Atsitiktinio miško modelio taikymas užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 lygmeniu

Irmantas Pilypas, Irma Šileikienė

Vilniaus universiteto Šiaulių akademija, Regionų plėtros institutas
Vytauto g. 84, Šiauliai
irmantas.pilypas@sa.stud.vu.lt, irma.sileikiene@sa.vu.lt

Santrauka. Nagrinėjamos trūkstamų ekonominių rodiklių reikšmių ES NUTS 2 regionų lygmeniu užpildymo galimybės, taikant atsitiktinio miško (angl. *Random Forest*) modelį. Trūkstamos reikšmės ekonominiuose duomenyse kelia iššūkių analizei ir gali mažinti analizės rezultatų patikimumą. Sudarytas duomenų rinkinys, apimantis 260 regionų ir 8 ekonominius rodiklius (1990–2023), surinktas iš Eurostat DB. Kiekvienam rodikliui sukurtas atskiras atsitiktinio miško modelis trūkstamų reikšmių užpildymui. Modelių tikslumas vertintas pagal 4 metrikas: RMSE, MAE, R^2, MAPE. Tyrimas atskleidė, kad atsitiktinio miško metodas ypač tiksliai užpildo užimtumo, nedarbo ir ilgalaikio nedarbo rodiklių reikšmes ($R^2 > 0,9$), išlaikydamas tarpusavio ryšius tarp kintamųjų. Gautas užpildytas duomenų rinkinys gali būti taikomas tolesnėje regioninės ekonomikos rodiklių analizėje.

Raktiniai žodžiai: atsitiktinis miškas, mašininis mokymasis, trūkstamos reikšmės, ekonominiai rodikliai, NUTS 2 regionai, hiperparametrų optimizavimas.

1 Įvadas

Europos Sąjungos regioninės ekonominės analizės kokybė ir tikslumas dažnai priklauso nuo turimų statistinių duomenų išsamumo. NUTS 2 (angl. *Nomenclature of Territorial Units for Statistics*) regionų duomenų rinkiniuose neretai pasitaiko trūkstamų reikšmių, ypač mažiau išsivysčiusiuose regionuose ar tam tikrais laiko periodais. Tai sukelia analitikams ir tyrėjams reikšmingų iššūkių, nes trūkstami duomenys riboja galimybes atlikti patikimą ekonominių rodiklių analizę ir palyginimą [10].

Trūkstamų reikšmių užpildymas – tai metodas, skirtas trūkstamoms reikšmėms užpildyti statistiniais ar mašininio mokymosi metodais [13]. Tradiciniai užpildymo metodai, tokie kaip vidurkio ar medianos naudojimas, neišsaugo kintamųjų tarpusavio ryšių, todėl gali iškreipti duomenų struktū-

DOI: https://doi.org/10.15388/LMITT.2025.20

rą [6]. Pažangesni metodai, kaip MICE (angl. *Multiple Imputation by Chained Equations*) metodas, geriau išsaugo kintamųjų tarpusavio ryšius [25], tačiau jų taikymas didelės apimties daugiamačiams duomenims gali kelti skaičiavimo iššūkių [11].

Šiame tyrime analizuojamas atistiktinio miško metodo taikymas, užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 regionų lygmeniu. Atsitiktinio miško metodas pasižymi gebėjimu modeliuoti kompleksinius netiesinius ryšius tarp kintamųjų ir atsparumu išskirtims [4].

Tyrimo tikslas yra ištirti atsitiktinio miško metodo efektyvumą užpildant įvairių ekonominių rodiklių NUTS 2 lygmeniu trūkstamus duomenis.

Tikslui pasiekti iškelti tokie **uždaviniai**:

1. Išanalizuoti ankstesniuose tyrimuose naudotus trūkstamų reikšmių duomenų rinkiniuose užpildymo metodus ir tikslumo metrikas;
2. Sudaryti empirinio tyrimo metodiką trūkstamų reikšmių duomenų rinkiniuose užpildymui;
3. Įvertinti atlikto trūkstamų reikšmių duomenų rinkiniuose užpildymo gautus rezultatus, naudojant atsitiktinio miško bazinį metodą ir su optimizuotais hiperparametrais.

Šiame straipsnyje yra analizuojama ir išbandoma trūkstamų reikšmių užpildymo proceso metodologija, analizuojant įvairių ekonominių rodiklių užpildymo rezultatus ir formuluojant rekomendacijas būsimiems tyrimams šioje srityje.

Tyrimui atlikti naudoti mokslinės literatūros analizės ir empirinių tyrimų metodai.

2 Literatūros analizė

Trūkstamų reikšmių užpildymo metodai plačiai taikomi įvairiose srityse, įskaitant statistiką, ekonometriją, socialinių mokslų ir biomedicininius tyrimus. Trūkstamų reikšmių užpildymas gali būti skirstomas į vienmatį ir daugiamatį, priklausomai nuo to, ar atsižvelgiama į kintamųjų tarpusavio ryšius [9].

Tradiciniai užpildymo metodai, tokie kaip vidurkio, medianos ar modos naudojimas, yra paprasti, tačiau jie neatsižvelgia į kintamųjų tarpusavio ryšius ir gali iškraipyti duomenų struktūrą [12]. Statistiniai metodai, tokie kaip regresinis užpildymas ar k-artimiausių kaimynų (k-NN) metodas, geriau išsaugo kintamųjų tarpusavio ryšius, tačiau jie taip pat turi trūkumų, ypač dirbant su didelės apimties daugiamačiais duomenimis [1].

MICE (Multiple Imputation by Chained Equations) metodas tapo populiarus dėl savo universalumo ir gebėjimo išsaugoti kintamųjų tarpusavio ryšius [2]. MICE metodas skaičiuoja kiekvieno kintamojo prognostinį modelį, naudodamas kitus kintamuosius kaip prediktorius, ir iteratyviai atnaujina užpildytas reikšmes, kol pasiekiama konvergencija [27]. Tačiau MICE metodas gali būti, skaičiavimo prasme, imlus dideliems duomenų rinkiniams.

Mašininio mokymosi metodai, tokie kaip atsitiktinių miškų, neuroniniai tinklai ir gradientinis stiprinimas (gradient boosting), vis dažniau taikomi trūkstamų reikšmių užpildymui [23]. Tang ir Ishwaran [24] parodė, kad atsitiktinio miško metodas gali efektyviai užpildyti trūkstamas reikšmes, išsaugodamas kintamųjų tarpusavio ryšius. Shah ir kt. [21] nustatė, kad atsitiktinio miško metodas dažnai pranoksta tradicinius užpildymo metodus, ypač kai duomenys yra susiję netiesiškai.

Atsitiktinio miško metodas turi keletą privalumų trūkstamų reikšmių užpildymo kontekste: (1) jis gali modeliuoti kompleksinius netiesinius ryšius tarp kintamųjų; (2) jis yra atsparus išimtims ir triukšmui duomenyse; (3) jis gali apdoroti didelės apimties daugiamačius duomenis [26]. Tačiau atsitiktinio miško metodo efektyvumas priklauso nuo tinkamo hiperparametrų parinkimo, kuris gali būti sudėtingas uždavinys [19].

Hiperparametrų optimizavimas yra esminė atistiktinio miško metodo taikymo dalis. Dažniausiai optimizuojami hiperparametrai yra medžių skaičius, maksimalus medžio gylis, minimalus pavyzdžių skaičius šakos mazge ir minimalus pavyzdžių skaičius lape [3]. Probst ir kt. [18] nustatė, kad medžių skaičius ir maksimalus medžio gylis turi didžiausią įtaką atistiktinio miško metodo efektyvumui.

Europos Sąjungos NUTS 2 regionų lygmens kontekste, ekonominių rodiklių trūkstamų reikšmių užpildymo tyrimų, atlikta nedaug. García-Laencina ir kt. [8] taikė įvairius mašininio mokymosi metodus trūkstamų reikšmių užpildymui fundamentaliuose tyrimuose neprisirišant prie duomenų pobūdžio, todėl rezultatai gali būti pritaikomi ir ekonominių rodiklių analizei. Aljinbaz ir kt. [15] analizavo nedarbo lygio prognozavimą įvairiose šalyse, pasitelkiant duomenų struktūravimo metodus ir neuroninius tinklus. Atliekant literatūros analizę tyrimų, kuriuose analizuojami bendri, Europos sąjungos NUTS 2 regionų lygmenyje ekonominių rodiklių, trūkstamų reikšmių užpildymo metodai, nebuvo aptikta.

3 Duomenys ir metodologija

3.1 Duomenų aprašymas

Tyrime naudojamas NUTS 2 regionų ekonominių rodiklių duomenų rinkinys, apimantis 1990–2023 metų laikotarpį. Duomenys apima 260 unikalius NUTS 2 regionus (iš viso 6984 stebėjimai) ir 8 ekonominius rodiklius [22]:

1. **Employment_rate** - užimtumo lygis (%);
2. **Rd_expenditure_pct_gdp** - išlaidos moksliniams tyrimams ir technologinei plėtrai, procentais nuo BVP;
3. **Unemployment_rate** - nedarbo lygis (%);
4. **Tertiary_education_pct** - aukštąjį išsilavinimą turinčių gyventojų dalis (%);
5. **Youth_unemployment_rate** - jaunimo nedarbo lygis (%);
6. **Population_density** - gyventojų tankis (gyv./km^2);
7. **Long_term_unemployment_share** - ilgalaikio nedarbo dalis (%);
8. **Female_employment_rate** - moterų užimtumo lygis (%).

Tolesnėje analizėje, siekiant aiškesnio rodiklių identifikavimo, jie bus žymimi ROD1-ROD8 indeksais, atitinkančiais aukščiau pateiktą numeraciją.

Visi tyrimo duomenys buvo surinkti iš Europos Sąjungos statistikos tarnybos (Eurostat) oficialios duomenų bazės, užtikrinant patikimumą ir nuoseklumą tarp skirtingų regionų ir laikotarpių [22].

Duomenų rinkinyje yra reikšmingas trūkstamų reikšmių skaičius, kuris skiriasi tarp rodiklių. Trūkstamų reikšmių procentas svyruoja nuo 10 % (population_density) iki 53,8 % (rd_expenditure_pct_gdp), kaip parodyta 1 lentelėje.

1 lentelė. Trūkstamų reikšmių procentas pagal ekonominį rodiklį

Rodiklio ID	Rodiklis	Trūkstamos reikšmės (%)
ROD1	employment_rate	28,15
ROD2	rd_expenditure_pct_gdp	53,81
ROD3	unemployment_rate	29,04
ROD4	tertiary_education_pct	29,70
ROD5	youth_unemployment_rate	36,78
ROD6	population_density	10,01
ROD7	long_term_unemployment_share	35,09
ROD8	female_employment_rate	28,15

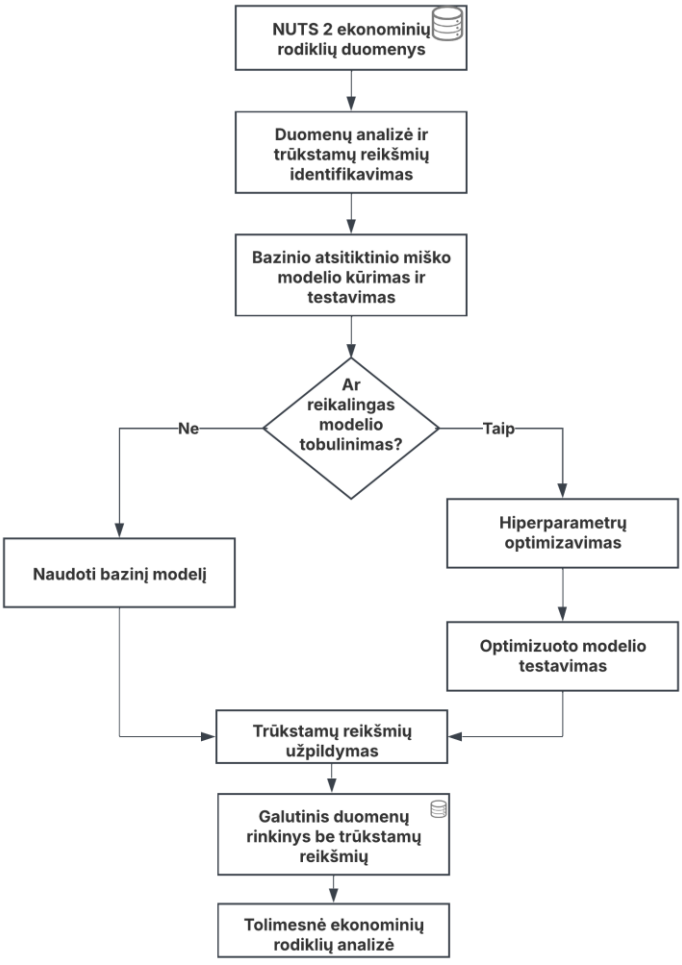
Tyrimo duomenų rinkinys pateiktas Github saugykloje: https://github.com/Pilypas/Random_Forest-nuts2-ekonominiu-rodikliu-uzpildymas

3.2 Metodologija

Tyrimo metodologija apima kelis esminius žingsnius:

1. **Duomenų analizė ir paruošimas** - atliekama pradinė duomenų analizė, identifikuojami kintamųjų pasiskirstymai, tarpusavio ryšiai bei trūkstamos reikšmės.
2. **Dirbtinių testavimo duomenų kūrimas** - sukuriami sintetiniai testavimo duomenys užpildymo metodų tikslumui įvertinti. 20 % netrūkstamų reikšmių atsitiktinai pašalinama iš duomenų rinkinio, o vėliau šios pašalintos reikšmės naudojamos užpildymo tikslumo vertinimui [14].
3. **Bazinio atsitiktinio miško modelio kūrimas** - sukuriamas bazinis atsitiktinio miško modelis naudojant Python scikit-learn bibliotekos numatytąsias (angl. *default*) hiperparametrų reikšmes: 100 medžių, maksimalus gylis neribojamas, minimalus pavyzdžių skaičius šakos mazge - 2, minimalus pavyzdžių skaičius lape - 1.
4. **Hiperparametrų optimizavimas** - bazinio modelio hiperparametrai optimizuojami naudojant *Grid Search* metodą su 5 sluoksnių kryžminiu patikrinimu. Paieškos aibę sudaro šie hiperparametrų rinkiniai [20]:
   - n_estimators (medžių skaičius): {50, 100, 200};
   - max_depth (maksimalus medžio gylis): {None, 10, 20, 30};
   - min_samples_split (minimalus pavyzdžių skaičius šakos mazge): {2, 5, 10};
   - min_samples_leaf (minimalus pavyzdžių skaičius lape): {1, 2, 4}.
5. **Modelių efektyvumo vertinimas -** bazinio ir optimizuotų modelių efektyvumas vertinamas pagal šias metrikas [17]:
   - RMSE (angl. *Root Mean Square Error*) - vidutinis kvadratinis nuokrypis;
   - MAE (angl. *Mean Absolute Error*) - vidutinė absoliutinė paklaida;
   - R^2 (angl. *Determination Coefficient*) - determinacijos koeficientas;
   - MAPE (angl. *Mean Absolute Percentage Error*) - vidutinė absoliutinė santykinė paklaida.
6. **Požymių svarbos analizė** - analizuojama požymių svarba kiekvieno ekonominio rodiklio prognozavimui, identifikuojant svarbiausius prediktorius.

1 pav. pavaizduotas nuoseklus trūkstamų reikšmių užpildymo procesas, kurio metu priklausomai nuo bazinio modelio tikslumo sprendžiama, ar reikalingas papildomas hiperparametrų optimizavimas, siekiant užtikrinti geriausią įmanomą kiekvieno ekonominio rodiklio trūkstamų reikšmių užpildymo tikslumą.



1 pav. Ekonominių rodiklių NUTS 2 regionų lygmeniu trūkstamų reikšmių užpildymo proceso blokinė schema

Tyrimas atliktas naudojant Python programavimo kalbą ir scikit-learn bibliotekos atsitiktinio miško regressor modelį trūkstamų reikšmių užpildymo algoritmo įgyvendinimui.

Atsitiktinio miško metodo apibendrintas algoritmas trūkstamų reikšmių užpildymui:

1. Kiekvienam ekonominiam rodikliui, turinčiam trūkstamų reikšmių, sukuriamas atskiras atsitiktinio miško modelis.
2. Šis modelis apmokomas naudojant stebėjimus, kuriuose tikslinė reikšmė yra žinoma.
3. Mokymo metu kiti ekonominiai rodikliai naudojami kaip prediktoriai (nepriklausomi kintamieji).
4. Apmokytasis modelis naudojamas prognozuoti trūkstamas reikšmes, remiantis kitų rodiklių vertėmis.

Pavyzdžiui, jei konkrečiame regione trūksta nedarbo lygio duomenų, modelis apsimoko, nustatydamas ryšius tarp nedarbo lygio ir kitų rodiklių (užimtumo lygio, išsilavinimo lygio, gyventojų tankio ir kt.) naudodamas kitų regionų turimus nedarbo lygio duomenis. Vėliau šis modelis naudojamas įvertinti trūkstamą reikšmę, remiantis kitais to regiono rodikliais.

4 Rezultatai

4.1 Bazinio atsitiktinio miško modelio rezultatai

Pirmiausia sukurtas bazinis atsitiktinio miško modelis ir įvertintas modelio efektyvumas. Bazinis modelis sukurtas su standartiniais hiperparametrais, naudojant 100 medžių. 2 lentelėje pateiktos bazinio atsitiktinio miško modelio tikslumo vertinimo metrikos kiekvienam ekonominiam rodikliui.

2 lentelė. Bazinio atsitiktinio miško modelio metrikos

Rodiklio ID	MSE	RMSE	MAE	R^2	MAPE (%)
ROD1	1,9420	1,3936	1,0516	0,9739	1,5586
ROD2	0,4557	0,6751	0,4096	0,7180	-
ROD3	2,2041	1,4846	1,0418	0,9344	13,2975
ROD4	23,9705	4,8960	3,5438	0,7819	15,0099
ROD5	20,1910	4,4934	3,1664	0,8828	17,6417
ROD6	369181,5051	607,6031	273,5064	0,4586	323,1901
ROD7	1,8774	1,3702	0,8884	0,8862	26,3958
ROD8	4,9056	2,2148	1,5780	0,9668	2,8183

Rezultatai rodo, kad bazinis atsitiktinio miško modelis pasiekia gana aukštą tikslumą daugumai rodiklių. Ypač aukšti R^2 rodikliai pasiekiami (ROD1) užimtumo lygio (0,9739), (ROD3) nedarbo lygio (0,9344) ir (ROD8)

moterų užimtumo lygio (0,9668) rodikliams, o tai rodo stiprų sąryšį tarp rodiklių. Tačiau (ROD6) gyventojų tankio (population_density) rodiklio užpildymo tikslumas yra žymiai mažesnis (R^2 = 0,4586), o MAPE vertė labai aukšta (323,19 %), kas rodo, kad šio rodiklio užpildymas yra labai netikslus.

4.2 Hiperparametrų optimizavimo rezultatai

Siekiant padidinti užpildymo tikslumą, atliktas atsitiktinio miško modelio hiperparametrų optimizavimas naudojant Grid Search metodą. 3 lentelėje pateikiami gauti geriausi hiperparametrų rinkiniai kiekvienam ekonominiam rodikliui.

3 lentelė. Optimalūs hiperparametrai kiekvienam rodikliui

Rodiklio ID	max_depth	min_samples_leaf	min_samples_split	n_estimators
ROD1	10	1	2	100
ROD2	None	4	2	200
ROD3	10	4	10	100
ROD4	None	4	2	200
ROD5	10	4	2	200
ROD6	10	2	2	100
ROD7	20	4	10	100
ROD8	20	2	10	200

Optimizuoti hiperparametrai skiriasi lyginant skirtingus rodiklius, kas rodo, kad skirtingiems ekonominiams rodikliams reikalingi skirtingo sudėtingumo modeliai. 4 lentelėje pateiktos optimizuotų atsitiktinio miško modelių tikslumo vertinimo metrikos ir kryžminio patikrinimo rezultatai.

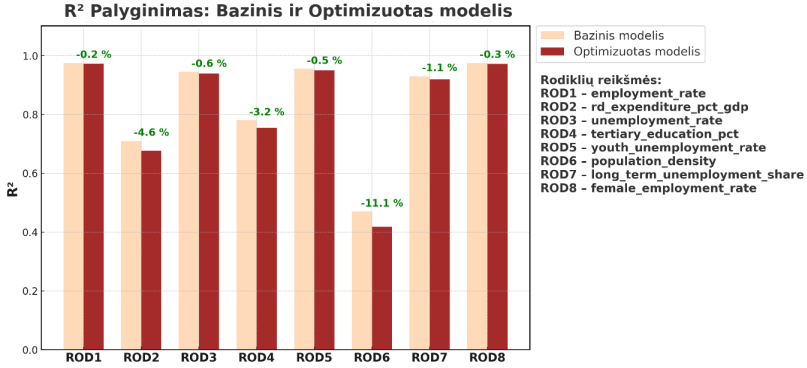
Pažymėtina, kad optimizuoti modeliai ne visada pagerina užpildymo tikslumą, lyginant su baziniu modeliu. Kai kuriais atvejais, pavyzdžiui, (ROD1) užimtumo lygio (employment_rate) ir (ROD3) nedarbo lygio (unemployment_rate) rodikliams, bazinis modelis pasiekia šiek tiek geresnį tikslumą. Tai gali būti susiję su persimokymo (overfitting) problemomis, kai optimizuoti modeliai pernelyg prisitaiko prie mokymo duomenų ir prasčiau apibendrina naujus duomenis [7].

4 lentelė. Optimizuotų atsitiktinio miško modelių tikslumo vertinimo metrikos ir kryžminio patikrinimo rezultatai

	Optimizuotų rodiklių tikslumo metrikos					Kryžminio patikrinimo rezultatai		
Rodiklio ID	**MSE**	**RMSE**	**MAE**	**R^2**	**MAPE (%)**	**RMSE (vid±std)**	**MAE (vid±std)**	**R^2 (vid±std)**
ROD1	2,0841	1,4436	1,0967	0,9720	1,6261	2,395±0,8633	1,8377±0,5638	0,9168±0,0406
ROD2	0,5094	0,7137	0,4339	0,6848	-	1,0577±0,1715	0,7104±0,1073	0,0673±0,1068
ROD3	2,3870	1,5450	1,0976	0,9290	14,2330	1,9987±0,3797	1,4957±0,2731	0,8472±0,0635
ROD4	26,6963	5,1668	3,7738	0,7571	15,9998	8,4068±1,1812	6,8070±0,9329	0,2749±0,1369
ROD5	20,9363	4,5756	3,2666	0,8785	18,4185	5,9680±1,0842	4,4952±0,8218	0,7347±0,0955
ROD6	403778,6331	635,4358	287,2524	0,4078	364,9937	876,4285±333,1299	438,1537±112,5270	-0,5673 ±0,5968
ROD7	2,0408	1,4286	0,9218	0,8763	27,3312	2,0400±0,4970	1,3785±0,3726	0,7495±0,1054
ROD8	5,2680	2,2952	1,6518	0,9643	2,9731	3,8652±1,1000	2,8837±0,7635	0,8824±0,0330

4.3 Modelių efektyvumo vertinimas

Įvertinus bazinio ir optimizuotų atsitiktinio miško modelių tikslumą (2 pav.), svarbu palyginti jų efektyvumą įvairiems ekonominiams rodikliams. Šis palyginimas padeda nustatyti, ar hiperparametrų optimizavimas iš tiesų pagerino trūkstamų reikšmių užpildymo tikslumą, ir kokius modelius geriausia naudoti galutiniam duomenų užpildymui.



2 pav. R² palyginimas - bazinio ir optimizuotų modelių

2 paveikslėlyje galima pastebėti, kad hiperparametrų optimizavimas daugeliu atvejų neduoda reikšmingo tikslumo pagerėjimo, o kai kuriais atvejais netgi šiek tiek pablogina rezultatus. Tai gali būti susiję su tuo, kad bazinis atsitiktinio miško modelis jau buvo pakankamai gerai pritaikytas duomenims, o optimizavimo metu galėjo įvykti pernelyg didelis prisitaikymas prie mokymo duomenų. Remiantis kryžminio patikrinimo rezultatais (žr. 4 lentelę), kur matomas žymus skirtumas tarp testavimo metrikų ir standartinių nuokrypių, ypač (ROD6) gyventojų tankio (population_density) rodikliui, kurio R^2 vidurkis yra -0,5673 su dideliu standartiniu nuokrypiu (±0,5968). Tokie dideli standartiniai nuokrypiai ir neigiamos R^2 vertės rodo, kad optimizuoti modeliai skirtingose kryžminio patikrinimo sluoksniuose veikia pakankamai netolygiai, kas yra būdinga persimokymo atvejams.

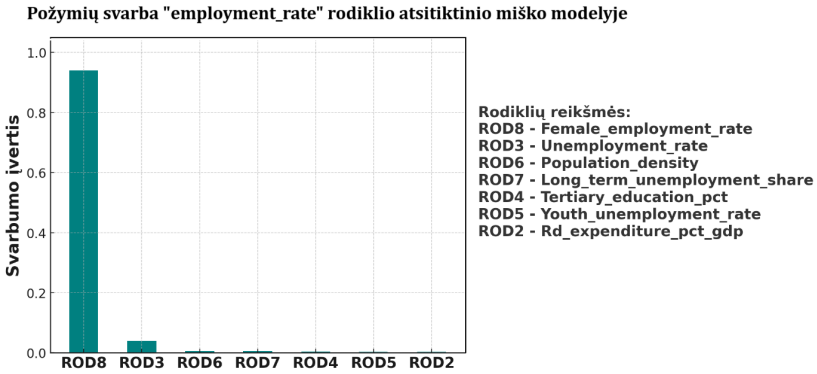
Lyginant su kitais literatūroje aprašytais metodais, tyrime analizuotas atsitiktinio miško metodas pasiekia panašius ar geresnius rezultatus. Pavyzdžiui, Aljinbaz ir kt. [15] taikydami neuroninius tinklus nedarbo lygio prognozavimui įvairiose šalyse pasiekė (R^2 apie 0,87), o šiame tyrime analizuotas metodas pasiekia (R^2 apie 0,93) nedarbo lygio rodikliui. Panat ir Chandra

[16] savo tyrime apie Globalų Gyvenimo Kokybės Indeksą pasiekė pakankamai gerus tikslingumo rezultatus ekonominiams rodikliams (R^2 daugiau nei 0,7), tačiau nedarbo lygio rodiklio užpildymas buvo mažiau tikslus (R^2 apie -0,25), kai tuo tarpu šiame tyrime, šiam rodikliui pasiektas gerokai aukštesnis tikslumas. Tačiau, kaip ir minėtuose tyrimuose, šiame darbe taip pat buvo pastebėta, kad kai kurių rodiklių (ypač gyventojų tankio) prognozavimas yra žymiai sudėtingesnis.

Šio tyrimo rezultatus tiesiogiai lyginti su kituose tyrimuose gautais rezultatais negalima, nes naudoti skirtingi duomenų rinkiniai, tačiau šio tyrimo metu gauti atsitiktinio miško modelio tuščių reikšmių užpildymo rezultatai yra panašiame tikslumo lygmenyje.

4.4 Požymių svarbos analizė

Analizuojant požymių svarbą, galima identifikuoti, kurie ekonominiai rodikliai yra svarbiausi prognozuojant kitus rodiklius. Požymių svarbos vizualizacija užimtumo lygio (employment_rate) rodikliui pateikiama 3 pav.



3 pav. Požymių svarba užimtumo lygio (employment_rate) rodikliui

Požymių svarbos analizė rodo, kad (ROD8) moterų užimtumo lygis (female_employment_rate) yra svarbiausias požymis prognozuojant bendrą užimtumo lygį, kas nėra netikėta, atsižvelgiant į tai, kad moterų užimtumo lygis sudaro reikšmingą dalį bendro užimtumo lygio.

5 Išvados

Atlikto tyrimo rezultatai rodo, kad atsitiktinio miško metodas gali būti efektyviai naudojamas užpildant trūkstamas ekonominių rodiklių NUTS 2 regionų lygmeniu reikšmes. Tačiau metodų efektyvumas reikšmingai skiriasi tarp skirtingų ekonominių rodiklių, o tai rodo, kad kiekvieno rodiklio trūkstamų reikšmių užpildymo uždavinys turėtų būti vertinamas atskirai.

Geriausiai atsitiktinio miško metodas veikia užpildant užimtumo rodiklius (bendrą užimtumo lygį ir moterų užimtumo lygį), nedarbo lygį ir ilgalaikio nedarbo dalį. Šiems rodikliams pasiekiami aukšti determinacijos koeficientai R^2 didesnis nei 0,8, o tai rodo, kad atsitiktinio miško modeliai gali paaiškinti didžiąją dalį šių rodiklių variacijos. Tai gali būti susiję su tuo, kad šie rodikliai turi stiprius tarpusavio ryšius ir taip pat yra susiję su kitais socialiniais-ekonominiais rodikliais, tokiais kaip išsilavinimo lygis.

Gyventojų tankio ir išlaidų moksliniams tyrimams bei technologinei plėtrai rodiklių užpildymas yra mažiau tikslus. Gyventojų tankio atveju tai gali būti susiję su tuo, kad šis rodiklis labiau priklauso nuo geografinių, istorinių ir urbanistinių veiksnių, kurie nėra tiesiogiai susiję su kitais ekonominiais rodikliais duomenų rinkinyje [5]. Išlaidų moksliniams tyrimams ir technologinei plėtrai atveju, mažą tikslumą gali lemti tai, kad šis rodiklis labai priklauso nuo konkrečios valstybės mokslo ir inovacijų politikos, taip pat nuo regiono specializacijos ir pramonės struktūros, kurios nėra niekaip susietos su kitais duomenų rinkinio rodikliais.

Pastebėtina, kad hiperparametrų optimizavimas nedavė reikšmingų tikslumo pagerėjimų, o kai kuriais atvejais netgi šiek tiek pablogino rezultatus. Tai gali būti susiję su persimokymo problema, kai modeliai pernelyg prisitaiko prie mokymo duomenų ir prasčiau apibendrina naujus duomenis arba sąryšiai su kitais ekonominiais rodikliais iš duomenų rinkinio per silpni. Šie rezultatai primena paprastų modelių (parsimony) principo svarbą: sudėtingesni modeliai ne visada yra geresni, ypač kai turimų duomenų kiekis yra ribotas [28].

Požymių svarbos analizė išryškino ekonominių rodiklių tarpusavio ryšius. Pvz., moterų užimtumo lygis yra svarbiausias požymis prognozuojant bendrą užimtumo lygį, o nedarbo lygis ir jaunimo nedarbo lygis taip pat yra svarbūs. Tai atitinka ekonominę intuiciją ir patvirtina, kad atsitiktinio miško modeliai sugeba surasti ir panaudoti prasmingus ryšius tarp ekonominių rodiklių.

Remiantis atlikta analize, ekonominių rodiklių NUTS2 regionų lygmeniu trūkstamų reikšmių užpildymui, pasirinktas bazinis atsitiktinio miško mo-

delis, kadangi gautas pakankamas tikslumas daugumai rodiklių. Užpildžius visas trūkstamas reikšmes, sukurtas pilnas duomenų rinkinys, kuris gali būti panaudojamas tolimesnei ekonominei analizei tikintis tikslesnių ekonominės analizės ir prognozavimo rezultatų, naudojant kitus mašininio mokymosi algoritmus.

Tolimesniuose tyrimuose planuojama ištirti kitus mašininio mokymosi metodus, tokius kaip neuroniniai tinklai ar gradient boosting, trūkstamų ekonominių rodiklių reikšmių užpildymui. Taip pat būtų vertinga ištirti užpildymo metodus atsižvelgiant į laiko dimensiją, t.y. naudojant laiko eilučių metodus. Galiausiai, būtų naudinga ištirti, kaip trūkstamų reikšmių užpildymo metodai padeda užtikrinti tikslesnius ekonominės analizės rezultatus.

Planuojama atlikti ekonominių rodiklių analizę su pirminiu ir papildytu duomenų rinkiniu, naudojant kitus mašininio mokymosi algoritmus, siekiant nustatyti su kuriais duomenimis gaunami tikslesni ekonominės analizės rezultatai.

Literatūra

[1] Andridge, R. R., & Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. International statistical review = Revue internationale de statistique, 78(1), 40. https://doi.org/10.1111/J.1751-5823.2010.00103.X

[2] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research, 20(1), 40. https://doi.org/10.1002/MPR.329

[3] Bergstra, J., Ca, J. B., & Ca, Y. B. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. Journal of Machine Learning Research, 13, 281–305. http://scikit-learn.sourceforge.net.

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

[5] Cui, C., Hu, Y., Bao, Y., & Li, H. (2024). Population Density Prediction at Township Scale Supported by Machine Learning Method: A Case Study in Inner Mongolia. ISPRS International Journal of Geo-Information, 13(12). https://doi.org/10.3390/IJGI13120426

[6] Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: a gentle introduction to imputation of missing values. Journal of clinical epidemiology, 59(10), 1087–1091. https://doi.org/10.1016/J.JCLINEPI.2006.01.014

[7] Fatima, S., Hussain, A., Amir, S. Bin, Ahmed, S. H., & Aslam, S. M. H. (2023). XGBoost and Random Forest Algorithms: An in Depth Analysis. Pakistan Journal of Scientific Research, 3(1), 26–31. https://doi.org/10.57041/PJOSR.V3I1.946

[8] García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. Neural Computing and Applications, 19(2), 263–282. https://doi.org/10.1007/S00521-009-0295-6

[9] Graham, J. W. (2008). Missing Data Analysis: Making It Work in the Real World. https://doi.org/10.1146/annurev.psych.58.110405.085530

[10] Guide to statistics in European Commission development co-operation 2017 edition. (s.a.). https://doi.org/10.2785/30851
[11] Ieva Ivanauskienė. (2022). ES šalių konkurencingumo vertinimas ir klasterizavimas Baigiamasis magistro studijų projektas. Kauno technologijos universitetas.
[12] Jönsson, P., & Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using Ilkert data. Proceedings - International Software Metrics Symposium, 108–118. https://doi.org/10.1109/METRIC.2004.1357895
[13] Little, R. J. A., & Rubin, D. B. (2014). Statistical analysis with missing data. Statistical Analysis with Missing Data, 1–381. https://doi.org/10.1002/9781119013563
[14] Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2023). Machine Learning for Synthetic Data Generation: A Review. https://arxiv.org/abs/2302.04062v9
[15] Monir Aljinbaz, A. M., Mahmoud, M., & Rahhal, A. (2024). Forecasting Unemployment Rate for Multiple Countries Using a New Method for Data Structuring. IJACSA) International Journal of Advanced Computer Science and Applications, 15(12). www.ijacsa.thesai.org
[16] Panat, T., & Chandra, R. (s.a.). Global Ease of Living Index: a machine learning framework for longitudinal analysis of major economies.
[17] (PDF) A Consolidated Overview of Evaluation and Performance Metrics for Machine Learning and Computer Vision. (s.a.). Gauta 2025 m. balandžio 1 d., https://www.researchgate.net/publication/374558675_A_Consolidated_Overview_of_Evaluation_and_Performance_Metrics_for_Machine_Learning_and_Computer_Vision
[18] Probst, P., Bischl, B., & Boulesteix, A.-L. (2018). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. http://arxiv.org/abs/1802.09596
[19] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301. https://doi.org/10.1002/WIDM.1301
[20] RAMADHAN, M. M., SITANGGANG, I. S., NASUTION, F. R., & GHIFARI, A. (2017). Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency. DEStech Transactions on Computer Science and Engineering, cece. https://doi.org/10.12783/DTCSE/CECE2017/14611
[21] Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. American Journal of Epidemiology, 179(6), 764–774. https://doi.org/10.1093/AJE/KWT312
[22] Statistics | Eurostat. (s.a.). Gauta 2025 m. balandžio 1 d., https://ec.europa.eu/eurostat/databrowser/view/TGS00042/default/table
[23] Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118. https://doi.org/10.1093/BIOINFORMATICS/BTR597
[24] Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. https://doi.org/10.1002/sam.11348
[25] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/JSS.V045.I03
[26] Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. R. (2013). Comparison of imputation methods for missing laboratory data in medicine. BMJ Open, 3(8), e002847. https://doi.org/10.1136/BMJOPEN-2013-002847

[27] White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30(4), 377–399. https://doi.org/10.1002/SIM.4067

[28] Zhang, X., Chen, S., Xue, J., Wang, N., Xiao, Y., Chen, Q., Hong, Y., Zhou, Y., Teng, H., Hu, B., Zhuo, Z., Ji, W., Huang, Y., Gou, Y., Richer-de-Forges, A. C., Arrouays, D., & Shi, Z. (2023). Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping. Geoderma, 432, 116383. https://doi.org/10.1016/J.GEODERMA.2023.116383