specializaciją, matuojant žemės ūkio, miškininkystės ir žuvininkystės veiklų (A sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_EU_02	Parodo regiono ekonominę specializaciją, vertinant pramonės (sektorių, išskyrus statybą) sukuriamos bendrosios pridėtinės vertės dalį (proc.)			
SCH_IS_EU_04	Įvardija regiono ekonominę specializaciją, matuojant statybos sektoriaus sukuriamos bendrosios pridėtinės vertės dalį (proc.).			

Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 115 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	Rodiklio aprašymas	Užpildytų kiekis	Trūkstamų kiekis	Trūkstamų (%)
SCH_IS_EU_06	Pažymi regiono ekonominę specializaciją, vertinant didmeninės ir mažmeninės prekybos bei susijusių paslaugų (sektorių) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5601	499	8,18
SCH_IS_EU_09	Nusako regiono ekonominę specializaciją, vertinant finansų, draudimo ir verslo paslaugų (sektorių) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
SCH_IS_EU_13	Įvertina regiono ekonominę specializaciją, matuojant viešosios administracijos, gynybos, švietimo, sveikatos apsaugos ir socialinio darbo veiklų (sektorių) sukuriamos bendrosios pridėtinės vertės dalį (proc.).			
HC(Lf)Q_EDUC_06	Parodo regiono žmogiškojo kapitalo potencialą, matuojant asmenų, turinčių aukštąjį išsilavinimą ir/arba dirbančių mokslo bei technologijų srityse, dalį nuo visų gyventojų (proc.).	5 632	468	7,67
Y_02	Nusako regiono darbo rinkos situaciją, matuojant 15-74 metų gyventojų užimtumo lygį (proc.).	5 658	442	7,25
SAN_ED_01	Apibūdina regiono darbo rinką, fiksuodamas 15-74 metų užimtų asmenų skaičių (tūkst. žmonių).			
SAN_UR_01	Apibūdina regiono urbanizacijos lygį, matuojant miestuose gyvenančių gyventojų dalį nuo visos populiacijos (proc.).	5663	437	7,16

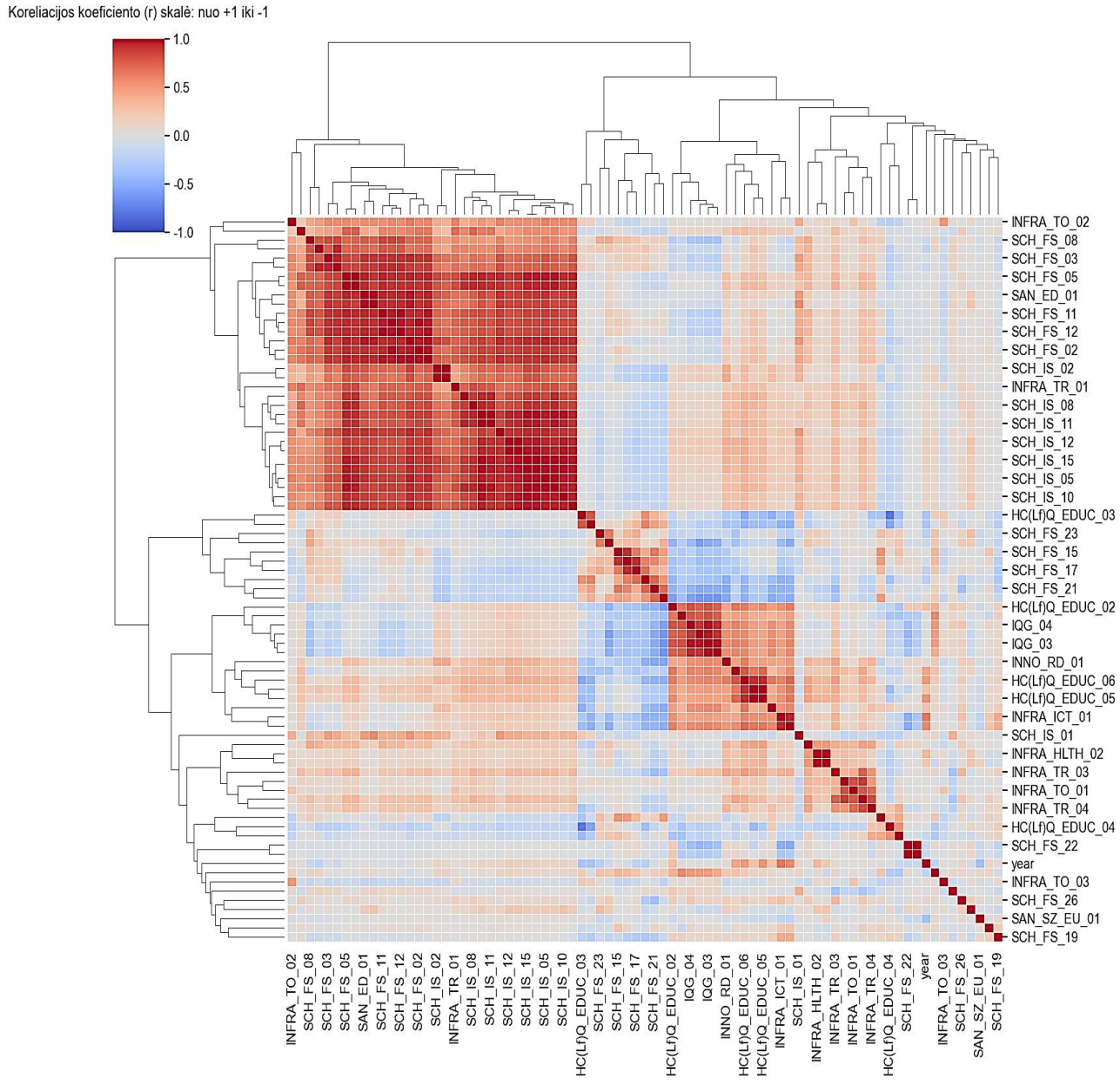
Šaltinis: sudaryta darbo autoriaus remiantis [53]. Priedo tęsinys 116 puslapyje.

4 priedo tęsinys. Analizuojamų ekonominių rodiklių sąrašas įvesties duomenų faile

Rodiklio kodas	**Rodiklio aprašymas**	**Užpildytų kiekis**	**Trūkstamų kiekis**	**Trūkstamų (%)**
SCH_IS_15	Parodo bendrą regiono ekonominės struktūros vaizdą, vertinant visų NACE veiklų (bendrojo ekonomikos sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.).	5698	402	6,59
SCH_IS_EU_15	Įvertina regiono ekonominę specializaciją, matuojant visų NACE veiklų (bendrojo ekonomikos sektoriaus) sukuriamos bendrosios pridėtinės vertės dalį (proc.)			
SAN_SZ_02	Nusako regiono ekonominį dydį, matuojant bendrąjį vidaus produktą (BVP) einamosiomis kainomis (mln. eurų).			
year	Nurodo, už kokius metus yra prieinami duomenys tam rodikliui.	6100	0	0
geo	Geografinis kodas, nusakantis regioną (NUTS 2), kuriam priklauso pateikti duomenys.			

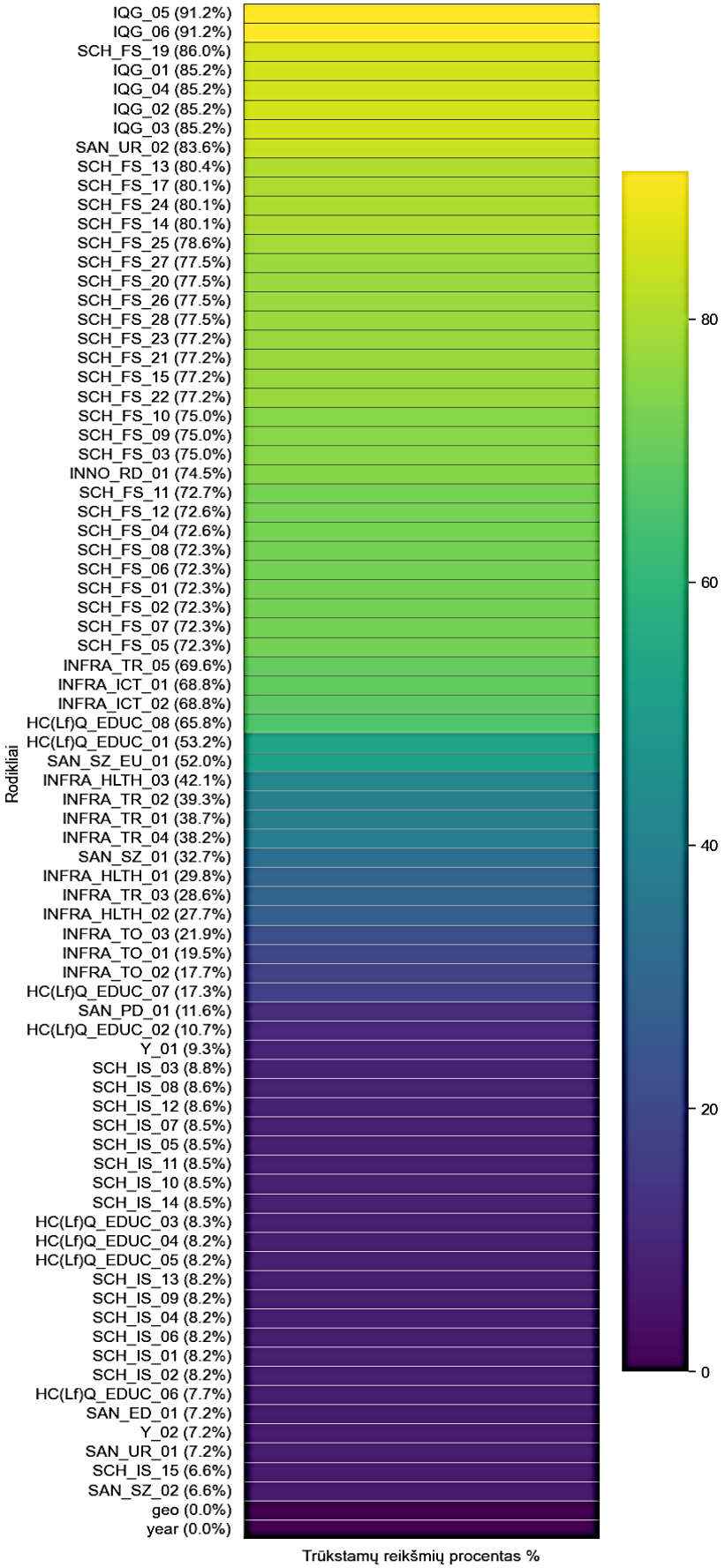
Šaltinis: sudaryta darbo autoriaus remiantis [53].

5 priedas. Koreliacijų matrica su hierarchiniu klasterizavimu tarp analizuojamų ekonominių rodiklių



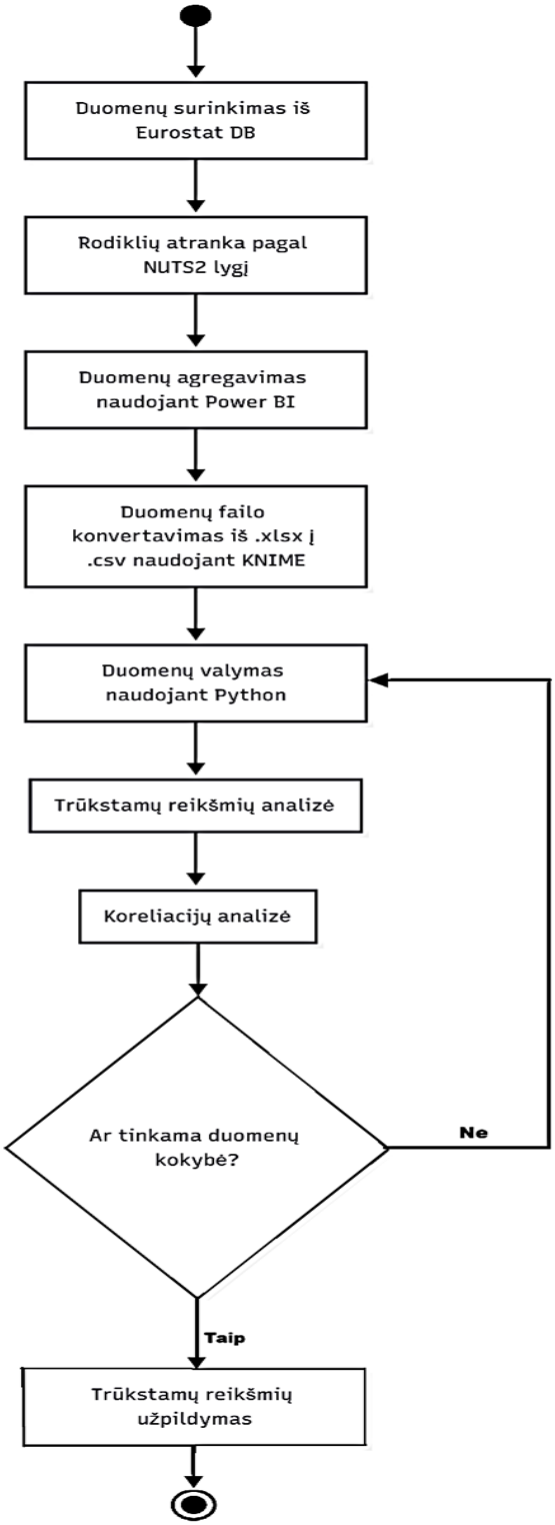
Šaltinis: sudaryta autoriaus.

6 priedas. Trūkstamų reikšmių procentinė analizė po duomenų valymo



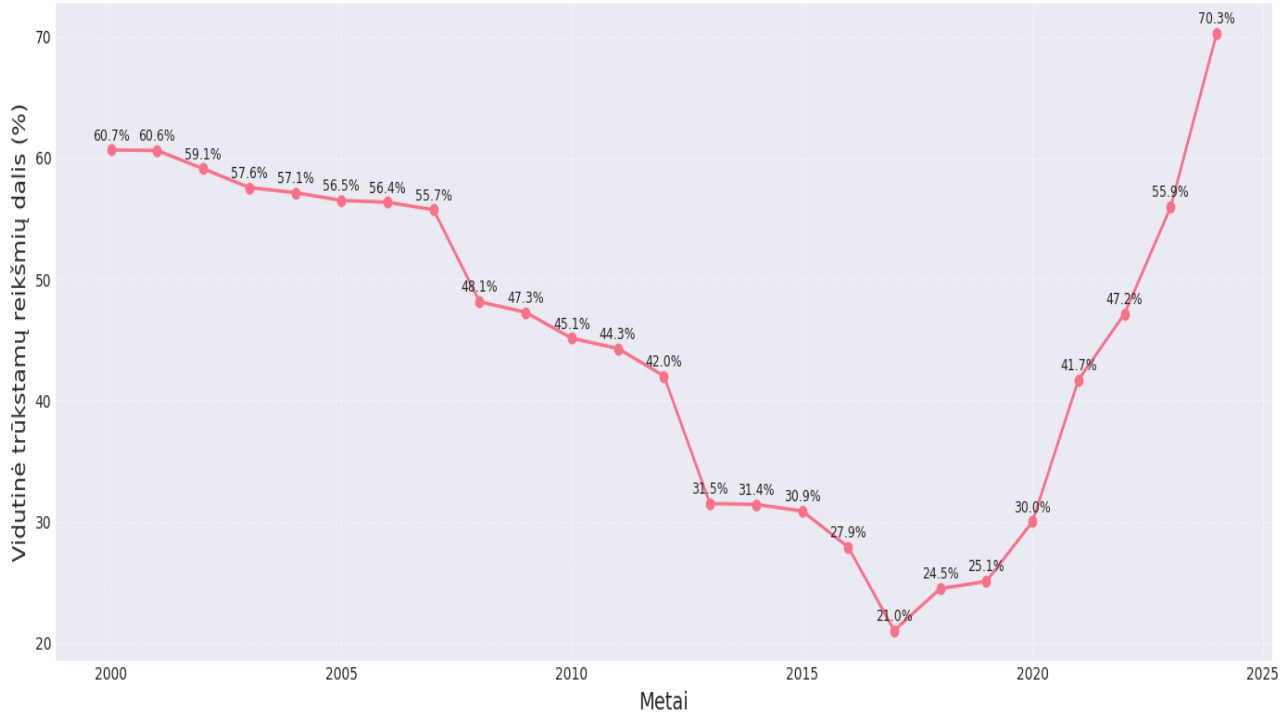
Šaltinis: sudaryta autoriaus.

7 priedas. Duomenų paruošimo metodinė eiga



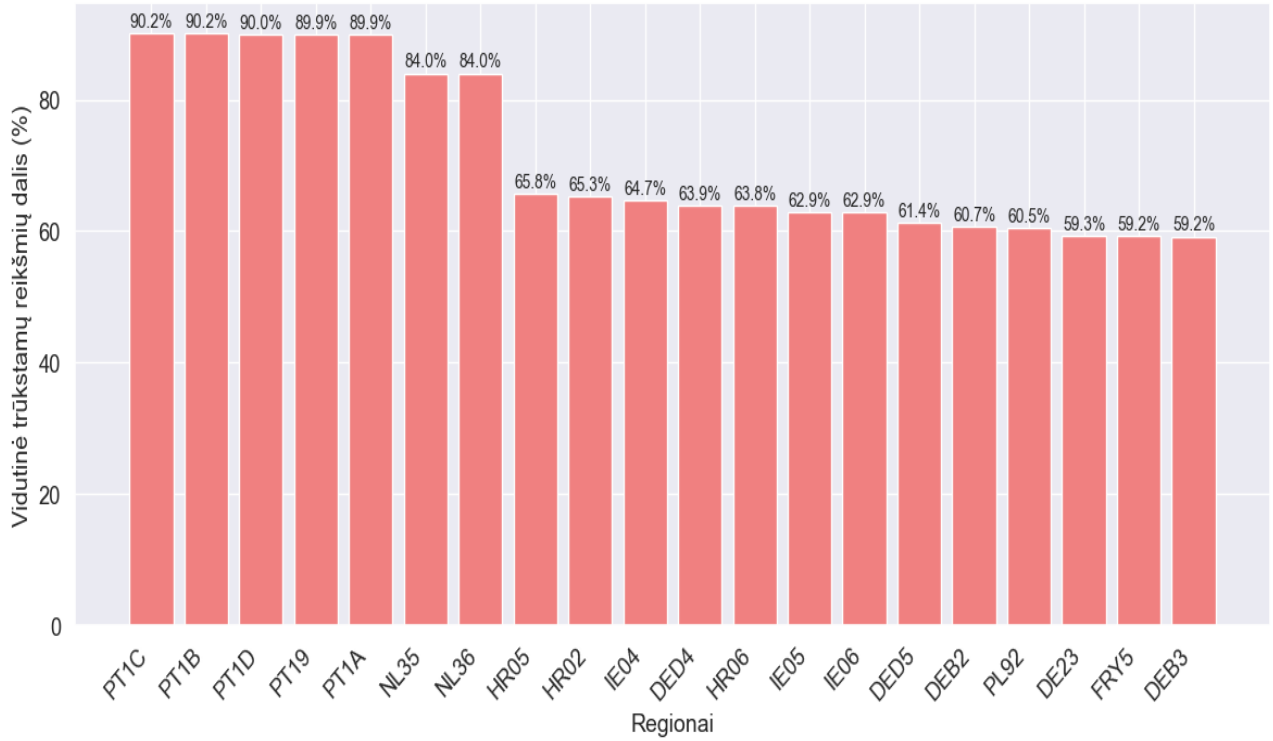
Šaltinis: sudaryta autoriaus.

8 priedas. Vidutinė trūkstamų reikšmių dalis pagal metus



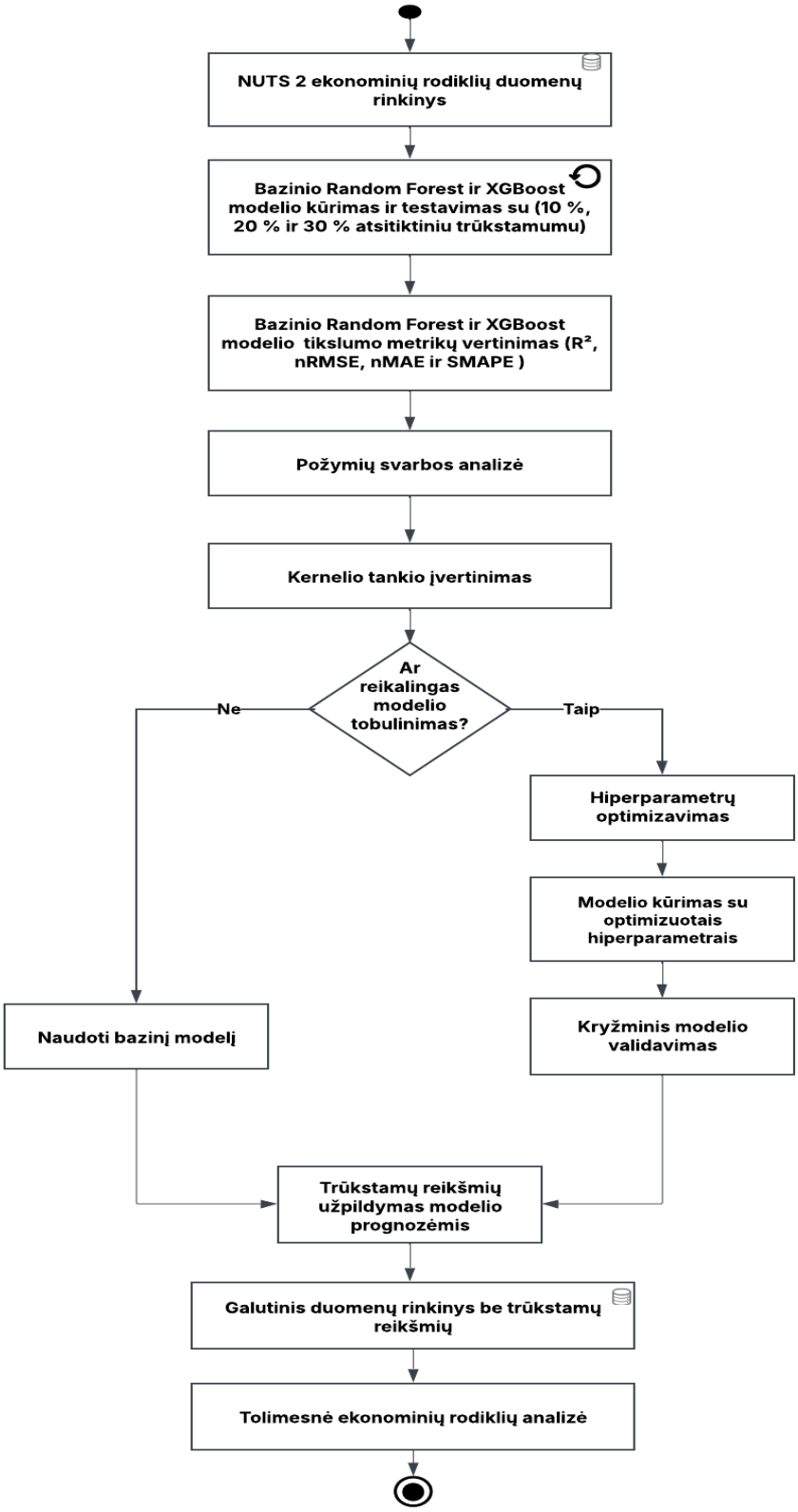
Šaltinis: sudaryta autoriaus.

9 priedas. Vidutinė trūkstamų reikšmių dalis pagal NUTS 2 regionus



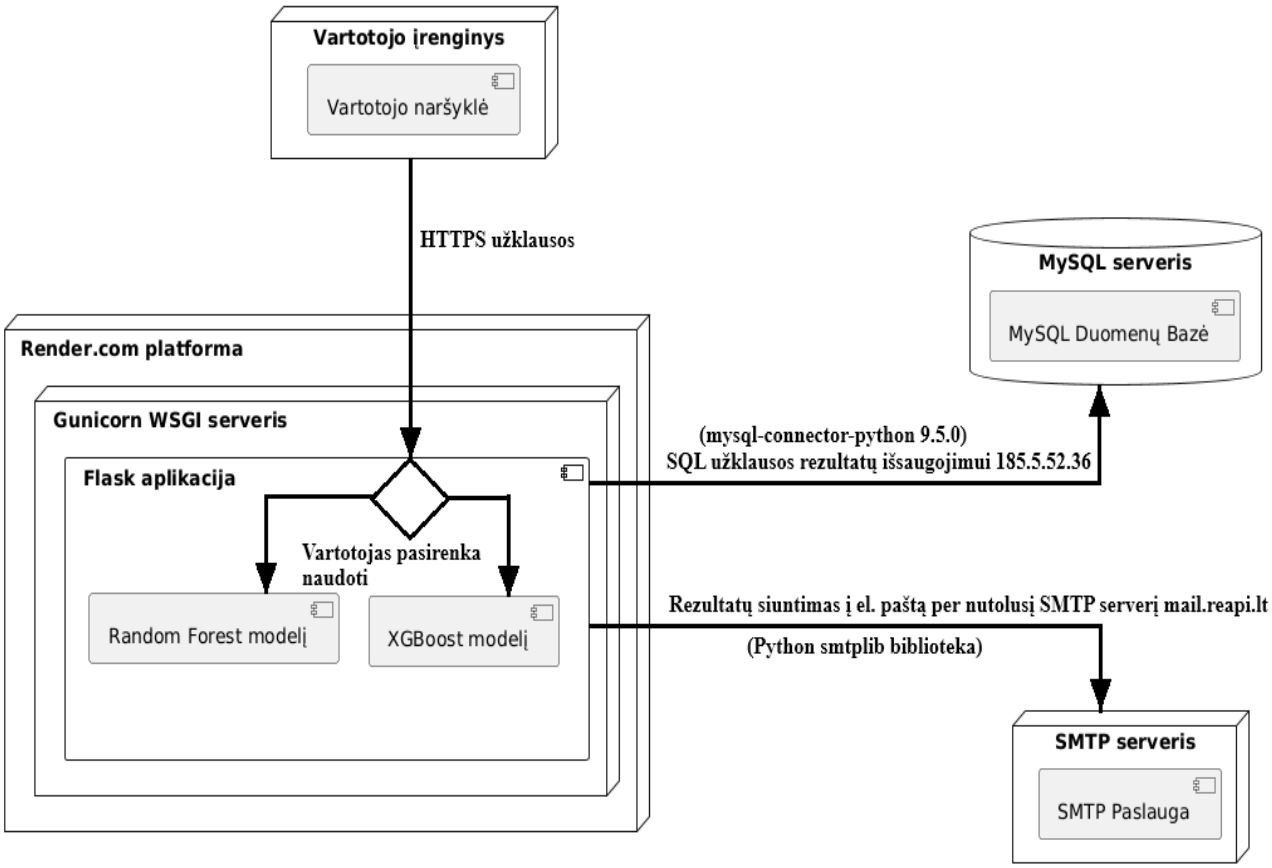
Šaltinis: sudaryta autoriaus.

10 priedas. Trūkstamų reikšmių užpildymo proceso blokinė schema NUTS 2 ekonominių rodiklių duomenyse



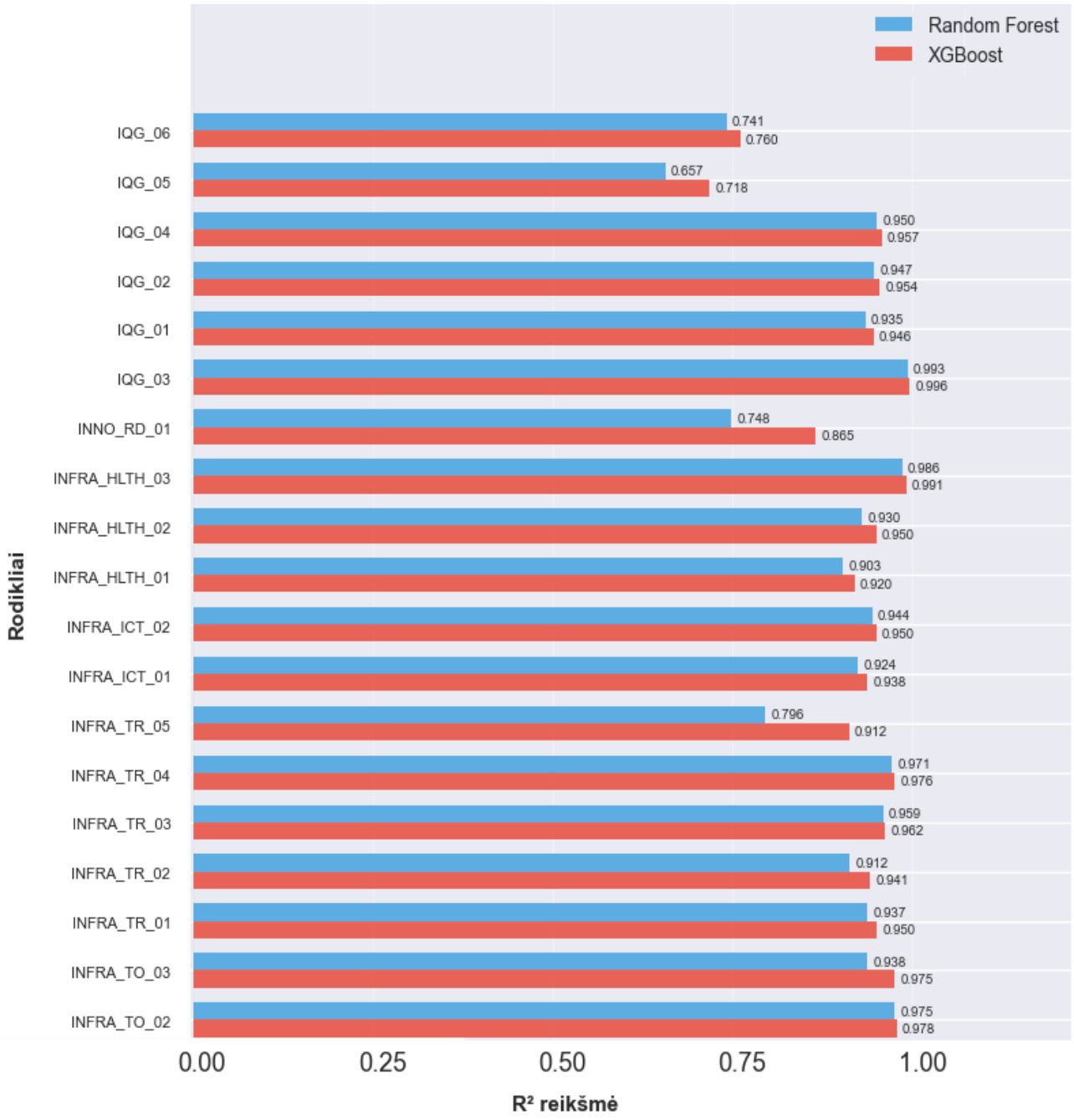
Šaltinis: sudaryta autoriaus.

11 priedas. Flask aplikacijos architektūra ir komponentų sąveika



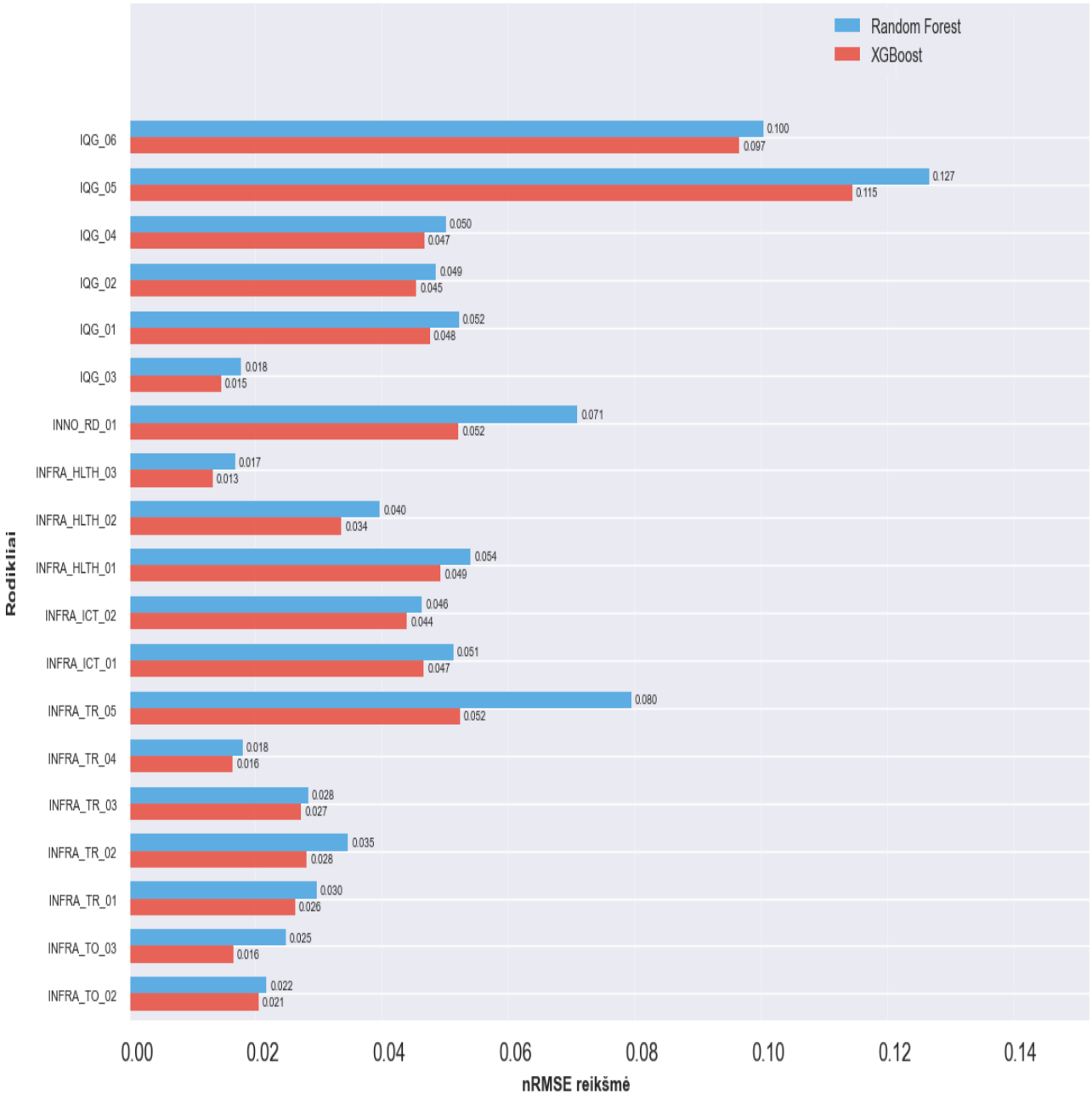
Šaltinis: sudaryta autoriaus.

12 priedas. Random Forest ir XGBoost optimizuotų modelių R^2 metrikų palyginimas pagal atskirus ekonominius rodiklius



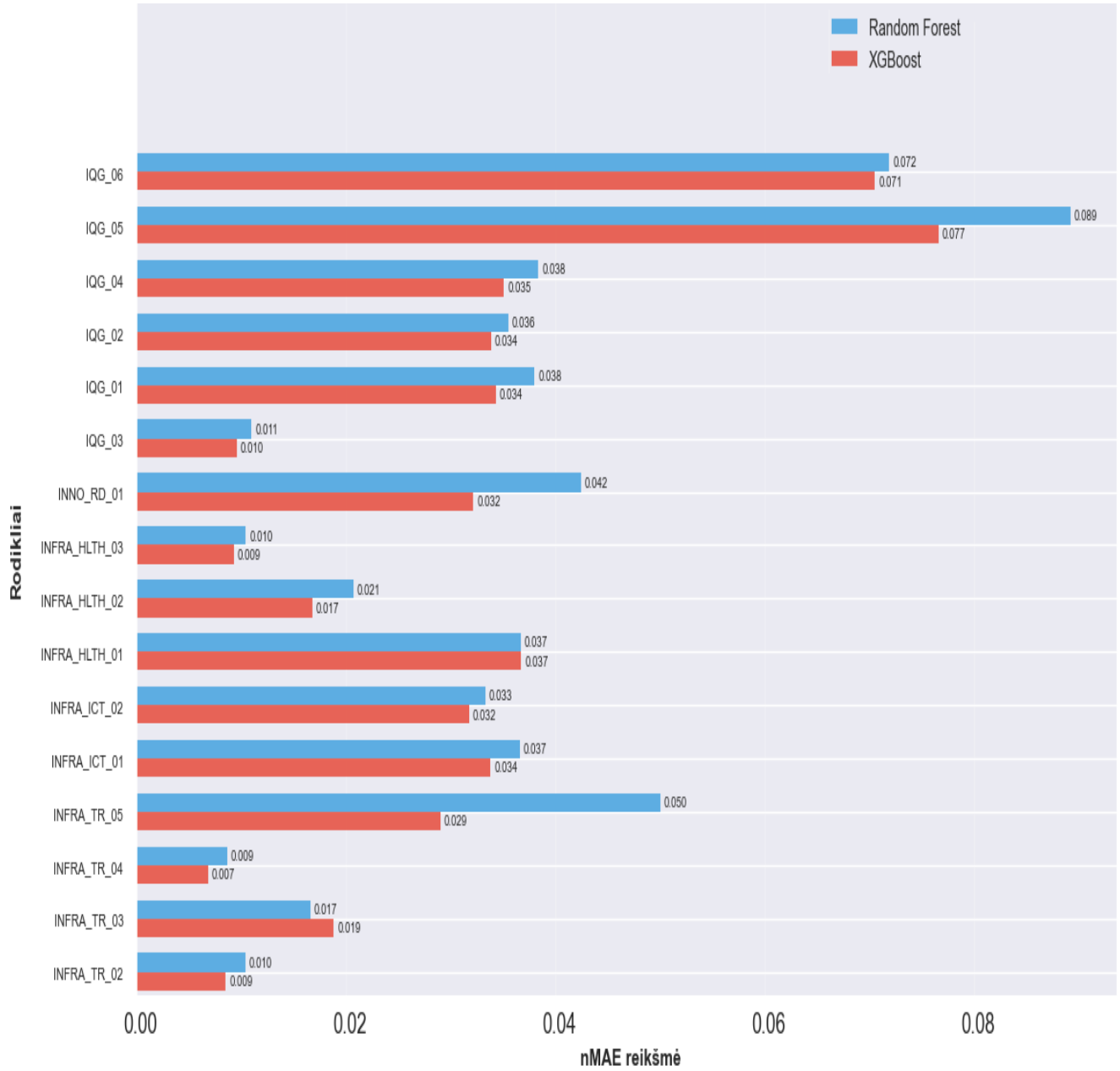
Šaltinis: sudaryta autoriaus.

13 priedas. Random Forest ir XGBoost optimizuotų modelių nRMSE metrikų palyginimas pagal atskirus ekonominius rodiklius



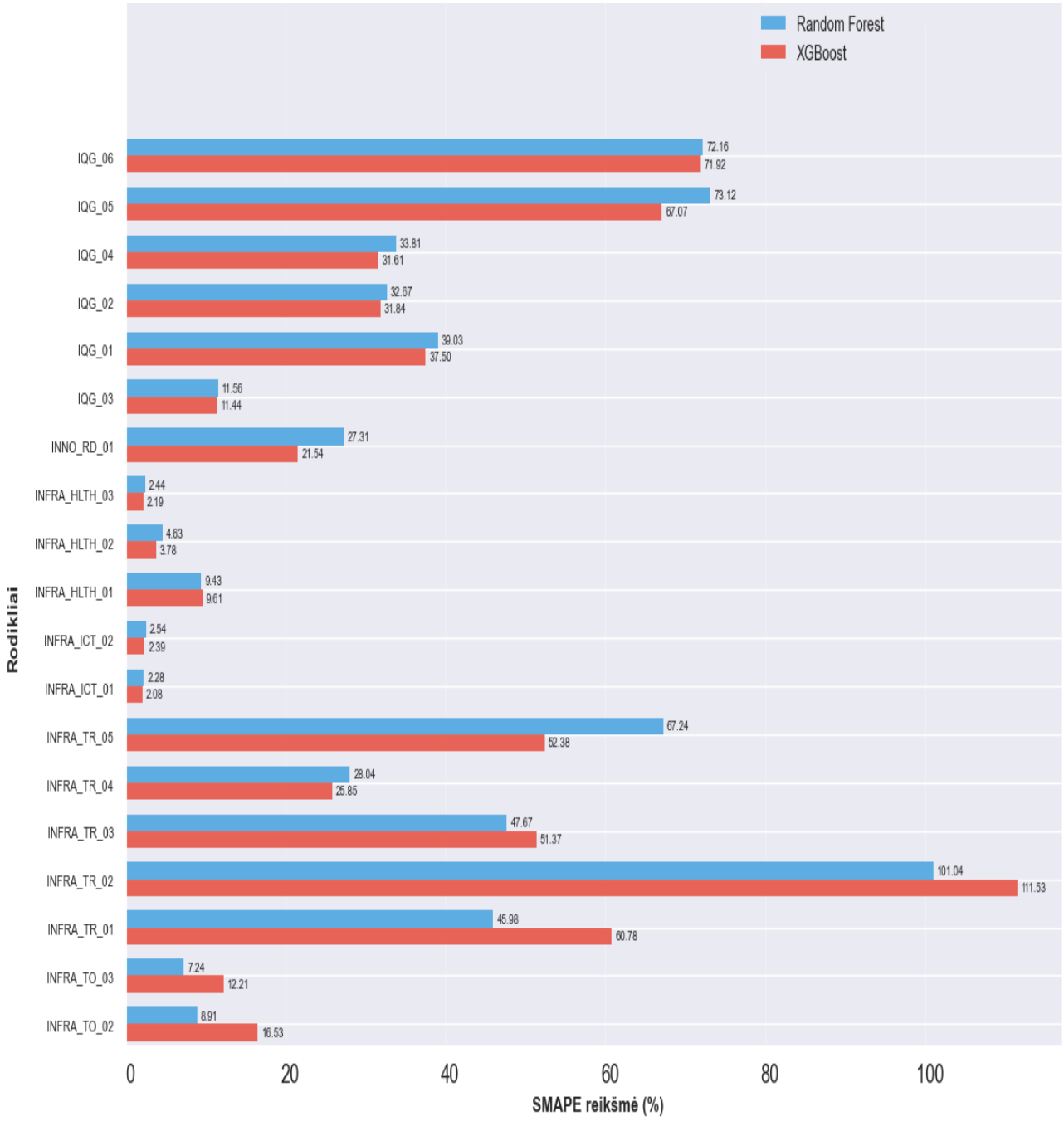
Šaltinis: sudaryta autoriaus.

14 priedas. Random Forest ir XGBoost optimizuotų modelių nMAE metrikų palyginimas pagal atskirus ekonominius rodiklius



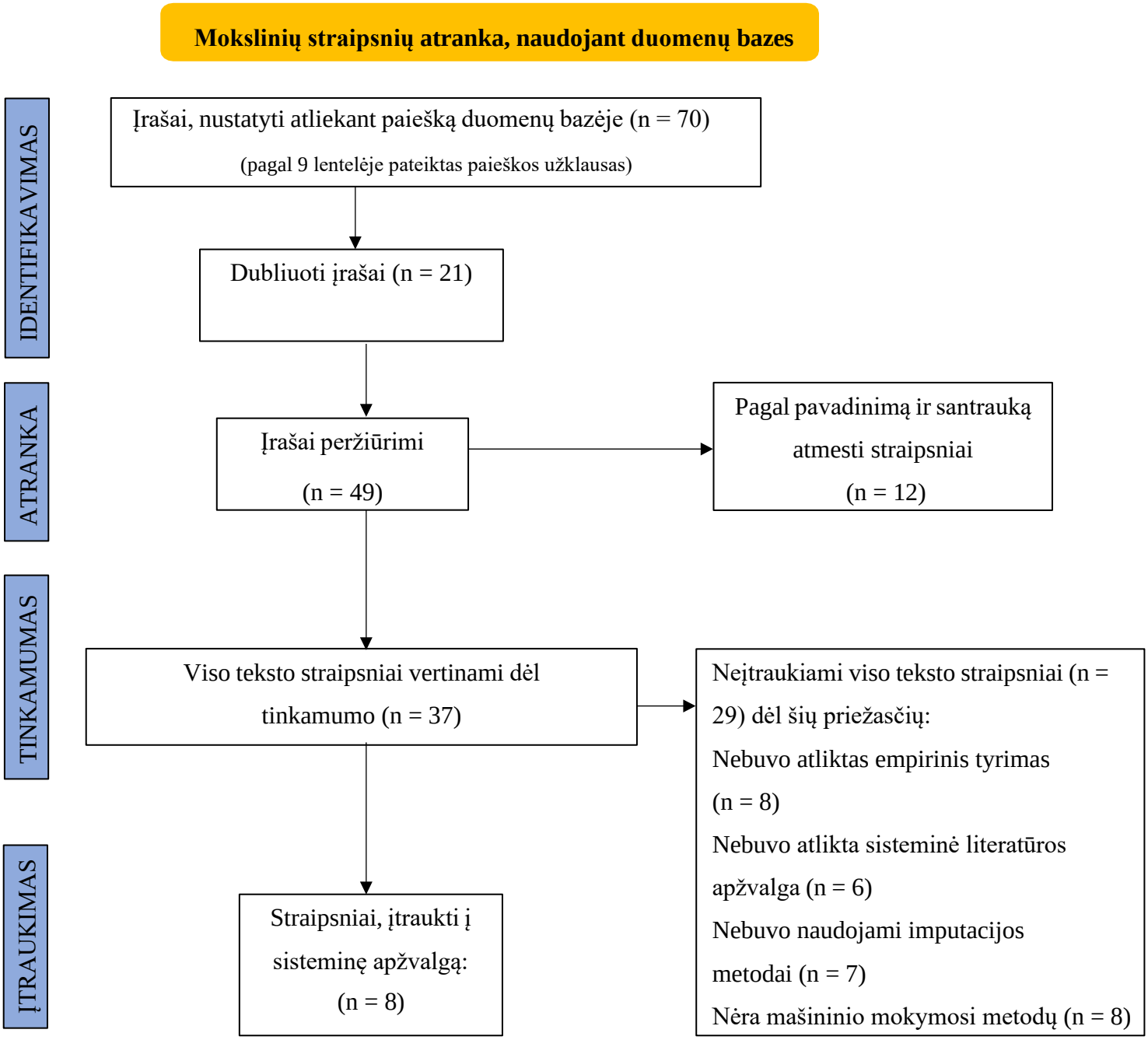
Šaltinis: sudaryta autoriaus.

15 priedas. Random Forest ir XGBoost optimizuotų modelių sMAPE metrikų palyginimas pagal atskirus ekonominius rodiklius



Šaltinis: sudaryta autoriaus.

16 priedas. Mokslinių publikacijų atrankos schema



Šaltinis: sudaryta autoriaus.

17 priedas. NUTS 2 regionų kodai ir atitinkamos šalys

NUTS 2 regiono kodas	Šalis
AT11	Austrija
AT12	Austrija
AT13	Austrija
AT21	Austrija
AT22	Austrija
AT31	Austrija
AT32	Austrija
AT33	Austrija
AT34	Austrija
BE10	Belgija
BE21	Belgija
BE22	Belgija
BE23	Belgija
BE24	Belgija
BE25	Belgija
BE31	Belgija
BE32	Belgija
BE33	Belgija
BE34	Belgija
BE35	Belgija
BG31	Bulgarija
BG32	Bulgarija
BG33	Bulgarija
BG34	Bulgarija
BG41	Bulgarija
BG42	Bulgarija
CY00	Kipras
CZ01	Čekija
CZ02	Čekija
CZ03	Čekija
CZ04	Čekija
CZ05	Čekija
CZ06	Čekija
CZ07	Čekija
CZ08	Čekija
DE11	Vokietija
DE12	Vokietija
DE13	Vokietija
DE14	Vokietija
DE21	Vokietija
DE22	Vokietija
DE23	Vokietija
DE24	Vokietija
DE25	Vokietija
DE26	Vokietija
DE27	Vokietija
DE30	Vokietija
DE40	Vokietija
DE50	Vokietija
DE60	Vokietija
DE71	Vokietija
DE72	Vokietija

NUTS 2 regiono kodas	Šalis
DE73	Vokietija
DE80	Vokietija
DE91	Vokietija
DE92	Vokietija
DE93	Vokietija
DE94	Vokietija
DEA1	Vokietija
DEA2	Vokietija
DEA3	Vokietija
DEA4	Vokietija
DEA5	Vokietija
DEB1	Vokietija
DEB2	Vokietija
DEB3	Vokietija
DEC0	Vokietija
DED2	Vokietija
DED4	Vokietija
DED5	Vokietija
DEE0	Vokietija
DEF0	Vokietija
DEG0	Vokietija
DK01	Danija
DK02	Danija
DK03	Danija
DK04	Danija
DK05	Danija
EE00	Estija
EL30	Graikija
EL41	Graikija
EL42	Graikija
EL43	Graikija
EL51	Graikija
EL52	Graikija
EL53	Graikija
EL54	Graikija
EL61	Graikija
EL62	Graikija
EL63	Graikija
EL64	Graikija
EL65	Graikija
ES11	Ispanija
ES12	Ispanija
ES13	Ispanija
ES21	Ispanija
ES22	Ispanija
ES23	Ispanija
ES24	Ispanija
ES30	Ispanija
ES41	Ispanija
ES42	Ispanija
ES43	Ispanija
ES51	Ispanija
ES52	Ispanija
ES53	Ispanija
ES61	Ispanija

NUTS 2 regiono kodas	Šalis
ES62	Ispanija
ES63	Ispanija
ES64	Ispanija
ES70	Ispanija
FI19	Suomija
FI1B	Suomija
FI1C	Suomija
FI1D	Suomija
FI20	Suomija
FR10	Prancūzija
FRB0	Prancūzija
FRC1	Prancūzija
FRC2	Prancūzija
FRD1	Prancūzija
FRD2	Prancūzija
FRE1	Prancūzija
FRE2	Prancūzija
FRF1	Prancūzija
FRF2	Prancūzija
FRF3	Prancūzija
FRG0	Prancūzija
FRH0	Prancūzija
FRI1	Prancūzija
FRI2	Prancūzija
FRI3	Prancūzija
FRJ1	Prancūzija
FRJ2	Prancūzija
FRK1	Prancūzija
FRK2	Prancūzija
FRL0	Prancūzija
FRM0	Prancūzija
FRY1	Prancūzija
FRY2	Prancūzija
FRY3	Prancūzija
FRY4	Prancūzija
FRY5	Prancūzija
HR02	Kroatija
HR03	Kroatija
HR05	Kroatija
HR06	Kroatija
HU11	Vengrija
HU12	Vengrija
HU21	Vengrija
HU22	Vengrija
HU23	Vengrija
HU31	Vengrija
HU32	Vengrija
HU33	Vengrija
IE04	Airija
IE05	Airija
IE06	Airija
ITC1	Italija
ITC2	Italija
ITC3	Italija
ITC4	Italija

NUTS 2 regiono kodas	Šalis
ITF1	Italija
ITF2	Italija
ITF3	Italija
ITF4	Italija
ITF5	Italija
ITF6	Italija
ITG1	Italija
ITG2	Italija
ITH1	Italija
ITH2	Italija
ITH3	Italija
ITH4	Italija
ITH5	Italija
ITI1	Italija
ITI2	Italija
ITI3	Italija
ITI4	Italija
LT01	Lietuva
LT02	Lietuva
LU00	Liuksemburgas
LV00	Latvija
MT00	Malta
NL11	Nyderlandai
NL12	Nyderlandai
NL13	Nyderlandai
NL21	Nyderlandai
NL22	Nyderlandai
NL23	Nyderlandai
NL32	Nyderlandai
NL34	Nyderlandai
NL35	Nyderlandai
NL36	Nyderlandai
NL41	Nyderlandai
NL42	Nyderlandai
PL21	Lenkija
PL22	Lenkija
PL41	Lenkija
PL42	Lenkija
PL43	Lenkija
PL51	Lenkija
PL52	Lenkija
PL61	Lenkija
PL62	Lenkija
PL63	Lenkija
PL71	Lenkija
PL72	Lenkija
PL81	Lenkija
PL82	Lenkija
PL84	Lenkija
PL91	Lenkija
PL92	Lenkija
PT11	Portugalija
PT15	Portugalija
PT19	Portugalija
PT1A	Portugalija

NUTS 2 regiono kodas	Šalis
PT1B	Portugalija
PT1C	Portugalija
PT1D	Portugalija
PT20	Portugalija
PT30	Portugalija
RO11	Rumunija
RO12	Rumunija
RO21	Rumunija
RO22	Rumunija
RO31	Rumunija
RO32	Rumunija
RO41	Rumunija
RO42	Rumunija
SE11	Švedija
SE12	Švedija
SE21	Švedija
SE22	Švedija
SE23	Švedija
SE31	Švedija
SE32	Švedija
SE33	Švedija
SI03	Slovėnija
SI04	Slovėnija
SK01	Slovakija
SK02	Slovakija
SK03	Slovakija
SK04	Slovakija

Šaltinis: sudaryta darbo autoriaus remiantis [151]

18 priedas. Skaityto pranešimo konferencijoje pažymėjimas *„Didžiųjų duomenų kokybės gerinimas: atsitiktinio miško modelio taikymas trūkstamų reikšmių užpildymui“*

VILNIAUS UNIVERSITETO

ŠIAULIŲ AKADEMIJA

PAŽYMĖJIMAS

Nr. MVG-VUŠA-2025-325

(4.16 E) 850000-V-189

IRMANTAS PILYPAS

dalyvavo jaunųjų tyrėjų tarptautinėje mokslinėje konferencijoje
„JAUNASIS TYRĖJAS IŠMANIAJAI VISUOMENEI“

Ir skaitė pranešimą tema:

**„Didžiųjų duomenų kokybės gerinimas:
atsitiktinio miško modelio taikymas trūkstamų reikšmių užpildymui“**

Direktoriaus pavaduotoja
studijoms dr. Regina Karvelienė

2025 m. gegužės 08 d.

19 priedas. Skaityto pranešimo konferencijoje pažymėjimas *„Random Forest modelio taikymas užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 lygmeniu“*

Pažymėjimas

Irmantas Pilypas

Vilniaus universiteto Šiaulių akademija

dalyvavo VI-oje konferencijoje

Lietuvos magistrantų informatikos ir IT tyrimai

vykusioje 2025 m. gegužės 13 dieną
ir skaitė pranešimą

Random Forest modelio taikymas užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 lygmeniu

Programinis komitetas:

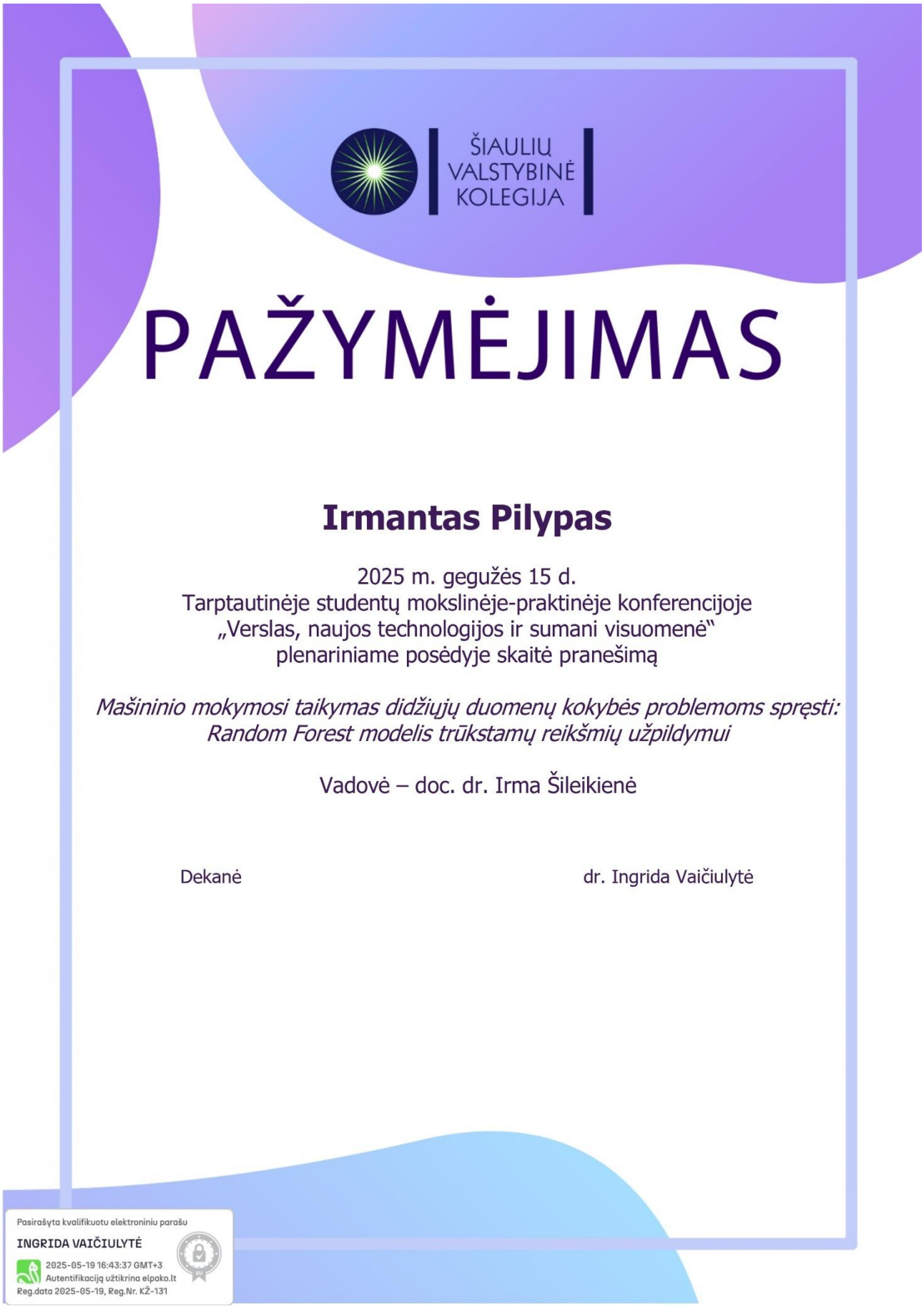
Prof. Gintautas Dzemyda

Prof. Olga Kurasova

Prof. Julius Žilinskas

Vilnius, 2025-05-13

20 priedas. Skaityto pranešimo konferencijoje pažymėjimas *„Applying machine learning to solve big data quality problems: random forest model for filling missing values“*



ŠIAULIŲ VALSTYBINĖ KOLEGIJA

PAŽYMĖJIMAS

Irmantas Pilypas

2025 m. gegužės 15 d.
Tarptautinėje studentų mokslinėje-praktinėje konferencijoje
„Verslas, naujos technologijos ir sumani visuomenė“
plenariniame posėdyje skaitė pranešimą

Mašininio mokymosi taikymas didžiųjų duomenų kokybės problemoms spręsti: Random Forest modelis trūkstamų reikšmių užpildymui

Vadovė – doc. dr. Irma Šileikienė

Dekanė dr. Ingrida Vaičiulytė

Pasirašyta kvalifikuotu elektroniniu parašu
INGRIDA VAIČIULYTĖ
2025-05-19 16:43:37 GMT+3
Autentifikaciją užtikrina elpako.lt
Reg.data 2025-05-19, Reg.Nr. KŽ-131

21 priedas. Prototipinės sistemos reliacinis duomenų bazės modelis