Samdomojo darbo pajamų praleistų reikšmių įrašymas taikant daugiareikšmį įrašymą

Guoda Puslytė, Rūta Levulienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Taikomosios matematikos institutas, Naugarduko g. 24, Vilnius
guoda.puslyte@mif.stud.vu.lt

Santrauka. Analizuojant apklausų duomenis, dažnai tenka spręsti neužpildytų reikšmių problemą. Įprastai respondentai vengia atsakyti į klausimus, kurie yra susiję su asmenine ar finansine informaciją. Šiame darbe palyginami keturi daugiareikšmio įrašymo metodai. Rezultatai atskleidžia, kad atsitiktinių miškų bei klasifikavimo ir regresijos medžių metodai užtikrina mažiausią paklaidą.

Raktiniai žodžiai: Praleistos reikšmės, daugiareikšmis įrašymas, samdomojo darbo pajamos.

1 Įvadas

Praleistos reikšmės yra dažna problema, su kuria susiduriama analizuojant apklausų metu surinktus duomenis. Dėl įvairių priežasčių respondentai gali neatsakyti į tam tikrus klausimus arba pateikti netikslią informaciją, ypač kai klausimai susiję su jautriomis temomis, pavyzdžiui, pajamomis. Tiksli informacija apie gyventojų pajamas yra svarbi norint įvertinti skurdo lygį šalyje, ekonominius skirtumus tarp valstybių bei valstybės institucijoms priimant politinius sprendimus.

Pastaruoju metu, siekiant tiksliau užpildyti praleistas reikšmes, vis plačiau taikomi mašininio mokymosi metodai, tačiau dauguma tyrimų ir metodų apsiriboja vieno tipo kintamaisiais, dažniausiai – kiekybiniais [1, 2]. Vis dėlto, apklausų duomenyse vyrauja kategoriniai kintamieji. Vienas iš nedaugelio metodų, kuris yra tinkamas tiek kategorinių, tiek kiekybinių kintamųjų reikšmių įrašymui yra daugiareikšmis įrašymas [3, 4].

Šio tyrimo tikslas – įvertinti, kaip praleistų reikšmių užpildymas taikant daugiareikšmį įrašymą veikia samdomojo darbo pajamų įrašymo tikslumą bei parinkti tinkamiausią įrašymo metodą.

DOI: https://doi.org/10.15388/LMITT.2025.21

2 Duomenys

Šiame tyrime naudojami viešai prieinami duomenys iš 2023 m. metinio pajamų ir gyvenimo sąlygų statistinio tyrimo [5]. Analizei buvo atrinkti 28 kintamieji, suteikiantys informaciją apie respondento lytį, amžių, gyvenamąją vietą, šeimyninę padėtį, išsilavinimą, sveikatos būklę, užimtumą, darbo patirtį ir laisvalaikio įpročius, iš kurių 22 – kategoriniai ir 6 – kiekybiniai. Iš viso 16 kintamųjų turėjo praleistų reikšmių, kurių dalis sudarė nuo 0,04 iki 27 procentų. Pašalinus nežinomas reikšmes, galutinėje duomenų aibėje liko 2378 stebėjimai.

3 Metodologija

Pagrindinis tikslas yra palyginti daugiareikšmio įrašymo metodus ir pateikti rekomendacijas, todėl buvo paimtas pilnas, be praleistų reikšmių, duomenų rinkinys, pašalinta dalis stebėtų reikšmių, atliktas įrašymas ir palyginti rezultatai.

Kadangi didžioji dalis tiriamojo duomenų rinkinio kintamųjų yra kategoriniai, nebuvo galimybės taikyti „Little's MCAR" testo, skirto nustatyti, ar nežinomos reikšmės yra praleistos visiškai atsitiktinai (angl. Missing Completely At Random), todėl šiame tyrime daroma prielaida, kad nežinomos reikšmės yra praleistos atsitiktinai (angl. Missing At Random), t. y., praleistos reikšmės priklauso nuo kitų kintamųjų stebėtų reikšmių ir nepriklauso nuo nežinomų reikšmių [6, 7].

Galutinei duomenų aibei buvo sugeneruotos sintetinės nežinomos reikšmės taip, kad jų pasiskirstymas kuo tiksliau atspindėtų originalių duomenų nežinomų reikšmių pasiskirstymą bei nežinomų reikšmių dalis kiekvienam kintamajam išliktų tokia pati. Dirbtinėms praleistoms reikšmėms sugeneruoti buvo naudojama „R" paketo „missMethods" funkcija *delete_MAR_1_to_x()*. Ši funkcija padalina duomenų rinkinį į dvi grupes pagal parametre *cols_ctrl* nurodyto kintamojo reikšmių medianą ir sugeneruoja nežinomas reikšmes grupėse parametre *cols_mis* nurodytiems kintamiesiems santykiu 1:x (šiuo atveju $x = 5$). Be to, parametre *p* nurodoma, kokią duomenų dalį sudarys praleistos reikšmės kiekviename iš pasirinktų kintamųjų [7]. Geriausiai pavyko atkurti nežinomų reikšmių kombinacijas, kurios originalių duomenų rinkinyje pasikartoja dažniausiai, iš viso buvo atkurta apie 65% kombinacijų. Siekiant įvertinti įrašytų reikšmių poveikį samdomojo darbo pajamų įrašymo tikslumui, buvo sukurta antra duomenų aibė, kurioje nežinomos reikšmės buvo sugeneruotos tik samdomojo darbo pajamų kintamajam.

Daugiareikšmis įrašymas atliktas naudojant „R“ paketo „mice“ [4] funkciją *mice()* su parametru *m = 5*, t. y., kiekviena praleista reikšmė yra įrašoma penkis kartus, o galutinei reikšmei apskaičiuoti naudotas įrašytų reikšmių vidurkis. Praleistų reikšmių įrašymas pradedamas nuo mažiausiai praleistų reikšmių turinčio kintamojo. Daugiareikšmis įrašymas buvo atliekamas keturiais metodais, kurie yra tinkami tiek kiekybiniams, tiek kategoriniams kintamiesiems.

Vienas iš plačiausiai taikomų daugiareikšmio įrašymo metodų yra prognozuojamo vidurkio atitikimo metodas (PVA, angl. predictive mean matching). Šio metodo esmė – pasitelkiant regresijos modelį, sudarytą remiantis pilnais duomenimis, apskaičiuojama prognozė kiekvienam nežinomą reikšmę turinčiam stebėjimui. Tuomet kiekvienai prognozuotai reikšmei randamos kelios, šiuo atveju – 5, artimiausios stebėtos reikšmės, iš kurių atsitiktinai parenkama viena ir priskiriama praleistą reikšmę turėjusiam įrašui.

Atsitiktinės imties iš stebėtų reikšmių (AISR, angl. random sample from observed value) atvejų praleistų stebėjimų įrašymui yra naudojama stebėtų to paties kintamojo reikšmių aibė, iš kurios atsitiktinai parenkama imtis ir įrašoma vietoje praleistų reikšmių.

Klasifikavimo ir regresijos medžių (KRM, angl. classification and regression trees) metodas veikia panašiai kaip PVA tačiau vietoje regresijos modelio yra sudaromas sprendimų medis. Praleistą reikšmę turintis stebėjimas yra priskiriamas vienam iš sprendimo medžio lapų (angl. terminal node), tuomet iš lapui priklausančių stebėtų reikšmių atsitiktinai išrenkama viena ir įrašoma.

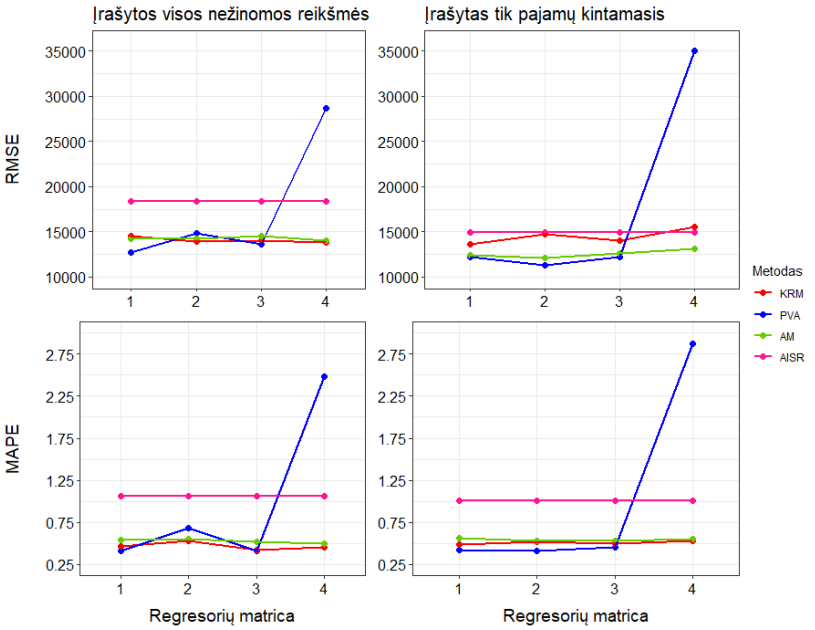
Atsitiktinių miškų (AM, angl. random forest) metodas praleistų reikšmių įrašymui taiko atsitiktinių miškų modelį, sudaryta iš 10 sprendimo medžių. Praleistų reikšmių įrašymui kiekybiniams kintamiesiems naudojamas sprendimo medžių prognozių vidurkis, o kategoriniams kintamiesiems – dažniausiai pasikartojanti prognozė.

Be to, siekiant įvertinti metodų jautrumą regresorių parinkimui, kiekvienam metodui buvo naudojamos keturios skirtingos regresorių matricos. Trys iš jų buvo sukurtos naudojant „mice“ paketo funkciją quickpred(), kuri leidžia greitai atrinkti regresorius kintamiesiems su praleistomis reikšmėmis, atsižvelgiant į kintamųjų tarpusavio koreliaciją. Šios regresorių matricos buvo sudarytos remiantis Spirmeno, Kendalo ir Pirsono koreliacijos koeficientais. Nors šis regresorių parinkimo metodas nėra pritaikytas kategoriniams kintamiesiems, šiame tyrime jis buvo naudojamas įvertinti daugiareikšmio įrašymo metodų jautrumą skirtingoms regresorių kombinacijoms, o ne parinkti optimalius regresorius kiekvienam kintamajam. Ketvirtojoje regresorių matricoje kiekvienam kintamajam kaip regresoriai buvo priskirti visi likę kintamieji.

Rezultatų palyginimui naudojamos trys metrikos: šaknis iš vidutinės kvadratinės paklaidos (RMSE) ir vidutinė absoliutinė procentinė paklaida (MAPE).

4 Rezultatai

Rezultatai pateikti 1 pav. Abiejų duomenų aibių atveju gauti panašūs rezultatai, kas rodo, jog praleistų reikšmių įrašymas neblogina samdomojo darbo pajamų įrašymo tikslumo. Prasčiausi rezultatai gauti taikant PVA metodą kartu su ketvirtąja regresorių matrica, kur kiekvieno kintamojo praliestoms reikšmėms įrašyti kaip regresoriai yra naudojami visi likę kintamieji, kas leidžia teigti, jog naudojant šį metodą svarbu tinkamai atrinkti regresorius kiekvienam praleistų reikšmių turinčiam kintamajam. Stabiliausi ir patikimiausi rezultatai gauti taikant AM ir KRM metodus. AISR atveju, visiems regresorių matricos atvejams gauta vienoda paklaida, taip yra todėl, nes šiuo metodu įrašymo reikšmės nepriklauso nuo kitų kintamųjų reikšmių ir priklauso tik nuo pačio įrašomo kintamojo nepraleistų reikšmių aibės, vis dėlto šio metodo atveju įrašomos reikšmės yra parenkamos atsitiktinai, todėl įrašytų reikšmių paklaida yra didelė.



1 pav. RMSE ir MAPE rodiklių palyginimas skirtingiems praleistų reikšmių įrašymo metodams.

5 Išvados

Atsižvelgiant į gautus rezultatus, galima teigti, kad praleistų reikšmių užpildymas taikant daugiareikšmį įrašymą tinkamas samdomojo darbo pajamų įrašymui. Geriausi rezultatai gauti taikant AM ir KRM metodus, kadangi įrašytų reikšmių paklaida nepriklausomai nuo regresorių matricos ir duomenų aibės išliko panaši visais atvejais. Panaši paklaida buvo gauta ir PVA atveju, tačiau taikant šį metodą svarbu tinkamai parinkti regresorius kiekvienam praleistų reikšmių turinčiam kintamajam.

Literatūra

[1] Md. Kamrul Hasan, Md. Ashraful Alam, Shidhartho Roy, Aishwariya Dutta, Md. Tasnim Jawad, Sunanda Das. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). Informatics in Medicine Unlocked, Volume 27, 100799.

[2] Lin, WC., Tsai, CF. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 53, 1487–1509.

[3] Van Buuren, S., & Oudshoorn, K. (1999). Flexible multivariate imputation by MICE. (pp. 1-20). Leiden: TNO.

[4] van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition (2nd ed.). Chapman and Hall/CRC.

[5] Valstybės duomenų agentūra. 2023 m. atlikto metinio pajamų ir gyvenimo sąlygų statistinio tyrimo asmenų duomenys.

[6] Roderick J. A. Little. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association, 83(404), 1198–1202.

[7] Santos, M. S., Pereira, R. C., Costa, A. F., Soares, J. P., Santos, J., & Abreu, P. H. (2019). Generating Synthetic Missing Data: A Review by Missing Mechanism. IEEE Access, 7, 11651-11667